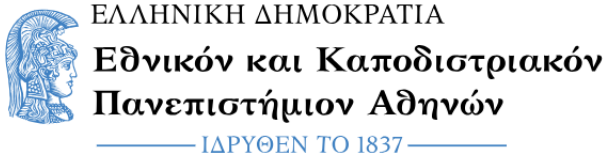
National and Kapodistrian University of Athens



Joint Models for Longitudinal and Survival data

Author:

Nteits Nikolaos

March 10, 2021

To my family and friends

Acknowlegdements

I would like to express my deepest appreciation and sincere gratitude to the leading supervisor professor Siannis Fotios for his trust and useful guidance without which this dissertation would not have been realized. I would also like to express my thanks and gratitude to all my fellow colleagues with whom I collaborated during my latest academic endeavor. Finally, my deepest love and appreciation is extended to my family and friends for their unending support and confidence in me which helped me achieve my goals.

Abstract

Longitudinal data and survival data frequently arise together in practice and are associated in many ways. A common objective in longitudinal studies is to characterize the relationship between a longitudinal response process and a time-to-event since separate analyses of longitudinal data and survival data may lead to inefficient or biased results. Joint models for longitudinal and survival data aim to incorporate all information simultaneously and provide valid and efficient inferences by connecting one or more longitudinal trajectories to the risk for an event. This class of models opens the door of personalized inference in modern biostatistics, something invaluable in the pursuit of personalized medicine. Personalized medicine allows us to tailor a drug or medication specificaly for the needs of the patient on their predicted response or risk of disease or an event. What follows is a presentation of the standard Joint model for Longitudinal and Survival data, which is the building block of this rich class of models, along with extensions that prove usefull in certain situations. Diagnostics are thoroughly presented as the procendure deviates from the standard diagnostics procendure in statistics. As a final step predictions are being discussed along with their assessment tool. All required knowledge is presented in the first two chapters.

Contents

Chapter 1

Analysis of Longitudinal Data

In longitudinal studies we perform multiple measurements on the same individuals aiming to study within-individual change through time, as well as the factors that influence this change. In longitudinal studies measurements of the same subject are correlated and this correlation must be taken into account.

There are two broad classes of models for longitudinal data proposed in the literature: *Linear Mixed Effects Models* and *Marginal Models*, with the former focusing on individualised inference and the later on population wide inference.
As Joint Modeling for Longitudinal and Survival data focuses on individualised inference *Linear Mixed Effects Models* are used for the longitudinal submodel .

1.1 Formulation of a Mixed Effects Model

The motivation behind a Linear Mixed Effects Model in the context of biostatistics is the construction of a model that not only does the job of a regression model but distinguishes between patients as well. The defining feature of this class of mod-

els is that some of the regression parameters can vary randomly from one individual to another, thereby accounting for sources of natural heterogeneity in the population. As a result individuals in the population are assumed to have their own subject specific mean response trajectories over time, something more natural when it comes modeling participants in a clinical trial. This is achieved through the introduction of random effects in the Standard Regression model which account for the uniqueness of each subject. In Linear Mixed Effects models the mean response is modeled as a combination of population characteristics, β (fixed effects) that are assumed to be shared by all individuals and the subject specific effects that are unique to a particular individual (random effects).
The model is formulated as follows :

$$y_i = X_i\beta + Z_i b_i + e_i$$

$$e_i \sim \mathcal{N}(0, \sigma^2 I_n)$$

In principle any multivariate distribution can be assumed for b_i but in practice , b_i is assumed to have multivariate normal distribution. Where :
n_i: The number of longitudinal measurements for patient i.
y_i: The $n_i \times 1$ longitudinal responce vector for patient i.
X_i: The $n_i \times p$ design matrix for the fixed effects for patient i.
Z_i: The $n_i \times q$ design matrix for the random effects for patient i.
β : The $q \times 1$ vector of fixed effects.
b_i : The $q \times 1$ vector of random effects for patient i.
$\theta_b = \text{vech}(D)$ (for later use).

Fixed effects are being interpreted exactly as in standard linear regression. The interpretation of b_i ,$i = 1, \ldots, p$ is the impact in the longitudinal response of the ith

patient for a unit change in one element of Z_i with all others held constant.

Random effects b_i account for the correlation between the longitudinal responses for the ith subject. Aditionally they are assumed independent of the error terms so that $\mathrm{Cov(e_i, b_i)} = 0$. When b_i cannot capture the correlation between the longitundinal responses of the ith subject, a change in the e_i's covariance matrix structure need's to be done. This can be due to use of few random effects or misspecification of their covariace matrix. Standard options are: $\sigma^2 I_n$ (non-correlated), AR ,or completely unspecified.

The longitudinal responses of the ith subject are independent conditionally on the random effects b_i :

$$Pr(y_i \mid b_i, \theta) = \prod_{j=1}^{n} \Pr(y_{ij} \mid b_i; \theta).$$

1.2 Estimation of a Linear Mixed Effects Model

Estimation of a Mixed Effects Model is based on Maximum Likelihood principles. The marginal density of the ith subject is:

$$\Pr(y_i) = \int \Pr(y_i \mid b_i) \Pr(b_i) \mathrm{d}b_i,$$

with : $y_i \mid b_i \sim \mathcal{N}(X_i\beta, V_i)$ where $V_i = ZiDZi^T + \sigma^2 I_{n_i}$.
Assuming intersubject independence log-likelihood takes the form :

$$\ell(\theta) = \sum_{i=1}^{n} log(\Pr(y_i;\theta)) = \sum_{i=1}^{n} log(\int \Pr(y_i \mid b_i;\beta,\sigma^2) \Pr(b_i;\beta,\sigma^2) \mathrm{d}b_i,$$

with :

$$\Pr(y_i;\theta) = (2\pi)^{-n_i/2}|V_i|^{-1/2} \exp\left(-\frac{1}{2}(Y_i - X_i\beta)^T V_i^{-1}(Y_i - X_i\beta)\right).$$

Where θ denotes the full parameter vector so that $\theta^T = (\beta^T, \sigma^2, \theta_b^T)$, with θ_b=$vech(D)$.

If V_i is known $\Pr(y_i;\theta)$ has a closed form and $\hat{\beta} = (\sum_{i=1}^{n} X_i^T V_i^{-1} X_i)^{-1} \sum_{i=1}^{n} X_i^T V_i^{-1} Y_i$ which is the GLS estimator.

If V_i is unknown it is replaced by $\hat{V}_i$ and Maximum Likelihood estimation is being applied with V_i being asymptotically unbiased.

For small samples the ML estimator : $\hat{\sigma^2} = \frac{\sum_{i=1}^{n}(Y_i - X_i^T\hat{\beta})^2}{n}$ is biased as in simple linear regression. We can instread use $\hat{\sigma^2} = \frac{\sum_{i=1}^{n}(Y_i - X_i^T\hat{\beta})^2}{n-p}$ the REML (Restricted Maximum Likelihood) estimator which is unbiased.

The main idea behind REML estimation is to separate the data in two parts, the data used for the estimation of V_i and the data used for the estimation of β. So, in practice, for the estimation of V_i we need to express the likelihood in therms of V_i. We then proceed maximizing the modified log-likelihood function :

$$\ell(\theta_b;\sigma^2) = -\frac{n-p}{2}\log(2\pi) + \frac{1}{2}\log|\sum_{i=1}^{n} X_i^T X_i| - \frac{1}{2}\log|\sum_{i=1}^{n} X_i^T V_i^{-1} X_i| -$$

$$-\frac{1}{2}\sum_{i=1}^{n}[\log|V_i| + (Y_i - X_i\hat{\beta})^T V_i^{-1}(Y_i - X_i\hat{\beta})] \propto$$

$$\propto -\frac{1}{2}\sum_{i=1}^{n}\log|V_i| - \frac{1}{2}\sum_{i=1}^{n}[(Y_i - X_i\hat{\beta})^T V_i^{-1}(Y_i - X_i\hat{\beta}) - \frac{1}{2}\log|\sum_{i=1}^{n} X_i^T V_i X_i|,$$

with : $\hat{\beta} = (\sum_{i=1}^{n} X_i^T V_i^{-1} X_i)^{-1} \sum_{i=1}^{n}(X_i^T V_i^{-1} Y_i)$.

To get the REML estimate $\hat{V}_i$ we maximize the above function with the use of EM or Newton-Rhapson algorithm .

Standard errors for the fixed-effects regression coefficients can be directly obtained by calculating the variance of the generalized least squares estimator.

$$\hat{\mathrm{Var}}(\hat{\beta}) = (\sum_{i=1}^{n} X_i^T \hat{Q}_i X_i)^{-1}(\sum_{i=1}^{n} X_i^T \hat{Q}_i \hat{\mathrm{Var}}(Y_i) Q_i X_i)(\sum_{i=1}^{n} X_i^T Q_i X_i)^{-1} =$$

$$= (\sum_{i=1}^{n} X_i^T \hat{Q}_i X_i)^{-1}$$

with $\hat{Q}_i = \hat{V_i^{-1}}$.

It is always preferable for the model to be correctly specified, but in practice we cannot be sure about a possible misspecification of the model. This makes us want to built models and choose estimators that are robust to misspecification. Such an estimator is $\text{Var}(y_i) = (y_i - X_i\hat{\beta})(y_i - X_i\hat{\beta})^T$ which is based on the famous so called sandwich estimator. If the standard error structure is correctly specified the model is always more efficient with standard error of the unique parameters in V_i :

$$\text{Var}(\hat{\theta_{b,\sigma}}) = [\text{E}(-\sum_{i=1}^{n} \frac{\partial^2 \ell_i(\theta)}{\partial \theta_{b,\sigma}^T \partial \theta_{b,\sigma}}|_{\theta_{b,\sigma}=\hat{\theta}_{b,\sigma}})]^{-1}.$$

1.3 Generalized Linear Mixed Effects models

The models for longitudinal data discussed above can only handle continuous responses. There are cases though that we want to incorporate a binary longitundinal responce ,or a longitudinal responce for counts in our joint model. To be able to incorporate such longitundinal responses in our joint models, Generalized Linear Mixed effects Models (GLMM) are used to handle the longitudinal part of the joint model. There are two main reasons for the widespread applicability of GLMMs, first GLMMs are a straightforward extension of the Generalized Linear Models to multivariate data, and it is currently possible to fit this type of models in a wide range of software packages.

Generalized linear mixed effects models, as Generalized linear models require a three part formulation :

Step 1 We assume that the conditional distribution of each of the Y_{ij}s given the random effects ,belongs to the exponential family of distributions with
$\text{Var}(Y_{ij}|b_i) = u\{E(Y_{ij}|b_i)\}\phi$, where $u(\cdot)$ is a known variance function of the conditional mean. Given the random effects the longitudital responses are independent of one another.

Step 2 We proceed to specify the conditional mean of Y_{ij}. It is assumed to depend upon fixed and random effects through :

$$\text{g}\{E(Y_{ij}|b_i)\} = n_{ij} = X_{ij}^{'}\beta + Z_{ij}^{'}b_i,$$

where $g(\cdot)$ is the link function.

Step 3 We assume a distribution over the random effects b_i. Any multivariate distribution can be assumed with most common choice being the normal with zero mean and a $q \times q$ covariance matrix G. The random effects are assumed independent of the covariates X_i.

In the joint modeling setting, two main Generalized Linear Mixed Effects models are used, *GLME for counts* and *GLME for a Binary responce*. These models are presented below.

Generalized Linear Mixed Effects Models for Counts

Generalized Linear Mixed Effects Models for Counts are used in Joint models when our aim is to quantify the contribution to the risk for an event of a marker that takes only integer values. For example if we want to quantify the contribution to the risk for an event of the number of seasures a patient has experienced, a Generalized Linear Mixed Effects Models for Counts would be an appropriate choice to handle the longitudinal submodel of the joint model .
Suppose that Y_{ij} is a count. We proceed in the usual three part specification :

Step 1 The $Y_{ij}|b_i$ are independent and have a *Poisson* distribution.
Hence : $\mathrm{E}(Y_{ij}|b_i) = \mathrm{Var}(Y_{ij}|b_i)$.

Step 2 The conditional mean of the longitundinal responce depends upon fixed and random effects. It has the form :

$$\log\{E(Y_{ij}|B_i)\} = n_{ij} = X_{ij}^{'}\beta + Z_{ij}^{'}b_i,$$

where $X_{ij}^{'} = Z_{ij}^{'} = (1, t_{ij})$, or any other suitable form.
We usually choose the log-link function since it is the canonical link for the *Poisson*

distribution.

Step 3 The random effects are assumed to have a bivariate or multivariate normal distribution.

Generalized Linear Mixed Effects Models for a Binary responce

Suppose we want to quantify the change in the risk for an event of a patient in the presence or not of a desease say rheumatoid arthritis in a joint model having the information provided by X_i.

We would introduce a binary logitundinal response Y_i which would take the value 1 in the presence of the desease and 0 otherwise . For this purpose we would construct a Generalized Linear Mixed Effects Models for a Binary responce to handle the logitudinal part of the joint model in which $Y_{ij} = 1$ when the patient suffers from rheumatoid arthritis and $Y_{ij} = 0$ otherwise.
We proceed in the three part specification :

Suppose Y_{ij} ,a binary responce with values 0 or 1.

Step 1 Conditional on the random effects Y_{ij}s are independent and have a Bernouli distribution.

Step 2 The Conditional mean depends on fixed and random effects and is defined as :

$$\log\{\frac{\Pr(Y_{ij}=1|b_i)}{\Pr(Y_{ij}=0|b_i)}\} = n_{ij} = X_{ij}^{'}\beta + Z_{ij}b_i,$$

where $Z_{ij} = (1, t_{ij})$, or any other suitable form.
We choose the logit-link function since it is the canonical link for the *Bernoulli* distribution.

Step 3 Every random effect is assumed to have a normal distribution.

1.4 Modeling the mean

1.4.1 Parametric Curves

In the sections above, formulation and estimation of the Linear Mixed Effects Model was discussed, Generalized Linear Mixed Effects Models have also been briefly presented. One major advantage of the Linear Mixed Effects Model is it's flexibility in the specification of the mean structure . The mean response over time can often be discribed by parametric or semi-parametric curves when the data allow for it. Options for this specification will be explored in this section.

Linear Trends The simplest and most used parametric curve. If the plot of the mean response over time is a straight line the Linear Trends option is the appropriate one. Suppose in a two-group study that we have two groups, 0 and 1 whose mean responses diverge in a straight line fashion with different intercept and slope. We can fit the parametric curve :

$$\mathrm{E}(Y_{ij}) = \beta_0 + \beta_1 t_{ij} + \beta_2 G_i + \beta_3 (t_{ij} \times G_i).$$

where t_{ij} denotes the jth measurent of the ith individual and G_i is 0 if the ith individual belongs to group 0 and else is 1. This model for the mean gives distinct

intercepts and slopes to the two groups with :

$$\mathrm{E}(Y_{ij}) = \beta_0 + \beta_1 t_{ij}.$$

for group 0 and :

$$\mathrm{E}(Y_{ij}) = \beta_0 + \beta_2 + (\beta_2 + \beta_3) t_{ij}.$$

for group 1.

A major advantage of the linear model for the mean is it's interpretability.

Quadratic Trends Changes in the mean will not always be linear. When not , different order polynomials can be considered. Suppose in a two group study, when plotted, the change in the mean response of the two groups are curves. Assuming that the changes in the mean response can be approximated by quadradic trends we have :

$$\mathrm{E}(Y_{ij}) = \beta_0 + \beta_1 t_{ij} + \beta_2 t_{ij}^2 + \beta_3 G_i + \beta_4 (t_{ij} \times G_i) + \beta_5 (t_{ij}^2 \times G_i),$$

with t_{ij}, G_i same as above. The mean response over time for the subjects in group 0 is :

$$\mathrm{E}(Y_{ij}) = \beta_0 + \beta_1 t_{ij} + \beta_2 t_{ij}^2.$$

and the mean response over time for the subjects in group 1 is :

$$\mathrm{E}(Y_{ij}) = \beta_0 + \beta_3 + (\beta_1 + \beta_4) t_{ij} + (\beta_2 + \beta_5) t_{ij}^2.$$

It is important to note that if we are going to fit a 2nd degree polynomial we need to include the 1st degree terms in the model too and possible interactions. This way the model remains invariant under linear tranformation, a very usefull property in the joint modeling setting.

Linear Splines

In some cases mean response over time cannot be characterized by first or second degree polynomials. In some cases we have a linear trend that changes at some point in time or we have multiple points in time where the intercept and slope changes.

The idea behind a linear spline is the following: We divide the time axis into segments and build a linear model in every one of these segments that join together at the timepoints in which division happened (knots).
Suppose we have a two group study as before with the mean response of the groups changing slope at time t^*. We fit a linear spline with one knot at t^* as follows :

$$\mathrm{E}(Y_{ij}) = \beta_0 + \beta_1 t_{ij} + \beta_2 (t_{ij} - t^*)_+ + \beta_3 G_i + \beta_4 [t_{ij} \times G_i] + \beta_5 [(t_{ij} - t^*)_+ \times G_i].$$

where t_{ij}, G_i as above. In this model for the mean and slope changes happen at time t^* for both groups.
So the model for subjects in group 0 becomes :

$$\mathrm{E}(Y_{ij}) = \beta_0 + \beta_1 t_{ij} \quad \text{for } t_{ij} \leq t^*$$

$$\mathrm{E}(Y_{ij}) = \beta_0 + \beta_1 t_{ij} + \beta_2 (t_{ij} - t^*) \quad \text{for } t_{ij} \geq t^*.$$

$$\mathrm{E}(Y_{ij}) = \beta_0 - \beta_2 t^* + (\beta_1 + \beta_2) t_{ij} \quad \text{for } t_{ij} \geq t^*.$$

And for subjects in group 1 :

$$\mathrm{E}(Y_{ij}) = \beta_0 + \beta_3 + (\beta_1 + \beta_4) t_{ij} \quad \text{for } t_{ij} \leq t^*$$

$$\mathrm{E}(Y_{ij}) = \beta_0 + \beta_3 + (\beta_1 + \beta_4) t_{ij} + (\beta_2 + \beta_5)(t_{ij} - t^*) \quad \text{for } t_{ij} \geq t^*.$$

$$\mathrm{E}(Y_{ij}) = \beta_0 + \beta_3 + (\beta_1 + \beta_4 + \beta_2 + \beta_5) t_{ij} - (\beta_2 + \beta_5) t^* \quad \text{for } t_{ij} \geq t^*.$$

Linear splines are a very useful and flexible way to accommodate non-linear trends when polynomials fail to do so adequatelly. Another attribute of Linear Splines is their interpretability as they are linear models afterall.

Highed Order Polynomial Splines Spline models can become more complicated by using piece-wise quadratic or cubic models instead. There are two parameters that characterize spline models ,the order of the piece-wise polynomial on each interval and the number of knots. If a spline is of the k th-order,then for each knot there is a covariate that allows the coeficient of the kth-order term t_{ij}^K to change. When we choose to fit a k-th order spline we need to include all the terms up to that degree. So, the number of parameters in a higher order polynomial spline is equal to the degree of the polynomial plus the number of knots plus one. This makes us want to use as parsimonous structures as possible in terms of knots and degree.

1.5 Modeling Covariance

When we have missing data, a common situation in longitudinal studies, the model is not robust to misspecification of the Covariance matrix, so Covariance must be handled with care. Mean and Covariance are interdependent based on the fact that the Covariance of any pair depends on the model of the mean. As a result, we model the Covariance matrix based on the model selected for the mean.
The most common choices for the Covariance model follow :

Unstructured In the Unstructured setting we fit a symmetric positive definite matrix to the Covariance. This choice is considered reasonable when all subjects are measured at the same timepoints and the number of occasions is relatively small. One major advantage of this approach is that we give no structure to the Covariance matrix. However leaving the Covariance matrix unspecified creates a lot of parameters to be estimated ($\frac{n(n+1)}{2}$) and the approach does not work when we have mistimed measurements :

$$\mathrm{Cov}(Y_i) = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2n} \\ \dots & \dots & \dots & \dots \\ \sigma_{n1} & \sigma_{n2} & \dots & \sigma_{nn} \end{pmatrix}$$

Compound Symmetry We assume constant Variance and Covariance between any pair of Y_{ij}s so that $\mathrm{Cov}(Y_{ij}, Y_{ik}) = \lambda$. In this setting we have only two parameters to estimate but the assumptions are somewhat unrealistic.

$$\mathrm{Cov}(Y_i) = \sigma^2 \begin{pmatrix} 1 & \lambda & \dots & \lambda \\ \lambda & 1 & \dots & \lambda \\ \dots & \dots & \dots & \dots \\ \lambda & \lambda & \dots & 1 \end{pmatrix}$$

Toeplitz This model for the Covariance is appropriate when measurements are performed at approximately equal intervals. We assume that equally separated in time measurements have the same Covariance and Variance is constant. The parameters for estimation are n.

$$\mathrm{Cov}(Y_i) = \sigma^2 \begin{pmatrix} 1 & \lambda_1 & \dots & \lambda_{n-1} \\ \lambda_1 & 1 & \dots & \lambda_{n-2} \\ \dots & \dots & \dots & \dots \\ \lambda_{n-1} & \lambda_{n-2} & \dots & 1 \end{pmatrix}$$

Autoregressive In this setting we assume the Variance is constant and Covariance decays over time at a rate to be estimated. A very appealing assumption as is usually the case in longitudinal studies. Another appealing feature is that the parameters to be estimated are only two.

$$\mathrm{Cov}(Y_i) = \sigma^2 \begin{pmatrix} 1 & \lambda & \dots & \lambda^{n-1} \\ \lambda & 1 & \dots & \lambda^{n-2} \\ \dots & \dots & \dots & \dots \\ \lambda^{n-1} & \lambda^{n-2} & \dots & 1 \end{pmatrix}$$

Banded The Covariance is assumed to be zero after a specific period of time has passed from the measurement. A very strong assumption to be made but gives flexibility over the number of the parameters.

Exponential We can generalize the Autoregressive model for the case in which the measurement are not fixed or equally placed, something very common in clinical studies as subjects often show up late for measurement.
We assume:

$$\mathrm{Cov}(Y_{ij}, Y_{ik}) = \lambda^{|t_{ij}-t_{ik}|},$$

with $\lambda > 0$.

1.6 Missing Data

Missing data arises in almost all real life statistical analyses. Especially in longitudinal studies where the longitudinal marker of a subject is being collected at a set of a prespecified follow-up times, it is very common for a subject not to show up (missing value) or even to drop-out completely from the study for personal or health reasons. Handling missing values is no easy task and if improperly handled missing values can lead to misleading inferences .

Missing data can be monotone or non-monotone. Monotone missing data is the result of attrition ,drop-out or late entry. Non-monotone missing data arise when the subject misses one or more measurements and show's up for a later measurement.

The forms of missingness take different types, with different impacts on the validity of conclusions from research: Missing completely at random, missing at random, and missing not at random.

At first we introduce the missing data indicator to distinguish the longitudinal responses collected from the ones not collected:

$$r_{ij} = \begin{cases} 1 & \text{if } y_{ij} \text{ is observed} \\ 0 & \text{otherwise} \end{cases}$$

So we separate Y_i into two subvectors:

Y_i^a : the observed data.

Y_i^m : the missing data.

The $r_i = (r_{i1}, r_{i2}, \ldots, r_{in_i})$ and the generating process r_i are refered as the missing data process.

1.6.1 Missing Completely at Random

In the Missing Completely at Random (MCAR) setting the probability that the responses are missing is assumed unrelated to the specific values that would have beed obtained and to the observed resposes. As a result r_i is independent of Y_i^o and Y_i^m. Hence $\Pr(r_i|Y_i^o, Y_i^m; \theta_r) = \Pr(r_i; \theta_r)$.

When data is MCAR the distribution of the observed data does not differ from the distribution of the complete data since the data are missing completely at random. As a result we can obtain valid inferences by analysing just the observed data.

1.6.2 Missing at Random

In the Missing at Random (MAR) setting the probability of missingness is assumed to depend on the observed data but not on the outcomes that would have been obtained (missing data). Hence: $\Pr(r_i|Y_i^o, Y_i^m; \theta_r) = \Pr(r_i\ |Y_i^o; \theta_r)$,
which is equivalent to:

$$\Pr(Y_i^m|Y_i^o, r_i; \theta_r) = \frac{\Pr(Y_i^m, Y_i^o, r_i; \theta_r)}{\Pr(Y_i^o, r_i; \theta_r)} = \frac{\Pr(r_i|Y_i^o, Y_i^m; \theta_r)\Pr(Y_i^o, Y_i^m; \theta_y)}{\Pr(r_i|Y_i^o; \theta_r)\Pr(Y_i^o; \theta_y)} =$$

$$= \frac{\Pr(Y_i^o, Y_i^m; \theta_y)}{\Pr(Y_i^o; \theta_y)} = \Pr(Y_i^m|Y_i^o; \theta_y),$$

where:

θ : the parameter vector of the joint distribution of the measurements and missing process.

θ_y : the parameter vector of the measurement model .

When data is MAR, the distribution of the observed data is not the same as the distribution of the complete data due to the fact that the probability of missingness depends on the set of observed responses. The estimation is being carried out based on observed data and can provide valid inferences even if we ignore the contribution of r_i provided that the measurement process is correctly specified.

The likelihood takes the form:

$$\mathcal{L}_i(\theta) = \int_\Omega \Pr(Y_i, r_i; \theta) dY_i^m = \int_\Omega \Pr(r_i | Y_i^o, Y_i^m; \theta_r) \Pr(Y_i^o, Y_i^m; \theta_y) dY_i^m =$$

$$= \int_\Omega \Pr(r_i | Y_i^o; \theta_r) \Pr(Y_i^o, Y_i^m; \theta_y) dY_i^m = \Pr(Y_i^o; \theta_y) \Pr(r_i | Y_i^o, \theta_r) = \mathcal{L}_i(\theta_y) \times \mathcal{L}_i(\theta_r),$$

where :

Ω is the parameter space of Y_i^m .

If θ_y,θ_r are disjoint $\theta = (\theta_y^T, \theta_r^T)^T$ equalts to the product of the parameter spaces θ_y and θ_r. Inference on θ_y can be based on $\Pr(Y_i^o; \theta_y)$ ignoring the likelihood of the misssingness process.

Although MCAR and MAR data arise in real life problems , in clinical trials we usually have Missing Not at Random data. MNAR data needs greater care since it can contain information that if ignored can produce bias in the analysis.

1.6.3 Missing Not at Random

In the Missing Not at Random (MNAR) setting the probability that longitudinal responses are missing depends on a subset of the responses that would have been observed.
In particular the distribution of r_i depends on atleast some elements of Y_i^m.

Missing Not at Random model families

The most complicated type of missing data and the one nearest to real life problems is Missing Not at Random (MNAR). There are three main families of models used to handle this type of missing data ,Selection Models ,Pattern Mixture Models ,Shared Parameter Models. In practice what distinguishes the following families of models is what do we condition the joint distribution of the complete data and the missing mechanism on.

Selection Models

$$\Pr(Y_i^o, Y_i^m, r_i; \theta_y) = \Pr(Y_i^o, Y_i^m; \theta_y) \Pr(r_i | Y_i^o, Y_i^m; \theta_r),$$

where $\Pr(Y_i^o, Y_i^m; \theta_y)$ being the marginal density for the measurement process and $\Pr(r_i | Y_i^o, Y_i^m; \theta_r)$ being the density of the missingness process conditioned on the longitudinal responses.

Pattern Mixture Models

$$\Pr(Y_i^o, Y_i^m, r_i; \theta_y) = \Pr(Y_i^o, Y_i^m | r_i; \theta_y) \Pr(r_i; \theta_r),$$

where $\Pr(Y_i^o, Y_i^m | r_i; \theta_y)$ is the conditional density given the missingness process and $\Pr(r_i; \theta_r)$ the marginal distribution for the missingness process.

Shared Parameter Models

$$\Pr(Y_i^o, Y_i^m, r_i; \theta_y) = \int_\Omega \Pr(Y_i^o, Y_i^m | b_i; \theta_y) \Pr(r_i, b_i; \theta_r) \Pr(b_i; \theta_b) db_i.$$

Here we integrate out the random effects. Given the random effects missingness and measurement processes are independent.

1.7 Imputation

There are some ways to draw inference from missing data proposed in the literature. In joint models we use imputation, a very popular technique since we have the power to choose between simplicity and preciseness of the imputation algorithm.

In this approach we construct the full dataset by filling the missing data with values according to some prespecified rule. A variety of rules have been proposed, from simple rules that are easy to implement, to complex rules that are computionally intensive. When choosing the "rule" it is advised to be aware of the tradeoff between richfulness and applicability.

It is important to note that imputation is a tool used by modern statisticians to overcome issues presented by missing data and nothing more, there are imputation approaches in which data are imputed in ways that are not mathematically solid for the sake convinience and numerical stability. Some imputation approaches are presented below. For simplicity fixed times of observation are assumed, something not realistic in the longitudinal setting that will be loosened up later.

Mean Imputation The simplest approach is to replace each missing value with the mean of the observed values for that variable. This approach usually produces distorted measures and hence not advised.

Last value carried forward In this approach we replace the missing value with the last observed measurement for that variable or even a pre-treatment measurement. This approach is considered conservative or anticonservative depending on the nature of the illness we study.

Using information from related observations In this approach we replace the missing value with the estimate of the measurement of someone close to the subject. For example if we have missing data regarding the income of the fathers of children in a school, we can fill the missing value with the mothers report of the fathers income.

Indicator variables for missingness When dealing with missing unordered categorical predictors ,a simple and often usefull way to impute the data is to add an extra category for the variable indicating missingness.

Imputation based on logical rules Sometimes we impute the data based on logical rules. If in a company survey for example,all the observed emploees have had 22 days off work in the previous year, it is safe to assume that this is the policy of the company and as a result the missing observations for the days off work will be 22 too.

Random Imputation

The techniques presented above although easy to implement are rarely close to reality since they make strong assumpionts about the missing data. A more formal way to impute data is Random Imputation. There are many ways that we can randomly impute missing data, some of them are presented below in a way that leads the reader to understand the imputation scheme used in the *Imputation in Joint Models* section.

Simple Random Imputation The simplest approach is to impute missing values of a variable based on the observed data for this variable. In this approach if in a

survey for example height value is missing from a subject, we draw a height value from the observed heights and impute it. This approach ignores the usefull information from any other measurement for the same patient (sex,weight,etc).

Using regression predictors A simple and general imputation technique which uses the observed information to perform regression predictions for the missing values. The predicted values are then imputed. Imputation can be deterministic or random, by ignoring or not the random term of the prediction. For example we want to study the income of college graduates across Europe. We observe income, age, sex and region for 1000 subjects. If income is missing for some subjects, we can perform this technique to impute the missing values. It is important to note that in order to improve the predictive power of the model usually extremely high values are top-coded to some relatively managable number. For example if a graduate in Switzerland somehow makes 10 million per year, this will affect the model tremendously, so we would top-code the value to 200.000 in order for his effect to the model to become managable.

Predictors used in the imputation model One question that arises when using regression is which of the predictors will be used in the regression. Imputation is no different although some standard regression rules can be violated. For example if we fit a regression model of earnings on sex, ethnicity, nationality, education, the number of months worked in the previous year, hours worked per week and indicators for whether the respondent's family receives any forms of income support. Generally it is not valid to use income support to predict income since income support depends on income, but for the purposes of imputation it is considered acceptable.

Matching Another way to impute data is to impute the missing value with the observed value of a subject that has the same (or nearly the same) predictors.

Routine multivariate imputation This approach is used when several variables are missing. We allow the outcome Y as well as the predictors X to be vectors and we fit a multivariate model to all the variables that have missingness.

Iterative Regression imputation A way to generalize imputation using regression predictors is to apply the method iteratively. Assume that we have Y_1, Y_5, Y_{12} missing and a complete vector X. To use Iterative regression imputation we first impute the missing data using the approach presented above and then we reimpute Y_1 given the observed and the imputed data. This procendure is repeated until convergence of the imputed data.

1.7.1 Model-based imputation

Another approach, the one that will be used for imputation in Joint models is to fully specify the distribution of the missing values. The idea has its roots on Bayesian grounds.

We draw sample from the posterior distribution of the missing data given the observed data. This is relatively hard to do in a straightforward fashion so an iterative scheme is needed. Technical information on the scheme are presented in the Imputation in Joint Models section.

1.7.2 Multiple imputation

In order to reduce the uncertainty produced by a single random imputed value for every missing data value, we can produce multiply imputed data values for every missing value. These values are used to recreate the complete dataset and standard analysis tools can be then applied. Inferences are then combined to draw an overall conclusion.

For example, suppose we want to draw inference about a regression coefficient β, but we have some missing values Y_i. As a first step we multiply impute every missing value with M values using any of the techniques presented above and obtain estimates $\hat{\beta}_m$ in each of the M datasets as well as standard errors, $s_1, ..., s_M$. As a second step we obtain an overall point estimate by simply averaging over the estimates from the separate imputed datasets, thus $\hat{\beta} = \frac{1}{m}\sum_{m=1}^{M}\hat{\beta}_m$. As a last step we obtain the variance estimate V_β that reflects variation within and between imputations:

$$V_\beta = W + (1 + \frac{1}{m})B,$$

where $W = \frac{1}{m}\sum_{m=1}^{M} s_m^2$ and $B = \frac{1}{m-1}\sum_{m=1}^{M}(\hat{\beta}_m - \hat{\beta})^2$.

Chapter 2

Time to Event Data Analysis

The first feature that must be taken into account when it comes to analysis of failure times is the shape of their distribution. Event times must be positive and usually have skewed distribution, hence statistical methods based on normality are not direclty applicable. We can overcome this problem though, through the use of suitable transformations.

The most important characteristic that distinguishes the analysis of time to event data from other areas in statistics is censoring. The defining feature of censored data is that the event of interest is not observed on all study subjects. Implications that arise when we use standard tools in censored data :

- Standard tools such as sample average, standard dev., t-test, linear regression cannot be used as they assume complete data and therefore will produce biased estimates for the distribution of event times and related quantities.
- Inferences can be more sensitive to missespification of survival times compared to complete data.

2.1 Censoring

There are three types of censoring based on when did the censoring happen :

- **Right Cesoring**: For a subset of the subjects under study the event of interest in only known to occur after a certain point in time.
- **Left Cesoring**: For a subset of the subjects under study the event of interest in only known to occur before a certain point in time.
- **Interval Cesoring**: For a subset of the subjects under study the event of interest in only known to occur between two points in time.

A second classification of censoring has to do with whether or not the probability of a subject being censored depends on the failure process. Two types arise :

- **Informative Censoring**: Censoring occurs for reasons directrly related to the study.A censoring mechanism is informative if at any time t, the failure rates that apply to the subject still in the study are different from those that apply to subjects who have dropped out of the study
- **Non-Informative Censoring**: Censoring occurs for reasons not related to prognosis or the study in general.

Depending on censoring type differend inferencial approaches should be considered. The majority of the literature has focused on methods that can handle right censored data because these are the most frequently encountered. When we have informative censoring little can be done because the data do not contain enough

information to account for the informativeness so external information should be sought.

There are three main categories of tools for analysis based on the distributional assumptions made :

- **Non-parametric**: No distributional assumptions on any component of the model.
- **Semi-parametric**: Distributional assumptions on some components of the model.
- **Parametric**: Full distributional specification of the model.

2.2 Basic Functions in Survival Analysis

Let $T^\star$ denote the random variable of failure times. If the time event is death, the survival funtion expresses the probability of survival beyong time t.
Assuming $T^\star$ is continuous, the survival function is defined as:

$$\Pr(T^\star > t) = \int_t^\infty f(s)ds,$$

where $p(s)$ denotes the coresponding density function. The survival function must be non-increasing as t increases and $S(t = 0) = 1$.

Another building block of survival analysis is the hazard function :

$$h(t) = \lim_{dt \to\ 0} \frac{\Pr(t \leq T^\star < t + dt | T^\star > t)}{dt},$$

for $t \geq 0$. This is the risk of an event occuring in the interval $[t, t + dt]$ as $dt \to 0$.

A related quantity is the integrated or cumulative hazard function:

$$H(t) = \int_0^t h(s)ds.$$

$H(\cdot)$ describes the accumulated risk until time t.

The survival function can also be expresed in terms of risk as :

$$S(t) = \exp(-H(t)) = \exp(-\int_0^t h(s)ds).$$

When estimating any characteristic of the event time distribution, censoring must be taken into to account or inferences will be biased.
Let T_i be the observed event time for the ith subject defined as the minimum of the real event time and the censoring time c_i. Let $\delta_i = \mathbb{1}(T^\star \leq c_i)$ be an indicator function which takes the value 1 if the observed event time corresponds to a true event and 0 otherwise. We will be using (T_i,δ_i) to estimate characteristics for the distribution of $T^\star$.

2.3 Estimators

2.3.1 The Kaplan and Meyer estimator

The Kaplan–Meier (KM) estimator ,also known as the product limit estimator is the most well known and used survival funtion estimator. A non-parametric estimator which makes no assumption on the underlying distribution of the failure times.

The probability of surviving beyong any given timepoint t can be written (Total probability Theotem) :

$$\Pr(T^\star > t) = \Pr(T^\star > t | T^\star > t-1) \Pr(T^\star > t | T^\star > t-2) \dots$$

The above expression is used to estimate survival probabilities and leads to the estimator:

$$\widehat{S_{KM}(t)} = \prod_{i:t_i \leq t} \frac{r_i - d_i}{r_i},$$

where r_i denotes the number of subjects at risk at the unique timepoint t_i and d_i denotes the number of events at t_i.

The estimator's variance can be calculated using Greenwood's formula(See end of the section). Using asymptotic normality, a confidence interval for S_i can be constructed. A better approach though would be to derive an asymmetric confidence interval for S_t based on a symmetric confidence interval for $\log(H(t))$. This ensures that the boundaries of the confidence interval won't cross the boundaries of $[0, 1]$.

The variance of $\log(\widehat{H_{KM}(t)})$ is derived using similar arguments as in Greenwood's formula for $\widehat{S_{KM}(t)}$:

$$\hat{\text{Var}}(\log(\widehat{H_{KM}(t)})) = \frac{\sum_{i:t_i \leq t} \frac{d_i}{r_i(r_i - d_i)}}{(\sum_{i:t_i \leq t} \log(\frac{r_i - d_i}{r_i}))^2}.$$

The Kaplan-Meyer estimate for the cumulative hazard function is derived using the estimation $S_{KM}(t)$ and the fact that $\widehat{H_{KM}(t)} = -\log(\widehat{S_{KM}(t)})$.

2.3.2 The Nelson-Aalen estimator

Another similar estimator for the cumulative hazard function is the Nelson–Aalen (NA) estimator. The Nelson–Aalen (NA) estimator is a non-parametric estimator of the cumulative hazard function used in the case of censored or incomplete data :

$$H_{NA}(t) = \sum_{i:t_i \leq t} \frac{d_i}{ri},$$

with d_i,ri the same as noted above.

Breslow estimator

Based on the NA estimator we can derive the following estimator for the survival function known as the *Brieslow* estimator:

$$\widehat{S_B(t)} = \exp[-H(t)_{NA}(t)] = \prod_{i:t_i \leq t} (-\frac{d_i}{r_i}).$$

To derive a confidence interval for $S(t)$ based on this estimator we estimate the variance of $\log \hat{H}_{NA}(t)$ using a formula similar to Greenwood's formula for $\log(\widehat{H}_{KM}(t))$.

The Breslow estimator has uniformly lower variance than the Kaplan-Meyer but is biased upward. For small samples we have $\hat{S}_{KM}(t) \leq \hat{S}_B(t)$ but the two estimators are asymptoticaly equivalent.

Greenwood's formula In the presence of censoring, Greenwood suggested the following estimate for the variance of the Kaplan-Meier estimate:

$$\widehat{V(t)} = \mathrm{Var}(\hat{\mathrm{S}}(\mathrm{t})) \approx \widehat{S(t)}^2 \sum_{t(i)\leq t} \frac{d_i}{r_i(r_i - d_i)},$$

This leads to standard error :

$$s.e\{\hat{S}(t)\} \approx \widehat{S(t)}\{\sum_{t(i)\leq t} \frac{d_i}{r_i(r_i - d_i)}\}^{1/2}.$$

2.4 Likelihood for Censored data

When survival data is assumed to be of a specific form, estimation of the parameters is often based on Maximum Likelihood. Let $[T_i, \delta_i], i = 1, 2, \ldots, n$ the survival information in a random sample from a distribution function $\mathcal{P}$, parametrized by θ with probability density function $\Pr(t; \theta)$.

In the Likelihood construction we need to account for censoring .A subject i for whom an event is observed at timepoint T_i contributes $\Pr(t; \theta)$ to the likelihood. A subject who is censored at time T_i contributes $S_i(T_i; \theta)$, since all we know about him is that he survived timepoint t.
Thus the log-likelihood takes the form:

$$\ell(\theta) = \sum_{i=1}^{n} \delta_i \log(\Pr T_i; \theta) + (1 - \delta_i) \log(S_i(T_i; \theta),$$

with $h(t) = \frac{\Pr(t)}{S(t)}$ and $S(t) = \exp(-H(t))$. Hence:

$$\ell(\theta) = \sum_{i=1}^{n} \delta_i \log(h_i T_i; \theta) - \int_0^{T_i} h_i(s; \theta) ds.$$

All subjects contribute an amount equal to the negative of the cumulative hazard function to the log-likelihood evaluated at their corresponding event time. Subjects who exprerienced an event contribute more than subjects whose events are being censored. Once the log-likelihood has been formulated iterative optimization procendures can be used to obtain the $\hat{\theta}$ estimate such as EM algorithm or Newton-Rhapson.

So far non-parametric estimators and the Likelihood construction in the presence of censoring have been presented .Usually survival data comes with supplemen-

tary information though. Information recorded throughout the study such as age, sex, smoking status, alcohol abuse history or treatment group. Our primary interest is to explore the relationship of those markers with the risk for death of the subject. To do so Cox proposed a class of models, the Relative Risk Models.

2.5 Relative Risk Models

Relative Risk Models are semi-parametric models and the idea behind them is that the risk for death for the group of people that is on a standard treatment (S) is proportional to the risk for death of the group that is on a new treatment (N).
The model in it's simplest form is expressed as:

$$\mathrm{h}_N(t) = \psi \mathrm{h}_S(t),$$

with t non-negative ,ψ constant and $\mathrm{h}_S(t)$, unspecified. The assumption that ψ is constant is an oversimplification which we will loosen up next.

If $\psi < 1$,then the hazard of death of a patient on the new treatment is smaller than the hazard of death of a patient on the normal treatment according to the model and vice versa. The value of ψ which is the hazard ratio between standard and new treatment will always be possitive so a natural reparametrization would be to express it as $\psi = \exp(\beta)$. Hence $\beta = \log(\psi)$ is the log-hazard ratio.

Let $x_1, x_2, \ldots, x_p$ be the values of the explanatory variables,$X_1, X_2, \ldots, X_p$ recorded at the time the study starts. A natural extension would be to form a model that captures the impact of those measurements in the risk for death of a patient. Let $x_i = (x_{1i}, x_{2i}, x_{3i}, \ldots, x_{pi})'$ be the vector of measurements of the ith subject and $\beta = (\beta_1, \beta_2, \beta_3, \ldots, \beta_p)$ the vector of coefficients of the explanatory variables $x_{1i}, x_{2i}, x_{3i}, \ldots, x_{pi}$.

We denote: $\eta_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \cdots + \beta_p x_{pi}$ or $\eta_i = \beta' x_i$ in matrix notation, the linear component of the model for the ith individual(prognostic score).

This class of models assume that covariates have a multiplicative effect on the hazard for an event. Then the model is formulated as :

$$\mathrm{h}_i(t) = \mathrm{h}_0(t)\exp(\eta_i) \quad \Rightarrow$$

$$\Rightarrow \quad \log(\frac{\mathrm{h}_i(t)}{\mathrm{h}_0(t)}) = \eta_i \quad \Rightarrow$$

$$\Rightarrow \quad \log(\frac{\mathrm{h}_i(t)}{\mathrm{h}_0(t)}) = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \cdots + \beta_p x_{pi}.$$

There is no constant term in the linear component of the model, if there was one, say β_0, it would be absorbed by the baseline hazard function. We observe that a regression coefficient β_i for the predictor x_i denotes the change in the log hazard ratio at any fixed timepoint for a unit increase in x_i while all other predictors are held constant. Hence, the hazard ratio change is $\exp(\beta_i)$ for a unit increase in x_i while all other predictors are held constant.

The Relative Risk model is an extremely flexible model so that we can include variates, factors, random effects and Stochastic processes in the linear component. The flexibility and broad applicability of the Relative Risk model makes it a complelling choice among the models for event time data. It must be used with caution though since it is not always the appropriate choice for the analysis. Preliminary analysis is advised to check if the model is appropriate. As a rule of thumb:When plotting the survival curves of the two groups if the survival curves are parallel to each other or they cross then the Relative Risk model is not appropriate choice for analysis.

2.5.1 Estimation in a Relative Risk model

Our aim fitting a Relative Risk Model is to estimate the coefficients in the linear component $\beta_1, \beta_2, \beta_3, \ldots, \beta_p$ and in some cases the baseline survival function. At first we estimate the $\beta_i' s$ and then we use these estimates to estimate the baseline hazard function. We can estimate $\beta_i' s$ with standard full likelihood approach, Cox in 1972 proposed a way easier method through the Partial likelihood function though.

Suppose that we have a study with n individuals participating, r distinct deaths and $n - r$ censored survival times. Assuming that only one individual dies at each death time we can order the death times so that :
$t_{(1)} < t_{(2)} < t_{(3)} < \cdots < t_{(r)}$.
$R(t_{(j)})$: the number of individuals at risk at time $t_{(j)}$(Risk set).

Let :

$$\delta_i = \begin{cases} 1 & \text{if event of } i \text{ is observed} \\ 0 & \text{otherwise} \end{cases}$$

Partial likelihood function for the Relative Risk model is :

$$\mathcal{L}_(\beta) = \prod_{j=1}^{n} \left(\frac{\exp \beta' x_{(j)}}{\sum_{l \in R(t_{(j)})} \exp (\beta' x_l)} \right)^{\delta_i},$$

with $x_{(j)}$: the vector of covariates for the subject that dies at $t_{(j)}$. Individuals for whom the survival times are censored do not contribute to the numerator but they contribute to the denominator. The partial likelihood only takes under consideration the order in which the subjects experiece the event.

Log-likelihood takes the form :

$$\ell(\beta) = \sum_{i=1}^{n} \delta_i(\beta' x_i - \log(\sum_{l \in R(t_{(j)})} \exp(\beta' x_l)).$$

In clinical studies events are observed or censored in discrete timepoints, so ties in the event order are rather usual. We will extend the likelihood function so that it takes into consideretion this usual feature.

Let s_j be the vector of sums of each of the covariates of the individuals who die at the jth event time $t_{(j)}$, $j = 1, 2, 3, \ldots, r$. Suppose there are d_j events at timepoint $t_{(j)}$ the hth element of s_j is $s_{hj} = \sum_{k=1}^{d_j} x_{hjk}$ where x_{hjk} is the hth explanatory variable, $h = 1, 2, 3, \ldots, p$ for the kth subject who dies at $t_{(j)}$.
An approximation to the likelihood function often used (Breslow 1974) :

$$\prod_{j=1}^{r} \frac{\exp \beta' s_{(j)}}{(\sum_{l \in R(t_{(j)})} \exp(\beta' x_l))^{\delta_j}},$$

with d_j being the deaths at timepoint $t_{(j)}$ (considered distinct). We maximize log likelihood with Newton-Rhapson approach.

After we have maximazed the log-likelihood function and have derived the estimates of the β's, the $\hat{\beta}_i$'s with $i = 1, 2, 3, \ldots, p$ we are ready to estimate the baseline hazard function.
The estimated hazard function for the ith individual is:

$$\hat{\mathrm{h}}_i(t) = \exp(\hat{\beta}' x_i)\hat{\mathrm{h}}_0(t).$$

Let d_j , n_j be the deaths and the number of individuals at risk at time $t_{(j)}$. The estimated baseline hazard function at $t_{(j)}$ is:

$$\hat{\mathrm{h}}_0(t_{(j)}) = 1 - \hat{\xi}_j \quad (4.1.1.1),$$

where $\hat{\xi}_j$ being the solution of the equation:

$$\sum_{l \in D(t_{(j)})} \frac{\exp(\beta' x_l)}{1 - \hat{\xi}_j^{\exp(\beta' x_l)}} = \sum_{l \in R(t_{(j)})} \exp(\hat{\beta}' x_l),$$

with $D(t_{(j)})$,$j = 1, 2, 3, \ldots, r$, the set of individuals who died in the timepoint $t_{(j)}$.

When there are tied death times this equation doesn't have a solution in explicit form and an iterative scheme is required. Assuming that the hazard of death is constant between event timepoints we can estimate the baseline hazard function by dividing (4.1.1.1) by the time interval.
Hence :

$$\hat{\mathrm{h}}_0(t_{(j)}) = \frac{1 - \hat{\xi}_j}{t_{(j+1)} - t_{(j)}},$$

for $t_{(j)} \leq t < t_{(j+1)}$,$j = 1, 2, 3, \ldots, r-1$ and $\hat{\mathrm{h}}_0(t) = 0$ for $t < t_{(1)}$.

Given that we have estimated the baseline hazard funtion we can proceed with the estimation of numerous important functions in survival analysis using the known relationship between them such as:

The baseline survivor function : $\mathrm{S}_{(0)}(t) = \prod_{j=1}^{k} \hat{\xi}_j$.

The baseline cumulative hazard function: $\mathrm{H}_{(0)}(t) = \sum_{=1}^{k} \log(\hat{\xi}_j)$.

In the model formulation of the Relative Risk model the baseline hazard function is completely unspecified. That means that it can be modeled by a scarar function, spline or any distribution that models failure times. That gives the model tremendous flexibity, something very usefull in the joint modeling setting.

2.5.2 Semi Parametric Specification of the baseline Hazard function.

One simple yet satisfactory in practice option would be the piecewise-constant model.

The baseline hazard function takes the form:

$$h_0(t) = \sum_{q=1}^{Q} \xi_q I(u_{q-1} < t < u_q),$$

where $0 = u_0 < u_1 < u_2 < \cdots < u_Q$ is a split of the timescale with u_Q being larger that the largest time observed and ξ_q is the value of the baseline hazard function in the interval (u_{q-1}, u_q). Increase in the number of knots will lead to increase in flexibility and better fit but also an increase in parameter count. In the special case that every interval contains only one event the models is equivalent to the completely unspecified model.

Another choice proposed in the literature is the regression spline. For the regression spline model, the log-baseline function is expanded to:$\log(h_0)(t) = k_0 + \sum_{d=1}^{m} k_d B_d(t, q)$ with $k^T = (k_0, k_1, \ldots, k_m)$ are the spline coefficients, q the degree of the B-spline's basis function $B.$ and $m = m' + q - 1$ denoting the number of knots. Again, increasing the number of knots increases flexibility and the number of parameters that need estimation.

2.5.3 Parametric Specification of the baseline Hazard function.

Non-parametric and semi-parametric techniques are used in order to avoid full specification of the baseline hazard function. Sometimes though, when past research or diagnostics suggest a specific form for the baseline hazard function then fully specifying the baseline hazard function is preferable since inference will be more precise. Below are the two most frequently used models for the baseline hazard function in Relative Risk models and their corresponding quantities of interest :

- **Exponential** : The hazard of death for the subject has the memoryless property. Under this model the baseline hazard function is : $\mathrm{h}_0(t) = \lambda$ for $0 \leq t < \infty$.
  Probability density function of the survival times: $f(t) = \lambda e^{-\lambda t}$.
  Survival function: $\mathrm{S}(t) = \exp[-\int_0^t \lambda du] = e^{-\lambda t}$ with λ a parameter that needs estimation.

- **Weibull** : A more general hazard function: $\mathrm{h}_0(\mathrm{t}) = \lambda\gamma t^{\gamma-1}$, for $0 \leq t < \infty$ with survivor function $\mathrm{S}(t) = \exp[-\int_0^t \lambda\gamma u^{\gamma-1} du] = e^{-\lambda t^\gamma}$ and probability density function $f(t) = \lambda\gamma t^{\gamma-1}\exp(-\lambda t^\gamma)$. With $\lambda > 0$ and $\gamma > 0$ scale and shape parameter to be estimated.

2.5.4 Estimation in parametric and semi parametric models

On the downside, in parametric and semi-parametric models we can't use partial likelihood and full likelihood approach is needed.
Assume that on n subjects we observe r events and $n - r$ events are censored. Let $t_1, t_2, t_3, \ldots, t_r$ be the event times and $t_1^\star, t_2^\star, t_3^\star, \ldots, t_{n-r}^\star$ be the censoring times (right).

The death times contribution to the likelihood is: $\prod_{j=1}^{r} \mathrm{f}(t_j)$.
For the censored individuals the only thing we know is that they survived timepoint $t^\star$, so their contribution to the likelihood will be : $\prod_{l=1}^{n-r} \mathrm{S}(t_j^\star)$.

Let δ_i be an indicator as specified above. We can consider the data as pairs of (t_i, δ_i) with $i = 1, 2, 3, \ldots, n$. Again the likelihood function can be written as :

$$\mathcal{L}(\beta) = \prod_{i=1}^{r} [\mathrm{f}(t_i)]^{\delta_i} [\mathrm{S}(t_i)]^{(1-\delta_i)} =$$

$$= \prod_{i=1}^{r} [h(t_i) S(t_i)]^{\delta_i} [\mathrm{S}(t_i)]^{(1-\delta_i)} =$$

$$= \prod_{i=1}^{r} [h(t_i)^{\delta_i} S(t_i)].$$

Hence the log-likelihood is :

$$\ell(\beta) = \sum_{i=1}^{n} \delta_i \log\{h(t_i)\} + \sum_{i=1}^{n} \log\{S(t_i)\} =$$

$$= \sum_{i=1}^{n} \delta_i \log\{h(t_i)\} - \sum_{i=1}^{n} H(t_i),$$

where $H(\cdot)$ is the cumulative hazard function. This function should then be maximised with respect to the unknown parameters.

2.5.5 Time-Dependent Covariates

In the Relative Risk setting introduced above we assumed that the covariates in the linear component of the model are constant throughout the study. However usually there is interest on whether or not, time-dependent covariates such as clinical parameters or longitudinal markers are associated with the risk for an event. There

are two categories of time-dependent covariates:*Exogenous* and *endogenous*.

Exogenous is a covariate that is associated with the rate of failures over time, but its future path is not affected by failure at time t. In a sense when a covariate is *exogenous* it affects the risk for an event, but the occurence of an event does not affect the covariate, something that is not the case with *endogenous* covariates.*Endogenous* is a covariate that is associated with the risk for an event and the occurence of an event affects the covariate back. For example the weather of a clinical trial's location is likely to affect the risk for death of the patients involved in the study. This would be an exogenous covariate as death of a subject will not affect the locations weather. On the other hand, consider the blood pressure of a subject in a clinical trial, certain levels of blood pressure affect the risk for death of the subject and death of a subject drops its pressure to zero, so pressure is an endogenous covariate. It is important to distinguish the two types since different models are appropriate for every type. *Exogenous* covariates can be handled with extending the standard Cox model whereas *Endogenous* covariates are handled with Joint Models for Longitundinal and Survival data .

2.6 Residuals

We will proceed to the formulation of the Martingale residuals for the Cox model as Martingale residuals will be used in the diagnostics of the joint model.
Suppose we have n subjects with events observed for the r of them and censored for the $n - r$. Suppose we fit a Cox regression model to the survival times with p explanatory variables. The fitted hazard function for the ith subject is:

$$\hat{h}_i(t) = \hat{h}_0(t)exp(\hat{\beta}' x_i),$$

with $\hat{\beta}' x_i$ being the estimated linear component and $\hat{h}_0(t)$ the estimated hazard function.

2.6.1 Cox-Snell residuals

Cox-Snell residuals are the most widely used residuals. They are not used in the joint modeling setting but are used to formulate the Martingale residuals. The Cox-Snell residual for the ith subject is:

$$r_{C_i} = \hat{H}_0(t_i) \exp(\hat{\beta}' x_i)$$

with $i = 1, 2, 3, \ldots, n$ and $\hat{H}_0(t_i)$ being the estimate of the cumulative hazard function. Usually the Nelson-Aalen estimate is being used.

Furthermore we have :

$$r_{C_i} = \hat{H}_i(t_i) = -\log \hat{S}_i(t_i).$$

If the model is correctly specified the Cox-Snell residuals will have a unit exponential distribution. If a survival time is censored the residual corresponding to the observation will also be censored .

2.6.2 Martingale residuals

The Martingale residual for the ith subject is defined as :

$$r_{M_i} = \delta_i - r_{C_i}.$$

With δ_i being an indicator of whether the event was observed or censored and r_{C_i} being the Cox-Snell residual. Martingale residuals take values between $-\infty$ and one, with the residuals of the censored observations being negative. Martingale residuals have similar properties with the residuals encountered in linear regression with expected value of zero and non-correlation.

2.7 Accelerated Failure Time models

The proportional hazards model is widely used due to it's extendability and simplicity. This simplicity as it is a major advantage sometimes becomes a big disadvantage. The distributions that are used in the estimation of the baseline function of a relative risk model are easy to handle but restrictive. The Accelerated Failure Time models offer greater flexibility since more distributions can be used. Some options are:

- **Gamma distribution** : The gamma distribution with parameters λ and κ has a survivor function : $S(t) = 1 - I_k(\lambda t)$ with $I_k(t) = \frac{\int_0^x \lambda^{\kappa-1} e^{-x} dx}{\Gamma(k)}$.
  The is no closed formula for the survival hazard functions but can be computed numerically . The gamma distribution can be reparametrized in terms of the distribution of log-time. By a simple change of variables it can be shown that : $T \sim \Gamma(\lambda, \kappa) \Leftrightarrow \log(T) = \alpha + W$.
  With $f_W(w) = \frac{e^{\kappa w - e^w}}{\Gamma(\kappa)}$ theGeneralized Extreme value distribution.

- **Generalized Gamma**: Generalized Gamma adds a scale parameter in the expression for $\log(T)$ above, making it extremely flexible as it includes the distributions Gamma($p = 1$),Weibull($k = 1$) and exponential($p = k = 1$).
  So we have : $\log(T) = \alpha + \sigma W$,with W having the Generalized Extreme value distribution with probability density function : $f_W(w) = \frac{\lambda p (\lambda t)^{p\kappa - 1} e^{(\lambda t)^p}}{\Gamma(\kappa)}$ with $p = 1/\sigma$

- **Log-Normal** : The hazard function of the log-normal distribution starts from 0 reaches a maximum and then descends to 0 again as $t \to \infty$. That is a very usefull feature in clinical studies as often discribes the recovery mechanism of

the body after an invasive therapy. The probability distribution for T is :

$$f_T(t) = \frac{1}{\sigma\sqrt{2\pi}} t^{-1} exp(\frac{-(log(t)-\mu)^2}{2\sigma^2}),$$

with $0 \leq t < \infty, \sigma > 0$. Survival function is given by :

$$S(t) = 1 - \Phi(\frac{\log(t)-\mu}{\sigma}).$$

- **Log-logistic** : Another distribution with a mode that is non monotonic is the log-logistic. With hazard function

$$\mathrm{h}(t) = \frac{e^{\theta}\kappa t^{\kappa-1}}{1+e^{\theta}t^{\kappa}},$$

where $0 \leq t < \infty, \kappa > 0$, survival function :

$$S(t) = (1+e^{\theta}t^{\kappa})^{-1}$$

and probability density function :

$$f(t) = \frac{e^{\theta}\kappa t^{\kappa-1}}{(1+e^{\theta}t^{\kappa})^2}.$$

Standard options of the Cox model are also available.

In the Accelerated Failure Time model setting predictors act multiplicatively on failure time. The idea is that predictors alter the time rate at which the subject proceeds along the time axis. This means that the model can be interpreted in terms of the speed of progression of a disease. An interpretation very appealing in the biostatistics context .

For example, let there be a group of people assigned to a standard treatment (S) and a group of people assigned to a new treatment (N). In the Accelerated Failure

Time setting a subject assigned the new treatment will survive time multiple to the time a subject assigned the standard treatment. Hence the probability that a subject on the new treatment survives timepoint t is the propability that a subject on the standard treatment survives timepoint $\frac{t}{\phi}$ with ϕ a positive constant to be estimated .

Let $\mathrm{S}_S(t)$ and $\mathrm{S}_N(t)$ be the survival functions of the subjects in the standard and new treatment respectively. Then:

$$\mathrm{S}_N(t) = \mathrm{S}_S(\frac{t}{\phi}).$$

The interpretation of ϕ is the effect of the new treatment in the timescale. How much it "slows" or "speeds up" time for a subject on the new treatment so if $\phi < 1$ we need to avoid the new treatment since it "speeds up" time .
From the relasonship between the hazard, survivor and probability density functions we get that :

$$\mathrm{f}_N(t) = \phi^{-1}\mathrm{f}_S(\frac{t}{\phi}).$$

and

$$\mathrm{h}_N(t) = \phi^{-1}\mathrm{h}_S(\frac{t}{\phi}),$$

with $\phi > 0$. A natural generalisation would be to write the parameter ϕ in the same way done in the Relative Risk setting :

$$\phi = e^{n_i}$$

with

$$n_i = \alpha_1 x_{1i} + \alpha_2 x_{2i} + \cdots + \alpha_p x_{pi}$$

the linear component of the model.

So the model is naturally generalised to handle richer data and becomes:

$$\mathrm{h}_N(t) = e^{-n_i}\mathrm{h}_S(\frac{t}{e^{n_i}}),$$

with x_{ij}: the value of the ith explanatory variable X_i for the jth individual. Baseline hazard function is handled as in Relative Risk models with the only difference being that we have more options if we choose the parametric path.

Writing the model in the log-linear scale:

$$\log(T_i) = \mu + \alpha_1 x_{1i} + \cdots + \alpha_p x_{pi} + \sigma e_p$$

where μ, σ, e_i a location parameter, a scale parameter and a random variable of a particular distribution used to model the deviation of $\log(T_i)$ from the linear part of the model.

Consider the survival time of the ith subject:

$$\mathrm{S}_i(t) = \Pr(T_i \geq t) = \Pr(\exp(\mu + \alpha' x_i + \sigma e_i) \geq t),$$

with $\alpha' x_i = \alpha_1 x_{1i} + \alpha_2 x_{2i} + \cdots + \alpha_p x_{pi}$ the linear component of the ith subject. Hence:

$$\mathrm{S}_i(t) = \Pr(T_i \geq t) = \Pr(\exp(\mu + \sigma e_i) \geq \frac{t}{\exp(\alpha' x_i)})$$

and the baseline survival function($x = 0$) is :

$$\mathrm{S}_0(t) = \Pr(\exp(\mu + \sigma e_i) \geq \ t) \quad \Rightarrow$$

$$\Rightarrow \quad \mathrm{S}_i(t) = \mathrm{S}_0(\frac{t}{\exp(\alpha' x_i)}),$$

which is the survivor function for the ith subject with acceleration factor $\exp(\alpha' x_i)$.

Using the log-linear formulation of the model in the survival function for the ith subject we obtain :

$$\mathrm{S}_i(t) = \Pr(T_i \geq t) = \Pr(\log(T_i) \geq \log(t)) = \Pr(\mu+\alpha_1 x_{1i}+\cdots+\alpha_p x_{pi}+\sigma e_p \geq \log(t)) =$$

$$= \Pr(e_i \geq \frac{\log(t) - (\mu + \alpha_1 x_{1i} + \cdots + \alpha_p x_{pi})}{\sigma}) = \mathrm{S}_{ei}(\frac{\log(t) - (\mu + \alpha_1 x_{1i} + \cdots + \alpha_p x_{pi})}{\sigma})\star.$$

Using the relationship of the survivor and hazard function :

$$\mathrm{h}_i(t) = \frac{1}{\sigma t}\mathrm{h}_{ei}(\frac{\log(t) - (\mu + \alpha_1 x_{1i} + \cdots + \alpha_p x_{pi})}{\sigma}).$$

Usefull results that will be used in the estimation process .

Estimation of Accelerated Failure Time models

Baseline hazard function is handled as in Relative Risk models with the only difference being that we have more options if we choose the parametric path.

Accelerated Failure Time models are fitted using full likelihood approach with the same way parametric Relative Risk models do.Baseline hazard function is handled as in Relative Risk models with the only difference being that we have more options if we choose the parametric path.

Let $t_1, t_2, \ldots, t_n$ be the n observed survival times.The likelihood function is :

$$\mathrm{L}(\alpha, \mu, \sigma) = \prod_{i=1}^{n} [f_i(t_i)]^{\delta_i} [S_i(t_i)]^{1-\delta_i},$$

where f_i, S_i and δ_i the same as in the Relative Risk model.
From $\star$ we have :

$$\mathrm{S}_i(t_i) = \mathrm{S}_{ei}(\frac{\log(t) - (\mu + \alpha_1 x_{1i} + \cdots + \alpha_p x_{pi})}{\sigma})$$

and differentiating with respect to t we obtain :

$$\mathrm{f}_i(t_i) = \frac{1}{\sigma t_i}\mathrm{f}_{ei}(\frac{\log(t) - (\mu + \alpha_1 x_{1i} + \cdots + \alpha_p x_{pi})}{\sigma}).$$

Let $z_i = \frac{\log(t) - (\mu + \alpha_1 x_{1i} + \cdots + \alpha_p x_{pi})}{\sigma}$.

The likelihood function then becomes :

$$\mathrm{L}(\alpha, \mu, \sigma) = \prod_{i=1}^{n} (\sigma t_i)^{-\delta_i} [\mathrm{f}_{e_i}(z_i)]^{\delta_i} [\mathrm{S}_{e_i}(z_i)]^{1-\delta_i}.$$

And the log-likelihood :

$$\ell(\alpha, \mu, \sigma) = \sum_{i=1}^{n} [-\delta_i \log(\sigma t_i) + \delta_i \log(\mathrm{f}_{ei}(z_i)) + (1 - \delta_i) \log(\mathrm{S}_{e_i}(z_i)].$$

Maximum likelihood estimates of the parameters $\mu, \sigma, \alpha_1 \ldots$ and α_p are obtained using Newton-Rhapson procedure.

2.7.1 Residuals

Assume an acceletated failure time model for T_i in the log-scale :

$$\log T_i = \mu + \alpha_1 x_{1i} + \cdots + \alpha_p x_{pi} + \sigma e_i.$$

with $\mu, \sigma, e_i, \alpha_i, T_i, i = 1, 2, 3 \ldots, p$ unknown parameters as above. If the survival time is censored the corresponding residual is also censored.

We will proceed in the formulation of the martingale residuals, since martingale residuals are used in the diagnostics of the joint models :

Standardised residual The standardised residual is defined as :

$$\mathrm{r}_{S_i} = \frac{\log t_i - \hat{\mu} - \hat{\alpha_1} x_{1i} - \cdots - \hat{\alpha_p} x_{pi}}{\hat{\sigma}},$$

with t_i the observed survival time of the ith individual and $\hat{\mu}, \hat{\alpha_i}, \hat{\sigma}$ $j = 1, 2, \ldots, n$ the estimated parameters.

If the model is correct, the estimated survivor function of the residuals would be similar to the survivor function of e_i,$S_{e_i}(e)$.
Hence if the model is correct, $-\log S_{e_i}(r_{S_i})$ will have a unit exponential distribution since $-\log S_{e_i}(e)$ has the unit exponential distribution.

Cox-Snell Residuals The estimated survivor function for the ith subject in an accelerated failure time model is given by the equation:

$$\hat{\mathrm{S}}_i(t) = \hat{\mathrm{S}}_{e_i}(\frac{\log t_i - \hat{\mu} - \hat{\alpha_1} x_{1i} - \cdots - \hat{\alpha_p} x_{pi}}{\hat{\sigma}}),$$

with $S_{e_i}(e), \hat{\mu}, \hat{\sigma}$ and $\hat{\alpha}_i$ for $i = 1, 2, 3, \ldots, p$ defined as above.

The Cox-Snell residuals are difined as :

$$r_{C_i} = \hat{H}_i(t_i) = -\log \hat{S}_i(t_i),$$

with $\hat{H}_i(t_i)$ the estimated hazard fuction for the ith subject and $\hat{S}_i(t_i)$ at t_i the estimated survivor function at t_i.

These residuals have a unit exponential distribution when the model is correctly specified. Cox-Snell residuals are closely related to standardised residuals since : $r_{C_i} = -\log S_{e_i}(r_{S_i})$.

Martingale residuals Martingale residuals measure the difference between the observed number of deaths and the predicted by the model number of deaths in any

given interval $(0, t_i)$ for the ith subject. Observations with large martingale residuals are not well fitted by the model. Martingale residuals are defined as :

$$r_{M_i} = \delta_i - r_{C_i},$$

with δ_i being the event indicator for the ith subject as defined above. Techniques of handling martingale residuals will be discussed in the Residuals section of the joint model.

2.8 Competing Risks Model

In the past section standard models of survival analysis have been presented. These models are used to quantify the contribution to the risk for an event of exogenous covariates. In all of the above models, subjects are considered to be at risk of a certain event. What if a subject is at risk of more than one types of events?

That is when a Competing Risks model is appropriate.The defining feature of a Competing Risks model is that it can handle more than one type of events which are mutually exclusive. For example a patient is at risk of death from more than one different causes. The basic idea behind a Competing Risks model is to run k different Proportional Hazards models for the ith patient given that he is alive at timepoint$t$. These generate the so called Cause Specific hazard fuction for every type of event :

$$h_k(t) = \lim_{\delta t \to 0} \{ \frac{\Pr(t \leq T < t + \delta t, K = k | T \geq t)}{\delta t} \},$$

Hence the Cause Specific survival function is :

$$S_k(t) = \exp(- \int_0^t h_k(u) du).$$

Although Cause specific quantities are easily interpretable the corresponding quantities are difficult to describe as a probability.

At this point basic knowledge required to formulate a Joint model for Longitudinal and Survival is covered and it is time to proceed in the formulaton of this broad class of models.

In the next chapter we proceed in a step by step formulation of a Joint model as well as presenting some of the extensions proposed in the literature along with their respective attributes and drawbacks.

Chapter 3

Joint Models for Longitudinal and Survival data

The Extended Cox model is only appropriate for handling exogenous time dependent covariates. When primary interest rests in the association between endogenous time dependent covariates and the risk for an event a Joint Model for Longitudinal and Survival data would be the appropriate model choice.

The idea behind these models is to couple the survival model with a suitable model for the repeated measurements of the endogenous covariate that will account for its special features. Our aim in the formulation of a Joint model is to measure the association of a longitudinal marker which is our endogenous covariate and the risk for an event.

A natural way to proceed is to make a two part model with a longitudinal and a survival submodel in which the longitudinal submodel will be plugged in the sur-

vival submodel. That way the correlation between the longitudinal responses will be carried inside the survival model. As a result, risk for an event will be personalized to every patient and so will confidence intervals or predictions.

Joint models can vary, from really simple to very complicated and computionally expensive. At first a standard simple Joint model will be presented, then we will work our way to more rich and sophisticated Joint models by presenting extensions in a step by step fashion.

There are two main classes of Joint models, the first distinguishes every subject from the rest by including random effects in the hazard function, the second not only distinguishes subjects from one another but induces dynamic in the way every subject behaves in time. This is achieved through the use of a Gaussian process on top of the random effects in the hazard fuction. Although a lot richer, a model with a Gaussian process in its survival function becomes extremely computionally demanding and difficult to interpret. The tradeoff between richness and complexity is common in every statistical model and Joint models are no exception.

3.1 Notation

We begin by denoting the standard quantities used in the formulation of a Joint model:

T_i : observed event time.

T_i^* : true event time.

C_i : minimum of the potential censoring time.

δ_i : event indicator as denoted earlier.

$y_i(t)$: the observed value of the longitudinal marker of the ith patient at timepoint t. Observations are taken at prespecified specific time occasions t_{ij}.

$m_i(t)$: The true unobserved value of the longitundinal marker for the ith subject at timepoint t.

$M_i(t) = \{m_i(s), 0 \leq s < t\}$: The history of the true unobserved longitudinal process up to timepoint t.

3.2 The Survival Submodel

We assume that the true value of the longitudinal marker $m_i(t)$ is never observed. What we observe is a measurement of the true value which is contaminated by measurement error $y_i(t)$. Association between $m_i(t)$ and the risk for an event is being quantified by a relative risk model of the form :

$$h_i(t|M_i(t), w_i) = \lim_{dt \to 0} \Pr(t \leq T_i^* < t + dt | T_i^*, M_i(t), w_i)/dt =$$

$$= h_0(t) \exp(\gamma^T w_i + \alpha m_i(t)), \quad t > 0,$$

where $h_0(\cdot)$ denotes the baseline risk function and w_i is a vector of baseline covariates with γ the corresponding vector of regression coeficients. Parameter α is the strength of the association of the longitudinal outcomes to the risk for an event, hence $\exp(\alpha)$ denotes the relative increase in the risk for an event for one unit increase in $m_i(t)$ at time t. As in the Relative Risk model $\exp(\gamma_i)$ is the change in ratio of hazards for one unit change in w_{ij} at timepoint t.

The model assumes that the risk for an event at time t depends only on the current true value of the longitudinal marker $m_i(t)$, something that does not hold for the Survival function.

Using the known relation between hazard and survival fuctions we have :

$$S_i(t|M_i(t), w_i) = \Pr(T_i^* > t | M_i(t), w_i) = \exp(-\int_0^t h_0(s) \exp(\gamma^T w_i + \alpha m_i(s)) ds).$$

which implies that the survival function depends on the whole history of the true longitudinal marker $M_i(t)$.

As a final step in the formulation of the survival submodel we need to choose the formulation of the baseline risk function. One choice is to leave $h_0(t)$ completely unspecified in order to avoid misspecification. Such an approach will lead to underestimation of the standard errors of the parameters of the joint model (Hsieh et al., 2006), an issue that will be addressed later. Another option is to assume strict parametric form and to use one of the appropriate distributions pressented in the Analysis of Time to Event data section. A third and more preferable option would be to use parametric but flexible specification for the baseline hazard function. Such options are step-functions , linear splines, quadratic trends, higher order polynomial splines or regression splines as in the standard Relative risk Model.

Choosing the third path will usually improve model fit significantly but will give rise to new parameters to be estimated. This can cause a very standard phenomenon in statistical analysis called overfitting. Overfitting is the production of an analysis that corresponds too closely or exactly to a particular dataset and may therefore fail to fit additional data or predict future observations reliably and occurs when we use more parameters in our model than the data justify. We should always keep balance between bias and variance in order to avoid overfitting. As a rule of thumb we keep the total number of parameters in the linear predictor and in the baseline risk function combined between 1/10 and 1/20 of the total number of events (Harrell, 2001, Section 4.4). After the number of knots has been decided their location is based on the precentiles of the observed times or only the true event times.

3.3 The Longitudinal Submodel

The aim of the longitudinal submodel in the Joint modeling setting is to successfully estimate the true longitudinal responses of the subjects and to reconstruct their complete true longitudinal history.

To achieve this we postulate a suitable Mixed-Effects model to describe the subject specific features and the correlation between the measurements of the same subject.
Assuming normally distributed outcomes, the longitudinal submodel takes the form:

$$y_i(t) = m_i(t) + e_i(t)$$

$$m_i(t) = x_i^T(t)\beta + z_i(t)^T b_i$$

$$b_i \sim \mathrm{N}(0, D),\ e_i \sim \mathrm{N}(0, \sigma^2)$$

where:
x_i : the fixed effects covariates (β).
z_i : the random effects covariates (b_i).

We assume that error terms are independent of each other, independent of the random effects and normally distributed with mean zero and variance σ^2 as in the Linear Mixed Effects model. The mixed model accounts for the measurement error through the random error term.

The time structure of $x_i(t), z_i(t)$ allows us to account for subject individuality and reconstruct each subjects complete longitudinal history, something needed for the survival function. The idea behind the model is as said above to associate the

true level of the longitudinal marker with the risk for an event.

The survival function depends on the whole history of the true longitudinal marker, so for a good estimation of $S_i(t)$ we need to obtain a good estimate of $M_i(t)$. This requires a correct enough specification of the time structure in $x_i(t)$, $z_i(t)$ and the possible interactions between baseline covariates and the time structure specified.

In applications that show highly non-linear longitudinal trajectories, flexible structures of $x_i(t), z_i(t)$ should be considered as high order splines, polynomials. Splines generally are considered the better choice due to better numerical and natural properties.

3.4 Estimation of Joint Models

In joint models two main strategies have been proposed for the estimation of the parameters. First is the two stage approach (Self and Pawitan 1992), in which the baseline risk function is left unspecified and the random effects are estimated first using the least squares method. These estimates are then used to impute appropriate values of $m_i(t)$ in the partial likelihood of the Cox model. This approach although relatively easy to implement found to produce biased results in many instances due to the existence of random terms (Dafni and Tsiatis 1998, Tsiatis and Davidian 2001, Ye et al. 2008, and Sweeting and Thompson 2011).

The second approach, the one that we will be using, is semiparametric maximum likelihood which was first proposed by Wulfsohn and Tsiatis in 1997. Most

modern Joint models adopt this approach although it makes the estimation process more computionally demanding. Athough more computionally demanding we still use semiparametric likelihood because estimators keep the asymptotic properties of maximum likelihood, something really usefull for the construction of confidence intervals and hypothesis testing.

We proceed in the formulation of the likelihood function. Let :

$\{T_i, \delta_i, y_i\}$: be the observed outcomes.

$b_i = (b_{i1}, b_{i2}, \ldots, b_{ip})$: be the vector of the time-independent random effects.

We assume that the vector of the time-independent random effects underlies both the longitudinal and the survival processes, hence accounts for both the association between the longitudinal measurements and event outcomes ,and the correlation between the repeated measurements.

We have :

$$\Pr(T_i, \delta_i, y_i|b_i; \theta) = \Pr(T_i, \delta_i|b_i, \theta) \Pr(y_i|b_i; \theta),$$

where : $\Pr(y_i|b_i; \theta) = \prod_j \Pr(y_i(t_{ij})|b_i; \theta)$.
with $\theta = \{\theta_t^T, \theta_y^T \theta_b^T\}$, the full parameter vector and
θ_t : parameters for the event time outcomes.
θ_y : parameters for the longitundinal outcomes.
θ_b : parameters for the random effects covariance matrix.

We assume that given the observed history, the censoring mechanism and the visiting process are independent of the true event times and the future longitudinal measurements. Both assumptions are violated if either of the processes depend on the random effects. In plain english these assumptions imply that subjects decide to appear for a longitudinal measurement (or not) based only their observed history and not on prognosis (non-informateveness). Evaluating the plausibility of this assumption (non-informativeness) for the visiting and censoring processes requires external information since the data do not contain information that can challenge this hypothesis.

Under these assumptions the log-likelihood for the ith subject is :

$$\ell(\theta) = \log(\Pr(T_i, \delta_i, y_i; \theta)) = \log\{\int \Pr(T_i, \delta_i, y_i, b_i; \theta) db_i\} =$$

$$= \log\{\int Pr(T_i, \delta_i | b_i; \theta_t, \beta)[\prod_j \Pr(y_i(t_{ij}) | b_i; \theta_y)] \Pr(b_i; \theta_b)\} db_i \star,$$

with the conditional density for the survival part being :

$$Pr(T_i, \delta_i | b_i; \theta_t, \beta) = h_i(T_i | M_i(T_i); \theta_t, \beta)^{\delta_i} S_i(T_i | M_i(T_i); \theta_t, \beta) =$$

$$= [h_0(T_i) \exp\{\gamma^T w_i + \alpha m_i(T_i)\}]^{\delta_i} \exp\{-\int_0^{T_i} h_0(s) \exp\{\gamma^T w_i + \alpha m_i(s)\} ds\},$$

where h_0 is specified with one of the approaches discussed above.

The second part inside the integral in $\star$ is :

$$\Pr(y_i | b_i; \theta_y) \Pr(b_i; \theta) = \prod_j \Pr(y_i(t_{ij}) | b_i; \theta_y) \Pr(b_i; \theta_b) =$$

$$= (2\pi\sigma^2)^{-n_i/2} \exp\{-||y_i - X_i\beta - Z_i b_i||^2 / 2\sigma^2\} \times (2\pi)^{-q_b/2} \det(D)^{-1/2} \exp\{-b_i^T D^{-1} b_i / 2\},$$

where q_b : the dimensionality of the random vector and $||\cdot||$ the Euclidean norm.

The global maximum of $\ell(\theta)$ with respect to θ is achieved using Expectation-Maximization, Newton-Rhapson or hybrid algorithms. An efficient algorithm that is used consists of one or two steps of E-M and Newton-Rhapson until convergence. We use a hybrid algorithm because E-M has exceptional performance in it's first steps but very poor performance near the global maximum. More on the exact scheme used later (Appendix).

Rizopoulos in 2009 has noted that the key function in either approach is the score vector which can be rewritten in the form :

$$S(\theta) = \sum_i \frac{\partial}{\partial \theta^T} \log\{\int \Pr(T_i, \delta_i | b_i; \theta) \Pr(y_i | b_i; \theta) \Pr(b_i; \theta) db_i\} =$$

$$= \sum_i \Pr(T_i, \delta_i, y_i; \theta)^{-1} \frac{\partial}{\partial \theta^T} \int \{\Pr(T_i, \delta_i | b_i; \theta) \Pr(y_i | b_i; \theta) \Pr(b_i; \theta)\} db_i =$$

$$= \sum_i \Pr(T_i, \delta_i, y_i; \theta)^{-1} \int \frac{\partial}{\partial \theta^T} \{\Pr(T_i, \delta_i | b_i; \theta) \Pr(y_i | b_i; \theta) \Pr(b_i; \theta)\} db_i =$$

$$= \sum_i \int [\frac{\partial}{\partial \theta^T} \log\{\Pr(T_i, \delta_i | b_i; \theta) \Pr(y_i | b_i; \theta) \Pr(b_i; \theta)\}] \times$$

$$\times \frac{\Pr(T_i, \delta_i | b_i; \theta) \Pr(y_i | b_i; \theta) \Pr(b_i; \theta)}{\Pr(T_i, \delta_i, b_i; \theta)^{-1}} db_i =$$

$$= \sum_i \int A(\theta, b_i) \Pr(b_i | T_i, \delta_i, y_i; \theta) db_i,$$

with $A(\theta, b_i) = \partial\{\log(\Pr(T_i, \delta_i | b_i; \theta)) + \log(y_i | b_i; \theta) + \log(b_i; \theta)\}/\partial\theta^T$ the complete data score vector. The observed data score vector is the expected value of the complete data score with respect to the posterior distribution of the random effects.

If the score equations are solved with respect to θ and $\Pr(b_i|T_i, \delta_i, y_i; \theta)$ is fixed at the value of θ of the previous iteration, this corresponds to a step of the E-M algorithm. Whereas if the score equations are solved with respect to θ considering $\Pr(b_i|T_i, \delta_i, y_i; \theta)$ as a function of θ this corresponds to direct maximization of $\ell(\theta)$. This can be used to a straight forward calculation of the standard errors.

To do so, can directly use the observed data vector to calculate the Hessian matrix and the standard errors using the observed information matrix.
We can rewrite the Hessian matrix as follows:

$$\frac{\partial S_i(\theta)}{\partial \theta} = \frac{\partial}{\partial \theta} \int A(\theta, b_i) \Pr(b_i|T_i, \delta_i, y_i; \theta) db_i =$$

$$= \int \frac{\partial A(\theta, b_i)}{\partial \theta} \Pr(b_i|T_i, \delta_i, y_i; \theta) db_i +$$

$$+ \int A(\theta, b_i) \frac{\partial \Pr(b_i|T_i, \delta_i, y_i; \theta)}{\partial \theta} db_i,$$

where :

$$\int A(\theta, b_i) \frac{\partial \Pr(b_i|T_i, \delta_i, y_i; \theta)}{\partial \theta} db_i =$$

$$= \int A(\theta, b_i) [\frac{\partial \log Pr(b_i|T_i, \delta_i, y_i; \theta)}{\partial \theta}]^T \Pr(b_i|T_i, \delta_i, y_i; \theta) db_i =$$

$$= \int A(\theta, b_i) [\frac{\partial \{\log Pr(\delta_i, T_i, |b_i; \theta) + \log Pr(y_i|b_i; \theta) + \log Pr(b_i; \theta)\}}{\partial \theta} -$$

$$- \frac{\partial \log Pr(T_i, \delta_i, b_i; \theta)}{\partial \theta}]^T \Pr(b_i|T_i, \delta_i, y_i; \theta) db_i =$$

$$= \int A(\theta; b_i) \{A(\theta, b_i) - S_i(\theta)\}^T \Pr(b_i|T_i, \delta_i, y_i; \theta) db_i.$$

However, in practice it is easier to use a numerical derivative routine such as forward or central difference approximation n (Press et al., 2007) and calculate the

Hessian matrix using the function that computes the score vector. After $I(\theta)$ is estimated, standard errors can be computed: $\text{Var}(\theta) = \{I(\hat{\theta})\}^{-1}$.
where : $I(\hat{\theta}) = -\sum_{i=1}^{n} \frac{\partial S_i}{\partial \theta}|_{\theta=\hat{\theta}}$.

The observed information matrix is preferable to the expected one due to dropout that is caused from the occurence of an event though.

3.4.1 Standard errors with unspecified baseline risk function

In Cox models, partial likelihood can be used even when $h_0(t)$ is left completely unspecified. On the other hand, in the Joint modeling framework the random effects force us to use the full likelihood approach. When full likelihood approach is used it is preferable to choose a parametric form for $h_0(t)$, or a flexible semi parametric form as discussed in earlier chapters. That is because, when we define a Joint model with an unspecified risk function, the calculation of the likelihood is based on non-parametric likelihood arguments under which the unspecified cumulative baseline hazard function $H_0(t) = \int_0^t h_0(s)ds$ is replaced by a step function with a jump at every event timepoint (van der Vaart, 1998). That makes the parameter vector θ of really high dimensionality, something that implicates the inversion of the Hessian matrix.

If plots strongly suggest that $h_0(t)$ needs to be left completely unspecified though,we use the following estimator from the M-step of the EM algorithm :

$$h_0(t) = \sum_{i=1}^{n} \frac{\delta_i I(T_i = t)}{\sum_{i=1}^{n} I(T_i = t) \int \exp\{\hat{\gamma}^T w_j + \hat{\alpha} m_j(t, b)\} \Pr(b_i | T_i, \delta_i, y_i; \hat{\theta}) db_i},$$

where $\Pr(b_i|T_i, \delta_i, y_i; \hat{\theta})$ the posterior distribution of the random effects. This estimator is a function of $h_0(t)$ (through $\Pr(b_i|T_i, \delta_i, y_i; \hat{\theta})$) for this reason standard errors for the remaining parameters that are based on the profile score vector

$$S(\beta, \sigma, \gamma, \alpha, \hat{S_0(t)}) = \frac{\partial}{\partial \theta^T_{-h}} \ell_p(\beta, \sigma, \gamma, \alpha, \hat{S_0(t)}),$$

where $\ell(\cdot)$ is the profile likelihood and will be underestimated. Bootstrapping can be used to overcome this issue (Hsieh et al. 2006), but the computional cost usually outweights the benefits .

3.4.2 Computional Issues

The main reason Joint models are not yet widely used, is the two integrals that arise in their likelihood. Usually these integrals don't have analytical solutions and as a result require numerical approximation. First, the integral with respect to time in the definition of the survival function :

$$\int_0^t h_0(s) \exp\{\gamma^T w_i + \alpha m_i(s)\} ds.$$

This is the least demanding of the two, since it is always unidimentional. It can be efficiently approximated using the 7 or 15 point Gauss-Kronrod rule (Press et al 2007).

The second integral is the integral with respect to the random effects in the specification of the score vector :

$$\int \partial\{\log \Pr(T_i, \delta_i|b_i; \theta) + \log \Pr(y_i|b_i; \theta) + \log \Pr(b_i; \theta)\}/\partial\theta^T \Pr(b_i|T_i, \delta_i, y_i; \theta) db_i.$$

This integral is multidimensional, when the number of random effects in the Joint model is small, it can be approximated using Gaussian quadrature rules and

Monte Carlo sampling. When random effects increase in number though its dimensionality increases and approximation becomes extremely challenging (Wulfsohn and Tsiatis, 1997; Henderson et al., 2000; Song et al., 2002)(Appendix).

Rizopoulos in 2012 proposed an approach that significantly decreases the computional burden of the approximation and is based on the adaptive Gauss-Hermit rule(Appendix) by exploiting the properties of the posterior distribution of the random effects.We first need to determine the mode $\hat{b}_i$ and second order derivative matrix $\hat{H}_i$. To do so , we first write the density in the log-scale and we have :

$$\log \Pr(b_i|T_i, \delta_i, y_i; \theta) \propto \sum_{j=1}^{n_i} \log \Pr\{y_i(t_{ij})|b_i; \theta_y\} + \log \Pr(b_i; \theta_b) + \log \Pr(T_i, \delta_i|b_i; \theta_t, \beta).$$

As n_i increases the leading term is the logarithm of the density of the linear mixed model, $\log \Pr(y_i|b_i; \theta_y) = \sum_j \log \Pr\{y_i(t_{ij})\}$, which is quadratic in b_i. Using a variant of the Bayesian central limit theorem and under certain regularity conditions we have:

$$\Pr(b_i|T_i, \delta_i, y_i; \theta) \xrightarrow{p} N(\tilde{b}_i, \tilde{H}_i^{-1}),$$

where $\tilde{b}_i$ is the mode of $\log \Pr(y_i|b; \theta_y)$ with re spect to b and $\tilde{H}_i = -\partial^2 \log \Pr(y_i|b; \theta_y)/\partial b \partial b^T|_{b=\tilde{b}_i}$. This suggests that as n_i increases, it is sifficient to re-center and re-scale the integrand for each subject by utilizing only the information that comes from the mixed-effects model for the longitudinal outcome.

In practice we first fit the linear mixed effects model for the longitudinal outcome and extract $\tilde{b}_i$, $\tilde{H}_{-1}$. Then instead of the tranformation used in the adaptive Gauss Hermit rule we recenter the integrand using the quantities extracted above. Hence:

$$E\{A(\theta, b_i)|T_i, \delta_i, y_i; \theta\} = \int A(\theta, b_i) \Pr(b_i|T_i, \delta_i, y_i; \theta) db_i \approx$$

$$\approx 2^{q_b/2} |\tilde{B}_t|^{-1} \sum_{t_1,..,t_q} \pi_t A(\theta, \tilde{r}_t) \Pr(\tilde{r}_t|T_i, \delta_i, y_i; \theta) \exp(||b_t||^2),$$

where $\tilde{r}_t = \tilde{b}_i + \sqrt{2}\tilde{B}_t b_t$ with $\tilde{B}_i$ denote the Choleski factor of $\tilde{H}_i$ and $\tilde{\theta}_y$ are the maximum likelihood estimates of the linear mixed effects model. This relocation procendure is implemented only once, hence the computional burden generated by the relocation procendure of the adaptive Gauss Hermite rule is completely dropped. This procendure also requires fewer quadrature points.

3.5 Inference in Joint Models

3.5.1 Hypothesis Testing

Since Maximum likelihood approach has been implemented, the standard asymptotic tests such as Likelihood ratio test, Score test and Wald test are all directly applicable for the test : $H_0 : \theta = \theta_0 \quad vs \quad H_1 : \theta \neq \theta_0$.
In particular :

Likelihood Ratio Test : Let $\hat{\theta_0}, \hat{\theta}$ be the maximum likelihood estimates under the null and the alternative hypothesis respectively. The test statistic is:

$$LRT = -2[\ell(\hat{\theta_0}) - \ell(\hat{\theta})].$$

Score Test : Let $S(\cdot)$ be the score function and $I(\cdot)$ be the information matrix under the alternative hypothesis. The test statistic is:

$$U = S(\hat{\theta_0})^T \{I(\hat{\theta_0})\}^{-1} S(\hat{\theta_0}).$$

Wald Test : Let $\hat{\theta_0}, \hat{\theta}$ defined as above. The test statistic is:

$$W = (\hat{\theta} - \theta_0)^T \{I(\hat{\theta})\}(\hat{\theta} - \theta_0).$$

Under H_0 the asymptotic distribution of all three tests is X_p^2 with p the parameters being tested. Likelihood ratio test is the most reliable although the most

computionaly expensive. On the other hand the Wald test is the least computionally demanding but the least reliable too. The Score test sits in the middle in both computional intensity and reliability. These tests are viable only when the models compared are nested. For non nested models information criteria are used as usual (AIC,BIC).

If we want to check whether a random effect should be included in the model or not, we set it's value and all of the related quantities to its parameters in the Covariance matrix to zero. The problem with that is that a diagonal element of the Covariance matrix becomes zero,which is the boundary of it's parameter space. In this case none of the tests discussed above follow the X_p^2 distribution. As a practical guideline, use of higher type I error is advised.

3.5.2 Confidence Intervals

Asymptotic Confidence intervals can be based on the Wald statistic :

$$[\hat{\theta} \pm 1.96\hat{se}(\hat{\theta})].$$

Similarly Confidence intervals for the fitted values can be based on the asymptotic normal distribution of the Maximum Likelihood estimator. For example, for the average longitudinal evolutions $\mu = X\beta$ in the longitudinal process we can construct a Confidence Interval :

$$\hat{\mu} \pm 1,96\hat{se}(\hat{\mu}) \Rightarrow$$

$$\Rightarrow X\hat{\beta} \pm 1,96[diag\{X\hat{Var}(\hat{\beta})X^T\}].$$

X denotes the design matrix of interest and $\hat{Var}(\hat{\beta})$ the block of the Hessian matrix

corresponding to $\hat{\beta}$.

3.5.3 Estimation of Random Effects

In most ocassions we build a joint model to draw subject specific inference. To achieve this, estimation of random effects is required. We will use a Bayesian approach to estimate the random effects. At first we will compute their posterior distribution and then we will derive an estimation using standard Bayesian procendures.

Let $\Pr(b_i;\theta)$ denote the prior distribution of the random effects and $\Pr(T_i,\delta_i|b_i;\theta)\Pr(y_i|b_i;\theta)$ denote the conditional likelihood.
The posterior distribution takes the form:

$$\Pr(b_i|T_i,\delta_i y_i;\theta) \propto \frac{\Pr(b_i;\theta)\Pr(T_i,\delta_i|b_i;\theta)\Pr(y_i|b_i;\theta)}{\Pr(T_i,\delta_i,y_i;\theta)} \propto$$

$$\propto \Pr(b_i;\theta)\Pr(T_i,\delta_i|b_i;\theta)\Pr(y_i|b_i;\theta).$$

Which is not a multivariate normal distribution and has to be numerically computed. As n_i increases however, it converges to normal.

To describe the posterior distribution summary measures will be used. Namely location measures that will be used are :

$$\overline{b_i} = \int b_i \Pr(b_i|T_i,\delta_i,y_i;\theta)db_i \quad : \text{posterior mean.}$$

$$\hat{b_i} = \text{argmax}_{\text{b}}\{\log(b_i|T_i,\delta_i,y_i;\theta)\} \quad : \text{posterior mode.}$$

Scale measures :

$$\mathrm{Var}(b_i) = \int (b_i - \overline{b_i})^2 \Pr(b_i|T_i, \delta_i, y_i; \theta) db_i \quad : \text{posterior variance.}$$

$$H_i = \{\frac{-\partial^2 \log(b_i|T_i, \delta_i, y_i; \theta)}{\partial b^T \partial b}|_{b=\hat{b_i}}\}^{-1} \quad : \text{the posterior inverted Hessian matrix.}$$

Estimation of the above is made with Empirical Bayes approach replacing θ with $\hat{\theta}$.

3.6 Missing Data

If the longitudinal outcome is of primary interest, the occurence of an event usually corresponds to discontinuation of the longitudinal process. For example, a subject dies, after the occurence of the event his longitudunal measurements cannot be collected. Thus we can draw a connection between missing data and the longitudinal process.

We first split the longitudinal data vector in two parts, the observed part, which is the part until droppout : $y_i^o = \{y_i(t_{ij}) : t_{ij} < T^*, j = 1, 2, \ldots, n_i\}$ and the missing part, which is the part after droppout : $y_i^m = \{y_i(t_{ij}) : t_{ij} > T^*, j = 1, 2, \ldots, n_i'\}$. The observed part cosinsts of measurements documented by the physician. The missing part consists of measurements that would have been collected until the end of the study had the event not occured. A grey arrea here forms when the event is death. It may not seem reasonable to consider the values of the longitudinal outcome after the event time. The Joint model though implicitly makes assumptions for the complete longitudinal response vector, including observations that would have been collected after the event or censoring. When this occurs, great caution is advised in the interpretation of the results.

We proceed to the formulation dropout process :

$$\Pr(T_i^*|y_i^o, y_i^m; \theta) = \int \Pr(T_i^*, b_i|y_i^o, y_i^m; \theta)db_i =$$

$$= \int \Pr(T_i^*|b_i, y_i^o, y_i^m; \theta) \Pr(b_i|y_i^o, y_i^m; \theta)db_i =$$

$$= \int \Pr(T_i^*|b_i; \theta) \Pr(b_i|y_i^o, y_i^m; \theta)db_i.$$

The third to fourth equality holds due to the conditional independence assumption. The time to dropout depends on y_i^m through the posterior distribution of b_i. This corresponds to a Missing Not At Random mechanism. Under the simple random effects structure this missing mechanism implies that subjects that show steep changes in their longitudinal trajectory may have different probability to dropout from the subjects that do not. With $\alpha = 0$, if we condition the process upon the available covariates we get that the dropout process does not depend on neither the missing process not the observed longitudinal process. This coresponds to the Missing Completely At Random mechanism.

Under $\alpha = 0$ there is no longer association between the two submodels. The joint probability of the dropout and longitudinal processes can be factorized as:

$$\Pr(T_i, \delta_i, y_i; \theta) = \Pr(T_i, \delta_i; \theta_t) \Pr(y_i; \theta_y, \theta_b) =$$

$$= \Pr(T_i, \delta_i; \theta_t) \int \Pr(y_i|b_i; \theta_y) \Pr(b_i; \theta_b)db_i.$$

Hence the parameters of the two submodels can be separately estimated. The parameters estimated under this setting are valid under the Missing At Random setting too.

Discontinuation also occurs due to censoring. We have assumed that the censoring mechanism may depend on the observed history of the longitudinal responses

but is independent of future responses. Hence censoring corresponds to MAR mechanism.

Additionally this class of models can handle intermittent missingness and attrition.To see this we need the log-likelihood of the observed data under the complete data model. To derive this log-likelihood we integrate y_i^m out of the likelihood of the complete data under the complete data model. The log-likelihood of the observed data under the complete data model is :

$$\ell(\theta) = \sum_{i=1}^{n} \log \int \Pr(T_i, \delta_i, y_i^o, y_i^m; \theta) dy_i^m =$$

$$= \sum_{i=1}^{n} \log \iint \Pr(T_i, \delta_i, y_i^o, y_i^m | b_i; \theta) \Pr(b_i; \theta) db_i dy_i^m =$$

$$= \sum_{i=1}^{n} \log \int \Pr(T_i, \delta_i | b_i; \theta) \{ \int \Pr(y_i^o, y_i^m | b_i; \theta) dy_i^m \} \Pr(b_i; \theta) db_i =$$

$$= \sum_{i=1}^{n} \log \int \Pr(T_i, \delta_i | b_i; \theta) \Pr(y_i^o | b_i; \theta) \Pr(b_i; \theta) db_i.$$

Due to the assumption : $\Pr(T_i, \delta_i, y_i | b_i; \theta) = \Pr(T_i, \delta_i | b_i; \theta) \Pr(y_i | b_i; \theta)$, the missing longitudinal responses are only involved in the density of the longitudinal submodel.

Due to the assumption $\Pr(y_i | b_i; \theta) = \prod_j \Pr(y_i(t_{ij}) | b_i; \theta)$, the longitudinal responses are conditionally independent on each other. Hence the integral on the missing values can be dropped.

In this Section the standard Joint model was presented. Standard in the sense that we use the standard option for the Survival submodel which is the Relative

Risk model and the Linear Mixed Effects model for the Longitudinal submodel. As said previously, a standard Joint model is appropriate when we want to explore the association of the current true level of a longitudinal marker of a subject with the risk for an event. This enables us to draw personalized inference and make personalized predictions for every subject in a study. Although powerfull the Joint model presented above makes very strong assumptions regarding the association between the Longitudinal and the Survival process. The next section tackles some of these assumptions by presenting ways to Extend the model.

Chapter 4

Extensions of the Joint Model

The biggest advantage of the full likelihood approach in Joint modeling is it's extendability. When full likelihood approach is implemented the Joint model can be extended to fit numerous occasions according to the problem and still have the so desired asymptotic properties as they come with the use of the full likelihood approach. In this section extensions of the standard Joint model presented above are being presented.

There are two steps in the process of extending a Joint model. The first step is to extend the Joint model in the sense that it handles the same type of data more efficiently and the second is to extend it to handle richer data.

For example, we want to formulate a Joint model that is able to handle exogenous covariates in a study that has two types of hazard. As a first step we need to extend our standard Joint model to handle exogenous covariates, then proceed to extend the new model to handle multiple types of hazard.

4.1 Extensions of the Standard Joint Model

In this section extensions of the standard Joint Model will be presented. As a Standard Joint Model we define the Joint model presented in the previous section.

4.1.1 Reparametrization

The assosiation between the current level of the longitudinal marker and the risk for an event is captured by parameter a. This can be very limiting in occasions and a more general approach can be implemented. We first denote the vectors w_{i1}, w_{i2}. These are vectors of the covariates, they can contain all or some of the covariates and they can have some covariates in common.

We denote the risk for an event as :

$$h_i(t) = h_0(t) \exp\{\gamma^T w_{i1} + f(m_i(t-c), b_i, w_{i2}; \alpha)\},$$

where : $f(\cdot)$ is a function of the true level of the longitudinal marker, the random effects and extra covariates w_{i2}.
Under this formulation α can denote a vector of association parameters instead of a simple scalar

4.1.2 Including Factors

Sometimes the current risk for an event is dependent on a past value of the true longitudinal marker. For example the smoking status of a patient 10 or 5 years ago would be a more appropriate covariate to include in a lung cancer study than the current smoking status. In this case the standard approach should not be able

to come to logical conclusions. Lagged terms can be used to take account for this feature.

The model becomes :

$$h_i(t) = h_0(t) \exp\{\gamma^T w_i + \alpha m_i\{max(t-c, 0)\}\}.$$

The model above takes the value of the longitudinal marker lagged by c time units. We can further generalize the above model adding the slope of the longitudinal marker in the survival submodel (Ye et al. (2008b)). This generilization is really usefull in clinical trials since usually the risk for an event is often assosiated with the speed at which a desease progresses. When for example a patient has a severe worsening in his health condition, his longitudinal responses will usually reflect that, the biggest the change in his condition, the biggest usually is the change in his longitudinal marker levels. Hence a generalization like this will be really usefull in this context.

The extended model is then formulated :

$$h_i(t) = h_0(t) \exp\{\gamma^T w_i + \alpha_1 m_i(t) + \alpha_2 m_i'(t)\}.$$

where : $m_i(t)' = \frac{d}{dt} m_i(t) = \frac{d}{dt}\{x_i^T \beta + \alpha z_i^T b_i\}$.

In the extensions disscused above, the effect of the true longitudinal marker, current or lagged and the that of the slope is the same for all subgroups. This assumption might be untrue as different subgroups might have features which interact differently with their longitudinal marker. For example in a heart desease study suppose we have some diabetic patients, we would probably want to include the interaction of the longitudinal marker with diabetes as usually diabetic patients behave differently than non-diabetic patients.

To overcome this we can introduce interactions in the linear predictor. The model becomes :

$$h_i(t) = h_0(t) \exp\{\gamma^T w_{i1} + \alpha^T \{w_{i2} \times m_i(t)\}\},$$

where : w_{i1} are the direct effects of the baseline covariates to the risk for an event and w_{i2} are the interaction effects .

In the standard approach the risk for an event is associated with the current true level of the longitudinal marker. We added lagged effects, interactions and the effect of the slope of the longitudinal marker to the risk for an event. In practice even a model with all these extensions can be restricting since in most deseases the risk for an event depends on the history of the longitudinal response $M_i(t)$.
As a first step towards this direction (Sylvestre and Abrahamowicz, 2009; Hauptmann et al., 2000; Vacek, 1997) proposed models that allow risk to depend on a function of the longitudinal marker history.

One approach is to include in the linear predictor of the relative risk submodel the integral of the longitudinal trajectory, representing the cumulative effect of the longitudinal outcome up to time point t :

$$h_i(t) = h_0(t) \exp\{\gamma^T wi + \alpha \int_0^t m_i(s) ds\}.$$

As a step further we can adjust the integrand and multiply $m_i(t)$ with a chosen weight function. There are several weight functions used but in practice any weight function with good numerical properties can be used. A logical thing to consider would be to use a function that places high weights to recent values and low weights in longitudinal values further in the past. These integrals often do not have a closed form, and need to be computed numarically.

4.1.3 Exogenous Time-Dependent Covariates

The Standard Joint model approach accounts for endogeneiety as discussed above. Exogeneiety is often an issue to be investigated too. Suppose we have a multinational study on melanoma patients, we will need to take into account the climate of the country every patients lives in as sunshine greatly affects the desease. This can be in the form of, days with sunshine per year, town temperature or even a dummy variable denoting the country the patient lives in. We can extend the relative risk submodel to handle these situations just by including exogenous covariates in an augmented term m_i^* which takes the place of the longitudinal trajectory (endogenous).

The new model is formulated :

$$h_i(t) = h_0(t) \exp\{\gamma^T wi + \alpha m_i^*(t)\},$$

where $m_i^*(t)$ includes endogenous and exogenous time-dependent covariates.

We proceed to the estimation process as usual and all the extensions presented above can be applied. The only difference occurs in the calculation of the integral in the definition of the hazard function. For computional reasons we expand as:

$$S_i(t|M_i(t), w_i) = \exp\{-\int_0^t h_i(s)ds\} =$$

$$= \exp\{-\sum_{q=1}^{Q} \int_{\Omega_{iq}} h_0(s) \exp\{\gamma^T w_{iq}(s) + \alpha m_i(s)\}ds\}.$$

where $\{\Omega_{iq}, q = 1, 2, \ldots, Q_i\}$ denote the time intervals during which the exogenous time-dependent covariates $w_i(t)$ are assumed constant.

4.1.4 Stochastic Process Models

There are situations where serial correlation between the longitudinal measurements of a subject cannot be captured by the random effects. Our first action would be to change the Covariance matrix of the errors. If serial correlation persists we can add a Zero-mean Gaussian Process $V(\cdot)$ to capture the remaining serial correlation (Henderson, Diggle, Dobson 2000).

The longitudinal submodel becomes:

$$y_i(t) = m_i(t) + V(t) + e_i(t),$$

where $V(t)$ is a stationary Gaussian process independent of b_i that are used to capture local deviations.
We denote that $V(t) \sim N(0, \sigma_u^2)$ with $Corr(V(t), V(t-u)) = \exp(-|u|^v/\phi)$,where v is fixed and ϕ is to be estimated.

This extension can be implemented along with any of the extensions presented above and is incredibly powerfull when it comes to capturing serial correlation. This is a very usefull feature and its biggest drawback, since the comptutional complexity becomes too demanding at times.

In the standard approach the trajectory followed by each subject is dictated by time-indepentent random effects alone. The intuitive interpretation is that the characteristics of each subject are inherited and do not change over time. In applications that show highly non-linear longitudinal trajectories or serial correlation that cannot be captured by b_i's alone we can allow the random effects to vary over time with the introduction of this Gaussian zero mean process . The features that distinguish subjects from each other are now time dependent and hence are not inherited. This

although a more rich and natural approach as subjects change through time is more computionally demanding as the parameter space increases in dimensionality and difficult to interpret.

The parameter space for each subject becomes $(b_0, b_1, ..., b_k, V_1, ..., V_{n_i})$ where n_i is the number of measurements taken from the ith subject. As the number of measurements per subject increases the parameter space's dimensionality increases rapidly making it imposible to implement.

4.1.5 Accelerated Failure Time

When the proportionality assumption fails we can consider the use of the Accelerated Failure Time model. As presented in the first chapter, in Accelerated Failure Time models predictors act multiplicatively on the failure time. In a sense predictors, accelerate or decelerate time for the subject and hence the time for a event. There are situations that an Accelerated Failure Time model is more appropriate than a Relative Risk model. The same happens in the Joint modelling setting, there are situations where the true level of the longitudinal marker alters the flow of time for a patient. That is when we want to use an Accelerated Failure Time model as our Survival submodel (Tseng et al. 2005).

In Accelerated Failure Time models we have $\log T_i^* = \gamma^T w_i + \sigma_t e_{T_i}$, with σ_t a scale parameter, γ_j the change expected in log failure time for one unit change in the covariate w_{ij} and e_{it} the error terms with standard options for the distribution being the Normal,t or Extreme value.

In order to incorporate time dependent covariates in this setting we let :

$$S_0 \sim \int_0^{T^*} \exp\{\gamma^T w + \alpha m(s)\} ds,$$

the baseline survival function be absolutely continuous.
The risk function for the ith subject takes the form :

$$h_i(t|M_i(t), w_i) = h_0(V_i(t)) \exp\{\gamma^T w + \alpha m(s)\} \star,$$

where $V_i(t) = \int_0^t \exp\{\gamma^T w + \alpha m(t)\} ds$. As we can see, the subject specific hazard is assumed to be influenced by the entire covariate history, as $h_0(\cdot)$ is evaluated at $V_i(t)$. The baseline hazard function can be specified according to the options presented in the Accelerated Failure Time models section. An issue when introducing time-varying covariates in AFT models is that interpretation becomes more complicated because parameter α is involved in both terms in the right-hand side of $\star$. As a general interpretation guideline we have that in the accelerated failure time models setting the subject ages on $V_i(t)$ acceleration compared to the baseline S_0.

4.1.6 Generalized Linear Mixed Models

All the models presented until now attempt to quantify the association between a continuous longitudinal response or any of its features and the risk for an event. Sometimes though we want to explore the association between a categorical or a binary longitudinal response and the risk for an event. That's when Generalized Mixed Effects Models come in.

The idea is to formulate a Joint model by combining a Generalized Linear Mixed Effects model with a Relative Risk model under the standard independence assumptions. To do this we postulate two sepatare submodels as in the Linear

Mixed Models case with the only difference being that our longitudinal submodel is a Generalized Linear Mixed Effects model.

The models is:

$$\Pr(y_i(t)|b_i) = \exp\{\sum_{j=1}^{n_j}[y_{ij}\psi_{ij}(b_i) - c(\psi_{ij}(b_i))]/\alpha(\phi) - d(y_{ij},\phi)\}$$

$$m_i(t) = \mathrm{E}(y_i(t)|b_i) = g^{-1}\{x_i^T\beta + Z_i^T b_i\}$$

$$b_i \sim N(0, D)$$

$$h_i(t) = h_0(t)\exp\{\gamma^T w_{i1} + f(m_i(t-c, b_i, w_{i2}; \alpha))\}$$

where h_0 is the baseline hazard function, with the same options for it's specification. The interpretation for the vector of association parameters α under the different parameterizations remains the same as it was explained in the previous sections. The extensions presented in the sections above can be incorporated in the Generalized setting.

Maximum likelihood approach is implemented for parameter estimation:

$$\ell(\theta) = \sum_{i=1}^{n}\int \Pr(T_i, \delta_i|b_i;\theta)\{\prod_{i=1}^{n_i}\Pr(y_{ij}|b_i;\theta)\}\Pr(b_i;\theta)db_i =$$

$$= \sum_{i=1}^{n}\int [h_0(T_i)\exp\{\gamma w_{i1} + f(m_i(T_i-c), b_i, w_{i2};\alpha)\}]^{\delta_i}\times$$

$$\times\exp\{-\int_0^{T_i} h_0(s)\exp[\gamma^T w_{i1} + f(m_i(s-c), b_i, w_{i2};\alpha)]ds\}\times$$

$$\times\exp\{\sum_{j=1}^{n_i}\{y_{ij}\psi_{ij}(b_i) - c[\psi_{ij}(b_i)]\}/\alpha(\phi) - d(y_{ij},\phi)\}\times$$

$$\times(2\pi)^{-q_b/2}\det(D)^{-1/2}\exp\{-b_i^T D^{-1}b_i/2\}db_i.$$

A combination of numerical integration and maximization routines is used to find the global maximum .

Another use of Generalized Linear Mixed Effects in Joint models is handling non random dropout in discrete longitundinal responses (Pulkstenis et al., 1998; Albert and Follmann, 2000; Albert et al., 2002) or investigating the association structure between the categorical longitudinal process and the censored event time data (Faucett et al., 1998; Rizopoulos et al., 2008; Yao, 2008; Li et al., 2010) .

4.2 Further Extending The Joint model

In the previous section extensions of the Standard Joint model have been presented. In this section we further extend the Joint model making it able to handle non-homogenous population, recurrent and multiple types of events. It is worth mentioning that any of the extensions presented above can be applied to any of the extensions that will be presented below as a part of a two step extension.

4.2.1 Stratified Relative Risk

We start with homogeneity in the population, an assumption usually not realistic. For example, it is possible for women to be more prone to a disease than men.

We can extend the standard Joint model by altering the baseline hazard function for the different clusters of the population according to the feature we believe makes them fall to a certain category .

The survival submodel takes the form :

$$h_i(t) = h_{0k}(t) \exp\{\gamma^T w_i + \alpha m_i(t)\}.$$

with $h_{0k}(t)$ being the baseline hazard function for the k_{th} stratum.

We can further generalize the model by altering the effect of the true level of the longitudinal marker on the risk for an event for every stratum:

$$h_i(t) = h_{0k}(t) \exp\{\gamma_k^T w_i + \alpha_k m_i(t)\}.$$

in this model $\gamma_k, h_{0k}, \alpha_k$ all depend on the stratum of the subject.

To be able to use the Stratified Relative Risk extension we need to be able to assign every member of the population in the study to a cluster, something not always possible. Consider a quality that is not directly observable and has an effect to the risk for an event. In this situation the clusters are not observable.

What can we do when the clusters are not observed? That is when a Latent Class Model for the Survival submodel of the Joint model is appropriate.

4.2.2 Latent Class Models

Latent Class Joint Models (Proust-Lima et al., 2009; Lin et al., 2004, 2002) is a class of models related to the Stratified Relative Risk models. This class of models assumes that the heterogeneity of the population is latent and hence not captured by any of the observed covariates. Assume that we have G subpopulations in our population . Let $c_i = 1, ..., G$ be the class membership indicator of the ith subject.

These models work under the following conditional assumptions :

$$\Pr(T_i, \delta_i, y_i | c_i = g, b_i; \theta) = \Pr(T_i, \delta_i | c_i = g; \theta) \Pr(y_i | c_i = g, b_i; \theta)$$

$$\Pr(y_i | c_i = g, b_i; \theta) = \prod_j \Pr(y_i(t_{ij}) | c_i = g, b_i; \theta).$$

In this class of models we assume that the random effects account for the correlation between the repeated measurements and c_i's account for the association between the longitudinal process and the survival process. This enables the use of a more flexible association structure compared to the classical approach presented until now.

The model is specified as:

$$h_i(t | c_i = g) = h_{0g}(t) \exp\{\gamma_g^T w_i\}$$

$$\{y_i(t) | c_i = g\} = x_i^T \beta_g + Z_i b_{ig} + e_i(t)$$

$$\Pr(c_i = g) = \frac{\exp\{\lambda_g^T u_i\}}{\sum_{l=1}^{G} \exp\{\lambda_l^T u_i\}},$$

where : $e_i(t) \sim N(0, \sigma^2)$, $b_{ig} \sim N(\mu_g, \sigma_g^2 D)$ and $\lambda^T = (\lambda_1, \ldots, \lambda_G)$.

Patients are grouped up and groups are assumed to have different longitudinal evolutions and different risks for an event. The Covariance matrix is typically assumed to depend on c_i only via the scalar variance parameter σ_p^2.

The log-likelihood of the model is :

$$\ell(\theta) = \sum_{i=1}^{n} \log\{\sum_{g=1}^{G} \Pr(c_i = g; \theta) h_i(T_i | c_i = g; \theta)^{\delta_i} S_i(T_i | c_i = g; \theta) \times$$

$$\times \int [\prod_j \Pr(y_i(t_{ij}) | c_i = g, b_i; \theta)] \Pr(b_i | c_i = g; \theta) db_i\} =$$

$$= \sum_{i=1}^{n} \log\{\sum_{g=1}^{G} \Pr(c_i = g;\theta) h_i(T_i|c_i = g;\theta)^{\delta_i} S_i(T_i|c_i = g;\theta) \Pr(y_i|c_i = g;\theta)\}.$$

The integrals above can be computed in closed form. $\Pr(y_i|c_i = g, b)$ and $\Pr(b_i|c_i = g)$ lead to a multivariate Gaussian distribution under the assumption of normality, however $\ell(\theta)$ has multiple maxima which leads to the need to fit the model many times with different number of classes, since the number of classes in not known apriori.

4.2.3 Competing Risks

The techniques presented above try to capture the association between the true longitudinal marker level (or history) and the risk for an event. What if we wanted to distinguish the types of events that we anticipate happening? For example we have a patient with one tumor in his lungs and heart desease. Chemotherapy greatly affects the body of the patient and as a result his heart, so it would be standard to monitor both risks for an event. We want to explore the association between a longitudinal marker and the patients risk for death but also want to distinguish the risk induced by cancer and heart desease. To handle such a situation we introduce the Competing Risks model as an option for the survival submodel of the Joint model (Elashoff et al. 2008).

Assuming k types of events, we let $T_{i1}^*, \ldots, T_{ik}^*$ be the true occurence time of these events for the ith subject, $T_i = \min\{T_{i1}^*, \ldots, T_{ik}^*, C_i\}$ the observed event time for the ith subject with C_i denoting censoring. Let $\delta_i \in \{0, 1, \ldots, K\}$ with 0 corresponding to censoring and $1, \ldots, K$ denoting the event type. For each of the k cases we postulate a standard relative risk model in similar fashion to the standard

approach so that we have:

$$h_{ik}(t) = h_{0k}(t) \exp\{\gamma_k^T w_i + \alpha_k m_i(t)\},$$

with everything being the same as in the standard approach just for the kth case.

The longitudinal submodel remains unchanged so the likelihood becomes :

$$\Pr(T_i, \delta_i | b_i; \theta_t, \beta) = \prod_{k=1}^{K} [h_{0k}(T_i) \exp\{\gamma_k^T w_i + \alpha_k m_i(T_i)\}]^{I(\delta_i = k)} \times$$

$$\times \exp\{-\sum_{k=1}^{K} \int_0^{T_i} h_{0k}(s) \exp\{\gamma_k^T w_i + \alpha_k m_i(s)\} ds\}.$$

In a sense we have k different event types modeled with k different relative risk models that only the first one to occur for every subject contributes to the likelihood for the event occurence but all of them contribute for the time under risk.

4.2.4 Recurrent Events

Suppose that we want to quantify the association of the true level of a longitudinal marker with the risk for an event that is recurrent. Strokes for example. Strokes are brain attacks, they occur when the blood supply to the brain becomes blocked. A stroke is a medical emergency that needs immediate medical attention. If we wanted to explore the association of a longitudinal marker with the risk of suffering another stroke we would use a Recurrent Events model for our survival submodel of the Joint model (Liu , Huang 2009). This class of models aims to capture the association between the true level of a longitudinal marker and the risk for an event of a recurrent event.

We first distinguish events into two types, recurrent and terminal. Recurrent events are events that will not kill the subject, in our example a Stroke that will not kill the patient and terminal events are events that will kill the subject, in our example a Stroke that will kill the subject. We then postulate two separate Relative Risk models one of the recurrent and one for the terminal event.

Let U_{ik}, $k = 1, 2, 3, \ldots, k_i$ denote the recurrent event times for the ith subject.
d_{ik} : the indicator of the kth event for the ith subject.
T_i : the terminal event time for the ith subject with $T_i = \min(T_i^*, C_i)$ defined as above.
where T_i^* : the true terminal event time for the ith subject.
C_i : the censoring time for the ith subject.
$\delta_i = I(T_i^* \leq C_i)$: the terminal event indicator for the ith subject.

We postulate two standard relative risk models, one for the recurrent event and one for the terminal event :

$$r_i(t) = r_0(t) \exp\{\gamma_r^T w_{ri} + \alpha_r m_i(t) + v_i\}$$

$$h_i(t) = h_0(t) \exp\{\gamma_h^T w_{hi} + \alpha_h m_i(t) + \zeta v_i\},$$

where w_{ri} : the covariates affecting the risk for a recurrent event for the ith subject.
w_{hi} : the covariates affecting the risk for the terminal eventfor the ith subject.
γ_r : regression coefficients for a recurrent event.
γ_h : regression coefficients for the terminal event.
α_r : the measure of strenght between the current level of the true longitudinal marker and the risk for a recurrent event.
α_r : the measure of strenght between the current level of the true longitudinal marker

and the risk for the terminal event.

v_i : a random effect that accounts for the correlation between recurrent events.

ζ : measure of strenght between recurrent and terminal event process .

The $r_i(t), h_i(t), y_i(t)$ are assumed independent given the $\{b_i, v_i\}$. Recurent event times for subject i are assumed independent given v_i. Longitudinal responses between subjects are assumed independent given b_i.

Formally :

$$\Pr(T_i, \delta_i, U_i, d_i, y_i | b_i, v_i; \theta) = \Pr(T_i, \delta_i | b_i, v_i; \theta) \Pr(U_{ik}, d_{ik} | b_i, v_i; \theta) \Pr(y_i | b_i; \theta),$$

with $\Pr(U_i, d_i | b_i, v_i; \theta) = \prod_k \Pr(U_{ik}, d_{ik} | b_i, v_i; \theta)$,

and $\Pr(y_i | b_i; \theta) = \prod_j \Pr(y_{t_{ij}} | b_i; \theta)$

where $U_1^T = (U_{i1}, U_{i2}, \ldots, U_{ik})$ and $d_1^T = (d_{i1}, d_{i2}, \ldots, d_{ik})$.

So the likelihood contribution of the ith subject is :

$$\Pr(T_i, \delta_i, y_i; \theta) = \int \Pr(T_i, \delta_i | b_i, v_i; \theta) \Pr(U_{ik}, d_{ik} |; \theta) \Pr(y_i | b_i; \theta) \Pr(b_i; \theta) db_i$$

where :

$$Pr(T_i, \delta_i | b_i, v_i, \theta_t; \beta) = h_i(T_i | M_i(T_i), n_i; \zeta, \theta_t, \beta)^{\delta_i} S_i(T_i | M_i, v_i; \zeta, \theta_t, \beta) =$$

$$= [h_0(T_i) \exp\{\gamma_h^T w_i + \alpha_h m_i(T_i) + \zeta n_i\}]^{\delta_i} \times$$

$$\times \exp\{- \int_0^{T_i} h_0(s) \exp\{\gamma_h^T w_i + \alpha_h m_i(s) + \zeta n_i\} ds\}.$$

$$\Pr(y_i|b_i;\theta_y)\Pr(b_i;\theta_b) = \prod_j \Pr(y_i(t_{ij})|b_i;\theta_y)\Pr(b_i;\theta_b) =$$

$$= (2\pi\sigma^2)^{n_i/2}\exp\{-||y_i - X_i\beta - Z_i b_i||^2/2\sigma^2\}\times$$

$$\times(2\pi)^{-q_b/2}\det(D)^{-1/2}\exp\{-b_i^T D^{-1} b_i/2\},$$

$$\Pr(U_{ik}, d_{ik}|v_i;\theta) = \prod_{k=1}^{K_i}[r_0(U_{ik})\exp\{\gamma_r^T w_{ri} + \alpha_r m_i(U_{ik}) + v_i\}]^{\delta_{ik}}\times$$

$$\times\exp\{-\int_0^{U_{ik}} r_0(s)\exp\{\gamma_r^T w_{ri} + \alpha_r m_i(U_{ik}) + v_i\}ds\},$$

$$\Pr(U_{ik}, d_{ik}|;\theta) = \int \Pr(U_{ik}, d_{ik}|v_i;\theta)\Pr(v_i;\theta)dv_i.$$

As a last step we need to specify an appropriate distribution for the n_is. The standard choice would be to assume that $\log(v_i) \sim Gamma$ with mean 1 and variance σ_n. This choice leads to closed form for the marginal distribution of the event times on the separate analysis context. The integral with respect to time becomes:

$$\int_0^{U_{ik}} r_0(s)\exp\{\gamma_r^T w_{ri} + \alpha_r^T m_i(s) + v_i\}ds.$$

The integral with respect to v_i doesn't have an analytical solution and requires numerical integration. This is a problem that we can overcome with reparametrization of the random effects :

$$r_i(t) = r_0(t)\exp\{\gamma_r^T w_{ri} + \alpha_r^T b_i + v_i\}.$$

The integral with respect to time becomes :

$$\int_0^{U_{ik}} r_0(s)\exp\{\gamma_r^T w_{ri} + \alpha_r^T b_i(t) + v_i\}ds =$$

$$= R_0(U_{ik})\exp\{\gamma_r^T w_{ri} + \alpha_r^T b_i(t) + v_i\},$$

where $R(\cdot)$ is the baseline cummulative hazard funtion fot the recurrent event. This leads to a marginal distribution of the recurrent event process specified as:

$$\Pr(U_{ik}, d_{ik}|u_i;\theta) = \frac{\Gamma(d_i+\theta)}{\sigma_v^\theta \Gamma(\theta)} \frac{\prod_k [r_0(U_{ik}) \exp\{\gamma_r^T w_{ri} + \alpha_r^T b_i + v_i\}]^{d_{ik}}}{[\theta + \sum_k R_0(U_{ik}) \exp\{\gamma_r^T w_{ri} + \alpha_r^T b_i(t) + v_i\}]^{d_i+\theta}},$$

with $d_i = \sum_k d_{ik}$ and $\theta = \frac{1}{\sigma_v}$.

Extensions of the Standard Joint model can be easily applied in the Recurrent Events setting.

All the extensions presented above can be combined. This makes the Joint model extremely versatile while holding its interpretability. It is the full likelihood approach that enables us to do so. Although a liability at a first glance, as makes the estimation process more complicated, full likelihood proves to be this methods biggest attribute.

Chapter 5

Diagnostics

As in every statistical model Analysis of residuals gives us insight on the correctness of the model specification and for ways to better fit the data if some component of the model is not correctly specified. In a Joint model every submodel has it's own residuals, hence we have the residuals for the longitudinal submodel and the residuals for the survival submodel. For a Joint model to be correctly specified we need both of these types of residuals to fulfill certain properties which will be presented bellow.

5.1 Residuals for the Longitudinal Submodel

There are two types of residuals for the longitudinal submodel, the subject specific residuals and the marginal residuals.

Subject Specific Residuals The Subject Specific Residuals aim to validate assumptions of the hierarchical version of the model :

$$Y_i = X_i\beta + Z_i b_i + e_i$$

$$b_i \sim N(0, D), e_i \sim N(0, \sigma^2)$$

The Subject Specific Residuals are defined as:

$$r_i^{y_s}(t) = \{y_i(t) - x_i^T \hat{\beta} - Z_i^T(t)\hat{b}_i\},$$

where $\hat{\beta}$ is the Maximum Likelihood Estimator and $\hat{b}_i$ the empirical Bayes estimator. The residuals are then standardised :

$$r_i^{y_{ss}} = \frac{r_i^{y_s}(t)}{\hat{\sigma}}$$

where $\hat{\sigma}$ is the Maximum Likelihood Estimator of the Standard Error.

These residuals predict the conditional errors $e_i(t)$, and can be used for checking the homoscedasticity and normality assumptions. White noise in the Residuals vs Fitted plot implies correct model specification for the hierarchical version of the model, the same does a diagonal for the Standardized Vs Theoretical Q-Q plot.

Marginal Residuals The main focus here is the Marginal model implied :

$$Y_i = X_i\beta + e_i^\star$$

where $e_i^\star \sim N(0, Z_i D Z_i^T + \sigma^2 I_{n_i})$. The residuals are defined as :

$$r_i^{y_m} = y_i - X_i\hat{\beta},$$

with the standardized version :

$$r_i^{y_{sm}} = \hat{V}_i^{-1/2}(y_i - X_i\hat{\beta}),$$

where $\hat{V}_i = Z_i \hat{D} Z_i^T + \sigma^2 I_{n_i}$ is the estimated marginal Covariance matrix of y_i.

Marginal residuals predict Marginal errors ($y_i - X_i\beta = Z_i b_i + e_i$) that can be used to check misspecification of the mean structure $X_i\beta$ and to validate the within subject covariance matrix structure V_i. White noise in the Residuals vs Fitted plot implies correct model specification for the mean, the same does a diagonal for the Standardized Vs Theoretical Q-Q plot.
The marginal survival function can be derived by integrating out the random terms b_i and can be estimated using the approximation presented below:

$$S(t) = \int S_i(t|b_i;\hat{\theta})\Pr(b_i;\hat{\theta})db_i \approx n^{-1}\sum S_i(t|\hat{b}_i;\theta)$$

and therefore the estimation for the marginal cumulative hazard function is:

$$H(t) = -\log(S(t)).$$

These quantities are usually being plotted to get an overall idea about the behaviour of the sample.

5.2 Residuals for the Survival Part

Two types of residuals from the ones that are being presented in the first chapter will be used :

- Martingale Residuals
- Cox-Snell Residuals

Martingale Residuals for the Survival Part

The first type of residuals that is being used for the survival part is the Martingale residuals. The Martingale residual for the ith subject is:

$$r_i^{tM}(t) = N_i(t) - \int_0^t R_i(s)h_i(s|\hat{M}_i(s);\hat{\theta})ds =$$

$$= N_i(t) - \int_0^t R_i(s)\hat{h_0}(s)\exp\{\hat{\gamma}^T w_i + \hat{\alpha}\hat{m}_i(s)\}ds,$$

where $R_i(t)$ is left continuous, $R_i(t) = 1$ if the ith subject is at risk and $R_i(t) = 0$ otherwise.

$N_i(t)$ is the counting process for the ith subject.

$\hat{h}_0(t)$, the estimated baseline hazard function and $\hat{m}_i(s) = x_i^T(s)\hat{\beta} + z_i^T(s)\hat{b}_i$.

Martingale residuals $r_i^{tM}(t)$ can be viewed as the difference between the observed number of events and the expected by the model number of events at any given time t. This type of residuals are mainly used for identification of excess events and evaluation the functional form for a covariate of interest in the model.

Under certain conditions, the scatterplot of martingale residuals from a model versus a predictor of interest can reveal it's true functional form. Our predictor of interest is the longitudinal outcome so we will plot the martingale residuals against the subject specific fitted values of the longitudinal outcome. A null horizontal line implies that the functional form chosen in the model is correct.

5.2.1 Cox-Snell Residuals for the Survival Part

The second type of residuals that is being used for the survival part of the Joint model is the Cox-Snell residuals. These are calculated as the value of the estimated cumulative risk function for the ith subject evaluated at his observed event time T_i :

$$r_i^{tcs}(t) = \int_0^{T_i} h_i(s|\hat{M}_i(s); \hat{\theta})ds =$$

$$= \int_0^{T_i} \hat{h}_0(s) \exp\{\hat{\gamma} w_i + \hat{\alpha}\hat{m}_i(s)\}ds,$$

hence : $r_i^{tcs} = N_i(T_i) - r_i^{tM}$.

When the assumed Joint model fits the data well $S(t) \sim U(0,1) \Rightarrow H(t) = -\log(S(t)) \sim Exp(1)$.

As a result we can check for model fit by plotting r_i^{tcs} against the unit exponential distribution. An issue with this approach arises when T_i is censored and as a result the corresponding residual is censored. To overcome this issue we check goodness of fit by comparing the survival function of the unit exponential distribution against the Kaplan-Meyer estimate of the survival function of the r_i^{tcs}. Deviation of the two functions implies weak data fit .

5.3 Residuals and Dropout

Most of the time in the Joint modeling setting a systematic trend in the residuals of the Joint model is observed even if the model specification is correct. The reason this happens is that the dropout mechanism is non random. The implication of the nonrandom nature of the dropout mechanism is that the observed data, upon which the residuals are calculated, do not constitute a random sample of the target population. In the process described until now we analyse only the observed data and this usually leads to misleading diagnostics plots. To overcome this issue we augment the data with randomly imputed data under the complete data model, corresponding to the longitudinal outcome of the patients had they not dropped out (Rizopoulos et al. 2010).

The procendure is carried out as follows :
1) Missing values are filled in, M times to generate M datasets.
2) The complete datasets are being analyzed and parameters are being estimated using standard methods for the Joint modelling setting.
3) Results from the M analyses are combined to produce a single estimation and draw inference.

One important question that arises in the procendure described above is when these missing measurements took place. To adress this problem we will construct a suitable model for the visiting process and use it to generate "visit times" after dropout.

5.3.1 The visiting process

We assume that all subjects visit at least once and we let u_{ik} where $(k = 2, 3, \ldots, n)$ denote the time elapsed between visit $k-1$ and k for the ith subject. Let $Y_i^\star$ be the complete longitudinal responce vector. Under the above and the non-informativeness assumption we have:

$$\Pr(u_{ik}|u_{i2}, \ldots, u_{i(k-1)}, Y_i^\star; \theta_u) = \Pr(u_{ik}|u_{i2}, \ldots, u_{i(k-1)}, y_i(t_1), \ldots, y_i(t_{k-1}); \theta_u),$$

where θ_u is the parameter vector of the visiting process and $\{\theta, \theta_u\}$ have disjoint parameter spaces,
$u_i^T = (u_{i2}, \ldots, u_{in_i})$ are the times elapsed between each visit and are correlated for each patient.

We need to fully specify $\Pr(u_{ik}|u_{i1}, \ldots, u_{i(k-1)}, y_i(t_1), \ldots, y_i(t_{k-1}); \theta_u)$, hence a conditional model is appropriate.
We will use a Weibull model with a multiplicative Gamma distribution frailty term defined as:

$$\lambda(u_{ik}|x_{ui}, w_i) = \lambda_0(u_{ik}) w_i \exp\{x_{ui}^T \gamma_u\},$$

where, $w_i \sim Gamma(\sigma_w, \sigma_w)$,
$\lambda(\cdot)$: the risk function conditional on the frailty term w_i,
x_{ui}: the covariate vector that may or may not contain a functional form of $y_i(t_{i1}), \ldots, y_i(t_{i(k-1)})$,
γ_u: the regression coefficient vector,
$\lambda_0(\cdot)$: baseline risk function (Weibull)
σ_w^{-1} : the unknown variance of the w_i's,
t_{max}: the end of the study,

We use this model because it best trades off richness for computional complexity and the posterior distribution of the frailty term, given the observed data, is of standard form . We will use the scheme above to generate times between visits in the multiple imputation algorithm.

5.3.2 Multiple Imputation in Joint Models

Let : y^o be the observed and y^m the missing longitudinal responce vector. We will be multispampling from the posterior distribution of y^m given the observed data averaged over the posterior distribution of the parameters.The density for this distribution can be expressed as:

$$\Pr(y_i^m|y_i^o, T_i, \delta_i) = \int \Pr(y_i^m|y_i^o, T_i, \delta_i; \theta) \Pr(\theta|y_i^o, T_i, \delta_i) d\theta,$$

with:

$$\Pr(y_i^m|y_i^o, T_i, \delta_i; \theta) = \int \Pr(y_i^m|y_i^o, T_i, \delta_i, b_i; \theta) \Pr(b_i|y_i^o, T_i, \delta_i; \theta) db_i =$$

$$= \int \Pr(y_i^m|b_i; \theta) \Pr(b_i|y_i^o, T_i, \delta_i; \theta) db_i,$$

where $\delta_{u,ik}$ is the event indicator that corresponds to u_{ik}.

For the posterior distribution of the parameters given the observed data, we use arguments of standard asymptotic Bayesian theory and assume that the size is again sufficiently large for $\{\theta|y_i^o, T_i, \delta_i\}$ to be approximated by $N(\hat{\theta}, \mathrm{Var}(\hat{\theta}))$ and for $\{\theta_u|y_i^o, T_i, \delta_i\}$ to be approximated by $N(\hat{\theta_u}, \mathrm{Var}(\hat{\theta_u}))$. $\hat{\theta}, \hat{\theta_u}, \mathrm{Var}(\hat{\theta}), \mathrm{Var}(\hat{\theta_u})$ are the

maximum likelihood estimators and their corresponding variances respectivelly.

The idea behind the scheme is simple and standard in Bayesian Statistics.

In Step 1 we draw the θ and θ_u,θ will be used to draw from the poterior distribution of the random effects and θ_u will be used to draw a visiting time. Their distribution is normal so we can draw them using a standard Gibbs scheme.

In Step 2 we draw the frailty terms and the random effects. Frailty terms have a Gamma distribution so we can again draw them using a Gibbs sampler. The distribution of the random effects is not known though so a Metropolis-Hastings algorithm needs to be implemented.

In Step 3 we draw the next visit time. Then we draw a longitudinal outcome on that time if it is before the timepoint that the study ends .The distributions of the u_i and y_is are Weibull and Normal respectively so we can again use a Gibbs sampler .

When Gibbs sampler is used, the update mechanism is straightforward. For the Metropolis-Hastings algorithm in order to generate the $b_i^{(\ell)}$s, we propose from idependent multivariate t distribution, centered at $\hat{b}_i$, with scale matrix : $\hat{\text{Var}}(\hat{\beta}_i)$ and four degrees of freedom.

The simulation scheme is :

Step 1:

$$\text{Draw}: \ \theta_u^{(\ell)} \sim N(\hat{\theta}_u, \hat{\text{Var}}(\hat{\theta}_u))$$

$$\text{Draw}: \ \theta^{(\ell)} \sim N(\hat{\theta}, \hat{\text{Var}}(\hat{\theta}))$$

Step 2:

$$\text{Draw}: \ w_i^{(\ell)} \sim Gamma(\sigma_w^{(\ell)}, \sigma_w^{(\ell)})$$

If the subject visits more that once then:

$$\text{Draw}: \ w_i^{(\ell)} \sim Gamma(A, B)$$

$$\text{Where}: \ A = \sigma_w^{\ell} + \sum_{k=2}^{n_i} \delta_{u,ik}, \ \text{and} \ B = \sigma_w^{(\ell)} + \phi^{(\ell)} \sum_{k=2}^{n_i} u_{ik}^{\psi(\ell)} \exp(x_{ui}^T \gamma_u^{(\ell)})$$

$$\text{Draw}: \ b_i^{(\ell)} \sim \{b_i | y_i^o, T_i, \delta_i, \theta^{(\ell)}\}$$

Step 3: $\text{Draw}: \ u_i^{(\ell)} \sim \text{Weibull}\{\psi^{(\ell)}, \phi^{(\ell)} w_i^{(\ell)} \exp(x_{ui}^T \gamma_u^{(\ell)})\}$

$$\text{Set}: \ \tilde{t}_i = u_i^{(\ell)} + t_{in_i}$$

Where t_{in_i} is the last observed visit time for the ith subject.

$$\text{If}: \ \tilde{t}_i > t_{max} \ \text{then no} \ \ y_i^m \ \text{generation} .$$

$$\text{Else , set:} \ \ m_i^{(\ell)}(\tilde{t}_i) = x_i^T(t_{in_i})\hat{\beta}^{(\ell)} + z_i^T(\tilde{t}_i)\hat{\beta}^{(\ell)} \ \ \text{and}$$

$$\text{draw:} \ \ y_i^{m(\ell)}(\tilde{t}_i) \sim N(m_i^{(\ell)}(\tilde{t}_i), (\hat{\sigma}^{(\ell)})^2)$$

$$\text{Set } t_{in_i} = \tilde{t}_i \ \text{ and repeat until } \ \ t_{in_i} > t_{max} \ \text{ for all } i.$$

If $\Pr(u_{ik}|u_{i2},\ldots,u_{i(k-1)},Y_i^\star;\theta_u) = \Pr(u_{ik}|u_{i2},\ldots,u_{i(k-1)},y_i(t_1),\ldots,y_i(t_{k-1});\theta_u)$, is violated the models does not provide valid inferences.

A stronger and more plausible assumption is :

$$\Pr(u_{ik}|u_{i2},\ldots,u_{i(k-1)},Y_i^\star;\theta_u) = \Pr(u_{ik}|u_{i2},\ldots,u_{i(k-1)},y_i(t_{k-1});\theta_u),$$

and

$$\Pr(u_{ik}|u_{i2},\ldots,u_{i(k-1)},Y_i^\star;\theta_u) = \Pr(u_{ik}|y_i(t_{k-1});\theta_u),$$

which implies that the doctor decides the next visit time , based on the last measurement.

5.3.3 Distribution of Random Effects

The final assumption we need to check is that of the distribution of the random effects. The non-random dropout makes the joint model sensitive to misspecification of the random effects. Semi-parametric techniques have been proposed in order to make the specification more flexible and widen our choices, something that comes at a great computional cost. It has been proven though by Rizopoulos et al. 2008 and Huang et al. 2009 that as n_i increases the effect of distributional misspecification diminishes and hence the model becomes robust, a very usefull result as big data is becoming common in clinical research studies.

Chapter 6

Prediction

One of the main reasons we build a statistical model is to make predictions. In the Joint modeling setting predictions produced are personalized and dynamic. Dynamic in the sense that predictions for every subject are time dependent, as time passes by and new information arise the model uses this information and updates its prediction. Personalized in the sense that every subject has its own distinct prediction for the longitudinal response and every survival quantity of interest.

Prediction in Joint models consists of two parts, Predictions for the Survival Probabilities of the subjects and Predictions for the Longitudinal Outcomes of the subjects. Quality of the predictions will be assesed based on Receiver Operating Characteristic (ROC) discrimination measures which will be presented too.

6.1 Predicting the Survival Probability

Let $D_n = \{T_i, \delta_i, y_i, i = 1, 2, \ldots, n\}$, $Y_i(t) = \{y_i(s) : 0 \leq s < t\}$ and w_i be the baseline covariates vector for the ith patient.

As presented above $y_i(t)$ is directly related to the failure mechanism so it would be reasonable to focus on the conditional probability of survival beyond a timepoint $u > t$ given survival until t. We integrade with respect to the random effects as usual and we get (conditioning on the covariates w_i is assumed but omitted from the notation):

$$\pi(u|t) = \Pr(T_i^\star \geq u | T_i^\star > t, Y_i(t), w_i, D_n; \theta^\star) =$$

$$= \int \Pr(T_i^\star \geq u | T_i^\star > t, Y_i(t), b_i; \theta) \Pr(b_i | T_i^\star > t, Y_i(t); \theta) db_i =$$

$$= \int \Pr(T_i^\star \geq u | T_i^\star > t, b_i; \theta) \Pr(b_i | T_i^\star > t, Y_i(t); \theta) db_i =$$

$$= \int \frac{S_i\{u | M_i(u, b_i, \theta); \theta\}}{S_i\{t | M_i(u, b_i, \theta); \theta\}} \Pr(b_i | T_i^\star > t, Y_i(t); \theta) db_i,$$

where M_i is the longitudinal history as estimated by the linear mixed effects model with the Empirical Bayes estimates being used for the random effects.
θ^* : the true parameter values.
$S(\cdot)$: the Survival function.

We estimate the conditional probability of survival beyond $u > t$ given survival until t as:

$$\hat{\pi}(u|t) = \frac{S_i\{u | M_i(u, \hat{b}_i^{(t)}, \hat{\theta}); \hat{\theta}\}}{S_i\{t | M_i(t, \hat{b}_i^{(t)}, \hat{\theta}); \hat{\theta}\}} + O([n_i(t)]^{-1}),$$

where $\hat{\theta}$ is the vector with the maximum likelihood estimations of the fixed effects, $\hat{b}_i^{(t)}$ the mode of the posterior distribution of the b_i conditioned on survival until time-point t and $n_i(t)$ the number of the logitudinal responses by time t for the ith patient.

The estimator above works really well but implications arise when trying to derive it's standard error and hence the construction of confidence intervals. In order to overcome this issue, a Markov Chain Monte Carlo algorithm can be used.

The posterior expectation of $\pi(u|t)$ can be derived as :

$$\Pr(T_i^\star \geq u | T_i^\star > t, Y_i(t), D_n) =$$

$$\int \Pr(T_i^\star \geq u | T_i^\star > t, Y_i(t); \theta) \Pr(\theta | D_n) d\theta.$$

The fist part of the integrand as shown above is given by:

$$\Pr(T_i^\star \geq u | T_i^\star > t, Y_i(t), D_n) =$$

$$= \int \frac{S_i\{u | M_i(u, \hat{b}_i^{(t)}, \hat{\theta}); \hat{\theta}\}}{S_i\{t | M_i(t, \hat{b}_i^{(t)}, \hat{\theta}); \hat{\theta}\}} + O([n_i(t)]^{-1}).$$

We assume that n is sufficient enough for the $\{\theta | D_n\}$ to be approximated by $N(\hat{\theta}, \mathrm{Var}(\hat{\theta})\}$. Using these results we can postulate simulation scheme to estimate $\pi(u|t)$:

Step 1:

$$\text{Draw}: \ \theta^{(\ell)} \sim N(\hat{\theta}, \text{Var}(\hat{\theta}))$$

Step 2:

$$\text{Draw}: \ b_i^{(\ell)} \sim \{b_i | T_i^{\star} > t, Y_i(t); \theta^{(\ell)}\}$$

Step 3:

$$\text{Compute}: \ \pi_i^{(\ell)}(u|t) \sim \frac{S_i\{u|M_i(u, b_i^{(\ell)}, \theta^{(\ell)}); \theta^{(\ell)}\}}{S_i\{t|M_i(t, b_i^{(\ell)}, \theta^{(\ell)}); \theta^{(\ell)}\}}$$

with $\ell = 1, 2, \ldots, L$. We are using $b_i^{\ell}, \theta^{\ell}$ to account for the Bayes estimates and Maximum Likelihood uncertainty .

In Step 2 the b_is are being drawn using a Metropolis-Hastings algorithm with independent proposals from the Student-t distribution with four degrees of freedom centered at the Empirical Bayes estimates $\hat{b}_i^{(t)}$ and scale matrix:

$$\hat{\text{Var}}(\hat{b}_i^{(t)}) = \{-\partial^2 \log \Pr(T_i^{\star} > t, Y_i(t), b; \hat{\theta}) / \partial b^T \partial b|_{b=\hat{b}_i(t)}\}^{-1}.$$

The point estimates used are median and mean and noted as $\tilde{\pi}_i(u|t)$.

6.2 Predicting the Longitudinal outcome

Sometimes prediction of the longitudinal responce is of interest. In most cases when the level of the longitudinal response exceeds a certain threshold a differect treatment should be implemented or the risk for an event becomes too high. That makes a prediction of the longitudinal responce very valuable. Such predictions can be estimated in similar fashion as the predictions for the survival process. Predictions are again dynamic and personalized.

Suppose subject i is alive at timepoint t and we want to predict the value of his longitudinal outcome at timepoint $u > t$ given his longitudinal history $Y_i(t)$:

$$w_i(u|t) = E(y_i(u)|T_i^\star > t, Y_i(t), D_n; \theta^\star)$$

The parameter $\theta^\star$ is unknown so we proceed as above:

$$E(y_i(u)|T_i^\star > t, Y_i(t), D_n) =$$
$$= \int E(y_i(u)|T_i^\star > t, Y_i(t); \theta) \Pr(\theta|D_n) d\theta,$$

where :

$$E(y_i(u)|T_i^\star > t, Y_i(t); \theta) =$$
$$= \int E(y_i(u)|T_i^\star > t, Y_i(t), b_i; \theta) \Pr(b_i|T_i^\star > t, Y_i(t); \theta) db_i =$$
$$= \int E(y_i(u)|b_i) \Pr(b_i|T_i^\star > t, Y_i(t); \theta) db_i =$$
$$= \int (x_i^T(u)\beta + z_i^T(u)b_i) \Pr(b_i|T_i^\star > t, Y_i(t); \theta) db_i = x_i^T(u)\beta + z_i^T(u)\tilde{b}_i^t,$$

where $\tilde{b}_i^t = \int \Pr(b_i|T_i^\star > t, Y_i(t); \theta) db_i$.

Although the estimator of $w_i(u|t)$ is obtained by replacing θ with $\hat{\theta}$ and calculating the mean of the posterior distribution of the random effects the same problem arises as above. To overcome this issue we construct an MCMC algorithm to simulate a sample from which we can compute point estimates and construct confidence intervals.

Again we assume the sample sufficiently large sample in order for $\{\theta|D\}$ to be approximated by a normal distribution centered at $\hat{\theta}$ with covariance matrix $\hat{\text{Var}}(\hat{\theta}) = \{I(\hat{\theta})\}^{-1}$ the inverse or the observed information matrix.
The simulation scheme is :

Step 1:

$$\text{Draw : } \theta^{(\ell)} \sim N(\hat{\theta}, \text{Var}(\hat{\theta}))$$

Step 2:

$$\text{Draw : } b_i^{(\ell)} \sim \{b_i|T_i^{\star} > t, Y_i(t); \theta^{(\ell)}\}$$

Step 3:

$$\text{Compute : } w_i^{(\ell)} = x_i^T \beta^{(\ell)} + z_i^T b_i^{(\ell)}$$

Steps 1,2 are simulated as in the algorithm used for the prediction of the conditional probability of survival and are used to account for the variability in $\hat{\theta}$,$\hat{b}_i^{(t)}$.

Step 3 calculates the predicted value of the longitudinal outcome $y_i(u)$. Confidence intervals can be derived using the 2.5th and 97.5th percentile of the $\{w_i^{(\ell)}(u|t), \ell = 1, 2, \ldots, L\}$. The scheme can be modified to produce prediction intervals. To achieve this we substitute the $w_i^{(\ell)} = x_i^T\beta^{(\ell)} + z_i^T b_i^{(\ell)}$ in Step 3 with $w_i^{(\ell)} \sim N(x_i^T\beta^{(\ell)} + z_i^T b_i^{(\ell)}, [\sigma^{(\ell)}]^2)$. Estimations can be derived instead using the mean or the median of the sample ganarated by the scheme.

6.3 Assessing Prediction Accuracy

Different models and parametrizations wield different predictions. The problem that arises is: Which model do we choose for prediction assessment?

There are two factors that affect the quality of a prediction, first is the capacity of the longitudinal marker to predict future events and second the correct formulation of the Joint model in order to reveal the true predictive performance of the longitudinal marker. Studies have shown that although prediction for the longitudinal outcome is somewhat stable as the model changes, the same is not true for the prediction of the survival probability. Information criteria and likelihood ratio tests are always available, these methods rank the model fit but not the predictive potential of the model. A standard method for assessing how well the longitudinal marker can discriminate between patients with high and low probability is the Receiver Operating Characteristic (ROC).

6.3.1 The Receiving Operating Characteristic Curve

The Receiving Operating Characteristic Curve (ROC-Curve) is used to visualize the tradeoff between clinical sensitivity and specificity for every possible tresshold set.

Let d_i be the desease status of patient i where $d_i = 1$ if the patient is deseased and $d_i = 0$ otherwise.

Let y_i be the longitudinal response.

We set an arbitrary thresshold c where if $y_i > c$ the subject is considered diseased.

The probability of a true positive (correct classification of a diseased patient) is:

$$TP(c) = \Pr(y_i > c)|d_i = 1)$$

with : $\Pr(y_i > c)|d_i = 1) = 1 - FP(c)$.

where : $FP(c) = \Pr(y_i > c|d_i = 0)$.

We call TP sensitivity and 1-FP specificity.

The ROC-Curve is the plot of Sensitivity against 1- Specificity for all possible levels of c formally defined as :

$$ROC(p) = TP\{FP^{-1}(p)\}$$

where : $FP^{-1}(p) = inf_c\{c : FP(c) \leq p\}$ and p $\in [0, 1]$.

A summary of the predictive accuracy of the model for all the possible thresshold values is given by:

$$AUC = \int_0^1 ROC(p)dp$$

which in the area under the $ROC(p)$ curve. AUC will be between zero and unity. Higher levels of AUC indicate higher predictive accuracy of the model.

6.3.2 Discrimination Measures for Survival outcomes

Following the rationale presented in the section explaining the ROC-Curve we view the event as a time dependent binary outcome. A plethora of methodologies have been proposed but the idea behind them is the same so one of them is presented.

Let $N_i(t) = I(t \geq T_i^\star)$ be the counting process of the true event times.
we denote:

$$TP_t^C(c): \quad \Pr(y_i > c | T_i^\star \leq t), \text{ and}$$

$$1 - FP_t^D(c): \quad \Pr(y_i \leq c | T_i^\star > t).$$

At any given timepoint t the entire population is classified as either a case or a control according to their event status. Control for $t < T_i^\star$ and case for $t \leq T_i^\star$.
Sensitivity measures the fraction of diseased subjects among the patients that suffered an event at time t. Specificity measures the fraction of non deseased subjects among those who survive time t.

In order to assess the predictive capability of the model we compute $TP_t^C(c)$ and $FP_t^D(c)$ and draw to coresponding ROC-Curve. As a last step we compute the AUC and compare it to the AUCs of the other model candidates.

6.3.3 Discirimination Measures for the Longitudinal marker

It would be usefull for us to be able to assess the value of a prediction for the longitundinal marker as correct prediction of the longitudinal marker can lead us to information about the future risk for an event of a subject.

Following the notation used above, let: $\mathrm{P}_i^S(t,k,c) = \{y_i(s) \geq c_s; k \leq s \leq t\}$ be the instances at which the marker indicates that an event will occur (success) and $\mathrm{P}_i^f(t,k,c) = \mathrm{R}^{r(k,t)}/\{y_i(s) \geq c_s; k \leq s \leq t\}$ be the instances at which the marker indicates that an event will not (failure), where $r(k,t)$ is the number of longitundinal measurements taken in $[k,t]$.

The rule upon which we treat an outcome or a set of outcomes as a sucess should be based upon a phisicians suggestion. Rules are usually based upon the behaviour of the specific desease. Rules can be simple or not.
We denote :

$$TP_t^{\Delta t}(c) = \Pr\{\mathrm{P}_i^S(t,k,c) | T_i^\star > t, T^\star \in (t, t+\Delta t]; \theta^\star\}$$

and

$$1 - FP_t^{\Delta t}(c) = \Pr\{\mathrm{P}_i^f(t,k,c) | T_i^\star > t, T^\star, T_i > t + \Delta t; \theta^\star\}.$$

Sucesses and failures are not only time dependent, they are also dependent on the lenght of the interval. We proceed in the construction of the ROC-Curve as usual:

$$ROC_t^{\Delta t}(p) = TP^{\Delta t}\{[FP^{\Delta t}]^{-1}\}$$

where $[FP^{\Delta t}]^{-1}\} = \inf_c\{c : [FP^{\Delta t}](c) \leq p\}$. Again we compute the AUC and compare it to the AUCs of other model candidates.

6.3.4 Overall Discrimination

The measures presented above help us rank models based on their ability to predict an event or the longitundinal outcome for a patient. The main goal of this section is to assess a measure that helps us choose a model based on its overall

predictive potency. The measure that will be used can be seen as an extension of the AUC in the binary context.

For two subjects $\{i, j\}$ whose true event times are ordered $T_i < T_j$ we are interested in : $c_n = \Pr(y_i > y_j | T_i^\star < T_j^\star)$.

It has been shown that $c_n = \int_0^\infty AUC_t u(t) dt$ where $u(t) = 2p(t)S(t)$ and $AUC_t = \Pr(y_i > y_j | T_i^\star = t, T_j^\star > t)$.

6.3.5 Discrimination under the Joint modelling framework

As with the residuals ,censoring complicates the estimation of sensitivity, specificity and AUC in the survival setting because if we want to estimate the sensitivity for example at time t but the subject was censored in timepoint $t' < t$, we cannot know if the subject is case or control. So the counting process cannot be carried through.

To overcome this issue we need to estimate the distribution of $\{T_i^\star, y_i\}$ something that can be done in the joint modeling framework with the use of a MCMC algorithm.

We denote :

$$\Pr\{\mathrm{P}_i^S(t,k,c) | T_i^\star > t, T^\star \in (t, t+\Delta t]; \theta^\star\} =$$
$$= \frac{\Pr\{\mathrm{P}_i^S(t,k,c), T^\star \in (t, t+\Delta t] | T_i^\star > t; \theta^\star\}}{1 - \Pr\{T_i^\star > t, T^\star \in (t, t+\Delta t]; \theta^\star\}}$$

where $\theta^\star$ the true parameter vector. For the numerator :

$$\Pr\{\mathrm{P}_i^S(t,k,c), T^\star \in (t, t+\Delta t] | T_i^\star > t; \theta^\star\} =$$

$$= \int \Pr\{\mathrm{P}_i^S(t,k,c), T^\star \in (t, t+\Delta t] | T_i^\star > t, b_i; \theta^\star\} \Pr\{b_i | T_i^\star > t; \theta^\star\} db_i =$$

$$= \int \Pr\{\mathrm{P}_i^S(t,k,c)|b_i;\theta^\star\} \times \Pr\{T^\star \in (t,t+\Delta t]|T_i^\star > t, b_i;\theta^\star\} \times \Pr\{b_i|T_i^\star > t;\theta^\star\} db_i.$$

where:

$$\Pr\{\mathrm{P}_i^S(t,k,c)|b_i;\theta^\star\} = \prod_{s=k}^{t} \Phi\{\frac{c_s - m_i(s,b_i,\beta^\star)}{\sigma^\star}\}$$

$$\Pr\{T^\star \in (t,t+\Delta t]|T_i^\star > t, b_i;\theta^\star\} = 1 - \frac{S_i\{t+\Delta t|M_i(t+\Delta t,b_i);\theta^\star\}}{S_i\{t|M_i(t,b_i);\theta^\star\}}$$

and for the denominator we have :

$$\Pr\{T_i^\star > t+\Delta t|T_i^\star > t;\theta^\star\} = \int \Pr\{T_i^\star > t+\Delta t|T_i^\star > t, b_i;\theta^\star\} \Pr\{b_i|T_i^\star > t;\theta^\star\} db_i =$$

$$= \int \frac{S_i\{t+\Delta t|M_i(t+\Delta t,b_i);\theta^\star\}}{S_i\{t|M_i(t,b_i);\theta^\star\}} \Pr\{b_i|T_i^\star > t;\theta^\star\} db_i.$$

So let :

$$\epsilon_1(b_i;\theta) = [\prod_{s=k}^{t} \Phi\{\frac{c_s - m_i(s,b_i,\beta^\star)}{\sigma^\star}\}][1 - \frac{S_i\{t+\Delta t|M_i(t+\Delta t,b_i);\theta^\star\}}{S_i\{t|M_i(t,b_i);\theta^\star\}}]$$

and

$$\epsilon_2(b_i;\theta) = \frac{S_i\{t+\Delta t|M_i(t+\Delta t,b_i);\theta^\star\}}{S_i\{t|M_i(t,b_i);\theta^\star\}}$$

with respect to $\Pr\{b_i|T_i^\star > t;\theta^\star\}$. We should note that this posterior distribution is not the same as the one used in the derivation of the conditional survival probabilities or in the predictions for the longitudinal outcome.

So if we want to construct an algorithm similar to the ones we constructed above for prediction, first we need to express $\Pr\{b_i|T_i^\star > t;\theta^\star\}$ in terms of $\Pr\{b_i|T_i^\star > t, Y_i(t);\theta^\star\}$.

$$\Pr\{b_i|T_i^\star > t;\theta^\star\} \propto \Pr\{T_i^\star > t|b_i;\theta^\star\} \Pr\{b_i;\theta^\star\} =$$

$$= \int \Pr\{T_i^\star > t, Y_i(t)|b_i;\theta^\star\} \Pr\{b_i;\theta^\star\} dY_i(t) =$$

$$= \int \Pr\{Y_i(t)|b_i;\theta^\star\} S_i\{t|M_i(t,b_i);\theta^\star\} \Pr\{b_i;\theta^\star\} dY_i(t).$$

We assume the sample sufficiently large so that $\{\theta|D_n\}$ is approximated by a normal distribution centered at $\hat{\theta}$ with covariance matrix $\hat{\text{Var}}(\hat{\theta})$.
So the scheme is :

Step 1:

$$\text{Draw : } \theta^{(\ell)} \sim N(\hat{\theta}, \text{Var}(\hat{\theta}))$$

Step 2:

$$\text{Draw : } Y_i^{(\ell)}(t) \sim \{N(x_i\beta^{(\ell)} + Z_i b_i^{(\ell-1)}, [\sigma^{(\ell)}]^2)$$

Step 3:

$$\text{Draw : } b_i^{(\ell)} \sim \{b_i|T_i^\star > t, Y_i^{(\ell)}(t), \theta^{(\ell)}\}$$

Step 4:

$$\text{Compute : } \epsilon_1(b_i^{(\ell)};\theta^{(\ell)}), \epsilon_2(b_i^{(\ell)};\theta^{(\ell)})$$

In Steps 1,2 and 4 distributions are known so Gibbs sampler is used .In Step 3 we again use the Metropolis-Hastings approach.
The sensitivity estimate takes the form:

$$\hat{\Pr}\{\text{P}_i^S(t,k,c)|T_i^\star > t, T^\star \in (t, t+\Delta t];\theta^\star\} = \frac{\sum_\ell \epsilon_1(b_i^{(\ell)};\theta^{(\ell)})}{L - \epsilon_2(b_i^{(\ell)};\theta^{(\ell)})}.$$

With the corresponding standard error estimated using the Monte Carlo standard errors of $\epsilon_1(b_i^{(\ell)};\theta^{(\ell)})$, $\epsilon_2(b_i^{(\ell)};\theta^{(\ell)})$ and the Delta Method:

$$\text{s.e}(\hat{\Pr}\{\mathrm{P}_i^S(t,k,c)|T_i^\star > t, T^\star \in (t, t+\Delta t]; \theta^\star\}) =$$

$$= \{gVg^T\}^{1/2}$$

where:

$$g = L[1/\{1 - \sum \epsilon_2(b_i^{(\ell)};\theta^{(\ell)})\}, \frac{\sum_\ell \epsilon_1(b_i^{(\ell)};\theta^{(\ell)})}{L - \epsilon_2(b_i^{(\ell)};\theta^{(\ell)})}]$$

and

$$\text{vech}(V) = L^{-1}[\text{Var}\{\epsilon_1(b_i^{(\ell)};\theta^{(\ell)})\}, \text{Cov}(\epsilon_1(b_i^{(\ell)};\theta^{(\ell)}), \epsilon_2(b_i^{(\ell)};\theta^{(\ell)})), \text{Var}\{\epsilon_2(b_i^{(\ell)};\theta^{(\ell)})\}].$$

Specificity can be estimated according to the same rationale. Having sensitivity and specificity estimated we can draw the *ROC*-curve and compute the *AUC* without any problem.

Chapter 7

Numerical Results

What follows is an implementation of the procendure presented in the Chapters above in a simple simulated dataset. The aim is purely educational, for that reason certain assumptions have been made that are not realistic in a real life scenario.

The dataset consists of two parts , the longitudinal part and the survival part.

The longitudinal part comes in twelve columns. Id is the id of the patient, y is the longitudinal measurement at the time denoted in the time column. Intercept is a quantity equal across all patients. Ctsxl denotes a baseline measurement for every patient and hence is equal for rows with the same id. Binxl is binary and serves the purpose of treatment. In the survtimel column we can find the survival time of each patient. In the Cens column we have binary observations that denote whether the event was censored or not and finally Event is binary and denotes whether an Event has hapened between current and the next scheduled measurement time. Measurement times are fixed and discrete.

The survival part of the data consists of the columns id, survtimel,cens and binxl. Patients with survival time less that 0.001 have been omited for two reasons,

first , in order to make the dataset more realistic and second, for computional stability reasons. In Figures 7.1 and 7.2 one can find a small sample of the two parts of the data.

	id	y	time	intercept	ctsxl	binxl	ltime	survtimel	cens	start	stop	Event
2	3	10.349	0	1	-0.1635747110	0	0	1.719	1	0	1	1
3	3	21.062	1	1	-0.1635747110	0	1	1.719	1	0	0	1
4	4	10.336	0	1	-0.9829034759	1	0	3.718	1	0	1	0
5	4	20.078	1	1	-0.9829034759	1	1	3.718	1	1	2	0
6	4	30.009	2	1	-0.9829034759	1	2	3.718	1	2	3	1
7	4	39.757	3	1	-0.9829034759	1	3	3.718	1	0	0	1
8	6	4.090	0	1	-1.4363961037	1	0	0.424	0	0	1	1
12	13	8.913	0	1	-0.0591650614	0	0	14.154	0	0	1	0
13	13	16.242	1	1	-0.0591650614	0	1	14.154	0	1	2	0
14	13	23.475	2	1	-0.0591650614	0	2	14.154	0	2	3	0
15	13	30.746	3	1	-0.0591650614	0	3	14.154	0	3	4	1
16	13	38.002	4	1	-0.0591650614	0	4	14.154	0	0	0	1
17	14	5.917	0	1	-0.6062510008	0	0	20.268	0	0	1	0
18	14	15.704	1	1	-0.6062510008	0	1	20.268	0	1	2	0
19	14	25.553	2	1	-0.6062510008	0	2	20.268	0	2	3	0
20	14	35.268	3	1	-0.6062510008	0	3	20.268	0	3	4	1
21	14	45.209	4	1	-0.6062510008	0	4	20.268	0	0	0	1

Figure 7.1: Sample of the Longitudinal part of the data.

	id	survtime	cens	ctsx	binx
1	1	0.000000e+00	1	1.4089495334	1
2	2	3.808031e-12	1	0.7587348284	0
3	3	1.719401e+00	1	-0.1635747110	0
4	4	3.717563e+00	1	-0.9829034759	1
5	5	0.000000e+00	1	0.6198512481	1
6	6	4.237720e-01	0	-1.4363961037	1
7	7	3.609227e-09	1	0.5499722714	0
8	8	1.579091e-12	1	0.8430497003	0
9	9	1.573089e-13	1	1.1296915605	0
10	10	0.000000e+00	1	1.9290942832	0
11	11	0.000000e+00	1	1.9144469421	1
12	12	0.000000e+00	1	0.8293037328	1
13	13	1.415354e+01	0	-0.0591650614	0
14	14	2.026819e+01	0	-0.6062510008	0
15	15	9.679849e-04	1	0.5612763664	0
16	16	3.987953e-06	1	0.6442367745	0

Figure 7.2: Sample of the Survival part of the data.

Our strategy will be to first fit the longitudinal and the survival part separately, this will help us extract the starting values of the parameters in the estimation scheme

for the Joint model. This strategy also enables us to prevent possible errors in the Joint model selection process. For example , if an int-slope model is the appropriate choice for the longitudinal part of the model, this will be clear when we fit the longitudinal part of the data separately.

As a first step we choose 4 patients at random and plot their longitudinal trajectory against time :

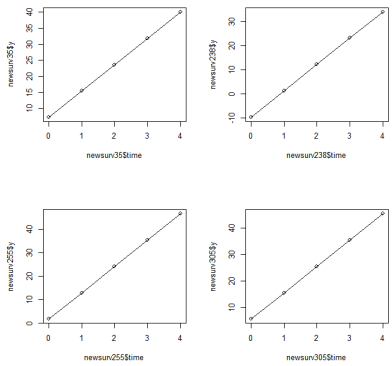


Figure 7.3: Y measurements of four random patients.

Looking at Figure 7.3, it is clear that the data have a linear upward trend. It is also clear that the starting point (intercept) and the slope is differs from patient to patient. Hence at first glance the ideal model choice seems to be a Linear Mixed Effects model with random intercept and slope.

In Figures 7.4 and 7.5 we plot the longitudinal trajectories of patients grouped up according to their treatment status.

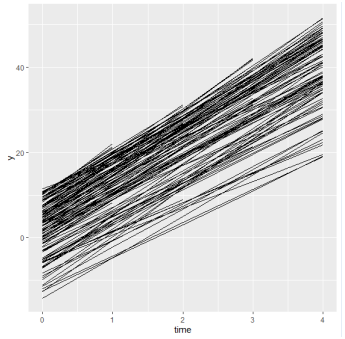


Figure 7.4: Binxl=0.

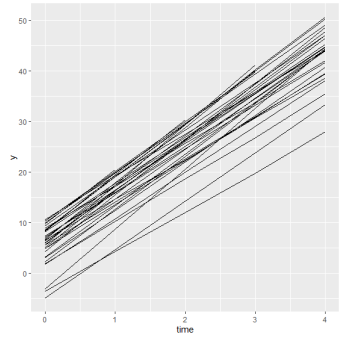


Figure 7.5: Binxl=1.

A difference in pattern between Figure 7.4 and Figure 7.5 is evident, this suggests that we need to include Binxl in our longitudinal model.

Keeping this in mind we proceed in a forward model selection. With BIC as guide we select a Mixed effects model with Fixxed effects: time, ctsxl and binxl and Random effects: intercept and slope. In figure 6 one can find the model output.

```
Linear mixed-effects model fit by REML
 Data: newlong
       AIC      BIC    logLik
  1894.025 1926.999 -940.0126

Random effects:
 Formula: ~time | id
 Structure: General positive-definite, Log-Cholesky parametrization
             StdDev     Corr
(Intercept) 6.25653058 (Intr)
time        1.40487930 -0.109
Residual    0.09918798

Fixed effects: y ~ time + binxl
               Value Std.Error  DF  t-value p-value
(Intercept) 3.876362 0.4925541 595  7.86992       0
time        9.784273 0.1079857 595 90.60714       0
binxl       4.432612 0.9060761 226  4.89210       0
 Correlation:
      (Intr) time
time  -0.085
binxl -0.541  0.011

Standardized Within-Group Residuals:
         Min           Q1          Med           Q3          Max
-2.449680993 -0.431703234  0.008323149  0.454174852  3.039500256
```

Figure 7.6: Binxl=1.

All selected parameters are statistically significant as p-value=0 for every parameter. The value of Binxl is 4.43 something that suggests a very strong connection between treatment and longitudinal measurement.

We proceed with the survival submodel. We first plot the Survival Probability for the two groups. As shown in Figure 7.7, the two curves neither do cross nor are parallel, so a Proportional Hazards model would probably be the appropriate choice.

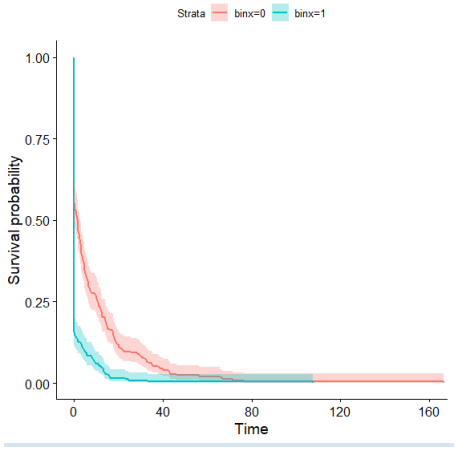


Figure 7.7: Survival Probability for the two groups.

We postulate a Cox Proportional Hazards model with a Weibull baseline risk fuction. In Figure 7.8 one can find the output of the model. We included binxl and ctsxl, both are statistically significant with coefficients 3.83 and 3.85 respectively. Hence the risk contribution for a patient in treatment is exp 3.83 and exp 3.85 for a unit change in ctsxl.

```
Call:
coxph(formula = Surv(survtime1, Event) ~ binxl + ctsxl, data = newsurvfitdata,
    x = "TRUE", timefix = TRUE)

  n= 228, number of events= 74

         coef exp(coef) se(coef)     z Pr(>|z|)
binxl  3.8360   46.3389   0.5263 7.289 3.13e-13 ***
ctsxl  3.8598   47.4538   0.5326 7.247 4.27e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

      exp(coef) exp(-coef) lower .95 upper .95
binxl     46.34     0.02158     16.52     130.0
ctsxl     47.45     0.02107     16.71     134.8

Concordance= 0.856  (se = 0.025 )
Likelihood ratio test= 104.7  on 2 df,   p=<2e-16
Wald test            = 55.75  on 2 df,   p=8e-13
Score (logrank) test = 59.57  on 2 df,   p=1e-13
```

Figure 7.8: Output of the survival model.

At this point we have both, longitudunal and survival submodels well defined and we are ready to postulate our Joint model. The Joint model will include every parameter in the two submodels plus the association parameter α. This parameter quantifies the contribution of the true longitudinal marker in the risk for an event. One can find the Joint model output in Figure 7.9.

```
Call:
jointModel(lmeObject = lmeFit1, survObject = survFit1, timeVar = "time")

Data Descriptives:
Longitudinal Process           Event Process
Number of Observations: 824    Number of Events: 74 (32.5%)
Number of Groups: 228

Joint Model Summary:
Longitudinal Process: Linear mixed-effects model
Event Process: Weibull relative risk model
Parameterization: Time-dependent

   log.Lik      AIC      BIC
 -1016.842 2057.685 2098.837

Variance Components:
            StdDev    Corr
(Intercept)  6.2295  (Intr)
time         1.4010 -0.1100
Residual     0.0992

Coefficients:
Longitudinal Process
            Value Std.Err z-value p-value
(Intercept) 3.8780  0.4905  7.9067 <0.0001
time        9.7841  0.1112 88.0019 <0.0001
binxl       4.4307  0.9023  4.9106 <0.0001

Event Process
             Value Std.Err z-value p-value
(Intercept)  0.2292  0.2624  0.8733  0.3825
binxl        3.8139  0.5395  7.0700 <0.0001
ctsxl        3.8395  0.5493  6.9894 <0.0001
Assoct      -0.0440  0.0146 -3.0170  0.0026
log(shape)  -1.0352  0.1666 -6.2140 <0.0001

Scale: 0.3551

Integration:
method: (pseudo) adaptive Gauss-Hermite
quadrature points: 5
```

Figure 7.9: Output of the Joint Model.

As we can in Figure 9 all parameters are statistically significant. The association parameter α is statistically significant, this implies the need of a Joint model. If α was not statistically significant, that would mean that we do not need a Joint model. There is a slight difference in parameter estimation between separate estima-

tion and Joint model estimation, we expected this since binxl parameter is included in both submodels. Estimation was achieved with the scheme presented earlier, with 5 quadrature points.

Figures 7.10,7.11 and 7.12 contain samples of the posterior mode, covariance matrix and inverse Hessian matrix respectively.

```
  (Intercept)        time
1   2.7794786 -0.09449434
2  -2.0200733  0.49784251
3  -7.6842860  1.16402036
4   8.2866640 -0.20624093
5   0.5192908  2.34021287
6   0.2577989  1.26417608
```

Figure 7.10: Sample of posterior modes.

```
$`1`
             (Intercept)          time
(Intercept)  0.0098350692 -0.0002457361
time        -0.0002457361  1.9390172293

$`2`
             (Intercept)          time
(Intercept)  0.005899986 -0.0019663543
time        -0.001966354  0.0009832161

$`3`
             (Intercept)          time
(Intercept)  0.005900004 -0.0019663600
time        -0.001966360  0.0009832181
```

Figure 7.11: Sample of covarience matrices.

```
$`1`
             (Intercept)      time
(Intercept)   72.088715 -2.002697
time          -2.002697 16.973395

$`2`
             (Intercept)      time
(Intercept)   44.575415 -9.380489
time          -9.380489  3.790471

$`3`
             (Intercept)      time
(Intercept)   36.079820 -5.278438
time          -5.278438  1.419307
```

Figure 7.12: Sample of inverse Hessian.

We proceed with the model diagnostics. Diagnostics procendure is similar to the diagnostics procendure in standard Survival analysis, the main diffence is that we expect the model to produce biased results since there dataset is incomplete by nature with the MNAR mechanism.

We start by plotting every martingale residual against its fitted value. If the model is correctly specified we expect the lowess smoother to be a null horizontal line. Figure 7.13 shows deviation of the lowess smoother, this is due to the omission of some of the patients.

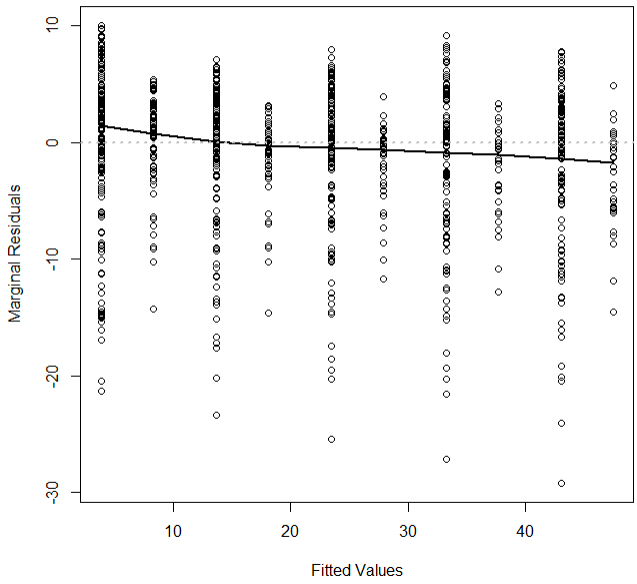


Figure 7.13: Martingale Residuals.

We proceed by plotting subject specific residuals against their fitted values and subject specific residuals for the two treatment groups, again a null horizontal line

implies correct model specification. Deviation from a null horizontal line is again observed.

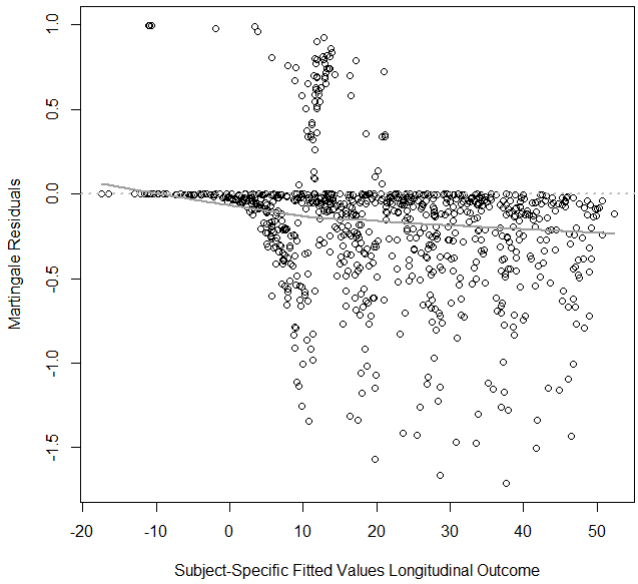


Figure 7.14: Subject specific residuals .

As a last step we plot the Kaplan-Meyer estimate of the Cox-Snell residuals. If the Kaplan-Meyer estimate of the Cox-Snell residuals is close to the unit exponential distribution, this implies strong model fit.

Diagnostics show that model fit is not ideal, this is due to the fact that some observations have been ommited. In real life problems part of the data will be missing data, this will result in diagnostics like the above even if the model specification is correct. To overcome this issue, we multiply impute the dataset according to the

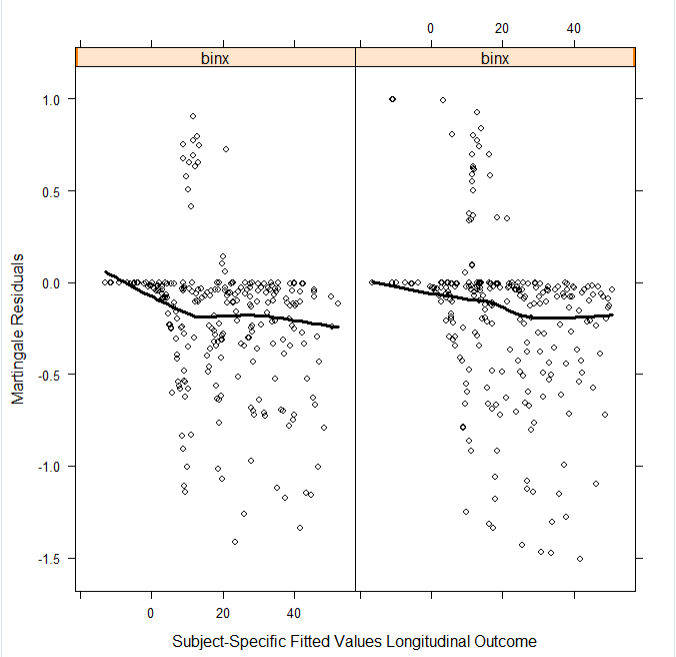


Figure 7.15: Subject specific residuals for the two treatment groups .

imputation schemes presented in the earlier chapters.

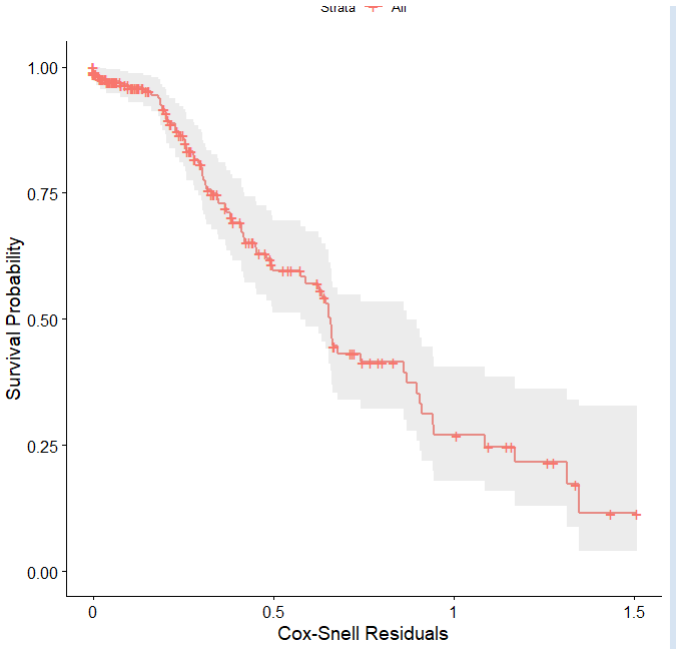


Figure 7.16: Kaplan-Meyer estimate for the Cox Snell Residuals .

Further Reading

This was a brief presentation of Joint models for Longitudinal and Survival data. In the last few years Joint models have come under the spotlight, new techniques and procendures have been explored. Steps towards robust analysis, sensitivity analysis and numerical stability have been explored as these issues keep Joint models from being widely applicable. The inclusion of multivariate longitudinal responses have been made with exceptional results. Furthermore, new choices of models have emerged as fully Bayesian models and even Deep neural Networks have beed utilized with good results. One who wants to seek further information about Joint models is advised to follow one of these paths.

Appendix A

Appendix

In this section technical information about the scheme used to maximize the log likelihood of Joint models is presented.

A.0.1 The E-M algorithm

The Expectation-Maximization (E-M) is an iterative algorithm used for maximum likelihood estimation in incomplete data problems .The idea behind the EM algorithm is that the log-likelihood corresponding to the complete data is typically much simpler to maximize, often in close form. To take advantage of this feature, the algorithm iterates between two steps: the Expectation (E) step and the Maximization (M) step. In the E-step we fill in the missing data and we replace, in fact, the log-likelihood of the observed data with a surrogate function which is then maximized in the M-step. This replacement creates the need for the algorithm to be iterative because the reconstruction of the missing data in the E-step is bound to be slightly wrong if the parameters do not already equal to their maximum likelihood estimates.

Briefly the algorithm proceeds as follows :

Let Y denote the complete data vector.

Let Y^o denote the observed part of the data vector.

Let Y^m denote the missing part of the data vector.

Our aim is to estimate the parameters θ of the complete data model, but using only the observed information. In the E-step we compute the expected value of the complete data log-likelihood:

$$Q(\theta|\theta^{it}) = E\{\log \Pr(y;\theta)|y^o;\theta^{it}\} =$$
$$= \int \log\{\Pr(y^o, y^m;\theta) \Pr(y^m|y^o;\theta^{it})dy^m,$$

and in the M-step we update the parameters by :

$$\theta^{(it+1)} = argmax Q(\theta|\theta^{(it)}).$$

At each iteration E-M leads to increase of the observed data likelihood (Dempster et al. 1977) i.e., $\log\{\Pr(y^o;\theta^{it+1}) \geq \log\{\Pr(y^o;\theta^{it+1})$ and avoids wildly overshooting or undershooting the maximum of the likelihood along its current direction of search. Another great advantage of the E-M algorithm is its numerical stability. However, an important drawback of the EM is its slow rate of convergence in a neighborhood of the maximum point. We will slide over this drawback by using the Newton - Rhapson algorithm for the final iterations of our maximization.

A.0.2 E-M for Joint models , E-step

We will illustrate the use of the EM algorithm to derive the maximum likelihood estimates of the standard joint model, all extensions presented above can be incorporated.

In particular, we consider the model:

$$h_i(t) = h_i(t)\exp(\gamma^T w_i + \alpha\{x_i^T\beta + z_i^T(t)b_i\}),$$

$$y_i(t) = x_i^T\beta + z_i^T b_i + \epsilon_i(t)$$

$$b_i \sim N(0, D), \quad \epsilon_i(t) \sim N(0, \sigma^2),$$

where $\theta = (\theta_t^T, \theta_y^T, \theta_b^T)^T$, with $\theta_y = (\theta^T, \sigma^2)^T$, $\theta_t = (\gamma^T, \alpha, \theta_{h_0}^T)^T$, θ_{h_0} denoting the parameters in the baseline risk function and $\theta_b = vech(D)$.

To apply EM in the Joint models we treat random effects as missing data. In particular, our aim is to find the parameter values that maximize the observed data log-likelihood by maximizing the expected value of the complete data log-likelihood instead :

$$Q(\theta|\theta^{(it)}) = \sum_i \int log\Pr(T_i, \delta_i, y_i, b_i; \theta)\Pr(b_i|T_i, \delta_i, y_i; \theta^{(it)})db_i =$$

$$\sum_i \int \{log\Pr(T_i, \delta_i|b_i; \theta_t, \beta) + log\Pr(y_i|b_i; \theta_y) +$$

$$+log\Pr(b_i; \theta_b)\}\Pr(b_i|T_i, \delta_i, y_i; \theta^{(it)}).$$

the integral with respect to the random effects as well as the integral with respect to time in the definition of the survival function involved in term $\Pr(T_i, \delta_i|b_i; \theta_t, \beta)$

do not have closed-form solutions therefore numerical integration procedures must be employed, such as Gaussian quadrature rules or Monte Carlo sampling.

A.0.3 E-M for Joint models , M-step

Due to the fact that the complete data log-likelihood is split into three parts, maximization of $Q(\theta|\theta^{(it)})$ with respect to θ involves three pieces in which the parameters on interest appear. The following expressions required in the specification of the M-step are presented using the integrals with respect to time and the random effects. For the actual calculation of these expressions, these integrals need to be approximated with the methods mentioned in the Numerical Integration section. More specifically, for the measurement error variance in the longitudinal measurement model and the covariance matrix of the random effects are updated in the M-step according to the closed-form expressions

$$\hat{\sigma}^2 = N^{-1} \sum_i \int (y_i - X_i\beta - Z_i b_i)^T (y_i - X_i\beta - Z_i b_i) \Pr(b_i|T_i, \delta_i, y_i; \theta) db_i =$$

$$= N^{-1} \sum_i (y_i - X_i\beta)^T (y_i - X_i\beta - 2Z_i\tilde{b}_i) + tr(Z_i^T Z_i \tilde{u}b_i) + \tilde{b}_i^T Z_i \tilde{b}_i,$$

$$\widehat{D} = n^{-1} \sum_i \tilde{u}b_i + \tilde{b}_i \tilde{b}_i^T,$$

where $N = \sum_i n_i$, $\tilde{b}_i = E(b_i|T_i, \delta_i, y_i; \theta^{(it)}) = \int b_i \Pr(b_i|T_i, \delta_i, y_i; \theta^{(it)}) db_i$ and $\tilde{ub}_i = Var(b_i|T_i, \delta_i, y_i; \theta^{(it)}) = \int (b_i - \tilde{b}_i^2) \Pr(b_i|T_i, \delta_i, y_i; \theta^{(it)}) db_i$. Under the above formulation of the joint model, we cannot obtain closed-form solutions of the score equations for the fixed effects β and the parameters of the survival submodel θ_t. Thus, for these parameters the M-step is implemented via a one-step Newton-Raphson update :

$$\hat{\beta}^{(it+1)} = \hat{\beta}^{(it)} - \{\partial S(\hat{\beta}^{(it)})/\partial\beta\}^{-1} S(\hat{\beta}^{(it)}),$$

$$\hat{\theta}^{(it+1)} = \hat{\theta}^{(it)} - \{\partial S(\hat{\theta}^{(it)})/\partial\theta\}^{-1} S(\hat{\theta}^{(it)}),$$

where $\hat{\beta}^{(it)}, \hat{\theta}^{(it)}$ denote the parameter values at each iteration and $\partial S(\hat{\beta}^{(it)})/\partial\beta, \hat{\theta}^{(it)})/\partial\theta$ denote the corresponding blocks of the Hessian matrix, evaluated at $\hat{\beta}^{(it)}, \hat{\theta}^{(it)}$ respectively. The components of the score vector corresponding to β and θ_t have the form

$$S(\beta) = \sum X_i^T \{y_i - X_i\beta - Z_i\tilde{b}_i\}/\sigma^2 + \alpha\delta_i x_i(T_i) -$$

$$- \exp(\gamma^T w_i) \int \int_0^{T_i} h_0(s)\alpha x_i(s) \exp[\alpha\{x_i^T(s)\beta + z_i^T(s)b_i\}]$$

$$\times \Pr(b_i|T_i, \delta_i, y_i; \theta) ds db_i,$$

$$S(\gamma) = \sum_i w_i[\ \delta_i - \exp(\gamma^T w_i) \int \int_0^{T_i} h_0(s) \exp[\alpha\{x_i^T(s)\beta + z_i^T(s)b_i\}] \times$$

$$\times \Pr(b_i|T_i, \delta_i, y_i; \theta) ds db_i\ ],$$

$$S(\alpha) = \int_i \delta_i \{x_i^T(T_i)\beta + z_i^T(T_i)\tilde{b}_i\} -$$

$$- \exp(\gamma^T w_i) \int \int_0^{T_i} h_0(s) \exp[\alpha\{x_i^T(s)\beta + z_i^T(s)b_i\}] \times$$

$$\times \Pr(b_i|T_i, \delta_i, y_i; \theta) ds db_i,$$

$$S(\theta_{h_0}) = \sum_i \delta_i \frac{\partial h_0(T_i;\theta_{h_0})}{\partial \theta_{h_0}^T} -$$

$$- \exp(\gamma^T w_i) \int \int_0^{T_i} \frac{\partial h_0(s;\theta_{h_0})}{\partial \theta_{h_0}^T} \exp[\alpha\{x_i^T(s)\beta + z_i^T(s)b_i\}] \times$$

$$\times \Pr(b_i|T_i, \delta_i, y_i; \theta) ds db_i.$$

The corresponding blocks of the Hessian matrix, respectively, can be computed using a central difference approximation (Press et al., 2007, Section 5.7).

Sources

Siannis F.: Introduction to Longitudinal Analysis.

Garrett Fitzmaurice, Marie Davidian, Geert Verbeke, Geert Molenberghs: Longitudinal Data Analysis.

Garrett Fitzmaurice, Nan M. Laird, James Ware : Applied Longitudinal Analysis.

Siannis F.: Introduction to Survival Analysis.

Collett D. : Modelling Survival Data in Medical Research.

Lawless J. F. : Statistical Models and Methods for Lifetime Data.

Kalbfleisch J. D., Prentice R.L. :The Statistical Analysis of Failure time Data.

Dimitris Rizopoulos : Joint Models for Longitudinal and Time-to-Event Data.

Christian, Brian; Griffiths, Tom (April 2017), "Chapter 7: Overfitting", Algorithms To Live By: The computer science of human decisions, William Collins.

Dafni, U. and Tsiatis, A. (1998). Evaluating surrogate markers of clinical outcome measured with error.

Tsiatis, A. and Davidian, M. (2001). A semiparametric estimator for the proportional hazards model with longitudinal covariates measured with error.

Ye, W., Lin, X., and Taylor, J. (2008a). A penalized likelihood approach to joint modeling of longitudinal measurements and time-to-event data.

Ye, W., Lin, X., and Taylor, J. (2008b). Semiparametric modeling of longitudinal measurements and time-to-event data – a two stage regression calibration approach.

Sweeting, M. and Thompson, S. (2011). Joint modelling of longitudinal and time-to-event data with application to predicting abdominal aortic aneurysm growth and rupture.

van der Vaart, A. (1998). Asymptotic Statistics.

Wulfsohn, M. and Tsiatis, A. (1997). A joint model for survival and longitudinal data measured with error.

Henderson, R., Diggle, P., and Dobson, A. (2000). Joint modelling of longitu-

dinal measurements and event time data.

Song, X., Davidian, M., and Tsiatis, A. (2002). A semiparametric likelihood approach to joint modeling of longitudinal and time-to-event data.

Sylvestre, M.-P. and Abrahamowicz, M. (2009). Flexible modeling of the cumulative effects of time-dependent exposures on the hazard.

Hauptmann, M., Wellmann, J., Lubin, J., Rosenberg, P., and Kreienbrock, L. (2000). Analysis of exposure-time-response relationships using a spline weight function.

Vacek, P. (1997). Assessing the effect of intensity when exposure varies over time.

Tseng, Y.-K., Hsieh, F., and Wang, J.-L. (2005). Joint modelling of accelerated failure time and longitudinal data.

Pulkstenis, E., Ten Have, T., and Landis, R. (1998). Model for the analysis of binary longitudinal pain data subject to informative dropout through remedication.

Albert, P. and Follmann, D. (2000). Modeling repeated count data subject to informative dropout.

Albert, P., Follmann, D., Wang, S., and Suh, E. (2002). A latent autoregressive model for longitudinal binary data subject to informative missingness.

Faucett, C., Schenker, N., and Elashoff, R. (1998). Analysis of censored survival data with intermittently observed time-dependent binary covariates.

Rizopoulos, D. (2010). JM: An R package for the joint modelling of longitudinal and time-to-event data.

Rizopoulos, D. (2011). Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data.

Rizopoulos, D. (2012a). Fast fitting of joint models for longitudinal and event time data using a pseudo-adaptive Gaussian quadrature rule.

Rizopoulos, D. (2012b). JM: Shared parameter models for the joint modelling of longitudinal and time-to-event data.

Rizopoulos, D. and Ghosh, P. (2011). A Bayesian semiparametric multivariate joint model for multiple longitudinal outcomes and a time-to-event.

Rizopoulos, D. and Lesaffre, E. (2012). Introduction to the special issue on joint modelling techniques.

Rizopoulos, D., Verbeke, G., and Lesaffre, E. (2008). A two-part joint model for the analysis of survival and longitudinal binary data with excess zeros.

Rizopoulos, D., Verbeke, G., and Lesaffre, E. (2009). Fully exponential Laplace

approximations for the joint modelling of survival and longitudinal data.

Rizopoulos, D., Verbeke, G., and Molenberghs, G. (2008). Shared parameter models under random effects misspecification.

Rizopoulos, D., Verbeke, G., and Molenberghs, G. (2010). Multiple imputation-based residuals and diagnostic plots for joint models of longitudinal and survival outcomes.

Yao, F. (2008). Functional approach of flexibly modelling generalized longitudinal data and survival time.

Proust-Lima, C. and Taylor, J. (2009). Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of posttreatment PSA: A joint modeling approach.

Lin, H., McCulloch, C., and Rosenheck, R. (2004). Latent pattern mixture models for informative intermittent missing data in longitudinal studies.

Data Analysis Using Regression and Multilevel/Hierarchical Models.