Politecnico di Milano

School of Industrial and Information Engineering
Master of Science in Mathematical Engineering

A BAYESIAN APPROACH TO BLOOD DONATION PROCESS

Advisor:
Prof. Ilenia Epifani

Co-Advisor:
Prof. Alessandra Guglemi

Author:
Ilaria Martinelli
Matricola 920862

Academic Year 2019-2020

Contents

List of Tables

List of Figures

Abstract

This thesis approaches the problem of blood donations, that is relevant to the health system, with Bayesian statistical modeling. The work focuses on the prediction of the number of donations in a blood collection centre. The data we have analyzed were registrered by Associazione Volontari Italiani Sangue (AVIS), in particular by the section of Lambrate in Milan. Blood donations are modeled as recurrent events under the Bayesian approach. Starting from the work previously done on this topic by Gianoli (2016) and Spinelli (2019), the thesis proposes a Bayesian model, suitably parameterized, of the *intensity function* of the process of recurrent events of blood donations (i.e., the instantaneous probability of the donation event occurrence); it depends on both time-dependent or time-fixed covariates (representing individual donor features) and on individual random frailties, that model the mean random heterogeneity among donors. The analysis highlights a decreasing trend of the *baseline intensity function* and identifies the significant covariates that influence the intensity function and hence determine the donors personal propensity to donate. Bayesian inference is promising, and the model could help to plan short, medium and long-term blood donations and to profile blood donors.

Sommario

Questa tesi affronta il problema delle donazioni di sangue, rilevante per il sistema sanitario, con la modellazione statistica bayesiana. Il lavoro è incentrato sulla previsione del numero di donazioni in un centro di raccolta del sangue. I dati analizzati sono stati estratti da due banche dati dell'Associazione Volontari Italiani Sangue (AVIS), in particolare dalla sezione di Lambrate di Milano. Le donazioni di sangue sono modellizzate come eventi ricorrenti tramite un approccio bayesiano. Partendo dal lavoro precedentemente svolto su questo tema da Gianoli (2016) e Spinelli (2019), la tesi propone un modello bayesiano, adeguatamente parametrizzato, della *funzione di intensità* del processo degli eventi ricorrenti delle donazioni di sangue (cioè la probabilità istantanea dell'accadimento della donazione); essa dipende sia da covariate tempo-dipendenti o fisse nel tempo (che rappresentano caratteristiche individuali del donatore) sia da effetti aleatori individuali (*random frailties*) che rappresentano l'eterogeneità non misurabile fra donatori. L'analisi evidenzia un andamento decrescente della *funzione di intensità di base* e identifica le covariate significative che influenzano la funzione di intensità e quindi determinano la propensione personale dei donatori a donare. I risultati ottenuti sono promettenti sull'utilizzo del modello allo scopo di pianificare le donazioni di sangue a breve, medio e lungo termine e di profilare i donatori di sangue.

Introduction

Human blood is needed to save lives, to improve their quality and to extend their lengths. It is essential in first aids, in emergency services, in organ transplants and much else. In Italy the acquisition of blood products relies on voluntary donations. The major organization in Italy that collects volunteer blood donors is Associazione Volontari Italiani Sangue (AVIS), founded in Milan in 1927. Today, thanks to its associates, it manages to ensure about 80% of the national blood needs. Overall, AVIS can count on over one million of members, who each year contribute to the collection of over two milions units of blood and its derivatives. AVIS is present throughout the country with over 3400 locations (19 locations of them founded in Switzerland by Italian emigrants in the Sixties).

The blood donation supply chain can be divided in four phases: collection, transportation, storage and utilization. In the collection phase, donor's eligibility to donate is checked and, if the donation occurred, blood is screened in laboratory to prevent infectious diseases and it is possibly fractionated in sub-components. Afterwards it is transported and stored to hospitals or transfusions centers. Finally, it is used for a transfusion. Collection is one of the most important phases of the blood donation supply chain. Blood has a shelf life and the demand of hospitals and transfusions centres has to be covered with the maximum precision, to avoid wastage of this resource. The storage should be planned to keep constant the number of blood units of each type across days in every centre. Moreover knowing in advance the number of incoming donors can lead to an optimal planning of the appointment scheduling system and of the amount of staff needed in a given period (next week, next month and so on). Another key aspect is profiling, which consists in carrying out effective acquisition campaigns of new donors.

The models studied in this thesis belong to the Bayesian statistics and a recurrent event approach is adopted. Blood donation event over time are modelled, with both time-dependent and not only time-fixed covariates as in Spinelli (2019); previously Gianoli (2016) modelled the waiting times between two successive blood donations.

Thanks to improvements of the performances of computing systems and to the spread of Markov Chain Monte Carlo (MCMC) methods, the Bayesian approach is spreading in the scien-

tific world: probabilistic estimates are exact, because they do not rely on a large sample theory and some tools like interval estimates have a clear meaning. Moreover, the Bayesian paradigm offers a natural way to do forecasting, by means of the predictive distributions. Apart from Gianoli (2016) and Spinelli (2019), in all the publications frequentist methods have been used, while the Bayesian approach is largely unexplored.

This work deals with real data provided by the AVIS section of Lambrate, in Milan and its original contribution is the proposal of a Bayesian model –suitable parameterized– for the blood donation event intensity function, with random frailties and time-dependent covariates. That model aims to explain the blood donor behaviour since his/her first donation, using his/her individual features, included in the model as covariates. It can also be used for purpose of planning blood donations in short, medium and long terms and of profiling blood donors.

The topics developed in the thesis are organized as follows. Chapter 1 is devoted to the theoretical background: it contains a general overview of recurrent event processes, some essential notes on the Bayesian approach and the Monte Carlo Markov Chains and a brief research on the state of the art of Bayesian model of recurrent events. Chapter 2 describes the raw data, downloaded from two databases in the AVIS' server using SQL queries, and their pre-processing. Then missing values' imputation is shown in Chapter 3. As a result a dataset of the times of donations and personal features of 5937 individuals has been created. In Chapter 4 three different Bayesian models of recurrent events for blood donations are formulated and described in details. Chapter 5 presents the posterior inference analysis, derived in the form of MCMC sampling via Stan, a `C++` open source software that provides MCMC output; in particular it contains: the model selection based on the Bayesian criteria discussed in Chapter 1, the posterior estimation of the selected model's parameters (along with some comments on them), and the robustness analysis. In Chapter 6 AVIS' needs are exploited, with a focus on the key aspects of planning and profiling. The last chapter concludes with a summary of the main points of the thesis.

Chapter 1

Theoretical background

Event history analysis is a statistical methodology used in many different settings where one is interested in the occurrence of events. Event histories are not restricted to humans. A sequence of events could also happen to animals, plants, cells and in general to anything that changes, develops, or decays. The purpose of this chapter is introducing some basic concepts and ideas in event history analysis. Most of the material here refers to Cook and Lawless (2007) and Song and Kuo (2013).

1.1 General overview of recurring events

Let us consider a single recurrent event process starting for simplicity at time $T_1 = 0$ and let $0 = T_1 < T_2 < T_3 < \dots$ denote the events' times, where T_k is the time of k-th event. The associated **counting process** $\{N(t), 0 \leq t\}$ records the cumulative number of events generated by the process. Specifically

$$N(t) = \sum_{k=1}^{\infty} I\{T_k \leq t\} \tag{1.1}$$

is the number of events occurring over the time interval $[0, t]$ and $N(s, t) = N(t) - N(s)$ represents the number of events occurring over the interval $(s, t]$. The **history of the process** at time t is

$$H(t) = \{N(s) : 0 \leq s < t\} \tag{1.2}$$

As defined here, counting processes are right continuous; that is, $N(t) = N(t^+)$. Models for recurrent events can be specified very generally by considering the probability distribution for the number of events $\Delta N(t) = N(t^-, t) = N(t + \Delta t^-) - N(t^-)$ in short intervals $[t, t + \Delta t)$,

given the history $H(t)$ of event occurrence before time t. The event **intensity function**

$$\lambda(t|H(t)) = \lim_{\Delta t \to 0} \frac{\mathbb{P}((N(t+\Delta t) - N(t)) = 1|H(t))}{\Delta t} \tag{1.3}$$

gives the instantaneous probability of an event occurring at t, conditional on the process history, and defines the process mathematically. Throughout this thesis, intensity functions are used to model event processes through event counts. The results that follow are essential to accomplish this purpose; they are borrowed from Cook and Lawless (2007). Let us consider two fixed times s_1 and s_2, with $s_1 < s_2$, then:

Theorem 1 *Conditionally on $H(s_1)$, the probability density of the outcome "$n > 0$ events, at times $T_1 < \ldots < T_n$ for a process with an integrable intensity $\lambda(t|H(t))$ over an interval $[s_1, s_2]$" is:*

$$\prod_{j=1}^{n} \lambda(t_j|H(t_j)) \exp\left\{-\int_{s_1}^{s_2} \lambda(u|H(u))Y(u)du\right\} \tag{1.4}$$

where $Y(t) = 1$ if an individual is "at risk" for experiencing the event at time t, 0 otherwise.
The knowledge of the intensity function allows us to write down both the probability of a specified event history and the conditional probabilities of the inter-event times (also called gap or waiting times), as made explicit in the following theorems.

Theorem 2 *For an event with integrable intensity $\lambda(t|H(t))$:*

$$\mathbb{P}(N(s_2) - N(s_1) = 0|H(s_1)) = \exp\left\{-\int_{s_1}^{s_2} \lambda(u|H(u))Y(u)du\right\} \tag{1.5}$$

Corollary 1 *Let $W_j = T_j - T_{j-1}$ be the waiting time between the events $(j-1)$ and j, then:*

$$\mathbb{P}(W_j > w|T_{j-1} = t_{j-1}, H(t_{j-1})) = \exp\left\{-\int_{t_{j-1}}^{t_{j-1}+w} \lambda(u|H(u))du\right\} \tag{1.6}$$

The mean function and variance function for the counting process $\{N(t), 0 \leq t\}$, given by:

$$\mu(t) = \mathrm{E}[N(t)] \quad \text{and} \quad V(t) = \mathrm{Var}(N(t)) \tag{1.7}$$

are difficult to determine for general intensity functions. The following sections describe some important families of recurrent event processes, which serve as a basis for modeling and data analysis.

1.2 Poisson and renewal processes

Two types of processes might be considered canonical in this context. One is the Poisson process, which describes situations where events occur randomly, in such a way that the numbers of events in non-overlapping time intervals are independent. The other is the renewal process, in which the waiting times between successive events are independent; that is, an individual is "renewed" after each event occurrence.

1.2.1 Methods based on event counts

The Poisson processes are the canonical framework for the analysis of event counts. Poisson models typically use the calendar time or the age of the process as the time scale. The independent increments property of a Poisson process states that $N(s_1, s_2)$ is independent of $N(s_3, s_4)$, provided $s_2 < s_3$. Hence, in the Poisson case, the history process $H(t)$ does not affect the instantaneous probability of events at time t, and in the absence of covariates the only factor determining the intensity is the current time t. Therefore Poisson processes are Markov, with intensity function of the form:

$$\lambda(t|H(t)) = \lim_{\Delta t \to 0} \frac{\mathbb{P}(\Delta N(t) = 1|H(t))}{\Delta t} = \lambda(t) \tag{1.8}$$

In Poisson case, $\lambda(t)$ is both the intensity function and the **rate function** giving the marginal (i.e. unconditional) instantaneous probability of an event at time t. Specifically, $\lambda(t)\Delta t = \mathrm{E}[N(t)] = \mathrm{Var}(N(t))$, and if $\mu(t)$ denotes the expected cumulative number of events at t, then:

$$\mu(t) = \mathrm{E}[N(t)] = \mathrm{Var}(N(t)) = \int_0^t \lambda(s)ds \tag{1.9}$$

and $\lambda(t) = \mu'(t) = d\mu(t)/dt$.
Extensions can be considered to accommodate the between-subject variability in events' rates through fixed or time-varying covariates, and random effects, as explained later.

1.2.2 Methods based on waiting times

Let $W_j = T_j - T_{j-1}$ be the waiting (or gap) time between the $(j-1)$-th and j-th event. Analyses based on waiting times are often useful when the events are relatively rare or when the prediction of the time to the next event is of interest. Analyses based on waiting times are natural in studies of system failures, where repairs, made at each failure, return the system to a working state. Other settings include studies of cyclical phenomena where characterization of cycle length is

of interest; for instance, in a case of recurrent infections, an individual return to a similar state after the infection has been cleared (Cook and Lawless (2007)). Other examples are recurrent episodes of hospitalization or disability. Renewal processes are the canonical models for waiting times and are defined by:

$$\lambda(t|H(t)) = h(t - T_{N(t^-)}) \tag{1.10}$$

where $N(t^-) = \lim_{\Delta t \to 0} N(t - \Delta t)$ and $h(\cdot)$ is the **hazard function** of the gap times between events, which are independent and identically distributed. In general, for an absolutely continuous gap time W:

$$h(t) = \lim_{\Delta t \to 0} \frac{\mathbb{P}(W > t + \Delta t | W \geq t))}{\Delta t} = \frac{f(t)}{S(t)} \tag{1.11}$$

where $f(t)$ is the common density function of the waiting times and $S(t) = \mathbb{P}(W \geq t)$ is the corresponding **survival function**. Also in this case, generalizations that accommodate within-subject association or trends in gap times are often useful.

1.3 Covariates

In many applications it is important to relate event occurrence to fixed or time-varying covariates. We typically use x to denote fixed covariates and $x(t)$ time-varying covariates. Time-varying covariates can be **external** or **internal**. An external covariate is one whose values are determined independently of the recurrent event process (fixed covariates are therefore external). A covariate that is not external is called internal. Thus, air pollution is an external covariate in a study on hospital visits due to breathing problems. Instead, the number of lines of code changed in a software debugging process is an internal covariate, because it depends on prior events (i.e. faults detected) in the process. To sum up, if observable fixed and/or time-varying covariates $x(t)$ are related to event occurrence, they may be incorporated in the model by redefining the process history to include covariate information. The covariates are all assumed to be external (exogenous) in the development that follows. Internal covariates are more difficult to deal with, in terms of both modeling and interpretation.

1.4 Periods at-risk

The **at-risk indicator** is useful for denoting which individuals can provide information about events occurrence at a given time. It is defined as a time function $Y(t)$ such that $Y_i(t) = 1$ if an individual i is at risk of being observed at time $t \in [0, \tau]$ and $Y_i(t) = 0$ otherwise. The time τ is sometimes referred to as a censoring or end-of-followup time for the observed event process.

More general observational or censoring patterns can arise if subjects temporarily cease to be under observation. This happens, for example, if individuals are asked to record events on daily diary cards, but they stop doing it for a period of time. It is also possible that an individual ceases to be at risk temporarily, because of the nature of the process.

1.5 The choice of time scale

The choice of an appropriate time scale is crucial. The time variable t is often chronological or calendar time, especially with processes that apply to humans or animals. The time scale also involves a choice of origin and this requires some care when multiple individuals are under study. Intensity-based analyses can adapt to the choice of a time origin through specification of the intensity, but it is nevertheless desirable to use an origin that is consistent across individuals and facilitates interpretation and analysis. In many contexts this may be clear: possible time origins include the time of birth of the patient (with age as the time scale), the time of disease onset, or the time of entry to the clinic. It should also be noted that once an underlying time scale is chosen, it is necessary to decide whether it is most suitable to develop models based on the cumulative time or on gap times between events. Although this could be viewed as a model specification decision, it affects the analysis and interpretation of results.

1.6 Heterogeneity among individuals

The heterogeneity among individuals can be modelled by taking to account both covariates and random effects. Random effects are related to unobserved heterogeneity and they denote variation not explained by covariates. A more popular term is **frailty**, indicating that some individuals are more frail than others, that is, the event in question is more likely to happen for them. More precisely, frailty shall mean a part of the unexplained variation (Aalen et al. (2008)). As the focus of this thesis is on event counts, hence the starting point are Poisson processes.

1.6.1 Covariates in Poisson processes

A vector of covariates $\mathbf{x}(t) = (x_1(t), \ldots, x_P(t))$ can be incorporated in a Poisson process by considering intensities of the form:

$$\lambda(t|\mathbf{x}(t)) = \lambda_0(t)g(\mathbf{x}(t); \boldsymbol{\beta}) \tag{1.12}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_P)$ is a vector of regression coefficients, of the same length as $\mathbf{x}(t)$. In case $g(\mathbf{x}(t); \boldsymbol{\beta}) = \exp\{\mathbf{x}'(t)\boldsymbol{\beta}\}$, we get the so called **multiplicative model** or **log-linear model**. The positive-valued function $\lambda_0(t)$ is called **baseline intensity function** and it is common to all individuals. An approach that is simple yet flexible enough considers a piecewise constant $\lambda_0(t)$:

$$\lambda_0(t) = \lambda_k \qquad \text{if } a_{k-1} < t \leq a_k \qquad \text{for } k = 1, \ldots, K$$

where K is the number of time-intervals and $0 = a_0, a_1, \ldots, a_{K-1}, a_K = \tau$ are the cut-points. This model requires to partition the time domain in a fixed number of intervals. In analogy to the homogeneous Poisson process, the parameter λ_k can be interpreted as the occurrence rate of the events in the k-th interval. Preliminary choices to be discussed are the type and the numbers of intervals. It is common to choose the quantiles of the event times as cut-points of the time domain. However this is a data driven choice and, by definition, it is not independent of the data that the model aims to fit. Therefore particular attention should be paid to this aspect.

1.6.2 Random effects

Sometimes, even after conditioning on covariates, there is more inter-individual variation in event occurrence than that accounted by a Poisson process. One sign of that over-variation is a $\text{Var}(N_i(t))$ that appears substantially larger than $\text{E}[N_i(t)]$; instead mean and variance are identical under a Poisson model.

If counts are of interest and Poisson processes are still thought to be reasonable models for individuals, then an unobservable subject-specific random effect u_i, also called frailty, for $i = 1, 2, \ldots, I$ can be included, such that, given u_i and covariates $\mathbf{x}_i(t)$, the process $\{N_i(t), 0 \leq t\}$ is Poisson with rate function:

$$\lambda(t|\mathbf{x}_i(t), u_i) = \lambda_0(t) u_i \exp\{\mathbf{x}_i(t)'\boldsymbol{\beta}\} \tag{1.13}$$

Typically, all the random effects u_i are modeled as i.i.d. Gamma-distributed random variables with mean equal to 1 and variance equal to $\phi > 0$. That model is equivalent to state that, conditionally to u_i, the stochastic process $\{N_i(t) : 0 \leq t\}$ is a Poisson process with intensity $\lambda_0(t) u_i \exp\{\mathbf{x}_i(t)'\boldsymbol{\beta}\}$. More details on this approach are provided in Spinelli (2019).

1.7 Bayesian approach

Let us be interested in estimating a parameter θ. From a Bayesian perspective, the unknown parameter is understood as a random variable with a prior distribution, say $\pi(\theta)$, and the

statistical problem consists of updating $\pi(\theta)$ by computing a posterior conditional probability, given data $\mathbf{y} = (y_1, y_2, \ldots, y_n)$. This is done by using the **Bayes' Theorem**:

$$\pi(\theta|\mathbf{y}) = \frac{\mathbb{P}(\mathbf{y}|\theta)\pi(\theta)}{\int \mathbb{P}(\mathbf{y}|\theta)\pi(\theta)d\theta} \tag{1.14}$$

where (the likelihood) $\mathbb{P}(\mathbf{y}|\theta)$ is a measure of belief that $\mathbf{y}$ would be sampled from the population if θ is the true parameter. The posterior joint distribution is then summarized in a simple way, typically with posterior means giving rise to a point estimate of the unknown parameters. Moreover, the associated posterior standard errors and a $\gamma 100$ percent credible interval for the unknown parameters are computed. In Bayesian statistics, a $\gamma 100$ percent credible interval for a parameter θ is given by

$$q_{\frac{1-\gamma}{2}} < \theta < q_{\frac{1+\gamma}{2}} \tag{1.15}$$

where $q_{\frac{1-\gamma}{2}}$, $q_{\frac{1-\gamma}{2}}$ are posterior quantiles of θ (Epifani and Nicolini (2013)). Usually, the posterior distributions in Formula (1.14) of all the unknown parameters do not have a closed form. Hence, we need another approach to deal with it, in particular we may use **Markov Chain Monte Carlo** (MCMC) algorithms to simulate and summarize them.

1.8 Monte Carlo Markov Chains

Monte Carlo Markov Chains (MCMC) techniques allow us to simulate samples from the posterior distribution; they aim to construct cleverly sampled chains, which (after a burn-in period) draw samples which are progressively more likely realizations of the distribution of interest. The more steps are included, the more closely the distribution of the sample matches the actual desired distribution. In particular, MCMC techniques generate an ergodic Markov Chain $\theta^{(1)}, \ldots, \theta^{(M)}$, i.e. $\theta^{(j)}$, conditionally on $\theta^{(j-1)}$, is independent of $\theta^{(1)}, \ldots, \theta^{(j-2)}$ and, for a measurable function $h(\theta)$, if $M \to \infty$ then:

$$\frac{1}{M}\sum_{j=1}^{M} h(\theta^{(j)}) \to \mathrm{E}_\pi[h(\theta)|\mathbf{y}] = \int_\Theta h(\theta)\pi(d\theta|\mathbf{y})$$

In this way, all the summarises of the posterior distribution can be approximated by averaging over the MCMC sample. The MCMC algorithm used in this thesis is efficiently implemented in the open source software **Stan**, written in `C++`, that can be integrated with the software `R` thanks to the package `rstan`.

1.9 Bayesian methods for goodness-of-fit and model selection

Bayesian models can be evaluated and compared in several ways. Most simply, all inference is summarized by the posterior distribution. The idea is to obtain an unbiased and accurate measure of the out-of-sample predictive error, through cross-validation. However especially in a Bayesian setting this could be a problem, because exact cross-validation requires re-fitting the model with different training sets (Vehtari et al. (2016)). Alternative methods aim to estimate the out-of-sample predictive error with the data, using a correction for the bias that arises from evaluating the model's prediction on the data used to fit it; some of these measures are the Akaike Information Criterion (AIC), the Deviance Information Criterion (DIC), or the Watanabe–Akaike information criterion (WAIC), which is a fully Bayesian method.

1.9.1 Log-Pseudo-Marginal Likelihood (LPML)

Let us start considering a generic dataset $\mathbf{y} = (y_1, \ldots, y_n)$, modeled as observations of $Y_1, \ldots, Y_n$ which are independent random variables given parameter θ, with prior density $\pi(\theta)$. The initial idea is to split $\mathbf{y}$ in

$$(\underbrace{y_1, \ldots, y_k}_{\text{training set}}, \underbrace{y_{k+1}, \ldots, y_n}_{\text{test set}}), \text{ for some } k < n$$

and use the training set to compute the posterior density $\pi(\theta|y_1, \ldots, y_k)$, and the test set to check the quality of the model. In this particular case, we consider n splittings of $\mathbf{y}$ in $\mathbf{y}_{-i} = (y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n)$ as training set, and y_i as test set, as i varies from 1 to n. Thus we define the **conditional predictive ordinate**:

$$\text{CPO}_i = f_i(y_i|\mathbf{y}_{i-1}) \tag{1.16}$$

which is the conditional distribution of Y_i computed in the observation y_i, given all the rest. The **Logarithm of the Pseudo-Marginal Likelihood** (LPML) is defined as:

$$\text{LPML} = \sum_{i=1}^{n} \log \text{CPO}_i \tag{1.17}$$

The larger the value of the CPO's (and hence the larger the value of the LPML), the better the fit of the model. In details, we have:

$$\text{CPO}_i = f_i(y_i|\mathbf{y}_{i-1}) = \frac{m(y_1,\ldots,y_n)}{m(\mathbf{y}_{i-1})} = \frac{\displaystyle\int_\Theta \prod_{j=1}^{n} f_j(y_j|\theta)\pi(d\theta)}{\displaystyle\int_\Theta \prod_{j\neq i} f_j(y_j|\theta)}$$

where $m(\cdot)$ is the marginal density. Now we should compute n different posterior distributions and this is very time consuming if n is large. In fact $Y_i|\theta \sim f_i(\cdot|\theta)$ are not i.i.d. for all i, because the distribution can depend on i (this happens in the presence of covariates, for example in the context of a regression problem with $f_i(\cdot|\theta) = f(\cdot|\mathbf{x}_i, \boldsymbol{\beta}, \sigma^2)$, where f is a Normal distribution and we can notice that the mean $\mathbf{x}_i\boldsymbol{\beta}$ depends on i, while the variance σ^2 does not). Here we use the trick of working with the posterior distribution $\pi(\theta|y_1,\ldots,y_n)$:

$$\frac{1}{\text{CPO}_i} = \frac{\displaystyle\int_\Theta \prod_{j\neq i} f_j(y_j|\theta)\pi(\theta)d\theta}{\displaystyle\int_\Theta \prod_{j=1}^{n} f_j(y_j|\theta)\pi(\theta)d\theta} = \frac{\displaystyle\int_\Theta \frac{1}{f_i(y_i|\theta)} \prod_{j=1}^{n} f_j(y_j|\theta)\pi(\theta)d\theta}{\displaystyle\int_\Theta \prod_{j=1}^{n} f_j(y_j|\theta)\pi(\theta)d\theta} = \int_\Theta \frac{1}{f_i(y_i|\theta)}\pi(\theta|y_1,\ldots,y_n)d\theta$$

So it is enough to sample from the posterior density $\pi(\theta|y_1,\ldots,y_n)$; if we perform M MCMC iterations that get the posterior samples $\{\theta^{(j)},\ j = 1,\ldots,M\}$, then an approximation for CPO_i is given by the harmonic mean of $f_i(y_i|\theta^{(j)})$'s, i.e.

$$\widehat{\text{CPO}}_i = \frac{M}{\displaystyle\sum_{j=1}^{M} \frac{1}{f_i(y_i|\theta^{(j)})}} \tag{1.18}$$

1.9.2 Watanabe–Akaike Information Criterion (WAIC)

WAIC (introduced in 2010 by Watanabe, who called it the widely applicable information criterion) is a fully Bayesian predictive goodness-of-fit tool which approximates the **expected log pointwise predictive density** (elppd):

$$\text{elppd} = \sum_{i=1}^{n} \text{E}_f[\ln(f_i(y_i^{\text{new}}|\mathbf{y})] \tag{1.19}$$

(Gelman et al. (2013)). We label f_i as the true model, $\mathbf{y} = (y_1,\ldots,y_n)$ as the observed data and y_i^{new} as future datum $\forall i = 1,\ldots,n$. In order to compute the WAIC, we evaluate it at

$y_i^{\text{new}} = y_i \ \forall i$, as

$$\text{WAIC} = \text{computed lppd} - \text{p}_{\text{WAIC}_2}$$

where, for the posterior MCMC sample $\{\theta^{(j)},\ j = 1, \ldots, M\}$, we obtain:

$$\text{computed lppd} = \sum_{i=1}^{n} \ln \left(\frac{1}{M} \sum_{j=1}^{M} f_i(y_i|\theta^{(j)}) \right)$$

and

$$\text{p}_{\text{WAIC}_2} = \sum_{i=1}^{n} V_{j=1}^{M}(f_i(y_i|\theta^{(j)}))$$

with

$$V_{j=1}^{M} a_j = \frac{1}{M-1} \sum_{j=1}^{M} (a_j - \overline{a})^2$$

thay represents the sample variance (Gelman et al. (2013)).

1.10 Bayesian recurrent events in literature

Song and Kuo (2013) presents a Bayesian analysis for recurrent events data using a non-homogeneous mixed Poisson point process with a dynamic subject-specific frailty function and a dynamic baseline intensity function. Implementation of Bayesian inference using a **Reversible Jump Markov Chain Monte Carlo** (RJMCMC) algorithm is developed to handle the change of the dimension in the parameter space for models with a random number of change points. The intensity function is denoted by $\lambda_i(t)$ for the i-th subject at time t, for $i = 1, \ldots, I$, $t = 1, \ldots, T$ and is modeled as:

$$\lambda_i(t) = \lambda_0(t) u_i(t) \exp\{\mathbf{x}'(t)\boldsymbol{\beta}\} Y_i(t) \tag{1.20}$$

In Formula (1.20):

- $\lambda_0(t)$ denotes the baseline intensity function at time t common to all subjects;
- $u_i(t)$ denotes the frailty of subject i at time t;
- $\mathbf{x}_i(t)$ is a P-dimensional vector of covariates evaluated for subject i at time t;
- $\boldsymbol{\beta}$ is the P-dimensional vector of regression coefficients;
- $Y_i(t)$ is an indicator function with value 1 when the subject i is at risk of experiencing an event at time t and 0 otherwise.

Therefore three components, $\lambda_0(t)$, $\boldsymbol{\beta}$ and $u_i(t)$ need to be estimated. To make the estimation of $\lambda_0(t)$ and $u_i(t)$ more computationally tractable and without sacrificing too much generality, we assume piecewise constant models as an intermediate between parametric and nonparametric models. The details are given in the following Subsection 1.10.1 for λ_0, u_i, number of cut-off K, Subsection 1.10.2 for likelihood and Subsection 1.10.3 for prior.

1.10.1 Models

In Song and Kuo (2013) several models are compared: including constant or piecewise constant subject-specific frailty and a fixed number or a random number of change points in the baseline.

- **Model I**: *dynamic baseline model with a fixed number of change points K in the baseline and constant subject-specific frailty.* In this case

$$u_i(t) = u_i \qquad i = 1, \ldots, I \tag{1.21}$$

and the the baseline intensity function is piecewise constant:

$$\lambda_0(t) = \sum_{k=1}^{K} \lambda_k \mathrm{I}_{(a_{k-1}, a_k]}(t) \qquad k = 1, \ldots, K \tag{1.22}$$

where $0 = a_0 < a_1 < \ldots < a_{k-1} < a_k < \ldots < a_K = \tau$ and $[0, \tau]$ is the whole observation period. The modelling of the baseline intensity λ_k at the k-th segment is extended as:

$$\lambda_k = \lambda_{k-1}\delta_k \quad \text{so that} \quad \lambda_k = \prod_{g=1}^{k} \delta_g \tag{1.23}$$

where δ_k is the **baseline innovation** at segment k and $\delta_1, \delta_2, \ldots, \delta_K \overset{iid}{\sim} \text{Gamma}(\nu_1, \nu_1)$. The autocorrelation between λ_k and λ_{k+d} is the following:

$$\rho(\lambda_k, \lambda_{k+d}) = \sqrt{\frac{\left(1 + \frac{1}{\nu_1}\right)^{k} - 1}{\left(1 + \frac{1}{\nu_1}\right)^{k+d} - 1}} \tag{1.24}$$

and it can be noticed that:

  - it is a rational function of $1 + 1/\nu_1$, where $1/\nu_1$ measures the temporal heterogeneity within each subject;
  - it is decreasing in the distance d between λ_k, λ_{k+d} and increasing in k;

  - when $\nu_1 \to 0$, then the autocorrelation approaches 0;
  - when $\nu_1 \to \infty$, then the autocorrelation approaches to its maximum value $\sqrt{k/(k+d)}$.

- **Model II**: *dynamic baseline model with a fixed number of change points K in the baseline (Formula* (1.22) *and Formula* (1.23)*) and dynamic frailty.* Here all subjects share the same pre-specified cut-points at a_k for $k = 0, \ldots, K$ (as in Model I), but with different unknown frailty magnitudes over the real time-intervals. We have:

$$u_i(t) = \sum_{k=1}^{K} u_{ik} \mathrm{I}_{(a_{k-1}, a_k]}(t) \qquad i = 1, \ldots, I \tag{1.25}$$

  The evolution of frailties over the k-th time-interval can be defined by:

$$u_{ik} = u_{i(k-1)} \phi_{ik} \quad \text{so that} \quad u_{ik} = \prod_{g=1}^{k} \phi_{ig} \tag{1.26}$$

  where $\phi_{ik} \overset{iid}{\sim} \text{Gamma}(\nu_2, \nu_2)$ is the **multiplicative frailty innovation** for the i-th subject over segment k. In particular, the autocorrelation between u_{ik} and $u_{i,k+d}$ is the following:

$$\rho(u_{ik}, u_{i,k+d}) = \sqrt{\frac{\left(1 + \frac{1}{\nu_2}\right)^{k} - 1}{\left(1 + \frac{1}{\nu_2}\right)^{k+d} - 1}} \tag{1.27}$$

  and comments on it are similar to those reported for the autocorrelation of the baseline intensity function (Formula (1.24)) with respect to the same lag d.

- **Model III**: *constant subject-specific frailty (Formula* (1.21)*) and dynamic baseline model with random K in the baseline (Formula* (1.22) *and Formula* (1.23)*).* Now K is not fixed anymore and by requiring the number of change points to be data-dependent, more flexibility and adaptability are obtained for model fitting. Song and Kuo (2013) develops a Reversible Jump Markov Chain Monte Carlo (RJMCMC) method to handle the change of the dimension of the parameter space, which is introduced in this chapter as an hint for future developments, even if it goes beyond the purpose of this thesis.

- **Model IV**: *dynamic frailty (Formula* (1.24) *and Formula* (1.25)*) and dynamic baseline model with random K in the baseline (Formula* (1.22) *and Formula* (1.23)*).* The baseline intensity function is modelled as in Model III, meaning that also in this case RJMCMC methods are used.

As a generalisation, the time-discretization for the baseline and for frailties respectively can be

different. In this thesis we will assume as simplifying hypothesis the same discretization for both and this is mainly due to computational reasons when performing the simulations of the models.

1.10.2 Likelihood

The ingredients required for the likelihood's formulation are the following:

- $(0, \tau]$ study period;
- $0 = a_0 < a_1 < \ldots < a_{k-1} < a_k < \ldots < a_K = \tau$ pre-specified cut-points;
- $\lambda_i(t) = \lambda_{ik}$ for $t \in (a_{k-1}, a_k]$ and $k = 1, \cdots, K$;
- $\lambda_0(t) = \sum_{k=1}^{K} \lambda_k \mathrm{I}_{(a_{k-1}, a_k]}(t) = \sum_{k=1}^{K} \left(\prod_{g=1}^{k} \delta_g \right) \mathrm{I}_{(a_{k-1}, a_k]}(t)$;
- $\boldsymbol{\Delta}$ vector of δ_g for $g = 1, \ldots, K$;
- $u_i(t) = \sum_{k=1}^{K} u_{ik} \mathrm{I}_{(a_{k-1}, a_k]}(t) = \sum_{k=1}^{K} \left(\prod_{g=1}^{k} \phi_{ig} \right) \mathrm{I}_{(a_{k-1}, a_k]}(t)$;
- $\boldsymbol{\Phi}$ vector of ϕ_{ig} for $i = 1, \ldots, I$, $g = 1, \ldots, K$;
- count information $\boldsymbol{N} = (\boldsymbol{n_1}, \boldsymbol{n_2}, \ldots, \boldsymbol{n_I})$, where $\boldsymbol{n_i} = \{n_{ik}\}_{k=1,\ldots,K}$ is the number of recurrences for the i-th subject in the k-th interval and I the number of subjects.

We describe the likelihood function for Model II first, under the assumption of non-informative censoring:

$$\mathscr{L}(\boldsymbol{\beta}, \boldsymbol{\Phi}, \boldsymbol{\Delta}; \boldsymbol{N}) = \prod_{i=1}^{I} \prod_{k=1}^{K} f(N_{ik} | \boldsymbol{\beta}, \boldsymbol{\Phi}, \boldsymbol{\Delta}) \tag{1.28}$$

$$= \prod_{i=1}^{I} \prod_{k=1}^{K} \frac{e^{-\lambda_{ik}} \lambda_{ik}^{n_{ik}}}{n_{ik}!} = \tag{1.29}$$

$$= \prod_{i=1}^{I} \prod_{k=1}^{K} \frac{e^{-\lambda_0(t) u_i(t) \exp\{\mathbf{x}'(t)\boldsymbol{\beta}\} Y_i(t)} \left(\lambda_0(t) u_i(t) e^{\mathbf{x}'(t)\boldsymbol{\beta}} Y_i(t) \right)^{n_{ik}}}{n_{ik}!} \tag{1.30}$$

The likelihood for Model I is a special case of Model II's likelihood, with $u_i(t) = u_i$. Finally, the likelihood for Model III and Model IV, conditioning on the given number K of change points in the baseline, is that of Model I and Model II respectively.

1.10.3 Priors

Song and Kuo (2013) for their Models I-IV proposed the following prior scheme: all parameters δ_k's, ϕ_{ik}'s, β_p's are independent each other and on $(a_1, \ldots, a_K, K)$ with:

- baseline innovation:
  $\delta_k \overset{iid}{\sim} \text{Gamma}(\nu_1, \nu_1) \quad k = 1, \ldots, K$

- multiplicative frailty innovation:
  $\phi_{ik} \overset{iid}{\sim} \text{Gamma}(\nu_2, \nu_2) \quad i = 1, \ldots, I, \quad k = 1, \ldots, K$

- regression coefficients:
  $\beta_p \overset{iid}{\sim} \mathcal{N}(0, \sigma^2) \quad p = 1, \ldots, P$

- change points:
  $a_1, a_2, \ldots, a_K \sim \mathcal{U}\{K; (0, \tau]\}$

- number of change points:
  - Model I/II: K pre-specified
  - Model III/IV: $K \sim \mathcal{P}(\xi)$

and $\nu_1, \nu_2, \sigma^2, \xi$ are preassigned hyperparameters. As regards the updating algorithm for Model III and Model IV. It works as follows: if at current iteration we have K change points, then the possible moves are

- $S =$ stay, with no changes of K in the baseline's time-dimension;
- $D =$ death, reducing K by 1;
- $B =$ birth, increasing K by 1.

The updating of the parameter is done by randomly selecting one of the three moves (S, D, B) with probabilities, respectively:

- $b_K = \gamma \cdot \min\left\{1, \dfrac{\mathcal{P}_\xi(K+1)}{\mathcal{P}_\xi(K)}\right\}$
- $d_K = \gamma \cdot \min\left\{1, \dfrac{\mathcal{P}_\xi(K-1)}{\mathcal{P}_\xi(K)}\right\}$
- $c_K = 1 - b_K - d_K$

γ is chosen as large as possible, such that $b_K + d_K \leq 0.9 \quad \forall K$.

Chapter 2

Data sources description

In this section, the data analysed are described; they come from two databases provided by the AVIS section of Lambrate in Milan. A preliminary data exploration is reported, performed with both summary statistics and visualization techniques. Finally, the chapter covers the data pre-processing needed to get a more suitable representation of some covariates.

2.1 Selection criteria

The data of AVIS section of Lambrate in Milan are collected from multiple tables of two databases:

- **EMONET** database: it contains information about donations and donors' personal data;
- **AVIS** database: it contains information about blood donors' habits and suspensions.

In particular, data are selected according to the following criteria:

- donations come from Lambrate collection centre;
- the observation period goes from 1^{st} January 2010 to 30^{th} June 2018;
- only "new" donors are considered, namely people who became donors in that period;
- only donations of whole blood are included in the study.

According to this selection criteria, some data preprocessing has been necessary. First of all, 3238 volunteers are not considered, because they donate only once. Each donor's observation time has a length that is generally different from the others and the number of donations is different, from a minimum of 2 to a maximum of 30 donations. Since the focus is on recurrences, first donations (corresponding to time $t = 0$) are removed. The final dataset consists of 25689

observations and 5937 unique donors, divided in 4005 men and 1932 women. All the covariates taken into account are reported below. The starting point is the dataset from Spinelli (2019) containing the time-fixed covariates listed in Tables 2.1 and 2.2:

Variable name	**type**	**Description**
CAI	cat	Donor unique ID
DTPRES	date	Data and time of the event
SESSO	cat	1 Man; 0 Woman
ETA_PRIMA	num	Age at first donation
ETA_DONAZ	num	Age at current donation
AB0	cat	Blood type: A, B, AB, 0
TIPO_RH	cat	Reshus factor: POS or NEG

Table 2.1: EMONET database, time-fixed covariates

Variable name	**type**	**Description**
CAI	cat	Donor unique ID
FUMO	cat	Smoking habits
ALCOOL	cat	Drinking habits
THE	cat	Tea consumption
CAFFE	cat	Coffee consumption
DIETA	cat	Diet type
STRESS	cat	Stress level
ATTIVITAFISICA	cat	Physical activity habits
ALTEZZA	num	Donor's height (m)
PESO	num	Donor's weight (kg)

Table 2.2: AVIS database, time-fixed covariates

By querying again the AVIS' database and then merging the new dataset with the previous one, the time-dependent covariates listed in Table 2.3 are added:

Variable name	**type**	**Description**
PMIN	num	Minimum pressure
PMAX	num	Maximum pressure
POLSO	num	Heart rate
EMOG	num	Hemoglobin

Table 2.3: AVIS database, time-dependent covariates

2.2 Italian donation rules

There are some rules that regulate blood donations' mechanism, meant to protect both the health of the patient who will receive the blood and the health of the donor himself:

- any candidate donor must be between 18 and 60 years, however the responsible physician can allow a candidate donor older than 60 years old to donate for the first time. The chronological age limit is increased to 65 years old for periodic donors. Even in this case the physician can allow a person to donate until 70 years old, after a clinical evaluation of the risks correlated to the age;
- every donor must weight more than 50 kg;
- the blood pressure, the heart rate and the level of hemoglobin must lie between certain ranges. As an example, for the hemoglobin they are:
  - male donors: 13-18 g/dl;
  - female donors: 12-16 g/dl;
- the yearly maximum number of donations for men and for women who are in menopause is 4, while for the other women is 2. This means that the minimum gap time between two consecutive donations is 90 days for men and 180 days for women, but a certain tolerance is considered: the gap time for people belonging to first category is reduced to 85, while for the second one to 150 (indeed there are donations that happen before, since a physician is allowed to move up donations).

Similar rules govern the blood donation process in other countries, for example in Spain (see Aldamiz-Echevarria and Aguirre-Garcia (2014))

2.3 Exploratory data analysis

Data visualization is one of the most powerful and appealing techniques for data exploration. Humans have a well-developed ability to analyze large amounts of information that is presented visually. This gives the possibility, on one side, to fastly detect general patterns and trends and, on the other, to discover the presence of outliers and unusual patterns.

2.3.1 Rate of donation and gap times

The total numbers of donations for each donor are reported in Figure 2.1.

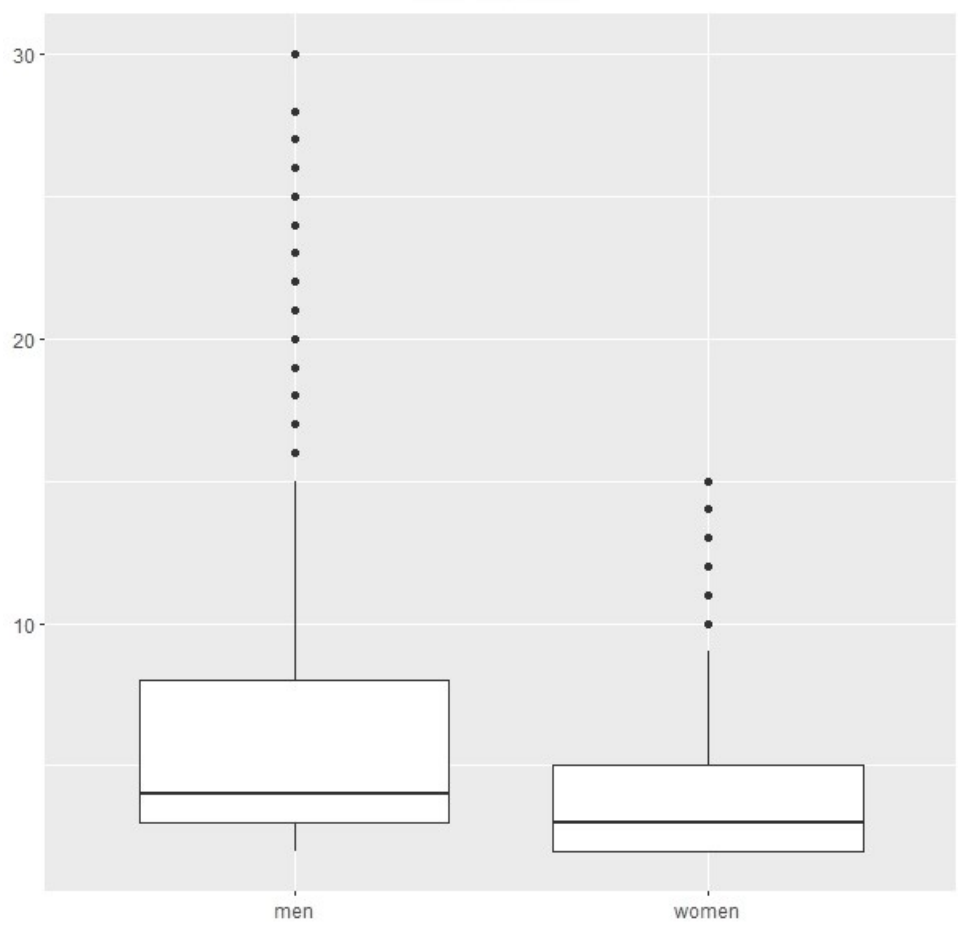

Figure 2.1: Boxplot of total donations for men and women

This number is not enough to understand how many donations are done in a certain time period. It should be related to the time in which each individual is observed, for example dividing it for the years of observation. The empirical rates of donation (number of donations divided for the years of observation) are computed and their histogram is shown in Figure 2.2. Notice that the empirical distribution of the yearly rate of donation is right-skewed: most of the donors did less than two donations per year. An interesting fact is the bimodality of the gap times' distribution, that reflects the difference of the donation rules between the two genders: men are allowed to donate twice than women. The red lines in Figure 2.3 correspond to the logarithms of 90 and 180, namely the minimum waiting times for men and women.

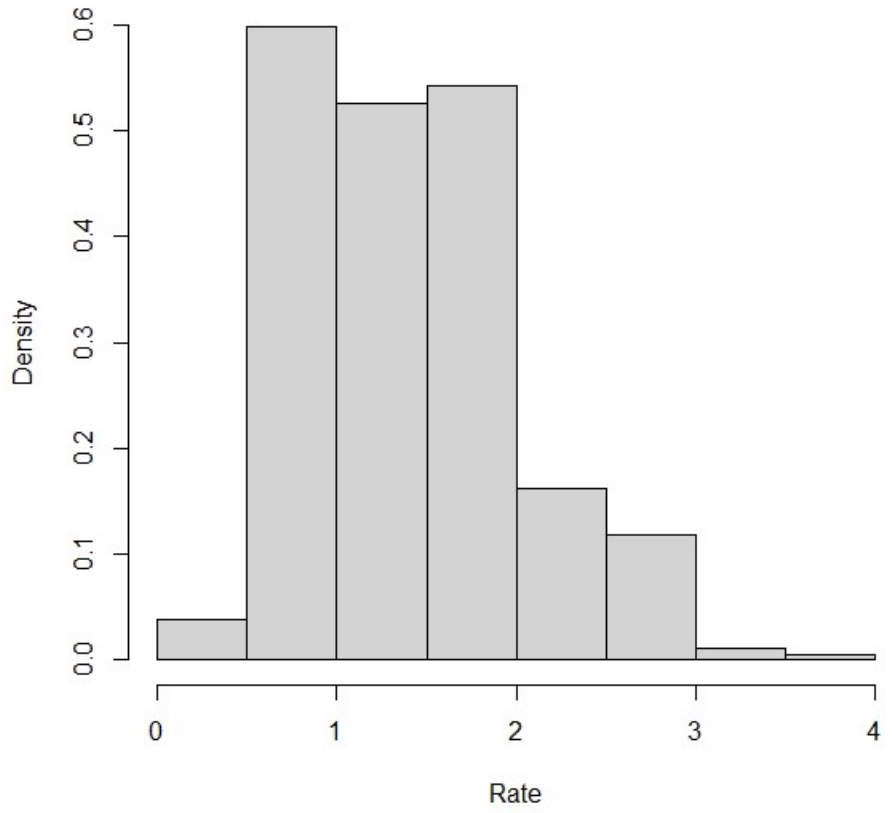

Figure 2.2: Histogram of the empirical yearly rates of donation

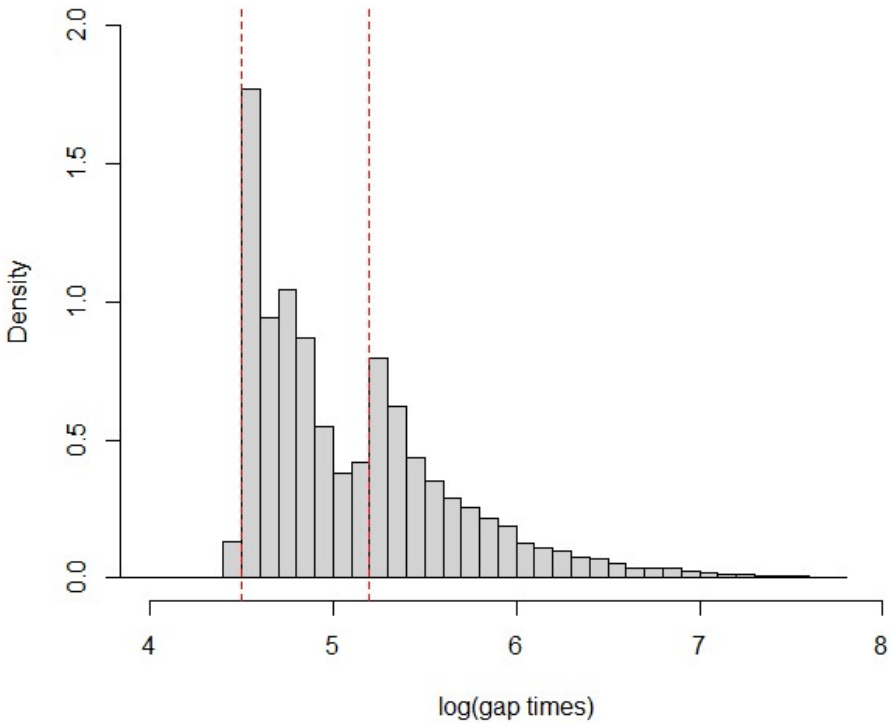

Figure 2.3: Histogram of the gap times on logaritmic scale. Red vertical lines correspond to the logarithms of 90 and 180, namely the minimum waiting times for men and women

2.3.2 Recurrences

As each donor's first donation is discarded in order to focus only on recurrences, the recurrences reach the maximum of 29 for men and 14 for women. A detailed analysis separates the donors' count into men and women respectively, as reported in Table 2.4 and the barplots in Figure 2.4 offer a graphical insights into this table. In particular, for women each bar represents:

$$\frac{\text{number of women who perform } i \text{ recurrences}}{(\text{total number of women}) * 14}$$

and for men:

$$\frac{\text{number of men who perform } i \text{ recurrences}}{(\text{total number of men}) * 29}$$

Normalizing frequencies with 14 and 29 is important, because it gives a fairer view of individuals' constancy in donating, regardless of their gender. It can be noticed that women outnumber men up to 8 total recurrences. After that threshold, men start to overcome women and this increase becomes gradually more marked. The strong peak in favor of men in the range $(13, 29)$ is justified simply by the fact that men can donate twice as much as women, therefore it cannot be seen as a greater propensity of men to donate. Overall, women seem to have more consistent behavior in the donation process and this fact will be confirmed by posterior analysis in Chapter 5.

Tot donations	Donors' count	Tot percentage	Sex	Donors' count	Percentage
2	1608	27.08%	F	708	36.65%
			M	900	22.47%
3	1101	18.54%	F	428	22.15%
			M	673	16.8%
4	723	12.18%	F	263	13.61%
			M	460	11.49%
5	555	9.35%	F	189	9.78%
			M	366	9.14%
6	417	7.02%	F	103	5.33%
			M	314	7.84%
7	292	4.92%	F	68	3.52%
			M	224	5.59%
8	262	4.41%	F	60	3.11%
			M	202	5.04%
9	178	3%	F	42	2.17%
			M	136	3.4%
10	160	2.64%	F	30	1.55%
			M	130	3.25%
11	119	2%	F	19	0.98%
			M	100	2.5%
12	101	1.7%	F	13	0.67%
			M	88	2.2%
13	753	12.68%	F	3	0.15%
			M	72	1.8%
14-30	346	7.09%	F	6	0.31%
			M	340	10.29%

Table 2.4: Distribution of total donations per individual

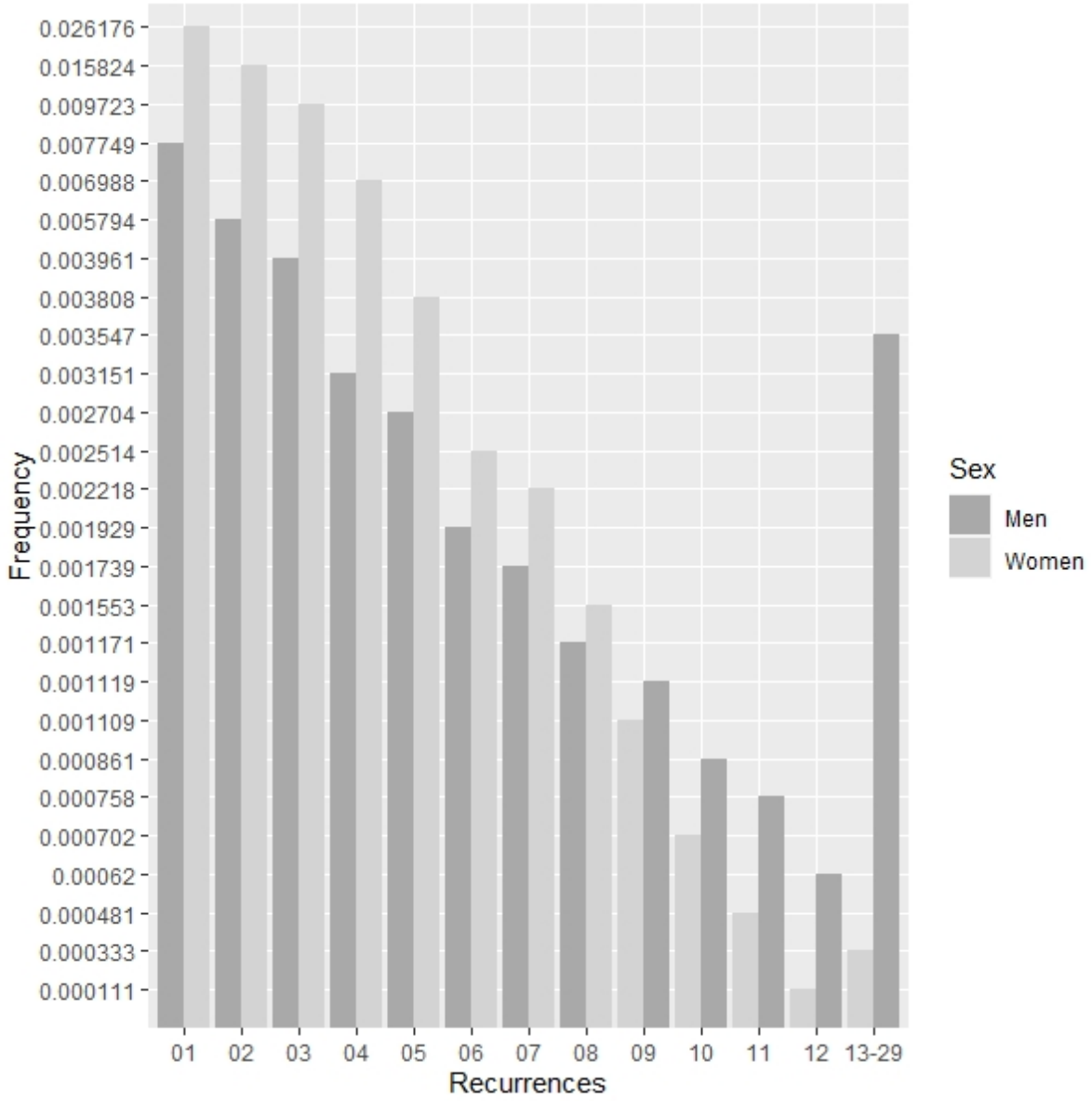

Figure 2.4: Barplot of total recurrences among male and female donors

2.3.3 Time-fixed covariates

In Table 2.5 all the categorical covariates are summarised with their sample frequencies. Some of them are objective (like sex, blood type or Rhesus factor), while the others are declared by the person her/him-self (smoke and alcohol habits and level of physical activity).

Variable	Value	Percentage
SESSO	F	32.52%
	M	67.48%
FUMO	Non-smoker	67.49%
	Smoker	32.51%
ALCOOL	Non-consumer	69.82%
	Consumer	30.18%
ATTIVITAFISICA	Active Life	76.03%
	Sedentary Life	23.97%
AB0	0	46.4%
	A	39.67%
	AB	1.36%
	B	12.57%
TIPO_RH	Positive	86.19%
	Negative	13.81%
DIETA	Balanced	97.17%
	Highly caloric	1.27%
	Lowly caloric	0.46%
	Vegetarian/Vegan	1.1%
STRESS	Absent	6.34%
	Negative 1	85.49%
	Negative 2	6.42%
	Negative 3	1.25%
	Positive	0.49%

Table 2.5: Sample frequencies of categorical variables

There are more men than women donors in the dataset, the majority of the population has blood type 0 and the positive Rhesus factor is more frequent than the negative one. Concerning living habits variables, it seems that donors have an healthy life. In fact there are more non-smokers than smokers, alcohol non-consumers than consumers and an active life is declared by most of the individuals. Covariates DIETA and STRESS do not seem to be useful for the analysis and they are immediately discarded, in fact almost all the donors declare to have a balanced diet (97.17%) and an absent level of stress (99.51%). Some descriptive statistics for continuous covariates are provided in Table 2.6.

Variable	Sample mean	Standard deviation	Min	Max
PESO (kg)	72.3502	13.3874	46	130
ALTEZZA (m)	1.74	0.0912	0.89	2.02
ETA_PRIMA (y)	32.20	9.9444	18	62
ETA_DONAZ (y)	35.79	10.4281	18	65.9

Table 2.6: Summary statistics of continuous variables

In Spinelli (2019) useful boxplots are done, in order to discover some correlation pattern between the continuous and categorical variables. However it is not clear from those graphs if a significant correlation exists. On the other hand, it can be clearly noticed that the rates' distribution reaches higher values in males than in female donors. This was expected since, according to law, men have the double of the possibilities to donate with respect to women. No other correlations are evident.

2.4 Data transformation

Most of the features are categorical variables with many levels. To be suitable to a statistical model, they are transformed into binary dummy variables:

- covariate FUMO takes the value 1 if the donor is a smoker, 0 if he/she is not;
- covariate ALCOOL takes value 1 if the donor declares to consume alcoholic beverages, 0 otherwise;
- covariate STRESS takes value 1 if the donor claims to be stressed, 0 otherwise;
- covariate ATTIVITAFISICA takes value 0 if the donor declares to have a sedentary lifestyle, or if his/her level of physical activity is low or irregular, 1 otherwise;
- blood type is represented by three dummy variables (TIPO 0, TIPO B, TIPO AB). For instance (1,0,0) is blood type 0 and so on. Blood type A is considered as baseline.

As regards the numerical features PESO (weight in kilograms) and ALTEZZA (height in meters), the Body Mass Index (BMI) is computed as BMI = PESO/ALTEZZA^2. For adults, BMI is interpreted using standard categories shown in Table 2.7, along with the corresponding donors' percentages in the dataset, stratified by the donors' gender. It can be noticed that women are the majority in the category "Underweight" and that men have highest percentages in all the others. This leads to believe that the addition in the model of the interaction between donor's gender and BMI could be significant.

BMI	Weight status	Tot percentage	Sex	Percentage
Below 18.5	Underweight	1.44%	F	64.71%
			M	35.29%
18.5 – 24.9	Healthy Weight	69%	F	32.29%
			M	67.70%
25.0 – 29.9	Overweight	24.84%	F	18.49%
			M	81.51%
30.0 and above	Obese	4.72%	F	27.70%
			M	72.30%

Table 2.7: Summary of donors' BMI

2.4.1 Time-dependent covariates and absurd values

Time-dependent covariates introduced in this thesis are: hemoglobin, heart rate, minimum pressure, maximum pressure. After a preliminary analysis, some absurd values have been discovered and replaced with "NA". The imputation of these values, along with the already existing missing values, will be discussed in Chapter 3. Absurd values do not comply with the general behavior of the data, so they appear as anomalous. They have been detected through manual inspection and knowledge of reasonable values. Table 2.8 summarizes the number of absurd values for each covariate and the summary statistics updated after absurd values' removal.

Variable	Number of absurd values	Sex	Mean	Standard deviation	Min	Max
EMOG	12	F	13.5528	0.7659	12.1	17.2
		M	15.2765	0.8717	13.1	18.9
POLSO	9	F	68.6634	7.3966	45	120
		M	66.7589	7.7300	40	125
PMIN	2	F	72.2098	7.0280	55	120
		M	75.8305	7.2481	50	120
PMAX	11	F	115.088	6.7889	100	165
		M	119.043	8.3295	100	186
BMI	16	F	22.4645	3.3140	13.3	44.6
		M	24.5583	3.1507	15.5	46.3

Table 2.8: Summary statistics after absurd values' removal

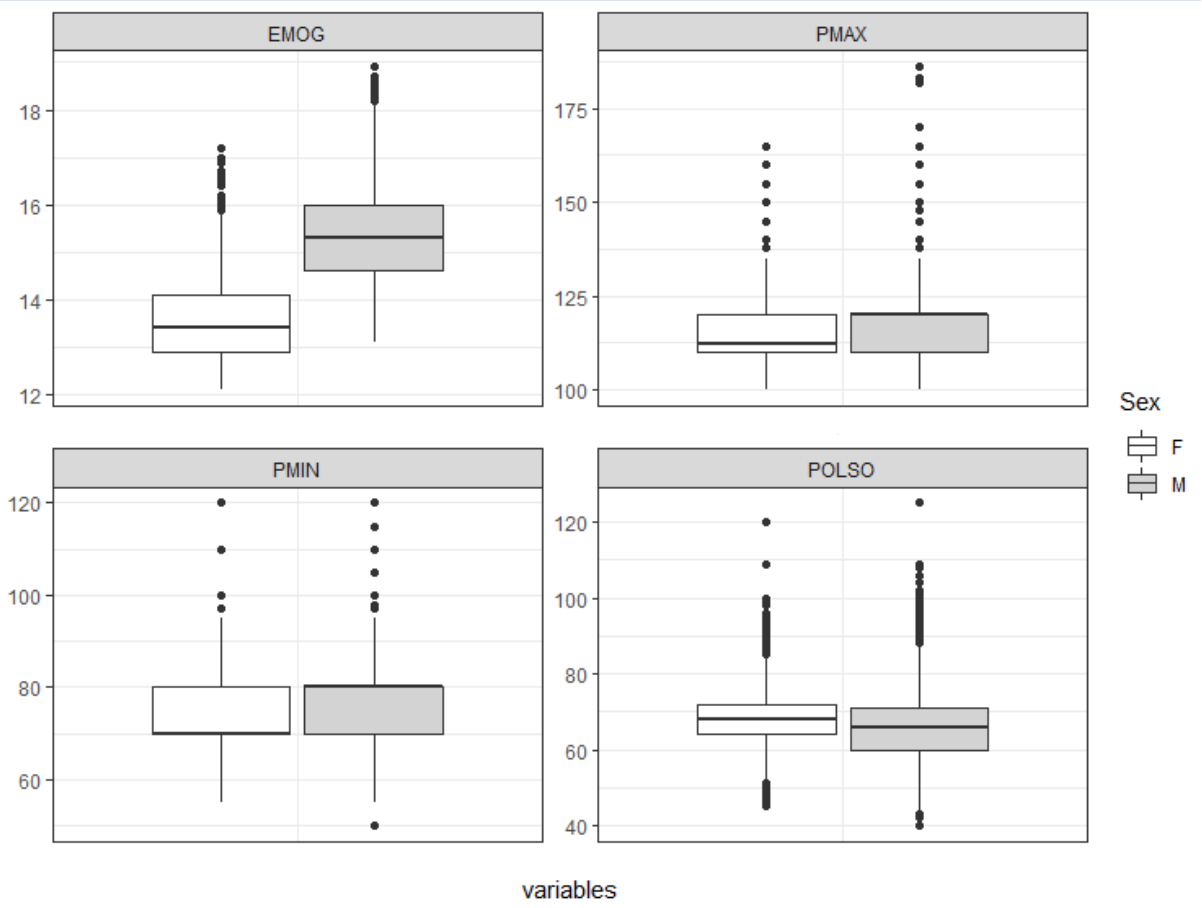

Figure 2.5: Boxplot of time-dependent covariates, divided in male and female donors

Finally, it is essential to understand whether a clear across-time variability in time-dependent covariates is present or not. Indeed if this is not true, it cannot be said that these covariates are really time-dependent and they should be considered time-fixed. Figure 2.6 and 2.7 show that women have a lower value of hemoglobin than men. Apart from this information, by overlapping the trends of all donors, they become indistinguishable and so difficult to interpret. Hence, less donors have been randomly selected several times, by repeating the drawing of the matplots each time; here just one of the resulting plots is reported, as a representative of the general discovered trends. By analyzing Figure 2.7, it is clear that the two covariates on the left, hemoglobin and minimum pressure, change more over time with respect to the other two. This fact has important implications on the modelling choices explained in Chapter 4.

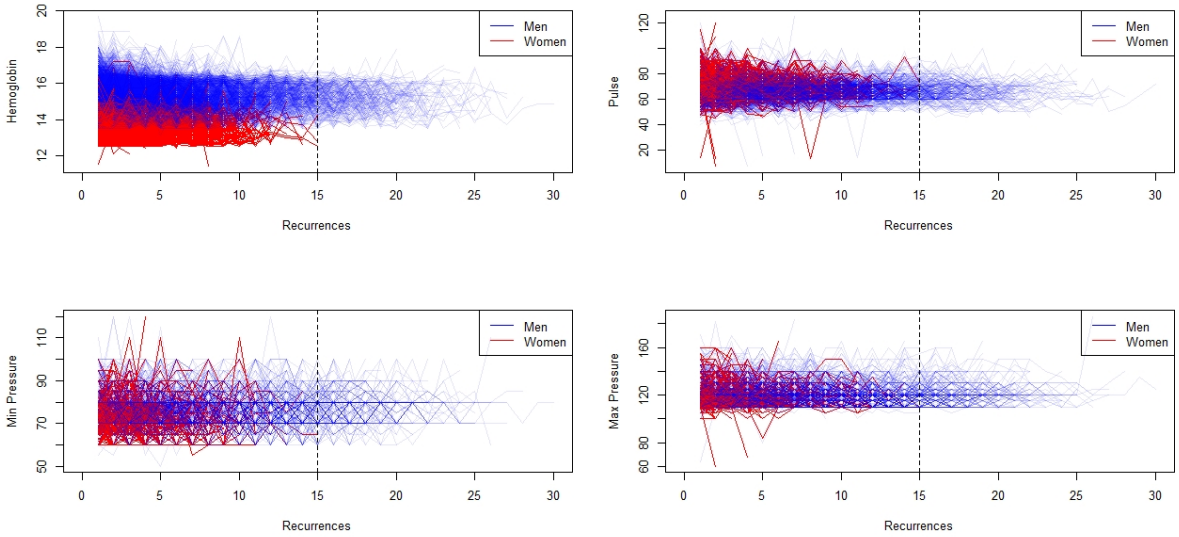

Figure 2.6: Matplot of time-dependent covariates considering all donors

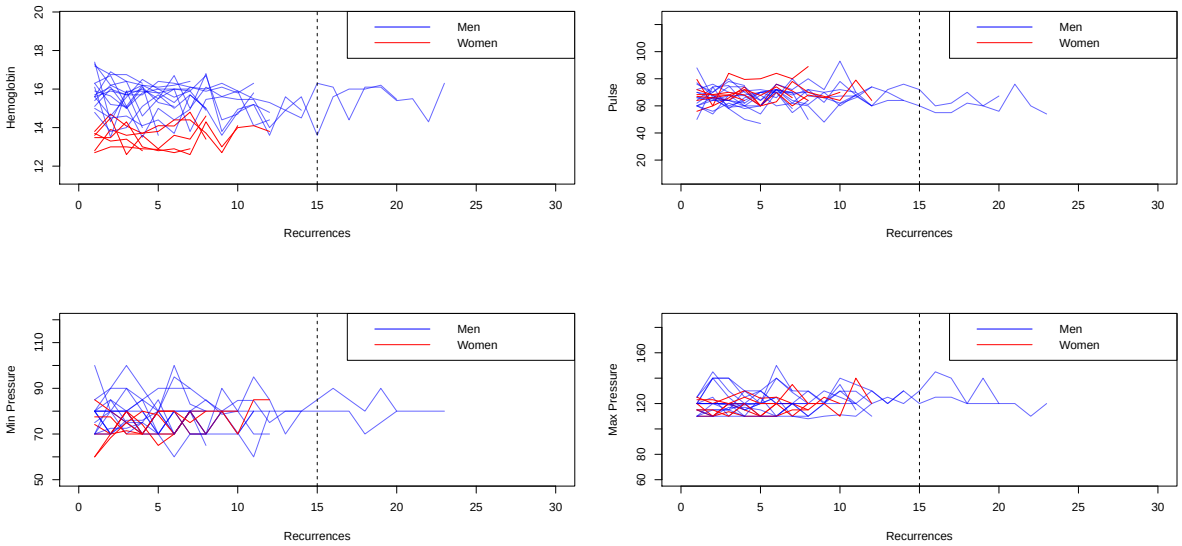

Figure 2.7: Matplot of time-dependent covariates for randomly selected donors

Chapter 3

Missing values

In this chapter, the missing values' imputation is performed. As a result, we obtain a complete dataset ready to be used.

3.1 Overview

When missing values are present in a dataset, it is important to understand why they are missing and their impact on the analysis. Sometimes ignoring missing data biases answers and potentially misleads to incorrect conclusions. Rubin (1976) differentiates between three types of missingness mechanisms:

1. **missing completely at random** (MCAR): the distribution of missing values does not depend on observed attributes or missing value (e.g., survey questions randomly sampled from larger set of possible questions);
2. **missing at random** (MAR): the distribution of missing values depends on observed attributes, but not on the missing value (e.g., men less likely than women to respond to question about mental health);
3. **missing not at random** (MNAR): the distribution of missing values depends on the missing value (e.g., respondents with high income less likely to report it). This is difficult to handle, because it requires strong assumptions about the patterns of missingness.

The most common techniques to deal with missing values are the following:

- **deletion methods**: listwise deletion, pairwise deletion;
- **single imputation methods**: mean/mode substitution, dummy variable method, single regression;

- **model-based methods**: maximum likelihood, multiple imputation.

Typically people delete all cases for which a value is missing. This method is called **complete case analysis** (CC). However, CC is valid only if data is MCAR and this case occurs rarely in practice. Another method is **multiple imputation** (MI), which is a Monte Carlo method that simulates multiple values to impute (fill-in) each missing value, then it analyses each imputed dataset separately and finally it pools the results together. Missing data are imputed multiple times to account for the uncertainty about the true (and unknown) values of the missing data. In theory, MI can handle all the three types of missingness. On the other hand, software packages that do MI are usually not designed for MNAR case. Missing data analysis of MNAR data is more complicated and often requires domain knowledge.

3.2 Missing values analysis on data

Figure 3.1 and Figure 3.2, obtained with `VIM` package in R, represent on the left the amount of missing values and on the right the way in which they are coupled together along the different observations. Missing values are represented in red and the non missing in blue.

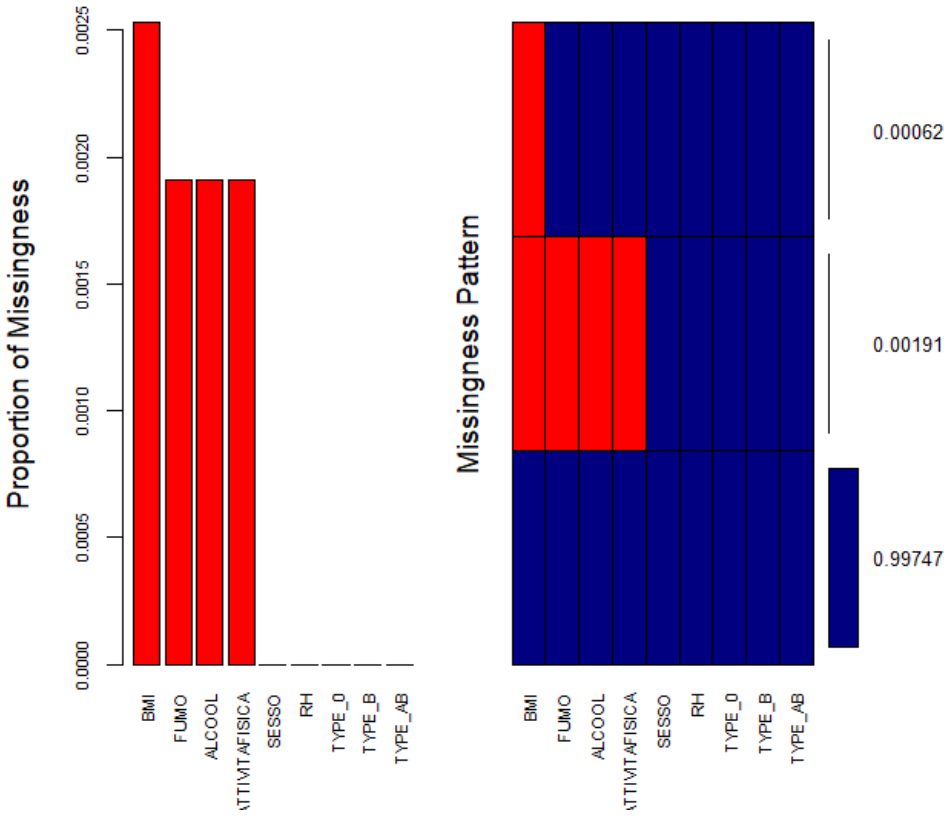

Figure 3.1: Missing values, part 1

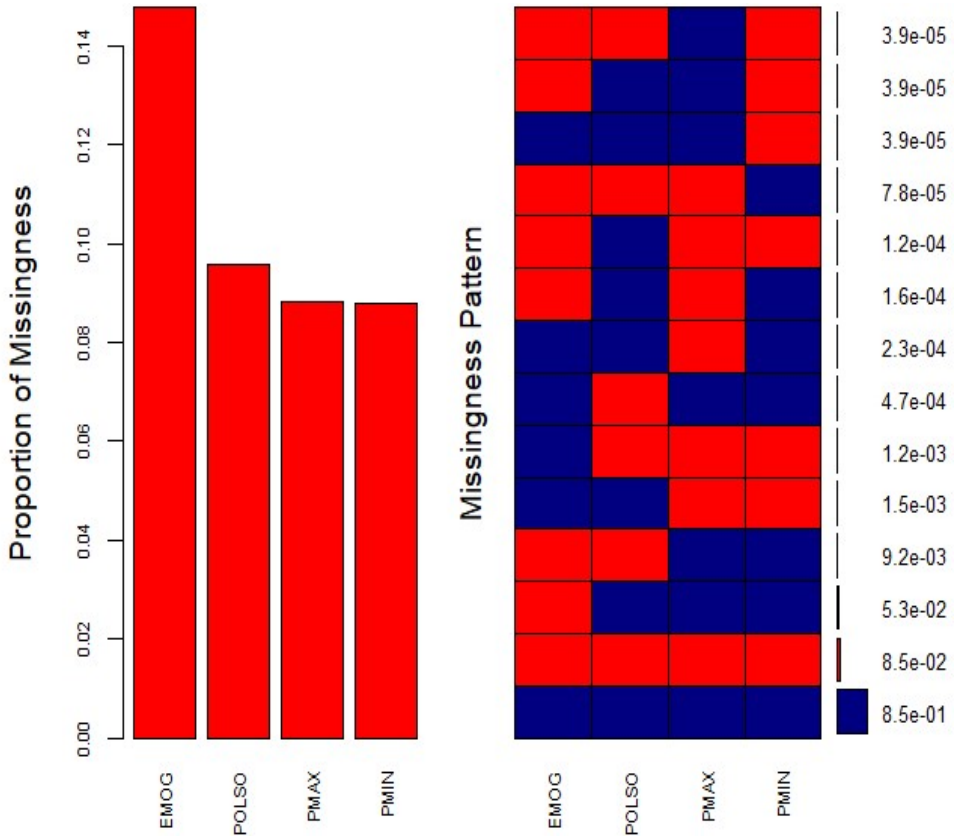

Figure 3.2: Missing values, part 2

Variables sorted by decreasing number of missing values are shown in Table 3.1:

Variable	**Sample percentage**	**Tot**	**Unique donors**	**Sex**	**Count**	**Mean age**
THE CAFFE	40.43%	25689	1832	F M	573 1259	33.4 34.5
EMOG	14.72%	3794	2370	F M	606 1764	33.8 34.9
POLSO	9.53%	2459	1617	F M	427 1190	32.8 34.6
PMIN	8.76%	2253	1480	F M	399 1081	32.7 34.5
PMAX	8.75%	2262	1487	F M	402 1085	32.7 34.6
BMI	0.25%	65	42	F M	17 25	33.3 29.8
FUMO ALCOOL ATTIVITAFISICA	0.19%	49	37	F M	15 22	34.5 29.6

Table 3.1: Missing values

The covariates SESSO, RH, ETA_PRIMA, ETA_DONAZ, TIPO_A, TIPO_0, TIPO_B, TIPO_AB have no missing values. Instead the covariates THE and CAFFE have too many missing values, so they are immediately discarded.

3.3 Assumptions' check for multiple imputation

It is important to understand why data are missing and the category they belong to among MCAR, MAR, MNAR. The focus is on hemoglobin, pressures and pulse, which contain a non-negligible number of missing values. As preliminary analysis, daily counts are performed. As shown in Figure 3.3, it can be noticed that all covarites have a peak of NA's in 2013 and hemoglobin has a lower peak even in 2018. Since measurements of these four values are mandatory at each donation, their lack may be due to data collection issues.

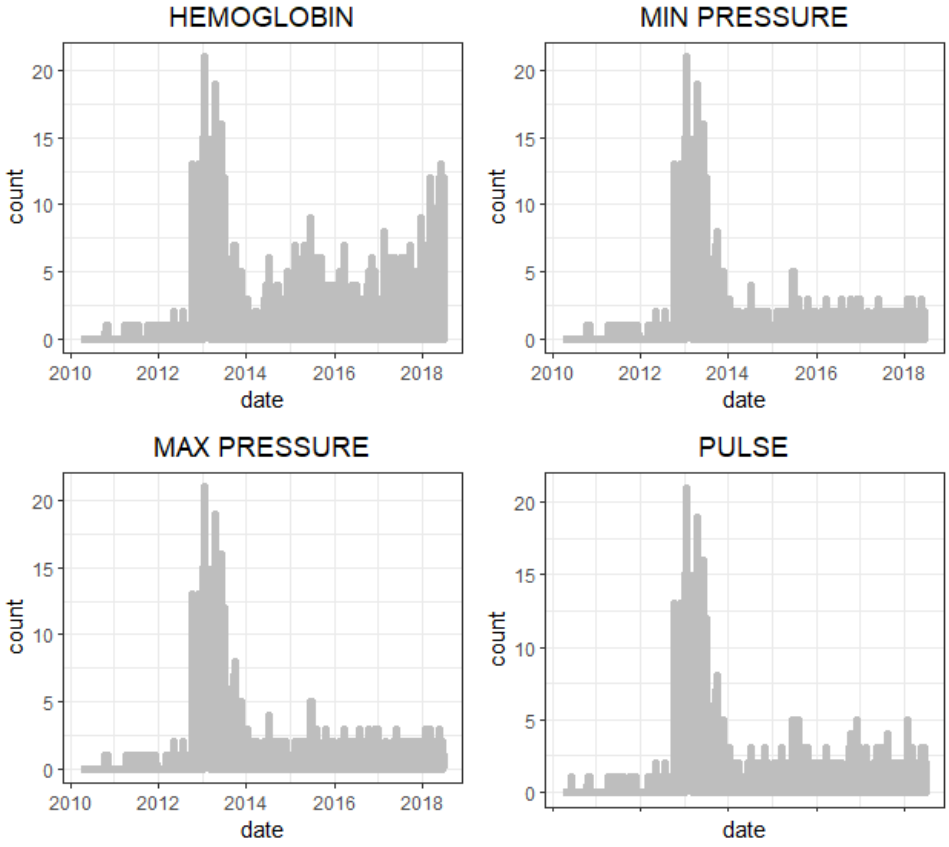

Figure 3.3: Daily counting of missing values

At first glance, this can be seen as a limiting aspect of a multiple imputation, since the presence of these NA's is closely linked to the year of observation. However, referring to recurrent events' framework, their randomness during time is restored. Indeed new donors are always allowed to arrive and time $t = 0$ refers to their first donation (that is the start of the counting process). Figure 3.4 shows the daily count of new donors. The 95.3% of the daily counts is below the

threshold (count = 6), represented by the dashed line.

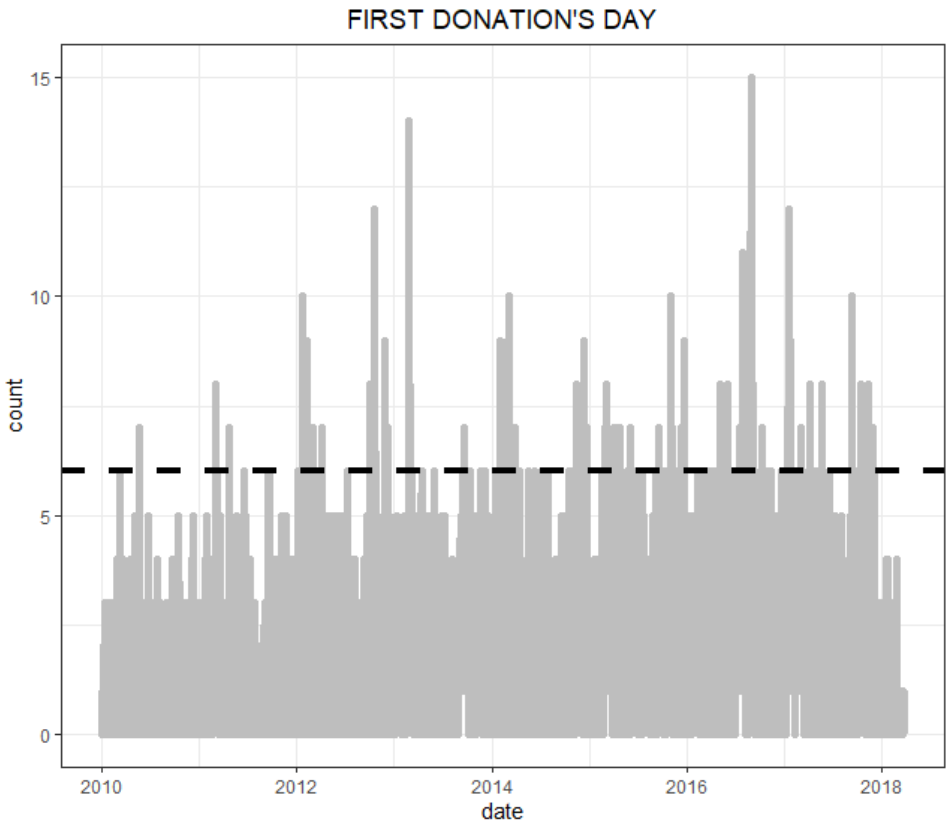

Figure 3.4: Daily counting of new arrivals

3.4 `MICE` package in R

R language has robust packages for missing value imputations. `MICE` (**Multivariate Imputation via Chained Equations**) is one of the most used and it assumes that the missing data are MAR. It imputes data by specifying an imputation model per variable and imputation is drawn from the conditional distribution by MCMC techniques (the details are explained in Appendix B). For example, let $X_1, X_2, \ldots X_k$ be random variables and suppose X_1 has missing values. Hence X_1 will be regressed on other variables X_2 to X_k and the missing values in X_1 will be replaced by the estimated predictive values. Similarly, if X_2 has missing values, then X_1, X_3 to X_k variables will be used in prediction model as independent variables; later, missing values will be replaced with predicted values and so on. The `mice()` function executes m streams in parallel, each of which generates one imputed dataset and these datasets are equal, except for imputed values. In particular, each imputed value is the last one of simulated chains. The final complete dataset is obtained by merging together the m imputed datasets. In order to address the issues posed by the real-life complexities of the data, it is convenient to specify the imputa-

tion model separately for each column in the data. Some of the imputation methods provided by `MICE` are reported in Table 3.2:

Method	**Type of variable**	**Description**
`pmm`	any	predictive mean matching
`cart`	any	classification and regression trees
`rf`	any	random forests imputation
`norm`	num	Bayesian linear regression
`norm.boot`	num	linear regression using bootstrap
`logreg`	num	logistic regression
`lda`	cat	linear discriminant analysis

Table 3.2: Imputation methods by `MICE`

To sum up, `MICE` package contains functions to:

1. inspect the missing data pattern;
2. impute the missing data m times, resulting in m completed datasets;
3. diagnose the quality of the imputed values;
4. analyze each completed dataset;
5. pool the results of the repeated analyses;
6. store and export the imputed data in various formats;
7. generate simulated incomplete data;
8. incorporate custom imputation methods.

3.5 Missing values imputation

In this thesis, $m = 5$ imputed datasets and a maximum number of 100 iterations are fixed. The covariates with missing values are imputed with the methods listed below, which have been selected after several trials. The remaining variables RH, TIPO_0, TIPO_B, TIPO_AB, SESSO, ETA_PRIMA and ETA_DONAZ have no missing values and they are used only to impute the others.

- BMI: random forest;
- FUMO: logistic regression;

- ALCOOL: logistic regression;
- ATTIVITAFISICA: logistic regression;
- EMOG: CART;
- POLSO: random forest;
- PMIN: random forest;
- PMAX: random forest.

Logistic regression is the go-to method for binary classification problems. **CART** is a useful nonparametric technique that can be used to explain a continuous or categorical dependent variable in terms of multiple independent variables. The independent variables can be continuous or categorical. CART employs a partitioning approach generally known as "divide and conquer". It is easy to use and can quickly provide valuable insights into massive amounts of data. **Random forest** creates multiple CART trees based on "bootstrapped" samples of data and then combines the predictions, where a bootstrap sample is a random sample conducted with replacement. Usually, the combination is an average of all the predictions from all CART models. Random Forest has better predictive power and accuracy than a single CART model (because random forest exhibit a lower variance), but it is more complex in terms of computations. Unlike the other continuous covariates, CART method works well for the hemoglobin, so it is enough for the goal set. This may be related to the fact that hemoglobin's values have a more regular density than the other continuous covariates, in particular maximum and minimum pressure, whose density is more complex and therefore its reproduction on missing values requires a more computationally expensive method.

3.6 Inspecting the distribution of original and imputed data

Let us now compare the distributions of original and imputed data using some useful plots. Three plots are considered: `densityplot()`, `stripplot()` and `xyplot()`. In particular:

- `densityplot()`: the density of the imputed data for each imputed dataset is showed in red, while the density of the observed data is showed in blue. Under previous assumptions, the distributions are expected to be similar;
- `stripplot()`: the function shows the distributions of the variables as individual points among the five imputed datasets. Notice that, in the plots that follow, the red points follow the blue points reasonably well;

- `xyplot()`: the function produces conditional scatterplots. It extends the usual features of `LATTICE` package and automatically separates observed (blue) from imputed (red) data.

The densityplots and stripplots in Figures 3.5-3.17 and the xyplots in Figures 3.18, 3.19 conclude the section.

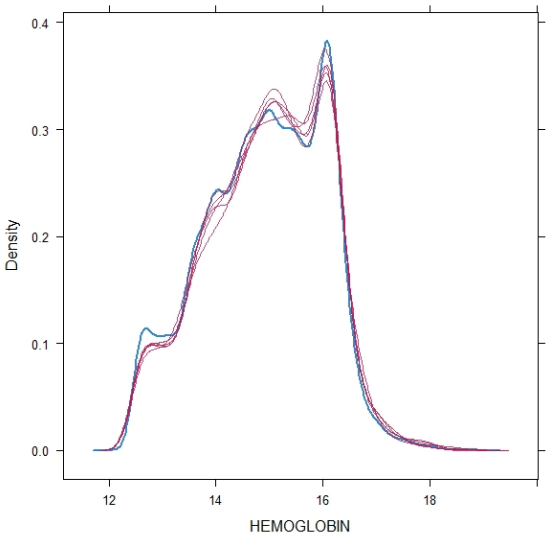

Figure 3.5: Densityplot hemoglobin

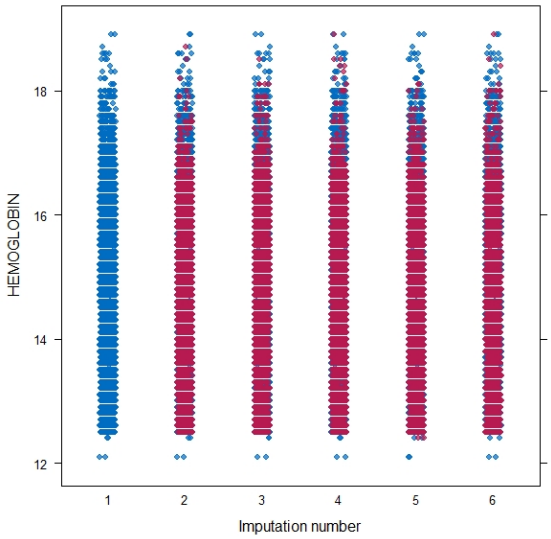

Figure 3.6: Stripplot hemoglobin

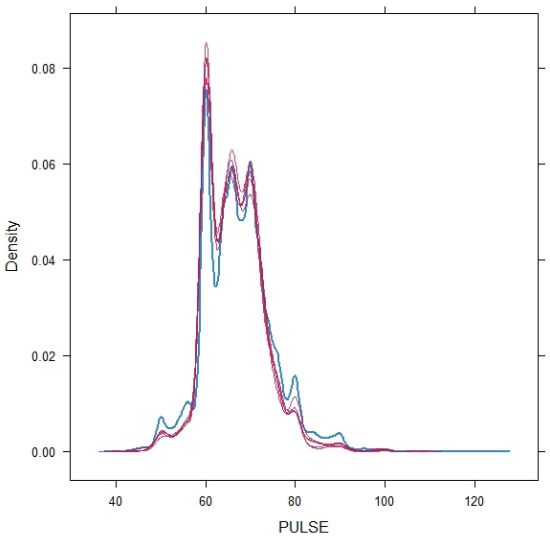

Figure 3.7: Densityplot pulse

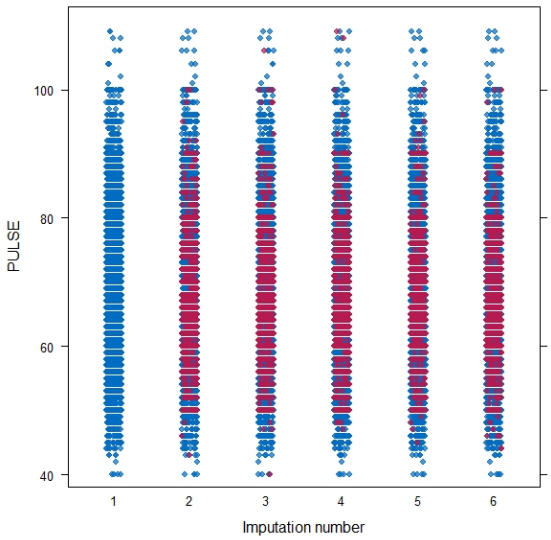

Figure 3.8: Stripplot pulse

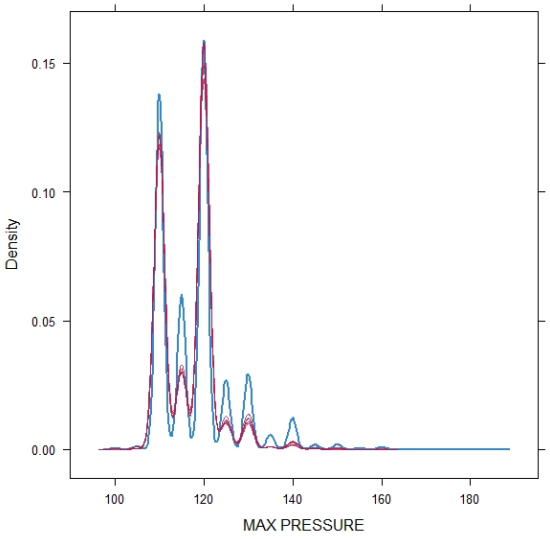

Figure 3.9: Densityplot max pressure

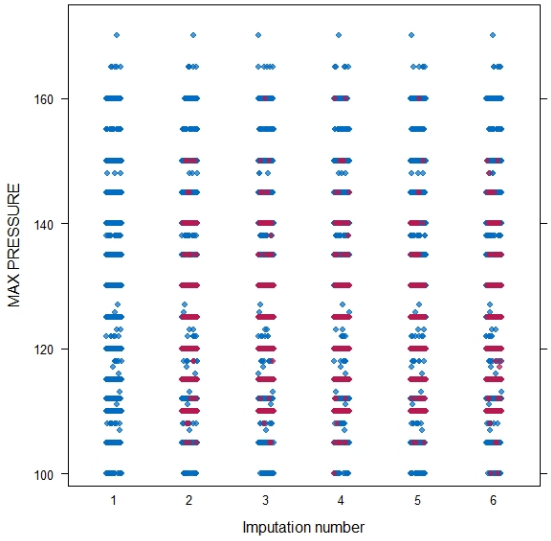

Figure 3.10: Stripplot max pressure

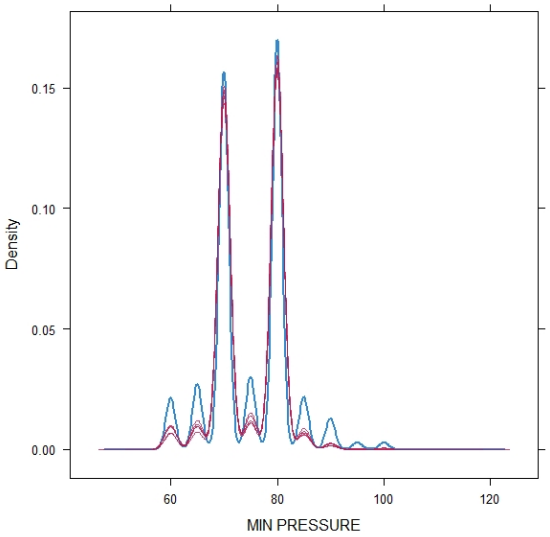

Figure 3.11: Densityplot min pressure

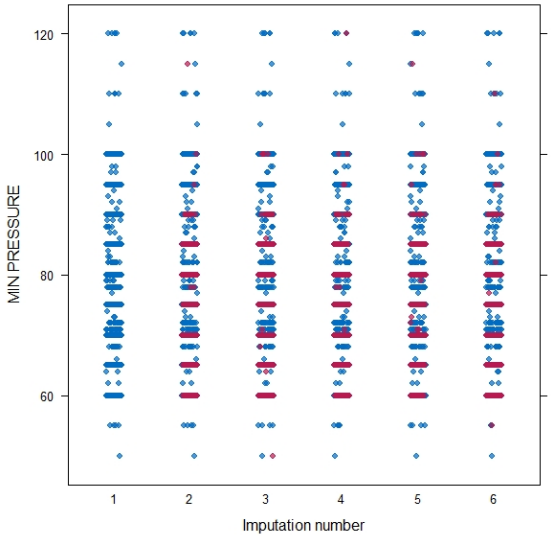

Figure 3.12: Stripplot min pressure

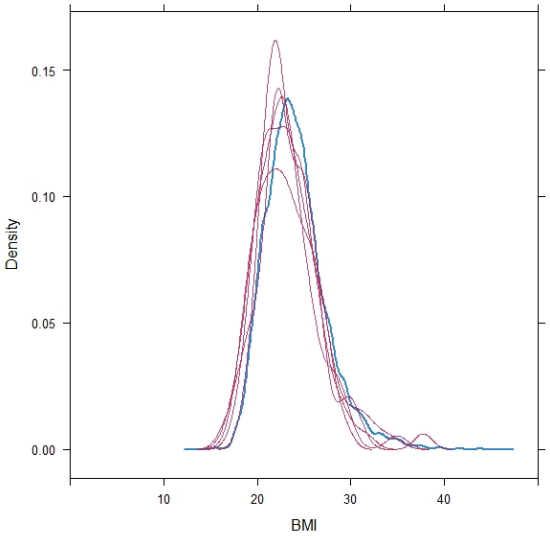

Figure 3.13: Densityplot BMI

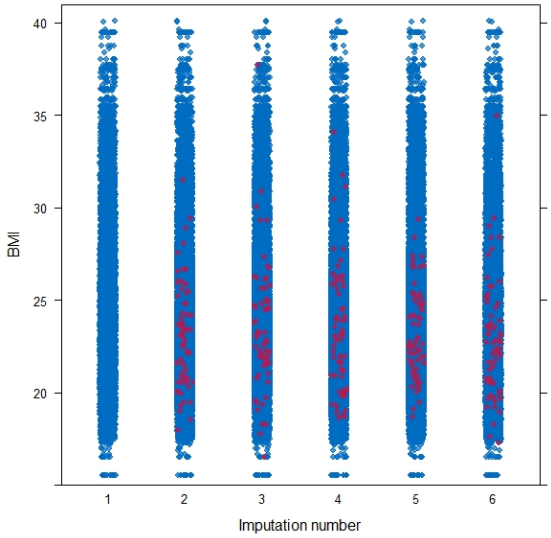

Figure 3.14: Stripplot BMI

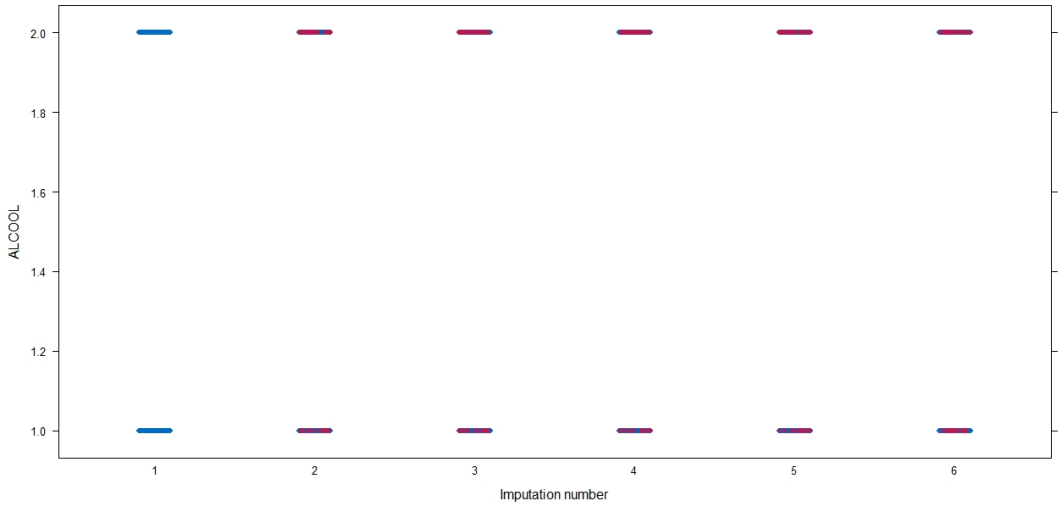

Figure 3.15: Stripplot alcool

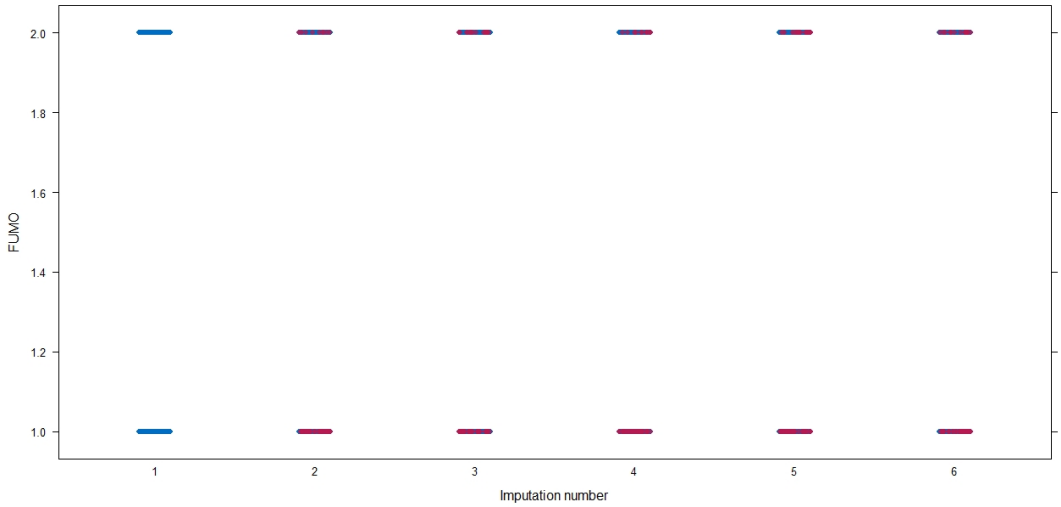

Figure 3.16: Stripplot fumo

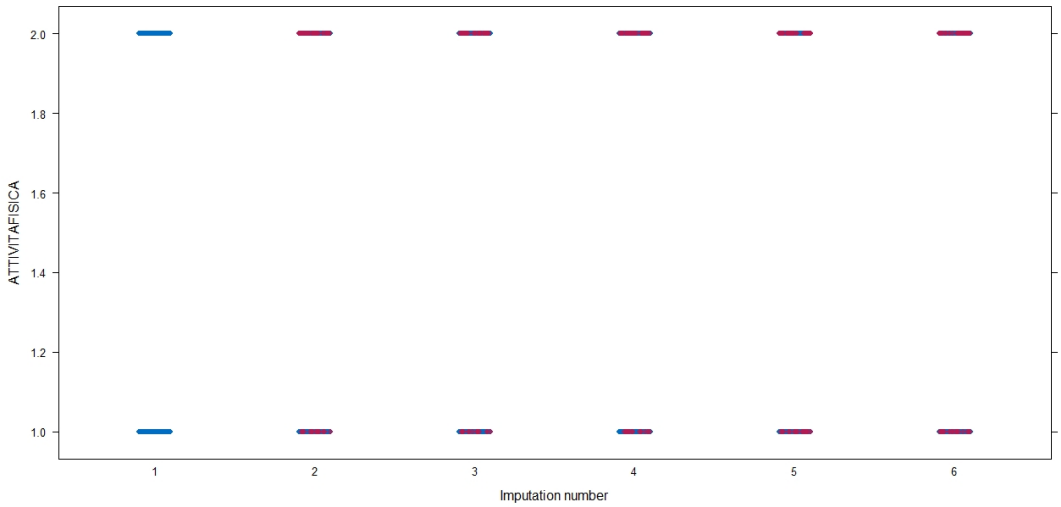

Figure 3.17: Stripplot attività fisica

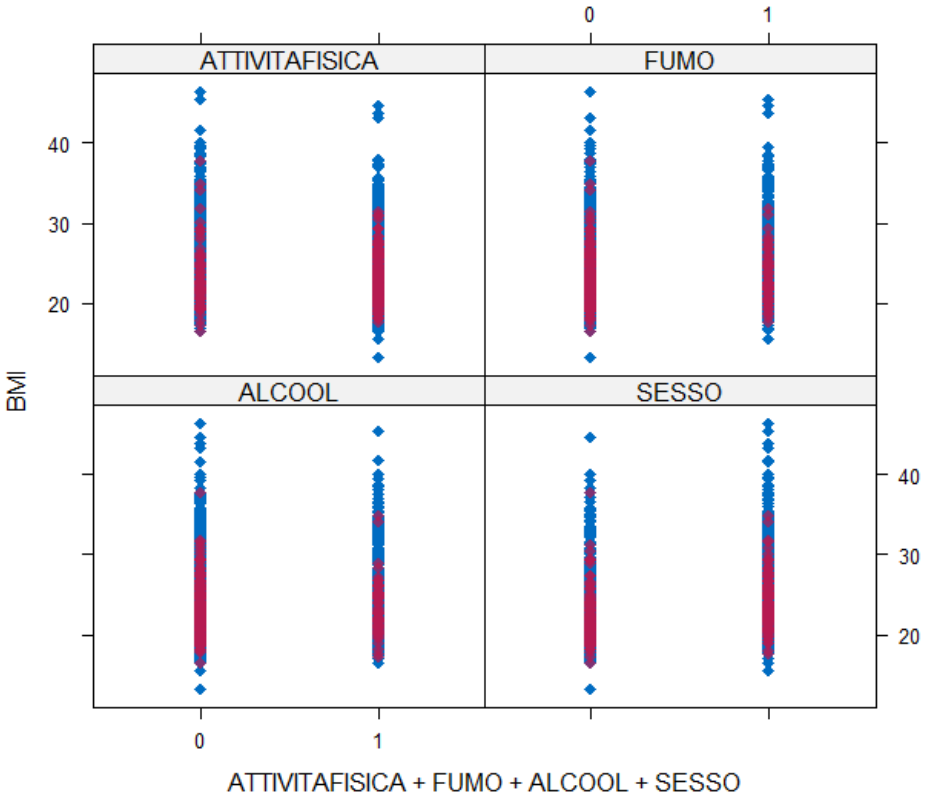

Figure 3.18: xyplot BMI

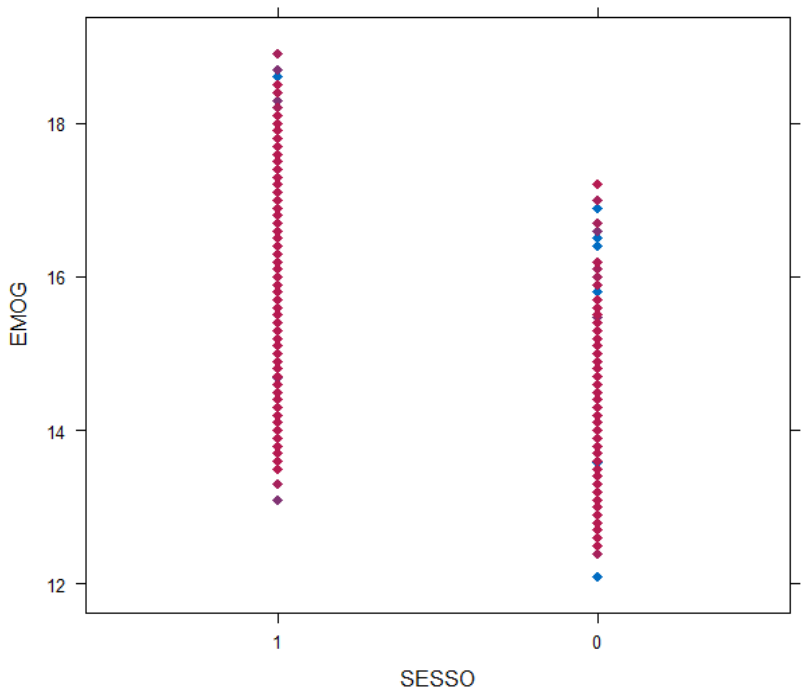

Figure 3.19: xyplot hemoglobin

3.7 Convergence monitoring

`MICE` runned $m = 5$ parallel chains, each with a certain number of iterations, and imputes values from the final iteration. To monitor convergence we use `traceplot()`, which plots estimates against the number of iteration. Figure 3.20 shows mean and standard deviation of the covariates through the 100 iterations for the 5 imputed datasets. An indicator of convergence is how well the 5 parallel chains mix.

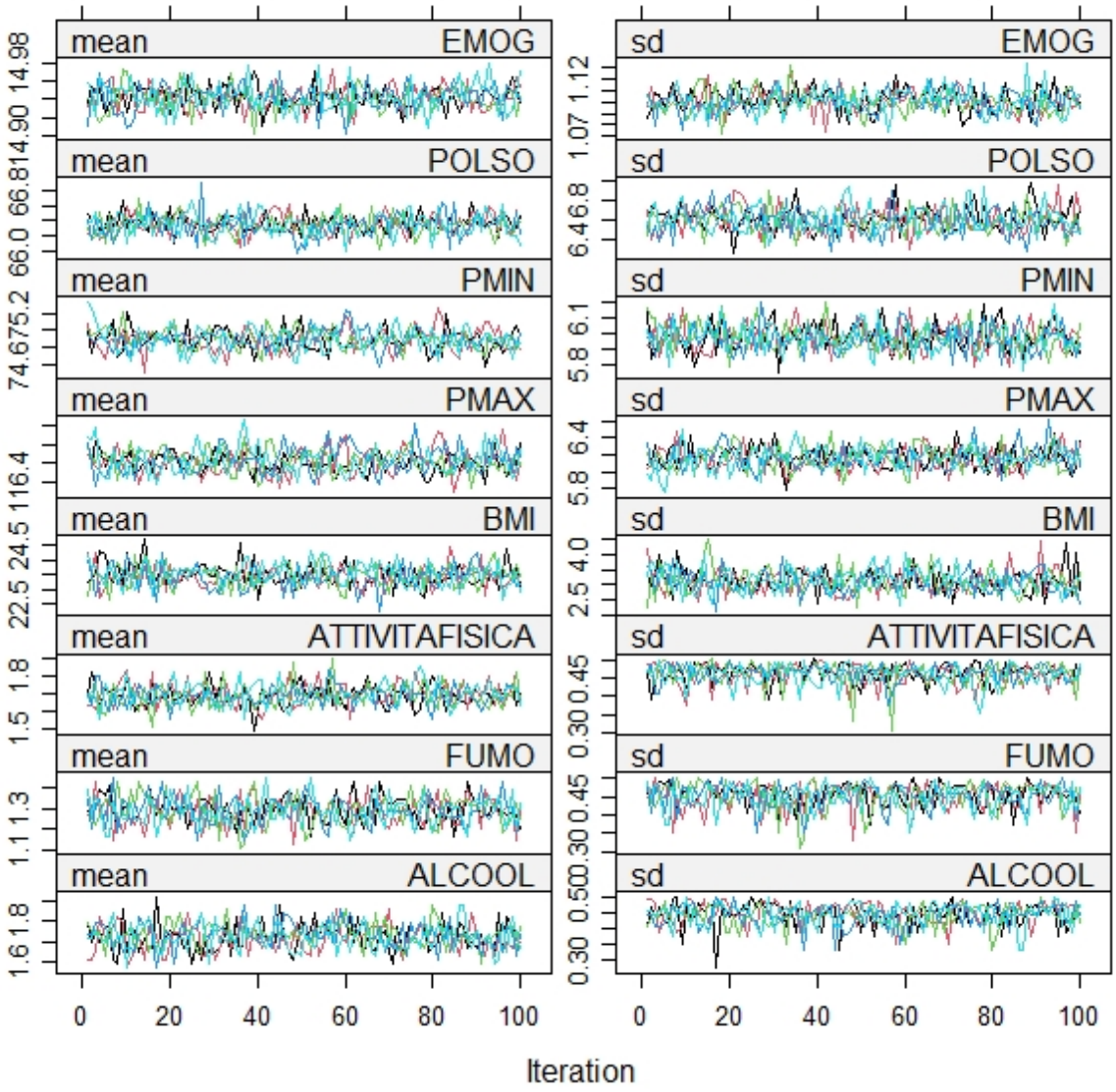

Figure 3.20: Traceplots for imputed covariates

Chapter 4

Bayesian models of recurrent events for blood donations

In this chapter we exploit how blood donations can modeled as recurrent events. Three different models are considered, starting from an initial "Model 0", two natural but different extensions of it, "Model 1" and "Model 2", are proposed.

4.1 Model 0

Model 0 is a base case, which prepares the ground for subsequent modeling extensions. It generalizes a model studied by Spinelli (2019), including time-dependent covariates in addition to time-fixed covariates.

4.1.1 Likelihood

Now we exploit the likelihood of *Model 0*. All the steps shown here are also a guideline for obtaining next models' likelihood, which may differ from *Model 0* for some secondary aspects. The theoretical background is here summarized.

- The time scale is:
  - $t = 0$: first whole blood donation;
  - t: number of days passed since the first donation.
- For any donor $i = 1, \dots, I$:
  - n_i is his/her total number of donations;
  - $t_{i,1}, \dots, t_{i,n_i}$ are the times at which the donations take place;

- $Y_i(t)$ is the at-risk indicator function;
- $N_i(t) = \sum_{k=1}^{\infty} I\{t_{ik} \leq t\}$ is the number of donations up to t (counting process);
- $\Delta N_i(t) = N(t + \Delta t^-) - N(t^-)$ is the number of donations in $[t, t + \Delta t)$;
- $H_i(t) = \{N_i(s) : 0 \leq s < t\}$ is the history up to time t;
- the time-varying covariates are simple/step functions:

$$\mathbf{x}_i(t) = \mathbf{x}_i(t_{ij}) \qquad t_{ij} \leq t < t_{i,j+1} \qquad j = 1, \ldots, n_i \tag{4.1}$$

  where $t_{i,n_i+1} = c_i$ and c_i is the censoring time for donor i;
- the intensity function of the recurrent process of donor i is:

$$\lambda_i(t|H_i(t)) = \lambda_0(t) u_i \exp\{\mathbf{x}_i(t)'\boldsymbol{\beta}\} \tag{4.2}$$

- By Theorem 1 of Chapter 1, which plays here an essential role, the likelihood of the single donor i is:

$$\begin{aligned}
\mathscr{L}_i &= \left(\prod_{j=1}^{n_i} \lambda_i(t_{ij}|H_i(t_{ij})) \right) \exp\left\{ -\int_0^{\infty} Y_i(s)\lambda_i(s|H(s))ds \right\} \\
&= \left(\prod_{j=1}^{n_i} \lambda_0(t_{ij}) u_i e^{\mathbf{x}_i'(t_{ij})\boldsymbol{\beta}} \right) \exp\left\{ -\int_0^{\infty} Y_i(s)\lambda_0(s) u_i e^{\mathbf{x}_i'(s)\boldsymbol{\beta}} ds \right\} \\
&= \left(\prod_{j=1}^{n_i} \lambda_0(t_{ij}) \right) u_i^{n_i} \exp\left\{ \sum_{j=1}^{n_i} \mathbf{x}_i'(t_{ij})\boldsymbol{\beta} - u_i \int_0^{\infty} Y_i(s)\lambda_0(s) e^{\mathbf{x}_i'(s)\boldsymbol{\beta}} ds \right\}
\end{aligned}$$

The total likelihood is obtained by product:

$$\mathscr{L} = \prod_{i=1}^{I} \mathscr{L}_i = \left(\prod_{i=1}^{I} \prod_{j=1}^{n_i} \lambda_0(t_{ij}) \right) \left(\prod_{i=1}^{I} u_i^{n_i} \right) \exp\left\{ \sum_{i=1}^{I} \sum_{j=1}^{n_i} \mathbf{x}_i'(t_{ij})\boldsymbol{\beta} - \sum_{i=1}^{I} u_i \int_0^{\infty} Y_i(s)\lambda_0(s) e^{\mathbf{x}_i'(s)\boldsymbol{\beta}} ds \right\}$$

Let us now define $w_k(t) := \mathrm{I}_{(a_{k-1}, a_k]}(t)$ to simplify the notation and substitute $\lambda_0(t) = \sum_{k=1}^{K} \lambda_k w_k(t)$ in $\mathscr{L}$. Hence:

- the first factor of $\mathscr{L}$ becomes:

$$\prod_{i=1}^{I} \prod_{j=1}^{n_i} \lambda_0(t_{ij}) = \prod_{i=1}^{I} \prod_{j=1}^{n_i} \sum_{k=1}^{K} \lambda_k w_k(t_{ij}) = \prod_{k=1}^{K} \lambda_k^{\sum_{i=1}^{I} \sum_{j=1}^{n_i} w_k(t_{ij})} = \prod_{k=1}^{K} \lambda_k^{n_{\cdot k}}$$

where $n_{\cdot k} = \sum_{i=1}^{M} \sum_{j=1}^{n_i} w_k(t_{ij})$ is the total number of donations in the interval $(a_{k-1}, a_k]$;

- the last factor (integral) in $\mathscr{L}$ becomes:

$$\sum_{i=1}^{I} u_i \int_0^{\infty} Y_i(s) \lambda_0(s) e^{\mathbf{x}_i'(s)\boldsymbol{\beta}} ds =$$
$$\sum_{i=1}^{I} u_i \int_0^{\infty} Y_i(s) \sum_{k=1}^{K} \lambda_k w_k(s) e^{\mathbf{x}_i'(s)\boldsymbol{\beta}} ds =$$
$$\sum_{i=1}^{I} \sum_{k=1}^{K} u_i \lambda_k \int_{a_{k-1}}^{a_k} Y_i(s) e^{\mathbf{x}_i'(s)\boldsymbol{\beta}} ds$$

In conclusion we obtain:

$$\mathscr{L} = \left(\prod_{k=1}^{K} \lambda_k^{n_{\cdot k}} \right) \left(\prod_{i=1}^{I} u_i^{n_i} \right) \exp \left\{ \sum_{i=1}^{I} \sum_{j=1}^{n_i} \mathbf{x}_i'(t_{ij}) \boldsymbol{\beta} - \sum_{i=1}^{I} \sum_{k=1}^{K} u_i \lambda_k \int_{a_{k-1}}^{a_k} Y_i(s) e^{\mathbf{x}_i'(s)\boldsymbol{\beta}} ds \right\}$$

4.1.2 Covariates

The covariates are included in the intensity function through a multiplicative model, defined as:

$$g(\mathbf{x}(t); \boldsymbol{\beta}) = \exp\{\mathbf{x}'(t)\boldsymbol{\beta}\} \tag{4.3}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_P)$ is a vector of regression parameters, with P equal to the number of covariates. In this analysis, both fixed-time and time-dependent donor-specific covariates are taken into account. The maximum number of covariates is 18 (including interactions), then any nested models (with less covariates) are compared thanks to goodness-of-fit indicators, eliminating those that are not relevant as a result of posterior analysis. The complete set of covariates is reported in Table 4.1 which includes also some possibly relevant interactions.
The most important steps performed at this stage are the following:

- as a consequence of observations in Section 2.4.1, only hemoglobin and minimum pressure are kept time-dependent, while all the other covariates are maintained time-fixed;
- all values of the maximum pressure and heart rate are replaced with the values measured at first donation;
- the numerical variables are standardized, in order to deal with tractable values and to avoid the unit of measure to play an ambiguous role in the obtained results;
- the categorical variable blood type with four levels (0, A, B, AB) is transformed in four

binary variables through one-hot encoding. Type 0 is chosen as a baseline, so its column is not included in the dataset for running the model;

- interactions:
  - between the donors' gender and the hemoglobin,
  - between the donors' gender and the Rhesus factor,
  - between the donors' gender and Body Mass Index,

  are added, in order to exploit their possible relevance in explaining the phenomenon in a different way for men and women.

Name	**Type**	**Description**
SESSO	binary	gender: 1 male, 0 female
ETA_PRIMA	num	age at the time of first donation
FUMO	binary	smoker: 1 yes, 0 no
ALCOOL	binary	alcohol consumption: 1 yes, 0 no
ATTIVITAFISICA	binary	active life: 1 yes, 0 no
RH	binary	rhesus factor: 1 positive, 0 negative
TIPO_0	binary	blood type 0
TIPO_A	binary	blood type A
TIPO_B	binary	blood type B
TIPO_AB	binary	blood type AB
BMI	num	body mass index
POLSO	num	heart rate
PMAX	num	maximum pressure
EMOG	num	hemogoblim
PMIN	num	minimum pressure
SESSO_EMOG	num	interaction gender-hemoglobin
SESSO_RH	binary	interaction gender-RH
SESSO_BMI	num	interaction gender-BMI

Table 4.1: Complete set of covariates

4.1.3 Baseline intensity function

The baseline intensity function $\lambda_0(t)$ is piecewise constant:

$$\lambda_0(t) = \lambda_k \qquad \text{if } a_{k-1} < t \leq a_k \qquad k = 1, \ldots, K \tag{4.4}$$

where $K = 10$ equi-spaced intervals have been chosen.

4.1.4 Random effects

In order to analyze data that occur repeatedly, it is necessary to account for subject-dependency in the multiple event times and this can be done by considering random effects u_i, also called frailties. They have a *multiplicative effect* on each donor's intensity function, in particular:

- $u_i > 1$ indicates more propensity to experience a donation;
- $u_i < 1$ indicates less propensity to experience a donation.

4.1.5 Priors

A priori the parameters λ_k's, β_p's and u_i's in *Model 0* are all independent with:

- $\lambda_k \overset{iid}{\sim} \text{Gamma}(\alpha_\lambda, \alpha_\lambda)$ $\qquad k = 1, \dots, K$
- $\beta_p \overset{iid}{\sim} \mathcal{N}(0, \sigma^2_\beta)$ $\qquad p = 1, \dots, P$
- $u_i|\beta_0 \overset{iid}{\sim} \text{Gamma}\left(\alpha_u, \dfrac{\alpha_u}{\beta_0}\right), \qquad \beta_0 \sim \text{Gamma}(\delta, \delta)$ $\qquad i = 1, \dots, I$

where $\alpha_\lambda = \delta = 2$, $\alpha_u = 0.01$ and $\sigma^2_\beta = 100$.

4.1.6 Prior moments as a function of the hyperparameters

First of all, we can notice that $\text{E}[u_i|\beta_0] = \beta_0$ and $\text{Var}[u_i|\beta_0] = \beta_0^2/\alpha_u$, so that the frailties are centered a priori in β_0 and α_u regulates the prior variance. For a better interpretation of the hyperparameters δ, α_u, marginal mean, variance and covariances of u_i's are computed:

$$\text{E}[u_i] = \text{E}[\text{E}[u_i|\beta_0]] = \text{E}[\beta_0] = 1$$

$$\begin{aligned}\text{Var}[u_i] &= \text{Var}[\text{E}[u_i|\beta_0]] + \text{E}[\text{Var}[u_i|\beta_0]] = \text{Var}[\beta_0] + \text{E}\left[\frac{\beta_0^2}{\alpha_u}\right] \\ &= \frac{1}{\delta} + \frac{1}{\alpha_u}\left(\frac{1}{\delta} + 1\right) = \frac{1}{\alpha_u} + \frac{1}{\delta} + \frac{1}{\alpha_u \delta}\end{aligned}$$

In particular, substituting $\alpha_u = 0.01$ and $\delta = 2$, we get: $\text{Var}[u_i] = 150.5$, by leading to a prior which seems explorative enough. Moreover, under the prior choice of Section 4.1.5, the frailties

of two donors i and j have covariance:

$$\begin{aligned}\mathrm{Cov}(u_i, u_j) &= \mathrm{E}[u_i u_j] - \mathrm{E}[u_i]\,\mathrm{E}[u_j] = \mathrm{E}[\mathrm{E}[u_i u_j|\beta_0]] - 1 \\ &= \mathrm{E}[\mathrm{E}[u_i|\beta_0]\,\mathrm{E}[u_j|\beta_0]] - 1 = E[\beta_0^2] - 1 \\ &= \left(\frac{\delta}{\delta^2} + 1\right) - 1 = \frac{1}{\delta}\end{aligned}$$

and correlation

$$\rho(u_i, u_j) = \frac{\alpha_u}{\alpha_u + \delta + 1}$$

so that α_u governs also the pairwise correlation between frailties.

4.1.7 At-risk indicator

The at-risk indicator function models the "risk" of experiencing an event. The general definition is the following:

$$Y_i(t) = \begin{cases} 1 & \text{individual } i \text{ can donate at time } t \\ 0 & \text{otherwise} \end{cases}$$

whereas in our specific problem, it assumes the following form:

$$\mathbf{1}_{\left\{(t - T_{N_i(t^-)}) \geq \Phi_i,\, t \leq c_i\right\}}(t) \tag{4.5}$$

We notice that it depends on the history of the process and it repeats itself equal after each event. In particular, it models the fact that a person cannot donate after his/her censoring time c_i and that for a certain period of time Φ_i he/she must wait after last donation, depending on the gender. The intensity is set equal to 0 for the next Φ_i days after every donation. In particular, according to AVIS rules, Φ_i is equal to 90 days for men and 180 days for women. However AVIS physicians might anticipate donations. Accordingly to AVIS rules, to represent the phenomenon in a more realistic way, we set:

$$\Phi_i = \begin{cases} 150 & \text{if } i \text{ is female donor} \\ 85 & \text{if } i \text{ is male donor} \end{cases}$$

The thresholds are fixed heuristically and the goal is to allow reasonable early donations, discarding as least as possible individuals from the study.

4.2 Model 1

Model 1 extends *Model 0* by considering time-autoregressive frailties. The reference paper for current section is Song and Kuo (2013). As a preliminary step, frailties are discretized becoming piecewise linear functions. Frailties and baseline intensity function have the same time step.

4.2.1 Likelihood

We take up the same general framework of Section 4.1.1, by replacing piecewise constant frailties for each donor $i = 1, \dots, I$:

$$u_i(t) = u_{ik} \qquad \text{if } a_{k-1} < t \leq a_k$$

or equivalently:

$$u_i(t) = \sum_{k=1}^{K} u_{ik} w_k(t) \qquad [\text{with } w_k(t) = \mathrm{I}_{(a_{k-1}, a_k]}(t)]$$

where K is both the number of time intervals for frailties and the number of time step for the baseline intensity function. Through mathematical steps similar to those previously performed to obtain the likelihood of *Model 0* (Formula (4.1)), the likelihood of *Model 1* is:

$$\begin{aligned}\mathscr{L} = &\left(\prod_{k=1}^{K} \lambda_k^{n_{\cdot k}}\right)\left(\prod_{k=1}^{K}\prod_{j=1}^{n_i} u_{ik}^{n_{ik}}\right) \times \\ &\times \exp\left\{\sum_{i=1}^{M}\sum_{j=1}^{n_i} \mathbf{x}_i'(t_{ij})\boldsymbol{\beta} - \sum_{i=1}^{M}\sum_{k=1}^{K} \lambda_k u_{ik} \int_{a_{k-1}}^{a_k} Y_i(s) e^{\mathbf{x}_i'(s)\boldsymbol{\beta}} ds\right\}\end{aligned}$$

where n_{ik} is the number of events experienced by individual i in the interval $(a_{k-1}, a_k]$.

4.2.2 Priors

In *Model 0*, a vector of frailties is considered and each cell represents the constant donor-specific value. Here instead we have a matrix $[M \times K]$, where each row represents the evolution of the donor's frailty along time. The evolution of frailties over segment k for donor i is defined by:

$$u_{ik} = u_{i(k-1)}\phi_{ik} \quad \Longrightarrow \quad u_{ik} = \prod_{g=1}^{k} \phi_{ig} \quad i = 1, \dots, I$$

where $\{\phi_{i1}, \phi_{i2}, \dots, \phi_{iK}\}$ constitute the *multiplicative frailty innovation* for donor i. Except for the frailties' modification, priors are kept equal to the previous *Model 0* and the whole picture is reported below:

- $\phi_{ik} \overset{iid}{\sim} \text{Gamma}(\psi, \psi) \qquad \psi \sim \text{Gamma}(a_\psi, a_\psi) \qquad i = 1, \dots, I,\ k = 1, \dots, K$

- $\lambda_k \overset{iid}{\sim} Gamma(a_\lambda, b_\lambda)$ $\qquad k = 1, \ldots, K$
- $\beta_p \overset{iid}{\sim} \mathcal{N}(0, \sigma^2_\beta)$ $\qquad p = 1, \ldots, P$

with $\alpha_\psi = \alpha_\lambda = 2$ and $\sigma^2_\beta = 100$.

4.2.3 Prior moments as a function of the hyperparameters

Figure 4.1 shows how $u(t)$ is distributed along time. We suppose that donor i's process is independent of donor j, $\forall i \neq j$. For donor i, we get:

$$\begin{aligned} u_{i1} &= \phi_{i1} \\ u_{i2} &= u_{i1}\phi_{i2} \\ &\vdots \\ u_{iK} &= u_{k-1}\phi_{iK} \end{aligned}$$

where $\phi_{i1}, \phi_{i2}, \ldots, \phi_{iK}$ are independent conditionally on ψ and s.t. $\mathrm{E}[\phi_{ik}] = 1$, $\forall k = 1, \ldots, K$.

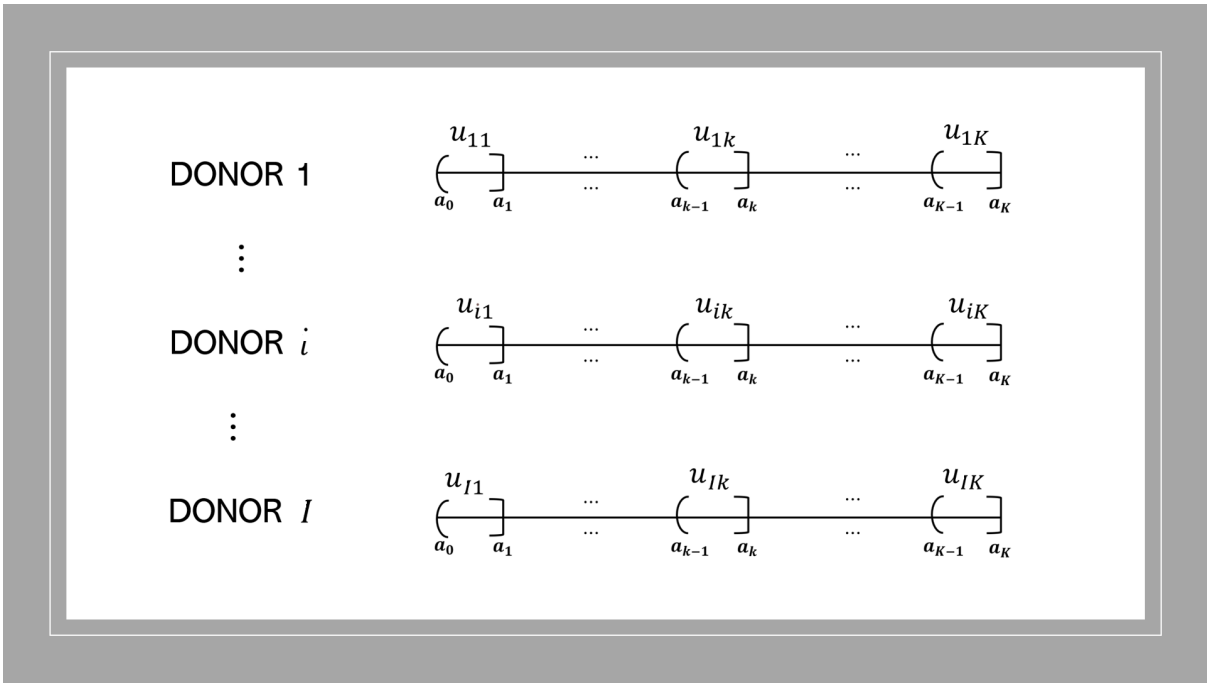

Figure 4.1: Representation of $u(t)$ during time

In the computations that follow we suppose ψ fixed and time interval $h < j$. We have:

$$\begin{aligned} \mathrm{E}[u_{ik}|u_{i1}, \ldots, u_{i,k-1}] &= \mathrm{E}[u_{ik}|u_{i,k-1}] = \mathrm{E}[u_{i,k-1}\phi_{ik}|u_{i,k-1}] \\ &= u_{i,k-1}\,\mathrm{E}[\phi_{ik}|u_{i,k-1}] = u_{i,k-1}\,\mathrm{E}[\phi_{ik}] = u_{i,k-1} \end{aligned}$$

$$\begin{aligned}
\mathrm{Var}[u_{ik}|u_{i1},\ldots,u_{i,k-1}] &= \mathrm{Var}[u_{ik}|u_{i,k-1}] = \mathrm{E}[u_{ik}^2|u_{i,k-1}] - (\mathrm{E}[u_{ik}|u_{i,k-1}])^2 \\
&= \mathrm{E}[u_{i,k-1}^2\phi_{ik}^2|u_{i,k-1}] - u_{i,k-1}^2 = u_{i,k-1}^2\,\mathrm{E}[\phi_{ik}^2] - u_{i,k-1}^2 \\
&= u_{i,k-1}^2(\mathrm{Var}[\phi_{ik}] + (\mathrm{E}[\phi_{ik}])^2) - u_{i,k-1}^2 \\
&= u_{i,k-1}^2\left(\frac{1}{\psi}+1\right) - u_{i,k-1}^2 = \frac{u_{i,k-1}^2}{\psi}
\end{aligned}$$

$$\begin{aligned}
\mathrm{Cov}(u_{ih},u_{ij}) &= \mathrm{Cov}(u_{ih}, u_{ih}\phi_{i,h+1}\ldots\phi_{i,j}) \\
&= \mathrm{E}[u_{ih}^2\phi_{i,h+1}\ldots\phi_{ij}] - \mathrm{E}[u_{ih}]\,\mathrm{E}[u_{ih}\phi_{i,h+1}\ldots\phi_{ij}] \\
&= \mathrm{E}[u_{ih}^2]\,\mathrm{E}[\phi_{i,h+1}]\ldots\mathrm{E}[\phi_{ij}] - \mathrm{E}[u_{ih}]\,\mathrm{E}[u_{ih}]\,\mathrm{E}[\phi_{i,h+1}]\ldots\mathrm{E}[\phi_{ij}] \\
&= \mathrm{E}[u_{ih}^2] - (\mathrm{E}[u_{ih}])^2 = \mathrm{Var}[u_{ih}]
\end{aligned}$$

Now, it is important to notice that:

- $\mathrm{E}[u_{i1}] = \mathrm{E}[\phi_{i1}] = 1$, $\mathrm{E}[u_{i2}] = 1$, …, $\mathrm{E}[u_{iK}] = 1$
- $\mathrm{E}[u_{i,h-1}^2] = \mathrm{E}[\phi_{i1}^2\phi_{i2}^2\ldots\phi_{i,h-1}^2] = \mathrm{E}[\phi_{i1}^2]\,\mathrm{E}[\phi_{i2}^2]\ldots\mathrm{E}[\phi_{i,h-1}^2] = \left(\frac{1}{\psi}+1\right)^{h-1}$

so that we can compute:

$$\begin{aligned}
\mathrm{Var}[u_{ih}] &= \mathrm{E}[\mathrm{Var}[u_{ih}|u_{i,h-1}]] + \mathrm{Var}[\mathrm{E}[u_{ih}|u_{i,h-1}]] \\
&= \mathrm{E}\left[\frac{u_{i,h-1}^2}{\psi}\right] + \mathrm{Var}[u_{i,h-1}] = \frac{1}{\psi}\left(\frac{1}{\psi}+1\right)^{h-1} + \left(\frac{1}{\psi}+1\right)^{h-1} - 1 \\
&= \left(\frac{1}{\psi}+1\right)^{h-1}\left(\frac{1}{\psi}+1\right) - 1 = \left(\frac{1}{\psi}+1\right)^{h} - 1 = \frac{(1+\psi)^h - \psi^h}{\psi^h}
\end{aligned}$$

In the end, we are able to obtain:

$$\rho(u_{ih},u_{ij}) = \frac{\mathrm{Var}[u_{ih}]}{\sqrt{\mathrm{Var}[u_{ih}]}\sqrt{\mathrm{Var}[u_{ij}]}} = \sqrt{\frac{\mathrm{Var}[u_{ih})}{\mathrm{Var}[u_{ij}]}} = \sqrt{\frac{\left(\frac{1}{\psi}+1\right)^h - 1}{\left(\frac{1}{\psi}+1\right)^j - 1}} = \sqrt{\frac{(1+\psi)^h - \psi^h}{(1+\psi)^j - \psi^j}\cdot\psi^{j-h}}$$

First, we noticed that the correlation is always strictly greater than zero and varies with ψ. In particular, we start considering the two extreme cases:

- CASE 1: $\psi \to \infty$

In this case, as $(1+f(\psi))^h - 1 \overset{f(\psi)\to 0}{\approx} hf(\psi)$, we obtain:

$$\lim_{\psi\to\infty}\rho(u_{ih},u_{ij}) = \lim_{\psi\to\infty}\sqrt{\frac{\left(\frac{1}{\psi}+1\right)^h - 1}{\left(\frac{1}{\psi}+1\right)^j - 1}} \longrightarrow \sqrt{\frac{h}{j}}$$

Hence, for large value of ϕ, the correlation decreases with increasing distance between h and j, moreover it is always less than $\sqrt{0.9}$, as $K = 10$.

- CASE 2: $\psi \to 0$

  Here it is enough to remember that $(1+\psi)^h \overset{\psi\to 0}{\approx} 1 + h\psi$ to obtain:

$$\lim_{\psi\to 0}\rho(u_{ih},u_{ij}) = \lim_{\psi\to 0}\sqrt{\frac{(1+\psi)^h - \psi^h}{(1+\psi)^j - \psi^j}\cdot\psi^{j-h}} = \lim_{\psi\to 0}\underbrace{\sqrt{\frac{1+h\psi-\psi^h}{1+j\psi-\psi^j}}}_{>0} \times \underbrace{\sqrt{\psi^{j-h}}}_{\to 0} \longrightarrow 0$$

As a priori $\mathrm{E}[\psi] = 1$, but $\mathrm{Var}[\psi] = 1/\alpha_\psi$, it follows that a small value of α_ψ allows ψ to range from small to large values and so a priori we have a correlation between 0 and 0.9, for $K = 10$.

4.3 Model 2

The formulation of *Model 2* allows us to overcome the identifiability problem that may arise in *Model 0* for the parameters λ_k, $k = 1,\dots,K$, in the baseline intensity function and the frailties u_i, $i = 1,\dots,I$. In particular, we introduce an unique step function $\widetilde{u}_i(t)$ for each donor i defined as:

$$\widetilde{u}_i(t) = \sum_{i=1}^{K}\widetilde{u}_{ik}\mathrm{I}_{(a_{k-1},a_k]}(t) \tag{4.6}$$

The function $\widetilde{u}_i(t)$ simultaneously model both the mean random heterogeneity among donors, i.e. their individual frailties and the baseline intensity function, that is common to all of them. We choosed the following prior structure for $\widetilde{u}_i(t)$, that takes account for the length of the time-interval $(a_{k-1}, a_k]$ (Christensen et al. (2011)):

$$\widetilde{u}_{ik}|\alpha_k, c \overset{iid}{\sim} \mathrm{Gamma}(l\alpha_k, l) \qquad \alpha_k \sim \mathrm{Gamma}(\delta,\delta) \tag{4.7}$$

where $\delta = 2$ and $l \propto (a_k - a_{k-1}) = 310$ for each k. Then $l = 310c$, where c quantifies the uncertainty on $\widetilde{u}_{ik}$, i.e. c is a measure of how widespread the prior of $\widetilde{u}_{ik}$ is. Indeed:

$$\mathrm{E}[\widetilde{u}_i(t)|\alpha_1, \ldots, \alpha_K, c] = \sum_{k=1}^{K} \mathrm{E}[\widetilde{u_{ik}}|\alpha_k, c]\mathrm{I}_{(a_{k-1},a_k]}(t) = \sum_{k=1}^{K} \alpha_k \mathrm{I}_{(a_{k-1},a_k]}(t) \tag{4.8}$$

and

$$\mathrm{Var}[\widetilde{u}_i(t)|\alpha_1, \ldots, \alpha_K, c] = \frac{1}{c} \sum_{k=1}^{K} \alpha_k \mathrm{I}_{(a_{k-1},a_k]}(t) \tag{4.9}$$

We choosed $c = 0.01$ and the robustness analysis for this model is to be made only with respect to c. It follows from Formula (4.6) that the mean step function

$$\alpha(t) = \sum_{k=1}^{K} \alpha_k \mathrm{I}_{(a_{k-1},a_k]}(t) \tag{4.10}$$

has an immediate interpretation in terms of baseline intensity function. Hence the hyperparameters $\{\alpha_1, \alpha_2, \ldots, \alpha_k\}$ provide - in a Bayesian way - a piecewise constant baseline intensity function. In that it lies the substantial difference between *Model 0* (where the baseline intensity function is included in the likelihood) and *Model 2* (where the baseline intensity function is a prior hyperparameter). As a consequence, the posterior mean of $\alpha(t)$ in *Model 2*, given by

$$\hat{\alpha}(t) := \sum_{k=1}^{K} \mathrm{E}[\alpha_k|Data]\mathrm{I}_{(a_{k-1},a_k]}(t) \tag{4.11}$$

is a Bayesian estimate of the baseline intensity function. In this parameterization, in accordance with the scale properties of the Gamma model, the parameter

$$v_{ik} := \frac{\widetilde{u_{ik}}}{\alpha_k} \tag{4.12}$$

represents the specific random effect of *Model 2*. Basically *Model 2* has the same likelihood of *Model 0*, but with a different parametrization of λ_k and u_i:

$$\mathscr{L} = \left(\prod_{k=1}^{K} \prod_{j=1}^{n_i} (\widetilde{u}_{ik})^{n_{ik}} \right) \exp\left\{ \sum_{i=1}^{M} \sum_{j=1}^{n_i} \mathbf{x}_i'(t_{ij})\boldsymbol{\beta} - \sum_{i=1}^{M} \sum_{k=1}^{K} \widetilde{u}_{ik} \int_{a_{k-1}}^{a_k} Y_i(s) e^{\mathbf{x}_i'(s)\boldsymbol{\beta}} ds \right\} \tag{4.13}$$

Chapter 5

Posterior analysis

In this chapter the posterior inference for the models described in Chapter 4 is presented. They are applied to the AVIS data presented in Chapter 2. The focus is on "Model 2" that, as a result of the simulations carried out, turns out to be the most performing model in terms of both indices WAIC and LPML.

5.1 Stan Software

Sampling from the posterior distribution is achieved via the software platform called Stan, which is a probabilistic programming language for statistical inference written in `C++`. Stan is a software for MCMC sampling more efficient than the ones written in the BUGS (Bayesian inference using Gibbs Sampling) language, like JAGS. For simple models there is little practical difference between the two platforms in the efficiency of the chains, but Stan outperforms BUGS as model size and complexity grow. In particular, Stan uses Hamiltonian Monte Carlo (HMC), a family of MCMC algorithms which promise improved efficiency and faster inference (Stan Development Team (2020)).

5.2 Sampling

Different sampling processes are considered, all with two chains having 5000 iterations, the first 3000 of warmup and the last 2000 of sampling. Hence all the results are MCMC samples of 2000 observations for each chain. As an example, the output for *Model 0*'s simulation is reported:

```
Inference for Stan model: time-dependent-MODEL0.
2 chains, each with iter=5000; warmup=3000; thin=1;
```

```
post-warmup draws per chain=2000, total post-warmup draws=4000
```

5.3 Model comparison

The convergence diagnostics of the different simulations have been checked, showing that all the MCMC chains reach stationarity, but in *Model 0* a problem of identifiability arises, due to the product between the baseline intensity function $\lambda_0(t)$ and the frailties u_i for $i = 1, \dots, I$ in the intensity function in the likelihood in Formula (4.2). Therefore the choice of the best model reduces between *Model 1* and *Model 2*. Table 5.1 shows that *Model 2* outperforms *Model 1* both in terms of WAIC and LPML. As explained in Section 4.3, *Model 2* overcomes the problem of identifiability, because the baseline intensity function is a prior hyperparameter and it doesn't directly appear in the likelihood.

	WAIC	**LPML**
Model 1	396574.9	-202183.4
Model 2	298826.8	-150630.6

Table 5.1: Goodeness of fit evaluation

5.4 Posterior inference for *Model 2*

Model 2 is summarised below.

- The likelihood is:

$$\mathscr{L} = \left(\prod_{k=1}^{K} \prod_{j=1}^{n_i} (\widetilde{u}_{ik})^{n_{ik}} \right) \exp \left\{ \sum_{i=1}^{M} \sum_{j=1}^{n_i} \mathbf{x}_i'(t_{ij}) \boldsymbol{\beta} - \sum_{i=1}^{M} \sum_{k=1}^{K} \widetilde{u}_{ik} \int_{a_{k-1}}^{a_k} Y_i(s) e^{\mathbf{x}_i'(s)\boldsymbol{\beta}} ds \right\}$$

  where

  - n_i is the total number of donations of individual i;
  - n_{ik} is the number of events experienced by individual i in the interval $(a_{k-1}, a_k]$;
  - $n_{\cdot k} = \sum_{i=1}^{M} \sum_{j=1}^{n_i} w_k(t_{ij})$ is the total number of donations in the interval $(a_{k-1}, a_k]$.

- The prior choice is:

  - $\beta_p \overset{iid}{\sim} \mathcal{N}(0, \sigma_\beta^2)$ $\qquad p = 1, \dots, P$
  - $\widetilde{u}_{ik} | \alpha_k, c \overset{iid}{\sim} \text{Gamma}(l\alpha_k, l)$ $\qquad i = 1, \dots, M$
  - $\alpha_k \sim \text{Gamma}(\delta, \delta)$ $\qquad k = 1, \dots, K$

where $\sigma_\beta^2 = 100$, $\delta = 2$, $l = 310c$ and $c = 0.01$ and β_p's are independent on $(\widetilde{u}_{ik}, \alpha_k)$'s.

5.4.1 Beta regression coefficients

By analyzing β_p's posterior densities, it is clear that the model can be simplified. As reported in Table 5.2, the parameters in grey color are not significant, because their 95% credibility intervals $(q_{0.025}, q_{0.095})$ contain 0.

Parameter	**Mean**	**Standard deviation**	$\mathbf{q_{0.025}}$	$\mathbf{q_{0.5}}$	$\mathbf{q_{0.095}}$
β_{SESSO}	-4.37	0.03	-4.44	-4.37	-4.31
β_{FUMO}	-0.23	0.02	-0.26	-0.23	-0.19
β_{ALCOOL}	-0.10	0.02	-0.13	-0.10	-0.07
$\beta_{\text{ATTIVITAFISICA}}$	-0.17	0.02	-0.20	-0.17	-0.14
β_{RH}	-2.98	0.03	-3.04	-2.98	-2.91
β_{PMIN}	0.03	0.01	0.02	0.03	0.05
$\beta_{\text{TIPO\_0}}$	-0.26	0.02	-0.29	-0.26	-0.22
$\beta_{\text{TIPO\_B}}$	-0.25	0.03	-0.30	-0.25	-0.20
$\beta_{\text{TIPO\_AB}}$	-0.44	0.06	-0.56	-0.45	-0.33
β_{EMOG}	1.12	0.02	1.07	1.11	1.16
$\beta_{\text{ETA\_PRIMA}}$	0.23	0.01	0.21	0.23	0.25
β_{BMI}	0.17	0.02	0.14	0.17	0.21
β_{PMAX}	0.00	0.01	-0.01	0.00	0.02
β_{POLSO}	-0.02	0.01	-0.03	-0.02	0.00
$\beta_{\text{SESSO\_EMOG}}$	-1.16	0.02	-1.21	-1.16	-1.11
$\beta_{\text{SESSO\_BMI}}$	-0.17	0.02	-0.21	-0.17	-0.14
$\beta_{\text{SESSO\_RH}}$	2.92	0.04	2.85	2.92	3.00

Table 5.2: Complete *Model 2*, summary of β_p's posterior densities (non significant β_p's in grey colour)

Figure 5.1 shows β_p's posterior densities. In particular, the covariates:

- in blue have a negative effect, i.e. reducing the risk to donate blood;
- in red have a positive effect, i.e. increasing the risk to donate blood;
- in white are not significant.

It is evident that the interactions $\beta_{\text{SESSO-RH}}$, $\beta_{\text{SESSO-EMOG}}$ and $\beta_{\text{SESSO-BMI}}$ are all significant, hence we have definitely found a way to distinguish the behavior of male and female donors.

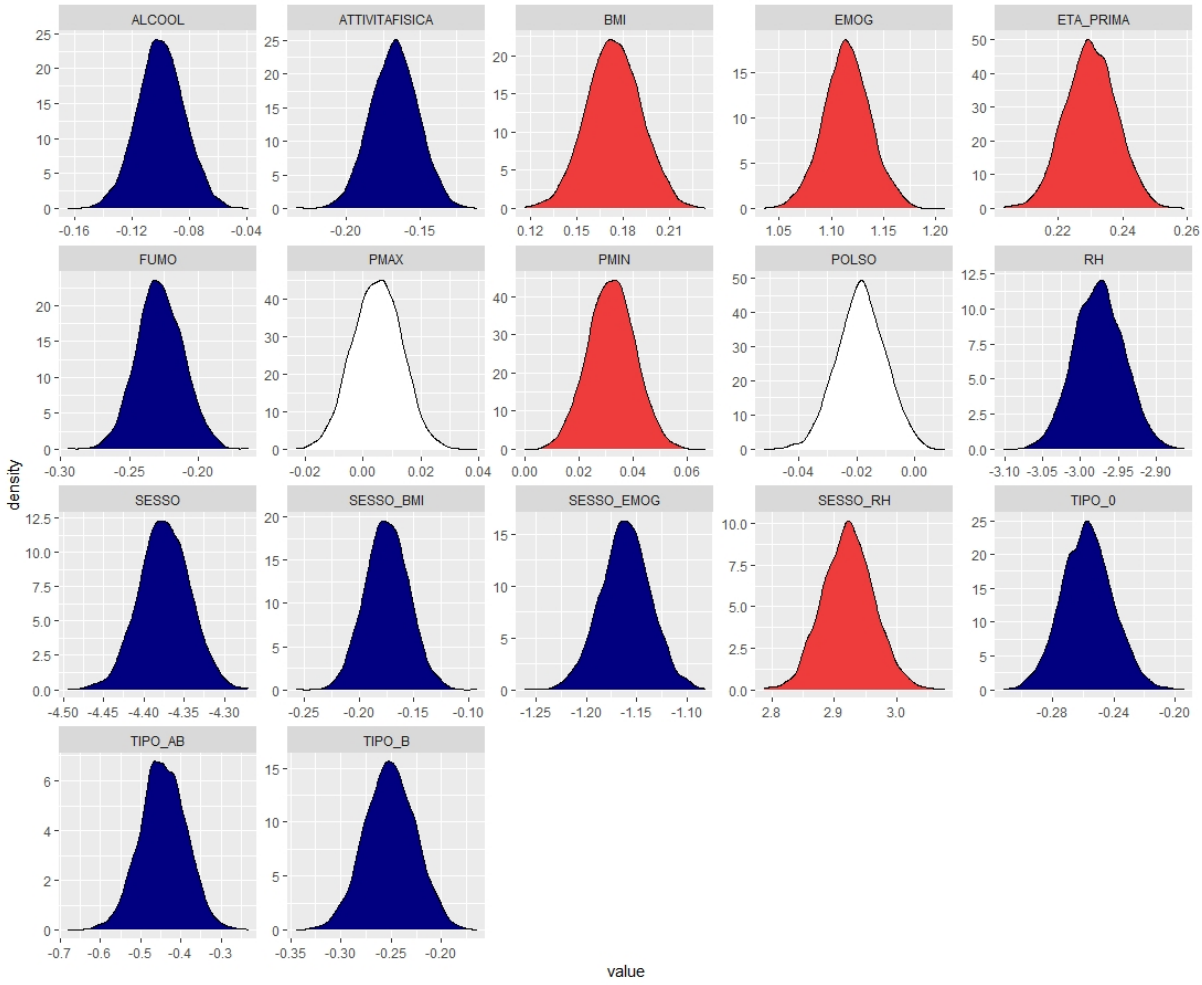

Figure 5.1: Complete *Model 2*, β_p's posterior densities. The covariates in blue have a negative effect, the covariates in red have a positive effect and the covariates in white are not significant

On the other hand, β_{PMAX} and β_{POLSO} are not significant, hence these covariates have been removed from the model and then the new model has been re-fitted. The results obtained for the reduced model are reported in Table 5.3 and the analysis of Figure 5.2 leads to the following comments:

- In Figure 2.4 (Chapter 2) we noticed that women outnumber men up to 8 total recurrences and that the strong peak in favor of men in the range (13, 29) is justified simply by the fact that men can donate twice as much as women. Now intuition that overall women have more consistent behavior in the donation process is confirmed by the negative value of β_{SESSO} (recalling that the gender is codified as 0 for women and as 1 for men);
- $\beta_{\text{TIPO\_0}}$, $\beta_{\text{TIPO\_B}}$ and $\beta_{\text{TIPO\_AB}}$ are significant with respect to the base case $\beta_{\text{TIPO\_A}}$;

Parameter	**Mean**	**Standard deviation**	$\mathbf{q_{0.025}}$	$\mathbf{q_{0.5}}$	$\mathbf{q_{0.095}}$
β_{SESSO}	-4.37	0.03	-4.43	-4.37	-4.31
β_{FUMO}	-0.23	0.02	-0.26	-0.23	-0.20
β_{ALCOOL}	-0.10	0.02	-0.13	-0.10	-0.07
$\beta_{\text{ATTIVITAFISICA}}$	-0.17	0.02	-0.20	-0.17	-0.13
β_{RH}	-2.98	0.03	-3.04	-2.98	-2.91
β_{PMIN}	0.03	0.01	0.02	0.03	0.05
$\beta_{\text{TIPO\_0}}$	-0.26	0.02	-0.29	-0.26	-0.22
$\beta_{\text{TIPO\_B}}$	-0.25	0.02	-0.30	-0.25	-0.20
$\beta_{\text{TIPO\_AB}}$	-0.44	0.06	-0.56	-0.44	-0.33
β_{EMOG}	1.11	0.02	1.07	1.11	1.16
$\beta_{\text{ETA\_PRIMA}}$	0.23	0.01	0.22	0.23	0.25
β_{BMI}	0.17	0.02	0.14	0.17	0.21
$\beta_{\text{SESSO\_EMOG}}$	-1.16	0.02	-1.21	-1.16	-1.11
$\beta_{\text{SESSO\_BMI}}$	-0.17	0.02	-0.21	-0.17	-0.14
$\beta_{\text{SESSO\_RH}}$	2.93	0.04	2.85	2.93	3.01

Table 5.3: Reduced *Model 2*, summary of β_p's posterior densities

- β_{FUMO} and β_{ALCOOL} have a negative effect, hence smokers and drinkers tend to donate less than non-smokers and non-drinkers; this is consistent with our intuition. On the other hand, $\beta_{\text{ATTIVITAFISICA}}$ has a negative effect, hence it seems that volunteers with a non-active life are slightly more likely to donate than people with an active life;

- $\beta_{\text{ETA\_PRIMA}}$ is positive, meaning that the older people are, the more likely they donate (it can be noticed that ETA_PRIMA is below 25 age for 25.4% of the donors);

- β_{RH} is negative. This was expected, because RH^- is much rarer than RH^+. It follows that volunteers with RH^- behave in a more responsible way than the others, by regularly donating.

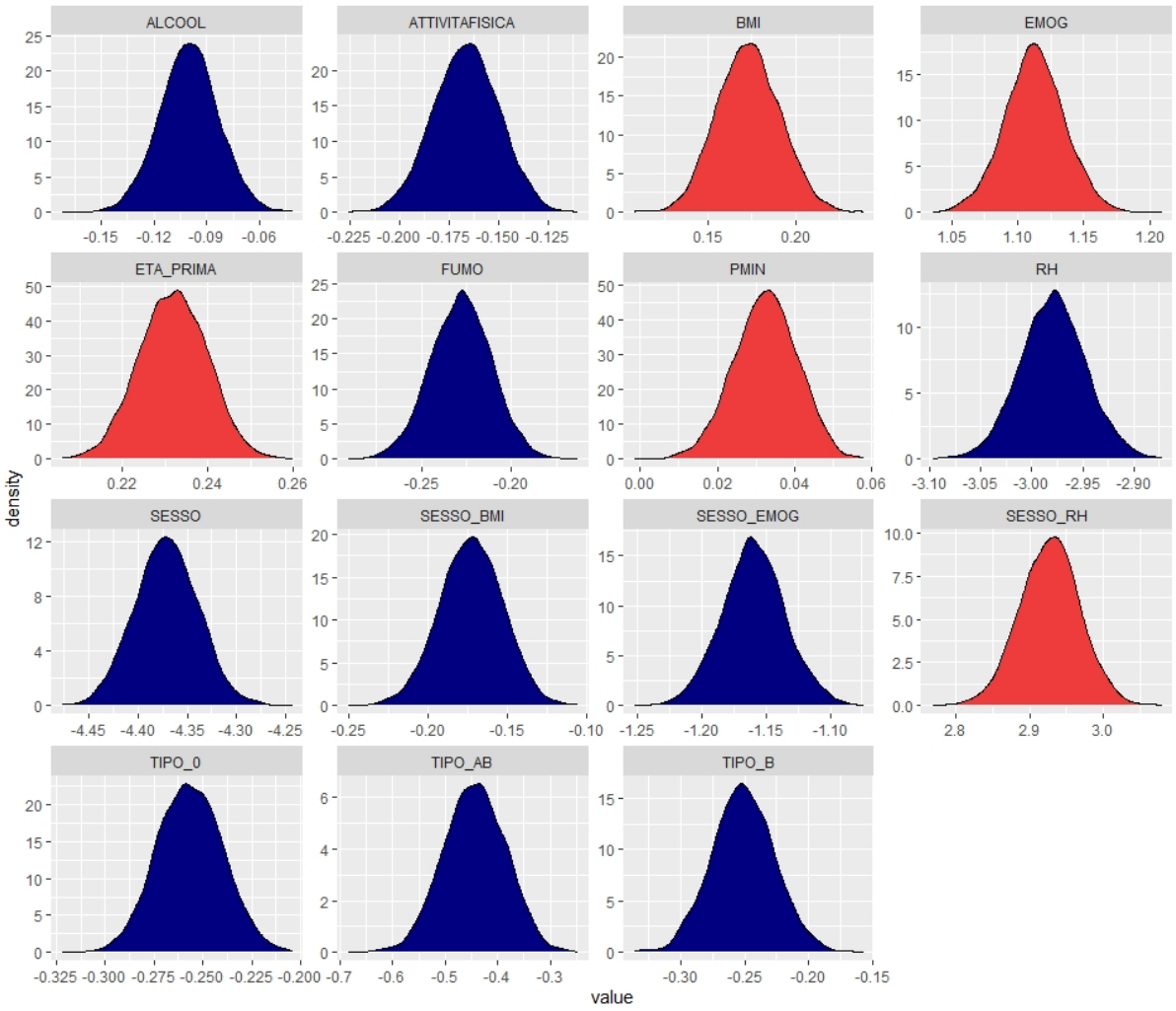

Figure 5.2: Reduced *Model 2*, β_p's posterior densities. The covariates in blue have a negative effect and the covariates in red have a positive effect

The MCMC convergence diagnostics such as those available in the R package `CODA` were computed for all parameters, indicating that convergence may have been achieved. All details concerning them are reported in Appendix E.

5.4.2 Alpha baseline intensity function

From now on, all analyses refer to reduced *Model 2*. The posterior summary for α_k, $k = 1, \ldots, K$ is reported in Table 5.4 and the posterior densities are shown in Figure 5.3. These parameters are centered around a value that decreases over time, except for the last time-intervals. The highest value corresponds to α_1, that refers to the first 310 days of the donation process; this means that volunteers are more likely to donate at the beginning and this propensity tends to decrease over time. As for β_p's parameters, convergence diagnostics are reported in Appendix E.

Parameter	Mean	Standard deviation	$q_{0.025}$	$q_{0.5}$	$q_{0.095}$
α_1	5.50	0.12	5.26	5.50	5.74
α_2	0.87	0.02	0.84	0.87	0.90
α_3	0.60	0.01	0.57	0.60	0.63
α_4	0.50	0.01	0.47	0.50	0.52
α_5	0.41	0.01	0.39	0.41	0.43
α_6	0.42	0.01	0.40	0.42	0.45
α_7	0.39	0.01	0.36	0.39	0.42
α_8	0.47	0.02	0.44	0.47	0.50
α_9	0.56	0.02	0.51	0.55	0.60
α_{10}	1.01	0.05	0.92	1.01	1.13

Table 5.4: Reduced *Model 2*, α_k's posterior summary

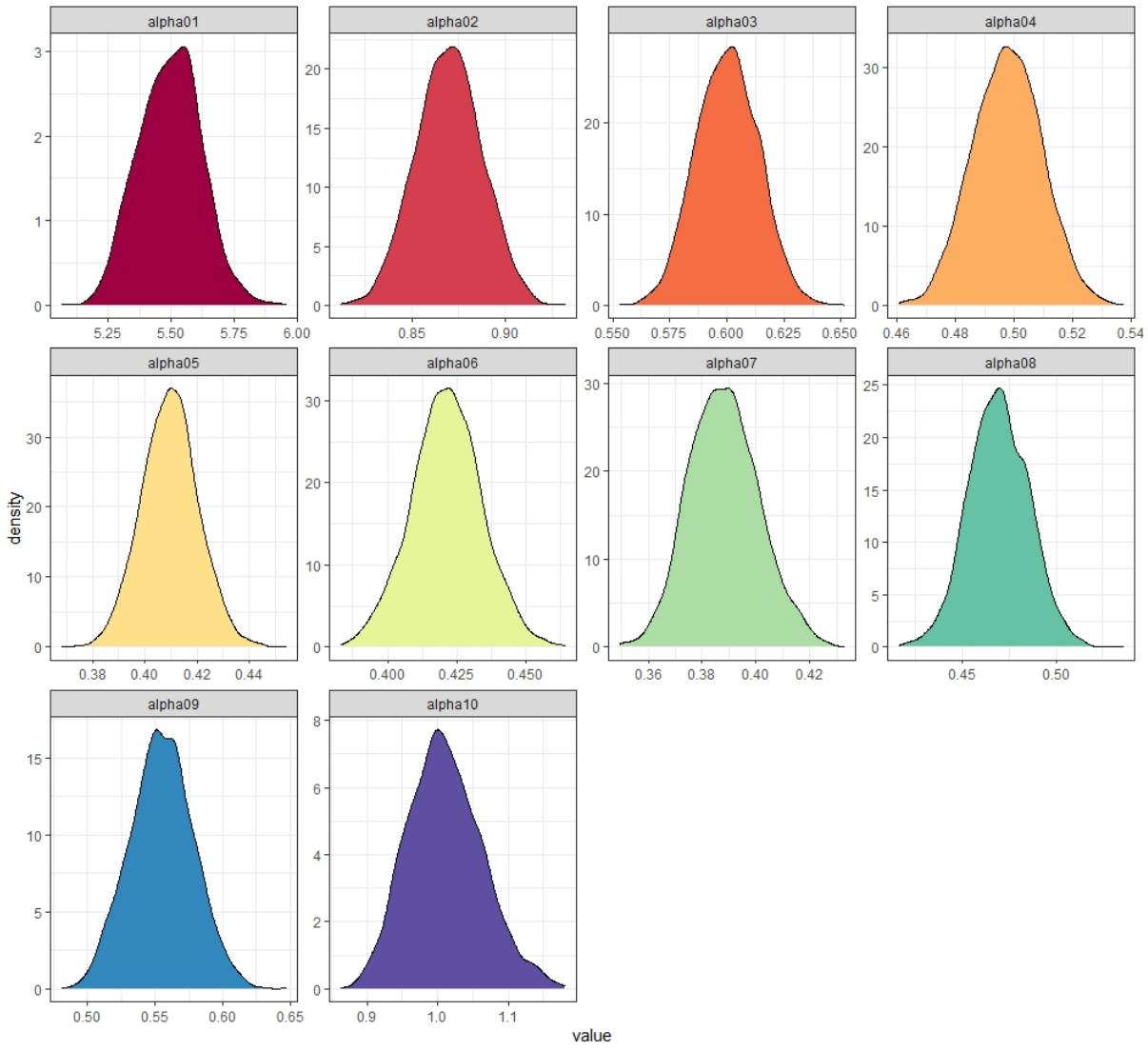

Figure 5.3: Reduced *Model 2*, α_k's posterior densities

5.4.3 Frailties

The number of frailties is very large: 59370 different parameters are simulated (ten for each donor, of which one for each time-interval); all the convergence criteria are met and there is nothing relevant to report concerning convergence diagnostics. Figure 5.4 and Figure 5.5 show frailties' trend over time for 100 randomly selected male donors (in blue) and 100 randomly selected female donors (in red). The fact that each $\{v_{ik}\}_{k=1,\dots,K}$ shows a trend varying with $i = 1, \dots, I$, i.e. from one donor to another, confirms the presence of a specific component in donors' behaviour, that the other parameters of the model are not able to capture.

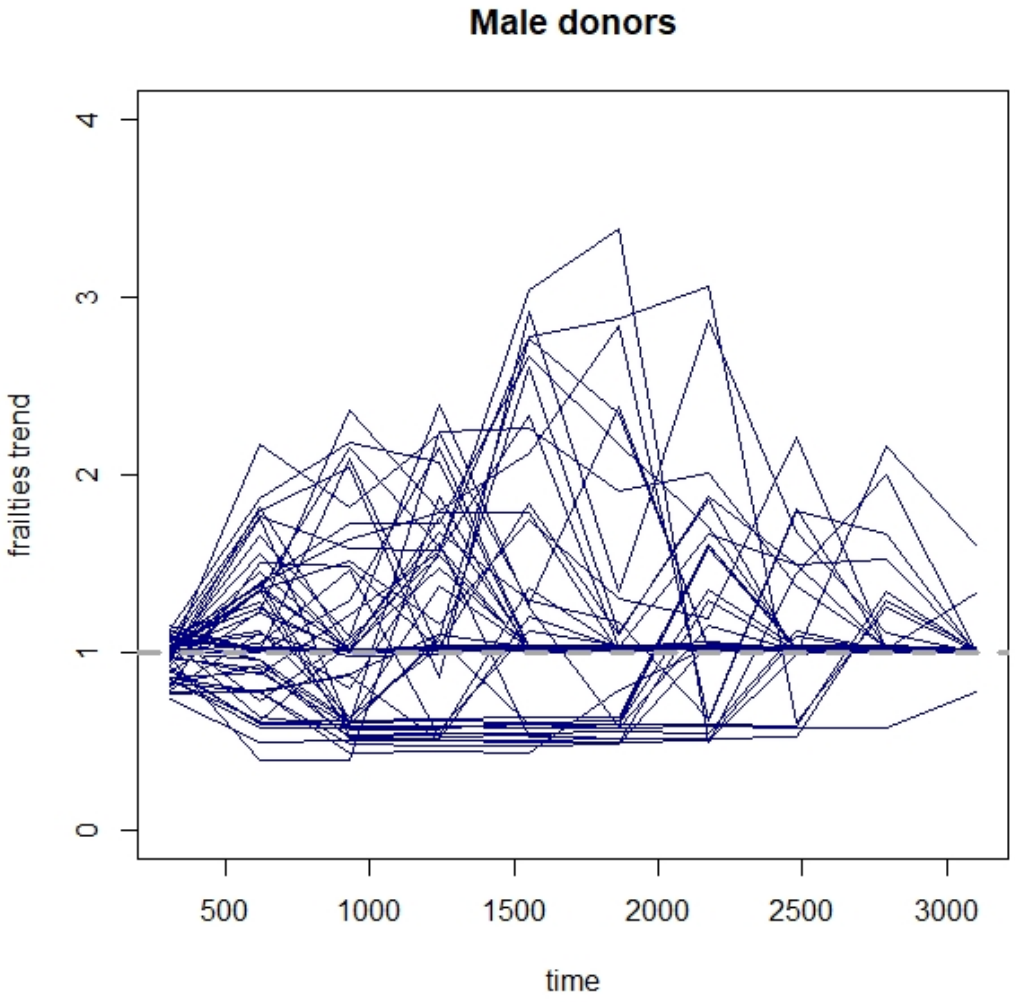

Figure 5.4: Reduced *Model 2*, v_{ik}'s trend over time for 100 randomly selected male donors

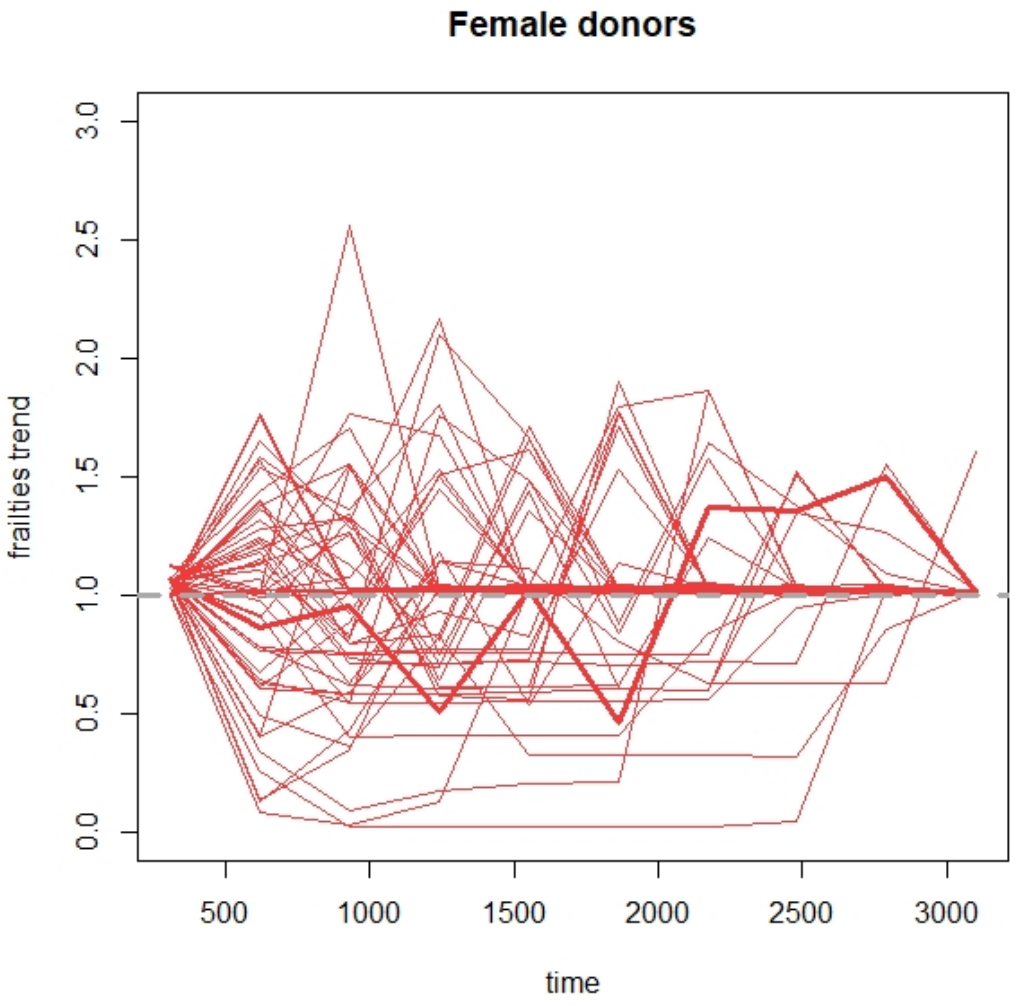

Figure 5.5: Reduced *Model 2*, v_{ik}'s trend over time for 100 randomly selected female donors

5.5 Robustness analysis

In this section, the robustness analysis for *Model 2* is presented. The analysis is to be made with respect the hyperparameter c, that regulates the variance of $\tilde{u}_i(t)$, as reported in Formula (4.9). The values considered are $c = 0.01$, $c = 1$, $c = 2$. Table 5.5 reports the goodness of fit evaluations in terms of both LPML and WAIC for the three simulations, along with the posterior means of the main parameters.

Parameter	**c = 0.01**	**c = 1**	**c = 2**
β_{SESSO}	-4.37	-3.15	-3.16
β_{FUMO}	-0.23	-0.40	-0.42
β_{ALCOOL}	-0.10	-0.23	-0.25
$\beta_{\text{ATTIVITAFISICA}}$	-0.17	-0.42	-0.45
β_{RH}	-2.98	-2.17	-2.19
β_{PMIN}	0.03	0.05	0.06
$\beta_{\text{TIPO\_0}}$	-0.26	-0.47	-0.50
$\beta_{\text{TIPO\_B}}$	-0.25	-0.49	-0.53
$\beta_{\text{TIPO\_AB}}$	-0.44	-0.77	-0.81
β_{EMOG}	1.11	1.08	1.09
$\beta_{\text{ETA\_PRIMA}}$	0.23	0.22	0.22
β_{BMI}	0.17	0.14	0.14
$\beta_{\text{SESSO\_EMOG}}$	-1.16	-1.13	-1.14
$\beta_{\text{SESSO\_BMI}}$	-0.17	-0.15	-0.15
$\beta_{\text{SESSO\_RH}}$	2.93	2.10	2.13
α_1	5.50	2.15	2.30
α_2	0.87	0.33	0.35
α_3	0.60	0.22	0.23
α_4	0.50	0.18	0.18
α_5	0.41	0.14	0.15
α_6	0.42	0.15	0.15
α_7	0.39	0.13	0.14
α_8	0.47	0.15	0.16
α_9	0.56	0.18	0.20
α_{10}	1.01	0.38	0.50
WAIC	**298843.9**	**330724.3**	**332461.6**
LPML	**-150621.2**	**-165206**	**-165935.9**

Table 5.5: Reduced *Model 2*, robustness analysis. The posterior mean for each parameter is reported

As c increases, a negligible change of performance can be observed both in terms of WAIC and LPML. The estimations of regression coefficients and baseline intensity function maintain unchanged sign and significance. Given these results, the model appears to be robust.

Chapter 6

Planning and profiling

The goal of this chapter is to consider AVIS' questions and needs and to answer to them thanks to the dataset that has been supplied through the selected "Model 2".

6.1 General overview

AVIS deals with **planning** and **profiling**. Planning is managerial, from the point of view of the internal organization and it answers to the question: "How much staff is needed at a certain moment?". In particular, knowing donors' characteristics and history: "How much blood is expected to be received in a certain amount of time?". AVIS collection center of Lambrate provides blood units to Niguarda hospital, that was opened on October 3, 1939 and today is one of the most important hospitals in Milan. Niguarda hospital's needs are described in terms of monthly blood units. Each donation corresponds to a unit, whose nominal weight is $430g$. One might ask: "Why people who weigh less than a certain threshold cannot donate?". The answer lies in the sustainability of the process: getting a unit has a fixed cost and it is not convenient to get half a unit from a person who weighs below the threshold, at the same cost as a full unit; moreover counting "half units" would become complicated. More is details, planning could be divided into:

- *monthly*: on average the AVIS center of Lambrate meets the demands of Niguarda, which are not fixed, but they can vary depending on the needs of that particular moment. Approximately 1500 donations are made per month. In this case, it is relevant to calculate the nominal level distinguishing among the different blood types;
- *weekly*: the planning level per week is considered for the sizing of the center, quantified in terms of staff needed for the blood drive. In this case, the distinction among different blood types is not necessary.

On the other hand, AVIS needs to carry out effective acquisition campaigns of new donors, which is the key aspect of profiling. The proposed approach is to define some typical profiles and to make predictions about their future donation history, in order to understand their propensity to donate. Through the discovery of the most effective profiles, AVIS will be able to decide where it is more likely to attract new donors, whether in high schools and/or universities and/or companies.

6.2 Profiling

Given a generic profile i, it is possible to forecast its propensity to donate thanks to Corollary 1 of Chapter 1. Assuming that the the first donation is made at time $T_0 = 0$ and by letting $W_1 = T_1 - T_0$ be the waiting time of the first recurrence (corresponding to the second donation), we compute the probability $\mathbb{P}(W_1 > t|T_0 = 0)$. We let t varying in the first three months in which profile i is allowed to donate: $t \in (\Phi_i + 1, \Phi_i + 90)$ where Φ_i is equal to 85 days for men and 150 for women. In this way, only $\widetilde{u}_{i1}$ is needed, because it covers the first 310 days of the donation process. We obtain:

$$\begin{aligned}\mathbb{P}(W_1 > t|T_0 = 0) &= \exp\left\{-\int_0^t \lambda_i(s)ds\right\} = \exp\left\{-\int_0^t \widetilde{u}_{i1}e^{\mathbf{x}_i\boldsymbol{\beta}}Y_i(s)ds\right\} = \\ &= \exp\left\{-\widetilde{u}_{i1}e^{\mathbf{x}_i\boldsymbol{\beta}}\int_0^t Y_i(s)ds\right\} = \begin{cases}1 & \text{if } t \leq \Phi_i \\ \exp\left\{-\widetilde{u}_{i1}e^{\mathbf{x}_i\boldsymbol{\beta}}(t-\phi_i)\right\} & \text{if } t > \Phi_i\end{cases}\end{aligned} \tag{6.1}$$

In order to provide a concrete example, some profiles have been randomly selected and they are reported in Table 6.1.

	Profile 1	**Profile 2**	**Profile 3**	**Profile 4**
Sex	*woman*	*man*	*woman*	*man*
Age at first donation	19	25	40	60
Weight	$55kg$	$75kg$	$60kg$	$80kg$
Heigth	$1.65m$	$1.80m$	$1.70m$	$1.75m$
Range BMI	*healthy weight*	*healthy weight*	*healthy weight*	*overweight*
Smoker	*no*	*yes*	*no*	*no*
Active Life	*yes*	*yes*	*no*	*no*
Rh	*negative*	*negative*	*positive*	*positive*
Blood type	0	A	B	AB

Table 6.1: Selected profiles

The hemoglobin and the minimum pressure of each profile are sampled from a Normal distribution with mean and standard deviation obtained by the data of similar donors in the dataset, with the same sex and the same age at first donation. After having sampled these values, they are kept constant during the 90 days analyzed; in fact, in *Model 2* - as in the other models presented - the time-dependent covariates vary in correspondence of each donation, but not in the time between two donations. Then the standardization of all covariates follows, referring to the means and standard deviations reported in Table 2.6 and Table 2.8. At this point, the four vectors of covariates $\mathbf{x}_i$ are complete and ready to be inserted in Formula (6.1). Moreover, recalling that

$$\widetilde{u}_{i1}|\alpha_1, c \overset{iid}{\sim} \text{Gamma}(l\alpha_1, l) \qquad \alpha_1 \sim \text{Gamma}(\delta, \delta)$$

where $\delta = 2$, $l = 310c$ and $c = 0.01$, then M values of $\widetilde{u}_{i1}$ have to be simulated from the sample $(\alpha_1^{(m)})_{m=1,\dots,M}$ of reduced *Model 2*. Figure 6.1 shows the posterior densities of $\widetilde{u}_{i1}$, $i = 1, \dots, 4$ for the profiles reported in Table 6.1.

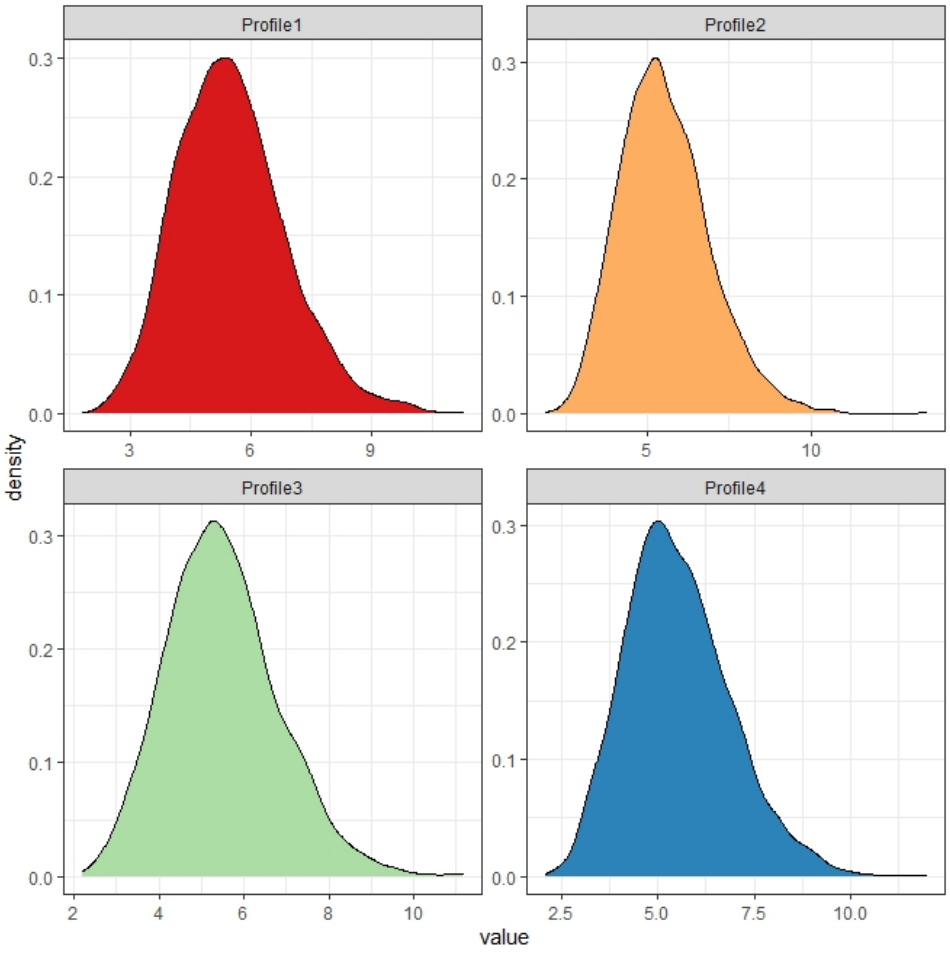

Figure 6.1: $\widetilde{u}_{i1}$'s posterior densities simulated for the profiles reported in Table 6.1

Hence, for each time $t \in (\Phi_i + 1, \Phi_i + 90)$, M different probabilities are computed. Figure 6.2

shows the posterior mean of the trends of $\mathbb{P}(W_1 > t|T_0 = 0)$ for four selected profiles, based on Formula (6.1).

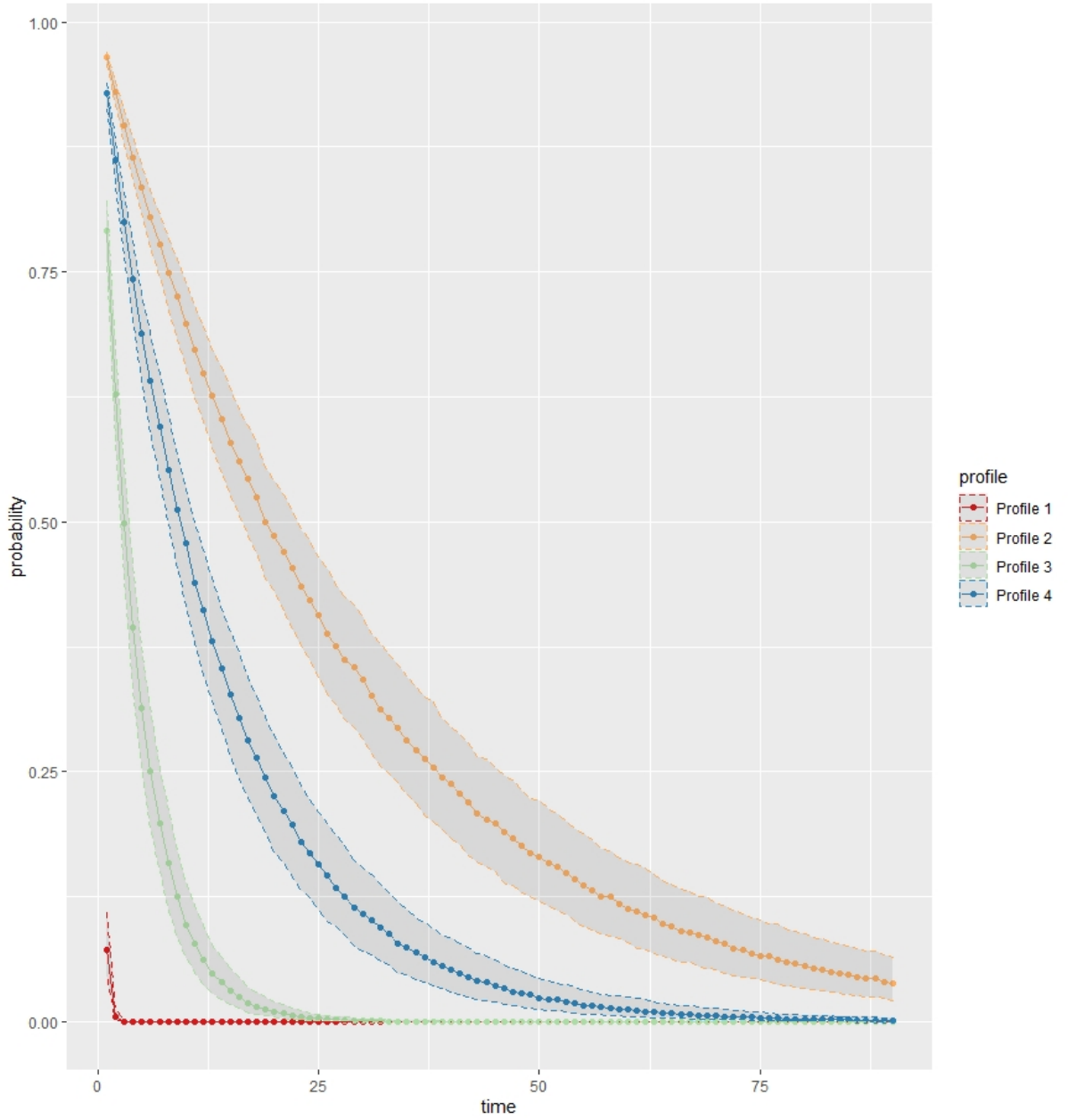

Figure 6.2: Representation of $\mathbb{P}(W_1 > t|T_0 = 0)$ in the first 90 days in which each selected profile is allowed to donate, after first donation. Credible intervals $(q_{0.25}, q_{0.75})$ are added as dashed lines

The faster these graphs decrease, the better it is, because it means that they are more likely to donate as soon as they can do it. It can be noticed that the best profiles are *Profile 1* and *Profile 3*, that represent both female donors; this reflects the analysis carried out, in which we noticed that women show a more consistent behaviour in the donation process. Even if *Profile 3*'s probability is high at the beginning, it decreases very rapidly in the first three weeks, after

which it closely approaches to zero. On the other hand, *Profile 2* and *Profile 4* show a lower propensity to donate and their probabilities decrease more slowly. At the end of the three months, all profiles, with the exception of *Profile 2*, have a probability that approaches to zero, meaning that they are very likely to have made the second donation within the time window analysed.

To conclude, even if this approach provides only a qualitative information, it can be considered a valid tool to make a first categorization of individuals who are more likely to donate, supporting the management of AVIS' donation campaigns.

6.3 Planning

Given a generic

$$N(t) = \sum_{i=1}^{\widetilde{I}} 1(N_i(t) = 1) \tag{6.2}$$

where $\widetilde{I}$ is the total number of donors (to avoid possible misunderstandings of the reader, it is specified that $\widetilde{I}$ is greater than $I = 5937$, which is the number of *new donors* entered in the study in the period considered in this thesis), the goal is to obtain an estimate of

$$\mu(t) = \mathrm{E}[N(t)] \tag{6.3}$$

for a future period of time t. For example, if we consider next month, then $\mu(t)$ represents the forecast of the average number of donations that will take place in next $t = 30$ days and it should be at least equal to the number of blood units requested by Niguarda hospital, otherwise it will be necessary to take appropriate actions. A number of donations grater than the one strictly needed is not a problem, because extra blood units can be stored for future needs or distributed to smaller neighboring centers. Niguarda hospital's demand varies from month to month and it may increase considerably in some particular periods. On average, the number of monthly donations performed by AVIS section of Lambrate is around 1500. More in details, considering

$$\mu(t) = \mathrm{E}\left[\sum_{i=1}^{\widetilde{I}} \mathbf{1}(N_i(t) = 1)\right] = \sum_{i=1}^{\widetilde{I}} \mathbb{P}(N_i(t) = 1) \tag{6.4}$$

the goal is to estimate the posterior mean

$$\mathrm{E}[\mu(t)|\mathrm{Data}] = \sum_{i=1}^{\widetilde{I}} \mathbb{P}(N_i(t) = 1|\mathrm{Data}) \tag{6.5}$$

that can be approximated using the posterior MCMC sample. In fact, it follows from Theorem 2 in Chapter 1 that

$$p_i(t) = \mathbb{P}(N_i(t) = 1|\text{Data}) = 1 - \exp\left\{-\int_0^t \widetilde{u}_i(t) e^{\mathbf{x_i}'\boldsymbol{\beta}} Y_i(s) ds\right\} \tag{6.6}$$

From now on we restrict the complete group of donors $\widetilde{I}$ to the I donors analyzed in this thesis, so that we can compute $\mu(t)$ by using the MCMC sample $(u_i^{(m)}, \boldsymbol{\beta}_i^{(m)})_{m=1,\dots,M}$, focusing on the following future periods:

- $t = 1$ week, for planning the amount of staff needed;
- $t = 1$ month, for understanding whether the demands of Niguarda hospital are supposed to be met or not, with the distinction among blood types.

In order to simplify the problem further, we suppose that the future time-interval $(0, t]$ is immediately after the period in which data are collected, meaning that $t = 0$ corresponds to 30^{th} June 2018. Moreover for a donor $i = 1, \dots, I$, we have that $Y_i(s)$, $s \in (0, t]$, can be easily computed as

$$Y_i(s) = \begin{cases} 1 & \text{if } s - \tau_i \geq \phi_i \\ 0 & \text{otherwise} \end{cases} \tag{6.7}$$

where τ_i is the time of last donation for donor i and ϕ_i is equal to 85 for men and 150 for women. Figure 6.3 provides a general idea of the time representation of this analysis.

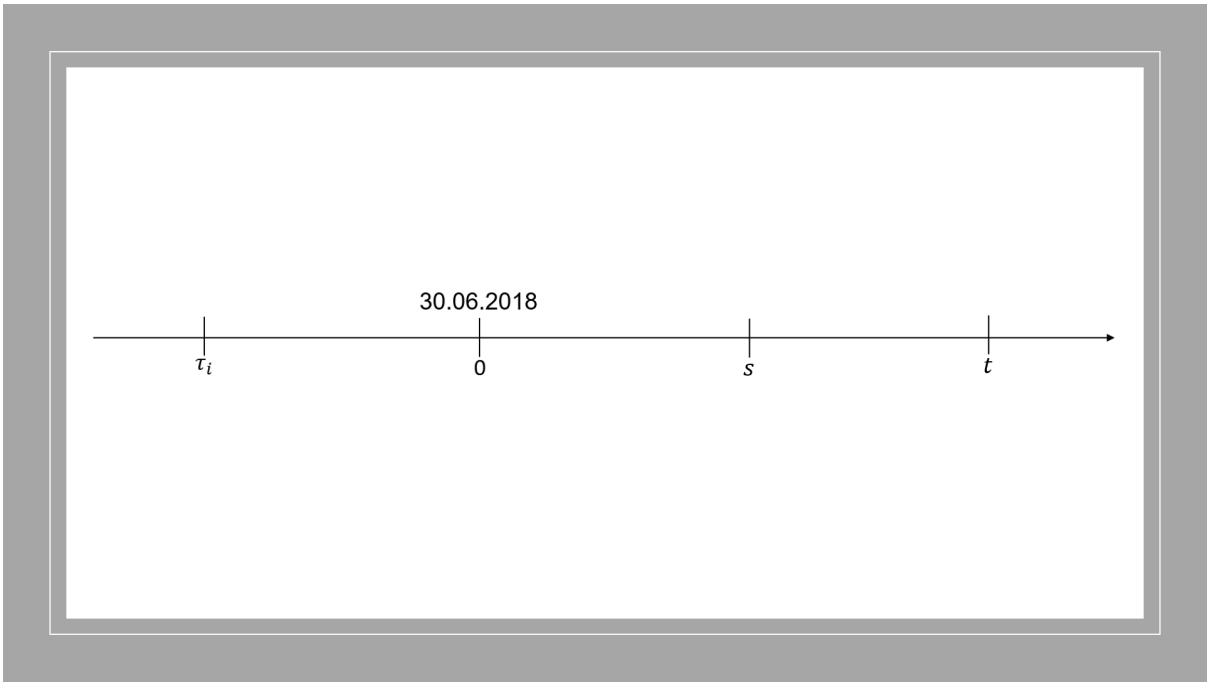

Figure 6.3: General time representation of the process, starting from donor i's last donation

In Formula (6.6) the lower bound of the integral, equal to "0" (30^{th} June 2018), is the same $\forall i$, but the distance from the initial time of each recurrent process, defined as τ_{0i}, is different for each donor. A more detailed representation of the whole process for each donor i is reported in Figure 6.4; in particular, s_i is the number of days between τ_{0i} and the 30^{th} June 2018.

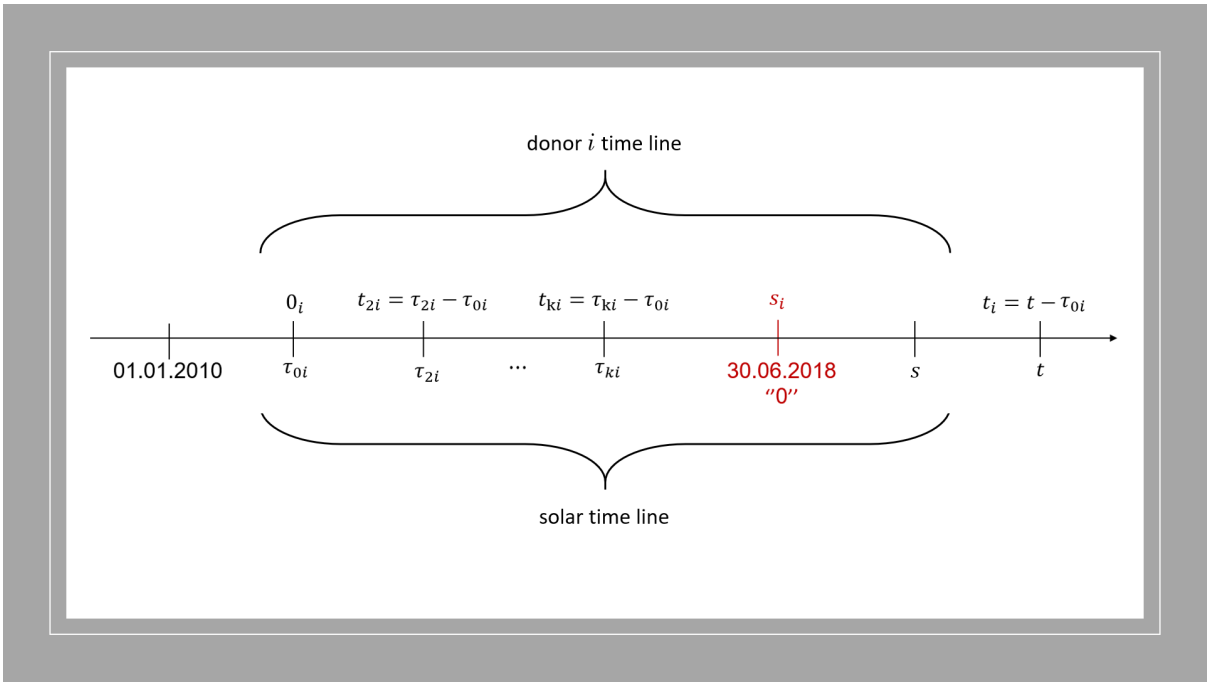

Figure 6.4: Detailed time representation of the process for donor i

At this point, it is clear that $N_i(t)$ has to be correctly specified as

$$\widetilde{N}_i(t) = N_i(t_i) - N_i(s_i) \tag{6.8}$$

and Formulae (6.4), (6.5) and (6.6) are adapted accordingly. As an intermediate step, we compute

$$\begin{aligned} \mathbb{P}(\widetilde{N}_i(t) = 1|\text{Data}) &= \mathbb{P}(N_i(t_i) - N_i(s_i) = 1) = 1 - \mathbb{P}(N_i(t_i) - N_i(s_i) = 0) \\ &= 1 - \exp\left\{ - \int_{s_i}^{t_i} \widetilde{u}_i(t) e^{\mathbf{x}_i'\boldsymbol{\beta}} Y_i(s) ds \right\} \end{aligned} \tag{6.9}$$

and we define

$$p_{it_i} = \mathbb{P}(\widetilde{N}_i(t_i) = 1|\text{Data}) \tag{6.10}$$

Now, first of all we deal with the computation of $Y_i(s)$:

- *case* 1: $s_i \geq t_{ki} + \phi_i \quad \longrightarrow \quad \begin{cases} Y_i(s) = 1 \quad \forall s \in [s_i, t_i] \\ p_{it_i} = 1 - \exp\left\{ - \int_{s_i}^{t_i} \widetilde{u}_i(t) e^{\mathbf{x}_i'\boldsymbol{\beta}} ds \right\} \end{cases}$

- *case* 2: $t_i \leq t_{ki} + \phi_i \quad \longrightarrow \quad \begin{cases} Y_i(s) = 0 \quad \forall s \in [s_i, t_i] \\ p_{it_i} = 0 \end{cases}$

- *case* 3: $s_i \leq t_{ki} + \phi_i \leq t_i \quad \longrightarrow \quad \begin{cases} Y_i(s) = 1 \quad \forall s \in [t_{ki} + \phi_i, t_i] \\ p_{it_i} = 1 - \exp\left\{ - \int_{t_{ki}+\phi_i}^{t_i} \widetilde{u}_i(t) e^{\mathbf{x}_i' \boldsymbol{\beta}} ds \right\} \end{cases}$

Second, we deal with the computation of the remaining part

$$\int_{b_i}^{t_i} \widetilde{u}_i(t) e^{\mathbf{x}_i' \boldsymbol{\beta}} ds \tag{6.11}$$

where the lower bound of the integral b_i varies accordingly to *case 1*, *case 2* and *case 3*. As reported in Figure 6.5, we organize the cut-points $0 = a_0 < a_1 < a_2 < \ldots < a_K$ in sequence starting from τ_{0i}, so that

$$\begin{aligned} a_{0i} &= \tau_{0i} \\ a_{1i} &= \tau_{0i} + a_1 \\ a_{2i} &= \tau_{0i} + a \\ &\vdots \\ a_{Ki} &= \tau_{0i} + a_K \end{aligned} \tag{6.12}$$

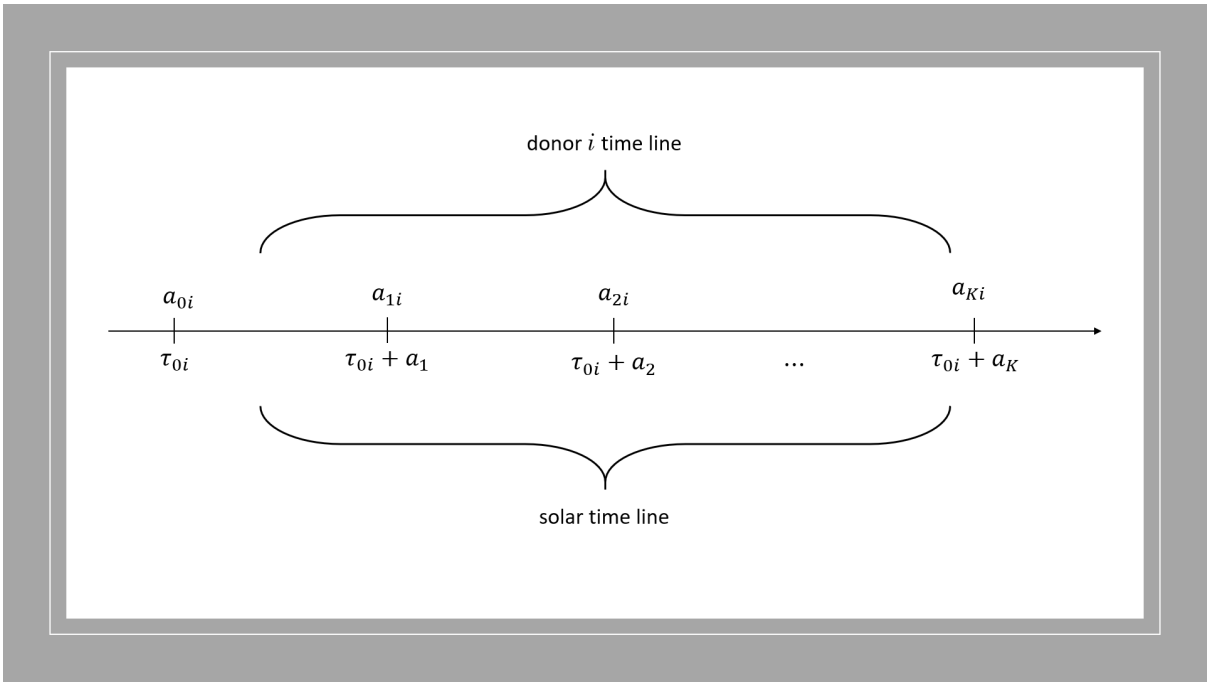

Figure 6.5: Time representation of the cut-points for donor i

If $b_i \in (a_{k_b-1,i}, a_{k_b,i}]$ and $t_i \in (a_{k_t-1,i}, a_{k_t,i}]$, then

$$\begin{aligned}\int_{b_i}^{t_i} \widetilde{u}_i(t) e^{\mathbf{x}_i'\boldsymbol{\beta}} ds = e^{\mathbf{x}_i'\boldsymbol{\beta}}[(a_{k_b,i} - b_i)\widetilde{u}_{k_b,i} + \\ + (a_{k_b+1,i} - a_{k_b,i})\widetilde{u}_{k_b+1,i} + \ldots + (a_{k_t,i} - a_{k_t-1,i})\widetilde{u}_{k_t,i}]\end{aligned} \tag{6.13}$$

where for the value of the hemoglobin and the minimum pressure for each donor has been computed as the average value among his/her donations.

Figure 6.6 shows the forecast of the average number of donors for next week and next month, where the present day is supposed to be the 30^{th} June 2018 (defined as day "0"). The number of forecasted montlhy donors is below the average threshold of 1500 individuals who are generally expected to donate each month at AVIS section of Lambrate. This gap can be explained by the fact that in this analysis we are not counting all donors, but only who has started his/her donation process not earlier than 1^{st} January 2010 (the reason for this choice is that for the previous period the information is not present on the two provided databases), with the additional assumption of having at least one recurrence (in fact all individuals who donated only once have been discarded, because their behavior was not significant to study the recurrences of the donation process). More in details, Figure 6.7 shows the distinction among blood types.

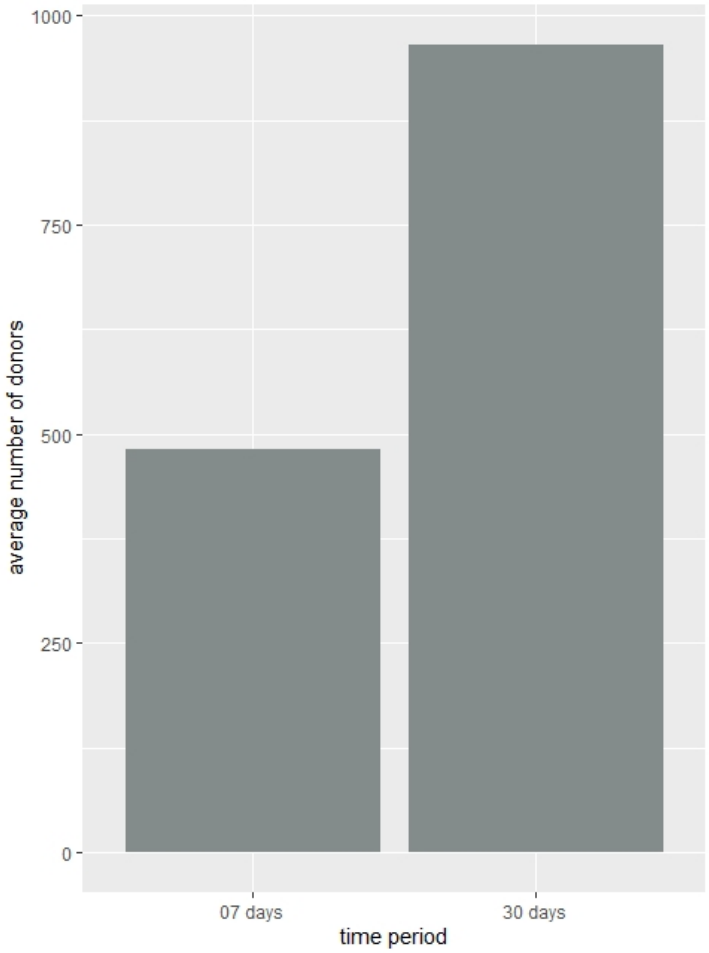

Figure 6.6: Forecast of the average number of individuals who are supposed to donate next week and next month, where the present is intended to be the last day of recorded data (30^{th} June 2018)

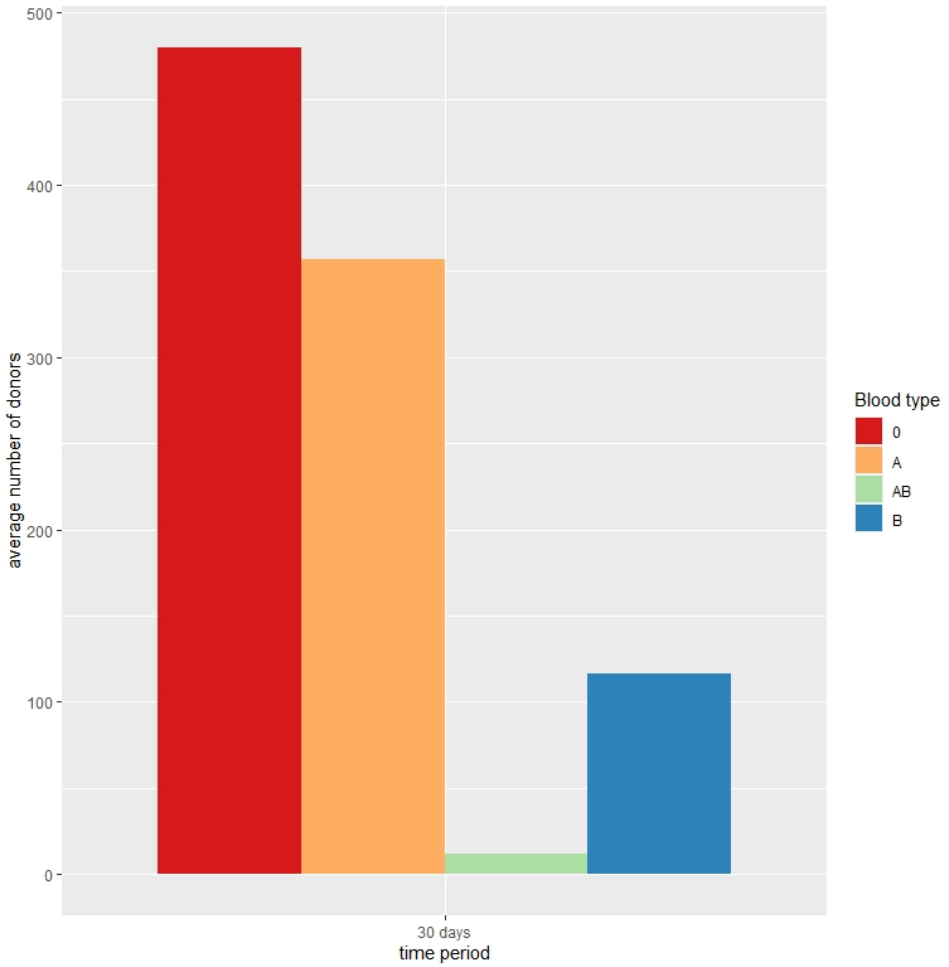

Figure 6.7: Forecast of the average number of individuals who are supposed to donate next month, divided by blood type, where the present is intended to be the last day of recorded data (30^{th} June 2018)

To conclude, it is interesting to notice that blood types are not all equally common. Some are very frequent, others very rare. In Italy group 0 is the most represented, followed by groups A, B and finally AB; it can be noticed that Figure 6.7 reflects this trend. If these proportions are on the whole true in Italy and in Europe, by examining the values of Asian and Indian populations, we would notice a clear prevalence of group B, sometimes even greater than group 0; these differences are related to our evolutionary history (Fondazione Gimema (2015)).

Conclusions and future developments

Starting from the works by Gianoli (2016) and Spinelli (2019), the thesis models the blood donation process as recurrent events in a Bayesian framework. Forecasting the number of donors' arrivals is essential to plan efficiently the storage of this important resource and to understand how much staff is needed in a future time window. Moreover, the model could be used to forecast the behaviour of donors with some specific profiles, in order to understand which categories of individuals are more likely to donate and to effectively carry out promotion campaigns to attract them.

The data are downloaded from two databases of the AVIS section of Lambrate, which is a very important blood collection centre in Milan and provides the blood bags to Niguarda hospital. In general, the data are obtained through questionnaires filled in by donors and through measurements of their vital parameters before each donation. There are several requirements to become a donor, such as age (between 18 and 65, or older with a medical permit), weight (exceeding 50 kilograms) and being in good medical condition. In addition to the safety of blood extracted, also the donors' health is important, so each time a person attempts to make a donation, the staff test his/her temperature, hemoglobin, blood pressure and pulse (Aldamiz-Echevarria and Aguirre-Garcia (2014)). Since the obtained dataset was not complete, a preliminary work for missing values imputation has been performed by means of the package `MICE` of the R statistical software.

Three different models are tested. *Model 0* is a natural extension of a model in Spinelli (2019), obtained with the inclusion of time-dependent covariates in addition to fixed ones and random frailties; all the formulas have been adapted accordingly. The time-dependent individual features measured at each donation are hemoglobin, pulse, minimum pressure and maximum pressure. *Model 1* and *Model 2* extend *Model 0* in two different ways: following Song and Kuo (2013), *Model 1* introduces an autoregressive behaviour for the random frailties, still keeping them separate from the baseline intensity function as in *Model 0*. Instead, *Model 2* thinks of the piecewise constant baseline intensity function as the common mean of the individual random frailties, i.e. as a prior hyperparameter; *Model 1* presented some identifiability problems that

have been overcome by the parameterization of *Model 2*. Sampling from the posterior distribution is achieved via the software platform Stan, that provides MCMC algorithms. After an accurate analysis, *Model 2* turns out to be the best one and its posterior analysis is presented in details in Chapter 5. The analysis highlights a decreasing trend of the baseline intensity function and identifies the individual features (donors gender, smoking habits, alcohol consumption, physical activity, BMI (Body Mass Index), Rh (Rhesus factor), blood type, age at first donation, hemoglobin and minimum pressure) as significant covariates that influence the intensity function and hence determine the donors personal propensity to donate. The addition of interactions between donors gender and hemoglobin, Rh and BMI respectively was found to be significant in differentiating male and female donors behaviour.

The focus of Chapter 6 is on planning and profiling. Planning is a key aspect for AVIS' internal organization. Knowing in advance the number of incoming donors can lead to an optimal planning of the appointment scheduling system and to an efficient sizing of the centre in a future time-window, in terms of staff required. On the other hand, profiling concerns the acquisition campaigns of new donors. Blood collection centres need to know both the internal and external factors that affect potential donors. The internal factors include personal characteristics, experiences, motivations, attitudes and perceived risks, while external factors include information channels as webpages, posters, ads and verbal communication. Groundless fears among non-donors - based on their misperceptions about what can happen to them if they donate (whether they can get a disease or any other fear) - has to be overcome by educational communication, convincing prospective donors of the safety of both collection practices and supply (Aldamiz-Echevarria and Aguirre-Garcia (2014)). The use of the model proposed in this thesis can facilitate this process of education, as it provides to the blood collection centres a priori knowledge of the profiles that are more likely to donate and it allows them to organize their acquisition campaigns accordingly.

In order to have a complete picture in the forecast of the number of donors' arrivals, also new incoming donors should be considered. The model developed in this thesis focuses only on donors already present (*insample prediction*). This choice was made because new donors are very few compared to the recurrent ones. A possible enrichment may consist in the addition of a predictive on new donors' arrivals (*outsample prediction*).

Another possible extension of the proposed model goes in the direction of including left-censoring times, in addition to the right-censure already present. In this way, the model can also be included information on "old" donors, which can not be traced at the beginning of their process of donation, because their first donation was done before recording regular entries in the

database.

In order to improve flexibility, parametric assumptions on the random frailties' distribution can be avoided by modelling them a priori with a Dirichlet Process. We mentioned this non-parametric Dirichlet choice in Section 1.10 and Song and Kuo (2013) was taken as a reference point. Also Pennell and Dunson (2006) deepens this approach. The structure proposed in the article is centered on a dynamic Gamma frailty model, but the true frailty distribution is allowed to deviate from the parametric form; the amount of uncertainty in the Gamma assumption is controlled by some hyperparameters. More in details, this method identifies clusters of subjects whose genetic traits convey a similar level of susceptibility at the outset of the study as well as clusters of subjects who experience similar increases in their susceptibility over each time-interval.

Appendix A

MICE notation and algorithm

A.1 Notation

Let Y_j with $j = 1, \ldots, p$ one of the p incomplete covariates, where $\boldsymbol{Y} = (Y_1, \ldots, Y_p)$. The observed and missing parts of Y_j are denoted by Y_j^{obs} and Y_j^{mis}, respectively, so $\boldsymbol{Y}^{obs} = (Y_1^{obs}, Y_2^{obs}, \ldots, Y_p^{obs})$ and $\boldsymbol{Y}^{mis} = (Y_1^{mis}, Y_2^{mis}, \ldots, Y_p^{mis})$ stand for the observed and missing data in $\mathbf{Y}$. The number of imputation is equal to $m \geq 1$. Finally let $Y_{-j} = (Y_1, \ldots, Y_{j-1}, Y_{j+1}, \ldots, Y_p)$ denote the collection of the $p - 1$ variables in $\mathbf{Y}$ except Y_j.

A.2 Algorithm

Let the hypothetically complete data Y be a partially observed random sample from the p-variate multivariate distribution $\mathbb{P}(\mathbf{Y}|\theta)$. We assume that the multivariate distribution of $\mathbf{Y}$ is completely specified by θ, a vector of unknown parameters. The problem is how to get the multivariate distribution of θ, either explicitly or implicitly. The `MICE` algorithm obtains the posterior distribution of θ by sampling iteratively from conditional distributions of the form

$$\mathbb{P}(Y_1|Y_{-1}, \theta_1)$$

$$\mathbb{P}(Y_2|Y_{-2}, \theta_2)$$

$$\vdots$$

$$\mathbb{P}(Y_p|Y_{-p}, \theta_p)$$

The parameters $\theta_1, \ldots, \theta_p$ are specific to the respective conditional densities and are not necessarily the product of a factorization of the 'true' joint distribution $\mathbb{P}(\mathbf{Y}|\theta)$. Starting from a

simple draw from observed marginal distributions, the t-th iteration of chained equations is a Gibbs sampler that successively draws

$$\theta_1^{*(t)} \sim \mathbb{P}(\theta_1 | Y_1^{obs}, Y_2^{(t-1)}, \ldots, Y_p^{(t-1)})$$

$$Y_1^{*(t)} \sim (Y_1^{obs}, Y_2^{(t-1)}, \ldots, Y_p^{(t-1)}, \theta_1^{*(t)})$$

$$\vdots$$

$$\theta_p^{*(t)} \sim \mathbb{P}(\theta_p | Y_p^{obs}, Y_1^{(t)}, \ldots, Y_{p-1}^{(t)})$$

$$Y_p^{*(t)} \sim (Y_p | Y_p^{obs}, Y_1^{(t)}, \ldots, Y_p^{(t)}, \theta_p^{*(t)})$$

where $Y_j^{(t)} = (Y_j^{obs}; Y_j^{*(t)})$ the j-th imputed variable at iteration t. Observe that previous imputations $Y_j^{*(t-1)}$ only enter $Y_j*(t)$ through its relation with other variables, and not directly. Convergence can therefore be quite fast, unlike many other MCMC methods.

Appendix B

Computations

All details reported in this section are referred to *Model 0*. The extensions for *Model 1* and *Model 2* can be easily derived. The first important step is to effectively represent the likelihood's integral in `C++`:

$$\tau_{ik} = u_i \int_{a_{k-1}}^{a_k} Y_i(s) e^{\mathbf{x}_i'(s)\boldsymbol{\beta}} ds$$

for each donor $i = 1, \ldots, M$ and time-interval $k = 1, \ldots, K$, where $a_0, a_1, \ldots, a_K$ are the cut-points. We assume the covariates to be step functions:

$$\mathbf{x}_i(t) = \mathbf{x}_i(t_{ij}) \qquad t_{ij} \leq t < t_{i,j+1} \quad j = 1, \ldots, n_i$$

where $t_{i,n_i+1} = c_i$ and c_i is the censoring time for donor i. Moreover n_i is his/her total number of donations. The result written in pseudo-code is

$$\tau_{ik} = \sum_{j=1}^{n_i} u_i e^{\mathbf{x}_i(t_{ij})'\boldsymbol{\beta}} ((t_{i,j+1} \wedge a_k) - ((t_{ij} + \Phi_i) \vee a_{k-1}) \vee 0) \tag{B.1}$$

$$\Phi_i = \begin{cases} 150 & \text{if } i \text{ is female donor} \\ 85 & \text{if } i \text{ is male donor} \end{cases}$$

The resulting log-likelihood is

$$\log(\mathscr{L}) = \sum_{k=1}^{K} n_{.k} \, log \, \lambda_k + \sum_{i=1}^{M} n_i \, \log \, u_i + \sum_{i=i}^{M} \sum_{j=1}^{n_i} \mathbf{x}_i(t_{ij})'\boldsymbol{\beta} - \sum_{i=1}^{M} \sum_{k=1}^{K} u_i \lambda_k \tau_{ik} \tag{B.2}$$

where $n_{.k}$ is the total number of donations in k-th interval.

Appendix C

The model written in C++ and functions for goodness of fit

The package `rstan` is the R interface to Stan and its source code is hosted on GitHub. The `stan` function does all of the work of fitting a Stan model and returning the results as an instance of `stanfit`. The main steps are the following:

- translate the Stan model to `C++` code;
- compile the `C++` code into a binary shared object, which is loaded into the current R session (an object `stanmodel` is created);
- draw samples and wrap them in a `stanfit` object.

The returned object can be used with methods such as *print*, *summary*, and *plot* to inspect and retrieve the results of the fitted model. The script written in Stan for *Model 0* is reported below (the details on `transformed data`'s initialization are reported in Appendix D).

```
data{
  int<lower=1> M;                  // number of donors
  int<lower=1> Nmax;               // max number of recurrences
  int<lower=2> K;                  // number of nodes (K-1 intervals)
  int<lower=0> P;                  // number of covariates
  int<lower=0> delay[2];           // rest time after last donation
  int<lower=1, upper=2> sex[M];    // donor's gender
  matrix[M,Nmax+1] times;          // events' times  + censoring time
                                   // negative time if event doesn't happen
  vector[K] nodi;                  // nodes
  int<lower=1> last[M];            // index of last observation
  matrix[Nmax, P] X[M];            // design array of matrices
```

```
  real<lower=0> alpha;              // u_i's shape parameter
  real<lower=0> a_lambda;           // lambda_k's shape parameter
  real<lower=0> b_lambda;           // lambda_k's scale parameter
  real<lower=0> sigma2_b;           // beta_p's variance
  real<lower=0> delta;              // beta_0's shape and scale parameter
 }

transformed data{
 matrix[K-1,M] n;              // number of events in each interval for each donor
 vector[K-1] npunto;           // number of events in each interval
 int nind[M];                  // number of events for each donor
 vector[M] n_donor;            // numer of donations for each donor
 matrix[Nmax, K-1] tau[M];     // K nodes and K-1 intervals
 int kappa[M, Nmax];           // indicate in which interval donations take place

 for(i in 1:M){
   nind[i]=0;
   for(t in 1:(Nmax)){
     if(times[i,t]>=0) nind[i] +=1;
   }
 }

 for(i in 1:M){
     for(k in 1:(K-1)){
       n[k,i] = 0;
     for(j in 1:nind[i]){
       if((times[i,j]>nodi[k]) && (times[i,j]<=nodi[k+1])){
          n[k,i] += 1;
          }
     }
   }
 }

 for(k in 1:(K-1)) {
   npunto[k] = sum(n[k,]);
 }

 for(i in 1:M){
   n_donor[i] = sum(n[,i]);
 }
```

```
 for(i in 1:M){
   for(j in 1:Nmax){
     for(k in 1:(K-1)){
       if(j<=last[i])
         tau[i,j,k] = fmax(fmin(times[i,j+1], nodi[k+1]) -
                    - fmax(times[i,j]+delay[sex[i]], nodi[k]), 0);
       else
         tau[i,j,k] = 0;
     }
   }
 }

 for(i in 1:M){
   for(j in 1:last[i]){
     for(k in 1:(K-1)){
       if(nodi[k]<times[i,j] && times[i,j]<=nodi[k+1])
         kappa[i,j] = k;
     }
   }
 }
}

parameters{
  vector<lower=0>[K-1] lambda;
  vector[P] beta;
  vector<lower=0>[M] u;
  real<lower=0> beta_0;
}

model{
 // log-likelihood 1st part
 target += sum(npunto.*log(lambda))+sum(n*log(u));
 // log-likelihood 2nd part
 for(i in 1:M){
   target += sum(X[i]*beta)-(u[i]*
             *((exp(X[i]*beta))'*tau[i]*lambda));
 }
 // log-priors
 target += gamma_lpdf(lambda|a_lambda, b_lambda);
 target += normal_lpdf(beta|0, sigma2_b);
 target += gamma_lpdf(beta_0|delta, delta);
```

```
 target += gamma_lpdf(u|alpha, alpha/beta_0);
}

generated quantities{
  vector[M] log_lik;
  for(i in 1:M){
    log_lik[i] = 0;
    for(j in 1:last[i]){
        log_lik[i] += log(lambda[kappa[i,j]]);
    }
    log_lik[i] += n_donor[i]*log(u[i])+sum(X[i]*beta)-(u[i]*
                  *((exp(X[i]*beta))'*tau[i]*lambda));
  }
}
```

It can be noticed that `generated quantities` are useful to compute the operators of goodness of fit, WAIC and LPML, thanks to the following R functions:

```
WAIC = function(fit, param){
llik  = rstan::extract(fit, param)[[1]]
p_WAIC = sum(apply(llik, 2, var))
lppd = sum(apply(llik, 2, function(x) log(mean(exp(x)))))
WAIC_score = -2 * lppd + 2 * p_WAIC
return(WAIC_score)
}

LPML = function(fit) {
llik = rstan::extract(fit, 'log_lik')[[1]]
CPO.inv = apply(llik, 2, function(x) mean(1/exp(x)))
LPML_score = sum(log( 1/CPO.inv ))
return(LPML_score)
}
```

where `fit` is the final `stanfit` object that constains all the simulation's results.

Appendix D

Data preparation in R

After the missing values' imputation, the dataset `avisComplete.R` is obtained (called `avis` in the script). This dataset is the starting point for data's preparation which are then transferred to Stan. The following code uses the R package `tidyverse`.

```
# CREATION STANDARDIZED DATASET
# 1) Time-fixed covariates
donors=avis %>% group_by(CAI) %>%
              summarise(ETA_PRIMA=first(ETA_PRIMA), BMI=first(BMI),
              PMAX=first(PMAX), POLSO=first(POLSO))
donors=donors %>% mutate(ETA_PRIMA = (ETA_PRIMA - mean(ETA_PRIMA))/sd(ETA_PRIMA),
                                BMI = (BMI - mean(BMI))/sd(BMI),
                                PMAX = (PMAX - mean(PMAX))/sd(PMAX),
                                POLSO = (POLSO - mean(POLSO))/sd(POLSO))

# 2) Time-dependent covariates
avis$PMIN=(avis$PMIN-mean(avis$PMIN))/sd(avis$PMIN)
avis$EMOG=(avis$EMOG-mean(avis$EMOG))/sd(avis$EMOG)
avis$ETA_PRIMA=NULL
avis$BMI=NULL
avis$POLSO=NULL
avis$PMAX=NULL
avis=left_join(avis, donors)

# 3) Interaction SEX-HEMOGLOBIN
temp=rep(1, dim(avis)[1])
idx=which(avis$SESSO == 0)
temp[idx]=0
avis$SESSO_EMOG=temp*(as.numeric(avis$EMOG))
```

```
# 4) Interaction SEX-BMI
avis$SESSO_BMI=temp*(as.numeric(avis$BMI))

# 5) Interaction SEX-RH
avis$SESSO_RH=temp*(as.numeric(avis$RH))

save(avis, file="avisStandard.R")
```

The dataset `avisStandard.R` is the final one that allows us to create all the needed data which are then transferred to Stan, in order to perform the MCMC simulations. It contains standardized numerical variables and their interactions.

It is now particularly important to understand how the matrix `times` and the array of matrixes `X` are obtained. Their creation requires the use of two temporary datasets: `donors`, created in previous script, which stores only time-fixed information and `num_repetition`, which stores the number of recurrences. Everything is grouped by CAI.

```
M = length(donors$CAI)
num_repetition = avis %>% group_by(CAI) %>% summarise(rep=n(),
                  censoring=first(censoring))
Nmax = max(num_repetition$rep)
tempi = avis[c('CAI', 'time')]
times = matrix(-1, nrow=M, ncol=Nmax+1)
for(i in 1:M){
  tempi_singolo = tempi %>% filter(CAI == num_repetition$CAI[i])
  for(j in 1:num_repetition$rep[i]){
    times[i,j] = as.numeric(tempi_singolo[j,2])
  }
  times[i, num_repetition$rep[i]+1] = num_repetition$censoring[i]
}
save(times, file='times.RData')
```

```
X = array(0, dim=c(M, Nmax, P))
for(i in 1:M){
  df_donor = avis %>% filter(CAI == num_repetition$CAI[i])
  n = num_repetition$rep[i]
  X[i, 1:n, 1:P] = array(as.matrix(df_donor[,c(col1,col2,...)]))
}
save(X, file = "designArray.RData")
```

Appendix E

Convergence diagnostics

Focusing on reduced *Model 2*, the analysis of the following diagnostic tools is performed:

- trace plots;
- potential scale reduction statistic, $\hat{R}$;
- effective sample size metric, n_{eff};
- autocorrelation plots.

E.1 Beta regression coefficients

Trace plots, which are shown in Figure E.1, are time series plots of the Markov chains and they show the evolution of the parameters over the 2000 sampling iterations for each chain (warm-up iterations are not reported). It can be noticed that all chains seem to explore the same region of their parameter values, which is a good sign. Another way to monitor whether a chain has converged to the equilibrium distribution is to compare its behavior to other randomly initialized chains. This is the motivation for the **potential scale reduction** statistic, $\hat{R}$. The $\hat{R}$ statistic measures the ratio of the average variance of draws within each chain to the variance of the pooled draws across chains; if all chains are at equilibrium, these will be the same and $\hat{R}$ will be one. If the chains have not converged to a common distribution, the $\hat{R}$ statistic will be greater than one. The **effective sample size**, denoted as n_{eff}, is an estimate of the number of independent draws from the posterior distribution. The n_{eff} metric used in Stan is based on the ability of the draws to estimate the true mean value of the parameter.

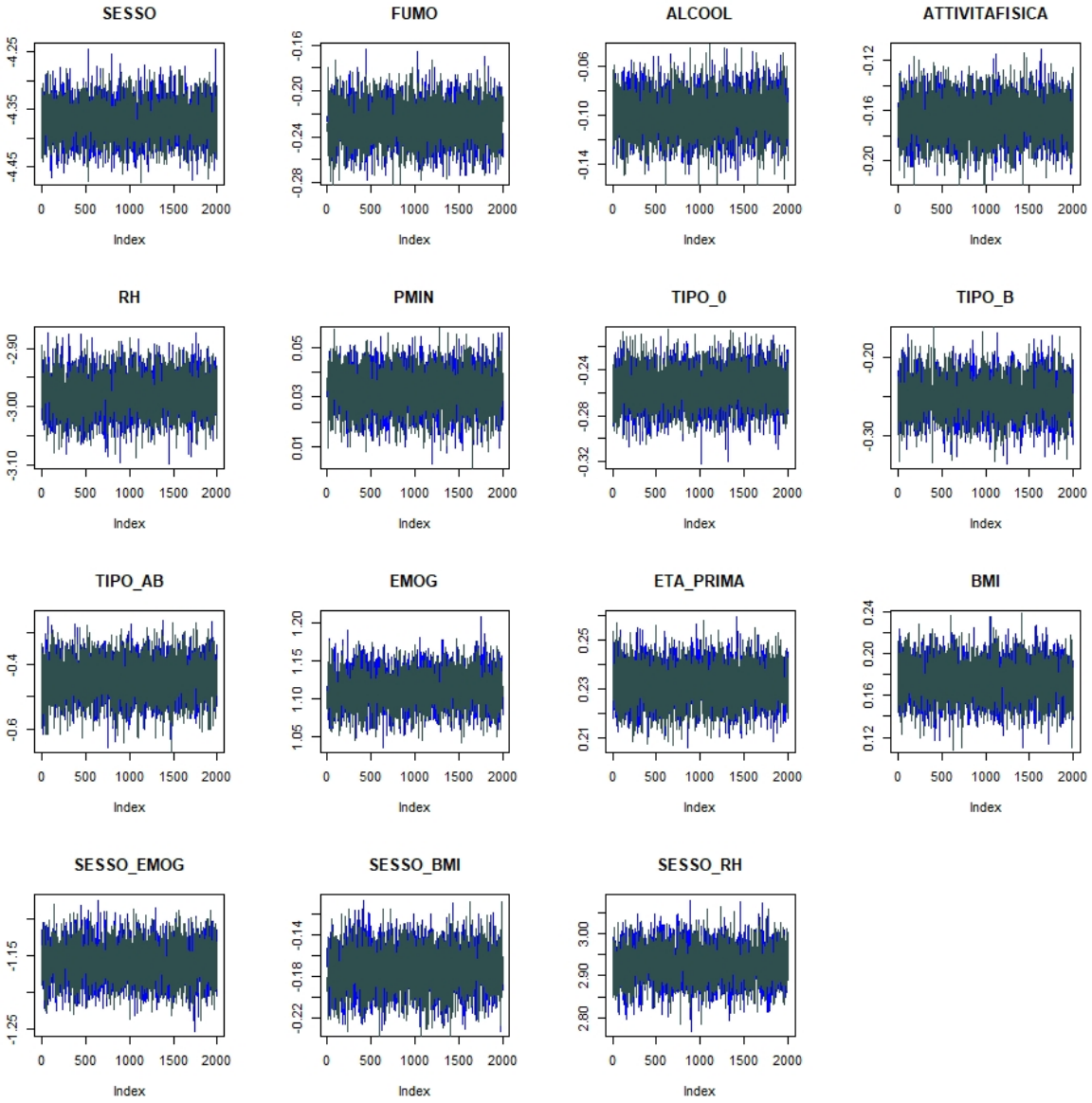

Figure E.1: Reduces *Model 2*, β_p's trace plots

In particular

$$n_{\text{eff}} = \frac{N}{1 + 2\sum_{k=1}^{\infty} \rho(k)} \tag{E.1}$$

where N is the total sample size and $\rho(k)$ is the **autocorrelation** of the chain at lag k. Since the draws from Markov chain are not independent, n_{eff} is usually smaller than N if autocorrelation is present. The larger the ratio of n_{eff} to N, the better the model is. These values are reported in Table E.1. Since n_{eff}/N decreases as autocorrelation becomes more extreme, it is useful to visualize also the autocorrelation (Figure E.2).

Parameter	n_{eff}	$\hat{R}$
β_{SESSO}	292	1.01
β_{FUMO}	3987	1.00
β_{ALCOOL}	5123	1.00
$\beta_{\text{ATTIVITAFISICA}}$	2527	1.00
β_{RH}	946	1.00
β_{PMIN}	4423	1.00
$\beta_{\text{TIPO\_0}}$	3254	1.00
$\beta_{\text{TIPO\_B}}$	4068	1.00
$\beta_{\text{TIPO\_AB}}$	4977	1.00
β_{EMOG}	577	1.00
$\beta_{\text{ETA\_PRIMA}}$	4441	1.00
β_{BMI}	2081	1.00
$\beta_{\text{SESSO\_EMOG}}$	692	1.00
$\beta_{\text{SESSO\_BMI}}$	2188	1.00
$\beta_{\text{SESSO\_RH}}$	1433	1.00

Table E.1: Reduced *Model 2*, diagnostic parameter for β_ps

In general, positive autocorrelation is bad, because it means the chain tends to stay in the same area between iterations, while the goal is it quickly drops to zero with increasing lags. Negative autocorrelation is possible and it is useful, as it indicates fast convergence of sample mean towards true mean.

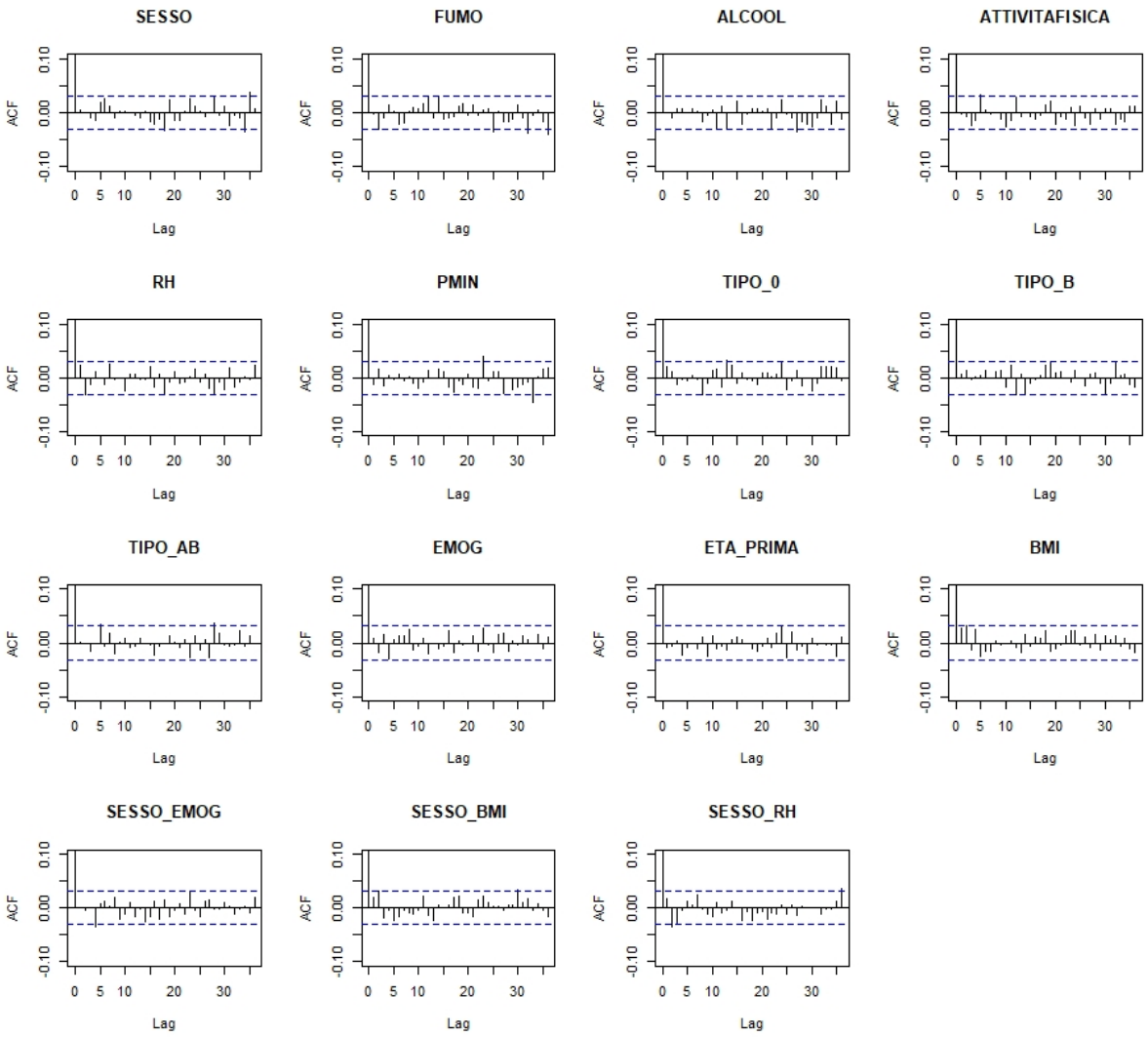

Figure E.2: Reduces *Model 2*, β_p's autocorrelation plots

E.2 Alpha baseline intensity function

Diagnostic parameters for α_k, $k = 1, \ldots, K$ are reported in Table E.2. Trace plots (Figure E.3) and autocorrelation plots (Figure E.4) follow, showing an overall good performance.

Parameter	n_{eff}	$\hat{R}$
α_1	79	1.02
α_2	115	1.02
α_3	140	1.01
α_4	155	1.02
α_5	205	1.02
α_6	250	1.01
α_7	210	1.01
α_8	233	1.00
α_9	125	1.00
α_{10}	113	1.00

Table E.2: Reduced *Model 2*, diagnostic parameters for α_k's

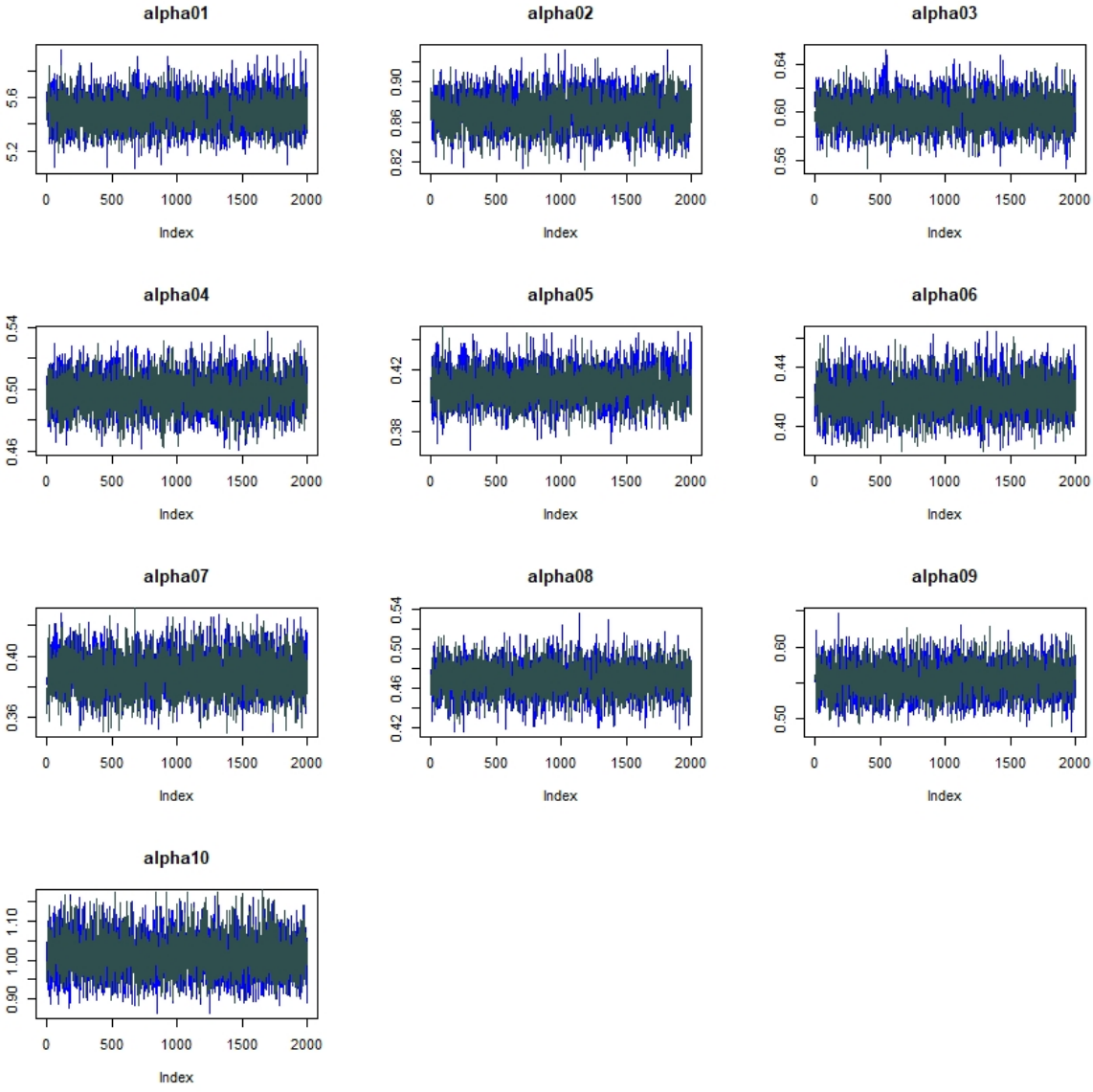

Figure E.3: Reduced *Model 2*, α_k's trace plots

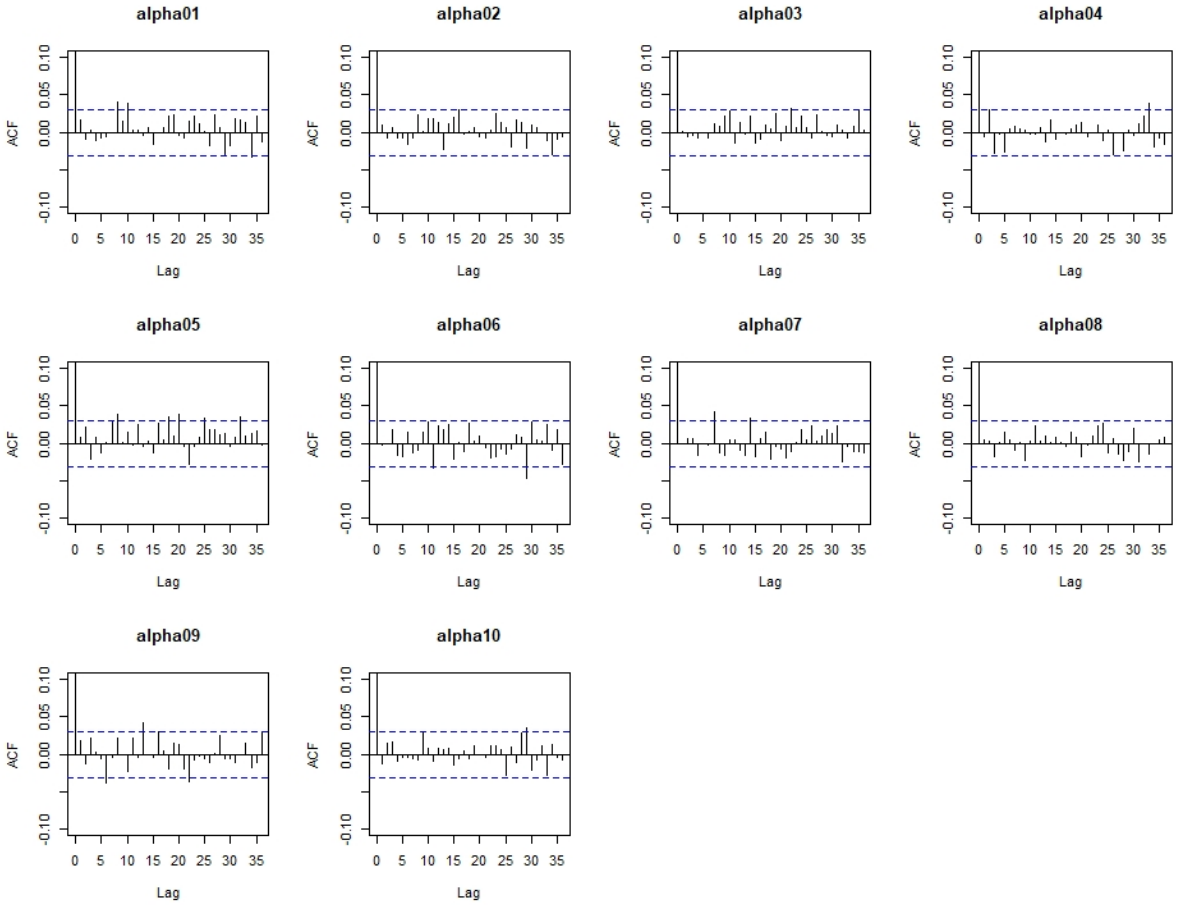

Figure E.4: Reduced *Model 2*, α_k's autocorrelation plots

Bibliography

Aalen, O., Borgan, O., & Gjessing, H. K. (2008). *Survival and event history analysis* (Vol. 1). Springer.

Aldamiz-Echevarria, C., & Aguirre-Garcia, M. S. (2014). A behavior model for blood donors and marketing strategies to retain and attract them. *Revista Latino-Americana de Enfermagem*, *22*(3), 467–475.

Christensen, R., Johnson, W., Branscum, A., & Hanson, T. (2011). Bayesian ideas and data analysis: An introduction for scientists and statisticians. *Chapman Hall/CRC Texts in Statistical Science*, 49–50.

Cook, R. J., & Lawless, J. (2007). *The statistical analysis of recurrent events* (Vol. 1). Springer.

Epifani, I., & Nicolini, R. (2013). On the population density distribution across space: A probabilistic approach. *Journal of Regional Science*, *53*(3), 481–510.

Fondazione Gimema. (2015). Diffusione dei gruppi sanguigni. https://www.gimema.it/diffusione-dei-gruppi-sanguigni/

Gelman, A., Hwang, J., & Vehtari, A. (2013). Understanding predictive information criteria for bayesian models. *Statistics and Computing*, *24*(6), 997–1016.

Gianoli, I. (2016). *Analysis of gap times of recurrent blood donations via bayesian nonparametric models* (Doctoral dissertation). Politecnico di Milano.

Pennell, M. L., & Dunson, D. B. (2006). Bayesian semiparametric dynamic frailty models for multiple event time data. *The International Biometric Society*, *62*(4), 1044–1052.

Rubin, D. B. (1976). Inference and missing data. *Journal Storage*, *63*(3), 581–592.

Song, C., & Kuo, L. (2013). Dynamic frailty and change point models for recurrent events data. *Journal of the Iranian Statistical Society*, *12*(1), 127–151.

Spinelli, E. (2019). *Count processes approach to recurrent event data: A bayesian model for blood donations* (Doctoral dissertation). Politecnico di Milano.

Stan Development Team. (2020). RStan: The R interface to Stan [R package version 2.21.2]. http://mc-stan.org/

Vehtari, A., Gelman, A., & Gabry, J. (2016). Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing*, *27*(5), 413–1432.

Ringraziamenti

Sono grata in primo luogo a tutti i professori che ho incontrato lungo il mio percorso accademico e a tutte le persone che ho avuto modo di conoscere. Questi anni al Politecnico di Milano sono stati in grado di plasmare la mia essenza, permettendomi di realizzare appieno come la determinazione, l'impegno e la perseveranza siano elementi essenziali per raggiungere i propri obiettivi.

Desidero ringraziare la Prof.ssa Ilenia Epifani, relatrice di questa tesi e la co-relatrice Prof.ssa Alessandra Guglielmi, per il loro costante aiuto e i loro preziosi consigli. In un momento delicato come quello che stiamo affrontando, realizzo a maggior ragione come siano riuscite a farmi sentire appieno parte di una realtà importante come quella del Politecnico di Milano, accorciando le distanze fisiche inevitabili grazie alla loro disponibilità, professionalità e umanità.

Ringrazio la mia famiglia per avermi sempre supportata, permettendomi di affrontare con serenità questo percorso accademico e ringrazio Paolo per avermi trasmesso una carica pazzesca negli ultimi mesi e per avermi sostenuta sin da subito. Ringrazio poi i miei amici di sempre, Ilaria, Jacopo, Silvia e Angela, che più di tutti conoscono le diverse sfacettature che hanno caratterizzato questi anni, dai momenti più difficili a quelli più felici.

Ringrazio infine, ma non per importanza, tutte le persone del gruppo "Poggio e Sambuca fa Pirez". Come dimenticare le pause pranzo al sole distesi in piazza Leo, le partite a carte, le piade dell'Harp, il Carrefour, le lezioni in Nave, il Trifoglio, le giornate in biblio, l'aperiparco, le serate e tutte le vacanze fatte insieme. Grazie ad Anna, Ele, Agne, Tappi, Pirez, Sampa, Spigolo, Franco, Fiore, Cerre, Paggia, Bole, Elisa, Giulia, Bitta, Albi, Ariel, Dave, Guenza. Ad maiora!