DECLARATORIA SULLA TESI DI DOTTORATO

Da inserire come prima pagina della tesi

Il/la sottoscritto/a

COGNOME Mura

NOME Fabrizio

Matricola di iscrizione al Dottorato 1194976

Titolo della tesi:

Birth and Death Models with Missing Data. An Application to the Survival of Firms

Dottorato di ricerca in Statistica

Ciclo XXII

Tutor del dottorando Marco Bonetti

Anno di discussione 2012

DICHIARA

sotto la sua responsabilità di essere a conoscenza:

1) che, ai sensi del D.P.R. 28.12.2000, N. 445, le dichiarazioni mendaci, la falsità negli atti e l'uso di atti falsi sono puniti ai sensi del codice penale e delle Leggi speciali in materia, e che nel caso ricorressero dette ipotesi, decade fin dall'inizio e senza necessità di nessuna formalità dai benefici previsti dalla presente declaratoria e da quella sull'embargo;

2) che l'Università ha l'obbligo, ai sensi dell'art. 6, comma 11, del Decreto Ministeriale 30 aprile 1999 prot. n. 224/1999, di curare il deposito di copia della tesi finale presso le Biblioteche Nazionali Centrali di Roma e Firenze, dove sarà consentita la consultabilità, fatto salvo l'eventuale embargo legato alla necessità di tutelare i diritti di enti esterni terzi e di sfruttamento industriale/commerciale dei contenuti della tesi;

3) che il Servizio Biblioteca Bocconi archivierà la tesi nel proprio Archivio istituzionale ad Accesso Aperto e che consentirà unicamente la consultabilità on-line del testo completo (fatto salvo l'eventuale embargo);

4) che per l'archiviazione presso la Biblioteca Bocconi, l'Università richiede che la tesi sia consegnata dal dottorando alla Società NORMADEC (operante in nome e per conto dell'Università) tramite procedura on-line con contenuto non modificabile e che la Società Normadec indicherà in ogni piè di pagina le seguenti informazioni:

- tesi di dottorato (titolo *tesi)* .... ;

- di *(cognome e nome del dottorando)* .... ;
- discussa presso l'Università commerciale Luigi Bocconi – Milano nell'anno .... (*anno di discussione);*
- La tesi è tutelata dalla normativa sul diritto d'autore (legge 22 aprile 1941, n.633 e successive integrazioni e modifiche). Sono comunque fatti salvi i diritti dell'Università Commerciale Luigi Bocconi di riproduzione per scopi di ricerca e didattici, con citazione della fonte;
- **solo nel caso sia stata sottoscritta apposita altra dichiarazione con richiesta di embargo**: La tesi è soggetta ad embargo della durata di .... mesi (indicare durata embargo);

5) che la copia della tesi depositata presso la NORMADEC tramite procedura on-line è del tutto identica a quelle consegnate/inviate ai Commissari e a qualsiasi altra copia depositata negli Uffici dell'Ateneo in forma cartacea o digitale e che di conseguenza va esclusa qualsiasi responsabilità dell'Ateneo stesso per quanto riguarda eventuali errori, imprecisioni o omissioni nei contenuti della tesi;

6) che il contenuto e l'organizzazione della tesi è opera originale realizzata dal sottoscritto e non compromette in alcun modo i diritti di terzi (legge 22 aprile 1941, n.633 e successive integrazioni e modifiche), ivi compresi quelli relativi alla sicurezza dei dati personali; che pertanto l'Università è in ogni caso esente da responsabilità di qualsivoglia natura, civile, amministrativa o penale e sarà dal sottoscritto tenuta indenne da qualsiasi richiesta o rivendicazione da parte di terzi;

7) **scegliere l'ipotesi 7a o 7b indicate di seguito**:

~~7a)~~ che la tesi di dottorato non è il risultato di attività rientranti nella normativa sulla proprietà industriale, non è stata prodotta nell'ambito di progetti finanziati da soggetti pubblici o privati con vincoli alla divulgazione dei risultati; non è oggetto di eventuali registrazioni di tipo brevettale o di tutela, e quindi non è soggetta a embargo;

Oppure

7b) che la tesi di Dottorato rientra in una delle ipotesi di embargo previste nell'apposita dichiarazione "**RICHIESTA DI EMBARGO DELLA TESI DI DOTTORATO**" sottoscritta a parte.

Data 31/01/2012_______

F.to (indicare nome e cognome) Fabrizio Mura

Birth and Death Models with Missing Data. An Application to the Survival of Firms

Fabrizio E. Mura

January 31, 2012

Contents

1 Introduction 7

2 Missing Data 9

2.1 Introduction and Notation 9

2.2 Missing Data Mechanism 10

2.3 Traditional Approaches 17

2.4 Imputation Methods 21

2.4.1 Single Imputation 22

2.4.2 Multiple Imputation 23

2.5 Example: Regionalization of Public Investment Expenditure 26

2.5.1 Objectives 26

2.5.2 Data Dimensions 27

2.5.3 The public capital expenditure indicator 28

2.5.4 A Proposal for an Imputation Approach 29

3 Complete Inference Approaches 33

3.1 Ignorable Missing Data Mechanism 33

3.1.1 Sampling Distribution Inference 34

3.1.2 Direct Likelihood Inference 36

3.2 Non Ignorable Missing Data Mechanism 39

4 A Birth and Death Model with Missing Data **43**

4.1 Model Definition 43

4.2 Birth and Death model simulations 47

4.3 Monza and Brianza Businesses Dataset 50

4.4 Birth and Death model: an application to Monza and Brianza firms data . 55

4.5 A Birth and Death Distribution 61

4.6 Results for the Kummer M Function 62

4.7 Special Cases 66

5 Lexis Point Process **71**

5.1 Poisson Point Process 71

5.1.1 Simulation of the Poisson Point Process 73

5.2 The Lexis Diagram and the Lexis Point Process 74

5.2.1 Time Window Sampling Scheme 78

5.2.2 Piecewise Constant Intensities 83

5.2.3 Monza and Brianza: Mortality in the Construction Firms 84

5.2.4 Model Diagnostics 91

6 Conclusion **97**

A Kummer M Function **99**

B R Code **101**

List of Figures

2.1 MDM examples: complete $N(0,3)$ random sample, $n = 1000$ 11

2.2 MDM examples: Random Selection (not variable dependent) 13

2.3 MDM examples: Deterministic Selection (variable dependent) 14

2.4 MDM examples: Random Selection (variable dependent) 15

3.1 the true density (solid), the density estimated not considering the MDM (dashed) and the one considering the MDM (dotted) 42

4.1 Monza and Brianza dataset: Year of birth distribution 52

4.2 Monza and Brianza dataset: Year of death distribution 53

4.3 Monza and Brianza dataset: Time of death distribution 54

4.4 Monza and Brianza dataset: Proportion of missing by year of death 55

4.5 Monza and Brianza dataset: Observed versus estimated proportion of missingness 60

5.1 Lexis point process: homogeneous Poisson birth process and $Exp(1/3)$ lifetime 76

5.2 Lexis point process: time window on the $[4, 8]$ interval. 79

5.3 Monza and Brianza construction firms: year of birth 85

5.4 Monza and Brianza construction firms: year of death 86

5.5 Monza and Brianza construction firms: observed proportion by year of death 87

5.6 Monza and Brianza construction firms: $\hat{\beta}$ (star) and $\hat{\lambda}$ (circle) estimates . . 90

5.7 Monza and Brianza construction firms: observed (solid) vs simulated (dashed) complete data Survival function 93

5.8 Monza and Brianza construction firms: 3rd cohort observed (solid) vs simulated (dashed) complete data Survival function 94

Chapter 1

Introduction

This Thesis is motivated by a dataset containing the birth and death times of a selection of construction firms in the Italian Monza and Brianza province. This dataset is interesting for the sampling scheme followed and for its missing data pattern. In particular, only death times within a specific time window are observed, and a relevant part of the birth times are missing. This leads to a thorough discussion of missing data mechanisms and to the various approaches to dealing with them. Two different models for the motivating problem are then developed and used to perform inference on the latent birth-and-death process.

In Chapter 2 the definitions of different cases of missing data are illustrated. The first class of methods, proposed as a solution to this lack of information, is presented: traditional methods and imputation methods. All of these aim to obtain complete data by discarding or filling in missing information. After this stage all the models available for complete data can be implemented. Regionalization of Public Investment Expenditure is a presented as an example of application of the multiple imputation technique.

Chapter 3 introduces a different approach to missing data: complete inference. In this case missing data is neither discarded nor replaced: the missing data mechanism which generated missing data is considered explicitly and included into the model.

Following the complete inference approach, a birth and death model is presented in

Chapter 4. This model could be useful for survival analysis applications where birth time and death time are the variables of interest. For this model, the likelihood function including the missing data mechanism is derived analytically and an application to the survival of firms is provided as an example.

As the analysed firms data, were not only affected by a missing data mechanism, but was also collected according to a particular sampling scheme, Lexis point process is finally proposed in 5. This stochastic process turns out to be a really powerful tool for many combinations of missing data mechanisms and sampling schemes as pointed out by the really good fit to firms survival data presented at the end of the chapter.

Chapter 2

Missing Data

In the present chapter the general problem of missing data is introduced. The typical approach in this situation, is to obtain a complete in dataset on which is possible to perform an *ordinary* analysis. This target is achieved in two ways: the first one is to follow some rules for discarding missing observation, presented in section 2.3; the second one is to assign likely values to missing observation, as discussed in section 2.4. In the last section of this chapter, it is shown how the well established multiple imputation technique, presented in section 2.4, could help in the relevant and difficult process of the aggregation of the Italian public expenditure.

2.1 Introduction and Notation

Missingness is a familiar feature of many datasets, as data can fail to be recorded for several reasons such as defects of the measuring instruments, errors in data entry operations and so on, missingness can also occur as a result of the data collection process. Usually missing data are regarded as a technical nuisance and the typical approach is to discard them, as done by default by most of the statistical packages. In some cases this apporach could lead to biased results, since the missing data mechanism has an informative content that should be considered explicitly when performing inference. A systematic approach to the missing data problem dates back to the 1970's. In [1] implications on likelihood inference of the presence of missingness were studied and later developed in [2]. In [3] a fundamental and widespread computational tool, the *EM algorithm*, was introduced to handle missing data in some models. Helpful reviews on the topic are [4] and more recently [5] and [6].

Let us now introduce the notation that will be used in this and in the following chapters. The main inferential focus will be on the investigating the relationship between a response variable Y and a predictor X as in the usual regression framework. The upper case letters Y and X will be used to represent both random variables and random vectors of observations whenever the choice is clear by the context. In the vector case the n elements will be indexed by an i subscript, and if Y contains some missing elements it will be assumed to be partitioned in two blocks, one for the observed part and one for the missing part, as $Y = (Y_{obs}, Y_{mis})$. Two sets of indices $\mathcal{O}$ and $\mathcal{M}$ will be useful to refer to the block of observed and missing observations explicitly: $\mathcal{O} = \{i : Y_i \in Y_{obs}\}$ and $\mathcal{M} = \{i : Y_i \in Y_{mis}\}$. An indicator variable R is defined on Y to describe the missingness as follows:

$$R_i = \begin{cases} 1 & \text{when } Y_i \text{ is observed (i.e. } Y_i \in Y_{obs}) \\ 0 & \text{when } Y_i \text{ is missing (i.e. } Y_i \in Y_{mis}). \end{cases}$$

By construction, $R_i = 1$ iff $i \in \mathcal{O}$, and $R_i = 0$ iff $i \in \mathcal{M}$.

2.2 Missing Data Mechanism

A fundamental issue when analyzing data generating processes in which some variables may present a missing data part is the *missing data mechanism* (MDM), that is the process that makes the data unobservable. The MDM can be defined as the probability distribution of the indicator variable R (i.e. the probability that Y is observed), conditioned on data X and Y and indexed by the parameter ϕ, $f(R|X, Y; \phi)$. Let us consider as an illustration of the main ideas an example of the case $f(R|X, Y; \phi) = f(R|Y; \phi)$.

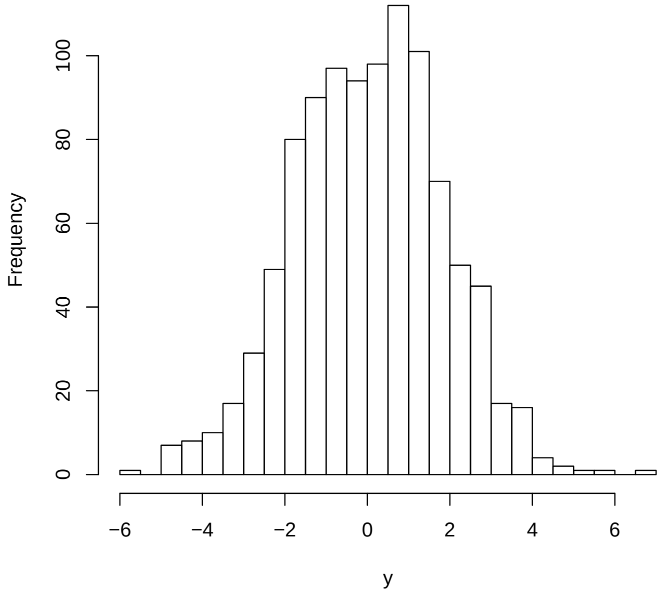

Figure 2.1: MDM examples: complete $N(0,3)$ random sample, $n = 1000$

Consider a sample of 1,000 observations from a Normal distribution with $\mu = 0$ and $\sigma^2 = 3$, represented in Fig. 2.1

Three different kinds of missing data mechanisms can be considered:

1. *Random selection (not variable dependent).* In this case each observation is exposed to a probability π (say 0.5) of being observed, the missing data mechanism can therefore be described as $f(R|Y;\phi) = \mathrm{P}\,(R = 1|Y;\phi) = 0.5$. In this case the inclusion probability is *independent* of the value of the variable Y itself. The observed data Y_{obs} is then a random sub sample of the complete data, and inference based on it is still valid. The distribution of the observed data resembles the original one (see Fig. 2.2) as

$$f(Y|R=1) = \frac{f(Y)\mathrm{P}\,(R=1|Y)}{\mathrm{P}\,(R=1)} = \pi\frac{f(Y)}{\mathrm{P}\,(R=1)} \propto f(Y)$$

	Original sample	Ex.1	Ex.2	Ex.3
n	1000	508	599	507
mean	0.0546	-0.0076	0.0663	1.0526
variance	3.2134	2.9996	5.1384	2.1187

Table 2.1: Univariate MDM examples

2. *Deterministic selection (variable dependent).* While in the previous example the missing data mechanism was not dependent on the variable value, in this case only the values outside the interval $[-1, 1]$ are observed, leading to the model for the mechanism $f(R|Y; \phi) = \mathbb{1}(Y \notin [-1, 1])$. The sampled vector distribution is shown in Fig. 2.3, and in this case it is clear that the likelihood model for the observed data has to take into account the missing data mechanism.

3. *Random selection (variable dependent).* In this last example, a mix of two main features of the previous ones is presented. As in the first case, a random selection acts on data but not indistinctly on all observations: the probability of observing an observation depends on the value Y through $f(R|Y; \phi) = \Phi(Y/2)$ where $\Phi(y)$ is the cumulative distribution function of a standard normal. Having from Fig. 2.4 it seems clear that making inference from this sample ignoring the missingness mechanism would not be correct as that would prevent one from recognizing even the most basic features of the distribution (e.g. the skewness).

Table 2.1 reports the sample mean and variance for the sample data described above. It is easy to notice as the simple method of moments inference made on the first subsample would lead to valid results. In the second case, the estimate of the mean is still valid and this is because the censoring interval is centered on the mean of a Gaussian, a symmetric distribution (in other cases the estimate would have been biased). For the second case, the MDM acts allowing the observation of only the tails of the distribution, resulting in a greater sample variance. In the last example, the MDM acts with an

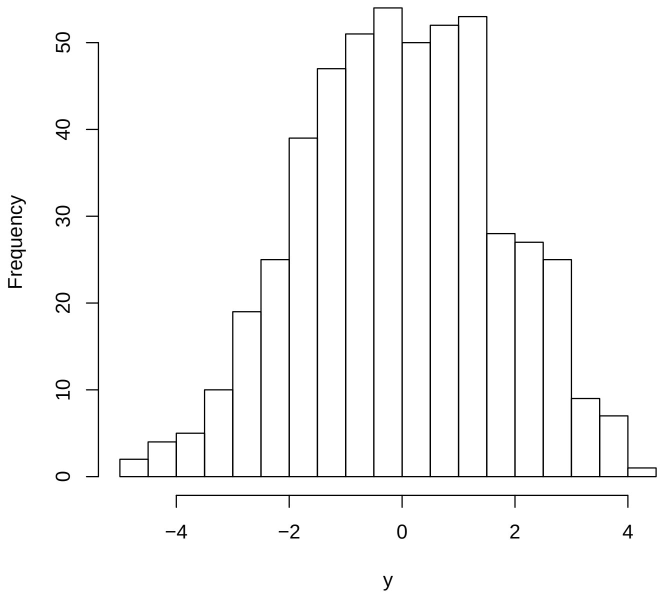

Figure 2.2: MDM examples: Random Selection (not variable dependent)

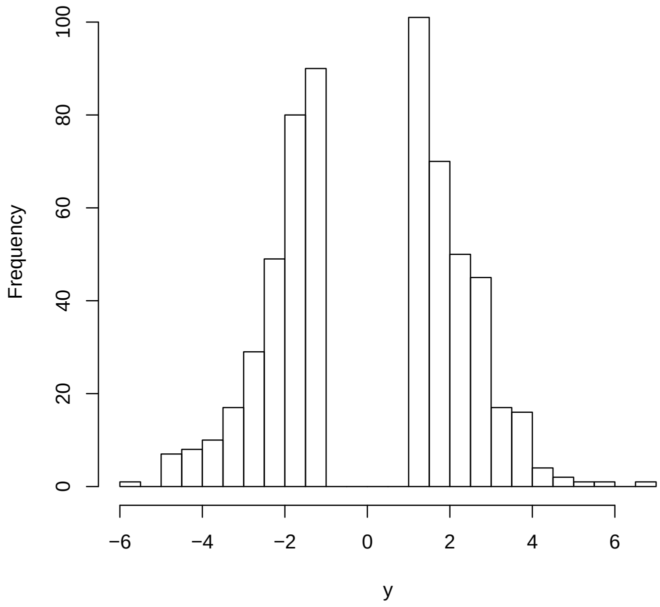

Figure 2.3: MDM examples: Deterministic Selection (variable dependent)

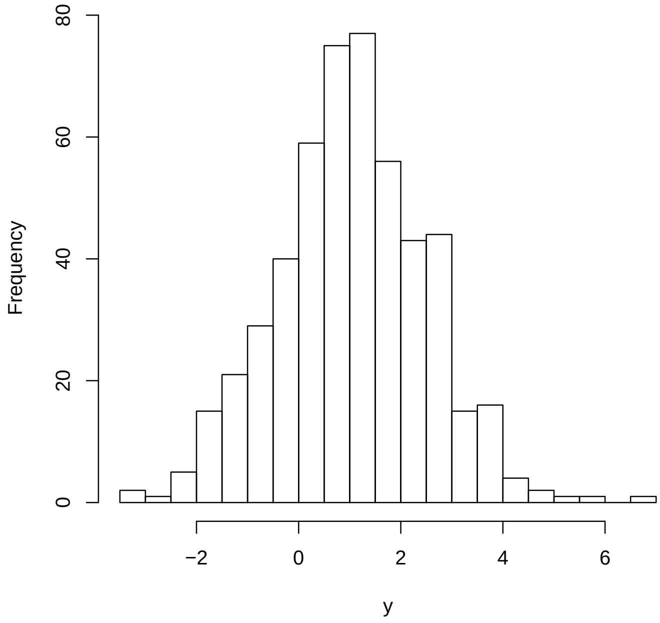

Figure 2.4: MDM examples: Random Selection (variable dependent)

increasing probability of observation as y grows, thus deflating the left tail of the observed distribution, and explaining the higher mean and the lower variance.

Let us now state the main definitions for missingness mechanism as introduced in [1]. Let θ be the parameter (real or vector valued) of the density $f(Y;\theta)$ of Y. Then

Definition 2.2.1. Missing data are said to be **missing at random** (MAR) if for all values of ϕ, the probability $f(R|Y;\phi)$ is the same for all possible values of Y_{mis}.

Definition 2.2.2. Observed data are said to be **observed at random** (OAR) if for all values of ϕ, the probability $f(R|Y;\phi)$ is the same for all possible values of Y_{obs}.

Definition 2.2.3. The parameters θ and ϕ are said to be **distinct** if the joint parameter space factorizes into a θ-space and a ϕ-space. In a Bayesian context this definition holds if the *prior* joint distribution factorizes into two independent distributions.

Consider the case when the MDM is completely random and is not influenced by any data, whether or not observed

$$f(R|Y;\phi) = f(R;\phi).$$

then, Y_{obs} are OAR and Y_{mis} are MAR. This situation is referred to in the literature as *missing completely at random* (MCAR), and observed data can be considered as a sample without replacement from the original data Y. When the MDM depends on the data and when it is influenced only by observed data, i.e. when

$$f(R|Y;\phi) = f(R|Y_{obs};\phi), \tag{2.1}$$

then the definition is the same as in [1] and this situation is simply called MAR. It is fundamental to state that (2.1) must hold only for the observed pattern of R and not for any of its possible values. In a likelihood setting, if the sample would be completely observed, than the possibility that some observations could have been missing would not alter inferences. The last and more problematic situation occurs when (2.1) does not hold and the MDM is allowed to depend on unobserved data: this case is referred as *missing-not-at-random* (MNAR). In this case the missingness of an observation is *informative* since it depends on the unobserved value itself. In this situation the MDM is defined *nonignorable.*

2.3 Traditional Approaches

In this section will be presented methods traditionally applied to dataset affected by missing data. These methods provide criteria for discarding missing observation, thus obtaining a complete subset of the original dataset. As it will turn out these techniques can be applied only when some conditions on the missing data mechanism hold, otherwise they could lead to stronlgy biased results. The first method presented is the *complete case analysis*, the easiest way to deal with missing data. Suppose we have a data matrix Y with n observations and k variables, and the $n \times k$ matrix R in which each element is an indicator defined as in Sec.2.1. In the complete case methods only observations observed for all variables are retained i.e., the set of the indices of the observations included in the analysis is

$$\Big\{i : \sum_{j=1}^{k} R_{ij} = k\Big\}.$$

Acting in this way would produce a loss of information, because some observed variables would be discarded just because some other variables are missing for the same observation. This can result in a large portion of the data being ignored. Consider for example a situation where $k = 10$ and a constant random selection acts independently on each variable with a 0.9 probability of being observed. Then, the expected proportion of complete cases is $0.9^{10} = 0.3487$. However, what really matters when analyzing data *selected* in this way is whether or not this technique biases the estimation process. In the MCAR case, the complete case sub sample is still a random sample of the observations and thus estimation is not biased.

In other situations, however, missingness could affect results and the bias would depend on the MDM. Suppose for example that $k = 2$, and the variables Y_1 and Y_2 are age and income respectively. Let the MDM be defined as follows

$$\begin{aligned} f(R_{i1} = 0, R_{i2} = 1|Y_1 = y_{i1}, Y_2 = y_{i2}) &= \phi_{01}(y_{i2}) \\ f(R_{i1} = 1, R_{i2} = 0|Y_1 = y_{i1}, Y_2 = y_{i2}) &= \phi_{10}(y_{i2}) \\ f(R_{i1} = 1, R_{i2} = 1|Y_1 = y_{i1}, Y_2 = y_{i2}) &= 1 - \phi_{01}(y_{i2}) - \phi_{10}(y_{i2}), \end{aligned} \tag{2.2}$$

where ϕ_{01} and ϕ_{10} are two generic function which take values in $[0, 1]$ depending on the income level. In this situation the missingness can affect both Y_1 and Y_2, but is influenced only by income. Indeed under-reporting is quite common for this variable. Suppose that ϕ_{01} and ϕ_{10} in (2.2) are such that middle incomes are more likely to be observed. In this case, the marginal distributions of age and income are distorted since the middle class is over-represented, and the correlation and regression coefficients of $Y_2|Y_1$ may be biased. Nevertheless, complete case analysis is still a widespread practice for its ease of implementation and for its use in statistical packages. Further references can be found in [7] where the inference on odds ratio in 2 × 2 counts tables is discussed, and in the more recent [8] biases on the estimates of the proportional hazards model are investigated under different MDMs.

One of the main problems with the complete case approach is the waste of information resulting from the restrictive discarding rule. In the *available case analysis* all the available observations are retained. For example, when the interest is on the marginal distributions for each variable all the observed values are considered. For example, the sample variance of the j^{th} variable Y_j would be calculated as

$$s_{jj}^{(j)} = \sum_{(j)} \frac{\left(y_{ij} - \bar{y}_j^{(j)}\right)^2}{n^{(j)} - 1}$$

where (j) is the set

$$(j) = \{i : R_{i,j} = 1\} \tag{2.3}$$

and using it as a superscript (as in $\bar{y}_j^{(j)}$) it indicates the statistic calculated on the observations corresponding to that set. The term $n^{(j)}$ indicates the cardinality of the set in (2.3). When estimation involves more than one variable, as for example for the covariance between Y_j and Y_k, are considered only the observations for which both variables are observed, i.e. the ones whose index belongs to the set

$$(jk) = \{i : R_{i,j} = R_{i,k} = 1\}.$$

In this case the sample covariance is

$$s_{jk}^{(jk)} = \sum_{(jk)} \frac{\left(y_{ij} - \bar{y}_j^{(jk)}\right)\left(y_{ij} - \bar{y}_k^{(jk)}\right)}{n^{(jk)} - 1}. \tag{2.4}$$

An alternative version of the sample covariance 2.4 that uses more observations in the sample means is the following

$$\tilde{s}_{jk}^{(jk)} = \sum_{(jk)} \frac{\left(y_{ij} - \bar{y}_j^{(j)}\right)\left(y_{ij} - \bar{y}_k^{(k)}\right)}{n^{(jk)} - 1} \tag{2.5}$$

and was suggested in [9]. Having defined sample covariances, it is possible to estimate the correlation coefficient

$$r_{jk}^{*} = \frac{s_{jk}^{(jk)}}{\sqrt{s_{jj}^{(j)} s_{kk}^{(k)}}}. \tag{2.6}$$

Even if all the available information is preserved by 2.6 the three statistics that appears in its definition are based on three possibly distinct subsets of data i.e., (jk), (j) and (k). For this reason (2.6) could not fall in the usual range $[-1, +1]$. To overcome this issue a modification of the (2.6) is available

$$r_{jk}^{(jk)} = \frac{s_{jk}^{(jk)}}{\sqrt{s_{jj}^{(jk)} s_{kk}^{(jk)}}}, \tag{2.7}$$

its properties are discussed in [10]. An alternative definition of the covariance is available as well

$$s^*_{jk} = r^{(jk)}_{jk} \sqrt{s^{(j)}_{jj} s^{(k)}_{kk}}. \tag{2.8}$$

With respect to the complete case method, the available case approach does not introduce great complications, and it has the appealing property of preserving a larger amount of information. However, only in the MCAR situation it produces unbiased results, and when correlations between variables are high it can result in poor performance, as shown in [11]. Another issue that has great practical impact is that none of the proposed definitions of covariance guarantees that the covariance matrix will be positive definite.

The last class of methods proposed in this section are the *weighting methods*. The methods described above require the often unrealistic MCAR property to hold, as removing some observations could introduce bias. To reduce this bias we may introduce weights that allow the observations retained after case deletion to represent those removed. This technique is closely related to weighting in sample surveys, where the weight of the i-th observation is π_i^{-1}, the inverse of the sample inclusion probability. In the case of missing values, the relevant inclusion probability is the probability of being observed and it has to be estimated from auxiliary variables, for example through a logit or a probit regression model. In [2] the case of missing data in sample surveys and related results are presented within the general framework of the Horwitz-Thompson estimation [12]. The estimator for the total T of a variable Y from a stratified random sample is

$$\hat{T} = \sum_{i=1}^{N} y_i I_i \hat{\pi}_i^{-1}$$

where I_i is the sample indicator function

$$I_i = \begin{cases} 1 & \text{if } Y_i \text{ is in the sample} \\ 0 & \text{if } Y_i \text{ is not in the sample} \end{cases}$$

and $\hat{\pi}_i = n_i/N_i$ is the estimating sampling probability for a unit that falls in the i-th stratum of the sample.

The following example will show how to use these estimators in presence of missing data. Suppose that it is possible to identify J groups in the population, whose individuals are indexed by $i = 1, \ldots, N$, henceforth called *adjustment cells*, within which the missingness variable R is (conditionally) independent of values taken by variable Y (not MCAR) and of the sampling mechanism described by I. Define C as the adjustment cell variable which takes value j for all units in the j-th cell (stratum), with $j = 1, 2, \ldots, J$. The missingness probability is then

$$f(R|I,Y,C) = \begin{cases} \prod_{j=1}^{J} \binom{N_j}{M_j}^{-1} & \sum_{i:c_i=j} R_i = M_j \text{ for all } j \\ 0 & \sum_{i:c_i=j} R_i \neq M_j \text{ for any } j \end{cases}$$

where N_j and M_j are the number of units and the number of respondents sampled inside the j-th cell. Having defined the response rate as $\phi_j = M_j/N_j$, if these values were known then the Horwitz-Thompson estimator could be easily obtained by using weights $(\pi_i\phi_i)^{-1}$. In practice, however, ϕ_j are replaced by their sample estimates $\hat{\phi}_j$, resulting in a mean estimate

$$\frac{\sum_{j=1}^{J} \sum_{i \in R(j)} y_i \hat{\pi}_i^{-1} \hat{\phi}_j{-1}}{\sum_{j=1}^{J} \sum_{i \in R(j)} \hat{\pi}_i^{-1} \hat{\phi}_j{-1}}. \tag{2.9}$$

A common choice for $\hat{\phi}_j$ is the sample estimate $\hat{\phi}_j = m_j/n_j$, so that (2.9) becomes

$$\bar{y}_{wc} = \frac{\sum_{j=1}^{J} n_j \bar{y}_{jR}}{n}$$

where $\bar{y}_{jR}$ is the sample mean for respondents in cell j. In [13] the mean and the variance of $\bar{y}_{wc}$ are derived. More recent developments on weighting techniques can be found in [14]. Interesting developments involving semiparametric estimators, based on inverse probability weighted estimating equations are presented in [15] and [16].

2.4 Imputation Methods

The methods presented so far require the partial or complete elimination of the observations which present missing data. Imputation techniques, on the other hand, replace missing values with one or more imputed values, so that one or more complete datasets are made available for the application of standard analysis techniques. The main drawback of imputation, as we will discuss below, is the risk of distorting the distribution of individual variables, and the relationships among variables.

2.4.1 Single Imputation

The first and simplest imputation method is *mean imputation*, which consists of the substitution of each missing value with the mean of the observed values for the same variable. This solution is widespread for its ease of implementation, and preserves the observed mean of the distribution. However, the dispersion of the distribution is severely altered: for each imputed value the sample variance is reduced, and quantiles become more concentrated. Using the notation introduced in the previous sections, in single imputation each missing value of variable Y_j is substituted by the observed mean value $\bar{y}_j$, and this produces the overall estimated sample mean for the imputed Y_j and the estimated sample variance of $s_{jj}^{(j)}[(n^{(j)}-1)/(n-1)]$. In case of imputation of two variables Y_j and Y_k, the completed sample covariance is $\tilde{s}_{jk}^{(jk)}[(n^{(jk)}-1)/(n-1)]$, where $\tilde{s}_{jk}^{(jk)}$ is as defined in (2.5). Under MCAR, s_{jj} and $\tilde{s}_{jk}^{(jk)}$ are consistent estimates of the variance and of the covariance and hence those statistics are underestimated by a factor $(n^j-1)/(n-1)$ for the former and a factor $(n^{jk}-1)/(n-1)$ for the latter. In this case the correction factors $(n-1)/(n^j-1)$ and $(n-1)/(n^{jk}-1)$ should be applied to the estimates (see [2]).

An extension of single imputation that allows to preserve better the dispersion of the distribution is called *hot deck imputation*. In this case, instead of replacing all missing values with the sample mean, each missing is replaced with a random pick from the observed values for that variable. However, this procedure leads to correct results only in the MCAR situation.

2.4.2 Multiple Imputation

Multiple imputation is a popular technique designed to tackle the missing data problem in many fields. Many references can be found on this subject, among them the early [17], [18] and the two monographs [19] and [20]. A more recent review on the subject is contained in [21]. Multiple imputation is intended for situations in which the database constructor and the ultimate users are distinct entities. It is a method intended for the following achievable basic objective: each analytical tool available to the ultimate user such as linear regression, logistic regression, factor analysis etc. should be readily applied to any dataset with or without missing data. Any method presented in the previous section such as available case and complete case satisfy this basic objective, and this is the reason why these methods are appealing and widely used. However, this objective is not the only relevant one as another critical goal, discussed in [19], is to produce *statistically valid* answers for *scientific estimands* and this is not the case for the previous methods.

A *scientific estimand* is intended to be a quantity (calculated on a population) of scientific interest, that can be calculated and that is not influenced by the data collection process that is used to measure it. To clarify, assume that X is the set of predictor variables and are fully observed, and Y the set of variables of interest, then a scientific estimand is a function $Q = Q(X, Y)$. Following this broad definition, many common statistics are scientific estimands (e.g. population mean and variance, regression coefficients, factor loadings etc.), but for example the sampling variance of a sample mean is not a scientific estimand since this quantity depends upon the data collection method, i.e. the sampling technique. Even the expectation of the complete data sample mean cannot in general be considered a scientific estimand. *Statistically valid* in Rubin's opinion has to be a frequentist concept based upon the idea of averaging over randomization distributions generated by the sampling mechanism, and over distributions arising from the missing data mechanism. To have statistical validity, point estimation of scientific estimands has to be approximately unbiased, whilst when dealing with interval estimation and hypothesis testing, the correct nominal levels should be achieved. In particular, *Randomization* *validity* means that for interval estimates, the actual interval coverage equals the nominal interval coverage. For hypothesis testing the actual rejection rate must equal the nominal rejection rate. Randomization validity is a standard requirement in most survey applications. In such cases, typically a complete data estimate $\hat{Q}$ of an estimand Q follows a normal distribution centered at Q with a sampling variance estimated consistently (and possibly unbiasedly) by another statistic U; the randomization distribution is generated by the sampling indicator I, which is in general a function of (X, Y). Then

$$\mathrm{E}(\hat{Q}|X, Y) = Q$$

and

$$\mathrm{E}(U|X, Y) = \mathrm{Var}\,(\hat{Q}|X, Y). \tag{2.10}$$

U^{-1} is known as the precision of $\hat{Q}$. A more general objective, however, is the so called *confidence validity*, which requires less restrictive conditions: for interval estimates the actual interval coverage should to be greater than or equal to the nominal coverage, and for hypothesis testing the actual rejection rate should be lower than or equal to the nominal rejection rate. In this case (2.10) could be replaced by the requirement

$$\mathrm{E}(U|X, Y) \geq \mathrm{Var}\,(\hat{Q}|X, Y). \tag{2.11}$$

The distinction between randomization validity and confidence validity is important when dealing with approximate procedures, typically applied to non-response in surveys as described in [22]. The definitions of confidence intervals, confidence coefficients and confidence limits, still remain a pillar in the foundations of mathematical statistics [23].

A fundamental form of multiple imputation is called *repeated imputation*, which is a very intuitive approach. The main results related to this method can be found in [24]. Having defined a particular Bayesian model for the data generating process and for the missing data mechanism, repeated imputations are draws from the *posterior predictive distribution* of the missing values. If we produce m samples for the set of missing values, it is then possible to produce m complete data sets on which m complete data analysis can be performed, leading to m statistics $\hat{Q}_{*1}, \ldots, \hat{Q}_{*m}$ and the associated variance estimates $U_{*1}, \ldots, U_{*m}$. These m statistics can be combined into one, and the so-called repeated-imputation inference then appropriately adjusts the estimate under the postulated non-response mechanism. Let us define the related distribution of Q as $\mathrm{P}\left(Q|Y_{obs}, Y_{mis}\right)$ and the predictive posterior distribution for the missing data as $\mathrm{P}\left(Y_{mis}|Y_{obs}\right)$. The posterior distribution of the statistics Q is

$$\mathrm{P}\left(Q|Y_{obs}\right) = \int \mathrm{P}\left(Q|Y_{obs}, Y_{mis}\right) \mathrm{P}\left(Y_{mis}|Y_{obs}\right) \mathrm{d}Y_{mis}. \tag{2.12}$$

We can thus obtain an estimate of the posterior distribution of Q from samples of imputed values generated from the posterior predictive distribution $\mathrm{P}\left(Y_{mis}|Y_{obs}\right)$,by averaging the distribution of the statistics Q over the imputed values. Two main results follow [24], the posterior mean of Q is

$$\mathrm{E}(Q|Y_{obs}) = \mathrm{E}[\mathrm{E}(Q|Y_{obs}, Y_{mis})|Y_{obs}] \tag{2.13}$$

while the posterior variance of Q is

$$\begin{aligned} \mathrm{Var}\left(Q|Y_{obs}\right) = {} & \mathrm{E}[\mathrm{Var}\left(Q|Y_{obs}, Y_{mis}\right)|Y_{obs}] \\ & + \mathrm{Var}\left[\mathrm{E}(Q|Y_{obs}, Y_{mis})|Y_{obs}\right]. \end{aligned} \tag{2.14}$$

Then we combines the m complete datasets to obtain the repeated imputation estimate as

$$\bar{Q}_m = \frac{\sum_{i=1}^{m} \hat{Q}_{*i}}{m} \tag{2.15}$$

and its variance

$$T_m = \bar{U}_m + \frac{m+1}{m} B_m, \tag{2.16}$$

where

$$\bar{U}_m = \frac{\sum_{i=1}^{m} U_{*i}}{m} \tag{2.17}$$

is the *within* imputation variability and

$$B_m = \frac{\sum_{i=1}^{m} (\hat{Q}_{*i} - \bar{Q}_m)^2}{m-1} \tag{2.18}$$

as m tends to infinity, the standardized random variable Q converges in distribution to a standard normal variable:

$$\frac{Q - \bar{Q}_m}{T_m} \overset{\mathrm{d}}{\rightarrow} N(0,1) \tag{2.19}$$

(see [21]).

2.5 Example: Regionalization of Public Investment Expenditure

2.5.1 Objectives

The Regional Public Accounts System (Sistema dei Conti Pubblici Territoriali) is the result of a project started in the 1994 and conducted by the Economic Development Department of the Ministry for the Economy and Finance. The main objectives of the project were the following:

- identification of the financial flows involving all government entities in the individual regional areas, with the maximum institutional and territorial detail that the accounting documentation allows;

- reconstruction of the consolidated accounts for total expenditures (current and capital) in the public sector in the twenty Italian regions;

- reconstruction of revenue flows and territorial financial balances, thereby achieving, inter alia, the following additional objectives:
  - contribute to the verification of compliance with the principle enunciated in Article 119(5) of the Italian Constitution (”In order to promote economic development, cohesion and social solidarity, to remove economic and social imbalances, to foster the effective exercise of the rights of the person or to achieve purposes other than the normal performance of their functions, the State shall provide additional resources and undertake special actions in favour of certain municipalities, provinces, metropolitan cities and regions”) by providing information to assess whether and to what extent the expenditure financed with additional EU funds (Structural Funds and national co-financing) and additional national funds (funds for underdeveloped areas) should be considered to be *additional* with respect to ordinary spending;
  - satisfy the European Community rules regarding compliance with the principle of additionality (Article 11 of Regulation (EC) 1260/99);
  - measure and analyze the allocation of public expenditure between the South and Centre-North areas of the country and among individual regions;
  - measure and analyze the composition of public capital expenditure between investments and transfers;
  - measure and analyze the sectoral composition of capital expenditure for investment and provide a reference base for analyzing the effectiveness of that expenditure.

2.5.2 Data Dimensions

The geographical reference universe for the Regional Public Accounts (RPA) consists of the 19 regions and the autonomous provinces of Trento and Bolzano. Each territorial area is identified using the standard ISTAT (Italian National Institute for Statistics) codes in order to facilitate consultation and ensure comparability with other databases. The regional detail of the RPA database is a unique information resource within the Italian statistical system. One of the most complex aspects of constructing the RPAs was the development of sound criteria for the regional allocation of data that, for certain entities, is often available only at the national level. The most common approach is the aggregation of the regions into the five macro-areas adopted by ISTAT:

- **North-west**: Piedmont, Val d'Aosta, Lombardy and Liguria;
- **North-east**: Autonomous Province of Trento, Autonomous Province of Bolzano, Veneto, Friuli Venezia Giulia and Emilia Romagna;
- **Centre**: Tuscany, Umbria, Marche and Lazio;
- **South**: Abruzzo, Molise, Campania, Puglia, Basilicata and Calabria;
- **Islands**: Sicily and Sardinia

Consistently with the classification system adopted for the national accounts and, therefore, with the classification of the functions of government (COFOG), the Regional Public Accounts are constructed on the basis of a 30 main sectors (e.g. education, welfare etc.). This is especially useful in meeting the needs of European Community for programming and analyzing public expenditure. Data are collected on a yearly basis since 1996 to 2007.

2.5.3 The public capital expenditure indicator

The RPA database now provides annual data with a lag of about 12-18 months due to the activation of the performance reserve mechanism. Since 2003, the RPA Project has developed (for general government capital expenditure only) a leading indicator, which initially provided regionalised estimates for the reference year subsequently reached by the RPAs for the various categories of expenditure. However, in order to analyse current economic developments and have a stable and effective monitoring tool to support public investment decisions, even more timely information is necessary. Accordingly, the scope of application of the indicator has now been expanded, with the generation of interim expenditure estimates (quarterly and annual estimates produced during the year) and the production of annual expenditure forecasts for the year following the current year. The indicator follows a bottom-up approach, obtaining the result for the broadest level by aggregating the data from narrower levels. The indicator of total expenditure on capital account for General Government as a whole is therefore obtained by aggregating indicators for the individual expenditure items (investments, transfers to enterprises, transfers to households) and the individual segments of General Government (Regions, Provinces, Municipalities, the State, ANAS, etc.).

2.5.4 A Proposal for an Imputation Approach

Within the setting of multiple imputation a great attention has been put on *categorical data*. A reference for this area is [25], while a specific treatment of missing data problems in this framework can be found in [26].

Let $Y_1, Y_2, \dots, Y_p$ be p categorical variables. Without loss of generality, each of these variables can be thought of taking a finite number of levels, that conventionally will be represented by the natural $1, 2, \dots, d_i$, for $i = 1, 2, ..., p$. Sampling n units for each variable then results in an $n \times p$ data matrix that can be reduced to a contingency table with $D = \prod_{i=1}^{p} d_i$ cells. In some cases some of these cells could be constrained to be zero since the combination of levels that they represent does not make sense: think for example of the combination age < 5 years and smoker $=$ TRUE. These cases are called *structural zeroes*, and they are handled simply by excluding such cells from analysis. Reducing to a D cells contingency table results in D frequency counts that can be represented by the vector $x = (x_1, x_2, \dots, x_D)$. By construction, $n = \sum_{i=1}^{D} x_i$, and any cell count can be expressed as the complement to n of the sum of all the other counts. In particular one has that $x_D = n - \sum_{i=1}^{D-1} x_i$. Having defined the probability that each observation be included in the D cells as the vector $\theta = (\theta_1, \theta_2, \ldots, \theta_D)$, a natural way to define the joint probability of the observed table is by $x|\theta \sim Multinomial(n|\theta)$, so that

$$\mathrm{P}\left(x|\theta\right) = \frac{n!}{x_1!x_2!\ldots x_D!}\theta_1^{x_1}\theta_2^{x_2}\ldots\theta_D^{x_D}. \tag{2.20}$$

The parameter space is represented by the simplex

$$\Theta = \{\theta : \theta_i \geq 0 \text{ for all } i \text{ , } \sum_{i=1}^{D}\theta_i = 1\}$$

since for the θ vector each element can be rewritten as the complement to 1 of the sum of the other elements. In the Bayesian context, a *prior distribution* which is typically used for the vector of probabilities θ is the Dirichlet density

$$\mathrm{P}\left(\theta|\alpha\right) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\ldots\Gamma(\alpha_D)}\theta_1^{\alpha_1-1}\theta_2^{\alpha_2-1}\ldots\theta_D^{\alpha_D-1}, \tag{2.21}$$

where the vector $\alpha = (\alpha_0, \alpha_1, \ldots, \alpha_D)$ the vector of the *hyper-parameters*, with $\alpha_0 = \sum_{i=1}^{D}\alpha_i$.

Some methodological suggestions within this framework could help tackle the problem of allocating State capital expenditure on a regional basis [27]. Indeed, there is a lack of reliable information on the territorial location of each expenditure operation registered in the payment orders database SIRGS (State General Accounting Department Information System). This database contains data on the *orders to pay* (OP) and on the *orders to credit* (OA). For the first kind of expenditure operations, the variable RZ *zone of intervention* is available and identifies in a reliable way the region of interest, but such information is missing in more than 50% of the cases. This can be considered as the first variable of interest

$$Y_1 = RZ. \tag{2.22}$$

For the regional imputation of OP's some proxy variables are available:

- the region identified by the treasury section issuing the order ($Y_2 = RT$);

- the region identified by the tax code of the enterprise, or the person from whom the expenditure was commissioned ($Y_3 = RC$);
- the region identified by the bank branch code of the enterprise or person receiving the payment ($Y_4 = RA$);
- the region extracted from the description string of the beneficiaries themselves ($Y_5 = RB$);

The first variable is partially observed, so we use the indicator variable

$$R = \begin{cases} 1 & \text{when } Y_1 \text{ is observed} \\ 0 & \text{when } Y_1 \text{ is missing.} \end{cases}$$

Given this indicator, it is possible to partition the n dimensional sampled vector for each variable as the union of the observed part and the missing part. Y_{obs} is the set of elements of the Y_1 vector for which $R = 1$ (for which the zone of intervention was available), while Y_{mis} contains the remaining elements of Y_1. We let X be the matrix composed by the elements of the Y_2, Y_3, Y_4, Y_5 that can be used as a proxy for the cases in which the location is not available (i.e. when $R = 0$). Eq.(2.12) thus applies as

$$\mathrm{P}\left(Q|Y_{obs}, X\right) = \int \mathrm{P}\left(Q|Y_{obs}, Y_{mis}\right)\mathrm{P}\left(Y_{mis}|X\right) \mathrm{d}Y_{mis}. \tag{2.23}$$

The predictive probability $\mathrm{P}\left(Y_{mis}|X\right)$ can be easily estimated in a regression setting such as the *Multinomial Probit Regression* [28]. Within this framework the predictive probability takes the general form

$$\mathrm{P}\left(Y_{mis}|X\right) = H(\sum_{i=1}^{k} \beta_i X_i) \tag{2.24}$$

being H a known cumulative density function linking the linear predictor $\sum_{i=1}^{k} \beta_i X_i$ to the probability. In particular for the probit model H is the standard Gaussian cdf. In the next chapter will be shown how to include explicitly the missing data mechanism into the likelihood function.

Chapter 3

Complete Inference Approaches

One of the most important advantages of imputation methods is the ease of the subsequent analyses, which could be based on *traditional* approaches and models. When direct inference has to be performed on missing data, on the other hand, a lot of care is required to make sure that sampling distribution inference or direct likelihood inference results are valid. In this chapter the main results on likelihood inference in presence of missing data are presented, justifying then the model formulated in the following chapter, for the analysis of the Monza and Brianza firms data.

3.1 Ignorable Missing Data Mechanism

A relevant issue is whether or not we can safely ignore the missing data mechanism (MDM). In most of the cases MDM is unknown, and is then preferable not to consider it whenever possible. We refer to the notation introduced in [2]. By ignoring the MDM we mean that (i) the random variable R is considered as fixed at the observed pattern of observed-missing data in the sample; and (ii) the observed data y_{obs} are assumed to come from the marginal density of Y_{obs}, which can be written as

$$f(Y_{obs};\theta) = \int f(Y;\theta) dY_{mis}. \tag{3.1}$$

3.1.1 Sampling Distribution Inference

Sampling distribution inference, together with likelihood inference and Bayesian is one of the three kinds of inference discussed in [1]. In [29] we can find a review of these approaches, and in particular a discussion of the fact that the second and the third approaches are in general to be preferred to the first one and in some sense this will be the case even in presence of missing data. Sampling distribution inference is an approach to inference in which, taking the parameters θ and ϕ as fixed, the observed value of a statistic is compared with its sampling distribution under several hypothesized distribution. This is for the example the case of the classical confidence interval testing approach.

The critical assumption when ignoring the MDM is to consider Y_{obs} as following the distribution in (3.1), while the conditional distribution given the missingness state is

$$\begin{aligned} f(Y_{obs}|R=r;\theta,\phi) =& \frac{\int f(Y,R=r;\theta,\phi)dY_{mis}}{k_{\theta,\phi}(r)} \\ =& \frac{\int f(Y;\theta)f(R=r|Y;\phi)dY_{mis}}{k_{\theta,\phi}(r)}, \end{aligned} \tag{3.2}$$

with $k_{\theta,\phi}(r)$ the marginal distribution for the observed R:

$$k_{\theta,\phi}(r) = \mathrm{P}\left(R=r;\theta,\phi\right) = \int f(Y,R=r;\theta,\phi)dY$$

A complete data sampling distribution inference about the true value of θ would compare the observed value of a statistic $S(y)$ to its distribution as derived from $f(Y;\theta)$. Since we have missing data, the statistic of interest can only be based on the observed data $S(y_{obs},r)$, and ignoring the MDM would mean to fix r and assume that the sampling distribution follows from (3.1) instead of the correct (3.2), which takes into account the specified model $f(R|Y;\phi)$. In some cases, however, the MDM can safely be ignored. Consider the following

Theorem 3.1.1. *Suppose that*

1. *Missing data are* missing at random

2. Observed data are observed at random

being hence in the MCAR situation, then the sampling distribution of $S(Y_{obs}, r)$ under the density $f(Y;\theta)$ ignoring the missing data mechanism i.e. calculated from (3.1) equals the sampling distribution under the density $f(Y;\theta)f(R|Y;\phi)$ i.e. calculated from (3.2), provided that $k_{\theta,\phi}(r) > 0$.

The proof of this and the following two theorems can be found in [1]. In this case the results follows straightforwardly from the fact that when data are MCAR, then for each value of ϕ the probability distribution $f(R|Y;\phi)$ takes the same value for any Y. Another important result is the following

Theorem 3.1.2. *The sampling distribution of $S(Y_{obs}, r)$ under $f(Y;\theta)$ ignoring the missing data mechanism equals the correct sampling distribution under $f(Y;\theta)f(R|Y;\phi)$ given the observed missingness pattern r if and only if*

$$\mathrm{E}_{y_{mis}}\{f(R|Y;\phi)|r, y_{obs}, \theta, \phi\} = k_{\theta,\phi}(r). \tag{3.3}$$

This theorem essentially states that the ignorability of the MDM in sampling distribution inference is equivalent to requiring that the expected value of the probability function of R conditionally on Y and computed with respect to all possible unobserved values Y_{mis}, is equal to the marginal distribution of R.

The last necessary and sufficient condition for the ignorability of the missing data mechanism is given by the following

Theorem 3.1.3. *The sampling distribution of $S(Y_{obs}, r)$ under $f(Y;\theta)$ ignoring the missing data mechanism equals the correct sampling distribution under $f(Y;\theta)f(R|Y;\phi)$ given the observed pattern r if and only if*

$$f(R|Y;\phi) = 1. \tag{3.4}$$

therefore using (3.1) instead of (3.2) for sampling distribution inference is equivalent requiring that deterministic missing data mechanism applies.

3.1.2 Direct Likelihood Inference

Direct likelihood inference is an inferential approach in which the only relevant evidence follows from ratios of likelihood functions for different values of the parameters [30]. Likelihood principle found its earlier applications in [31], and was formally defined in [32]. For this kind of inference sufficient and necessary conditions for the ignorability of MDM are described in [1]. These are less restrictive compared to the ones which are required for sampling distribution inference.

In this case, ignoring the missing data is equivalent to considering for inference the observed marginal likelihood function

$$L(\theta|Y_{obs}) = \mathbb{1}(\theta \in \Theta) \int f(Y;\theta) dY_{mis}, \tag{3.5}$$

hence ignoring the information conveyed by the variable R. The correct likelihood function to be used in this case should be

$$L(\theta,\phi|Y_{obs},R) = \mathbb{1}((\theta,\phi) \in \Omega_{\theta,\phi}), \int f(Y;\theta) f(R|Y;\phi) dY_{mis} \tag{3.6}$$

where $\Omega_{\theta,\phi}$ is the joint parameter space of (θ, ϕ).

Theorem 3.1.4. *Assume that:*

1. *Missing data are missing at random*

2. *θ is distinct from ϕ*

then the missing data mechanism can be ignored i.e. the likelihood ratio $L(\theta_1|Y_{obs})/L(\theta_2|Y_{obs})$ equals the correct ratio $L(\theta_1,\phi|Y_{obs},R)/L(\theta_2,\phi|Y_{obs},R)$ for all $\phi \in \Phi$ such that $f(R|Y;\phi) > 0$.

This theorem, analogous to theorem 3.1.1 for sampling distribution inference, poses less restrictive requirements for MDM ignorability. Observed data do not have to be observed at random anymore, the more common MAR situation will suffice, provided that parameters *distinctness* holds.

A different condition for ignorability can be obtained, involving the expected value taken over the all possible values of Y_{mis} of the probability function defining the MDM. The following result is the counterpart for direct likelihood methods of theorem 3.1.2

Theorem 3.1.5. *Assume that* $L(\theta|Y_{obs}, R) > 0$ *for all* $\theta \in \Theta$*. Then any likelihood ratio for all* $\theta \in \Theta$ *ignoring the missing data mechanism provides correct inference if and only if:*

1. $\Omega_{\theta,\phi} = \Theta \times \Phi$ *(parameters distinctness)*

2. for each $\phi \in \Phi$*,* $\mathrm{E}_{y_{mis}}\{f(R|Y;\phi)|r, y_{obs}, \theta, \phi\}$ *takes the same positive value for all* $\theta \in \Theta$*.*

The second condition in theorem 3.1.5 means that for any value of ϕ, the expected value of $f(R|Y;\phi)$ given the observed data y_{obs} and r, taken over Y_{mis}, does not change. This means that missing data are not relevant ”on average” for the MDM. Proofs of the theorems in this and the previous section can be found in the original paper [1].

However, following the approach of [2] we can easily show sufficiency for theorem 3.1.4. Indeed, recalling from (3.6) that

$$f(Y_{obs}, R|\theta, \phi) = \int f(Y;\theta)f(R|Y;\phi)dY_{mis},$$

by MAR assumption we have that $f(R|Y;\phi) = f(R|Y_{obs};\phi)$. Then,

$$f(Y_{obs}, R|\theta, \phi) = f(R|Y;\phi)\int f(Y;\theta)dY_{mis} = f(R|Y;\phi)f(Y_{obs};\theta),$$

and assuming then parameters distinctness we have that ϕ is not a function of θ producing the final result

$$L(\theta, \phi | Y_{obs}, R) \propto f(R|Y; \phi) f(Y_{obs}; \theta) \propto f(Y_{obs}; \theta) \propto L(\theta | Y_{obs}). \tag{3.7}$$

Let us now show an example which illustrates this point.

Example 3.1.1. Suppose that n subsequent lifetimes X_i's are collected but that after a total observation time T is reached, any remaining X_i's are not observed, thus becoming missing. This is an example for which observed data is not observed at random and missing data is missing at random, being the MDM influenced only by the first (random) $k = \max\{j : \sum_{i=1}^{j} x_i < T\}$ observed x_i's. To illustrate this example, consider a sample of dimension 100 generated from the gamma distribution

$$f(x; \alpha, \lambda) = \frac{\lambda^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x},\ x > 0 \tag{3.8}$$

where α is taken to be equal to 20 and λ is equal to 2. We set T equal to 800 resulting in a 78% proportion of observed data in the generated data, close to the theoretical 80% proportion (being $\mathrm{E}(X) = 10$). The missing data mechanism can be easily written as

$$f(r|x) = \prod_{i=1}^{k} r_i \prod_{i=k}^{n} (1 - r_i)$$

Two considerations should be made here.

(i) the model (3.1.1) which describes the MDM depends on data only through k, which is a function of the observed data only

(ii) since there is no presence of the parameter ϕ in (3.1.1), parameter distinctness clearly holds.

By theorem 3.1.4, the missing data mechanism can be ignored, and inference lead by the observed likelihood (3.7) are valid. In this case the likelihood function is

$$L(\alpha, \lambda; \mathbf{x}) = \prod_{i=1}^{k} \frac{\lambda^{\alpha}}{\Gamma(\alpha)} x_i^{\alpha-1} e^{-\lambda x_i}$$

which when maximized leads to the parameter estimates $\hat{\alpha} = 23.72$ and $\hat{\lambda} = 2.38$.

This example presents *optional stopping* problem from a missing data likelihood inference perspective. What is relevant in this case for inference, is the fact that the likelihood can be based on the first k observations, the remaining $n-k$ observations can then be discarded.

3.2 Non Ignorable Missing Data Mechanism

In the previous section we described conditions for the ignorability of the missing data mechanism. In many real cases, however, missing data are not missing at random, and they require a precise definition of the MDM. Whenever conditions presented in the previous sections do not hold using the direct likelihood may produce incorrect results. Consider the Example 3.1.1: the MDM depended only on observed data since the missing data were missing at random. Consider now the following example

Example 3.2.1. Let n lifetimes $X_1, X_2, \ldots, X_n$ be observed. For some reason each lifetime x_i is reported (i.e. $r_i = 1$) only if its value is not too large, say if it is not above a known threshold ω. This is a case of the X's being *missing-not-at-random.* Suppose for example that the X_i's are gamma distributed having density (3.8). What is different here is the fact that the missing data mechanism that will depend on both the observed and on the missing data:

$$f(r|x) = \begin{cases} 1 & \text{if } (r = 1 \text{ and } x \le \omega) \text{ or } (r = 0 \text{ and } x > \omega) \\ 0 & \text{otherwise.} \end{cases}$$

The likelihood function for this problem is

$$\begin{aligned} L(\alpha, \lambda; \mathbf{x}) = & \prod_{i \in \mathcal{O}} \left[\frac{\lambda^\alpha}{\Gamma(\alpha)} x_i^{\alpha-1} e^{-\lambda x_i} \mathbb{1}(x_i \in [0, \omega]) \right] \times \\ & \prod_{i \in \mathcal{M}} \left[\frac{\lambda^\alpha}{\Gamma(\alpha)} x_i^{\alpha-1} e^{-\lambda x_i} \mathbb{1}(x_i \in (\omega, +\infty)) \right]. \end{aligned}$$

Since the likelihood function contains unobserved values, to produce valid inferences we need to integrate out the missing data X_{mis}:

$$
\begin{aligned}
L(\alpha,\lambda;\mathbf{x_{obs}}) =& \int \Bigg\{ \prod_{i\in\mathcal{O}} \left[\frac{\lambda^\alpha}{\Gamma(\alpha)} x_i^{\alpha-1} e^{-\lambda x_i} \mathbb{1}(x_i \in [0,\omega]) \right] \times \\
& \prod_{i\in\mathcal{M}} \left[\frac{\lambda^\alpha}{\Gamma(\alpha)} x_i^{\alpha-1} e^{-\lambda x_i} \mathbb{1}(x_i \in (\omega,+\infty)) \right] \Bigg\} d\mathbf{x_{mis}} \\
=& \prod_{i\in\mathcal{O}} \left[\frac{\lambda^\alpha}{\Gamma(\alpha)} x_i^{\alpha-1} e^{-\lambda x_i} \mathbb{1}(x_i \in [0,\omega]) \right] \times \\
& \prod_{i\in\mathcal{M}} \left[\int \frac{\lambda^\alpha}{\Gamma(\alpha)} x_i^{\alpha-1} e^{-\lambda x_i} \mathbb{1}(x_i \in (\omega,+\infty)) dx_i \right] \\
=& \prod_{i\in\mathcal{O}} \left[\frac{\lambda^\alpha}{\Gamma(\alpha)} x_i^{\alpha-1} e^{-\lambda x_i} \mathbb{1}(x_i \in [0,\omega]) \right] \left[\frac{\Gamma(\alpha,\lambda\omega)}{\Gamma(\alpha)} \right]^{\#\mathcal{M}}
\end{aligned}
$$

where

$$
\Gamma(\alpha,\omega) = \int_\omega^\infty t^{\alpha-1} e^{-t} dt
$$

the *upper incomplete gamma function.*

The log-likelihood function for this problem easily follows:

$$
\begin{aligned}
l(\alpha,\lambda;\mathbf{x_{obs}}) =& \log \left[\prod_{i\in\mathcal{O}} \left[\frac{\lambda^\alpha}{\Gamma(\alpha)} x_i^{\alpha-1} e^{-\lambda x_i} \mathbb{1}(x_i \in [0,\omega]) \right] \left[\frac{\Gamma(\alpha,\lambda\omega)}{\Gamma(\alpha)} \right]^{\#\mathcal{M}} \right] \\
=& \sum_{i\in\mathcal{O}} \left[\alpha \log\lambda - \log\Gamma(\alpha) + (\alpha-1)\log x_i - \lambda x_i \right] + \\
& \#\mathcal{M} \left[\log\Gamma(\alpha,\lambda\omega) - \log\Gamma(\alpha) \right].
\end{aligned}
$$

Taking the partial derivative with respect to α will produce:

$$
\begin{aligned}
\frac{\partial}{\partial\alpha} l(\mathbf{x_{obs}};\alpha,\lambda) =& \sum_{i\in\mathcal{O}} \left[\log\lambda - \psi^{(0)}(\alpha) + \log x_i \right] + \\
& \#\mathcal{M} \left[\frac{\log(\lambda\omega)\Gamma(\alpha,\lambda\omega) + \omega T(3,\alpha,\lambda\omega)}{\Gamma(\alpha,\lambda\omega)} - \psi^{(0)}(\alpha) \right],
\end{aligned}
$$

with

$$\frac{\partial}{\partial\alpha}\Gamma(\alpha,x) = \log x\Gamma(\alpha,x) + xT(3,\alpha,x)$$

and $\psi^{(0)}(\alpha)$ the *digamma function* (the derivative of the logarithm of the gamma function) and $T(k,\alpha,x)$ a special case of the *Meijer G-function* ([33]). The likelihood equation obtained differentiating with respect to the other parameter λ is:

$$\frac{\partial}{\partial\lambda}l(\alpha,\lambda;\mathbf{x_{obs}}) = \sum_{i\in\mathcal{O}}\left[\frac{\alpha}{\lambda} - x_i\right] - (\#\mathcal{M})\omega\frac{(\lambda\omega)^{\alpha-1}/\mathrm{e}^{(\lambda\omega)}}{\Gamma(\alpha,\lambda\omega)}$$

To illustrate the importance of considering the missing data mechanism in estimation, consider an $n = 1000$ sample from the $Gamma(10,2)$ distribution. From such a generated dataset and taking a threshold equal to 6 we obtain a proportion of observed data equal to 75.9% (the number of observed is $\#\mathcal{O} = 759$). As a first step we estimate the parameters without taking into account the MDM, optimizing the log-likelihood of the complete observations. The estimates for the parameters were $\hat{\alpha} = 16.22$ and $\hat{\lambda} = 3.78$. Since none of the observations was greater than 6, the fitted distribution assigns only a low probability mass to the $(6,+\infty)$ half line. A more satisfying result was achieved by optimizing the log-likelihood explicitly including the MDM. The estimated parameters obtained were $\hat{\alpha} = 9.75$ and $\hat{\lambda} = 1.95$, are much closer to the true ones. Note that in this case deviance is a misleading indicator, since in the first case it is equal to 2220.11 while in the second case it is 3150.27. This is because the first deviance is the result of an estimation on a wrongly reduced dataset. In Fig. 3.1 the true and the estimated densities are reported.

In the following chapter the complete inference approach presented here will be applied to formulate a model in which birth time and death time are the variables interest. Such a model will be designed to fit Monza and Brianza data and the assumed missing data pattern.

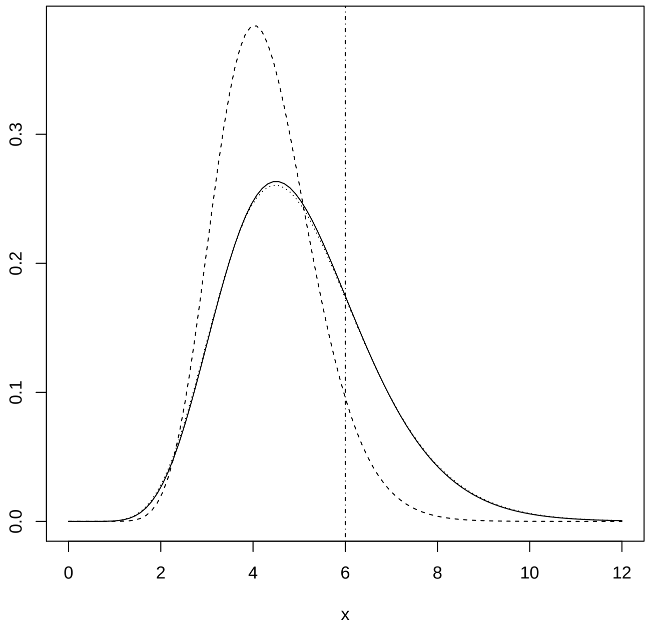

Figure 3.1: the true density (solid), the density estimated not considering the MDM (dashed) and the one considering the MDM (dotted)

Chapter 4

A Birth and Death Model with Missing Data

Following the results on the complete inference approach presented in the previous chapter, is now possible to deal with situations in which the missing data mechanism is not ignorable. Having assumed a form for the missing data mechanism, based on the empirical evidence, will be possible to define a model for the Monza and Brianza birth and death latent process. This model will turn out to be satisfactory from a likelihood inference point of view, but still lacking to consider the sampling mechanism i.e., the time window on the observed deaths. This limit will be finally overcome by the model proposed in the following chapter.

4.1 Model Definition

Suppose to be interested in estimating the length of life of a set of individuals for which the death time is observed. For some individuals the birth time is missing, and the probability of being missing is connected with the birth date itself. For example, more recent births have an higher probability of being missing.

Let us assume that birth times X_i's are distributed according to a $Gamma(\alpha, \beta)$ distribution, life lengths $(Y_i - X_i)$ follow a $Gamma(\gamma, \delta)$ distribution, with the Y_i's representing the death times (equivalently Y_i is a translated Gamma random variable). The probability that birth times are observed decreases monotonically and smoothly, and can be approximated by an exponential decay function with parameter ρ, $f(r=1|x) = \mathrm{e}^{-\rho x}$. For this model, the complete likelihood, as usual conditioned on the observed missing data pattern, is

$$\begin{aligned} L(\alpha,\beta,\gamma,\delta,\rho;\mathbf{x},\mathbf{y}) = \prod_{i\in\mathcal{O}} \Bigg[& \frac{\beta^\alpha}{\Gamma(\alpha)} x_i^{\alpha-1} \mathrm{e}^{-\beta x_i} \frac{\delta^\gamma}{\Gamma(\gamma)} (y_i - x_i)^{\gamma-1} \mathrm{e}^{-\delta(y_i - x_i)} \\ & \mathrm{e}^{-\rho x_i} \mathbb{1}(x_i \in [0, y_i)) \Bigg] \times \\ \prod_{i\in\mathcal{M}} \Bigg[& \frac{\beta^\alpha}{\Gamma(\alpha)} x_i^{\alpha-1} \mathrm{e}^{-\beta x_i} \frac{\delta^\gamma}{\Gamma(\gamma)} (y_i - x_i)^{\gamma-1} \mathrm{e}^{-\delta(y_i - x_i)} \\ & (1 - \mathrm{e}^{-\rho x_i}) \mathbb{1}(x_i \in [0, y_i)) \Bigg]. \end{aligned} \tag{4.1}$$

in order to produce valid inference we integrate out the X_i's whenever not observed, thus obtaining the likelihood function

$$\begin{aligned} L(\alpha,\beta,\gamma,\delta,\rho;\mathbf{x_{obs}},\mathbf{y}) = \prod_{i\in\mathcal{O}} \Bigg[& \frac{\beta^\alpha}{\Gamma(\alpha)} x_i^{\alpha-1} \mathrm{e}^{-\beta x_i} \frac{\delta^\gamma}{\Gamma(\gamma)} (y_i - x_i)^{\gamma-1} \mathrm{e}^{-\delta(y_i - x_i)} \\ & \mathrm{e}^{-\rho x_i} \mathbb{1}(x_i \in [0, y_i)) \Bigg] \times \\ \prod_{i\in\mathcal{M}} \Bigg[& \int \frac{\beta^\alpha}{\Gamma(\alpha)} x_i^{\alpha-1} \mathrm{e}^{-\beta x_i} \frac{\delta^\gamma}{\Gamma(\gamma)} (y_i - x_i)^{\gamma-1} \mathrm{e}^{-\delta(y_i - x_i)} \\ & (1 - \mathrm{e}^{-\rho x_i}) \mathbb{1}(x_i \in [0, y_i)) dx_i \Bigg]. \end{aligned} \tag{4.2}$$

The integral in (4.2) can be split into two integrals as follows:

$$\begin{aligned}
&\int \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} \mathrm{e}^{-\beta x} \frac{\delta^{\gamma}}{\Gamma(\gamma)} (y-x)^{\gamma-1} \mathrm{e}^{-\delta(y-x)} (1-\mathrm{e}^{-\rho x}) dx &= \\
&\frac{\beta^{\alpha}}{\Gamma(\alpha)} \frac{\delta^{\gamma}}{\Gamma(\gamma)} \mathrm{e}^{-\delta y} \int_0^y x^{\alpha-1} \mathrm{e}^{-\beta x} (y-x)^{\gamma-1} \mathrm{e}^{\delta x} (1-\mathrm{e}^{-\rho x}) dx &= \\
&\frac{\beta^{\alpha}}{\Gamma(\alpha)} \frac{\delta^{\gamma}}{\Gamma(\gamma)} \mathrm{e}^{-\delta y} \Big\{ \underbrace{\int_0^y x^{\alpha-1} (y-x)^{\gamma-1} \mathrm{e}^{(\delta-\beta)x} dx}_{A} &+ \\
&- \underbrace{\int_0^y x^{\alpha-1} (y-x)^{\gamma-1} \mathrm{e}^{(\delta-\beta-\rho)x} dx}_{B} \Big\}. & (4.3)
\end{aligned}$$

Note that the two integrals A and B in (4.3) are essentially the same (apart from their parameters). In particular A has been evaluated through the change of variable $t = \frac{x}{y}$

$$\begin{aligned}
A := &\int_0^y x^{\alpha-1} (y-x)^{\gamma-1} \mathrm{e}^{(\delta-\beta)x} dx &= \\
&\int_0^1 (yt)^{\alpha-1} y^{\gamma-1} (1-t)^{\gamma-1} \mathrm{e}^{y(\delta-\beta)t} y dt &= \\
&y^{\alpha+\gamma-1} \int_0^1 t^{\alpha-1} (1-t)^{\gamma-1} \mathrm{e}^{y(\delta-\beta)t} dt. & (4.4)
\end{aligned}$$

Now, recalling the integral representation of the *Kummer M function* (Confluent Hypergeometric function) Def.13.2.1 in [33]

$$\frac{\Gamma(b-a)\Gamma(a)}{\Gamma(b)} M(a,b,z) = \int_0^1 \mathrm{e}^{zt} t^{a-1} (1-t)^{b-a-1} dt \qquad (4.5)$$

considering the parameters

$$z = y(\delta-\beta); \quad a = \alpha; \quad b = \gamma + \alpha,$$

the solution of (4.4) is

$$A = y^{\alpha+\gamma-1} \frac{\Gamma(\gamma)\Gamma(\alpha)}{\Gamma(\gamma+\alpha)} M(\alpha, \gamma+\alpha, y(\delta-\beta)). \qquad (4.6)$$

The solution of B is analogous, using $z' = y(\delta - \beta - \rho)$

$$B = y^{\alpha+\gamma-1}\frac{\Gamma(\gamma)\Gamma(\alpha)}{\Gamma(\gamma+\alpha)}M(\alpha, \gamma+\alpha, y(\delta-\beta-\rho)). \tag{4.7}$$

By combining the results of (4.6) and (4.7) with (4.3) we obtain the likelihood form for the observations for which the birth time is missing i.e., for all y_i's such that $i \in \mathcal{M}$:

$$\begin{aligned} L(\alpha, \beta, \gamma, \delta, \rho; y) = & \frac{\delta^\gamma}{\Gamma(\alpha+\gamma)}\mathrm{e}^{-\delta y}y^{\alpha+\gamma-1}\beta^\alpha[M(\alpha, \gamma+\alpha, y(\delta-\beta))+ \\ & - M(\alpha, \gamma+\alpha, y(\delta-\beta-\rho)]. \end{aligned} \tag{4.8}$$

We are finally able to write the observed likelihood function descending from (4.2) as

$$\begin{aligned} & L(\alpha, \beta, \gamma, \delta, \rho; \mathbf{x_{obs}}, \mathbf{y}) = \\ & \prod_{i\in\mathcal{O}}\left[\frac{\beta^\alpha}{\Gamma(\alpha)}x_i^{\alpha-1}\mathrm{e}^{-\beta x_i}\frac{\delta^\gamma}{\Gamma(\gamma)}(y_i-x_i)^{\gamma-1}\mathrm{e}^{-\delta(y_i-x_i)}\mathrm{e}^{-\rho x_i}\mathbb{1}(x_i\in[0,y_i))\right]\times \\ & \prod_{i\in\mathcal{M}}\left[\frac{\delta^\gamma}{\Gamma(\alpha+\gamma)}\mathrm{e}^{-\delta y_i}y_i^{\alpha+\gamma-1}\beta^\alpha[M(\alpha,\gamma+\alpha,y_i(\delta-\beta))+\right. \\ & \left. - M(\alpha,\gamma+\alpha,y_i(\delta-\beta-\rho)]\mathbb{1}(y_i\in[0,\infty))\right] = \qquad (4.9) \\ & \prod_{i=1}^{n}\left[\beta^\alpha\delta^\gamma\mathrm{e}^{-\delta y_i}\right]\times \\ & \prod_{i\in\mathcal{O}}\left[\frac{1}{\Gamma(\alpha)\Gamma(\gamma)}x_i^{\alpha-1}(y_i-x_i)^{\gamma-1}\mathrm{e}^{(\delta-\beta-\rho)x_i}\mathbb{1}(x_i\in[0,y_i))\right]\times \\ & \prod_{i\in\mathcal{M}}\left[\frac{1}{\Gamma(\alpha+\gamma)}y_i^{\alpha+\gamma-1}[M(\alpha,\gamma+\alpha,y_i(\delta-\beta))+\right. \\ & \left. - M(\alpha,\gamma+\alpha,y_i(\delta-\beta-\rho)]\mathbb{1}(y_i\in[0,\infty)\right], \qquad (4.10) \end{aligned}$$

from (4.10) the log-likelihood easily follows

$$\begin{aligned} l(\alpha,\beta,\gamma,\delta,\rho;\mathbf{x_{obs}},\mathbf{y}) = n[\alpha\log\beta+\gamma\log\delta]-\delta\sum_{i=1}^{n}y_i+ \\ -\#\mathcal{O}[\log\Gamma(\alpha)+\log\Gamma(\gamma)]+(\alpha-1)\sum_{i\in\mathcal{O}}\log x_i+ \\ +(\gamma-1)\sum_{i\in\mathcal{O}}\log(y_i-x_i)+(\delta-\beta-\rho)\sum_{i\in\mathcal{O}}x_i+ \\ -\#\mathcal{M}\log\Gamma(\alpha+\gamma)+(\alpha+\gamma-1)\sum_{i\in\mathcal{M}}\log y_i+ \\ -\sum_{i\in\mathcal{M}}\log[M(\alpha,\gamma+\alpha,y_i(\delta-\beta))-M(\alpha,\gamma+\alpha,y_i(\delta-\beta-\rho))] \end{aligned} \tag{4.11}$$

This log-likelihood can then be used for inference.

4.2 Birth and Death model simulations

In order to explore the behaviour of the model and the sensitivity to some changes on the basic assumptions, we performed some simulations. In the first simulation a sample of size $n = 1,000$ was generated using $X \sim Gamma(5,3)$ and $Y - X \sim Gamma(8,2)$, the missing data mechanism was set as $f(r = 1|x;\rho) = \mathrm{e}^{-\rho x} = \mathrm{e}^{-0.2x}$. Maximum likelihood estimation was performed on two different models: the first one considering the missing data mechanism (i.e. considering the likelihood in (4.10)), while the second one ignored the MDM and considered only the likelihood of complete observations.

In Table 4.1 the estimates of the two models are reported. As expected, the estimation conducted considering the complete likelihood function led to estimates close to the true values, while the estimation based on the complete data technique led to biased results. In particular, looking at the estimated moments of the two distributions (e.g., $\hat{\mathrm{E}}(X) = \hat{\alpha}/\hat{\beta}$) it can be noticed that the distribution of X is more contracted toward the origin i.e., a lower mean and variance are estimated as a consequence of the *censoring* acted by the missing data mechanism on the original data: the lower the value is the higher is the chance for it to be observed. This departure from the true distribution does not seem

model/parameters	α	β	γ	δ	ρ
true	5,00	3,00	8,00	2,00	0,20
MDM considered	5,41	3,31	7,88	1,96	0,21
MDM not considered	5,29	3,41	7,66	1,88	-
model/moments	$\mathrm{E}(X)$	$\mathrm{Var}\,(X)$	$\mathrm{E}(Y-X)$	$\mathrm{Var}\,(Y-X)$	
true	1,67	0,56	4,00	2,00	
MDM considered	1,63	0,49	4,02	2,05	
MDM not considered	1,55	0,46	4,08	2,17	

Table 4.1: Birth and Death Model. Estimation results for model considering the MDM and the model that ignores the MDM ($n = 1,000$ and $\rho = 0.2$)

however to be too serious in these simulated data, and this is due to the fact that the small value of $\rho = 0.2$ leads to probabilities of being observed close to 1 for any value of x. An analogous simulation was performed with $\rho = 1$ and leaving the other parameters unchanged.

This change in the MDM reduced abruptly the observed data fraction from 72% to the 25%, and produced the results presented in Table 4.2.

Looking at the results, it is easy to notice how the more incisive selection effect, introduced by the increased value of ρ, acts on the difference between the two models, leading for X to a great reduction on the mean estimate and to a variance estimate which is nearly half of the true one. At the same time, it has to be noticed that the estimation of the parameters of the $Y - X$ distribution is not affected, since the MDM is defined on the X random variable. To have an insight on the error arising under the complete observations technique, could be useful to consider the likelihood function. The inference in this case is based only on the fully observed samples (and hence on the observations $i \in \mathcal{O}$)

model/parameters	α	β	γ	δ	ρ
true	5,00	3,00	8,00	2,00	1,00
MDM considered	5,10	3,05	8,16	2,04	0,94
MDM not considered	5,16	4,19	7,87	1,96	-
model/moments	E(X)	Var (X)	E($Y-X$)	Var ($Y-X$)	
true	1,67	0,56	4,00	2,00	
MDM considered	1,67	0,55	3,99	1,95	
MDM not considered	1,23	0,29	4,01	2,05	

Table 4.2: Birth and Death Model. Estimation results for model considering the MDM and the model that ignores the MDM ($n = 1,000$ and $\rho = 1$)

$$\begin{aligned} f(x;\alpha,\beta)f(r=1|x;\rho) =& \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} e^{-\rho} \\ =& \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-(\beta+\rho)x}. \end{aligned}$$

To obtain the density of X when it is observed we need to calculate the integration constant

$$\frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty x^{\alpha-1} e^{-(\beta+\rho)x} dx = \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha)}{(\beta+\rho)^\alpha} = \left(\frac{\beta}{\beta+\rho}\right)^\alpha,$$

so that under complete observation the density function of the observed data is

$$\begin{aligned} f(x|r=1;\alpha,\beta,\rho) =& \left(\frac{\beta}{\beta+\rho}\right)^{-\alpha} \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-(\beta+\rho)x} \\ =& \frac{(\beta+\rho)^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-(\beta+\rho)x}, \end{aligned}$$

which is the density of a $Gamma(\alpha, \beta+\rho)$.

4.3 Monza and Brianza Businesses Dataset

This dataset contains 25,000 records extracted from the Infocamere database detained by Unioncamere, an association of the Italian Chambers of Commerce. Infocamere provides also, on a quarterly base, statistics about businesses births and death aggregated by economic sector and province. This information is available through the web archive Movimprese; however, the available data comes from an *ad hoc* detailed extraction in which the population considered is composed by businesses registered in the newborn Monza and Brianza province and the information regard the main aspects of the businesses' life such as birth time, death time, sector of activity etc.

The complete list of variables is as follows:

- type of company;
- town;
- economic sector;
- registration time (*birth time*);
- cancellation time (*death time*).

Since the original dataset included 15 sectors of activity, businesses were aggregated into a broader classification as shown in Table 4.3.

For ease of calculation two variables, representing time (birth and death) were converted from date to numeric format assuming 1st January 1900 as the origin and using year as measurement unit, hence converting e.g. 1st July 2000 to 100.5; birth and death time, represented in this way, will be called henceforth X and Y. At the moment of the data extraction, a particular type of *selection* acted: only those businesses which ended their activity in the 1996-2006 period were included. Furthermore, two variables in this dataset present *missing values*, and precisely 1,059 for the economic sector and 10,426 for the birth time. While one thousand missing records is not a big proportion among twenty

Sector	Description	Macro Sector
A	agriculture	O
B	fishing	O
C	mining	O
D	industry	D
E	energy	O
F	construction	F
G	commerce	G
H	hotels and restaurants	H
I	logistics	I
J	finance	J
K	other activities	O
L	public	O
M	education	O
N	health care	O

Table 4.3: Monza and Brianza dataset: economic sectors

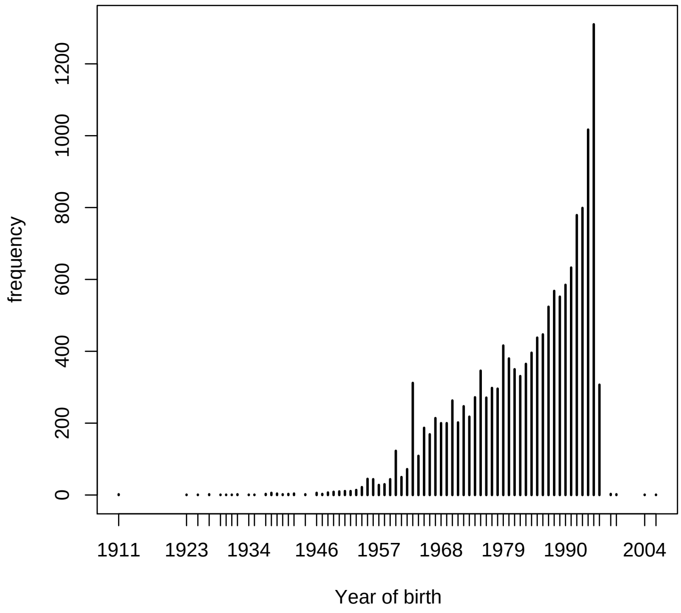

Figure 4.1: Monza and Brianza dataset: Year of birth distribution

five thousands observations, ten thousands is a rather pathological quota which requires modelling to be handled.

The main problem connected with the missingness of the birth time is that it makes it impossible to calculate the lifetimes and hence to perform a *survival analysis* to investigate the mortality patterns. However, even with standard tools such as graphical representation it is possible to perform a preliminary exploratory analysis to investigate the *missing data mechanism*.

Fig. 4.1 shows the distribution of the observed years of birth. This variable shows quite a regular distribution with all births concentrated against the left extreme of the window in which the deaths were observed (1996-2006). Looking at the right side of the birth time distribution we note that the last visible stick, corresponding to the year 1996, seems to correspond to observations for which there is a *partial observation*.

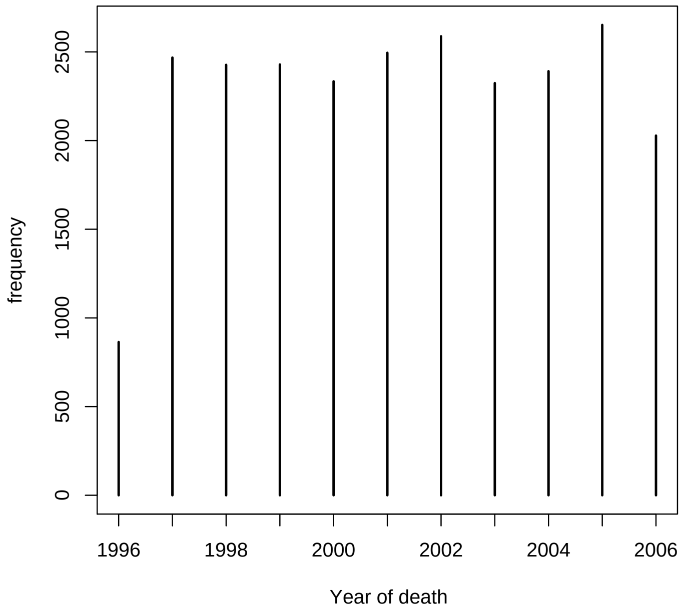

Figure 4.2: Monza and Brianza dataset: Year of death distribution

year	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006
deaths	864	2468	2427	2429	2334	2495	2588	2324	2391	2652	2028

Table 4.4: Monza and Brianza dataset: Deaths by year

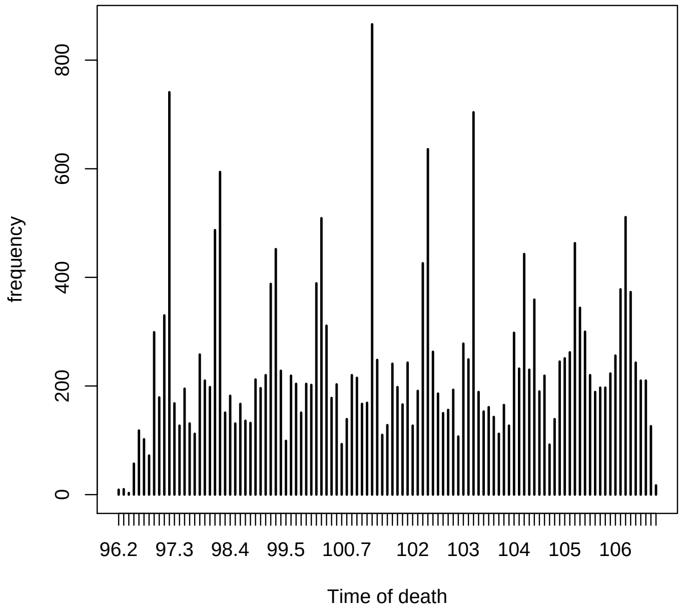

Figure 4.3: Monza and Brianza dataset: Time of death distribution

On the other hand, looking at the year of death distribution as shown by Fig. 4.2 (death counts are presented in Table 4.4) we notice that the number of businesses which ended their activity is roughly constant, except for 1996. As for the case of the births, this year seems to be anomalous from the *data collection* viewpoint. Another feature of the death time variable Y is due to *data registration*: looking at Fig. 4.3 it is easy to recognize high peaks corresponding to particular days of the year. This fact seems to be due to the registration process.

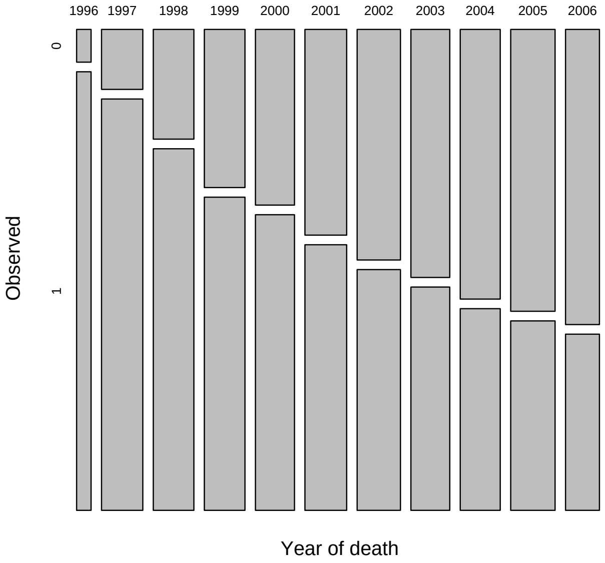

Figure 4.4: Monza and Brianza dataset: Proportion of missing by year of death

4.4 Birth and Death model: an application to Monza and Brianza firms data

All the subjects in the observed data experienced death between the years 1996 and 2006, and what was missing were birth times. Looking back at Fig. 4.1 births seem to concentrate close to the window in which deaths were observed suggesting a possible exponential decay in these life times. In Fig. 4.4 the proportion of missing data by year of death is reported.

The first possible and direct explanation is that the probability for a birth time of being observed is linked to the death time and decays as this last one increases. Another possibility is that the birth dates included in the 1996-2006 period were not registered in the dataset. As noted before, lifetimes seem to decay exponentially, and hence as death time increases, birth time is more likely to be included in the death observation period.

In the first case, a modification of the MDM function is required. A possible formulation is

$$f(r = 1|y; \omega) = \min(\mathrm{e}^{-\rho(y-\omega)}, 1), \tag{4.12}$$

where ω regulates the time from which the probability of being observed decays. In this case, after dividing times by ten to make the numerical computations more stable, we chose $\omega = 9.6$. By converse,

$$f(r = 0|y; \omega) = \max(1 - \mathrm{e}^{-\rho(y-\omega)}, 0). \tag{4.13}$$

In this situation it may seem to be in the missing at random case since the MDM depends only on the death time Y. However, since the distribution of this variable is tied to the one of X (being a translated gamma distribution), we are definitely in the missing-not-at-random situation. Recalling (4.1), the complete data likelihood needs to be modified

$$\begin{aligned} L(\alpha, \beta, \gamma, \delta, \rho; \mathbf{x}, \mathbf{y}, \omega) = & \prod_{i \in \mathcal{O}} \Bigg[\frac{\beta^\alpha}{\Gamma(\alpha)} x_i^{\alpha-1} \mathrm{e}^{-\beta x_i} \frac{\delta^\gamma}{\Gamma(\gamma)} (y_i - x_i)^{\gamma-1} \mathrm{e}^{-\delta(y_i - x_i)} \\ & \min(\mathrm{e}^{-\rho(y_i-\omega)}, 1) \mathbb{1}(x_i \in [0, y_i)) \Bigg] \times \\ & \prod_{i \in \mathcal{M}} \Bigg[\frac{\beta^\alpha}{\Gamma(\alpha)} x_i^{\alpha-1} \mathrm{e}^{-\beta x_i} \frac{\delta^\gamma}{\Gamma(\gamma)} (y_i - x_i)^{\gamma-1} \mathrm{e}^{-\delta(y_i - x_i)} \\ & \max(1 - \mathrm{e}^{-\rho(y_i-\omega)}, 0) \mathbb{1}(x_i \in [0, y_i)) \Bigg]. \end{aligned} \tag{4.14}$$

the observed data likelihood must be integrated, to obtain

$$\begin{aligned} L(\alpha,\beta,\gamma,\delta,\rho;\mathbf{x_{obs}},\mathbf{y},\omega) = & \prod_{i\in\mathcal{O}}\Bigg[\frac{\beta^\alpha}{\Gamma(\alpha)}x_i^{\alpha-1}\mathrm{e}^{-\beta x_i}\frac{\delta^\gamma}{\Gamma(\gamma)}(y_i-x_i)^{\gamma-1}\mathrm{e}^{-\delta(y_i-x_i)} \\ & \min(\mathrm{e}^{-\rho(y_i-\omega)},1)\mathbb{1}(x_i\in[0,y_i))\Bigg]\times \\ & \prod_{i\in\mathcal{M}}\Bigg[\int\frac{\beta^\alpha}{\Gamma(\alpha)}x_i^{\alpha-1}\mathrm{e}^{-\beta x_i}\frac{\delta^\gamma}{\Gamma(\gamma)}(y_i-x_i)^{\gamma-1}\mathrm{e}^{-\delta(y_i-x_i)} \\ & \max(1-\mathrm{e}^{-\rho(y_i-\omega)},0)\mathbb{1}(x_i\in[0,y_i))dx_i\Bigg]. \qquad (4.15) \end{aligned}$$

The integral in (4.15)

$$\begin{aligned} & \int\frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}\mathrm{e}^{-\beta x}\frac{\delta^\gamma}{\Gamma(\gamma)}(y-x)^{\gamma-1}\mathrm{e}^{-\delta(y-x)}\max(1-\mathrm{e}^{-\rho(y-\omega)},0)dx \quad = \\ & \frac{\beta^\alpha}{\Gamma(\alpha)}\frac{\delta^\gamma}{\Gamma(\gamma)}\mathrm{e}^{-\delta y}\max(1-\mathrm{e}^{-\rho(y-\omega)},0)\int_0^y x^{\alpha-1}\mathrm{e}^{-\beta x}(y-x)^{\gamma-1}\mathrm{e}^{\delta x}dx \quad = \\ & \frac{\beta^\alpha}{\Gamma(\alpha)}\frac{\delta^\gamma}{\Gamma(\gamma)}\mathrm{e}^{-\delta y}\max(1-\mathrm{e}^{-\rho(y-\omega)},0)\ A \qquad (4.16) \end{aligned}$$

is function of the part A of the integral in (4.3). Recalling (4.6) the likelihood for the missing observations (i.e. for $i\in\mathcal{M}$) is

$$\begin{aligned} L(\alpha,\beta,\gamma,\delta,\rho;y,\omega) = & \frac{\beta^\alpha\delta^\gamma}{\Gamma(\alpha+\gamma)}\mathrm{e}^{-\delta y}y^{\alpha+\gamma-1}M(\alpha,\gamma+\alpha,y(\delta-\beta))\times \\ & \max(1-\mathrm{e}^{-\rho(y-\omega)},0) \qquad (4.17) \end{aligned}$$

Finally the observed data likelihood follows as

$$\begin{align*}
&L(\alpha,\beta,\gamma,\delta,\rho;\mathbf{x_{obs}},\mathbf{y},\omega) = \\
&\quad \prod_{i\in\mathcal{O}}\Bigg[\frac{\beta^{\alpha}}{\Gamma(\alpha)}x_i^{\alpha-1}\mathrm{e}^{-\beta x_i}\frac{\delta^{\gamma}}{\Gamma(\gamma)}(y_i-x_i)^{\gamma-1}\mathrm{e}^{-\delta(y_i-x_i)} \\
&\quad \min(\mathrm{e}^{-\rho(y_i-\omega)},1)\mathbb{1}(x_i\in[0,y_i))\Bigg]\times \\
&\quad \prod_{i\in\mathcal{M}}\Bigg[\frac{\beta^{\alpha}\delta^{\gamma}}{\Gamma(\alpha+\gamma)}\mathrm{e}^{-\delta y_i}y_i^{\alpha+\gamma-1}M(\alpha,\gamma+\alpha,y(\delta-\beta)) \\
&\quad \max(1-\mathrm{e}^{-\rho(y-\omega)},0)\mathbb{1}(y_i\in[0,\infty)\Bigg] = \tag{4.18} \\
&\quad \prod_{i=1}^{n}\Bigg[\beta^{\alpha}\delta^{\gamma}\mathrm{e}^{-\delta y_i}\Bigg]\times \\
&\quad \prod_{i\in\mathcal{O}}\Bigg[\frac{1}{\Gamma(\alpha)\Gamma(\gamma)}x_i^{\alpha-1}(y_i-x_i)^{\gamma-1}\mathrm{e}^{(\delta-\beta)x_i} \\
&\quad \min(\mathrm{e}^{-\rho(y_i-\omega)},1)\mathbb{1}(x_i\in[0,y_i))\Bigg]\times \\
&\quad \prod_{i\in\mathcal{M}}\Bigg[\frac{1}{\Gamma(\alpha+\gamma)}y_i^{\alpha+\gamma-1}M(\alpha,\gamma+\alpha,y_i(\delta-\beta)) \\
&\quad \max(1-\mathrm{e}^{-\rho(y-\omega)},0)\mathbb{1}(y_i\in[0,\infty)\Bigg] \tag{4.19}
\end{align*}$$

It is easy to write down the log-likelihood as

α	β	γ	δ	ρ	ω
71.06	8.37	2.63	1.56	1.01	9.60
E(X)	Var(X)	E(Y-X)	Var(Y-X)		
8.49	1.02	1.68	1.08		

Table 4.5: Birth and Death Model. Estimation results for the Monza and Brianza dataset

$$\begin{aligned}
l(\alpha,\beta,\gamma,\delta,\rho;\mathbf{x_{obs}},\mathbf{y},\omega) &= n[\alpha\log\beta+\gamma\log\delta]-\delta\sum_{i=1}^{n}y_i+\\
&-\#\mathcal{O}[\log\Gamma(\alpha)+\log\Gamma(\gamma)]+(\alpha-1)\sum_{i\in\mathcal{O}}\log x_i+\\
&+(\gamma-1)\sum_{i\in\mathcal{O}}\log(y_i-x_i)+(\delta-\beta)\sum_{i\in\mathcal{O}}x_i+\\
&+\sum_{i\in\mathcal{O}}\log\min(\mathrm{e}^{-\rho(y_i-\omega)},1)-\#\mathcal{M}\log\Gamma(\alpha+\gamma)+\\
&+(\alpha+\gamma-1)\sum_{i\in\mathcal{M}}\log y_i-\sum_{i\in\mathcal{M}}\log M(\alpha,\gamma+\alpha,y_i(\delta-\beta))+\\
&+\sum_{i\in\mathcal{M}}\log\max(1-\mathrm{e}^{-\rho(y-\omega)},0).
\end{aligned}\tag{4.20}$$

The log-likelihood in (4.20) has been maximized in order to fit the Birth and Death model to the Monza and Brianza dataset. The results of estimation are reported in Table 4.5.

Based on these results, the mean birth year of firms that died in the 1996-2006 decade is around 1985, which is slightly larger than the sample mean (calculated on complete data) of about 1983. This is due to the fact that the missing birth years were implicitly imputed by the model to more recent times. Analogously, the mean lifetime of 16.8 years is larger than the complete data sample mean of 12.9 because the missing data mechanism of this model, as expressed in (4.12), assigns higher probability of being missing to observations with higher death time, and consequently with a longer lifetime.

Furthermore as can be seen in Fig. 4.5 the estimated exponential decay parameter

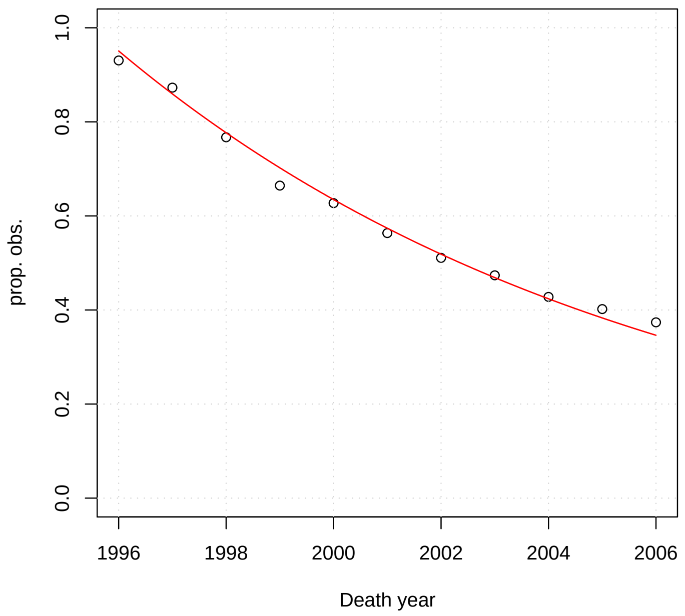

Figure 4.5: Monza and Brianza dataset: Observed versus estimated proportion of missingness

($\hat{\rho} = 1.01$) leads to a good fit to the observed sample rate. However, what can be asserted by looking at this model is that it is adequate to *mimic* the data. The estimation is, in fact, influenced by the MDM that has been assumed.

The main feature of these data that the model fails to take into account is that lifetimes should be truncated, since for all observations the death time Y cannot exceed 10.7 (corresponding to the end of the year 2006). This leads to a truncation on the lifetime distribution, and for each observation truncation is conditional on the birth time X. This truncation could be applied to the fully observed data, since it will require the use of the well known *incomplete gamma function*.

4.5 A Birth and Death Distribution

Let us recall the A integral from (4.4)

$$\int_0^y x^{\alpha-1}(y-x)^{\gamma-1}\mathrm{e}^{(\delta-\beta)x}dx \tag{4.21}$$

X is a gamma distributed random variable and Y is a gamma distributed random variable translated by X from the origin, and hence it is the sum of two independent gamma random variables.

Since the joint density of X and Y is

$$\begin{aligned} f(x,y;\alpha,\beta,\gamma,\delta) =& \frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}\mathrm{e}^{-\beta x}\frac{\delta^\gamma}{\Gamma(\gamma)}(y-x)^{\gamma-1}\mathrm{e}^{-\delta(y-x)} \\ & \mathbb{1}(y\in[0,\infty))\mathbb{1}(x\in[0,y)) \end{aligned} \tag{4.22}$$

(4.21) can then be used to obtain the density for Y, as the convolution kernel of the two random variables.

Using (4.6), as the expression of the A, the density of Y easily follows as

$$f(y;\alpha,\beta,\gamma,\delta) = \frac{\beta^\alpha\delta^\gamma}{\Gamma(\alpha+\gamma)}y^{\alpha+\gamma-1}\mathrm{e}^{-\delta y}M(\alpha,\alpha+\gamma,y(\delta-\beta)) \tag{4.23}$$

From (4.23) it is possible to obtain a first integral relation, given by the integration constant:

$$\int_0^\infty \frac{\beta^\alpha\delta^\gamma}{\Gamma(\alpha+\gamma)}y^{\alpha+\gamma-1}\mathrm{e}^{-\delta y}M(\alpha,\alpha+\gamma,y(\delta-\beta))dy = 1, \tag{4.24}$$

it must be that

$$\int_0^\infty y^{\alpha+\gamma-1}\mathrm{e}^{-\delta y}M(\alpha,\alpha+\gamma,y(\delta-\beta))dy = \frac{\Gamma(\alpha+\gamma)}{\beta^\alpha\delta^\gamma} \tag{4.25}$$

This four-parameters density allows one to make inference onto two, subsequent, independent time-related phenomena having observed just the latter of the two. Having fixed a time origin, the first phenomenon starts to act at different times (that will not be observed), once the first event happens the second phenomenon can act, but what is observed is only the sum of these two times or, analogously, the time of the second event passed since the time origin. Examples of possible applications are the reconstruction of birth and death distributions when births are not observed, but their distribution can be assumed to start from a certain point in time (same breed), or the estimation of a contagion time when some deadly disease starts to spread (time origin) and only death times are observed.

4.6 Results for the Kummer M Function

Some distributional properties of the proposed model are easy to derive by recalling that the death time random variable is the convolution of two gamma random variables. Statistical results involving the density function (4.23) can be used to obtain analytical results (such as integral relations) on the *Confluent Hypergeometric function.* The main properties of the Confluent Hypergeometric function that will be recalled in this section can be found in Appendix A.

The first question that will be addressed is whether or not the convolution is *log concave.* A function f on $\mathbb{R}^d$ is said to be log concave if it is of the form

$$f(x) = \exp\{\phi(x)\} \tag{4.26}$$

for some concave function $\phi : \mathbb{R}^d \to (-\infty, +\infty)$. The class of log concave distributions includes the most widely used parametric distributions and has some interesting properties descending from the shape of the density which always admits a global maximum, thus allowing maximization using MLE algorithms. A recent review of this subject can be found in [34].

Also, a convolution of two log concave distributions is still log concave. This comes from the fact that log concave functions are PF_2 i.e. *Polya Frequency functions* of order two [35] [36] and by the *Basic Composition Formula*, the convolution of Polya frequency functions of different order is still a Polya frequency function of order equal to the lower of the two.

The moments of the distribution are easy to calculate, recalling that the expected value for a gamma random variable X is

$$\begin{aligned} \mathrm{E}(X) =& \int_0^\infty x\frac{\lambda^\alpha}{\Gamma(\alpha)}x^{\alpha-1}\mathrm{e}^{-\lambda x}dx = \frac{\lambda^\alpha}{\Gamma(\alpha)}\int_0^\infty x^\alpha \mathrm{e}^{-\lambda x}dx = \\ =&\frac{\Gamma(\alpha+1)}{\lambda^{\alpha+1}}\frac{\lambda^\alpha}{\Gamma(\alpha)} = \frac{\alpha}{\lambda}, \end{aligned} \tag{4.27}$$

Since Y is the sum of two independent gamma random variables it is immediate to obtain $\mathrm{E}(Y) = \alpha/\beta + \gamma/\delta$. From this result, another integral relation can be derived for the Kummer M function, being

$$\begin{aligned} \mathrm{E}(Y) =& \int_0^\infty y\frac{\beta^\alpha\delta^\gamma}{\Gamma(\alpha+\gamma)}y^{\alpha+\gamma-1}\mathrm{e}^{-\delta y}M(\alpha,\alpha+\gamma,y(\delta-\beta))dy = \\ =&\frac{\beta^\alpha\delta^\gamma}{\Gamma(\alpha+\gamma)}\int_0^\infty y^{\alpha+\gamma}\mathrm{e}^{-\delta y}M(\alpha,\alpha+\gamma,y(\delta-\beta))dy = \\ =&\frac{\alpha}{\beta}+\frac{\gamma}{\delta} = \frac{\alpha\delta+\gamma\beta}{\beta\delta} \end{aligned} \tag{4.28}$$

which leads to

$$\int_0^\infty y^{\alpha+\gamma}\mathrm{e}^{-\delta y}M(\alpha,\alpha+\gamma,y(\delta-\beta))dy = \frac{\Gamma(\alpha+\gamma)}{\beta^\alpha\delta^\gamma}\frac{\alpha\delta+\gamma\beta}{\beta\delta}. \tag{4.29}$$

Another integral relationship, arising from the probabilistic properties of the density function descends from the *moment generating function* (mgf). The mgf of a random variable X is

$$M_X(t) = \mathrm{E}(\mathrm{e}^{tx}) \text{ for any t in a neighbourhood of 0.} \tag{4.30}$$

recall that $M_X(t) = 1$.

The moment generating function is important because when it exists, and when is available in closed form, it allows one to obtain the moments of any order for the random variable X. Simply differentiating it k times and computing the derivative at $t = 0$:

$$m_k = \frac{\mathrm{d}^k}{\mathrm{d}t^k} M_X(t)\bigg|_{t=0}. \tag{4.31}$$

Also, recall that moment generating function the convolution of two or more independent random variables, can be obtained immediately as for $Y = X + Z$, X and Z independent

$$M_Y(t) = M_X(t)M_Z(t)$$

In the case of the Gamma density function, the moment generating function is

$$\begin{aligned} M_X(t) = \int_0^\infty \mathrm{e}^{tx} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \mathrm{e}^{-\lambda x} dx = \int_0^\infty \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \mathrm{e}^{-(\lambda-t)x} dx = \\ = \left(\frac{\lambda}{\lambda - t}\right)^\alpha = \left(1 - \frac{t}{\lambda}\right)^{-\alpha}, \end{aligned} \tag{4.32}$$

and (4.31) can be used for example to easily obtain the first moment

$$\frac{\mathrm{d}}{\mathrm{d}t} M_X(t)\bigg|_{t=0} = \frac{\alpha}{\lambda}. \tag{4.33}$$

Now, let us go back to the distribution of the sum Y of two gamma random variables. Applying the (4.32) it is easy to obtain its moment generating function

$$\begin{aligned} M_Y(t) = M_X(t)M_Z(t) = & \left(1 - \frac{t}{\beta}\right)^{-\alpha} \left(1 - \frac{t}{\delta}\right)^{-\gamma} = \\ = & \left(\frac{\beta}{\beta - t}\right)^\alpha \left(\frac{\delta}{\delta - t}\right)^\gamma. \end{aligned} \tag{4.34}$$

another integral relationship then follows since

$$\begin{aligned} M_Y(t) = &\int_0^\infty \mathrm{e}^{ty} \frac{\beta^\alpha \delta^\gamma}{\Gamma(\alpha+\gamma)} y^{\alpha+\gamma-1} \mathrm{e}^{-\delta y} M(\alpha, \alpha+\gamma, y(\delta-\beta)) dy = \\ = &\frac{\beta^\alpha \delta^\gamma}{\Gamma(\alpha+\gamma)} \int_0^\infty y^{\alpha+\gamma-1} \mathrm{e}^{-(\delta-t)y} M(\alpha, \alpha+\gamma, y(\delta-\beta)) dy = \\ = &\left(\frac{\beta}{\beta-t}\right)^\alpha \left(\frac{\delta}{\delta-t}\right)^\gamma, \end{aligned} \tag{4.35}$$

and

$$\int_0^\infty y^{\alpha+\gamma-1} \mathrm{e}^{-(\delta-t)y} M(\alpha, \alpha+\gamma, y(\delta-\beta)) dy = \frac{\Gamma(\alpha+\gamma)}{(\beta-t)^\alpha (\delta-t)^\gamma} \tag{4.36}$$

The following distributional property confirms a well known special value for the Kummer M function. It is well know that the sum of two independent gamma random variables with the same rate parameter is still a gamma random variable

$$Y = Gamma(\alpha, \lambda) + Gamma(\gamma, \lambda) \sim Gamma(\alpha+\gamma, \lambda).$$

The birth and death density, in our proposed model, in the case in which $\beta = \delta = \lambda$, should then reduce to a $Gamma(\alpha+\gamma, \lambda)$ density

$$f(y; \alpha, \gamma, \lambda) = \frac{\lambda^{\alpha+\gamma}}{\Gamma(\alpha+\gamma)} \mathrm{e}^{-\lambda y} y^{\alpha+\gamma-1} M(\alpha, \alpha+\gamma, 0), \tag{4.37}$$

and this is true since $M(a, b, 0) = 1$ for any a and b.

This last property shows how from simple probabilistic considerations can descend a fundamental property of the Kummer M function. Since the proposed distribution is the result of the convolution of two random variables, the order of the sum should not be relevant. For this reason the following equality must hold:

$$f(y; \alpha, \beta, \gamma, \delta) = f(y; \gamma, \delta, \alpha, \beta),$$

this means that

$$\frac{\beta^\alpha \delta^\gamma}{\Gamma(\alpha+\gamma)} y^{\alpha+\gamma-1} \mathrm{e}^{-\delta y} M(\alpha, \alpha+\gamma, y(\delta-\beta)) = \\ \frac{\beta^\alpha \delta^\gamma}{\Gamma(\alpha+\gamma)} y^{\alpha+\gamma-1} \mathrm{e}^{-\beta y} M(\gamma, \alpha+\gamma, y(\beta-\delta)), \tag{4.38}$$

then the following must be true

$$\mathrm{e}^{-\delta y} M(\alpha, \alpha+\gamma, y(\delta-\beta)) = \mathrm{e}^{-\beta y} M(\gamma, \alpha+\gamma, y(\beta-\delta)), \tag{4.39}$$

and finally

$$\mathrm{e}^{y(\beta-\delta)} M(\alpha, \alpha+\gamma, y(\delta-\beta)) = M(\gamma, \alpha+\gamma, y(\beta-\delta)). \tag{4.40}$$

(4.40) is known as the *Kummer relation* and is usually [33] reported in the form

$$\mathrm{e}^x M(a, b, -x) = M(b-a, b, x). \tag{4.41}$$

4.7 Special Cases

Recall that the gamma distribution includes some special distributions. The best known cases are the *exponential* $Exp(\beta) = Gamma(1, \beta)$ and the *chi squared* with ν degrees of freedom $\chi^2(\nu) = Gamma(\nu/2, 1/2)$. Since these two densities are special cases of the gamma density, it is possible to obtain the convolution density for any pair of these distributions from the general results above.

The first combination considered is the Gamma-Exponential, in this case we have

$$\begin{aligned} \alpha &= \alpha \\ \beta &= \beta \\ \gamma &= 1 \\ \delta &= \delta \end{aligned}$$

with these parameters (4.23) is

$$f(y;\alpha,\beta,1,\delta)=\frac{\beta^{\alpha}\delta}{\Gamma(\alpha+1)}y^{\alpha}\mathrm{e}^{-\delta y}M(\alpha,\alpha+1,y(\delta-\beta)) \tag{4.42}$$

in this situation when $M(a,a+1,-x)$ with x strictly positive ($\beta>\delta$) the Kummer M function reduces to

$$M(a,a+1,-x)=ax^{-a}\gamma(a,x) \tag{4.43}$$

where $\gamma(a,x)$ is the *Lower Incomplete Gamma Function* defined as

$$\gamma(\alpha,x)=\int_0^x t^{\alpha-1}\mathrm{e}^{-t}\,\mathrm{d}t \tag{4.44}$$

and is the complement to 1 of the *Upper Incomplete Gamma Function*

$$\Gamma(\alpha,x)=\int_x^{\infty} t^{\alpha-1}\mathrm{e}^{-t}\,\mathrm{d}t \tag{4.45}$$

these two functions compose the gamma function $\Gamma(\alpha)=\gamma(\alpha,x)+\Gamma(\alpha,x)$

Applying (4.43) to the density (4.42) leads to

$$\begin{aligned}f(y;\alpha,\beta,1,\delta)=&\frac{\beta^{\alpha}\delta}{\Gamma(\alpha+1)}y^{\alpha}\mathrm{e}^{-\delta y}M(\alpha,\alpha+1,y(\delta-\beta))\\=&\frac{\beta^{\alpha}\delta}{\Gamma(\alpha+1)}y^{\alpha}\mathrm{e}^{-\delta y}\alpha[y(\beta-\delta)]^{-\alpha}\gamma(\alpha,y(\beta-\delta))\\=&\frac{\beta^{\alpha}\delta(\beta-\delta)^{-\alpha}}{\Gamma(\alpha)}\mathrm{e}^{-\delta y}\gamma(\alpha,y(\beta-\delta))\\=&\left[\frac{\beta}{\beta-\delta}\right]^{\alpha}\frac{\gamma(\alpha,y(\beta-\delta))}{\Gamma(\alpha)}\delta\mathrm{e}^{-\delta y},\end{aligned} \tag{4.46}$$

recalling that the cumulative distribution function for a gamma random variable is

$$\begin{aligned}F(y;\alpha,\lambda)=&\int_0^y\frac{\lambda^{\alpha}}{\Gamma(\alpha)}x^{\alpha-1}\mathrm{e}^{-\lambda x}\,\mathrm{d}x=\frac{1}{\Gamma(\alpha)}\int_0^{\lambda y}t^{\alpha-1}\mathrm{e}^{-tx}\,\mathrm{d}t=\\&\frac{\gamma(\alpha,\lambda y)}{\Gamma(\alpha)}.\end{aligned} \tag{4.47}$$

By the change of variable $t = \lambda x$, the (4.46) is the product of a $Gamma(\alpha, \beta - \delta)$ CDF evaluated at y and of an $Exp(\delta)$ density evaluated at the same point.

This means that evaluating the probability that the sum Y of a gamma random variable $X \sim Gamma(\alpha, \beta)$ and of an (independent) exponential random variable $Z \sim Exp(\delta)$, $\beta > \delta$, lies in the infinitesimal interval centered on y, is the same as evaluating the probability that in the same interval an event governed by the $Y' \sim Exp(\delta)$ random variable happens, and conditional on this the random variable $X' \sim Gamma(\alpha, \beta - \delta)$ takes a value lower than y. This simpler form and its probabilistic interpretation seems to be available only for the case in which the risk rate of the gamma distribution is higher than the one of the exponential distribution.

A second combination considered is the convolution of two exponential distributions with different risk rates.

$$\begin{aligned} \alpha &= 1 \\ \beta &= \beta \\ \gamma &= 1 \\ \delta &= \delta \end{aligned}$$

In this case the density of the convolution reduces to

$$f(y; 1, \beta, 1, \delta) = \beta\delta\, y\, \mathrm{e}^{-\delta y} M(1, 2, y(\delta - \beta)) \tag{4.48}$$

and even for this situation a special case of the M function can be considered. In particular,

$$M(1, 2, 2z) = \frac{\mathrm{e}^z}{z} \sinh z, \tag{4.49}$$

which implies

$$M(1, 2, y(\delta - \beta)) = 2\frac{\exp\left\{y\, \frac{\delta-\beta}{2}\right\}}{y(\delta - \beta)} \sinh\left(y\, \frac{\delta - \beta}{2}\right) \tag{4.50}$$

leading to the closed form density

$$f(y;1,\beta,1,\delta) = 2\frac{\beta\delta}{\delta-\beta}\exp\left\{-y\,\frac{\delta+\beta}{2}\right\}\sinh\left(y\,\frac{\delta-\beta}{2}\right). \tag{4.51}$$

Another particular form of this distribution arises when the shape parameters of the two gamma are equal

$$\begin{aligned}\alpha &= \alpha\\ \beta &= \beta\\ \gamma &= \alpha\\ \delta &= \delta\end{aligned}$$

With these parameters the density becomes

$$f(y;\alpha,\beta,\alpha,\delta) = \frac{(\beta\delta)^{\alpha}}{\Gamma(2\alpha)}y^{2\alpha-1}\mathrm{e}^{-\delta y}M(\alpha,2\alpha,y(\delta-\beta)), \tag{4.52}$$

and the special case of the Kummer M function useful in this situation is the following

$$M(\nu+\frac{1}{2},2\nu+1,2z) = \Gamma(1+\nu)\,\mathrm{e}^{z}\left(\frac{z}{2}\right)^{-\nu}I_{\nu}(z), \tag{4.53}$$

where $I_\nu(z)$ is the *modified Bessel function of first kind.* Since in this case $\nu = \alpha - \frac{1}{2}$ and $z = y\frac{\delta-\beta}{2}$, the closed form of the density becomes

$$f(y;\alpha,\beta,\alpha,\delta) = \frac{\Gamma(\alpha+\frac{1}{2})}{\Gamma(2\alpha)}\frac{(\beta\delta)^{\alpha}}{\left(\frac{\delta-\beta}{4}\right)^{\alpha-\frac{1}{2}}}y^{\alpha-\frac{1}{2}}\,\mathrm{e}^{-y\frac{\beta+\delta}{2}}I_{\alpha-\frac{1}{2}}\left(y\frac{\delta-\beta}{2}\right). \tag{4.54}$$

While this birth and death model efficiently incorporates the missing data mechanism into the likelihood function, it fails to consider the sampling mechanism (time window). The Lexis point process is introduced in the next chapter, and the sampling mechanism will be effectively included as well into the likelihood function.

Chapter 5

Poisson Point Process in the Lexis Diagram

In this chapter a general model for handling birth and death data is presented. This model, descending from the Poisson Point Process, provides a very flexible way to handle several sampling schemes and missing data patterns. For these reasons, this model represents a more powerful tool as compared to the birth and death model presented in the previous chapter. The likelihood function for the Monza and Brianza sampling scheme and missing data pattern is obtained within this framework. The fit to the empirical data turns out to be extremely satisfactory as confirmed by the model diagnostic.

5.1 Poisson Point Process

The Poisson Point Process is a stochastic point process [37] widely used in many fields such as survival analysis, signal theory, queueing theory etc. to describe the occurrence of independent events over time. The Poisson Point Process belongs to the class of Counting Processes [38] and can be defined as the number of events $N(t)$ occurred in the time interval $[0, t]$. Let us briefly recall the main properties of this process.

(i) $N(0) = 0$

(ii) $N(t) \geq 0$

(iii) $N(t) \in \mathbb{N}\ \forall t$

(iv) $N(t) \geq N(s)$ for any $t \geq s$

(v) $\lim_{h\to 0} \frac{\mathrm{P}\,(N(t+h)-N(t)\geq 2)}{h} = 0.$

The Poisson Point Process in particular is characterized by the following distributional properties. After defining the *intensity function* $\lambda(t),\ t \geq 0$ and the *mean function*

$$m(t) = \int_0^t \lambda(s)\,\mathrm{d}s,\ t > 0, \tag{5.1}$$

then

(i) the number of events in disjoint intervals are independent random variables

(ii) the number of events occurring in the $(t, t+s]$ interval is a Poisson random variable with parameter $m(t+s) - m(t)$

(iii) $\lim_{h\to 0} \frac{\mathrm{P}\,(N(t+h)-N(t)=1)}{h} = \lambda(t).$

.

When $\lambda(t) = \lambda,\ t \geq 0$ the process is called *homogeneous Poisson process*, and in that case several additional properties hold

(iv) increments are stationary i.e. the probability distribution of the number of occurrences counted in any time interval depends only on the length of the interval

(v) the probability distribution of the length of the time interval between two events is exponential with parameter λ.

5.1.1 Simulation of the Poisson Point Process

While simulating the homogeneous Poisson process is straightforward, through the simulation of a sequence of Exponential random variables by (v) above, that is not the case for the nonhomogeneous case. Several algorithms have been proposed for simulating the Poisson process in this case. Having defined $\Lambda(t) = \int_0^t \lambda(s)\,\mathrm{d}s$ as the *intensity measure*, an algorithm belonging to the class of the *inversion methods* can be implemented, through the following result [39]:

Theorem 5.1.1. *Let $\Lambda(t)$, $t \geq 0$ be a non negative valued, continuous, nondecreasing function. Then the random variables $T_1, T_2, \dots$ are event times corresponding to a nonhomogeneous Poisson process with intensity measure $\Lambda(t)$ if and only if $\Lambda(T_1), \Lambda(T_2), \dots$ are the event times corresponding to a homogeneous Poisson process with rate 1.*

This theorem provides a simple method to generate a nonhomogeneous Poisson process: first, an homogeneous Poisson process is simulated, then the event times are transformed via the inverse of $\Lambda(\cdot\,)$. The efficiency of the Cinlar's method is strictly connected with the ease of the inversion of the intensity measure.

An alternative simulation technique bases on *order statistics* descends by the following theorem [40]

Theorem 5.1.2. *Let $T_1, T_2, \dots$ be random variables representing the event times of a nonhomogeneous Poisson process having intensity measure $\Lambda(t)$ and let $N(t)$ be the corresponding counting process as defined above. Then, conditional on the number of events $N(t_0) = n$, the event times $T_1, T_2, \dots, T_n$ are distributed as order statistics from a sample with distribution function $F(t) = \frac{\Lambda(t)}{\Lambda(t_0)}$ for $t \in [0, t_0]$.*

Even in this case the simulation is simple: the number of events n occurring in $[0, t_0]$ is drawn from a $Poisson(\Lambda(t_0))$ distribution, then n events $t_1, t_2, \dots, t_n$ are generated by the distribution $F(t) = \frac{\Lambda(t)}{\Lambda(t_0)}$. At last take the ordered events $t_{(1)}, t_{(2)}, \dots, t_{(n)}$ as the event times of the nonhomogeneous Poisson process on $[0, t_0]$. This algorithm depends critically on the ease of the random number generation from the cdf $F(t)$.

However, the most popular method for generating nonhomogeneous Poisson processes is the *acceptance-rejection* method called *thinning* [41].

Theorem 5.1.3. *Consider a nonhomogeneous Poisson process with intensity function* $\lambda_u(t)$, $t \geq 0$. *Suppose that* $T_1^*, T_2^*, \ldots, T_n^*$ *are random variables representing event times for such process over the interval* $[0, t_0]$. *Let* $\lambda(t)$, $t \geq 0$ *be another intensity functions such that* $0 \leq \lambda(t) \leq \lambda_u(t)$ $\forall t \in [0, t_0]$. *Then if each* i*th event time* T_i^* *is independently rejected with probability* $1 - \lambda(T_i^*)/\lambda_u(T_i^*)$ *then the retained event times follow a nonhomogeneous Poisson process with intensity function* $\lambda(t)$ *over the interval* $[0, t_0]$.

In practice, a constant dominating function $\lambda_u(t) = \lambda_u$ is usually chosen, leading to the acceptance probability $\lambda(t)/\lambda_u$. The efficiency of this method critically depends on how the dominating function $\lambda_u(t)$ closely fits the intensity function $\lambda(t)$. High rejectance rates leads in general to inefficient simulations. To overcome this issue, a piecewise constant modification for the dominating function has been proposed in [42].

5.2 The Lexis Diagram and the Lexis Point Process

The Lexis Diagram is a widespread tool for representing event times in Survival Analysis. Lifetimes are represented by segments of unit slope connecting the birth time point $(\sigma, 0)$ to the death time $(t, x) = (\sigma + x, x)$ being x the age at death. In [43] the history of this diagram and the inference on point process defined by the death points (t, x) is discussed. A first representation of the death process according to a Poisson point process can be found in [44]. In particular, births are assumed to follow a nonhomogeneous Poisson process having intensity function $\varphi(\sigma) = \varphi(t - x)$, while life lengths are assumed to be (conditionally) independent random variables described by the death intensity (hazard rate) $\mu(t, x)$ defined as

$$\mathrm{P}\left(x \leq X < x + h | X \geq x\right) \approx \mu(t, x)h, \tag{5.2}$$

where X is the life length of an individual born at time σ and h is small enough. Under these assumptions, the bivariate point process of deaths in the Lexis diagram can be shown to be a Poisson process having intensity function

$$\begin{aligned}\lambda(t,x) &= \varphi(t-x)\mu(t,x)\exp\left\{-\int_0^x \mu(t-x+y,y)\,\mathrm{d}y\right\}\\ &= \varphi(\sigma)\mu(\sigma+x,x)\exp\left\{-\int_0^x \mu(\sigma+y,y)\,\mathrm{d}y\right\} = \lambda_\sigma(t). \qquad (5.3)\end{aligned}$$

The nonparametric estimation of the intensity function and extensions in presence of morbidity are described in [43]. From a missing data point of view it is interesting to consider how data selection influences inference on the point process. In [45] several different sampling patterns on lifetimes are considered, and in particular the *time window* case i.e., the pattern including all deaths in the set $[t_1, t_2] \times \mathbb{R}_0$. Note that this case is common in register based studies such as the Monza and Brianza firms lifetimes. However, as discussed earlier, these data present an additional source of missingness acting on birth times.

A modification of the point process on the Lexis diagram can be constructed. Let $\eta = \sum_{i\in I}\varepsilon_{\sigma_i}$, where ε_{σ_i} is the Dirac measure at σ_i, be the counting measure associated to the birth process $(\sigma_i)_{i\in I}$ having intensity function $\varphi(\sigma)$ and corresponding intensity measure $\Phi(B) = \int_B \varphi(\sigma)\,\mathrm{d}\sigma = \mathrm{E}\eta(B) < \infty$ with B a Borel set $B \in \mathcal{B}$. Let the lifetimes $(X_i)_{i\in I}$ be independent conditionally on the process η of birth times. Also let P be a Markov kernel $(\mathbb{R},\mathcal{B}) \to (\mathbb{R}_0,\mathcal{B}_0)$ describing the distribution of X_i given σ_i. Then $\mathrm{P}\,(\sigma,\cdot)$ or alternatively $\mathrm{P}_\sigma(\cdot)$ is a probability measure on $(\mathbb{R}_0,\mathcal{B}_0)$ for all $\sigma\in\mathbb{R}$, and define F_σ and $\bar{F}_\sigma$ as the corresponding distribution function and survival function. A sample realization of this process is shown in Fig. 5.1.

The following results have great importance in the remainder of this work.

Definition 5.2.1. The *Lexis point process* is the point process $\mu = \sum_{i\in I}\varepsilon_{(\sigma_i,X_i)}$ on $(\mathbb{R}\times\mathbb{R}_0, \mathcal{B}\otimes\mathcal{B}_0)$

A theorem, known as *positioning dependent marking* [46] states that when the birth

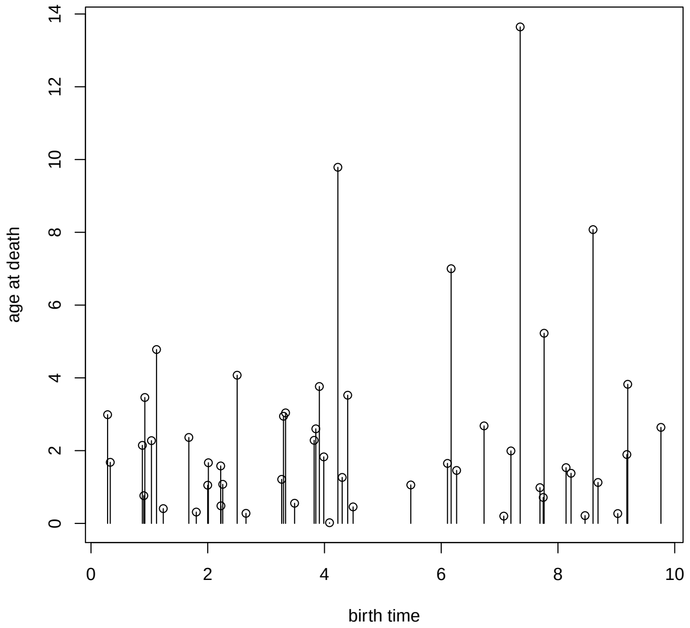

Figure 5.1: Lexis point process: homogeneous Poisson birth process and $Exp(1/3)$ lifetime

process is a Poisson point process then the Lexis point process is a (planar) Point process itself.

Theorem 5.2.1. *Let η be a Poisson process on $(E, \mathcal{E})$ having intensity measure Φ and let P be a Markov kernel $(E, \mathcal{E}) \to (G, \mathcal{G})$. Assume that, given $\eta = \sum_{i \in I} \varepsilon_{\sigma_i}$, $(X_i)_{i \in I}$ is a family of independent random variables, and that each X_i has distribution $\mathrm{P}\,(\sigma_i, \cdot)$ on $(G, \mathcal{G})$ for all $i \in I$. Then the resulting Lexis point process $\mu = \sum_{i \in I} \varepsilon_{(\sigma_i, X_i)}$ is a Poisson process on $(E \times G, \mathcal{E} \otimes \mathcal{G})$ with intensity measure*

$$\Lambda(A \times B) = \int_A \mathrm{P}\,(\sigma, B)\Phi(\,\mathrm{d}\sigma) \tag{5.4}$$

for any $A \in \mathcal{E}$ and $B \in \mathcal{G}$.

According to Theorem 5.2.1 then the Lexis point process μ has intensity $\lambda(\sigma, x) = \varphi(\sigma) f_\sigma(x)$ being $f_\sigma(\cdot)$ the density w.r.t. the Lebesgue measure of $\mathrm{P}_{\,\sigma}(\cdot)$.

As the intensity function has now been defined, the related distribution and likelihood have to be determined in order to perform inference on the data. The following fundamental results can be found in [37]. Let μ be a Poisson process on $(S, \mathcal{S})$ with intensity measure Λ, consider a set $A \in \mathcal{S}$ such that $0 < \Lambda(A) < \infty$. Then, given $\mu(A) = n$, the points $(y_i)_{i \in I} = (\sigma_i, x_i)_{i \in I}$ of μ on A are i.i.d. with distribution $\pi(\cdot) = \Lambda(\cdot \cap A)/\Lambda(A)$. Let μ have intensity λ. Then the likelihood for the observations $(y_i)_{i \in I}$ of μ on A is

$$L = \exp(-\Lambda(A)) \prod_{i \in I} \lambda(y_i), \tag{5.5}$$

and considering that the probability of having no points in an infinitesimal set $a \in A$ is $\exp(-\Lambda(a))$, then $\exp(-\Lambda(A))$ represent the likelihood term for all the points in the set A in which no events happened, recalling that $\mathrm{e}^{-\lambda}$ is the probability that 0 events occur for a Poisson random variable with parameter λ. Then, this factor controls for those individuals whose death was not observed because not included in the considered set A.

5.2.1 Time Window Sampling Scheme

A particular sampling scheme often arising in register based studies is the *time window* scheme. In this particular situation, only deaths occurred in the time period $[t_1, t_2]$ are reported, as shown for example in Fig. 5.2. The area of interest described by this sampling scheme is

$$C = \{(\sigma, x) : \max(0, t_1 - \sigma) < x < \max(0, t_2 - \sigma),\ \ 0 < \sigma < t_2\}, \tag{5.6}$$

which can be decomposed as

$$\begin{aligned} C = A \cup B = &\{(\sigma, x) : t_1 - \sigma < x < t_2 - \sigma,\ \ \sigma < t_1\} \cup \\ &\{(\sigma, x) : 0 < x < t_2 - \sigma,\ \ t_1 \leq \sigma < t_2\} \end{aligned} \tag{5.7}$$

Once assumptions are made on the birth process and on the time to death density, the likelihood in (5.5) can be obtained explicitly. Having defined a suitable time origin, births can be thought of realizations of a Poisson point process having intensity function

$$\varphi(\sigma) = \alpha \mathrm{e}^{\beta\sigma} \tag{5.8}$$

and associated intensity measure

$$\Phi(\sigma) = \int_0^\sigma \alpha \mathrm{e}^{\beta t}\, \mathrm{d}t = \frac{\alpha}{\beta}\mathrm{e}^{\beta\sigma} - \frac{\alpha}{\beta}. \tag{5.9}$$

Assuming that lifetimes follow an exponential distribution

$$f_\sigma(x) = \lambda \mathrm{e}^{-\lambda x} \tag{5.10}$$

then the intensity function of the Lexis process follows easily as

$$\lambda_\sigma(x) = \alpha\lambda \mathrm{e}^{\beta\sigma - \lambda x}. \tag{5.11}$$

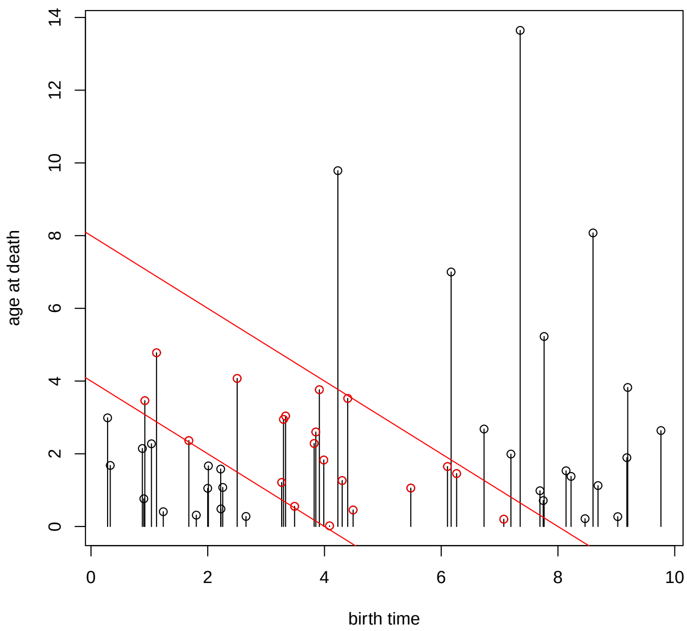

Figure 5.2: Lexis point process: time window on the $[4, 8]$ interval.

Now that the intensity function is explicit it is possible to determine the expected value and the variance of the number of deaths for the time window, that is $\Lambda(C) = \Lambda(A)+\Lambda(B)$. This value will enter in the survival control factor $\exp(-\Lambda(C))$. The first term is

$$\begin{aligned}
\Lambda(A) =& \int_0^{t_1} \int_{t_1-\sigma}^{t_2-\sigma} \lambda_\sigma(x)\,\mathrm{d}x\,\mathrm{d}\sigma = \alpha\lambda \int_0^{t_1} \mathrm{e}^{\beta\sigma} \left[\int_{t_1-\sigma}^{t_2-\sigma} \mathrm{e}^{-\lambda x}\,\mathrm{d}x \right] \mathrm{d}\sigma \\
=&\alpha \int_0^{t_1} \mathrm{e}^{\beta\sigma} \left[\mathrm{e}^{-\lambda(t_1-\sigma)} - \mathrm{e}^{-\lambda(t_2-\sigma)} \right] \mathrm{d}\sigma \\
=&\alpha \left[\mathrm{e}^{-\lambda t_1} \int_0^{t_1} \mathrm{e}^{(\beta+\lambda)\sigma}\,\mathrm{d}\sigma - \mathrm{e}^{-\lambda t_2} \int_0^{t_1} \mathrm{e}^{(\beta+\lambda)\sigma}\,\mathrm{d}\sigma \right] \\
=&\alpha \left[\int_0^{t_1} \mathrm{e}^{(\beta+\lambda)\sigma}\,\mathrm{d}\sigma \right] \left[\mathrm{e}^{-\lambda t_1} - \mathrm{e}^{-\lambda t_2} \right] \\
=&\frac{\alpha}{\beta+\lambda} \left[\mathrm{e}^{t_1(\beta+\lambda)} - 1 \right] \left[\mathrm{e}^{-\lambda t_1} - \mathrm{e}^{-\lambda t_2} \right], \qquad (5.12)
\end{aligned}$$

and the second term is

$$\begin{aligned}
\Lambda(B) =& \int_{t_1}^{t_2} \int_0^{t_2-\sigma} \lambda_\sigma(x)\,\mathrm{d}x\,\mathrm{d}\sigma = \alpha\lambda \int_{t_1}^{t_2} \mathrm{e}^{\beta\sigma} \left[\int_0^{t_2-\sigma} \mathrm{e}^{-\lambda x}\,\mathrm{d}x \right] \mathrm{d}\sigma \\
=&\alpha \int_{t_1}^{t_2} \mathrm{e}^{\beta\sigma} \left[1 - \mathrm{e}^{-\lambda(t_2-\sigma)} \right] \mathrm{d}\sigma \\
=&\alpha \left[\int_{t_1}^{t_2} \mathrm{e}^{\beta\sigma}\,\mathrm{d}\sigma - \mathrm{e}^{-\lambda t_2} \int_{t_1}^{t_2} \mathrm{e}^{(\beta+\lambda)\sigma}\,\mathrm{d}\sigma \right] \\
=&\frac{\alpha}{\beta} \left[\mathrm{e}^{\beta t_2} - \mathrm{e}^{\beta t_1} \right] - \frac{\alpha}{\beta+\lambda} \mathrm{e}^{-\lambda t_2} \left[\mathrm{e}^{(\beta+\lambda)t_2} - \mathrm{e}^{(\beta+\lambda)t_1} \right] \\
=&\frac{\alpha}{\beta} \left[\mathrm{e}^{\beta t_2} - \mathrm{e}^{\beta t_1} \right] - \frac{\alpha}{\beta+\lambda} \left[\mathrm{e}^{\beta t_2} - \mathrm{e}^{(\beta+\lambda)t_1-\lambda t_2} \right]. \qquad (5.13)
\end{aligned}$$

Adding the two terms finally leads to

$$\begin{aligned}\Lambda(C) =&\Lambda(A)+\Lambda(B)=\frac{\alpha}{\beta+\lambda}\left[\mathrm{e}^{\beta t_1}-\mathrm{e}^{(\beta+\lambda)t_1-\lambda t_2}+\mathrm{e}^{-\lambda t_2}-\mathrm{e}^{-\lambda t_1}\right]\\ &+\frac{\alpha}{\beta}\left[\mathrm{e}^{\beta t_2}-\mathrm{e}^{\beta t_1}\right]-\frac{\alpha}{\beta+\lambda}\left[\mathrm{e}^{\beta t_2}-\mathrm{e}^{(\beta+\lambda)t_1-\lambda t_2}\right]\\ =&\frac{\alpha}{\beta+\lambda}\left[\mathrm{e}^{-\lambda t_2}-\mathrm{e}^{-\lambda t_1}\right]+\mathrm{e}^{\beta t_1}\left[\frac{\alpha}{\beta+\lambda}-\frac{\alpha}{\beta}\right]\\ &+\mathrm{e}^{\beta t_2}\left[\frac{\alpha}{\beta}-\frac{\alpha}{\beta+\lambda}\right]\\ =&\alpha\left[\frac{1}{\beta+\lambda}\left[\mathrm{e}^{-\lambda t_2}-\mathrm{e}^{-\lambda t_1}\right]+\left[\frac{1}{\beta}-\frac{1}{\beta+\lambda}\right]\left[\mathrm{e}^{\beta t_2}-\mathrm{e}^{\beta t_1}\right]\right]. \qquad (5.14)\end{aligned}$$

It is clear from 5.14 that α represents a scale parameter as it controls the initial birth intensity. The likelihood contribution from the observed deaths is

$$\begin{aligned}\prod_{i=1}^{n}\lambda_{\sigma_i}(x_i)=&\prod_{i=1}^{n}\alpha\lambda\mathrm{e}^{\beta\sigma_i-\lambda x_i}=\\ &=(\alpha\lambda)^n\exp\left\{\beta\sum_{i=1}^{n}\sigma_i\right\}\exp\left\{-\lambda\sum_{i=1}^{n}x_i\right\} \qquad (5.15)\end{aligned}$$

Recalling (5.5), the log-likelihood for the time window sampling scheme is

$$\begin{aligned}l(\alpha,\beta,\lambda;\sigma,\mathbf{x})=&\sum_{i=1}^{n}\log\lambda_{\sigma_i}(x_i)-\Lambda(C)=n[\log\alpha+\log\lambda]+\\ &+\beta\sum_{i=1}^{n}\sigma_i-\lambda\sum_{i=1}^{n}x_i-\frac{\alpha}{\beta+\lambda}\left[\mathrm{e}^{-\lambda t_2}-\mathrm{e}^{-\lambda t_1}\right]+\\ &-\mathrm{e}^{\beta t_1}\left[\frac{\alpha}{\beta+\lambda}-\frac{\alpha}{\beta}\right]-\mathrm{e}^{\beta t_2}\left[\frac{\alpha}{\beta}-\frac{\alpha}{\beta+\lambda}\right] \qquad (5.16)\end{aligned}$$

In order to take into account the missingness in the birth times it is possible to focus the analysis over sets of the plane (σ, x) in which the single values of the variables birth time σ and age at death x are irrelevant and for example where only the sum of both (death time) matters. If, for example, we consider a situation where only death times are registered, an *indifference set* $C_k(t),\ t \in (t_1+k, t_2)$ can be constructed as

$$C_k(t) = \{(\sigma, x) : \sigma \in (0,t), \max(0, t-\sigma-k) \leq x < t-\sigma\}. \tag{5.17}$$

the likelihood function within this framework is easy to obtain since the observed count for each of the $(t_2 - t_1)/k$ indifference sets is distributed as a Poisson random variable and the parameter can be easily determined by (5.14) as being

$$\begin{aligned}\Lambda(C_k(t)) =& \mathrm{E}_k(t) = \mathrm{Var}_k(t) = \alpha\left[\frac{1}{\beta+\lambda}\left[\mathrm{e}^{-\lambda t} - \mathrm{e}^{-\lambda(t-k)}\right] + \right. \\ & \left.\left[\frac{1}{\beta} - \frac{1}{\beta+\lambda}\right]\left[\mathrm{e}^{\beta t} - \mathrm{e}^{\beta(t-k)}\right]\right].\end{aligned} \tag{5.18}$$

Considering only the death times reduces the dimensionality of the problem from the two variables σ and x to the single sum of both $t = \sigma + x$, and thus a one dimensional Poisson point process can be studied. Its intensity can be obtained by considering the integrated intensity over the one dimensional collapsed version of the set in (5.17)

$$C_0(t) = \{(\sigma, x) : \sigma \in (0,t), x = t - \sigma\}. \tag{5.19}$$

The intensity function of the nonhomogeneous Poisson process of the death times is then

$$\begin{aligned}\lambda_0(t) =& \int_{C_0(t)} \alpha\lambda \mathrm{e}^{\beta\sigma - \lambda x}\,\mathrm{d}\sigma\,\mathrm{d}x = \alpha\lambda\mathrm{e}^{-\lambda t}\int_0^t \mathrm{e}^{(\beta+\lambda)\sigma}\,\mathrm{d}\sigma = \\ =& \frac{\alpha\lambda}{\beta+\lambda}\mathrm{e}^{-\lambda t}\left[\mathrm{e}^{(\beta+\lambda)t} - 1\right] = \frac{\alpha\lambda}{\beta+\lambda}\left[\mathrm{e}^{\beta t} - \mathrm{e}^{-\lambda t}\right].\end{aligned} \tag{5.20}$$

Since the survival control factor is still defined over the set C, the log-likelihood function for the case of death times only over the time window is the following:

$$\begin{aligned}l(\alpha,\beta,\lambda;\mathbf{t}) =& \sum_{i=1}^n \log\lambda_0(t_i) - \Lambda(C) = n[\log\alpha + \log\lambda - \log(\beta+\lambda)] + \\ & + \sum_{i=1}^n \log\left[\mathrm{e}^{\beta t_i} - \mathrm{e}^{-\lambda t_i}\right] - \frac{\alpha}{\beta+\lambda}\left[\mathrm{e}^{-\lambda t_2} - \mathrm{e}^{-\lambda t_1}\right] + \\ & - \mathrm{e}^{\beta t_1}\left[\frac{\alpha}{\beta+\lambda} - \frac{\alpha}{\beta}\right] - \mathrm{e}^{\beta t_2}\left[\frac{\alpha}{\beta} - \frac{\alpha}{\beta+\lambda}\right].\end{aligned} \tag{5.21}$$

5.2.2 Piecewise Constant Intensities

Within this framework it is possible to consider a partition on the birth time axis σ in order to introduce piecewise constant time varying intensities. Individuals are then divided into K cohorts identified for simplicity by the sequence of $K+1$ points $\{k_0 \equiv 0, k_1, \dots, k_{K-1} \equiv t_1, k_K \equiv t_2\}$. In the generic j-th cohort, birth and death intensities are assumed to be constant and equal to β_j and λ_j respectively, while the survival control factor can be easily obtained by (5.12) as

$$\begin{aligned}\Lambda(A_j) = \alpha \left[\int_{k_{j-1}}^{k_j} \mathrm{e}^{(\beta_j+\lambda_j)\sigma}\, \mathrm{d}\sigma\right]\left[\mathrm{e}^{-\lambda_j t_1} - \mathrm{e}^{-\lambda_j t_2}\right] = \\ \frac{\alpha}{\beta_j+\lambda_j}\left[\mathrm{e}^{(\beta_j+\lambda_j)k_j} - \mathrm{e}^{(\beta_j+\lambda_j)k_{j-1}}\right]\left[\mathrm{e}^{-\lambda_j t_1} - \mathrm{e}^{-\lambda_j t_2}\right]\end{aligned} \tag{5.22}$$

for $j = 1, \dots, K-1$.

The last cohort window is, by construction, $A_K \equiv B$. Thus

$$\begin{aligned}\Lambda(A_K) \equiv \Lambda(B) =& \frac{\alpha}{\beta_K}\left[\mathrm{e}^{\beta k_K} - \mathrm{e}^{\beta k_{K-1}}\right] - \\ & \frac{\alpha}{\beta_K+\lambda_K}\left[\mathrm{e}^{\beta_K k_K} - \mathrm{e}^{(\beta_K+\lambda_K)k_{K-1}-\lambda_K k_K}\right]\end{aligned} \tag{5.23}$$

Define for each cohort the index set $\mathcal{I}_j = \{i : \sigma_i \in A_j\}$ including the individuals belonging to it, and its corresponding cardinality $n_j = \#\mathcal{I}_j$. Then the likelihood term for the observed deaths is

$$\begin{aligned}\prod_{j=1}^{K}\prod_{i\in\mathcal{I}_j} \lambda_{\sigma_i}^{(j)}(x_i) = \prod_{j=1}^{K}\prod_{i\in\mathcal{I}_j} \alpha\lambda_j \mathrm{e}^{\beta_j\sigma_i - \lambda_j x_i} = \\ \alpha^n\left[\prod_{j=1}^{K}\lambda_j^{n_j}\right]\exp\left\{\sum_{j=1}^{K}\beta_j\sum_{i\in\mathcal{I}_j}\sigma_i - \sum_{j=1}^{K}\lambda_j\sum_{i\in\mathcal{I}_j}x_i\right\}.\end{aligned} \tag{5.24}$$

The log-likelihood for the piecewise constant intensities under the time window sampling scheme is finally obtained as being

$$
\begin{aligned}
l(\alpha,\beta,\lambda;\sigma,\mathbf{x}) = {} & n\log(\alpha) + \sum_{j=1}^{K} n_j \log(\lambda_j) + \sum_{j=1}^{K} \beta_j \sum_{i\in\mathcal{I}_j} \sigma_i - \sum_{j=1}^{K} \lambda_j \sum_{i\in\mathcal{I}_j} x_i \\
& - \sum_{j=i}^{K-1} \frac{\alpha}{\beta_j+\lambda_j} \left[\mathrm{e}^{(\beta_j+\lambda_j)k_j} - \mathrm{e}^{(\beta_j+\lambda_j)k_{j-1}} \right] \left[\mathrm{e}^{-\lambda_j k_{K-1}} - \mathrm{e}^{-\lambda_j k_K} \right] \\
& - \frac{\alpha}{\beta_K} \left[\mathrm{e}^{\beta_K k_K} - \mathrm{e}^{\beta_K k_{K-1}} \right] + \frac{\alpha}{\beta_K+\lambda_K} \left[\mathrm{e}^{\beta_K k_K} - \mathrm{e}^{(\beta_K+\lambda_K)k_{K-1}-\lambda_K k_K} \right]. \qquad (5.25)
\end{aligned}
$$

Inference can then proceed as usual by maximization of (5.25).

5.2.3 Monza and Brianza: Mortality in the Construction Firms

The Lexis point process is a powerful framework for analyzing mortality patterns and in particular sampling scheme situations such as the Monza and Brianza business dataset presented in section 4.3. We illustrate the application of the model and in particular we focus on the subset of the construction firms has been considered.

The complete set is composed by 4,450 observations, and the year of death ranges, as for the complete dataset over the period 1996-2006. However, it is not clear how the death times were selected in the year 1996: as can be seen in Fig. 5.3 and Fig. 5.4 the observations are clearly incomplete. Note that within this framework, excluding firms which were born and died in 1996 means to narrow by one year the death time window, that becomes 1997-2006, and to exclude the 1996 cohort. The area on the (σ, x) plane considered and the survival control factor, will then change consequently. From Fig. 5.3 it seems clear that the cohorts experiencing death within the considered time windows were those starting from the 1950 year, the time origin for the birth process.

These considerations lead us to discard:

- 1 firm born in 1939
- 42 firms born in 1996
- 1 firm born in 1998

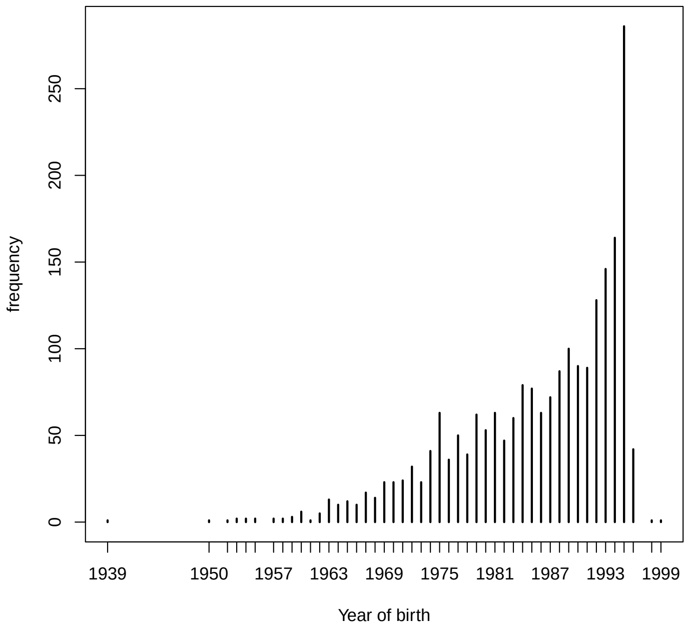

Figure 5.3: Monza and Brianza construction firms: year of birth

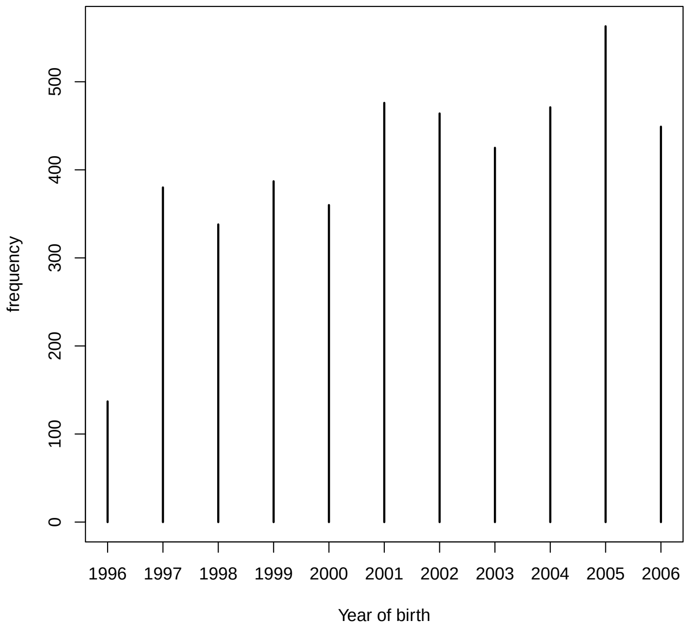

Figure 5.4: Monza and Brianza construction firms: year of death

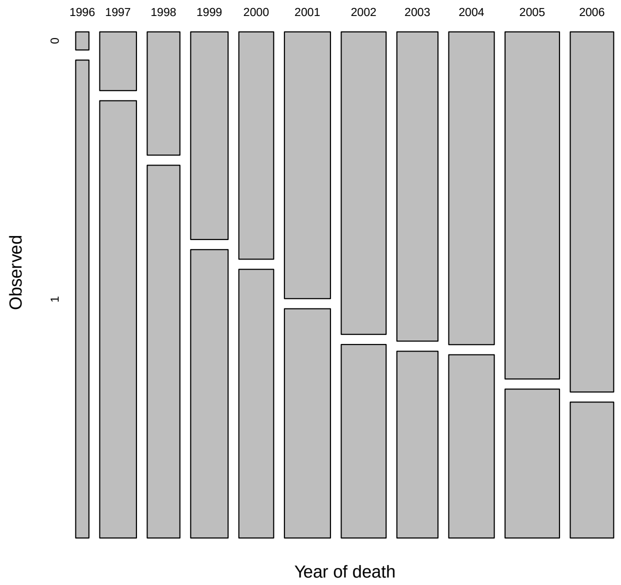

Figure 5.5: Monza and Brianza construction firms: observed proportion by year of death

- 1 firm born 1999
- 137 firms died in 1996

resulting in a final sample size $n = 4,268$. Because of the exponentially decreasing proportion of the observed birth times, (see Fig. 5.5) and of the exponentially distributed lifetime, missing births times were assumed to occur during the time window 1997-2006.

Firms have then been grouped in the following $K = 5$ birth cohorts

- 122 born in [1950-1970)
- 372 born in [1970-1980)
- 662 born in [1980-1990)
- 835 born in [1990-1996)

- 2277 born in [1997-2007)

Since the time origin has been set on 1950, the set of points that define the cohort partitions is $\mathbf{k} = \{k_0 = 0, k_1 = 20, k_2 = 30, k_3 = 40, k_4 = 46, k_5 = 47, k_6 = 57\}$. The deaths in the time window considered, that correspond to birth times in 1996 are those that have been discarded; the set A_5 then will not be included in the log-likelihood. In time window A_6 only death times are recorded, then the point process will be considered unidimensional as in (5.20). The integral in this case will range from $k_{K-1} = t_1$ to t.

Define the set of the considered cohorts as $\mathcal{C} = \{1, 2, 3, 4, 6\}$, it is now possible to define the log-likelihood

$$\begin{aligned}
l(\alpha, \beta, \lambda; \sigma, \mathbf{x}) &= n \log(\alpha) + \sum_{j \in \mathcal{C}} n_j \log(\lambda_j) - n_K \log(\beta_K + \lambda_K) + \\
&\sum_{i \in \mathcal{I}_K} \log(\mathrm{e}^{\beta_K t_i} - \mathrm{e}^{\beta_K k_{K-1} - \lambda_K (t_i - k_{K-1})}) + \sum_{j=1}^{K-2} \beta_j \sum_{i \in \mathcal{I}_j} \sigma_i - \sum_{j=1}^{K-2} \lambda_j \sum_{i \in \mathcal{I}_j} x_i + \\
&- \sum_{j=i}^{K-2} \frac{\alpha}{\beta_j + \lambda_j} \left[\mathrm{e}^{(\beta_j + \lambda_j) k_j} - \mathrm{e}^{(\beta_j + \lambda_j) k_{j-1}} \right] \left[\mathrm{e}^{-\lambda_j k_{K-1}} - \mathrm{e}^{-\lambda_j k_K} \right] + \\
&- \frac{\alpha}{\beta_K} \left[\mathrm{e}^{\beta_K k_K} - \mathrm{e}^{\beta_K k_{K-1}} \right] + \frac{\alpha}{\beta_K + \lambda_K} \left[\mathrm{e}^{\beta_K k_K} - \mathrm{e}^{(\beta_K + \lambda_K) k_{K-1} - \lambda_K k_K} \right] \qquad (5.26)
\end{aligned}$$

The log-likelihood in (5.26) has been maximized (code included in Appendix B). The parameter α represents the initial (1950) intensity of the birth process. The two vectors of parameters β and λ represent the birth intensity and the force of mortality experienced by the cohort identified. It is important to remind that force of mortality was defined as cohort-wise for the conditional independence assumption to hold. The results of estimation are reported in the Table 5.1

The estimates show that the second and the third cohorts [1970,1990) experienced a low mortality, especially if compared to that experienced in the following considered period. As for the birth intensity, it has been quite stable apart from the higher rate of the first period (right after World War II).

Table 5.1: Monza and Brianza construction firms: Estimation results for the Lexis point process parameters

	Estimate	Std. Error	z value	Pr(z)
α	132.8800	0.0000	1.58E+07	<2.2E-16
λ_1	0.1114	0.0079	1.41E+01	<2.2E-16
λ_2	0.0479	0.0122	3.94E+00	8.167E-05
λ_3	0.0473	0.0098	4.81E+00	1.485E-06
λ_4	0.1610	0.0108	1.49E+01	<2.2E-16
λ_6	0.2173	0.0260	8.34E+00	<2.2E-16
β_1	0.0850	0.0136	6.27E+00	3.498E-10
β_2	0.0290	0.0030	9.65E+00	<2.2E-16
β_3	0.0235	0.0018	1.28E+01	<2.2E-16
β_4	0.0202	0.0008	2.45E+01	<2.2E-16
β_6	0.0209	0.0012	1.68E+01	<2.2E-16

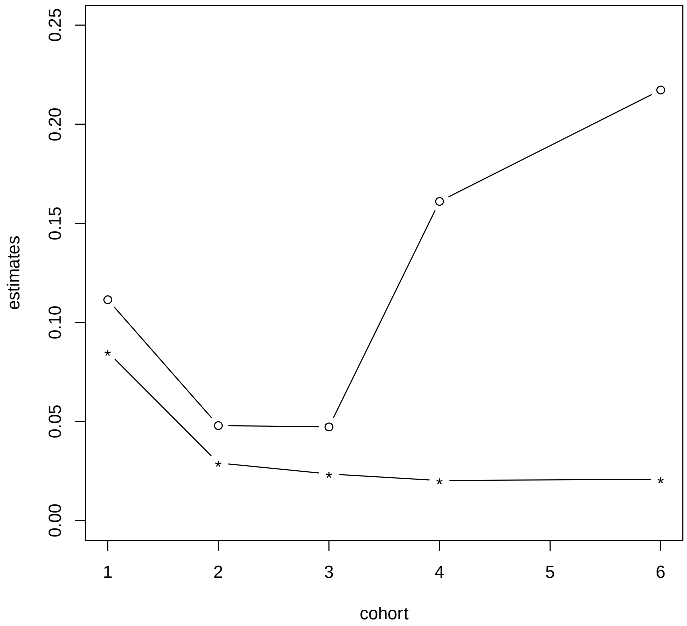

Figure 5.6: Monza and Brianza construction firms: $\hat{\beta}$ (star) and $\hat{\lambda}$ (circle) estimates

Table 5.2: Monza and Brianza construction firms: expected versus observed deaths

Cohort	Expected deaths	Observed Deaths
1950-1970	120.5	122
1970-1980	373.4	372
1980-1990	659.7	662
1990-1996	837.6	835
1997-2007	2276.7	2277

5.2.4 Model Diagnostics

In order to assess the model goodness of fit we follow two different approaches. The first one is based on observed moments. Since $\Lambda(A_k)$ is equal to the expected number of deaths for the k-th cohort in the time window, it is possible to compare this expected value as estimated by the model $\Lambda(\hat{A}_k)$ with the number of deaths actually observed. For the Monza and Brianza dataset, estimates of the expected values were obtained by plugging fitted parameters presented in Table 5.1 into (5.12) and (5.13).

As shown in Table 5.2 the fit excellent. Note that expected values are functions of the process parameters, and that by the *functional invariance* and *consistency* properties of the MLEs, closeness equivalence between MLE based quantities and the observed moments should be expected to hold, but only when the model is correctly specified. The results are therefore very encouraging.

The second approach is based on simulation and it relies on Theorem 5.1.1. In a first stage, the complete data (latent) distribution previously estimated for the proposed model is simulated, then the selection is applied to produce observed data (i.e., time window and missingness inside the window itself will be applied on the simulated data). Finally, the survival functions of the observed and simulated complete data can be compared. Let us introduce the integrated intensity of the birth process over the j-th cohort

$$\Phi_j = \frac{\alpha}{\beta_j}\mathrm{e}^{\beta_j k_j} - \frac{\alpha}{\beta_j}\mathrm{e}^{\beta_j k_{j-1}} = \frac{\alpha}{\beta_j}\left[\mathrm{e}^{\beta_j k_j} - \mathrm{e}^{\beta_j k_{j-1}}\right] \tag{5.27}$$

and the related cumulative sum

$$\Psi_c = \sum_{i=1}^{c-1} \Phi_i, \ \Psi_1 := 0 \tag{5.28}$$

As for Theorem 5.1.1 the first set of values $\tilde{\sigma}$ will be a realization of an homogeneous Poisson point process with parameter 1. Since the simulated birth times σ will be obtained by the inversion of the intensity measure $\Phi(\cdot)$, values will be simulated until the limit Ψ_{K+1} is reached. Each $\tilde{\sigma}_i$ value will then be assigned to the c-th cohort if

$$\Psi_c < \tilde{\sigma}_i \leq \Psi_{c+1} \tag{5.29}$$

In order to obtain a realization from the required stepwise non homogeneous Poisson process the inversion of the intensity function is required. Each σ_i birth time is obtained as

$$\begin{aligned}
\tilde{\sigma}_i =& \Psi_c + \frac{\alpha}{\beta_c}\mathrm{e}^{\beta_c \sigma_i} - \frac{\alpha}{\beta_c}\mathrm{e}^{\beta_c k_{c-1}} \\
\mathrm{e}^{\beta_c \sigma_i} =& \frac{\beta_c}{\alpha}[\tilde{\sigma}_i - \Psi_c] + \mathrm{e}^{\beta_c + k_{c-1}} \\
\sigma_i =& \frac{1}{\beta_c}\log\left[\frac{\beta_c}{\alpha}[\tilde{\sigma}_i - \Psi_c] + \mathrm{e}^{\beta_c k_{c-1}}\right]
\end{aligned}$$

The age at death of each individual is then simulated from an Exponential distribution with the parameter correspondent to its assigned cohort. Finally, only individuals whose death happened within the time window are retained. For the Monza and Brianza dataset, only the first 4 cohorts were simulated. The 5-th cohort was not considered in the model, and the 6-th one was the one including all the missing birth times. Therefore, it would not contribute to the complete data survival function (see Appendix B for R code).

Fig 5.7 shows the comparison between the survival functions of the observed and the simulated complete data i.e., individuals with both birth and death times reported. The

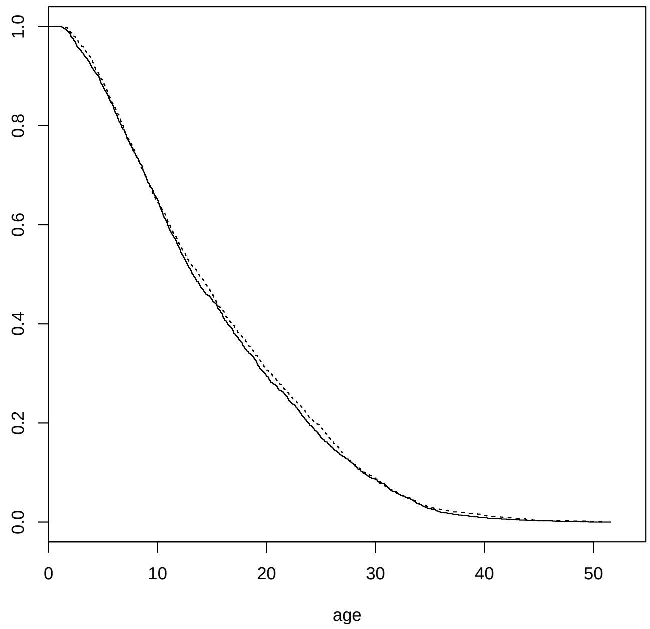

Figure 5.7: Monza and Brianza construction firms: observed (solid) vs simulated (dashed) complete data Survival function

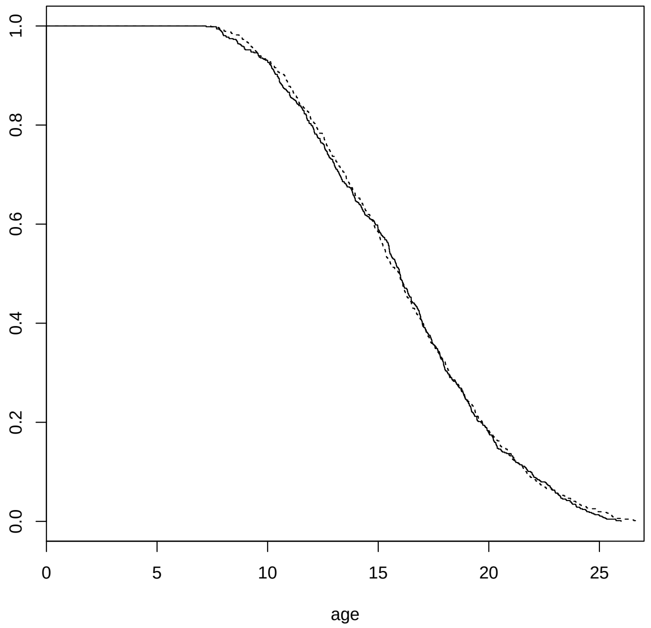

Figure 5.8: Monza and Brianza construction firms: 3rd cohort observed (solid) vs simulated (dashed) complete data Survival function

fit is very satisfactory. Fig. 5.8 shows the same survival functions on a data subset, the 3rd cohort, and also in this case the fit is satisfactory. Note that the flat beginning of both functions is clearly due to the time window sampling scheme.

Chapter 6

Conclusion

A great advantage of the two models presented, with respect to the other methods presented in Chapter 2, is that there is no loss of information since all observations are retained in the analysis. However, compared to the complete likelihood approach, presented in Chapter 4, the Lexis point process approach provides a more flexible framework for handling several sampling and missing data patterns. In both cases, the likelihood function is determined analytically. Thus, all asymptotic MLE results for the parameters estimators hold. However, in the first case there was no control factor for the firms not dying within the time window, being excluded by the sampling scheme. In the second model the $\mathrm{e}^{-\Lambda(A)}$ factor, of the Lexis point process likelihood function, controls for the surviving firms.

As often happens in missing data problems, the missing data mechanism is unknown. This is the case of our analysed data, and since the correction factors and the modified unidimensional likelihood have been calibrated on the assumed sampling and missingness pattern, wrong assumptions could lead to bias in the results. In the Monza and Brianza data only birth times are missing, empirical evidence seems to suggest that these births are concentrated within the last cohort.

A cohort-wise analysis as the one that we have proposed can give an indication of the evolution of the mortality over time. As a limit of the proposed approach it has to be remarked that time varying coefficients cannot be defined because of the conditional independence of deaths on births requirement.

However, our model shows very encouraging fitting properties being the underlying time to event distribution a mixture of truncated exponential functions: birth intensity represents the weight of the mixture and conditionally on the realized birth time the life time variable is left-truncated because of the time window scheme. If an additional insight is desired on the birth process, possible extensions to this approach could include the use of different functions for the birth intensity (such as the Gompertz function) or a more flexible spline function approach. For the age at death distribution some other integrable densities could be feasible.

Appendix A

Kummer M Function

The *Kummer M Function* is the first of the two linearly independent solutions of *Kummer's equation*

$$z\frac{\mathrm{d}^2w}{\mathrm{d}z^2} + (b-z)\frac{\mathrm{d}w}{\mathrm{d}z} - aw = 0. \tag{A.1}$$

This equation has a regular singularity at $z = 0$ and an irregular singularity at ∞. This solution is also known with the name *Confluent Hypergeometric Function of the First Kind*, and is one of the *Generalized Hypergeometric Series* introduced in [47]:

$$M(a,b,z) = \sum_{n=0}^{\infty} \frac{(a)_n z^n}{(b)_n n!}, \tag{A.2}$$

with $(a)_n = a(a+1)\dots(a+n-1)$ the rising factorial and $(a)_0 = 1$. Other alternate notations are $\Phi(a,b,z)$ and ${}_1F_1(a,b,z)$. The second solution of (A.1) was studied by [48], it is usually indicated as $U(a,b,z)$ and can be expressed as a function of the first one:

$$U(a,b,z) = \frac{\pi}{\sin \pi b}\left\{\frac{M(a,b,z)}{\Gamma(1+a-b)\Gamma(b)} - z^{1-b}\frac{M(1+a-b,2-b,z)}{\Gamma(a)\Gamma(2-b)}\right\}. \tag{A.3}$$

The Kummer M function has the following useful integral representations:

$$\frac{\Gamma(b-a)\Gamma(a)}{\Gamma(b)}M(a,b,z) = \int_0^1 \mathrm{e}^{zt}t^{a-1}(1-t)^{b-a-1}\,\mathrm{d}t \tag{A.4}$$

$$\frac{\Gamma(b-a)\Gamma(a)}{\Gamma(b)}M(a,b,z) = 2^{1-b}\mathrm{e}^{\frac{1}{2}z}\int_{-1}^{+1}\mathrm{e}^{-\frac{1}{2}zt}(1+t)^{b-a-1}\theta \\ (1-t)^{a-1}\,\mathrm{d}t \tag{A.5}$$

$$\frac{\Gamma(b-a)\Gamma(a)}{\Gamma(b)}M(a,b,z) = 2^{1-b}\mathrm{e}^{\frac{1}{2}z}\int_{0}^{\pi}\mathrm{e}^{-\frac{1}{2}zt\cos\theta}\sin^{b-1}\theta \\ \cot^{b-2a}\left(\frac{1}{2}\right)\,\mathrm{d}\theta \tag{A.6}$$

$$\frac{\Gamma(b-a)\Gamma(a)}{\Gamma(b)}M(a,b,z) = \mathrm{e}^{-Az}\int_{A}^{B}\mathrm{e}^{zt}(t-A)^{a-1}(B-t)^{b-a-1}\,\mathrm{d}t \\ A = B-1. \tag{A.7}$$

Lastly, the following equations are known as *Kummer Transformations*:

$$M(a,b,z) = \mathrm{e}^{z}M(b-a,b,-z) \tag{A.8}$$

$$z^{1-b}M(1+a-b,2-b,z) = z^{1-b}\mathrm{e}^{z}M(1-a,2-b,-z) \tag{A.9}$$

Appendix B

R Code

R script for Lexis point process maximum likelihood estimation and model diagnostic (moments comparison and simulation) are shown below.

```
library(bbmle)

#
# Model fitting
#

# Poisson point process loglikelihood - time window and missing birth times

nllik<-function(a,l1,l2,l3,l4,l6,b1,b2,b3,b4,b6,n,k,t,S,SX){
nllik<- -(sum(n)*log(a)+n[1]*log(l1)+n[2]*log(l2)+n[3]*log(l3)+n[4]*log(l4)
+n[6]*log(l6)-n[6]*log(b6+l6)+ sum(log(exp(b6*t)-exp(b6*k[6]-l6*(tk[6]))))
+b1*S[1]+b2*S[2]+b3*S[3]+b4*S[4]-l1*SX[1]-l2*SX[2]-l3*SX[3]-l4*SX[4]
-a/(b1+l1)*(exp((b1+l1)*k[2])-exp((b1+l1)*k[1]))*(exp(-l1*k[6])
-exp(-l1*k[7]))-a/(b2+l2)*(exp((b2+l2)*k[3])-exp((b2+l2)*k[2]))*
(exp(-l2*k[6])-exp(-l2*k[7]))-a/(b3+l3)*(exp((b3+l3)*k[4])
-exp((b3+l3)*k[3]))*(exp(-l3*k[6])-exp(-l3*k[7]))-a/(b4+l4)*
(exp((b4+l4)*k[5])-exp((b4+l4)*k[4]))*(exp(-l4*k[6])-exp(-l4*k[7]))
-a/b6*(exp(b6*k[7])-exp(b6*k[6]))+a/(b6+l6)*(exp(b6*k[7])
-exp((b6+l6)*k[6]-l6*k[7]))
)
return(nllik)}

# initial values

a<-20
b<-5
l<-1

fit<-mle2(minuslogl=nllik,start=list(a=a,l1=l,l2=l,l3=l,l4=l,l6=l
,b1=b,b2=b,b3=b,b4=b,b6=b),data=list(n=n,k=k,t=t,S=S,SX=SX),trace=T)

summary(fit)

#
# Diagnostics
#

# fitted parameters

a<-132.88285025
l1<-0.11143718
l2<-0.04792029
l3<-0.04725934
l4<-0.16103929
l6<-0.21725727
b1<-0.08504902
```

```
b2<-0.02903951
b3<-0.02350445
b4<-0.02017609
b6<-0.02087073

# estimated expected values

a/(b1+l1)*(exp((b1+l1)*k[2])-exp((b1+l1)*k[1]))*
(exp(-l1*k[6])-exp(-l1*k[7]))
a/(b2+l2)*(exp((b2+l2)*k[3])-exp((b2+l2)*k[2]))*
(exp(-l2*k[6])-exp(-l2*k[7]))
a/(b3+l3)*(exp((b3+l3)*k[4])-exp((b3+l3)*k[3]))*
(exp(-l3*k[6])-exp(-l3*k[7]))
a/(b4+l4)*(exp((b4+l4)*k[5])-exp((b4+l4)*k[4]))*
(exp(-l4*k[6])-exp(-l4*k[7]))
a/b6*(exp(b6*k[7])-exp(b6*k[6]))-a/(b6+l6)*
(exp(b6*k[7])-exp((b6+l6)*k[6]-l6*k[7]))

#
# Fitted model simulation
#

library(survival)

# cohorts intensity measure

ci<-numeric()
ci[1]<-a/b1*exp(b1*20)-a/b1
ci[2]<-a/b2*exp(b2*30)-a/b2*exp(b2*20)
ci[3]<-a/b3*exp(b3*40)-a/b3*exp(b3*30)
```

```
ci[4]<-a/b4*exp(b4*46)-a/b4*exp(b4*40)

# previous cohorts cumulated intensities

Ac<-c(0,cumsum(ci))
sim_tmax<-max(Ac)
sim_ttot<-0
sim_t1<-numeric()

# homogeneous Poisson point process 1

while (sim_ttot<sim_tmax){
  x<-rexp(1)
  sim_t1<-c(sim_t1,x)
  sim_ttot<-sim_ttot+x
}

sim_t1<-cumsum(sim_t1)
sim_t1<-sim_t1[1:(length(sim_t1)-1)]
nsim<-length(sim_t1)

# cohort assignment

coho<-sim_t1*0
for (i in 1:nsim){
  coho[i]<-sum(Ac<sim_t1[i])
}

sim_torig<-sim_t1*0
b_est<-c(b1,b2,b3,b4)
```

```
l_est<-c(l1,l2,l3,l4)

# birth and death simulation

sim_torig<-1/b_est[coho]*log(exp(b_est[coho]*k[coho])
+b_est[coho]/a*(sim_t1-Ac[coho]))
sim_deaths<-sim_t1*0

for (i in 1:nsim){
sim_deaths[i]<-rexp(1,l_est[coho[i]])
}

plot(sim_torig,sim_deaths)
sim_tdeath<-sim_torig+sim_deaths

# Time window selection and missingness in tw

sim_tw<-sim_tdeath>47 & sim_tdeath<57 & sim_torig<56

# number per cohort check

sum(sim_tw[coho==1])
sum(sim_tw[coho==2])
sum(sim_tw[coho==3])
sum(sim_tw[coho==4])

# survival functions comparison

data_surv<-Surv(ttd[ttd>=0])
data_survfit<-survfit(data_surv~1)
```

```
plot(data_survfit,conf.int=F,xlab="age")
sim_surv<-Surv(sim_deaths[sim_tw])
sim_survfit<-survfit(sim_surv~1)
lines(sim_survfit,lty=2)

# 3rd cohort

data_surv3<-Surv(t3)
data_survfit3<-survfit(data_surv3~1)
plot(data_survfit3,conf.int=F,xlab="age")
sim_surv3<-Surv(sim_deaths[sim_tw & coho==3])
sim_survfit3<-survfit(sim_surv3~1)
lines(sim_survfit3,lty=2)
```

Bibliography

[1] D.B. Rubin. Inference and missing data. *Biometrika*, 1976.

[2] R.J.A. Little and D.B. Rubin. *Statistical analysis with missing data.* John Wiley and Sons, 1987.

[3] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood estimation from incomplete data via the em algorithm (with discussion). *Journal of the Royal Statistical Society*, 1977.

[4] A.A. Afifi and R.M. Elashoff. Missing observations in multivariate statistics: Review of the literature. *Journal of the American Statistical Association*, 1966.

[5] X.L. Meng. Missing data:dial m for??? *Journal of the American Statistical Association*, 2000.

[6] J.L. Shafer and J.W. Graham. Missing data: Our view of the state of the art. *Psychological Methods*, 2002.

[7] D.G. Kleinbaum, H. Morgenstern, and L.L. Kupper. Selection bias in epidemiologic studies. *American Journal of Epidemiology*, 1981.

[8] S. Demissie, M.P. LaValley, N.J. Horton, R.J. Glynn, and L.A. Cupples. Bias due to missing exposure data using complete-case analysis in the proportional hazards regression model. *Statistics in Medicine*, 2003.

[9] S.S. Wilks. Moments and distributions of estimates of population parameters from fragmentary samples. *Annals of Mathematical Statistics*, 1932.

[10] A. Matthai. Estimation of parameters from incomplete data with application to design of sample surveys. *Sankhya: The Indian Journal of Statistics*, 1951.

[11] Yoel Haitovsky. Missing data in regression analysis. *Journal of the Royal Statistical Society*, 1968.

[12] D.G. Horvitz and D.J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 1952.

[13] H.L. Oh and F.J. Scheuren. *Incomplete Data in Sample Surveys.* John Wiley and Sons, 1983.

[14] J.G. Ibrahim. Incomplete data in generalized linear models. *Journal of the American Statistical Association*, 1990.

[15] J.M. Robins, A. Rotnitzky, and L.P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 1994.

[16] J.M. Robins and A. Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 1995.

[17] D.B. Rubin. Multiple imputation in sample surveys - a phenomenological bayesian approach to nonresponse. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 1978.

[18] D.B. Rubin and N. Schenker. Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, 1986.

[19] D.B. Rubin. *Multiple Imputation for Nonresponse in Surveys.* John Wiley, 1987.

[20] J.L. Schafer. *Analysis of Incomplete Multivariate Data by Simulation.* John Wiley, 1987.

[21] D.B. Rubin. Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 1996.

[22] J. Neyman. On the two different apects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 1934.

[23] E.L. Lehmann. *Testing Statistical Hypotheses.* John Wiley, 1959.

[24] D.B. Rubin. Interval estimation from multiply imputed data: A case study using agriculture industry codes. *Journal of Official Statistics*, 1987.

[25] A. Agresti. *Categorical Data Analysis.* John Wiley, 1990.

[26] J.L. Schafer. *Analysis of Incomplete Multivariate Data.* Chapman & Hall, 1997.

[27] F. Barbaro, C. Carlucci, F. David, S. De Luca, R. Di Manno, F. Nusperli, A. Tancredi, F. Utili, and M. Volpe. L'indicatore anticipatore della spesa pubblica in conto capitale: la stima regionale annuale. Metodi 1, Ministero dello Sviluppo Economico - Dipartimento per lo Sviluppo e la Coesione Economica, 2004.

[28] J.H. Albert and S. Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 1993.

[29] D.B. Rubin. A note on bayesian, likelihood, and sampling distribution inferences. *Journal of Educational Statistics*, 1978.

[30] A.W.F. Edwards. *Likelihood.* Cambridge, 1972.

[31] R.A. Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society*, 1922.

[32] L.J. Savage. *The Foundations of Statistical Inference.* Methuen, 1962.

[33] M. Abramowitz and I.A. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables.* Dover, 1964.

[34] G. Walther. Inference and modeling with log-concave distributions. *Statistical Science*, 2009.

[35] S. Karlin. *Total Positivity.* Stanford University Press, 1968.

[36] S. Dharmadhikari and K. Joag-Dev. *Unimodality, convexity, and applications.* Academic Press Boston, 1988.

[37] D.J. Daley and D. Vere-Jones. *An Introduction to the Theory of Point Processes.* Springer, 1988.

[38] K. Andersen, O. Borgan, R.D. Gill, and N. Keiding. *Statistical Models Based on Counting Processes.* Springer, 1995.

[39] E. Cinlar. *Introduction to Stochastic Processes.* Prentice Hall, 1997.

[40] D.R. Cox and P.A.W. Lewis. *The Statistical Analysis of Series of Events.* Methuen, 1966.

[41] P.A.W. Lewis and G.S. Shedler. Simulation of nonhomogeneous poisson processes by thinning. Research Report RJ 2286, IBM, 1978.

[42] S. Ross. *Simulation.* Elsevier, 2006.

[43] N. Keiding. Statistical inference in the lexis diagram. *Philosophical Transactions: Physical Sciences and Engineering*, 1990.

[44] D.R. Brillinger. The natural variability of vital rates and associated statistics. *Biometrics*, 1986.

[45] J. Lund. Sampling bias in population studies. how to use the lexis diagram. *Scandinavian Journal of Statistics*, 2000.

[46] A.F. Karr. *Point Processes and Their Statistical Inference.* Marcel Dekker, 1991.

[47] E.E. Kummer. De integralibus quibusdam definitis et seriebus infinitis. *Journal fur die reine und angewandte Mathematik*, 1837.

[48] F. Tricomi. Sulle funzioni ipergeometriche confluenti. *Annali di Matematica Pura ed Applicata*, 1947.