Statistical analysis of matched survival data in national health databases.

Vanessa CHEZEU[1] & Jean-Francois DUPUY[2]& Valerie GARES[3]
& Samuel BOWONG[4]& Andre NANA[5]

[1] *Univ Rennes, INSA Rennes, CNRS, IRMAR - UMR 6625, F-35000 Rennes, France/ Univerty of Douala, Cameroun; Fidelette.Chezeu-Toumeni@insa-rennes.fr*
[2] *Univ Rennes, INSA Rennes, CNRS, IRMAR - UMR 6625, F-35000 Rennes, France; Jean-Francois.Dupuy@insa-rennes.fr*
[3] *Univ Rennes, INSA Rennes, CNRS, IRMAR - UMR 6625, F-35000 Rennes, France, France; Valerie.Gares@insa-rennes.fr*
[4] *University of Douala, Cameroun; sbowong@gmail.com*
[5] *University of Douala, Cameroun; nanayakam@yahoo.com*

Résumé.

Nous nous intéressons à l'estimation des paramètres du modèle à risques proportionnels de Cox, à partir de bases de données de santé appariées. Nous considérons la situation dans laquelle les variables explicatives et les durées de vie des individus ne sont pas rapportées dans la même base de données. Un processus préalable de couplage probabiliste d'enregistrements (chaînage) est donc nécessaire pour obtenir une base de données complète. Dans ce travail, nous proposons une équation d'estimation des paramètres du modèle de Cox, adaptée à ce cadre. Cette équation est obtenue en adaptant la fonction de score partiel "usuelle" du modèle de Cox, afin de prendre en compte le processus préalable de couplage des données. Nous décrivons une première étude de simulation, dans laquelle les probabilités d'appariement de chaque paire couplée sont déjà disponibles. Au travers de cette étude, nous évaluons les propriétés des estimateurs proposés. Dans une seconde approche, nous simulons deux bases de données, puis estimons les probabilités d'appariement des individus de ces bases (à l'aide d'un modèle "recordlinkage"), avant d'appliquer la méthodologie d'estimation proposée. Nous décrivons également les résultats de cette étude.

Mots clé. Couplage d'enregistrements ; Données censurées ; Durées de vie ; Santé publique ; Simulations numériques.

Abstract.

In this work, we investigate estimation in the Cox proportional hazards model from matched health databases. We consider the situation where the explanatory variables and individual lifetimes are not reported in the same database. A prior process of probabilistic record linkage is therefore necessary to obtain a complete database. We propose an estimating equation for the Cox model, adapted to this framework. This equation is obtained by adapting the "usual" partial score function, in order to take account of the prior linkage data process. We assess the properties of the resulting estimate via simulations. In the first simulation study, we assume that the matching probabilities of each linked pair are already available. In a second study, we simulate two databases, then we estimate the matching

probabilities of their respective individuals (using a record linkage model), and we apply the proposed estimation methodology. Results are described.

Keywords. Record linkage ; Censored data ; Duration data ; Public health ; Numerical simulations.

1 Introduction

Survival analysis refers to the set of statistical methods used to analyse data where the outcome of interest is the time of occurrence of some event (such as death, relapse...). Survival data occur in a wide range of fields: economy (unemployment duration), finance (time until repayment of a loan), insurance (duration between the beginning of long term care and death), engineering and reliability (duration until failure of an engine). In this work, we consider the statistical analysis of survival data arising from matched health databases.

The National Health Data System ('SNDS') is a large health database, which is often used to enrich existing medical cohorts and registries. This enrichment allows to recover as much information as possible on the evolution of patients health status. This, in turn, allows to make more robust analysis, but taking advantage of complementary data.

The enrichment of health databases can be done through a linkage data process between two databases. This linkage is simple if one has access to some unique patients identifiers (such as a unique code assigned definitively to a patient from his first contact within the establishment). However the use of this identifier may not be permitted for ethical reasons, or an identifier may simply not be available. In this case, we may only use partial identifiers, which are common to these databases (such as gender, postal code, dates of treatment...) to identify matched pairs from both databases.

Probabilistic record linkage method was first developed by Fellegi and Sunter (1969). Record linkage is a process of combining information about an individual or event in two or more databases. That is, the data from one source is joined with the data from another source that describes the same entity. For each pair of records, this method provides a score (or matching probability) which makes it possible to take into account, in the statistical analysis, the errors related to the matching process. To date, there have been several applications and improvements of the Fellegi and Sunter method, see for example Thanh et al. (2022) and Danhyang et al. (2022). Several authors have considered survival analysis from linked databases. For example, Thanh et al. (2023) propose an estimating equation in Cox proportional hazards (PH) model when no information on matching variables is available to the analyst, and the linkage errors are estimated from a validation file.

In the present work, we consider the situation where two databases are available: one of them contains the patients survival times (and censoring information), the second one contains information on the patients covariates. Matching probabilities between pairs of patients in both databases are available. We propose some estimation methods in the PH model, adapted

to this setting, which has yet received little attention (see Lahiri and Larsen (2005), Hof et al. (2017), Ying and Partha (2019)).

Our work is structured as follows: in section 2, we describe our problem. In section 3, we define the record linkage model. In section 4, we formulate the model and we propose an estimating equation for the parameters of the Cox model, based on the linked data. In section 5, we assess, via numerical simulations, the properties of the proposed estimate.

2 Problem description

Let us consider two databases A and B with respective sizes n_A and n_B such that $n_A \leq n_B$, and all individuals of database A are included in database B. The database B contains information on n_B independent individuals, such that for each individual $j = 1, \ldots, n_B$, we observe a pair $\{\boldsymbol{X}_j, \boldsymbol{Y}_j\}$, where $\boldsymbol{X}_j = (X_j^1, X_j^2, \ldots, X_j^P)^\top$ is a vector containing measures of the P covariates (e.g. blood group, monthly income), and $\boldsymbol{Y}_j = (Y_j^1, Y_j^2, \ldots, Y_j^K)^\top$ is a vector containing the measures of K variables, considered as partial identifiers of individual j (e.g. name, age, postal code). These variables are usually called matching variables.
The database A contains survival data of n_A independent individuals, such that for each individual $i = 1, \ldots, n_A$, we observe the triplet $\{T_i, \delta_i, \boldsymbol{Y}_i\}$, where $\boldsymbol{Y}_i$ is the same K-vector of partial identifier variables as in database B, $T_i = \min(\tilde{T}_i, C_i)$ is the observed survival time, $\tilde{T}_i$ is the event time of interest (e.g., time between inclusion in a trial and death), C_i is a random right-censoring time, and $\delta_i = 1_{\tilde{T}_i \leq C_i}$ is the censoring indicator. The event of interest is observed for an individual i if it occurs before the censoring time C_i. If for an individual i, we did not observe the event before C_i, this individual is considered to be censored (we have no information about its state after this period).

We assume that a unique individual j in B corresponds to each individual i in A, and that the record linkage errors are non-informative of the regression model (that is, matching errors can depend only on errors in the matching process, but not on covariates, nor on survival time data, see Thanh Vo and al (2022)).

Our objective is to assess the relationship between the patients lifetimes (which are recorded in database A) and the covariates (which are recorded in the database B but not in A).

3 Probabilistic record linkage model

In order to constitute a complete health database which contains information from both databases A and databases B, it is necessary to identify which measurements from databases A and databases B belong to the same individual. Due to the absence of unique identifiers in most situations, we must use the partial identifier variables $\boldsymbol{Y} = (Y^1, Y^2, \ldots, Y^K)^\top$ common to both databases.

Let, $\Omega = A \times B = \{(i, j) : i \in A \text{ and } j \in B; i = 1, \ldots, n_A; j = 1, \ldots, n_B\}$ the space which contains all the comparison pairs of units from A and B. For each pair of individuals $(i, j) \in$

Ω, the value of the matching variables are compared to each other. In this context, we use the binary comparison method proposed by Fellegy and Sunter (1969). Let the comparison function for each variable Y^k defined as,

$$\forall k = \{1, \ldots, K\}, \Gamma_{ij}^k = \begin{cases} 1 \text{ if } Y_i^k = Y_j^k, \\ \\ 0 \text{ if } Y_i^k \neq Y_j^k. \end{cases}$$

For each pair $(i, j) \in \Omega$, one obtains a comparison vector,

$$\mathbf{\Gamma}_{ij} = \left(\Gamma_{ij}^1, \ldots, \Gamma_{ij}^k, \cdots, \Gamma_{ij}^K\right) = (\Gamma_{ij}^k)_{1 \leq k \leq K}.$$

According to Fellegi and Sunter, every possible pair of individuals belongs either to the "matched" set M or to the "unmatched" set U such that,

$$M = \{(i, j); i = j, i \in A, j \in B\}, \quad \text{and} \quad U = \{(i, j); i \neq j, i \in A, j \in B\}.$$

If there is no error in the partial identifier variables, only the record pairs which have a comparison vector $\mathbf{\Gamma}_{ij} = (1, \ldots, 1)$ will be observed in the set M. Also, if the partial identifier variables were unique for each individual, only pairs which have a comparison vector $\mathbf{\Gamma}_{ij} = (0, \ldots, 0)$ will be observed in the set U. Because of errors generally present in databases (often due to bad recording of information, error in coding, transcription), we could observe pairs in the set M with different comparison vectors (e.g. $\mathbf{\Gamma}_{ij} = (0, 1, \ldots, 1)$).

Considering each comparison vector $\mathbf{\Gamma}_{ij}$, the probabilistic record linkage associates to each pair a probability of belonging to one of the two subsets of Ω.

Let $\boldsymbol{\gamma}_{ij} = \left(\gamma_{ij}^1, \ldots, \gamma_{ij}^K\right) \in \{0; 1\}^K$ be the realization set of the random variable $\mathbf{\Gamma}_{ij}$. The distribution of the comparison vector $\mathbf{\Gamma}_{ij}$ for each pair (i, j) is given by the following model:

$$\mathbb{P}\left(\mathbf{\Gamma}_{ij} = \boldsymbol{\gamma}_{ij}\right) = \mathbb{P}(\mathbf{\Gamma}_{ij} = \boldsymbol{\gamma}_{ij} \mid (i, j) \in M)\mathbb{P}((i, j) \in M) + \mathbb{P}(\mathbf{\Gamma}_{ij} = \boldsymbol{\gamma}_{ij} \mid (i, j) \in U)\mathbb{P}((i, j) \in U). \tag{1}$$

In practice, the number of distinct values of $\mathbf{\Gamma}_{ij}$ can be large, and estimation of the probabilities becomes complicated. Fellgi and Sunter (1969) and some authors (Andersen, and Gill (1982)) have therefore consider the hypothesis that the components of the vector $\mathbf{\Gamma}_{ij}$ can be reorganized, and are mutually statistically independent. According to the independence between the components of the vector $\boldsymbol{\gamma}_{ij}$, one has $\forall i = 1, \cdots n_A; j = 1, \cdots, n_B$:

$$\mathbb{P}\left(\mathbf{\Gamma}_{ij} = \boldsymbol{\gamma}_{ij} \mid (i, j) \in M\right) = \prod_{k=1}^{K} \mathbb{P}(\Gamma_{ij}^k = \gamma_{ij}^k \mid (i, j) \in M),$$

and

$$\mathbb{P}\left(\mathbf{\Gamma}_{ij} = \boldsymbol{\gamma}_{ij} \mid (i, j) \in U\right) = \prod_{k=1}^{K} \mathbb{P}(\Gamma_{ij}^k = \gamma_{ij}^k \mid (i, j) \in U),$$

Let us define by: $m^k = \mathbb{P}(\Gamma_{ij}^k = 1 \mid (i,j) \in M)$, $u^k = \mathbb{P}(\Gamma_{ij}^k = 1 \mid (i,j) \in U)$ and $\pi_M = \mathbb{P}((i,j) \in M)$; $\theta = (u^k, m^k, \pi_M; k = 1, \cdots, K)$ the set of all parameters to be estimated. We have a total of $(2K+1)$ parameters.

The final objective is to estimate $q_{ij} = \mathbb{P}((i,j) \in M \mid \mathbf{\Gamma}_{ij})$ (the probability for having match knowing the comparison vector) and $\mathbb{P}((i,j) \in U \mid \mathbf{\Gamma}_{ij}) = 1 - q_{ij}$.

Let be the matrix $\mathbf{\Gamma} = \{\boldsymbol{\gamma}_{ij}; i = 1, \cdots n_A; j = 1, \cdots n_B\}$. Considering the independence hypothesis for all comparison vectors, the likelihood function bored on $\mathbf{\Gamma}$ is defined by:

$$L(\theta \mid \mathbf{\Gamma})) = \prod_{i=1}^{n_A} \prod_{j=1}^{n_B} \left[\pi_M \mathbb{P}(\mathbf{\Gamma}_{ij} = \boldsymbol{\gamma}_{ij} \mid (i,j) \in M) \mathbb{P}((i,j) \in M)\right]^{z_{ij}} \times \left[\pi_U \mathbb{P}(\mathbf{\Gamma}_{ij} = \boldsymbol{\gamma}_{ij} \mid (i,j) \in U) \mathbb{P}((i,j) \in U)\right]^{1-z_{ij}},$$

where, $\pi_U = 1 - \pi_M$, and z_{ij} the indicator function such that:

$$z_{ij} = \begin{cases} 1 \text{ if } (i,j) \in M, \\ \\ 0 \text{ otherwise .} \end{cases}$$

The problem is then to maximize $L(\theta \mid \mathbf{\Gamma})$ under the constraint $\pi_U + \pi_M = 1$.

The EM algorithm (Meng and Rubin (1993), F.Santos (2015)) consists in estimating the parameters θ from a set of incomplete data. This algorithm makes it possible to determine the parameter θ which maximizes the likelihood function of the considered model.

Once all parameters are estimated using EM algorithm, the posterior probabilities are estimated for all pair (i,j) by the Bayesian formula

$$\begin{aligned} q_{ij} &= \mathbb{P}((i,j) \in M \mid \mathbf{\Gamma}_{ij} = \boldsymbol{\gamma}_{ij}) \\ &= \frac{\pi_M \mathbb{P}(\mathbf{\Gamma}_{ij} = \boldsymbol{\gamma}_{ij} \mid (i,j) \in M)}{\pi_M \mathbb{P}(\mathbf{\Gamma}_{ij} = \boldsymbol{\gamma}_{ij} \mid (i,j) \in M) + (1 - \pi_M) \mathbb{P}(\mathbf{\Gamma}_{ij} = \boldsymbol{\gamma}_{ij} \mid (i,j) \in U)} \end{aligned} \tag{2}$$

4 Cox regression with linked data

4.1 Cox proportional hazards model

Cox proportional hazards (PH) model (1972) is the most widely used survival regression model. It allows to estimate the effect of covariates $\boldsymbol{X}$ on patients lifetimes T. It is defined through the conditional hazard function:

$$\lambda(t \mid \boldsymbol{X}) = \lim_{dt \to 0} \frac{\mathbb{P}\left[t \leq T < t + dt \mid T \geq t, \boldsymbol{X}\right]}{dt}.$$

In Cox PH model, the conditional hazard function is specified as:

$$\lambda(t \mid \boldsymbol{X_i}) = \lambda_0(t) \exp(\boldsymbol{\beta}^\top \boldsymbol{X}_i), \tag{3}$$

where $\boldsymbol{X}_i = (X_i^1, \cdots, X_i^p)^\top$ is a vector of covariates for individual i, $\lambda_0(t)$ is an unknown non-negative function of time (the so-called baseline hazard function) and $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)$ is a p-vector of unknown parameters to be estimated.

An estimator of β is obtained by maximizing the partial likelihood, given by:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left(\frac{\exp(\boldsymbol{\beta}^\top \boldsymbol{X}_i)}{\sum_{j=1}^{n} Y_j(T_i) \exp(\boldsymbol{\beta}^\top \boldsymbol{X}_j)} \right)^{\delta_i}, \tag{4}$$

where $Y_j(t) = 1_{T_j \geq t}$ is the indicator that individual j is still at risk at time t. Differentiating $\log L(\boldsymbol{\beta})$ with respect to β yields the following estimating equation:

$$H_{\text{Cox}}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \left(\boldsymbol{X}_i - \frac{\sum_{j=1}^{n} Y_j(T_i) \exp(\boldsymbol{\beta}^\top \boldsymbol{X}_j) \boldsymbol{X}_j}{\sum_{j=1}^{n} Y_j(T_i) \exp(\boldsymbol{\beta}^\top \boldsymbol{X}_j)} \right) = 0.$$

The solution of this equation is called the maximum partial likelihood estimator of β. It is consistent and asymptotically normal, see Andersen and Gill (1982).

4.2 Estimation with linked data

We wish to estimate $\boldsymbol{\beta}$ in model (3), based on databases A and B described in section 2. Let i be some individual in database A. Recall that we do not observe $\boldsymbol{X}_i$. We only know that the covariate vector for individual i takes one value from the set $\{\boldsymbol{X}_1, \cdots \boldsymbol{X}_{n_B}\}$ in B.
Let us denote by $\boldsymbol{Z}_i$ the covariate vector that will be affected to individual i. From the record linkage process described above, we only know that

$$\mathbb{P}(\boldsymbol{Z_i} = \boldsymbol{X_j}) = q_{ij}, \quad j = 1, \ldots, n_B,$$

where the $\{q_{ij},\, i = 1, \ldots, n_A$ and $j = 1, \ldots, n_B\}$ are defined by (2). We normalize the q_{ij} by

$$\frac{q_{ij}}{\sum_{j=1}^{n_B} q_{ij}}$$

so that they sum to 1. We propose several estimation methods for $\boldsymbol{\beta}$.

Method 1: a naive approach. A first idea is to affect, to every individual i, the covariate vector $\boldsymbol{X}_j$ in B which has the largest posterior probability q_{ij}, that is:

$$\boldsymbol{Z}_i = \boldsymbol{X}_j \quad \text{where} \quad j = \text{argmax}_{1 \leq j \leq n_B}(q_{ij})$$

By naively treating the linked covariates $\boldsymbol{Z}_i$ as the true covariates $\boldsymbol{X}_i$, the partial likelihood (4) becomes:

$$L_{\text{naive}}(\boldsymbol{\beta}) = \prod_{i=1}^{n_A} \left(\frac{\exp(\boldsymbol{\beta}^\top \boldsymbol{Z}_i)}{\sum_{j=1}^{n_B} Y_j(T_i) \exp(\boldsymbol{\beta}^\top \boldsymbol{Z}_j)} \right)^{\delta_i},$$

from which we can estimate $\boldsymbol{\beta}$.

Method 2: a weighted partial likelihood. In order to reduce the bias of the naive estimator in method 1, we propose to modify the partial likelihood, by taking into account the probabilistic aspect of the linked covariate $\boldsymbol{Z}_i$. The basic idea is as follows.
Consider the i-th individual in database A. We successively affect each of the $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{n_B}\}$ to $\boldsymbol{Z}_i$, thus creating n_B fictional individuals. Each of these individuals enters the partial likelihood, but is suitably weighted by the matching probability of $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$. This yields the following weighted partial likelihood:

$$L_{\text{weighted}}(\boldsymbol{\beta}) = \prod_{i=1}^{n_A} \left(\prod_{j=1}^{n_B} \left[\frac{\exp(\boldsymbol{\beta}^\top \boldsymbol{X}_j)}{\sum_{p=1}^{n_A} Y_p(T_i) \left(\sum_{r=1}^{n_B} \exp(\boldsymbol{\beta}^\top \boldsymbol{X}_r) q_{pr} (\tilde{q}_p^{-1}) \right)} \right]^{q_{ij}(\tilde{q}_i^{-1})} \right)^{\delta_i},$$

where $\tilde{q}_i = \min_{1 \le j \le n_B} q_{ij}$, and $\tilde{q}_p = \min_{1 \le j \le n_B} q_{pj}$.

Maximizing L_{weighted} provides a second estimate of $\boldsymbol{\beta}$.

Method 3: complete partial likelihood with unobserved variables. Let $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n_B}\}$ be the set of possible values of $\boldsymbol{Z}_i$. If $\boldsymbol{Z}_i$ were observed, we could write the following likelihood, based on the observations $\{t_i, \delta_i, \boldsymbol{Z}_i\}, i = 1, \ldots, n_A$:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n_A} \mathbf{f}_{(T,\delta,\boldsymbol{Z})} \left(t_i, \delta_i, \boldsymbol{Z}_i\right).$$

In fact, $\boldsymbol{Z}_i$ is not observed, therefore, we propose to maximize the conditional expectation of the log-likelihood for complete data, given the observations. This is the idea of the EM algorithm. We have:

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{Z}|T,\delta} \left(\log \left(L(\boldsymbol{\beta}) \mid (T, \delta) \right) \right) &= \sum_{i=1}^{n_A} \mathbb{E}_{\boldsymbol{Z}|T,\delta} \left(\log \left(\mathbf{f}_{(T,\delta,\boldsymbol{Z})} \left(t_i, \delta_i, \boldsymbol{Z}_i\right) \right) \mid (t_i, \delta_i)_{1 \le i \le n_A} \right), \\
&= \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} \log \left(\mathbf{f}_{(T,\delta,\boldsymbol{Z})} \left(t_i, \delta_i, \boldsymbol{Z}_i = \boldsymbol{x}_j\right) \right) \left(\mathbb{P} \left(\boldsymbol{Z}_i = \boldsymbol{x}_j \mid (t_i, \delta_i) \right) \right),
\end{aligned}$$

where

$$\log \left(\mathbf{f}_{(T,\delta,\boldsymbol{Z})} \left(t_i, \delta_i, \boldsymbol{Z}_i = \boldsymbol{x}_j\right) \right) = \log \left[\mathbf{f}_{(T,\delta|\boldsymbol{Z})} \left(t_i, \delta_i \mid \boldsymbol{Z}_i = \boldsymbol{x}_j\right) \mathbb{P} \left(\boldsymbol{Z}_i = \boldsymbol{x}_j \right) \right],$$

$$\mathbb{P} \left(\boldsymbol{Z}_i = \boldsymbol{x}_j \mid (t_i, \delta_i) \right) = \frac{\mathbf{f}_{(T,\delta|\boldsymbol{Z})} \left(t_i, \delta_i \mid \boldsymbol{Z}_i = \boldsymbol{x}_j\right) \mathbb{P} \left(\boldsymbol{Z}_i = \boldsymbol{x}_j \right)}{\sum_{j=1}^{n_B} \mathbf{f}_{(T,\delta|\boldsymbol{Z})} \left(t_i, \delta_i \mid \boldsymbol{Z}_i = \boldsymbol{x}_j\right) \mathbb{P} \left(\boldsymbol{Z}_i = \boldsymbol{x}_j \right)},$$

with

$$\begin{aligned}
\mathbf{f}_{(T,\delta|\boldsymbol{Z})} \left(t_i, \delta_i \mid \boldsymbol{Z}_i = \boldsymbol{x}_j\right) &= f(t_i \mid \boldsymbol{x}_j)^{\delta_i} S(t_i \mid \boldsymbol{x}_j)^{1-\delta_i} \\
&= \left[\lambda_0(t_i) \exp \left(\boldsymbol{\beta}^\top \boldsymbol{x}_j \right) \right]^{\delta_i} S(t_i \mid \boldsymbol{x}_j),
\end{aligned}$$

and

$$\begin{aligned} S(t_i \mid \boldsymbol{x}_j) &= \exp\left(-\int_0^{t_i} \lambda_0(s)\exp\left(\boldsymbol{\beta}^\top \boldsymbol{x}_j\right)\, ds\right) \\ &= \exp\left(-\Lambda_0(t_i)\exp(\boldsymbol{\beta}^\top \boldsymbol{x}_j)\right) \quad \text{where} \quad \Lambda_0(t_i) = \int_0^{t_i} \lambda_0(s) ds. \end{aligned}$$

Replace the above expressions in the conditional expectation of the log-likelihood, and let

$$G_{ij}(\boldsymbol{\beta}) = \mathbb{P}\left(\boldsymbol{Z}_i = \boldsymbol{x}_j \mid (t_i, \delta_i)\right) = \frac{q_{ij}\left[\lambda_0(t_i)\exp\left(\boldsymbol{\beta}^\top \boldsymbol{x}_j\right)\right]^{\delta_i} S(t_i \mid \boldsymbol{x}_j)}{\sum_{j=1}^{n_B} q_{ij}\left[\lambda_0(t_i)\exp\left(\boldsymbol{\beta}^\top \boldsymbol{x}_j\right)\right]^{\delta_i} S(t_i \mid \boldsymbol{x}_j)}.$$

Then we obtain:

$$\begin{aligned} \mathbb{E}\left(\log\left(L(\boldsymbol{\beta}) \mid (T,\delta)\right)\right) &= \sum_{i=1}^{n_A}\sum_{j=1}^{n_B} \log\left[q_{ij}\left[\lambda_0(t_i)\exp\left(\boldsymbol{\beta}^\top \boldsymbol{x}_j\right)\right]^{\delta_i} S(t_i \mid \boldsymbol{x}_j)\right] \times G_{ij}(\boldsymbol{\beta}), \\ &= \sum_{i=1}^{n_A}\sum_{j=1}^{n_B} \left[\log(q_{ij}) + \delta_i \log(\lambda_0(t_i)\exp\left(\boldsymbol{\beta}^\top \boldsymbol{x}_j\right)) + \log(S(t_i \mid \boldsymbol{x}_j))\right] \times G_{ij}(\boldsymbol{\beta}), \end{aligned}$$

5 Simulation studies

5.1 Data generation

We adapt the data generating process developped in Thanh et al (2022, 2023). We consider two databases A and B containing K matching variables. We first generate the observations in database B (which is of size n_B), and then, we extract a random subset of size n_A of B, to obtain the database A. In database B, each Y_j^k is simulated from a Bernoulli distribution with probability p^k, for $j = 1, \ldots, n_B$ and $k = 1, \ldots, K$ (here, we let $p^k = 0.2$ for every k). Since there are only binary matching variables, the record linkage methods require a large number of matching variables to get good performances. We choose $K = 40$. We consider $P = 2$ covariates, which are simulated as follows: $X_j^1 \simeq \mathcal{N}(0,1)$ and $X_j^2 \simeq \mathcal{B}(0.8)$. The true survival times $\tilde{T}_j$ are simulated as:

$$\tilde{T}_j = -\frac{\log(U_j)}{\lambda \exp(\boldsymbol{\beta}^\top \boldsymbol{X}_j)},$$

where U_j follow a standard uniform distribution. We set $\beta = (0.5, -0.5)^\top$ and $\lambda = 1$. Then, we obtain $T = \min(\tilde{T}, C)$ and $\delta = 1_{\tilde{T} \leq C}$ by using a fixed censoring time C, chosen to yield an approximate censoring rate equal to 0.25.
Finally, a random subset of B is selected to produce the database A. We observe (T, δ) in database A and only $\boldsymbol{X}$ in database B.

To account for possible errors in the matching variables, the Y_i^k in A (that is, for $i = 1, \ldots, n_A$) are obtained from the Y_i^k in B as:

$$Y_i^k = \begin{cases} Y_i^k & \text{with probability } 1 - e^k \\ 1 - Y_i^k & \text{with probability } e^k \end{cases} \qquad \text{for } k = 1, \ldots, K.$$

We choose $e^k = 0.04$. Let $\boldsymbol{Q} = (q_{ij})_{1 \leq i \leq n_A, 1 \leq j \leq n_B} \in \mathbb{M}_{(n_A, n_B)}$, be a matching probability matrix where q_{ij} is given by (2).

5.2 Methods

We consider two scenarios. We first assume that the $\boldsymbol{Q}$ matrix resulting from the record linkage process is known (here, we choose a set a random values for the q_{ij}). In a second step, we estimate $\boldsymbol{Q}$ for the data generated in section 5.1 by the record linkage method developed by Vo et al. (2023). For each of these scenarios, we compare the values and properties of the different estimators of $\boldsymbol{\beta}$ obtained with the methods 1, 2 and 3 proposed in section 4.2.

Bibliographie

Fellegi, I. and Sunter, A. (1969), A theory for record linkage, *Journal of the American Statistical Association*, 64, pp. 1183-1210.

Thanh, H., Chauvet, G., Happe, A., Oger, E., Paquelet, S., Garès, V. (2022), Extending the Fellegi-Sunter record linkage model for mixed-type data with application to the French national health data system, *Journal of Computational Statistics and data Analysis*, 179, n°107656 .

Cox, D. (1972), Regression models and life-tables, *Journal of the royal statistical society*, Series B (Methodological),ISSN 00359246, 34, pp. 187-220.

F.Santos. (2015), L'algorithme EM: une courte présentation, CNRS, UMR 5199 PACEA.

Meng, X. and Rubin, D. (1993), Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, 80, pp. 267-278.

Hof, M., Ravelli, A. and Zwinderman, A. (2017), A probabilistic record linkage model for survival data, *Journal of the american statistical association*, 112(520), pp. 1504-1515.

Vo, TH., Garès, V., Zhang LC, Happe, A., Oger, E., Paquelet, S., and Chauvet, G. (2023), Cox regression with linked data, *Statistics in medicine* 43(2), pp. 296-314.

Ying, H. and Lahiri, P. (2019), Statistical analysis with linked data, *International statistical review* , 87(S1), pp S139-S157.

Lahiri, P. and Larsen, D. (2005), Regression analysis with linked data, *Journal of the American statistical association* , 100(469), pp. 222-230.

Danhyang, L., Li-Chun, Z. and Jea, K. (2022), Maximum entropy classification for record linkage, *Survey methodology*, 48(1), pp. 1-23.

Andersen, P., and Gill, R., (1982), Cox's regression model for counting processes: A large sample study, *The annals of statistics*, 10(4), pp 1100-1120.