Proceedings of Statistics Canada Symposium 2018
Combine to Conquer: Innovations in the Use of Multiple Sources of Data

Pairwise Estimating Equations for the Primary Analysis of Linked Data

A. Dasylva[1]

Abstract

A new estimating equation methodology is proposed for the primary analysis of linked data, i.e. an analysis by someone having an unfettered access to the related microdata and project information. It is described when the data come from the linkage of two registers with an exhaustive coverage of the same population, or from the linkage of two overlapping probability samples, as when the said registers have some undercoverage. This methodology accounts for the uncertainty about the match status of the record pairs, with a mixture model for the distribution of the agreements in a pair. It relies on the conditional independence assumption between the agreements and the responses given the covariates.

Key words: record linkage, linkage errors, data matching

1. Introduction

Record linkage has become an important tool in official statistics. A common application is linking a file of responses to a file of covariates, to produce an analytical dataset with all the required variables. A good example is the cohort mortality study by Sanmartin, Decady, Trudeau, Dasylva, Tjepkema, Finés, Burnett, Ross, and Manuel (2016). However record linkage is susceptible to linkage errors, including false positives and false negatives. A false negative is not linking two records from the same individual, while a false positive is linking records from different individuals. These errors are a source of bias if ignored (Bohensky, Jolley, Sundararajan, Evans, Pilcher, Scott and Brand 2010). In general, the analysis of linked data raises three connected missing data problems. The first problem is the uncertainty about which records relate to the same individual and the resulting linkage errors. The second problem is about the missing covariates for some individuals. The third problem is about the missing responses for other individuals. Of course, these last two problems are caused by the selection mechanisms into the different files. In previous work, the focus has been on the first problem (Chipperfield, Bishop and Campbell 2011; Lahiri and Law 2015; Chambers and Kim 2016; Hof, Ravelli and Zwinderman 2017). This paper describes a comprehensive solution under conditional independence assumptions. It is a methodology for the *primary* analysis of linked data, i.e. when all the linkage microdata and project information are available to the analyst. It involves estimating equations that are called *pairwise*, because they are based on the conditional mean response given the observed agreements in a single record pair. Thus, it offers a convenient way to fully exploit the information from each pair, while dispensing with the joint distribution of agreements across many pairs, which is much harder to handle. The remaining sections are organized as follows. Section 2 describes the notation and assumptions. Section 3 provides expressions for the response conditional mean or distribution given the observed pair agreements and covariates. Section 4 describes the related estimation procedures. Section 5 applies the methodology to two regression problems in simulations, including a linear model and a survival model with proportional hazards. The last section gives the conclusion.

2. Notation and assumptions

This paper considers the analysis of linked data that come from a finite population of individuals clustered in blocks. Each individual is characterized by quasi-identifiers and analytical variables that comprise of

[1]Abel Dasylva, Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa ON, Canada, K1A 0T6, (abel.dasylva@canada.ca);

a response and covariates. The data sources are two files about two samples of individuals. The first file contains the quasi-identifiers and the covariates for the first sample while the second file contains the quasi-identifiers and the responses for the other sample. In both files, the variables are error-free except for the quasi-identifiers. A given individual is drawn by the first file *at random*, and by the second file in a manner that is possibly *informative* but conditionally independent of the inclusion in the first file given the covariates. One may also view the files as Bernoulli samples drawn from two notional registers. After their creation, the files are linked by comparing their records in pairs within the blocks. For each pair this comparison produces a vector of outcomes, which serves to make a linkage decision or model the uncertainty about the pair *match status*, i.e. whether the records are from the same individual. Ultimately, the linkage produces an analytical file that comprises of all the pairs from the blocks with the corresponding analytical variables and vectors of outcomes. To exploit these latter vectors in the analysis, one requires a model of their interaction with the responses and the file inclusion indicators. Hereafter, the assumed model is essentially one of conditional independence given the covariates. The following paragraphs details this situation with the proposed notation.

The finite population: It is comprised of H random size blocks, where block h has size N_h. The individuals are labeled from 1 to $N = N_1 + \ldots + N_H$, with individual i characterized by the related quasi-identifiers, the covariates $\boldsymbol{x}_i$ and the response y_i. In a semiparametric model, $E[y_i|\boldsymbol{x}_i] = \mu(\boldsymbol{x}_i;\boldsymbol{\beta})$ where $\boldsymbol{\beta}$ is to be determined. In a parametric model, $y_i\big|\boldsymbol{x}_i \sim f(.\big|\boldsymbol{x}_i;\boldsymbol{\beta})$. The analytical variables from the different individuals are independent and identically distributed given N, according to a distribution that is also independent of N.

The registers: Two registers A' and B' contain the quasi-identifiers and the block identifier[2] for each individual. Additionally, A' contains the covariates while B' contains the responses. In both registers the variables are error-free except for the quasi-identifiers that may have typos. It is also convenient to represent a register by the set $\{1,\ldots,N\}$, and the set of individuals in block h by a subset of this latter set, which is denoted by A'_h or B'_h. In A', the records are labeled according to the related individual, such that record i is from individual i. However, in B', the records are labeled after an independent random permutation within each block[3], such that the same individual is associated with record $j(i)$, where $j(.)$ is the corresponding permutation of $\{1,\ldots,N\}$. In the same register, z_j denotes the observed response on record j.

The files: They are created by drawing records from the registers. File A is created by drawing record i from A' *at random*, with the probability $\pi(\boldsymbol{x}_i)$ given $\boldsymbol{x}_i$, with A_h denoting the resulting subset in block h. As for file B, it is created by drawing record $j(i)$ from B' with the probability $\nu(\boldsymbol{x}_i)$ given $\boldsymbol{x}_i$, with B_h denoting the resulting subset in block h. However the selection of record $j(i)$ from B' is possibly *informative* but conditionnally independent of the inclusion in A given $\boldsymbol{x}_i$. Thus, $I(i \in A_h)$ and $\big(y_i, I(j(i) \in B_h)\big)$ are conditionally independent given $\boldsymbol{x}_i$.

The linkage: The two files are linked by forming all the pairs within the blocks and comparing the quasi-identifiers in these pairs. A pair is *matched* if its records are from the same individual, e.g. $(i,j(i)) \in A_h \times B_h$. Otherwise it is *unmatched*. Let m_{ij} denote the *match status* of the pair (i,j), which is set to 1 if the pair is matched and to 0 otherwise. For the same pair, the comparison of the records produces the *outcomes vector* γ_{ij}, which determines if the pair is linked. When there are K quasi-identifiers, γ_{ij} may be of the form $\big(\gamma_{ij}^{(1)},\ldots,\gamma_{ij}^{K}\big)$, where $\gamma_{ij}^{(k)}$ is the comparison outcome for the k-th quasi-identifier. The linkage aims at producing an analytical file for the estimation. This file may be limited to the linked pairs or include all the pairs that satisfy the blocking criteria as in this paper. In the latter case, the analysis can account for the uncertainty about the m_{ij}'s with the vectors of outcomes. To this end, it is assumed that $[(\boldsymbol{y}_i, I(j(i) \in B_h))]_{i\in A'_h}$, $[I(i \in A_h)]_{i\in A'_h}$ and $\big[\big(m_{ij},\gamma_{ij}\big)\big]_{(i,j)\in A'_h\times B'_h}$ are conditionally independent given N_h and $[\boldsymbol{x}_i]_{i\in A'_h}$.

3. Conditional mean response

In a standard regression problem, the parameters are estimated based on the vector of covariates for each response. With linked data, this vector is unknown and maybe outside the file. One way of accounting for

[2]Assuming that the blocks are numbered from 1 to H

[3]No generality is lost because it is always possible to apply such a permutation.

this uncertainty is finding the conditional mean response given the covariates and outcomes vectors, which are observed. Dasylva (2018) has provided such expressions in different scenarios depending on the types of sources, including two registers, a sample and a register, and two samples. The following paragraphs describe these results and illustrate them in two examples including a linear model and a survival model. For ease of presentation a single block is considered, i.e. $H = 1$.

Two registers: The conditional mean response given N, $[\boldsymbol{x}_{i'}]_{1\leq i'\leq N}$ and γ_{ij} results from Theorem 1 by Dasylva (2018, p. 32).

$$E\left[z_j \left| N, [\boldsymbol{x}_{i'}]_{1\leq i'\leq N}, \gamma_{ij}\right.\right] = q_{ij}\mu\left(\boldsymbol{x}_i\right) + \frac{1-q_{ij}}{N-1}\sum_{i'\neq i}\mu\left(\boldsymbol{x}_{i'}\right), \tag{3.1}$$

where $q_{ij} = E\left[m_{ij} \left| N, [\boldsymbol{x}_{i'}]_{1\leq i'\leq N}, \gamma_{ij}\right.\right]$. When γ_{ij} and $\left(N, [\boldsymbol{x}_{i'}]_{1\leq i'\leq N}\right)$ are conditionally independent given m_{ij},

$$E\left[m_{ij} \left| N, [\boldsymbol{x}_{i'}]_{1\leq i'\leq N}, \gamma_{ij}\right.\right] = \left(1 + (N-1)\frac{P\left(\gamma_{ij}|m_{ij}=0\right)}{P\left(\gamma_{ij}|m_{ij}=1\right)}\right)^{-1}. \tag{3.2}$$

In equation (3.1), the conditional mean response is a weighted sum over the possible choices of covariates, with weights according to the conditional match probability of the pair. The total weight is assigned to $\boldsymbol{x}_i$ if the pair is surely matched ($q_{ij} = 1$), and it is uniformly distributed among the other choices if the pair is surely unmatched ($q_{ij} = 0$).

The conditional mean response given $\boldsymbol{x}_i$ and γ_{ij} results from Theorem 2 by Dasylva (2018, p. 37), when assuming that $P\left(m_{i'j} = 1 \left| N, [\boldsymbol{x}_{i''}]_{1\leq i''\leq N}, \boldsymbol{\gamma}_{ij}\right.\right)$ is the same for all $i' \neq i$ and that $\left[\boldsymbol{x}_{i'}\right]_{i'\neq i}$ and γ_{ij} are conditionally independent given N and $\boldsymbol{x}_i$. Under these additional assumptions, the conditional mean response is

$$E\left[z_j \left| \boldsymbol{x}_i, \gamma_{ij}\right.\right] = q_{ij}\mu\left(\boldsymbol{x}_i\right) + \left(1-q_{ij}\right)E\left[\mu\left(\boldsymbol{x}_{i'}\right)\right], \tag{3.3}$$

where $q_{ij} = E\left[m_{ij} \left| \boldsymbol{x}_i, \gamma_{ij}\right.\right]$. When γ_{ij} and $\left(N, \boldsymbol{x}_i\right)$ are conditionally independent given m_{ij},

$$E\left[m_{ij} \left| \boldsymbol{x}_i, \gamma_{ij}\right.\right] = \sum_{n\geq 1} P\left(N=n\right)\left(1 + (n-1)\frac{P\left(\gamma_{ij}|m_{ij}=0\right)}{P\left(\gamma_{ij}|m_{ij}=1\right)}\right)^{-1}. \tag{3.4}$$

In equation (3.3) the conditional mean response is still a weighted sum over the possible choices, with weights according to q_{ij}. When the pair is surely matched, the total weight is assigned to $\boldsymbol{x}_i$. When the pair is surely unmatched each possible covariates value is given a weight equal to its probability.

A sample of covariates and a register with all the responses: When the first source is a sample, the vector of covariates associated with z_j may not be on the file. In this case, the conditional mean response given N, $[\boldsymbol{x}_{i'}]_{1\leq i'\leq N}$, $i \in A$ and γ_{ij} results from Theorem 5 by Dasylva (2018, p. 76).

$$\begin{aligned} E\left[z_j \left| N, [\boldsymbol{x}_{i'}]_{i'\in A}, i\in A, \gamma_{ij}\right.\right] &= q_{ij}\mu\left(\boldsymbol{x}_i\right) + \left(1-q_{ij}\right)\left(\frac{1}{N-1}\sum_{i'\in A-\{i\}}\mu\left(\boldsymbol{x}_{i'}\right) + \right. \\ &\left. \frac{N-|A|}{N-1}\frac{E\left[(1-\pi(\boldsymbol{x}_{i''}))\mu\left(\boldsymbol{x}_{i''}\right)\right]}{E\left[(1-\pi(\boldsymbol{x}_{i''}))\right]}\right), \end{aligned} \tag{3.5}$$

where $q_{ij} = E\left[m_{ij} \left| N, [\boldsymbol{x}_{i'}]_{i'\in A}, i\in A, \gamma_{ij}\right.\right]$. When γ_{ij} and $\left(N, [\boldsymbol{x}_{i'}]_{i'\in A}\right)$ are conditionally independent given $i \in A$ and m_{ij},

$$E\left[m_{ij} \left| N, [\boldsymbol{x}_{i'}]_{i'\in A}, i\in A, \gamma_{ij}\right.\right] = \left(1 + (N-1)\frac{P\left(\gamma_{ij}|i\in A, m_{ij}=0\right)}{P\left(\gamma_{ij}|i\in A, m_{ij}=1\right)}\right)^{-1}. \tag{3.6}$$

In equation (3.5) , as before, the conditional mean response is a weighted sum, where the total weight is assigned to $\boldsymbol{x}_i$ if the pair is surely matched. However,when the pair is surely unmatched, the weight is

distributed among the other observed vectors ($\boldsymbol{x}_{i'}$, $i' \in A - \{i\}$) and the possible values of the unobserved vectors ($\boldsymbol{x}_{i'}$, $i' \in A' - A$).
From Theorem 6 by Dasylva (2018, p. 86), one obtains the conditional mean response given $i \in A$, $\boldsymbol{x}_i$ and γ_{ij}, when assuming that $P\left(m_{i'j} = 1 \left| N, [\boldsymbol{x}_{i''}]_{i'' \in A}, \boldsymbol{\gamma}_{ij} \right.\right)$ is the same for all $i' \neq i$ and that $\left[\boldsymbol{x}_{i'}\right]_{i' \in A - \{i\}}$, γ_{ij} and $I(i \in A)$ are conditionally independent given N and $\boldsymbol{x}_i$.

$$E\Big[z_j \Big| i \in A, \boldsymbol{x}_i, \gamma_{ij}\Big] = q_{ij}\mu\left(\boldsymbol{x}_i\right) + \left(1 - q_{ij}\right) E\left[\mu\left(\boldsymbol{x}_{i''}\right)\right], \tag{3.7}$$

where $q_{ij} = E\left[m_{ij} \left| i \in A, \boldsymbol{x}_i, \gamma_{ij}\right.\right]$. When γ_{ij} and $\left(N, \boldsymbol{x}_i\right)$ are conditionally independent given that $i \in A$ and m_{ij},

$$E\left[m_{ij} \left| i \in A, \boldsymbol{x}_i, \gamma_{ij}\right.\right] = \sum_{n \geq 1} P\big(N = n\big) \left(1 + (n-1)\, \frac{P\left(\gamma_{ij} \big| i \in A, m_{ij} = 0\right)}{P\left(\gamma_{ij} \big| i \in A, m_{ij} = 1\right)}\right)^{-1}. \tag{3.8}$$

Equation (3.7) is explained as equation (3.3).

Two samples: When the second register is also a sample, one must account for the selection of the observed responses, especially if it is informative as in a mortality study. The conditional mean response is derived by adapting equations. (3.5) and (3.7) as follows. Let O_{ij} denote the conditioning event that is based on N, $[\boldsymbol{x}_{i'}]_{i' \in A}$, γ_{ij} and $(i,j) \in A \times B$ or on $\boldsymbol{x}_i$, γ_{ij} and $(i,j) \in A \times B$. Let O'_{ij} denote the related event such that $O_{ij} = O'_{ij} \cap \{j \in B\}$. Then

$$E\Big[z_j \Big| O_{ij}\Big] = \frac{E\left[I(j \in B) z_j \big| O'_{ij}\right]}{E\left[I(j \in B) \big| O'_{ij}\right]}. \tag{3.9}$$

In this equation, the numerator is found by replacing $\mu(.)$ with the function $\boldsymbol{x} \mapsto \nu(\boldsymbol{x}) E\big[y_i \big| j(i) \in B, \boldsymbol{x}_i = \boldsymbol{x}\big]$ in equations (3.5) and (3.7). As for the denominator, it is found by replacing $\mu(.)$ with $\nu(.)$ in the same equations. The resulting expressions are found in Corollary 5 and Corollary 8 from Dasylva (2018, p. 81, p. 89). They show that one must account for the inclusion probabilities of the responses, even if their selection is done at random, unlike in a standard regression problem.

Conditional variance: The conditional variance is easily obtained from the expressions for the conditional mean response. Let O_{ij} denote the conditioning event that is defined as above. Indeed, the conditional variance is $var\big(z_j \big| O_{ij}\big) = E\big[z_j^2 \big| O_{ij}\big] - E\big[z_j \big| O_{ij}\big]^2$, where the conditional expectation $E\big[z_j^2 \big| O_{ij}\big]$ is obtained by replacing z_j and $\mu(\boldsymbol{x}_i)$ by z_j^2 and $E\left[y_i^2 \left| \boldsymbol{x}_i \right.\right]$ respectively, in the corresponding expression of the conditional mean response. The details are found in Corollary 6 and Corollary 9 from Dasylva (2018, p. 83, p. 90).

Parametric problem: In this case, y_i has a parametric given $\boldsymbol{x}_i$. Let O_{ij} denote the conditioning event that is defined as above. The conditional distribution of z_j given O_{ij} is found as follows. For a categorical response having ξ as a possible value, the conditional probability $P\left(z_j = \xi \big| O_{ij}\right)$ is obtained simply by replacing z_j and $\mu(\boldsymbol{x}_i)$ by $I\left(z_j = \xi\right)$ and $P\left(y_i = \xi \left| \boldsymbol{x}_i\right.\right)$ respectively, in the expression of the conditional mean response. For a continuous response, the conditional probability $P\left(z_j \leq \xi \big| O_{ij}\right)$ is obtained by substituting z_j and $\mu(\boldsymbol{x}_i)$ with $I\left(z_j \leq \xi\right)$ and $P\left(y_i \leq \xi \left| \boldsymbol{x}_i\right.\right)$ respectively, in the same expression. The conditional response density is obtained as the derivative of the cumulative distribution. The resulting expressions are found in Corollaries 7 and 10 from Dasylva (2018, p. 84, p. 91).

Linear model example: Let us consider the model $E\big[y_i \big| \boldsymbol{x}_i\big] = \boldsymbol{x}_i^\top \boldsymbol{\beta}$, $var\big(y_i \big| \boldsymbol{x}_i\big) = \sigma^2$, with an inclusion at random in either file, i.e. the conditional independence of y_i, $I(i \in A)$ and $I(j(i) \in B)$ given $\boldsymbol{x}_i$. Then $E\big[z_j \big| O_{ij}\big] = \boldsymbol{w}_{ij}^\top \boldsymbol{\beta}$, where $\boldsymbol{w}_{ij}$ depends on O_{ij}. When O_{ij} is comprised of N, $[\boldsymbol{x}_{i'}]_{i' \in A}$, γ_{ij} and $(i,j) \in A \times B$,

$$\boldsymbol{w}_{ij} = \frac{E\left[m_{ij} \left| O_{ij}\right.\right] \nu\left(\boldsymbol{x}_i\right) \boldsymbol{x}_i + \left(1 - E\left[m_{ij} \left| O_{ij}\right.\right]\right) \left(\frac{\sum_{i' \in A - \{i\}} \nu(\boldsymbol{x}_{i'}) \boldsymbol{x}_{i'}}{N-1} + \overbrace{\frac{N - |A|}{N-1} \frac{E\left[\left(1 - \pi\left(\boldsymbol{x}_{i'}\right)\right) \nu\left(\boldsymbol{x}_{i'}\right) \boldsymbol{x}_{i'}\right]}{E\left[\left(1 - \pi\left(\boldsymbol{x}_{i'}\right)\right)\right]}}^{(I)}\right)}{E\left[m_{ij} \left| O_{ij}\right.\right] \nu\left(\boldsymbol{x}_i\right) + \left(1 - E\left[m_{ij} \left| O_{ij}\right.\right]\right) \left(\frac{\sum_{i' \in A - \{i\}} \nu(\boldsymbol{x}_{i'})}{N-1} + \overbrace{\frac{N - |A|}{N-1} \frac{E\left[\left(1 - \pi\left(\boldsymbol{x}_{i'}\right)\right) \nu\left(\boldsymbol{x}_{i'}\right)\right]}{E\left[\left(1 - \pi\left(\boldsymbol{x}_{i'}\right)\right)\right]}}^{(II)}\right)}, \tag{3.10}$$

and the corresponding conditional variance is

$$var\left(z_j\middle|O_{ij}\right) = \left(E\left[m_{ij}\middle|O_{ij}\right]\nu\left(\boldsymbol{x}_i\right)\left(\left(\boldsymbol{x}_i^\top\boldsymbol{\beta}\right)^2+\sigma^2\right)+\left(1-E\left[m_{ij}\middle|O_{ij}\right]\right)\left(\frac{\sum_{i'\in A-\{i\}}\nu\left(\boldsymbol{x}_{i'}\right)\left(\left(\boldsymbol{x}_{i'}^\top\boldsymbol{\beta}\right)^2+\sigma^2\right)}{N-1}+\right.\right.$$
$$\left.\left.\overbrace{\frac{N-|A|}{N-1}\frac{E\left[\left(1-\pi\left(\boldsymbol{x}_{i'}\right)\right)\nu\left(\boldsymbol{x}_{i'}\right)\left(\left(\boldsymbol{x}_{i'}^\top\boldsymbol{\beta}\right)^2+\sigma^2\right)\right]}{E\left[\left(1-\pi\left(\boldsymbol{x}_{i'}\right)\right)\right]}}^{(I)}\right)\right)\times$$
$$\left(E\left[m_{ij}\middle|O_{ij}\right]\nu\left(\boldsymbol{x}_i\right)+\left(1-E\left[m_{ij}\middle|O_{ij}\right]\right)\left(\frac{\sum_{i'\in A-\{i\}}\nu\left(\boldsymbol{x}_{i'}\right)}{N-1}+\right.\right.$$
$$\left.\left.\overbrace{\frac{N-|A|}{N-1}\frac{E\left[\left(1-\pi\left(\boldsymbol{x}_{i'}\right)\right)\nu\left(\boldsymbol{x}_{i'}\right)\right]}{E\left[\left(1-\pi\left(\boldsymbol{x}_{i'}\right)\right)\right]}}^{(II)}\right)\right)^{-1}-\left(\boldsymbol{w}_{ij}^\top\boldsymbol{\beta}\right)^2. \tag{3.11}$$

Equations (3.10) and (3.11) also apply when the covariate file is a register, with $(I)=(II)=0$. When O_{ij} is based on $\boldsymbol{x}_i$, γ_{ij} and $(i,j)\in A\times B$,

$$\boldsymbol{w}_{ij}=\frac{E\left[m_{ij}\middle|O_{ij}\right]\nu\left(\boldsymbol{x}_i\right)\boldsymbol{x}_i+\left(1-E\left[m_{ij}\middle|O_{ij}\right]\right)E\left[\nu\left(\boldsymbol{x}_{i'}\right)\boldsymbol{x}_{i'}\right]}{E\left[m_{ij}\middle|O_{ij}\right]\nu\left(\boldsymbol{x}_i\right)+\left(1-E\left[m_{ij}\middle|O_{ij}\right]\right)E\left[\nu\left(\boldsymbol{x}_{i'}\right)\right]}, \tag{3.12}$$

and the conditional variance is

$$var\left(z_j\middle|O_{ij}\right) = \frac{E\left[m_{ij}\middle|O_{ij}\right]\nu\left(\boldsymbol{x}_i\right)\left(\left(\boldsymbol{x}_i^\top\boldsymbol{\beta}\right)^2+\sigma^2\right)+\left(1-E\left[m_{ij}\middle|O_{ij}\right]\right)E\left[\nu\left(\boldsymbol{x}_{i'}\right)\left(\left(\boldsymbol{x}_{i'}^\top\boldsymbol{\beta}\right)^2+\sigma^2\right)\right]}{E\left[m_{ij}\middle|O_{ij}\right]\nu\left(\boldsymbol{x}_i\right)+\left(1-E\left[m_{ij}\middle|O_{ij}\right]\right)E\left[\nu\left(\boldsymbol{x}_{i'}\right)\right]}-$$
$$\left(\boldsymbol{w}_{ij}^\top\boldsymbol{\beta}\right)^2. \tag{3.13}$$

Equations (3.10) and (3.12) clearly demonstrate the need to account for the inclusion probabilities of the responses, even if they are drawn at random.

Survival model example: Let us consider a finite population of individuals, who are all born at time 0. The survival time y_i is such that $y_i\left|\boldsymbol{x}_i\right.\sim e^{\boldsymbol{x}_i^\top\boldsymbol{\beta}}\exp\left(-e^{\boldsymbol{x}_i^\top\boldsymbol{\beta}}y_i\right)$, i.e. it follows a proportional hazards model with a constant baseline hazard. The corresponding individual is included in the cohort with the probability $\mu(\boldsymbol{x}_i)$. The cohort is followed from time 0 to time T, and during that time interval, all occurring deaths (whether from the cohort or not) are recorded in the mortality file. Of course, no mortality record is created for the individuals, who are alive by time T. This means that $\nu(\boldsymbol{x}_i)=P\left(j(i)\in B\middle|\boldsymbol{x}_i\right)=P\left(y_i\leq T\middle|\boldsymbol{x}_i\right)$. There are many important differences with standard survival models (Cox 1972). Indeed the standard setting is characterized by the recording of deaths within the cohort, known survivors and known risk factors for each death. However, when linking imperfectly (i.e. with linkage errors) a mortality file to a health survey (Sanmartin et al. 2016), the mortality file includes deaths from outside the cohort, while the survivors are unknown like the covariates related to the recorded deaths. In this latter case, the analysis may rest on the conditional density of z_j that is

$$f_{ij}\left(z\middle|O_{ij};\boldsymbol{\beta}\right)=\frac{h_{ij}\left(z;\boldsymbol{\beta}\right)}{\int_0^T h_{ij}\left(t;\boldsymbol{\beta}\right)dt},\ z\leq T, \tag{3.14}$$

where the function $h_{ij}\left(.;\boldsymbol{\beta}\right)$ depends on the event O_{ij}. When O_{ij} is based on N, $\left[\boldsymbol{x}_{i'}\right]_{i'\in A}$, γ_{ij} and $(i,j)\in A\times B$,

$$h_{ij}(z;\boldsymbol{\beta}) = E\left[m_{ij}\left|O_{ij}\right.\right]e^{\boldsymbol{x}_i^\top\boldsymbol{\beta}-e^{\boldsymbol{x}_i^\top\boldsymbol{\beta}}z}+\left(1-E\left[m_{ij}\left|O_{ij}\right.\right]\right)\left(\frac{\sum_{i'\in A-\{i\}}e^{\boldsymbol{x}_{i'}^\top\boldsymbol{\beta}-e^{\boldsymbol{x}_{i'}^\top\boldsymbol{\beta}}z}}{N-1}+\right.$$
$$\left.\overbrace{\frac{N-|A|}{N-1}\frac{E\left[\left(1-\pi\left(\boldsymbol{x}_{i'}\right)\right)e^{\boldsymbol{x}_{i'}^\top\boldsymbol{\beta}-e^{\boldsymbol{x}_{i'}^\top\boldsymbol{\beta}}z}\right]}{E\left[\left(1-\pi\left(\boldsymbol{x}_{i'}\right)\right)\right]}}^{(I)}\right),\ z\leq T. \tag{3.15}$$

When the file with the risk factors is a register, equation (3.15) applies with $(I) = 0$. When O_{ij} is based on $\boldsymbol{x}_i$, γ_{ij} and $(i,j) \in A \times B$,

$$h_{ij}(z;\boldsymbol{\beta}) = E\left[m_{ij}\,|O_{ij}\right] e^{\boldsymbol{x}_i^\top \boldsymbol{\beta} - e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}} z} + \left(1 - E\left[m_{ij}\,|O_{ij}\right]\right) E\left[e^{\boldsymbol{x}_{i'}^\top \boldsymbol{\beta} - e^{\boldsymbol{x}_{i'}^\top \boldsymbol{\beta}} z}\right],\ z \leq T. \tag{3.16}$$

4. Estimation procedures

Two approaches are described depending on whether the regression is parametric or semiparametric.

Semiparametric problem: When $E\left[y_i\,|\boldsymbol{x}_i\right] = \mu(\boldsymbol{x}_i;\boldsymbol{\beta})$, the vector $\boldsymbol{\beta}$ may be estimated by

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{h=1}^{H} \sum_{(i,j)\in A_h \times B_h} \tau_{ij} \frac{\left(z_j - E\left[z_j|O_{ij}\right]\right)^2}{var\left(z_j|O_{ij}\right)}, \tag{4.1}$$

where τ_{ij} is a nonnegative and nondecreasing function of $E\left[m_{ij}|O_{ij}\right]$, e.g. $\tau_{ij} = I\left(E\left[m_{ij}|O_{ij}\right] \geq \theta\right)$. This latter variable is used to select the pairs that enter the estimation procedure. As for the variance components, they may be estimated from the expression of $var\left(z_j|O_{ij}\right)$.

Parametric problem: Let us suppose that $y_i\big|\boldsymbol{x}_i \sim f(.|\boldsymbol{x}_i;\boldsymbol{\beta})$, and denote by $f_{ij}\left(.\big|O_{ij};\boldsymbol{\beta}\right)$ the conditional distribution of z_j given O_{ij}. In this case, $\boldsymbol{\beta}$ may be estimated by maximizing the following *composite likelihood* (Varin, Reid and Firth 2011).

$$\widehat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \sum_{h=1}^{H} \sum_{(i,j)\in (A_h \times B_h)} \tau_{ij} \log f_{ij}\left(z_j\big|O_{ij};\boldsymbol{\beta}\right), \tag{4.2}$$

where τ_{ij} is defined as in the semiparametric case.

Linear model example: Let $\sigma_{ij}^2 = var\left(z_j|O_{ij}\right)$, where $var\left(z_j|O_{ij}\right)$ is according to equation (3.11) or equation (3.13). When σ^2 is known, the estimator is

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{h=1}^{H} \sum_{(i,j)\in A_h \times B_h} \frac{\tau_{ij}}{\sigma_{ij}^2} \boldsymbol{w}_{ij} \boldsymbol{w}_{ij}^\top\right)^{-1} \left(\sum_{h=1}^{H} \sum_{(i,j)\in A_h \times B_h} \frac{\tau_{ij}}{\sigma_{ij}^2} \boldsymbol{w}_{ij} z_j\right). \tag{4.3}$$

Survival model example: The estimator $\widehat{\boldsymbol{\beta}}$ is the solution of equation (4.2) where $f_{ij}\left(.\big|O_{ij};\boldsymbol{\beta}\right)$ is from equation (3.14). However this solution is computed numerically because it does not have a closed form.

5. Simulations

The proposed estimators are evaluated in simulations for the two examples. These simulations involve a population of 1,024 individuals distributed across $H = 128$ blocks of size $N_h = 8$. There are $K = 8$ dichotomous linkage variables with true values that are independently and identically distributed according to the $Bernoulli(1/2)$ distribution. With probability 0.1, a given linkage variable is recorded with an error in a given source, in a manner that is independent of the error on the same variable in the other source, the other variables, or the other individuals. For the linkage, the variables are compared based on exact agreement in the record pairs, where the resulting vectors of outcomes satisfy the conditional independence property. The linkage parameters are estimated by expectation maximization (Jaro 1989) and the pair (i,j) is linked if its estimated conditional match probability given γ_{ij} is at least 0.9. The simulations are based on 100 repetitions.

Linear model: For the simulations, $x_i \sim N(0,1)$, and a response $y_i\big|x_i \sim N\left(\beta_0 + \beta x_i, \sigma^2\right)$, where $[\beta_0\ \beta_1] = [0.5\ 1]$ and $\sigma^2 = 0.49$. The response y_i and the file inclusion indicators are conditionally independent given x_i. Individual i is excluded from a file according to the logistic model that is based on the covariate x_i and

the known coefficients $\boldsymbol{\beta}' = [-2\ 1]^\top$. Two estimators are considered that are called PW1[4] and PW2. PW1 is based on equation (4.3) where $\tau_{ij} = I\left(E\left[m_{ij}\left|O_{ij}\right.\right] \geq 0.9\right)$, O_{ij} is comprised of N_h, $\left[\boldsymbol{x}_{i'}\right]_{i' \in A_h}$, $(i,j) \times A_h \times B_h$ and γ_{ij}, with σ^2 and σ_{ij}^2 estimated based on $var\left(z_j\middle|O_{ij}\right)$ and the following estimator of $\boldsymbol{\beta}$, which does not involve σ^2.

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{h=1}^{H} \sum_{(i,j) \in A_h \times B_h} \tau_{ij} \boldsymbol{w}_{ij} \boldsymbol{w}_{ij}^\top\right)^{-1} \left(\sum_{h=1}^{H} \sum_{(i,j) \in A_h \times B_h} \tau_{ij} \boldsymbol{w}_{ij} z_j\right). \tag{5.1}$$

PW2 is similar to PW1 with O_{ij} based on $\boldsymbol{x}_i$, $(i,j) \times A_h \times B_h$ and γ_{ij}. These estimators are compared to the *naive* estimator and to the *complete data* estimator. The naive estimator is the least-squares estimator based on the *linked* pairs while ignoring the linkage errors. The compete data estimator is the least-squares estimator based on the *matched* pairs. The results are found in Table 5-1, where the mean squared error (MSE) ranks the different estimators from the least to the best performing in the following order: naive estimator, PW2, PW1 and complete data estimator. Although PW1 is more precise than PW2, it is also harder to use because it requires the block sizes in the population.

Survival model: The simulation scenario is based on the previously described example, except that the entire population is included in the cohort. The other parameters are $x_i \sim Bernoulli(1/2)$, $y_i\big|x_i \sim e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}} \exp\left(-e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}} y_i\right)$, where $\boldsymbol{\beta} = [0.5\ 1]^\top$ and $T = 2.0$. PW1 is based on equation (4.2), with O_{ij} based on N_h, $\left[\boldsymbol{x}_{i'}\right]_{i' \in A_h}$, $(i,j) \times A_h \times B_h$ and γ_{ij}. PW2 is similar to PW1, with O_{ij} based on $\boldsymbol{x}_i$, $(i,j) \times A_h \times B_h$ and γ_{ij}. For both estimators $\tau_{ij} = I\left(E\left[m_{ij}\left|O_{ij}\right.\right] \geq 0.9\right)$. These two estimators are compared to the naive estimator and to the complete data estimator, where the former is the maximum likelihood estimator based on the linked pairs, while ignoring the linkage errors, and the latter is the maximum likelihood estimator based on the matched pairs. The results are also found in Table 5-1. As with the linear model, the MSE

Table 5-1
Simulation results

		Linear model			Survival model		
Parameter	Method	Bias (%)	Variance	MSE	Bias (%)	Variance	MSE
β_0	Naive	-3.433	0.001622	0.001901	-57.384	10.318838	10.297974
	PW1	-0.551	0.001505	0.001498	-1.399	0.006326	0.006312
	PW2	-0.574	0.001563	0.001556	-8.099	0.133355	0.133661
	Complete	-0.121	0.00065	0.000644	-0.144	0.00256	0.002535
β_1	Naive	-6.649	0.002736	0.007129	7.438	2.578701	2.558447
	PW1	-0.515	0.002745	0.002744	0.167	0.002663	0.00264
	PW2	-0.551	0.002816	0.002818	1.78	0.033379	0.033362
	Complete	-0.287	0.000786	0.000786	-0.081	0.001022	0.001013

ranks the different estimators from the least to the best performing in the following order: naive estimator, PW2, PW1 and complete data estimator. PW1 is much more precise than PW2, but more difficult to use for the same reason as in the linear case.

6. Conclusion

This paper addresses the problem of regression with data based on the imperfect linkage of two files, including a file of responses and a file of covariates, both with a partial coverage of the population. The obtained results show that one must consider the uncertainty about the pair match, as well as the inclusion probabilities of the responses, even if the latter are drawn at random. They also show that the model parameters may be estimated with precision from the conditional expectation of the observed responses given the observed comparison vectors and covariates. The resulting estimator is more precise when conditioning with respect

[4]PW stands for pairwise.

to all the observed covariates. However, it is more convenient to condition only with respect to the covariates observed in a pair, thereby dispensing with the need to know the bloc sizes in the population.

Acknowledgements

I would like to express my sincere gratitude to Prof. S. Sinha and Prof. J.N.K. Rao for their interest, insights and support.

References

Bohensky, M., Jolley, D., Sundararajan, V., Evans, S., Pilcher, D., Scott, I., and Brand, C. (2010), "A powerful research tool with potential problems," *BMC Health Services Research*, 10, 1–7.

Chambers, R., and Kim, G. (2016), "Secondary analysis of linked data," in *Methodological Developments in Data Linkage*, eds. H. K., G. H., and D. C., Chichester: Wiley, pp. 83–108.

Chipperfield, J., Bishop, G., and Campbell, P. (2011), "Maximum likelihood estimation for contingency tables and logistic regression with incorrectly linked data," *Survey Methodology*, 37, 13–24.

Cox, D. (1972), "Regression Models and Life-Tables," *Journal of the Royal Statistical Society, Series B*, 34, 187–220.

Dasylva, A. (2018), Pairwise estimating equations for the analysis of linked data, PhD thesis, Carleton University, Ottawa.

Hof, M., Ravelli, A., and Zwinderman, A. (2017), "A Probabilistic Linkage Model for Survival Data," *Journal of the American Statistical Association*, 112, 1504–1515.

Jaro, M. (1989), "Advances in record linkage methodology to matching the 1985 census of Tampa, Florida," *Journal of the American Statistical Association*, 84, 414–420.

Lahiri, P., and Law, J. (2015), Analysis of statitical models with linked data,, in *4th Baltic-Nordic Conference on Survey Statistics (BANOCOSS2015)*.

Sanmartin, C., Decady, Y., Trudeau, R., Dasylva, A., Tjepkema, M., Finés, P., Burnett, R., Ross, N., , and Manuel, D. (2016), "Linking the canadian community health survey and the canadian mortality database: An enhanced data source for the study of mortality," *Health Reports*, 27, 1–11.

Varin, C., Reid, N., and Firth, D. (2011), "An overview of composite likelihood methods," *Statistica Sinica*, 21, 5–42.