Recueil du Symposium 2018 de Statistique Canada
Combiner pour vaincre : innovations dans l'utilisation de multiples sources de données

Équations estimantes par paire pour l'analyse primaire de données couplées

A. Dasylva[1]

Résumé

Une nouvelle méthodologie d'équation estimante est proposée pour l'analyse primaire de données couplées, c.-à-d. une analyse par quelqu'un ayant un accès total aux micro-données et informations de projet associées. Elle est décrite lorsque les données proviennent du couplage de deux registres ayant une couverture exhaustive de la même population, ou du couplage de deux échantillons probabilistes, qui se chevauchent, comme c'est le cas lorsque lesdits registres ont de la sous-couverture. Cette méthodologie prend en compte l'incertitude concernant le statut d'appariement des paires d'enregistrements, à partir d'un modèle de mélange de la distribution marginale du vecteur de concordances dans une paire. Elle s'appuie sur l'hypothèse d'indépendance conditionnelle entre les vecteurs de concordances et les réponses étant donné les variables explicatives.

Mots clés: couplage d'enregistrements, erreurs de couplage, appariement de données

1. Introduction

Le couplage est devenu un outil important pour les statistiques officielles. Une application fréquente est le couplage d'un fichier de réponses avec un fichier de variables explicatives, pour produire un fichier analytique contenant toutes les variables nécessaires. Un bon exemple est l'étude de mortalité de cohorte de Sanmartin, Decady, Trudeau, Dasylva, Tjepkema, Finés, Burnett, Ross, et Manuel (2016). Toutefois, le couplage d'enregistrements est vulnérable aux erreurs, incluant les faux positifs et les faux négatifs. Un faux négatif consiste à ne pas lier deux enregistrements du même individu, tandis qu'un faux positif consiste à lier des enregistrements d'individus différents. Ces erreurs sont une source de biais si on les ignore (Bohensky, Jolley, Sundararajan, Evans, Pilcher, Scott et Brand 2010). En général, l'analyse de données couplées soulève trois problèmes de données manquantes interreliés. Le premier problème est l'incertitude quant aux enregistrements relatifs au même individu, et les erreurs de couplage qui en découlent. Le deuxième problème concerne les variables explicatives manquantes pour certains individus. Le troisième problème concerne les réponses manquantes pour d'autres individus. Évidemment, ces deux derniers problèmes sont dus aux mécanismes de sélection des différents fichiers. Dans les travaux antérieurs, l'accent a été mis sur le premier problème (Chipperfield, Bishop et Campbell 2011; Lahiri et Law 2015; Chambers et Kim 2016; Hof, Ravelli et Zwinderman 2017). Cet article décrit une solution globale sous des hypothèses d'indépendance conditionnelle. C'est une méthodologie pour l'analyse primaire de données couplées, c.-à-d. lorsque toutes les micro-données et informations du projet sont disponibles pour l'analyste. Elle est fondée sur des équations estimantes, qui sont appelées *par paire* parce qu'elles se basent sur l'espérance d'une réponse observée conditionnellement aux concordances observées dans une seule paire. Ainsi elle offre une façon commode d'exploiter pleinement l'information de chaque paire, tout en se passant de la distribution conjointe des concordances de plusieurs paires, qui est beaucoup plus difficile à manier. Les autres sections sont organisées comme suit. La section 2 décrit la notation et les hypothèses. La section 3 fournit les expressions de l'espérance ou de la distribution d'une réponse observée conditionnellement aux concordances et variables explicatives observées. La section 4 décrit les procédures d'estimation qui en découlent. La section 5 applique la méthodologie à deux problèmes de régression dans des simulations, incluant un modèle linéaire et un modèle de survie avec des risques proportionnels. La dernière section donne la conclusion.

[1]Abel Dasylva, Statistique Canada, 100 allée du Pré Tunney, Ottawa ON, Canada, K1A 0T6, (abel.dasylva@canada.ca);

2. Notation et hypothèses

Cet article considère l'analyse de données couplées, qui proviennent d'une population finie d'individus regroupés dans des pochettes. Chaque individu est caractérisé par des quasi-identificateurs et des variables analytiques comprenant une réponse et des variables explicatives. Les sources de données sont deux fichiers concernant deux échantillons d'individus. Le premier fichier contient les quasi-identificateurs et les variables explicatives pour le premier échantillon, tandis que le deuxième fichier contient les quasi-identificateurs et les réponses pour l'autre échantillon. Dans les deux fichiers, les variables analytiques sont sans erreurs à la différence des quasi-identificateurs. Un individu est tiré par le premier fichier *au hasard*, et par le deuxième fichier d'une façon, qui est peut-être informative mais indépendante de son tirage dans le premier fichier conditionnellement aux variables explicatives. On peut aussi voir ces fichiers comme des échantillons de Bernoulli tirés de deux registres conceptuels. Après leur création, les fichiers sont appariés en comparant leurs enregistrements par paire à l'intérieur des pochettes. Pour chaque paire, cette comparaison produit un vecteur de résultats, qui sert à prendre une décision de couplage ou à modéliser l'incertitude concernant le *statut d'appariement* de la paire, c.-à-d. si les enregistrements proviennent du même individu. Finalement, le couplage produit un fichier analytique, qui comprend toutes les paires des pochettes, avec leurs variables analytiques et vecteurs de résultats. Pour exploiter ces derniers vecteurs dans l'analyse, il est nécessaire de modéliser leur interaction avec les réponses et les indicatrices d'inclusion dans les fichiers. Ci-après, le modèle proposé est essentiellement fondé sur l'indépendance conditionnelle étant donné les variables explicatives. Les paragraphes suivants donnent plus de détails avec la notation proposée.

La population finie: Elle comprend H pochettes de taille aléatoire, où la pochette h est de taille N_h. Les individus sont étiquetés de 1 à $N = N_1 + \ldots + N_H$, l'individu i étant caractérisé par les quasi-identificateurs associés, le vecteur de variables explicatives $\boldsymbol{x}_i$ et la réponse y_i. Dans un modèle semi-paramétrique, $E\left[y_i \left| \boldsymbol{x}_i \right.\right] = \mu(\boldsymbol{x}_i; \boldsymbol{\beta})$ où $\boldsymbol{\beta}$ est à déterminer. Dans un modèle paramétrique, $y_i \big| \boldsymbol{x}_i \sim f(. \big| \boldsymbol{x}_i; \boldsymbol{\beta})$. Les variables analytiques des différents individus sont indépendantes et identiquement distribuées conditionnellement à N, selon une distribution qui ne dépend pas de N.

Les registres: Les registres A' et B' contiennent les quasi-identificateurs et le numéro de pochette[2] de chaque individu. En outre, A' contient les variables explicatives tandis que B' contient les réponses. Dans les deux registres, les variables sont sans erreurs sauf les quasi-identificateurs. Il est aussi commode de représenter un registre par l'ensemble $\{1, \ldots, N\}$, et les individus de la pochette h par un sous-ensemble de ce dernier ensemble, que l'on désigne par A'_h ou B'_h. Dans A', les enregistrements sont étiquetés selon l'individu correspondant, de sorte que l'enregistrement i provient de l'individu i. Toutefois, dans B', les enregistrements sont étiquetés après une permutation aléatoire des individus dans chaque pochette[3], de sorte que le même individu est associé avec l'enregistrement $j(i)$, où $j(.)$ est la permutation correspondante de $\{1, \ldots, N\}$. Dans le même registre, z_j désigne la réponse de l'enregistrement j.

Les fichiers: Ils sont créés par tirage d'enregistrements dans les registres. Le fichier A est créé en tirant l'enregistrement i de A' *au hasard*, avec la probabilité $\pi(\boldsymbol{x}_i)$ conditionnellement à $\boldsymbol{x}_i$, avec A_h comme sous-ensemble résultant dans la pochette h. Quant au fichier B, il est créé en tirant l'enregistrement $j(i)$ de B' avec la probabilité $\nu(\boldsymbol{x}_i)$ conditionnellement à $\boldsymbol{x}_i$, avec B_h comme sous-ensemble résultant dans la pochette h. Toutefois, le tirage de l'enregistrement $j(i)$ de B' est peut-être *informatif* mais indépendant du tirage dans A conditionnellement à $\boldsymbol{x}_i$. En somme, $I(i \in A_h)$ et $\left(y_i, I\left(j(i) \in B_h\right)\right)$ sont indépendants conditionnellement à $\boldsymbol{x}_i$.

Le couplage: Les deux fichiers sont appariés en formant toutes les paires d'enregistrements dans les pochettes et en comparant les quasi-identificateurs dans ces paires. Une paire est *appariée* si ses enregistrements proviennent du même individu, par ex. $(i, j(i)) \in A_h \times B_h$. Sinon, elle est *non appariée*. Soit m_{ij} le *statut d'appariement* de la paire (i, j), qui est égal à 1 si la paire est appariée et à 0 sinon. Pour la même paire, la comparaison des enregistrements produit le *vecteur de résultats* γ_{ij} , qui sert à déterminer si la paire est liée. Lorsqu'il y a K quasi-identificateurs, γ_{ij} peut être de la forme $\left(\gamma_{ij}^{(1)}, \ldots, \gamma_{ij}^{K}\right)$, où $\gamma_{ij}^{(k)}$ est le résultat de la comparaison pour le k-ième quasi-identificateur. Le couplage vise à produire le fichier analytique pour l'estimation. Ce fichier peut se limiter aux paires liées ou inclure toutes les paires provenant des pochettes

[2] On suppose que les pochettes sont numérotées de 1 à H.

[3] Cela n'est pas restrictif puisqu'il est toujours possible d'appliquer une telle permutation.

comme dans cet article. Dans ce dernier cas, l'analyse peut prendre en compte l'incertitude concernant les m_{ij} à partir des vecteurs de résultats. À cette fin, on suppose que $[(\boldsymbol{y}_i, I\,(j(i) \in B_h))]_{i \in A'_h}$, $[I\,(i \in A_h)]_{i \in A'_h}$ et $\left[\left(m_{ij}, \gamma_{ij}\right)\right]_{(i,j) \in A'_h \times B'_h}$ sont indépendants conditionnellement à N_h et $[\boldsymbol{x}_i]_{i \in A'_h}$.

3. Réponse espérée conditionnelle

Dans un problème de régression classique, les paramètres sont estimés à partir du vecteur de variables explicatives associé à chaque réponse. Avec des données couplées, ce vecteur est inconnu et peut-être hors du fichier. Une manière de prendre en compte cette incertitude est de trouver la réponse espérée conditionnellement aux vecteurs de résultats et de variables explicatives, qui sont observés. Dasylva (2018) a fourni de telles expressions dans différents scénarios selon les types de sources, incluant deux registres, un échantillon et un registre, et deux échantillons. Les paragraphes suivants décrivent ces résultats et les illustrent avec deux exemples, dont un modèle linéaire et un modèle de survie. Pour faciliter la présentation, une seule pochette est considérée, c.-à-d. $H = 1$.

Deux registres: La réponse espérée conditionnellement à N, $[\boldsymbol{x}_{i'}]_{1 \le i' \le N}$ et γ_{ij} découle du Theorem 1 de Dasylva (2018, p. 32).

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

où $q_{ij} = E\left[m_{ij} \,\middle|\, N, [\boldsymbol{x}_{i'}]_{1 \le i' \le N}, \gamma_{ij}\right]$. Lorsque γ_{ij} et $\left(N, [\boldsymbol{x}_{i'}]_{1 \le i' \le N}\right)$ sont indépendants conditionnellement à m_{ij},

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

Dans l'équation (3.1), la réponse espérée conditionnelle est une somme pondérée des choix possibles pour le vecteur de variables explicatives, avec des poids selon la probabilité conditionnelle d'appariement de la paire. Le poids total est attribué à $\boldsymbol{x}_i$ lorsque la paire est sûrement appariée ($q_{ij} = 1$), et il est uniformément distribué parmi les autres choix si la paire est sûrement non appariée ($q_{ij} = 0$).
La réponse espérée conditionnellement à $\boldsymbol{x}_i$ et γ_{ij} découle du Théorème 2 de Dasylva (2018, p. 37) en supposant que $P\left(m_{i'j} = 1 \,\middle|\, N, [\boldsymbol{x}_{i''}]_{1 \le i'' \le N}, \boldsymbol{\gamma}_{ij}\right)$ est identique pour tous les $i' \ne i$ et que $\left[\boldsymbol{x}_{i'}\right]_{i' \ne i}$ et γ_{ij} sont indépendants conditionnellement à N et $\boldsymbol{x}_i$. Sous ces hypothèses supplémentaires, la réponse espérée conditionnelle est

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

où $q_{ij} = E\left[m_{ij} \,|\boldsymbol{x}_i, \gamma_{ij}\right]$. Lorsque γ_{ij} et $\left(N, \boldsymbol{x}_i\right)$ sont indépendants conditionnellement à m_{ij},

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

Selon l'équation (3.3) la réponse espérée conditionnelle est encore une somme pondérée des choix possibles, avec des poids selon q_{ij}. Lorsque la paire est sûrement appariée, le poids total est attribué à $\boldsymbol{x}_i$. Lorsque la paire est sûrement non appariée, le poids est attribué à une valeur possible des variables explicatives selon la probabilité correspondante.

Un échantillon et un registre: Lorsque la première source est un échantillon, le vecteur de variables explicatives correspondant à z_j est peut-être hors du fichier. Dans ce cas, la réponse espérée conditionnellement à N, $[\boldsymbol{x}_{i'}]_{1 \le i' \le N}$, $i \in A$ et γ_{ij} découle du Théorème 5 de Dasylva (2018, p. 76).

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

où $q_{ij} = E\left[m_{ij}\left|N, [\boldsymbol{x}_{i'}]_{i'\in A}, i \in A, \gamma_{ij}\right.\right]$. Lorsque γ_{ij} et $\left(N, [\boldsymbol{x}_{i'}]_{i'\in A}\right)$ sont indépendants conditionnellement à $i \in A$ et m_{ij},

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

Comme plus haut, la réponse espérée conditionnelle est une somme pondérée où la totalité du poids est attribué à $\boldsymbol{x}_i$ lorsque la paire est sûrement appariée. Par contre, lorsque la paire est sûrement non appariée, le poids est réparti parmi les autres vecteurs observés ($\boldsymbol{x}_{i'}$, $i' \in A - \{i\}$) et les valeurs possibles des vecteurs non observés ($\boldsymbol{x}_{i'}$, $i' \in A' - A$).
Á partir du Théorème 6 par Dasylva (2018, p. 86), on obtient la réponse conditionnelle espérée conditionnellement à $i \in A$, $\boldsymbol{x}_i$ et γ_{ij}, en supposant que $P\left(m_{i'j} = 1\left|N, [\boldsymbol{x}_{i''}]_{i''\in A}, \boldsymbol{\gamma}_{ij}\right.\right)$ est identique pour tous les $i' \neq i$ et que $\left[\boldsymbol{x}_{i'}\right]_{i'\in A-\{i\}}$, γ_{ij} et $I(i \in A)$ sont indépendants conditionnellement à N et $\boldsymbol{x}_i$.

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

où $q_{ij} = E\left[m_{ij}\left|i \in A, \boldsymbol{x}_i, \gamma_{ij}\right.\right]$. Lorsque γ_{ij} et $\left(N, \boldsymbol{x}_i\right)$ sont indépendants conditionnellement à $i \in A$ et m_{ij},

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

L'équation (3.7) s'explique comme l'équation (3.3).

Deux échantillons: Lorsque le second registre est aussi un échantillon, on doit prendre en compte le tirage des réponses observées, particulièrement s'il est informatif comme dans une étude de mortalité. La réponse espérée conditionnelle se dérive en adaptant les équations (3.5) et (3.7) de la façon qui suit. Soit O_{ij} l'événement conditionnant, qui est basé sur N, $[\boldsymbol{x}_{i'}]_{i'\in A}$, γ_{ij} et $(i,j) \in A \times B$ ou $\boldsymbol{x}_i$, γ_{ij} et $(i,j) \in A \times B$. Soit O'_{ij} l'événement associé tel que $O_{ij} = O'_{ij} \cap \{j \in B\}$. Alors

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

Dans cette équation, le numérateur s'obtient en remplaçant $\mu(.)$ par la fonction $\boldsymbol{x} \mapsto \nu(\boldsymbol{x})E\big[y_i\big|j(i) \in B, \boldsymbol{x}_i = \boldsymbol{x}\big]$ dans les équations (3.5) et (3.7). Quant au dénominateur, il s'obtient en remplaçant $\mu(.)$ par $\nu(.)$ dans les mêmes équations. Les expressions résultantes se trouvent dans les Corollaires 5 et 8 de Dasylva (2018, p. 81, p. 89). Elles montrent qu'il est nécessaire de prendre en compte les probabilités d'inclusion des réponses, même si leur tirage se fait au hasard, contrairement à un problème de régression classique.

Variance conditionnelle: La variance conditionnelle s'obtient facilement à partir des expressions de la réponse espérée conditionnelle. Soit O_{ij} l'événement conditionnant, qui est défini comme plus haut. En effet, la variance conditionnelle est $var\big(z_j\big|O_{ij}\big) = E\big[z_j^2\big|O_{ij}\big] - E\big[z_j\big|O_{ij}\big]^2$, où l'espérance conditionnelle $E\big[z_j^2\big|O_{ij}\big]$ s'obtient en remplaçant z_j et $\mu(\boldsymbol{x}_i)$ par z_j^2 et $E\left[y_i^2\left|\boldsymbol{x}_i\right.\right]$ respectivement, dans l'expression de la réponse espérée conditionnelle. Les détails sont fournis par les Corollaires 6 et 9 de Dasylva (2018, p. 83, p. 90).

Problème paramétrique: Dans ce cas, y_i a une distribution paramétrique conditionnellement à $\boldsymbol{x}_i$. Soit O_{ij} l'événement conditionnant tel que défini plus haut. La distribution conditionnelle de z_j s'obtient de la façon suivante. Pour une réponse catégorielle dont ξ est une valeur possible, la probabilité conditionnelle $P\left(z_j = \xi\middle|O_{ij}\right)$ est obtenue simplement en remplaçant z_j et $\mu(\boldsymbol{x}_i)$ par $I\left(z_j = \xi\right)$ et $P\left(y_i = \xi\middle|\,\boldsymbol{x}_i\right)$ respectivement, dans l'expression de la réponse espérée conditionnelle. Pour une réponse continue, la probabilité conditionnelle $P\left(z_j \leq \xi\middle|O_{ij}\right)$ est obtenue en remplaçant z_j et $\mu(\boldsymbol{x}_i)$ par $I\left(z_j \leq \xi\right)$ et $P\left(y_i \leq \xi\middle|\,\boldsymbol{x}_i\right)$ respectivement, dans la même expression. La densité conditionnelle de la réponse est obtenue en dérivant la distribution cumulative. Les expressions résultantes se trouvent dans les Corollaires 7 et 10 de Dasylva (2018, p. 84, p. 91).

Exemple de modèle linéaire: Considérons le modèle $E\big[y_i\big|\boldsymbol{x}_i\big] = \boldsymbol{x}_i^\top\boldsymbol{\beta}$, $var\big(y_i\big|\boldsymbol{x}_i\big) = \sigma^2$, avec un tirage au hasard dans chaque fichier, c.-à-d. l'indépendance de y_i, $I(i \in A)$ et $I(j(i) \in B)$ conditionnellement à $\boldsymbol{x}_i$. Alors $E\big[z_j\big|O_{ij}\big] = \boldsymbol{w}_{ij}^\top\boldsymbol{\beta}$, où $\boldsymbol{w}_{ij}$ dépend de O_{ij}. Lorsque O_{ij} est basé sur N, $[\boldsymbol{x}_{i'}]_{i'\in A}$, γ_{ij} et $(i,j) \in A \times B$,

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

et la variance conditionnelle correspondante est

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

Les équations (3.10) et (3.11) s'appliquent aussi quand le fichier de variables explicatives est un registre, avec $(I) = (II) = 0$. Quand O_{ij} est basé sur $\boldsymbol{x}_i$, γ_{ij} et $(i,j) \in A \times B$,

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

et la variance conditionnelle est

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

Les équations (3.10) et (3.12) démontrent clairement la nécessité de prendre en compte les probabilités de tirage des réponses même si celles-ci sont tirées au hasard.

Exemple de modèle de survie: Considérons une population d'individus, qui sont tous nés à l'instant 0. Le temps de survie y_i est tel que $y_i\big|\boldsymbol{x}_i \sim e^{\boldsymbol{x}_i^\top\boldsymbol{\beta}}\exp\left(-e^{\boldsymbol{x}_i^\top\boldsymbol{\beta}}y_i\right)$, c.-à-d. qu'il suit un modèle de risques proportionnels avec un risque de référence constant. L'individu correspondant est inclus dans la cohorte avec la probabilité $\mu(\boldsymbol{x}_i)$. Cette cohorte est suivie de l'instant 0 jusqu'à l'instant T, et durant cette intervalle de temps, tous les décès (qu'ils proviennent de la cohorte ou pas) sont enregistrés dans le fichier de mortalité. Évidemment, aucun enregistrement de décès n'est créé pour les individus qui sont vivants à l'instant T. Cela veut dire que $\nu(\boldsymbol{x}_i) = P\big(j(i) \in B\big|\boldsymbol{x}_i\big) = P\big(y_i \leq T\big|\boldsymbol{x}_i\big)$. Il y a plusieurs différences importantes avec les modèles de survie traditionnels (Cox 1972). En effet, le cadre habituel se caractérise par un enregistrement des décès limité à la cohorte, des survivants connus et des facteurs de risque connus pour chaque décès. Par contre, en appariant de façon imparfaite (c.-à-d. avec des erreurs de couplage) un fichier de mortalité et une enquête de santé (Sanmartin et coll. 2016), le fichier de mortalité inclut des décès hors cohorte, tandis que

les survivants sont incertains ainsi que les facteurs de risque relatifs aux décès enregistrés. Dans ce dernier cas, l'analyse peut s'appuyer sur la densité conditionnelle de z_j qui est

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

où la fonction $h_{ij}\left(.;\boldsymbol{\beta}\right)$ dépend de l'événement O_{ij}. Lorsque O_{ij} est basé sur N, $[\boldsymbol{x}_{i'}]_{i'\in A}$, γ_{ij} et $(i,j)\in A\times B$,

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

Quand le fichier des facteurs de risques est un registre l'équation (3.15) s'applique aussi avec $(I)=0$. Lorsque O_{ij} est basé sur $\boldsymbol{x}_i$, γ_{ij} et $(i,j)\in A\times B$,

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

4. Procédures d'estimation

Deux approches sont décrites, selon que la régression est paramétrique ou semi-paramétrique.

Problème semi-paramétrique: Lorsque $E\left[y_i\left|\boldsymbol{x}_i\right.\right]=\mu(\boldsymbol{x}_i;\boldsymbol{\beta})$, le paramétre $\boldsymbol{\beta}$ peut s'estimer par

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

où τ_{ij} est une fonction non négative et non décroissante de $E\big[m_{ij}\big|O_{ij}\big]$, par ex. $\tau_{ij}=I\big(E\big[m_{ij}\big|O_{ij}\big]\geq\theta\big)$. Cette derniére variable sert à choisir les paires, qui entrent dans la procédure d'estimation. Quant aux composantes de variance, elles peuvent s'estimer à partir de l'expression de $var\big(z_j|O_{ij}\big)$.

Problème paramétrique: Supposons que $y_i\big|\boldsymbol{x}_i\sim f(.|\boldsymbol{x}_i;\boldsymbol{\beta})$ et désignons par $f_{ij}\big(.\big|O_{ij};\boldsymbol{\beta}\big)$ la distribution de z_j conditionnellement à O_{ij}. Dans ce cas, $\boldsymbol{\beta}$ peut s'estimer en maximisant la *vraisemblance composite* suivante (Varin, Reid et Firth 2011).

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

où τ_{ij} est défini comme dans le cas semi-paramétrique.

Exemple de modèle linéaire: Soit $\sigma_{ij}^2=var\big(z_j|O_{ij}\big)$, où $var\big(z_j|O_{ij}\big)$ est selon l'équation (3.11) ou l'équation (3.13). L'estimateur des moindres carrés pondérés est

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

Exemple de modèle de survie: L'estimateur est la solution de l'équation (4.2) où $f_{ij}\big(.\big|O_{ij};\boldsymbol{\beta}\big)$ provient de l'équation (3.14). Toutefois cette solution se calcule numériquement parce qu'elle n'a pas une forme analytique.

5. Simulations

Les performances des estimateurs proposés sont évaluées par des simulations, pour les deux exemples. Ces simulations se basent sur une population de 1,024 individus répartis à travers $H = 128$ pochettes de taille $N_h = 8$. Il y a $K = 8$ quasi-identificateurs dichotomiques, dont les vraies valeurs sont indépendantes et identiquement distribuées selon une distribution de $Bernoulli(1/2)$. Avec la probabilité 0.1, un quasi-identificateur donné est enregistré avec une erreur dans chaque source, indépendamment de l'erreur sur la même variable dans l'autre source, des autres variables ou des autres individus. Pour le couplage, des comparisons exactes sont effectuées, qui produisent des vecteurs de résultats satisfaisant la propriété d'indépendance conditionnelle. Les paramètres de couplage sont estimés par espérance-maximisation (Jaro 1989) et la paire (i, j) est liée si sa probabilité d'appariement conditionnellement à γ_{ij} est au moins égale à 0.9. Les simulations sont basées sur 100 répétitions.

Modèle linéaire: Pour les simulations, $x_i \sim N(0, 1)$ et $y_i|x_i \sim N(\beta_0 + \beta x_i, \sigma^2)$, où $[\beta_0\ \beta_1] = [0.5\ 1]$ et $\sigma^2 = 0.49$. La réponse y_i et les indicatrices d'inclusion dans les fichiers sont indépendantes conditionnellement à x_i. L'individu i est exclu d'un fichier donné selon le modèle logistique, qui est basé sur la variable explicative x_i et les coefficients connus $\boldsymbol{\beta}' = [-2\ 1]^\top$. Les deux estimateurs considérés s'appellent PW1[4] et PW2. PW1 est fondé sur l'équation (4.3) où $\tau_{ij} = I\left(E\left[m_{ij}\,|O_{ij}\right] \geq 0.9\right)$, O_{ij} est basé sur N_h, $\left[\boldsymbol{x}_{i'}\right]_{i' \in A_h}$, $(i, j) \times A_h \times B_h$ et γ_{ij}, tandis que σ^2 et σ_{ij}^2 sont estimés à partir de l'équation (3.11) et de l'estimateur suivant de $\boldsymbol{\beta}$, qui ne fait pas intervenir 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 est similaire à PW1 avec O_{ij} basé sur $\boldsymbol{x}_i$, $(i, j) \times A_h \times B_h$ et γ_{ij}. Ces estimateurs sont comparés à un estimateur *naïf* et à un estimateur basé sur les *données complètes*. L'estimateur naïf est basé sur les moindres carrés et les paires *liées* tout en ignorant les erreurs de couplage. L'estimateur des données complètes (ci-après appelé estimateur complet) est basé sur les moindres carrés et les paires *appariées*. Les résultats se trouvent dans la Table 5-1, où l'erreur quadratique moyenne (EQM) classe les différents estimateurs du moins performant au plus performant, dans l'ordre estimateur suivant: naïf, PW2, PW1 et estimateur des données complètes. Bien que PW1 soit plus précis que PW2, il est aussi plus difficile à utiliser parce qu'il nécessite les tailles des pochettes dans la population.

Modèle de survie: Le scénario de simulation est basé sur l'exemple déjà décrit sauf que toute la population fait partie de la cohorte. Les autres paramètres sont $x_i \sim Bernoulli(1/2)$, $y_i|x_i \sim e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}} \exp\left(-e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}} y_i\right)$, où $\boldsymbol{\beta} = [0.5\ 1]^\top$ et $T = 2.0$. PW1 est basé sur l'équation (4.2) avec O_{ij} basé sur N_h, $\left[\boldsymbol{x}_{i'}\right]_{i' \in A_h}$, $(i, j) \times A_h \times B_h$ et γ_{ij}. PW2 est similaire à PW1 avec O_{ij} basé sur $\boldsymbol{x}_i$, $(i, j) \times A_h \times B_h$ et γ_{ij}. Pour les deux estimateurs $\tau_{ij} = I\left(E\left[m_{ij}\,|O_{ij}\right] \geq 0.9\right)$. Ces deux estimateurs sont comparés à un estimateur *naïf* et un estimateur basé sur les données *complètes*, où le premier est l'estimateur du maximum de vraisemblance basé sur les paires liées en ignorant les erreurs de couplage, et le deuxième est l'estimateur du maximum de vraisemblance basé sur les paires appariées. Les résultats se trouvent aussi dans la Table 5-1. Comme avec le modèle linéaire, l'erreur quadratique moyenne (EQM) classe les estimateurs du moins performant au plus performant, dans l'ordre suivant: naïf, PW2, PW1 et données complètes. PW1 est beaucoup plus précis que PW2, mais plus difficile à utiliser pour la même raison que dans le cas linéaire.

6. Conclusion

Cet article aborde le problème de la régression avec des donnée issues du couplage imparfait de deux fichiers, dont un fichier de réponses et un fichier de variables explicatives, tous les duex avec une couverture partielle de la population. Les résultats obtenus montrent qu'il est nécessaire de prendre en compte l'incertitude concernant le statut d'appariement des paires, et les probabilités de tirage des réponses, même si celles-ci sont tirées au hasard. Ils montrent aussi qu'il est possible d'estimer avec précision les paramètres du modèle

[4]PW est une abréviation de l'anglais *pairwise*.

Table 5-1
Résultats de simulation

		Modèle linéaire			Modèle de survie		
Paramètre	Méthode	Biais (%)	Variance	EQM	Bias (%)	Variance	EQM
β_0	Naïf	-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
	Complet	-0.121	0.00065	0.000644	-0.144	0.00256	0.002535
β_1	Naïf	-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
	Complet	-0.287	0.000786	0.000786	-0.081	0.001022	0.001013

á partir de l'espérance des réponses observées conditionnellement aux vecteurs de résultats et variables explicatives observés. L'estimateur résultant est plus précis en conditionnant par rapport à toutes les variables explicatives observés. Toutefois, il est plus commode de seulement conditionner par rapport aux variables explicatives observé dans une seule paire, pour ainsi se passer d'avoir à connaître les tailles des pochettes dans la population.

Remerciements

J'aimerais exprimer ma sincère gratitude envers Prof. S. Sinha et Prof. J.N.K. Rao pour leur interêt, perspicacité et soutien.

Bibliographie

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

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

Chipperfield, J., Bishop, G., et 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., et 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., et 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., , et 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., et Firth, D. (2011), "An overview of composite likelihood methods," *Statistica Sinica*, 21, 5–42.