Recueil du Symposium 2014 de Statistique Canada
Au-delà des méthodes traditionnelles d'enquêtes : l'adaptation à un monde en évolution

Modélisation des indicateurs de réponse d'autodénombrement et de suivi sous forme de survie en temps discret

A. Demnati[1]

Résumé

Recueillir des données par Internet ou par la poste auprès d'unités échantillonnées est plus économique que le faire par interview. Ces méthodes font de l'autodénombrement une approche de collecte des données intéressante pour les enquêtes et les recensements. En dépit de ses avantages, la collecte de données par autodénombrement, en particulier la collecte de données par Internet, peut produire des taux de réponse plus faibles que ceux obtenus par interview. Pour accroître les taux de réponse, on soumet les non-répondants à un mode mixte de traitements de suivi, qui influent sur la probabilité résultante de réponse, afin de les encourager à participer. Les analyses de régression comprennent habituellement des facteurs et des interactions qui ont une incidence importante sur l'interprétation des modèles statistiques. Comme l'occurrence d'une réponse est intrinsèquement conditionnelle, nous commençons par enregistrer l'occurrence des réponses en intervalles discrets, et nous caractérisons la probabilité de réponse comme étant un risque en temps discret. Cette approche facilite l'examen du moment où une réponse est la plus susceptible d'avoir lieu et de la façon dont la probabilité de réponse varie au fil du temps. Le biais de non-réponse peut être évité en multipliant le poids d'échantillonnage des répondants par l'inverse d'une estimation de la probabilité de réponse. Les estimateurs des paramètres du modèle, ainsi que des paramètres de la population finie sont présentés. Les résultats de simulations en vue d'évaluer la performance des estimateurs proposés sont également présentés.

Mots-clés : Analyse des biographies, données longitudinales, maximum de vraisemblance, enquêtes à mode de collecte mixte, unités partiellement classées.

1. Introduction

Combiner les modes de suivi et de collecte des données offre la possibilité de compenser les inconvénients d'un mode par les avantages de l'autre. Par exemple, ayant constaté qu'Internet, contrairement à l'envoi par la poste, permet de rapprocher la saisie et la vérification des données du répondant, de nombreux organismes statistiques offrent maintenant le choix d'utiliser des questionnaires électroniques en vue d'améliorer la qualité des processus statistiques tout en réduisant les coûts d'enquête. Cette amélioration potentielle de la qualité des enquêtes, conjuguée au fait que la collecte de données par Internet ou par la poste auprès des unités échantillonnées est beaucoup plus économique que la réalisation d'interviews, fait de l'autodénombrement une méthode de collecte des données intéressante pour les enquêtes et les recensements. Bien que l'autodénombrement, en particulier les enquêtes en ligne, offre des avantages et que l'on s'attend à voir s'étendre son application dans l'avenir, cela pose des difficultés particulières pour les enquêtes et les recensements. Les valeurs observées d'une variable d'intérêt type y pourraient dépendre de la variable y_m associée au mode m de collecte des données, $m = 1,..., M$, où M est le nombre de modes de collecte des données envisagé pour une enquête donnée. En principe, toute unité k de la population P de taille N peut donner toutes les réponses, c.-à-d. une réponse $y_{m;k}$ qu'elle donnerait si elle choisissait le mode de collecte m. Puisque chaque unité ne reçoit ou ne choisit qu'un seul mode, une seule réponse est observée. Si la variable d'intérêt est définie de manière unique et indépendamment de chaque mode, alors $y_{m;k}$ représente la valeur que l'unité considère comme étant la réponse correcte pour y_k à y, résultant du support d'information du mode m au moyen duquel la question est présentée à l'unité. Supposons que l'espérance de la réponse y_k sous le modèle est spécifiée par $E_M(y_k) = \mu_k(\boldsymbol{\chi}_k^T \boldsymbol{\Theta})$, où $\boldsymbol{\chi}_k = (\chi_{1k},..., \chi_{pk})^T$ est un vecteur de dimension $p \times 1$ de variables explicatives,

[1] A. Demnati, Division des méthodes d'enquêtes auprès des entreprises, Statistique Canada, Ottawa, Canada, K1A 0T6 (Abdellatif.Demnati@statcan.gc.ca).

$\boldsymbol{\Theta} = (\Theta_1, ..., \Theta_p)^T$ est le vecteur de dimension $p \times 1$ du paramètre du modèle et E_M désigne l'espérance sous le modèle. Nous supposons que le paramètre de population finie associé au paramètre vectoriel de modèle $\boldsymbol{\Theta}$ est défini comme la solution d'une équation estimante de la forme

$$\boldsymbol{S}(\boldsymbol{\Theta}) = \sum_k \boldsymbol{s}(y_k; \boldsymbol{\Theta}) - \boldsymbol{v}(\boldsymbol{\Theta}) = \boldsymbol{0}, \quad (1.1)$$

où $\sum_k$ est la somme sur l'ensemble des unités de la population finie, la fonction connue $\boldsymbol{s}(y_k; \boldsymbol{\Theta})$ est un vecteur - de y_k de taille p et la fonction connue $\boldsymbol{v}(\boldsymbol{\Theta})$ permet des paramètres explicitement définis. Pour les modèles de régression linéaire et logistique, $\boldsymbol{s}(y_k; \boldsymbol{\Theta}) = \boldsymbol{\chi}_k (y_k - \mu_k(\boldsymbol{\chi}_k^T \boldsymbol{\Theta}))$ et $\boldsymbol{v}(\boldsymbol{\Theta}) = \boldsymbol{0}$. Pour le cas particulier du total de population finie, $Y = \sum_k y_k$, $\boldsymbol{s}(y_k; \boldsymbol{\Theta}) = y_k$, $\boldsymbol{v}(\boldsymbol{\Theta}) = \Theta_N$ et $\Theta_N = Y$.

L'un des principaux objectifs du mode mixte de collecte des données est d'influencer l'unité en vue d'obtenir sa coopération, quel que soit le mode de collecte des données qu'elle préfère. Le taux de couverture global pour l'unité k pour les modes combinés peut être défini comme $r_k = 1 - \prod_{m=1}^{M} (1 - r_{m;k}) = \sum_{m=1}^{M} r_{m;k}$ et la probabilité globale de réponse peut être représentée sous la forme de mélange $\xi_k = \sum_{m=1}^{M} \phi_m \xi_{m;k}$, où ϕ_m est la proportion de la population utilisant le mode m, $\xi_{m;k} = E_r(r_{m;k})$ est la probabilité de réponse associée au mode m de collecte des données et E_r désigne l'espérance par rapport au mécanisme de réponse. Si le mode mixte peut accroître les taux globaux de réponse, nous serons, évidemment, heureux de quantifier et d'examiner la contribution de chaque mode à la probabilité de réponse. En réalité, l'autodénombrement peut produire des taux de réponse plus faibles que les interviews. Pour obtenir la collaboration des non-répondants et donc maximiser la qualité de l'enquête, les non-répondants sont affectés à un mode mixte de traitements de suivi. Différents coûts sont associés à différents traitements. Par exemple, le suivi sur place est plus coûteux que le suivi par téléphone. À l'heure actuelle, dans le cas de certaines enquêtes auprès des entreprises, afin de réduire le coût total de la collecte des données, le suivi des cas de non-réponse est effectué auprès d'une partie seulement des non-répondants. Ces unités sont souvent identifiées de manière déterministe en se basant, par exemple, sur leur contribution prévue à l'estimation. En outre, puisqu'un nombre important d'unités ne font jamais l'objet d'un suivi pour la non-réponse, le taux de réponse final est parfois très faible. Le biais de non-réponse peut être évité en multipliant le poids d'échantillonnage des répondants par l'inverse de la probabilité de réponse. Puisque la probabilité de réponse est inconnue, on se sert d'une probabilité estimée. Comme l'ont mentionné Rosenbaum (1987) et d'autres, les estimateurs utilisant la probabilité de réponse estimée peuvent être plus efficaces que ceux utilisant la probabilité de réponse réelle.

Compte tenu des problèmes susmentionnés, une question qui devrait intéresser tout particulièrement les organismes statistiques est la suivante : comment faut-il modéliser les probabilités de réponse sous un mode mixte ainsi que l'influence des traitements de suivi sur la probabilité de réponse résultante? Voici d'autres questions pertinentes : si un facteur du mode mixte est amélioré, quel sera l'effet sur la performance du mécanisme de réponse? Comment pouvons-nous estimer la probabilité de réponse due à un facteur d'intérêt particulier du mode mixte en présence des autres facteurs du mode mixte? Comme un suivi intensif est onéreux, une stratégie de suivi est nécessaire afin d'optimiser l'utilisation des ressources en prenant en considération la qualité des estimations. Puisqu'un traitement de suivi est susceptible de produire des estimations de meilleure qualité, la stratégie pourrait consister à affecter les non-répondants à différents traitements tout en tenant compte de l'effet des coûts de collecte des données. Afin de discuter de certaines de ces questions et d'autres, et compte tenu du nombre limité de pages du présent article, la première partie de notre exposé est présentée comme il suit : à la section 2, nous caractérisons la probabilité de réponse parun modèle de risque à temps discret, puis nous examinons les facteurs et les interactions dans une analyse de régression; à la section 3, nous étudions l'estimateur du paramètre de régression (ou de nuisance), ainsi que les estimateurs du paramètre d'intérêt dans les conditions mentionnées plus haut; enfin, à la section 4, nous présentons les résultats des simulations.

2. Formalisation d'un modèle de réponse

2.1 Risque à temps discret

Considérons un échantillon homogène d'unités exposées chacune au risque de connaître un événement cible unique : répondre. L'événement cible ne peut être répété. Pour enregistrer l'occurrence de la réponse en intervalles

discrets, nous divisons le temps continu en une série de périodes continues : 1, 2 et ainsi de suite. Supposons que la durée de la collecte des données soit constituée de I périodes. Soit t la variable aléatoire discrète qui indique la période i où la réponse a lieu pour une unité sélectionnée aléatoirement dans l'échantillon. Alors, chaque unité k est observée jusqu'à une certaine période I_k, avec $I_k \leq I$. L'observation de l'unité pourrait être interrompue pour deux raisons : 1) l'unité répond; ou 2) l'enquête se termine. Dans le premier cas, $t = I_k$. Dans le deuxième cas, on sait uniquement que $t > I$. Les unités pour lesquelles $t > I$ sont censurées à droite — on ne sait pas si elles auraient répondu. Comme l'occurrence de la réponse est intrinsèquement conditionnelle, nous caractérisons t par sa densité de probabilité conditionnelle — la distribution de la probabilité qu'une réponse ait lieu durant chaque période sachant qu'elle n'a pas déjà eu lieu durant une période antérieure — connue comme la fonction de risque à temps discret. Le risque à temps discret, $h_{ki}(\boldsymbol{x}_k, \boldsymbol{\beta})$, h_{ki} en abrégé, est défini comme étant la probabilité conditionnelle que l'unité k réponde à la période i, sachant que l'unité n'a pas répondu avant la période i :

$$h_{ki} = \Pr(t = i \mid t \geq i), \tag{2.1}$$

où $\boldsymbol{x}_k$ désigne les variables explicatives invariantes ainsi que variantes dans le temps et $\boldsymbol{\beta}$ est le paramètre vectoriel inconnu qu'il faut estimer. Pour l'unité pour laquelle $t = i$, la probabilité d'obtenir une réponse à la période i pourrait être exprimée en fonction du risque sous la forme

$$\Pr(t = i) = h_{ki} \prod_{j=1}^{i-1} (1 - h_{kj}). \tag{2.2}$$

Pour les unités pour lesquelles $t > i$, la probabilité d'obtenir une réponse peut être exprimée sous la forme

$$\Pr(t > i) = \prod_{j=1}^{i} (1 - h_{kj}). \tag{2.3}$$

Nous supposons que chaque unité de l'échantillon survit durant chaque période discrète successive jusqu'à ce qu'elle réponde ou qu'elle soit censurée parce que la collecte des données se termine. L'utilisation d'un mode de collecte mixte modifie l'expression de la fonction de risque (2.1) qui devient $h_{ki} = \sum_{m=1}^{M} \phi_m h_{ki|m}$, où $h_{ki|m}$ est la fonction de risque à temps discret pour le mode m. La probabilité marginale d'obtenir une réponse après I périodes est donnée par

$$\xi_k = 1 - \prod_{i=1}^{I} (1 - h_{ki}) = \sum_{i=1}^{I} \xi_k^{(i)}. \tag{2.4}$$

où $\xi_k^{(i)} = h_{ki} \prod_{j=1}^{i-1} (1 - h_{kj})$. Il est facile de voir d'après (2.4) que ξ_k augmente (ou reste la même) à mesure que le niveau d'effort augmente, où le niveau d'effort est considéré en termes de traitements de suivi et de périodes de collecte des données. Cela donne à penser qu'il faut explorer les coûts et les avantages de l'accroissement du niveau d'effort vu que, dans certaines circonstances, il existe un certain nombre de traitements de suivi pour lesquels un pourcentage élevé du coût est déployé pour obtenir les valeurs auprès de quelques non-répondants.

2.2 Influence du suivi sur la probabilité de réponse

Nous exprimons maintenant la fonction de lien inverse du taux de risque sous forme d'une fonction de variables explicatives $\boldsymbol{x}_{ki}$ et d'un paramètre vectoriel $\boldsymbol{\beta}$ à estimer. Pour les unités sous collecte des données par autodénombrement, la fonction de lien inverse du taux de risque est exprimée sous la forme

$$g^{-1}(h_{ki}) = \eta(\boldsymbol{x}_{ki}^{(0)}, \boldsymbol{\beta}^{(0)}), \tag{2.5}$$

pour une fonction connue $\eta(.)$, où $\boldsymbol{x}_{ki}^{(0)}$ est le vecteur de variables explicatives pour l'autodénombrement, $\boldsymbol{\beta}^{(0)}$ est le paramètre vectoriel inconnu associé qui doit être estimé, $\boldsymbol{x}_{ki} = \boldsymbol{x}_{ki}^{(0)}$, $\boldsymbol{\beta} = \boldsymbol{\beta}^{(0)}$ et $g(.)$ est une fonction de lien — quoique la fonction de lien soit habituellement utilisée pour transformer (ou pour lier) la moyenne conditionnelle au prédicteur linéaire $\boldsymbol{x}_{ki}^T \boldsymbol{\beta}$. Par exemple, $g(a) = a$ avec $\eta(\boldsymbol{x}_{ki}, \boldsymbol{\beta}) = \boldsymbol{x}_{ki}^T \boldsymbol{\beta}$ donne un modèle de régression linéaire et $g(a) = \exp(a)/\{1 + \exp(a)\}$ avec $\eta(\boldsymbol{x}_{ki}, \boldsymbol{\beta}) = \boldsymbol{x}_{ki}^T \boldsymbol{\beta}$ donne un modèle de régression logistique pour des réponses binaires r_{ki}, où r_{ki} représente une série d'indicateurs de réponse définie pour chaque unité k dont les valeurs sont fixées à $r_{ki} = 1$ si l'unité répond à la période i et $r_{ki} = 0$ si l'unité ne répond pas à la période i.

Des influences supplémentaires sur la probabilité de réponse peuvent être étudiées en ajoutant d'autres prédicteurs au modèle de risque à temps discret initial. Par exemple, le modèle qui suit se distingue du modèle (2.5) par l'inclusion du prédicteur du suivi variant dans le temps $\gamma_{ki}^{(1)} \boldsymbol{x}_{ki}^{(1)}$, dont l'influence est saisie par le paramètre $\boldsymbol{\beta}^{(1)}$:

$$g^{-1}(h_{ki}) = \eta(\boldsymbol{x}_{ki}^{(0)}, \boldsymbol{\beta}^{(0)}; \gamma_{ki}^{(1)} \boldsymbol{x}_{ki}^{(1)}, \boldsymbol{\beta}^{(1)}), \tag{2.6}$$

où la valeur de $\gamma_{ki}^{((1)}$ est fixée à 1 si le premier traitement de suivi a débuté, ou à 0 si ce n'est pas le cas, avec $\boldsymbol{x}_{ki} = (\boldsymbol{x}_{ki}^{(0)T}, \gamma_{ki}^{(1)}\boldsymbol{x}_{ki}^{(1)T})^T$ et $\boldsymbol{\beta} = (\boldsymbol{\beta}^{(0)T}, \boldsymbol{\beta}^{(1)T})^T$. Notons que (2.6) peut être utilisé pour définir différentes pentes et ordonnées à l'origine, auquel cas le paramètre $\boldsymbol{\beta}^{(1)}$ reflète les variations des ordonnées à l'origine et des pentes associées au changement pour passer de l'autodénombrement seulement à l'autodénombrement suivi du premier traitement de suivi. Par exemple, dans la spécification $\eta(\boldsymbol{x}_{ki}, \boldsymbol{\beta}) = \boldsymbol{x}_{ki}^{(0)T}\boldsymbol{\beta}^{(0)} + \boldsymbol{x}_{ki}^{(1)T}\boldsymbol{\beta}^{(1)}$, $i = 1,\ldots,I$, avec $\boldsymbol{x}_{ki}^{(0)T}\boldsymbol{\beta}^{(0)} = \alpha_{0i}^{(0)} + x_{ki}\alpha_{1i}^{(0)}$ et $\boldsymbol{x}_{ki}^{(1)T}\boldsymbol{\beta}^{(1)} = \gamma_{ki}^{(1)}(\alpha_{0i}^{(1)} + x_{ki}\alpha_{1i}^{(1)})$, les paramètres de régression $\alpha_{0i}^{(0)}$, $\alpha_{1i}^{(0)}$ et les valeurs x_{ki} représentent, respectivement, l'ordonnée à l'origine, la pente et le prédicteur associés à la collecte des données par autodénombrement à la période i. Nous avons $\boldsymbol{x}_{ki}^{(0)} = (D_{k1}^{(0)},..., D_{kI}^{(0)}, x_{k1}^{(0)},..., x_{kI}^{(0)})^T$ et $\boldsymbol{\beta}^{(0)} = (\alpha_{01}^{(0)},..., \alpha_{0I}^{(0)}, \alpha_{11}^{(0)},..., \alpha_{1I}^{(0)})^T$, où $D_{ki}^{(0)} = 1$, $x_{ki}^{(0)} = x_{ki}$, $D_{kj}^{(0)} = 0$ et $x_{kj}^{(0)} = 0$ pour $j \neq i$. Le prédicteur vectoriel de suivi est donné par $\boldsymbol{x}_{ki}^{(1)} = (D_{k1}^{(1)},..., D_{kI}^{(1)}, x_{k1}^{(1)},..., x_{kI}^{(1)})^T$ et les variations des ordonnées à l'origine et des pentes attribuables au suivi sont reflétées par le paramètre vectoriel $\boldsymbol{\beta}^{(1)} = (\alpha_{01}^{(1)},..., \alpha_{0I}^{(1)}, \alpha_{11}^{(1)},..., \alpha_{1I}^{(1)})^T$, où $D_{ki}^{(1)} = 1$, $x_{ki}^{(1)} = x_{ki}$, $D_{kj}^{(1)} = 0$ et $x_{kj}^{(1)} = 0$ pour $j \neq i$. Afin d'accroître les taux de réponse, les non-répondants sont soumis à de multiples traitements de suivi intensif par téléphone ou d'autres traitements pour les encourager à participer. Un traitement peut prendre la forme de rappels envoyés par la poste, de rappels envoyés par courriel, d'appels téléphoniques ou d'interviews sur place. Le processus de suivi selon ces traitements est mené en utilisant des calendriers de collecte des données assortis d'une stratégie spécifique pour chaque unité échantillonnée. Dans le cas de $1+J$ traitements de suivi, la forme de lien inverse du taux de risque peut être exprimée par $g^{-1}(h_{ki}) = \eta(\boldsymbol{x}_{ki}, \boldsymbol{\beta})$, où $\boldsymbol{x}_{ki} = (\boldsymbol{x}_{ki}^{(0)T}, \gamma_{ki}^{(1)}\boldsymbol{x}_{ki}^{(1)T},..., \gamma_{ki}^{(J)}\boldsymbol{x}_{ki}^{(J)T})^T$ et $\boldsymbol{\beta} = (\boldsymbol{\beta}^{(0)T}, \boldsymbol{\beta}^{(1)T},..., \boldsymbol{\beta}^{(J)T})^T$.

Considérons le cas de $J = 1$ où T_1 correspond à un suivi intensif et T_0 correspond à l'envoi du questionnaire, et supposons pour simplifier que le résultat en ce qui concerne la réponse est instantané. Après avoir recueilli les réponses des répondants par autodénombrement, le suivi est effectué de manière déterministe — les non-répondants pour lesquels $u_k \geq \kappa$ sont affectés au traitement T_1, où κ est une constante prédéterminée et u est une variable auxiliaire pour laquelle des valeurs sont disponibles pour toutes les unités échantillonnées. Supposons que toutes les unités sous T_1 ont répondu, tandis que les autres unités ne l'ont pas encore fait. Nous avons $\xi_k = h_{k1} + (1-h_{k1})1 = 1$ pour l'unité k pour laquelle $u_k \geq \kappa$ et $\xi_k = h_{k1} + (1-h_{k1})0 = h_{k1}$ pour les unités pour lesquelles $u_k < \kappa$. Cela met en relief l' importance du suivi sur les probabilités de réponse.

3. Estimation

3.1 Fonction de vraisemblance et l'algorithme Espérance-Maximisation

Le mode de collecte des données est connu pour une unité qui répond à un moment donné durant la collecte des données, tandis qu'il est inconnu pour une unité qui est censurée. Définissons un vecteur de variables indicatrices tel que $\ell_{m;k} = 1$ si l'unité k utilise le mode m, et $\ell_{m;k} = 0$ sinon, où $\ell_k = (\ell_{1k},..., \ell_{Mk})^T$ représente les réalisations des variables aléatoires indépendantes et identiquement distribuées suivant une loi multinomiale, $Mult_M(1, \boldsymbol{\varphi})$, et $\boldsymbol{\varphi} = (\phi_1,..., \phi_M)^T$ est le vecteur des proportions de population avec $\sum_{m=1}^{M}\phi_m = 1$. Par conséquent, la vraisemblance des données complètes est donnée par

$$L_c(\boldsymbol{\Phi}) = \prod_k \prod_{m=1}^{M}\{\phi_m f_m(t_k)\}^{\ell_{m;k}}, \tag{3.1}$$

et le logarithme du rapport de vraisemblance des données complètes est donnée par $l_c(\boldsymbol{\Phi}) = \sum_k \sum_{m=1}^{M} \ell_{m;k}\{\log\phi_m + \log f_m(t_k)\}$, où $\boldsymbol{\Phi} = (\phi_1,..., \phi_{M-1}, \boldsymbol{\beta}^T)^T$, $f_m(t_k)$ est $f(t_k)$ pour le mode m,

$$f(t_k) = \Pr(t = I_k)^{\delta_k}\Pr(t > I_k)^{1-\delta_k}, \tag{3.2}$$

$\delta_k = 1$ si l'unité k n'est pas censurée et $\delta_k = 0$ si l'unité k est censurée. Si l'unité k est censurée, soit elle répondra durant une période future $t > I$ ou ne répondra jamais. Pour calculer les estimations du maximum de vraisemblance, nous utilisons l'algorithme d'espérance-maximisation (EM) introduit par Hartley (1958) — et formalisé et nommé par Dempster et coll. (1977) — qui est devenu un outil important pour trouver les estimations du maximum de vraisemblance considérées insolubles en pratique, par exemple en présence de données manquantes. En utilisant une

valeur initiale pour $\boldsymbol{\Phi}$, disons $\boldsymbol{\Phi}^{(b)}$, l'étape E de l'algorithme EM requiert le calcul d'une fonction de $\boldsymbol{\Phi}$, $Q(\boldsymbol{\Phi},\boldsymbol{\Phi}^{(b)})$, telle que $Q(\boldsymbol{\Phi},\boldsymbol{\Phi}^{(b)}) = E\{l_c(\boldsymbol{\Phi}) \mid \boldsymbol{l}^{(o)},\boldsymbol{\Phi}^{(b)}\}$, où $\boldsymbol{\Phi}$ est le paramètre d'intérêt, $\boldsymbol{\Phi}^{(b)}$ est la valeur de $\boldsymbol{\Phi}$ à l'itération précédente, et $\boldsymbol{l}^{(o)}$ est la donnée observée. Alors, l'étape M de l'algorithme EM tente de choisir la valeur de $\boldsymbol{\Phi}$, disons $\boldsymbol{\Phi}^{(b+1)}$, qui maximise $Q(\boldsymbol{\Phi},\boldsymbol{\Phi}^{(b)})$. Si nous itérons l'étape E et l'étape M jusqu'à la convergence, sous des conditions de régularité, l'algorithme convergera vers l'estimation du maximum de vraisemblance. Nous avons

$$Q(\boldsymbol{\Phi},\boldsymbol{\Phi}^{(b)}) = \sum_k \sum_{m=1}^{M} \tau_{m;k}^{(b)} \log\phi_m + \sum_k \sum_{m=1}^{M} \tau_{m;k}^{(b)} \log f_m(t_k), \qquad (3.3)$$

c'est-à-dire que chaque variable indicatrice $\ell_{m;k}$ est remplacée par $\tau_{m;k}^{(b)} = E(\ell_{m;k} \mid \boldsymbol{l}^{(o)},\boldsymbol{\Phi}^{(b)})$, son espérance conditionnelle sur les données observées $\boldsymbol{l}_k^{(o)}$. L'étape E de l'algorithme comprend la création d'un ensemble de « pseudo-données » dans lequel les répondants sont laissés intacts et les non-répondants sont fractionnés en M pseudo-observations partiellement complètes. Le poids attribué à ces pseudo-observations est la probabilité conditionnelle que l'unité appartienne à une population associée. L'espérance conditionnelle $\tau_{m;k}$ est la probabilité que l'unité k utilise le mode m étant donné les données observées et avant l'estimation des paramètres. Une fois que $\boldsymbol{\Phi}^{(b)}$ a été obtenue, les estimations des probabilités conditionnelles peuvent être formées pour chaque unité k. La probabilité conditionnelle qu'un non-répondant k utilise le mode m est donnée par $\tau_{m;k}^{(b)} = \{\sum_n \phi_n^{(b)} f_n^{(b)}(t_k)\}^{-1} \phi_m^{(b)} f_m^{(b)}(t_k)$, où $f_m^{(b)}(t_k)$ est $f(t_k,\boldsymbol{\beta}^{(b)})$ pour le mode m. Les estimations de $\boldsymbol{\Phi}$ et de $\tau_{m;k}$ sont alternées de façon répétée, en remplaçant dans leur exécution subséquente la valeur prédite initiale $\boldsymbol{\Phi}^{(b)}$ par la valeur prédite courante $\boldsymbol{\Phi}^{(b+1)}$ pour $\boldsymbol{\Phi}$. Du premier terme du deuxième membre de (3.3), nous obtenons l'estimation subséquente $\phi_m^{(b+1)}$ de ϕ_m en utilisant

$$S_c(\phi_m) = \sum_k \{\tau_{m;k}^{(b)} - \phi_m \sum_m \tau_{m;k}^{(b)}\} = 0. \qquad (3.4)$$

Le dernier terme du deuxième membre de (3.3) correspond à la distribution des réponses selon la période (réponses-période) et ne fait intervenir que $\boldsymbol{\beta}$. En substituant (2.2) et (2.3) aux termes correspondants dans (3.2), et en prenant le logarithme, nous obtenons

$$\log f(t_k) = \delta_k \log\{h_{kd_k}/(1-h_{kd_k})\} + \sum_{i=1}^{I_k} \log(1-h_{ki}). \qquad (3.5)$$

En utilisant la série d'indicateurs de réponse r_{ki}, l'expression (3.5) peut être réécrite (Allison, 1982) sous la forme $\log f(t_k) = \sum_{i=1}^{I_k} r_{ki} \log\{h_{ki}/(1-h_{ki})\} + \sum_{i=1}^{I_k} \log(1-h_{ki})$. En prenant les dérivées, nous obtenons

$$\partial \log f(t_k)/\partial\boldsymbol{\beta} = \sum_{i=1}^{I_k} \dot{\boldsymbol{h}}_{ki}(\boldsymbol{\beta})(r_{ki} - h_{ki})\{h_{ki}(1-h_{ki})\}^{-1}, \qquad (3.6)$$

où $\dot{\boldsymbol{h}}_{ki}(\boldsymbol{\beta}) = \partial h_{ki}/\partial\boldsymbol{\beta}$. Sous un cas de recensement, l'estimation subséquente $\boldsymbol{\beta}^{(b+1)}$ de $\boldsymbol{\beta}$ est la valeur qui satisfait

$$\boldsymbol{S}_c(\boldsymbol{\beta}) = \sum_k \sum_{m=1}^{M} \tau_{m;k}^{(b)} \partial \log f_m(t_k)/\partial\boldsymbol{\beta} = \boldsymbol{0}, \qquad (3.7)$$

où $\partial \log f(t_k)/\partial\boldsymbol{\beta}$ est donné par (3.6).

3.2 Estimation des paramètres d'après des données d'enquête

Supposons qu'un échantillon probabiliste s est tiré de la population finie et désignons par $d_k(s)$ le poids d'échantillonnage appliqué à l'unité k avec $d_k(s) = 0$ si $k \notin s$. Nous utilisons les poids de sondage pour estimer les équations estimantes de l'algorithme EM. L'équation estimante pondérée associée à (3.4) et à (3.7) est donnée par

$$\hat{\boldsymbol{S}}_c(\boldsymbol{\Phi}) = \sum_k d_k(s) \boldsymbol{s}_{c;k}(\boldsymbol{\Phi}) = \boldsymbol{0},$$

où $\boldsymbol{S}_c(\boldsymbol{\Phi}) = (\boldsymbol{S}_c^T(\phi_1),..., \boldsymbol{S}_c^T(\phi_{M-1}), \boldsymbol{S}_c^T(\boldsymbol{\beta}))^T$. Les paramètres d'intérêt doivent encore être estimés. Un estimateur $\hat{\boldsymbol{\Theta}}$ du paramètre $\boldsymbol{\Theta}$ associé par (1.1) est la solution de l'équation estimante pondérée

$$\hat{\boldsymbol{S}}(\boldsymbol{\Theta}) = \sum_k d_k(s)(r_k/\hat{\xi}_k)\boldsymbol{s}(y_k;\boldsymbol{\Theta}) - \boldsymbol{v}(\boldsymbol{\Theta}) = \boldsymbol{0},$$

avec $r_k = 1 - \prod_{i=1}^{I_k}(1-r_{ki}) = \sum_{i=1}^{I_k} r_{ki}$ et $\hat{\xi}_k = \xi_k(\hat{\boldsymbol{\beta}})$. Puisque les répondants dans un mode ne peuvent pas être considérés comme un sous-échantillon aléatoire de l'ensemble de la population, un estimateur $\hat{\boldsymbol{\Theta}}_m$ du paramètre d'intérêt $\boldsymbol{\Theta}_m$ associé au mode m est la solution de l'équation estimante pondérée

$$\hat{\boldsymbol{S}}(\boldsymbol{\Theta}_m) = \sum_k d_k(s)(r_k/\hat{\xi}_k)(l_{m;k}/\hat{\phi}_{m;k})\boldsymbol{s}(y_{m;k};\boldsymbol{\Theta}_m) - \boldsymbol{v}(\boldsymbol{\Theta}_m) = \boldsymbol{0}.$$

4. Étude en simulation

Nous avons réalisé une petite étude par simulations pour illustrer les rendements des estimateurs des paramètres du modèle, ainsi que les estimateurs des totaux de population finie en rapport avec le mode mixte de collecte des données. Nous avons généré des valeurs pour chaque unité k d'une population finie de taille N = 2 000 indépendamment au moyen du modèle

$$\begin{pmatrix} y_k \\ u_k \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_y \\ \mu_u \end{pmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right),$$

où y est la variable d'intérêt, u est la variable auxiliaire, $\mu_y = \mu_u = 2$ et $\rho = 0{,}8$. Nous avons fixé $M = 2$ et nous avons créé les proportions de collecte par la poste et par Internet en utilisant $\ell_k = (\ell_{M;k}, \ell_{I;k})^T \sim Mult_2(1, \boldsymbol{\varphi})$ avec $\boldsymbol{\varphi} = (\phi_M, \phi_I)^T = (0{,}6; 0{,}4)^T$ et les valeurs de la variable d'intérêt associée à chaque mode de collecte des données ont été générées au moyen de

$$y_{m;k} \mid y_k \sim N\left(\mu_m + \rho_m (y_k - \mu_y), (1 - \rho_m^2)\right) \text{ pour } m \in \{M, I\},$$

avec $\mu_M = E_m(y_M) = 3$, $\mu_I = E_m(y_I) = 4$, $\rho_M = \rho(y_M, y) = 0{,}3$ et $\rho_I = \rho(y_I, y) = 0{,}7$. Nous avons maintenu les valeurs de population $(y_k, u_k, y_{M;k}, y_{I;k}, \ell_{M;k}, \ell_{I;k})$ fixes pour $k = 1, ..., N$, et nous avons tiré $A = 200$ échantillons de Bernoulli, chacun avec une fraction d'échantillonnage $f = 0{,}5$, de la population générale. Nous avons posé que $I = 7$ et $J = 2$. Les non-répondants à la période 2 pour lesquels $u_k \geq \mu_u$ ont été affectés à T_1, tandis que les non-répondants à la période 4 ont été affectés à T_2. Dans ce cas, nous avons en tout deux stratégies S_1 et S_2, où $S_1 = (0,0,0,T_2,0,0,0)$ et $S_2 = (0,T_1,0,T_2,0,0,0)$. Nous avons créé les répondants et les non-répondants en utilisant $r_{ki} \sim B(1, h_{ki})$, avec $logit(h_{ki}) = \boldsymbol{x}_{ki}^T \boldsymbol{\beta}$, $x_{ki}^T \boldsymbol{\beta} = F(T_0) + F(T_1) + F(T_2)$, $F(T_0) = (\ell_{M;k} \alpha_M^{(0)} + \ell_{I;k} \alpha_I^{(0)}) t$, $F(T_1) = 1(u_k \geq \mu_u) 1(t \geq 3)(\ell_{M;k} \alpha_M^{(1)} + \ell_{I;k} \alpha_I^{(1)}) t_1$ et $F(T_2) = 1(t \geq 5) \alpha^{(2)} t_2$, où $1()$ est la fonction de vérité et t_j est la période à laquelle le traitement T_j a débuté. Le tableau 4-1 donne les valeurs des paramètres du modèle de réponse. La première composante $\boldsymbol{\varphi}_N$ du paramètre vectoriel $\boldsymbol{\Phi} = (\boldsymbol{\varphi}_N^T, \boldsymbol{\beta}^T)^T$ correspond au mode de collecte des données et fait intervenir les proportions de population finie $\boldsymbol{\varphi}_N = (\phi_{M;N}, \phi_{I;N})^T = N^{-1} \sum_k (\ell_{M;k}, \ell_{I;k})^T$, tandis que la deuxième composante $\boldsymbol{\beta}$ du paramètre vectoriel correspond au modèle de réponse et fait intervenir $\boldsymbol{\beta} = (\alpha_M^{(0)}, \alpha_I^{(0)}, \alpha_M^{(1)}, \alpha_I^{(1)}, \alpha^{(2)})^T$. Nous avons estimé $\boldsymbol{\Phi}$ en utilisant l'algorithme EM. Soit $\hat{\theta}$ un estimateur du paramètre d'intérêt θ. Nous avons calculé $\hat{\theta}$, à partir de chacune des répétitions a ($a = 1, ..., A$) et de leurs moyennes $\bar{\hat{\theta}} = A^{-1} \sum_{a=1}^{A} \hat{\theta}_a$, où $\hat{\theta}_a$ est la valeur de $\hat{\theta}$ pour le a^e échantillon. Le biais de simulation (B), le biais relatif (BR) et l'erreur quadratique moyenne (EQM) de $\hat{\theta}$ sont calculés comme $B(\hat{\theta}) = (\bar{\hat{\theta}} - \theta)$, $BR(\hat{\theta}) = B(\hat{\theta}) / \theta$ et $EQM(\hat{\theta}) = A^{-1} \sum_{a=1}^{A} (\hat{\theta}_a - \theta)^2$ respectivement. Le taux de réponse global pour les 200 répétitions varie d'un minimum de 0,72 à un maximum de 0,79 avec une moyenne de l'ordre de 0,76. Le minimum, le maximum et la moyenne des itérations EM valent 12, 17 et 15, respectivement. Tous les cas ont convergé. Nous avons calculé $B(\hat{\theta})$ pour les paramètres de régression, et ces valeurs sont présentées au tableau 4-1. Ce tableau démontre clairement que le biais est faible pour chaque paramètre de régression. Nous avons également considéré l'estimation du total de population finie : $\boldsymbol{\Theta} = (\sum_k y_k, \sum_k y_{m;k}; m \in \{M, I\})^T$. Nous nous sommes servis de deux ensembles de poids : le premier ensemble de poids utilise les paramètres estimés du modèle, $(d_k(s)(r_k / \hat{\xi}_k), d_k(s)(r_k / \hat{\xi}_k)(l_{m;k} / \hat{\varphi}_{m;k}); m \in \{M, I\})^T$, tandis que le deuxième utilise les paramètres réels du modèle, $(d_k(s)(r_k / \xi_k), d_k(s)(r_k / \xi_k)(l_{m;k} / \phi_{m;k}); m \in \{M, I\})^T$. Nous avons calculé $BR(\hat{\theta})$ et les ratios de l'EQM pour chaque estimateur $\hat{\theta}$ avec $\hat{Y} = \sum_k d_k(s)(r_k / \hat{\xi}_k) y_k$ et ces valeurs sont présentées au tableau 4-2. Ce tableau indique clairement que tous les biais relatifs sont faibles. L'estimateur utilisant un paramètre de régression estimé est plus efficace qu'un estimateur utilisant le paramètre de régression réel. Aux fins de comparaison, le tableau 4-2 donne aussi les résultats pour les estimateurs calés sur la taille de population, qui indiquent que le calage sur la taille de la population est hautement efficace.

Tableau 4-1
Biais pour les estimations des paramètres du modèle

Paramètre θ :	$\phi_{M;N}$	$\phi_{I;N}$	$\alpha_M^{(0)}$	$\alpha_I^{(0)}$	$\alpha_M^{(1)}$	$\alpha_I^{(1)}$	$\alpha^{(2)}$
Valeur	0,61	0,39	-0,7	-0,4	0,7	0,1	0,1
Estimation	0,60	0,40	-0,7	-0,4	0,7	0,09	0,09
$B(\hat{\theta})$	-0,0016	0,0016	0,0007	-0,0003	0,0012	-0,0055	-0,006

Tableau 4-2
Biais relatif et ratios des erreurs quadratiques moyennes pour les totaux de population finie

Paramètre d'intérêt	Poids	Biais relatif et (ratios des EQM) Paramètre du modèle			
		Estimé		Réel	
$\theta = \sum_k y_k$	Sondage	0,0002	(1,00)	-0,0004	(1,56)
	Calage*	-0,0002	(0,32)	-0,0005	(0,35)
$\theta_M = \sum_k y_{M;k}$	Sondage	-0,0024	(2,76)	0,0063	(6,46)
	Calage*	0,0035	(0,81)	-0,0036	(0,83)
$\theta_I = \sum_k y_{I;k}$	Sondage	0,0031	(3,66)	-0,0111	(14,98)
	Calage*	0,0037	(0,85)	0,0036	(0,91)

*Calage sur la taille de population

5. Conclusion

Le présent article décrit l'introduction d'un modèle de risque à temps discret dans l'analyse des indicateurs de réponse dans les enquêtes et les recensements. L'approche proposée facilite l'examen de la forme de la fonction de risque. Puisque l'inspection de la forme de la fonction de risque indique quand une réponse est la plus susceptible d'avoir lieu et comment la probabilité varie au fil du temps, la description des formes de la fonction de risque joue un rôle important dans la qualité et le coût d'une enquête. Nous avons fait appel à l'analyse de régression pour étudier l'effet d'un mode de collecte mixte sur la probabilité de réponse. Les estimateurs des paramètres du modèle ainsi que les estimateurs de population finie des enquêtes à mode de collecte mixte ont été présentés.

Bibliographie

Allison, P. D. (1982). « Discrete-time methods for the analysis of event histories ». S. Leinhardt (sous la dir. de). *Sociological Methodology*. San Francisco: Jossey–Bass. p. 61 à 98.

Dempster, A. P., N.M. Laird et D.B. Rubin. (1977). « Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion) ». *Journal of the Royal Statistical Society B.* 39, p. 1 à 38.

Hartley, H. O. (1958). « Maximum likelihood estimation from incomplete data ». *Biometrics*. 4, p. 174 à 194.

Rosenbaum, P. R. (1987). « Model-based direct adjustment ». *Journal of the American Statistical Association*. 82, p. 387 à 394.