Composante du produit nº 12-001-X
Division des méthodes d'enquêtes auprès des entreprises

Article

Méthodes bayésiennes pour un tableau de contingence à double entrée incomplet avec application aux sondages électoraux de l'État de l'Ohio

par Bo-Seung Choi, Jai Won Choi et Yousung Park

Juin 2009

Statistique Canada Statistics Canada

Canada

Techniques d'enquête, juin 2009
Vol. 35, N° 1, pp. 41-56
Statistique Canada, N° 12-001-X au catalogue

Méthodes bayésiennes pour un tableau de contingence à double entrée incomplet avec application aux sondages électoraux de l'État de l'Ohio

Bo-Seung Choi, Jai Won Choi et Yousung Park [1]

Résumé

Nous appliquons une méthode bayésienne pour résoudre le problème des solutions limites de l'estimation du maximum de vraisemblance (MV) dans un tableau de contingence à double entrée incomplet en utilisant un modèle log-linéaire et des lois a priori de Dirichlet. Nous comparons cinq lois a priori de Dirichlet pour estimer les probabilités multinomiales par case sous un modèle de non-réponse non ignorable. Trois de ces lois a priori ont été utilisées dans le cas d'un tableau à simple entrée incomplet et les deux autres sont deux nouvelles lois a priori proposées afin de tenir compte de la différence entre les profils de réponse des répondants et des électeurs indécis. Les estimations bayésiennes obtenues à l'aide des trois premières lois a priori n'ont pas systématiquement de meilleures propriétés que les estimations du MV, contrairement à ce qu'indiquaient des études antérieures, tandis que les deux nouvelles lois a priori donnent de meilleurs résultats que les trois lois a priori antérieures et que les estimations du MV chaque fois qu'est obtenue une solution limite. Nous utilisons quatre jeux de données provenant des sondages électoraux réalisés en 1998 dans l'État de l'Ohio pour illustrer comment il convient d'utiliser et d'interpréter les résultats des estimations pour les élections. Nous procédons à des études par simulation pour comparer les propriétés de cinq estimations bayésiennes sous un modèle de non-réponse non ignorable.

Mots clés : Analyse bayésienne ; non-réponse non ignorable ; tableau de contingence ; solution limite ; algorithme EM.

1. Introduction

Fréquente dans la plupart des sondages, la non-réponse devient un problème sérieux à mesure que son taux augmente (De Heer 1999 ; Groves et Couper 1998). Si l'on résume des données de sondage dans un tableau de contingence à double entrée, ce dernier contient des effectifs de case entièrement classifiés, des effectifs partiellement classifiés (c'est-à-dire non-réponse partielle) et des effectifs non classifiés (c'est-à-dire non-réponse totale). Par exemple, dans le sondage électoral de l'Ohio (Chen et Stasny 2003), l'une des catégories est la préférence de l'électeur (candidat A, B ou C, ou indécis) et l'autre est la probabilité de voter (votera vraisemblablement, ne votera vraisemblablement pas, et indécis). La première marge supplémentaire contient uniquement des données sur les préférences des électeurs, la deuxième contient uniquement des données sur la probabilité de voter et la troisième donne uniquement le nombre de non-réponses totales (cas où les deux réponses sont inconnues). Nous désirons intégrer ces observations manquantes dans l'estimation de l'appui réel pour chaque candidat et présenter des modèles bayésiens pour prédire le gagnant.

Dans certains sondages, les réponses indécises sont traitées comme une catégorie de réponse valide si les répondants ne manifestent pas de préférence nette pour un candidat ni d'intention nette de voter (Smith 1984 ; Rubin, Stern et Vehovar 1995). Néanmoins, de nombreuses études ont révélé que le comportement de vote des électeurs indécis peut avoir une incidence importante sur le résultat final et que l'on peut améliorer l'exactitude de la prédiction des résultats de l'élection en tenant compte de ces électeurs (Perry 1979 ; Fenwick, Wiseman, Becker et Heiman 1982 ; Myers et O'Connor 1983 ; Kim 1995 ; Chen et Stasny 2003 ; Martin, Traugott et Kennedy 2005). Parmi ces auteurs, Perry (1979) a montré au moyen de données empiriques obtenues selon une approche de scrutin secret que le pourcentage d'électeurs indécis observé dans un sondage est vraisemblablement plus élevé que le pourcentage réel. Kim (1995) a également indiqué que le rôle de ces électeurs indécis est critique, surtout quand leur nombre est supérieur à l'écart entre les deux candidats en tête dans une course électorale. Trois de nos études empiriques décrites à la section 3 appartiennent à ce cas critique. Fenwick et coll. (1982) ainsi que Kim (1995) ont procédé à une analyse discriminante des données du sondage électoral d'octobre 1980 au Massachusetts et des données de l'élection présidentielle américaine de 1992 d'après laquelle ils ont réparti les électeurs indécis entre les candidats afin de montrer qu'en général, ces électeurs indécis ne se rendent pas aux urnes dans les mêmes proportions que leurs homologues décidés. Si l'on met l'accent sur le candidat pour lequel pourrait voter l'électeur indécis, il est préférable de traiter les réponses indécises comme des données manquantes (Myers et O'Connor 1983). Comme l'ont mentionné Flannelly, Flannelly et McLeod (2000) et Lau

1. Bo-Seung Choi, directeur de recherche, Institut d'économie, Université de la Corée, Séoul 136-701, Corée ; Jai Won Choi, professeur, Département de biostatistique, Medical College of Georgia, Augusta, GA 30912 ; Yousung Park, professeur, Département de statistique, Université de la Corée, Séoul 136-701, Corée. Courriel : yspark@korea.ac.kr.

(1994), l'erreur de prédiction des résultats réels de l'élection augmente parallèlement au taux d'électeurs indécis. Afin de surmonter ce problème, Monterola, Lim, Garcia et Saloma (2001) ont adopté une approche basée sur des réseaux neuronaux pour classer les électeurs indécis dans un sondage d'opinion publique. Smith, Skinner et Clarke (1999) et Molenberghs, Kenward et Goetghebeur (2001) ont utilisé des méthodes d'imputation fondées sur un modèle pour le British General Election Panel Survey de 1992 et pour le sondage d'opinion publique réalisé à l'occasion du plébiscite de 1991 en Slovénie. Notre objectif principal étant d'obtenir des prédictions plus exactes grâce à l'affectation des électeurs indécis aux cases appropriées, nous avons traité ces derniers comme des observations manquantes à l'instar des chercheurs susmentionnés.

Les non-réponses (ou, de manière équivalente, les électeurs indécis) peuvent être réparties en trois catégories (Little et Rubin 2002, page 11), à savoir les données manquant complètement au hasard (MCAR pour *missing completely at random*), ce qui signifie que la probabilité d'une non-réponse pour une variable d'intérêt est indépendante de toute variable étudiée, y compris la variable elle-même, les données manquant au hasard (MAR pour *missing at random*), ce qui signifie que la probabilité d'une non-réponse dépend uniquement des données observées, et les données ne manquant pas au hasard (MNAR pour *missing not at random*), ce qui signifie que la probabilité d'une non-réponse dépend des valeurs non observées. Les modèles qui correspondent aux cas MCAR ou MAR sont appelés modèles de non-réponse ignorable, tandis que ceux correspondant aux cas MNAR sont appelés modèles de non-réponse non ignorable. Par exemple, dans un sondage préélectoral, si les personnes qui répondent n'indiquent pas leur préférence pour un candidat bien qu'elles appuient un candidat particulier, le profil des préférences pour les candidats pourrait ne pas être le même pour les répondants et les non-répondants. Dans ces conditions, le mécanisme de non-réponse est non ignorable. Si l'on suppose que le mécanisme de création des données manquantes est MCAR, l'effet de la non-réponse peut être éliminé dans l'inférence de la vraisemblance (Little et Rubin 2002, page 11). Cependant, si le profil de réponse des non-répondants diffère de celui des répondants, le fait d'écarter les non-réponses ou de spécifier incorrectement le mécanisme de création de la non-réponse donne lieu à des estimations entachées d'une plus grand variance et d'un plus grand biais (Chen 1972 ; Park et Brown 1994).

Dans les tableaux de contingence, si la non-réponse est non ignorable, l'estimation du MV donne fréquemment des solutions limites où il est estimé que la probabilité de non-réponse est nulle dans certaines cases. Ces solutions limites fournissent souvent un maximum local de la fonction de vraisemblance. Dans ce cas, les estimations du maximum de vraisemblance (MV) des paramètres du modèle log-linéaire ne peuvent pas avoir de solution unique et présentent habituellement de grands écarts-types (voir la section 4 ou Baker, Rosenberger et Dersimonian (1992), ainsi que Park et Brown (1994) pour des discussions plus détaillées).

Les conditions dans lesquelles l'estimation du MV se situe sur la solution limite ont été proposées dans le cas d'un tableau de contingence à simple entrée (Baker et Laird 1988 ; Michiels et Molenbergs 1997). L'explication géométrique de la solution limite de l'estimation du MV a été présentée (Smith et coll. 1999 ; Clark 2002). Baker et coll. (1992) ont énoncé une condition suffisante et nécessaire sous laquelle l'estimation du MV peut avoir une solution limite dans un tableau de contingence à double entrée.

Afin de contourner ce genre de problème de solution limite dans l'estimation du MV en présence de non-réponse non ignorable, Park et Brown (1994) et Park (1998) ont proposé une approche bayésienne au moyen de lois a priori empiriques basées uniquement sur l'information fournie par les répondants. Clogg, Rubin, Schenker et Schultz (1991) ont utilisé une loi a priori constante pour un tableau de contingence à simple entrée incomplet. Ils ont montré qu'en cas de non-réponse non ignorable, les méthodes bayésiennes donnent des erreurs quadratiques moyennes (EQM) plus petites que l'estimation du MV pour les espérances par case, mais notre étude par simulation révèle que cela n'est généralement pas vérifié dans un tableau de contingence à double entrée incomplet. Donc, nous présentons deux modèles bayésiens dont les lois a priori dépendent de l'information provenant à la fois des répondants et des électeurs indécis. Puis, nous appliquons chacun d'eux pour analyser le tableau de contingence à double entrée incomplet. L'extension à un tableau à multiples entrées est simple. Nous pouvons appliquer facilement cette extension à des données pondérées issues d'un échantillonnage stratifié ou en grappes en utilisant les covariables appropriées (voir la section 2.2).

La suite de l'article est divisée en quatre sections. À la section 2, nous considérons des modèles bayésiens avec cinq lois a priori différentes et présentons un algorithme Espérance-Maximisation (EM) généralisé pour estimer les probabilités par case. À la section 3, nous appliquons les modèles bayésiens à quatre jeux de données empiriques provenant du sondage électoral de l'État de l'Ohio et nous comparons les estimations bayésiennes à l'estimation du MV ainsi qu'aux résultats réels de l'élection. À la section 4, nous recourons à des études par simulation pour comparer les EQM et les biais des estimations bayésiennes fondées sur différents pourcentages de données manquantes et profils de réponse des répondants et des non-répondants. Dans cette

section, nous calculons également la probabilité de couverture, afin d'examiner la performance des estimations bayésiennes. À la section 5, nous présentons certaines conclusions.

2. Modèles bayésiens

Nous discutons de cinq estimations bayésiennes pour traiter la non-réponse non ignorable dans un tableau de contingence à double entrée incomplet. À la section 2.1, nous présentons un algorithme EM pour aborder le problème de la non-réponse dans un tableau de contingence à double entrée. Puis, à la section 2.2, nous spécifions cinq lois a priori et étendons notre approche à un tableau de contingence à multiples entrées.

Soit X_1 et X_2 ayant pour indice I et J catégories respectivement, dans un tableau de contingence à double entrée. Posons aussi que $R_1 = 1$ quand X_1 est observée et $R_1 = 2$ quand X_1 manque. De même, $R_2 = 1$ quand X_2 est observée et $R_2 = 2$ quand X_2 manque. Alors, la série complète des X_1, X_2, R_1, et R_2 produit un tableau de contingence $I \times J \times 2 \times 2$ contenant des effectifs entièrement classifiés, des effectifs partiellement classifiés et des effectifs non classifiés. Afin de distinguer ces trois catégories d'observations, représentons par y_{ijkl} l'effectif appartenant à la i^{e} catégorie de X_1, la j^{e} catégorie de X_2, la k^{e} valeur de R_1 et la l^{e} valeur de R_2. Donc, nous utilisons y_{ij11} pour les effectifs complètement classifiés, y_{i+12} et y_{+j21} pour les marges supplémentaires de colonne et de ligne respectives, et y_{++22} pour les effectifs non classifiés. Nous supposons que ces trois catégories d'observations suivent une loi multinomiale afin d'obtenir la log-vraisemblance suivante :

$$l = \sum_i \sum_j y_{ij11} \cdot \log(\pi_{ij11}) + \sum_i y_{i+12} \cdot \log(\pi_{i+12}) + \sum_j y_{+j21} \cdot \log(\pi_{+j21}) + y_{++22} \cdot \log(\pi_{++22}) \qquad (1)$$

où $\pi_{ijkl} = \Pr[X_1 = i,\ X_2 = j,\ R_1 = k,\ R_2 = l]$ et $N = \sum_{i,j,k,l} y_{ijkl}$ est fixé.

Comme cette fonction de vraisemblance contient plus de paramètres que le nombre de degrés de liberté disponibles pour l'estimation, nous relions π_{ijkl} à des covariables pertinentes en utilisant une fonction log-linéaire. Puisque nous ne disposons pas de variables explicatives, nous n'en utilisons aucune. Cependant, des variables explicatives peuvent être intégrées facilement dans le modèle log-linéaire de la même façon qu'y sont intégrées les variables catégoriques (voir Baker et Laird 1988, ainsi que Park et Brown 1994 pour plus de précisions).

Nous définissons un modèle de non-réponse non ignorable pour toutes les variables X_1, X_2, R_1 et R_2 par

$$\log(m_{ijkl}) = \beta_0 + \beta^i_{X_1} + \beta^j_{X_2} + \beta^k_{R_1} + \beta^l_{R_2} + \beta^{ik}_{X_1R_1} + \beta^{jl}_{X_2R_2} + \beta^{ij}_{X_1X_2} + \beta^{kl}_{R_1R_2}$$

$$\text{pour } i = 1, \ldots, I,\ j = 1, \ldots, J,\ k = 1, 2, \text{ et } l = 1, 2 \qquad (2)$$

où $m_{ijkl} = N \cdot \pi_{ijkl}$ est la fréquence (ou effectif) par case attendue pour la $(i, j, k, l)^{\text{e}}$ catégorie et la somme de chaque terme β sur n'importe lequel de ses indices supérieurs respectifs est nulle.

Ce modèle log-linéaire est saturé, puisque le nombre de paramètres est exactement le même que le nombre de cases observées provenant du tableau de convergence à double entrée incomplet. Il s'agit aussi d'un modèle de non-réponse non ignorable à cause des termes d'interaction entre X_1 et R_1 et entre X_2 et R_2, ce qui sous-entend que, pour chaque variable de réponse, la non-réponse dépend de la situation de cette variable. Le modèle log-linéaire est un outil utilisé fréquemment pour analyser les tableaux de contingence incomplets avec non-réponse non ignorable. Soit p le nombre de paramètres (c'est-à-dire β) à estimer. Nous introduisons le $p \times 1$ vecteur de plan d'expérience $\mathbf{z}_{ijkl}$ pour préciser l'affiliation de l'observation appartenant à la $(i, j, k, l)^{\text{e}}$ catégorie. Alors, le modèle log-linéaire donné en (2) peut être réécrit sous la forme

$$\log \mathbf{m} = Z\boldsymbol{\beta} \qquad (3)$$

où le vecteur $\mathbf{m}$ de dimensions $I \times J \times 2 \times 2$ est l'espérance par case et $\boldsymbol{\beta}$ est la représentation vectorielle des β. Afin d'éviter une solution limite de l'estimation du MV dans le modèle (2), nous imposons aux probabilités par case $(\pi_{ij11}, \pi_{ij12}, \pi_{ij21}, \pi_{ij22})$ des lois a priori de Dirichlet donnés par

$$\prod_i \prod_j \pi_{ij11}^{\delta_{ij11}} \cdot \pi_{ij12}^{\delta_{ij12}} \cdot \pi_{ij21}^{\delta_{ij21}} \cdot \pi_{ij22}^{\delta_{ij22}} \qquad (4)$$

où les hyperparamètres δ_{ijkl} sont spécifiés à la section 2.2. Ces lois a priori de Dirichlet produisent une forme explicite et commode d'une loi a posteriori, parce qu'ils sont conjugués à une loi multinomiale (Clogg et coll. 1991 ; Park et Brown 1994 ; Forster et Smith 1998). Avec (3), la loi multinomiale de (1) pour les observations et la loi a priori (4), nous obtenons la log-distribution a posteriori :

$$l_{pos} = \sum_i \sum_j y_{ij11} \cdot (\mathbf{z}_{ij11} \cdot \boldsymbol{\beta}) - \sum_i \sum_j y_{ij11} \cdot \log\left(\sum_{i,j,k,l} \exp(\mathbf{z}_{ijkl} \cdot \boldsymbol{\beta}) \right) + \sum_i y_{i+12} \cdot \log\left(\sum_j \exp(\mathbf{z}_{ij12} \cdot \boldsymbol{\beta}) \right)$$

$$
\begin{aligned}
&- \sum_i y_{i+12} \cdot \log\Big(\sum_{i,j,k,l} \exp(\mathbf{z}_{ijkl} \cdot \boldsymbol{\beta})\Big) \\
&+ \sum_j y_{+j21} \cdot \log\Big(\sum_i \exp(\mathbf{z}_{ij21} \cdot \boldsymbol{\beta})\Big) \\
&- \sum_j y_{+j21} \cdot \log\Big(\sum_{i,j,k,l} \exp(\mathbf{z}_{ijkl} \cdot \boldsymbol{\beta})\Big) \\
&+ y_{++22} \cdot \log\Big(\sum_i \sum_j \exp(\mathbf{z}_{ij22} \cdot \boldsymbol{\beta})\Big) \\
&- y_{++22} \cdot \log\Big(\sum_{i,j,k,l} \exp(\mathbf{z}_{ijkl} \cdot \boldsymbol{\beta})\Big) \\
&+ \sum_{i,j,k,l} \delta_{ijkl} \cdot (\mathbf{z}_{ijkl} \cdot \boldsymbol{\beta}) \\
&- \sum_{i,j,k,l} \delta_{ijkl} \cdot \log\Big(\sum_{i,j,k,l} \exp(\mathbf{z}_{ijkl} \cdot \boldsymbol{\beta})\Big). \qquad (5)
\end{aligned}
$$

L'équation (5) est assez complexe et nous utilisons donc l'algorithme EM pour estimer les paramètres (c'est-à-dire $\boldsymbol{\beta}$).

2.1 L'algorithme EM

Nous maximisons la loi a posteriori donnée en (5) sur le paramètre $\boldsymbol{\beta}$ en utilisant l'algorithme Espérance-Maximisation généralisé (EMG) (Dempster, Laird et Rubin 1977) comportant les étapes E et M qui suivent.

Étape E : En utilisant les variables augmentées y_{ij12}, y_{ij21}, et y_{ij22} pour $i = 1, ..., I$ et $j = 1, ..., J$, la loi a posteriori (5) peut s'écrire sous la forme

$$
\begin{aligned}
l_{a.\text{pos}} &= \sum_i \sum_j (y_{ij11} + \delta_{ij11}) \log(\pi_{ij11}) \\
&+ \sum_i \sum_j (y_{ij12} + \delta_{ij12}) \log(\pi_{ij12}) \\
&+ \sum_i \sum_j (y_{ij21} + \delta_{ij21}) \log(\pi_{ij21}) \\
&+ \sum_i \sum_j (y_{ij22} + \delta_{ij22}) \log(\pi_{ij22}). \qquad (6)
\end{aligned}
$$

Pour déterminer l'espérance de la loi a posteriori du log donné par (6), nous calculons la moyenne sur les effectifs manquants y_{ij12}, y_{ij21} et y_{ij22}, sachant les estimations courantes des paramètres, π_{ijkl}^{old} et les sommes marginales y_{i+12}, y_{+j21} et y_{++22} :

$$
\begin{aligned}
E_{\text{old}}[l_{a.\text{pos}}] &= \sum_i \sum_j (y_{ij11} + \delta_{ij11}) \cdot \log(\pi_{ij11}) \\
&+ \sum_i \sum_j (E_{\text{old}}[y_{ij12} \mid \pi_{ijkl}^{\text{old}}, y_{i+12}] + \delta_{ij12}) \cdot \log(\pi_{ij12}) \\
&+ \sum_i \sum_j (E_{\text{old}}[y_{ij21} \mid \pi_{ijkl}^{\text{old}}, y_{+j21}] + \delta_{ij21}) \cdot \log(\pi_{ij21}) \\
&+ \sum_i \sum_j (E_{\text{old}}[y_{ij22} \mid \pi_{ijkl}^{\text{old}}, y_{++22}] + \delta_{ij22}) \cdot \log(\pi_{ij22}). \qquad (7)
\end{aligned}
$$

Puisque y_{ij12}, y_{ij21} et y_{ij22} sont des variables aléatoires multinomiales conditionnées sur les sommes marginales respectives y_{i+12}, y_{+j21} et y_{++22}, dans l'équation (7), les espérances conditionnelles sont données par

$$
E_{\text{old}}(y_{ij12} \mid \pi_{ijkl}^{\text{old}}, y_{i+12}) = y_{i+12} \frac{m_{ij12}^{\text{old}}}{m_{i+12}^{\text{old}}},
$$

$$
E_{\text{old}}(y_{ij21} \mid \pi_{ijkl}^{\text{old}}, y_{+j21}) = y_{+j21} \frac{m_{ij21}^{\text{old}}}{m_{+j21}^{\text{old}}},
$$

et

$$
E_{\text{old}}(y_{ij22} \mid \pi_{ijkl}^{\text{old}}, y_{++22}) = y_{++22} \frac{m_{ij22}^{\text{old}}}{m_{++22}^{\text{old}}}
$$

où $m_{ijkl}^{\text{old}} = N \cdot \pi_{ijkl}^{\text{old}}$.

Étape M : À cette étape, nous maximisons l'espérance de la loi a posteriori du log (7) en utilisant les pseudo-observations $\tilde{y}_{ij11} = y_{ij11} + \delta_{ij11}$, $\tilde{y}_{ij12} = y_{i+12}\, m_{ij12}^{\text{old}} / m_{i+12}^{\text{old}} + \delta_{ij12}$, $\tilde{y}_{ij21} = y_{+j21}\, m_{ij21}^{\text{old}} / m_{+j21}^{\text{old}} + \delta_{ij21}$ et $\tilde{y}_{ij22} = y_{++22}\, m_{ij22}^{\text{old}} / m_{++22}^{\text{old}} + \delta_{ij22}$. Nous imposons à ces pseudo-observations que leurs sommes marginales soient égales aux sommes marginales correspondantes des observations $\tilde{y}_{++11} = y_{++11}$, $\tilde{y}_{i+12} = y_{i+12}$, $\tilde{y}_{+j21} = y_{+j21}$ et $\tilde{y}_{++22} = y_{++22}$. Sous ces contraintes, les pseudo-observations sont alors

$$
y_{ijkl}^{*} = \begin{cases}
\tilde{y}_{ij11} \dfrac{y_{++11}}{y_{++11} + \delta_{++11}} & \text{pour k = 1 et l = 1} \\
\tilde{y}_{ij12} \dfrac{y_{i+12}}{y_{i+12} + \delta_{i+12}} & \text{pour k = 1 et l = 2} \\
\tilde{y}_{ij21} \dfrac{y_{+j21}}{y_{+j21} + \delta_{+j21}} & \text{pour k = 2 et l = 1} \\
\tilde{y}_{ij22} \dfrac{y_{++22}}{y_{++22} + \delta_{++22}} & \text{pour k = 2 et l = 2.}
\end{cases}
$$

Alors, l'espérance de la fonction a posteriori du log (7) devient

$$E_{\text{old}}[l_{a.\text{pos}}] = \sum_i \sum_j y^*_{ij11} \cdot \log(\pi_{ij11}) + \sum_i \sum_j y^*_{ij12} \cdot \log(\pi_{ij12}) + \sum_i \sum_j y^*_{ij21} \cdot \log(\pi_{ij21}) + \sum_i \sum_j y^*_{ij22} \cdot \log(\pi_{ij22}).$$

Cette équation a la même forme que la vraisemblance obtenue à partir d'un tableau de contingence à quatre entrées quand les effectifs par case y^*_{ijkl} sont entièrement observés. Donc, en utilisant la méthode itérative des moindres carrés repondérés (Agresti 2002, page 342), nous obtenons l'estimateur du maximum a posteriori (EMP) de $\boldsymbol{\beta}$ comme il suit :

$$\boldsymbol{\beta}^{(t+1)} = (Z^T \hat{V}_t^{-1} Z)^{-1} Z^T \hat{V}_t^{-1} \gamma^{(t)},$$

où $\gamma^{(t)}$ possède l'élément $\gamma^{(t)}_{ijkl} = \log m^{(t)}_{ijkl} + (y_{ijkl} - m^{(t)}_{ijkl})/m^{(t)}_{ijkl}$ et $\hat{V}_t = [\text{diag}(\mathbf{m}^{(t)})]^{-1}$. Enfin, nous itérons ces étapes E et M jusqu'à ce qu'un critère de convergence soit atteint. Nous choisissons comme critère de convergence $\varepsilon \leq 10^{-6}$, où ε est la différence entre deux fonctions a posteriori du log consécutives.

Soit $Y_{\text{obs}} = (y_{ij11}, y_{i+12}, y_{+j21}, y_{++22})$ et $Y_{\text{manq}} = (y_{ij12}, y_{ij21}, y_{ij22})$ pour $i = 1, \ldots, I$ et $j = 1, \ldots, J$ le vecteur des effectifs observés et le vecteur des effectifs manquants, respectivement. Alors, la log-distribution a posteriori (5) peut s'écrire

$$l_{\text{pos}} = l(\boldsymbol{\beta} \mid Y_{\text{obs}}) = l(\boldsymbol{\beta} \mid Y_{\text{obs}}, Y_{\text{manq}}) - \log f(Y_{\text{manq}} \mid Y_{\text{obs}}, \boldsymbol{\beta}). \quad (8)$$

Par double dérivation par rapport à $\boldsymbol{\beta}$, (8) donne

$$\frac{\partial^2 l(\boldsymbol{\beta} \mid Y_{\text{obs}})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} = \frac{\partial^2 l(\boldsymbol{\beta} \mid Y_{\text{obs}}, Y_{\text{manq}})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} - \frac{\partial^2 \log f(Y_{\text{manq}} \mid Y_{\text{obs}}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}$$

$$= -\mathbf{Z}^T [\text{diag}(\mathbf{m}) - \mathbf{m}\mathbf{m}^T/N] \mathbf{Z} + \mathbf{Z}^T [\text{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^T] AB\mathbf{Z}, \quad (9)$$

où $\boldsymbol{\pi}$ est l'expression vectorielle des probabilités par case π_{ijkl} et A, B sont données par

$$A = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & \text{diag}\left(\dfrac{y^2_{i+12}}{y_{i+12} + \delta_{i+12}} \dfrac{m_{ij12}}{m_{i+12}}\right) & 0 & 0 \\ 0 & 0 & \text{diag}\left(\dfrac{y^2_{+j21}}{y_{+j21} + \delta_{+j21}} \dfrac{m_{ij21}}{m_{+j21}}\right) & 0 \\ 0 & 0 & 0 & \text{diag}\left(\dfrac{y^2_{++22}}{y_{++22} + \delta_{++22}} \dfrac{m_{ij22}}{m_{++22}}\right) \end{pmatrix}$$

et

$$B = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & I_{IJ} - B^{12} & 0 & 0 \\ 0 & 0 & I_{IJ} - B^{21} & 0 \\ 0 & 0 & 0 & I_{IJ} - B^{22} \end{pmatrix}.$$

Ici, faute d'espace et puisque l'extension pour i et j généraux n'est pas difficile, nous illustrons B^{12}, B^{21} et B^{22} uniquement pour $I = 2$ et $J = 3$:

$$B^{12} = \begin{pmatrix} \frac{m_{1112}}{m_{1+12}} & \frac{m_{1212}}{m_{1+12}} & \frac{m_{1312}}{m_{1+12}} & 0 & 0 & 0 \\ \frac{m_{1112}}{m_{1+12}} & \frac{m_{1212}}{m_{1+12}} & \frac{m_{1312}}{m_{1+12}} & 0 & 0 & 0 \\ \frac{m_{1112}}{m_{1+12}} & \frac{m_{1212}}{m_{1+12}} & \frac{m_{1312}}{m_{1+12}} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{m_{2112}}{m_{2+12}} & \frac{m_{2112}}{m_{2+12}} & \frac{m_{2112}}{m_{2+12}} \\ 0 & 0 & 0 & \frac{m_{2212}}{m_{2+12}} & \frac{m_{2212}}{m_{2+12}} & \frac{m_{2212}}{m_{2+12}} \\ 0 & 0 & 0 & \frac{m_{2312}}{m_{2+12}} & \frac{m_{2312}}{m_{2+12}} & \frac{m_{2312}}{m_{2+12}} \end{pmatrix},$$

$$B^{21} = \begin{pmatrix} \frac{m_{1121}}{m_{+121}} & 0 & 0 & \frac{m_{2121}}{m_{+121}} & 0 & 0 \\ 0 & \frac{m_{1221}}{m_{+221}} & 0 & 0 & \frac{m_{2221}}{m_{+221}} & 0 \\ 0 & 0 & \frac{m_{1321}}{m_{+321}} & 0 & 0 & \frac{m_{2321}}{m_{+321}} \\ \frac{m_{1121}}{m_{+121}} & 0 & 0 & \frac{m_{2121}}{m_{+121}} & 0 & 0 \\ 0 & \frac{m_{1221}}{m_{+221}} & 0 & 0 & \frac{m_{2221}}{m_{+221}} & 0 \\ 0 & 0 & \frac{m_{1321}}{m_{+321}} & 0 & 0 & \frac{m_{2321}}{m_{+321}} \end{pmatrix},$$

et

$$B^{22} = \begin{pmatrix} \frac{m_{1122}}{m_{++22}} & \frac{m_{1222}}{m_{++22}} & \frac{m_{1322}}{m_{++22}} & \frac{m_{2122}}{m_{++22}} & \frac{m_{2222}}{m_{++22}} & \frac{m_{2322}}{m_{++22}} \\ \frac{m_{1122}}{m_{++22}} & \frac{m_{1222}}{m_{++22}} & \frac{m_{1322}}{m_{++22}} & \frac{m_{2122}}{m_{++22}} & \frac{m_{2222}}{m_{++22}} & \frac{m_{2322}}{m_{++22}} \\ \frac{m_{1122}}{m_{++22}} & \frac{m_{1222}}{m_{++22}} & \frac{m_{1322}}{m_{++22}} & \frac{m_{2122}}{m_{++22}} & \frac{m_{2222}}{m_{++22}} & \frac{m_{2322}}{m_{++22}} \\ \frac{m_{1122}}{m_{++22}} & \frac{m_{1222}}{m_{++22}} & \frac{m_{1322}}{m_{++22}} & \frac{m_{2122}}{m_{++22}} & \frac{m_{2222}}{m_{++22}} & \frac{m_{2322}}{m_{++22}} \\ \frac{m_{1122}}{m_{++22}} & \frac{m_{1222}}{m_{++22}} & \frac{m_{1322}}{m_{++22}} & \frac{m_{2122}}{m_{++22}} & \frac{m_{2222}}{m_{++22}} & \frac{m_{2322}}{m_{++22}} \\ \frac{m_{1122}}{m_{++22}} & \frac{m_{1222}}{m_{++22}} & \frac{m_{1322}}{m_{++22}} & \frac{m_{2122}}{m_{++22}} & \frac{m_{2222}}{m_{++22}} & \frac{m_{2322}}{m_{++22}} \end{pmatrix}.$$

Nous constatons que l'information provenant des données observées $\partial^2 l(\boldsymbol{\beta} \mid Y_{\text{obs}}) / \partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T$ est égale à la différence entre l'information provenant des données augmentées et celle provenant des données manquantes. Comme le montrent Gelman, Carlin, Stern et Rubin (2004, page 103), l'inverse de l'information provenant des données observées évalué à l'EMP de $\boldsymbol{\beta}$ est la variance de l'EMP de $\boldsymbol{\beta}$.

2.2 Spécification des lois a priori

Afin d'achever l'algorithme EM, nous devons déterminer les hyperparamètres δ_{ijkl}. Nous posons que la somme des lois a priori $\sum_{i,j,k,l} \delta_{ijkl}$ est égale au nombre de paramètres intervenant dans le modèle log-linéaire, p, comme l'ont suggéré Clogg et coll. (1991). Sous cette contrainte, nous proposons cinq types de lois a priori de la façon suivante. Nous commençons par attribuer δ_{ijkl} de manière que l'EMP de m_{ijkl} diminue et se rapproche de l'EMV obtenu sous non-réponse ignorable. Autrement dit, nous déterminons les lois a priori δ_{ijkl} uniquement en fonction des effectifs de réponses connues y_{ij11} et nous les appelons lois a priori axées sur les répondants.

Le premier type de loi a priori axée sur les répondants est, pour tout $i = 1, \ldots, I$ et $j = 1, \ldots, J$,

$$\begin{aligned} \delta_{ij11} &= \nabla_{11} \frac{y_{ij11}}{y_{++11}}, \delta_{ij12} = \nabla_{12} \frac{y_{ij11}}{y_{++11}}, \delta_{ij21} \\ &= \nabla_{21} \frac{y_{ij11}}{y_{++11}}, \text{ et } \delta_{ij22} = \nabla_{22} \frac{y_{ij11}}{y_{++11}} \end{aligned} \qquad (10)$$

où $\nabla_{kl} = p \cdot y_{++kl} / y_{++++}$ pour $k = 1, 2$ et $l = 1, 2$. Le deuxième type de loi a priori axée sur les répondants ne donne aucune loi a priori (c'est-à-dire, comme il est décrit plus bas, qu'aucune loi a priori n'est nécessaire) sur π_{ij11} dans le premier type de lois a priori. En d'autres termes, le deuxième type est identique au premier type excepté que $\delta_{ij11} = 0$ pour tout i et j. Dans le cas d'un tableau de contingence à simple entrée (c'est-à-dire que X_1 ou X_2 est entièrement observée sans information manquante) et $y_{++22} = 0$, le premier type se réduit aux lois a priori utilisées dans Park (1998), tandis que le deuxième type se réduit aux lois a priori utilisées dans Park et Brown (1994). Ces deux types de lois a priori axées sur les répondants pourraient être trop simplistes, parce que l'on suppose généralement que le profil de réponse des non-répondants diffère de celui des répondants sous un modèle de non-réponse non ignorable. Par exemple, le candidat préféré des non-répondants pourrait ne pas être le même que celui des répondants dans un sondage préélectoral.

Afin de définir le troisième type de loi a priori, désignons par $\hat{m}_{ijkl}$ l'EMV de m_{ijkl}. Nous pouvons tirer la forme explicite de $\hat{m}_{ijkl}$ de Baker et coll. (1992) où certains $\hat{m}_{ijkl}$ pourraient être nuls à cause des solutions limites. Par exemple, quand une marge supplémentaire de colonne possède une solution limite dans un tableau 2×2 incomplet, les EMV sont $\hat{m}_{1j11} = y_{1j11}$,

$$\hat{m}_{2j11} = \frac{y_{2+11}(y_{2j11} + y_{+j21})}{y_{2+11} + y_{++21}}, \; \hat{m}_{ij12} = \hat{m}_{ij11} b_j$$

où b_j est la solution de $\sum_{j=1}^{2} y_{ij11} b_j = y_{i+12}$, $\hat{m}_{1j21} = 0$,

$$\hat{m}_{2j21} = \hat{m}_{2j11} \frac{y_{++21}}{y_{2+11}},\ \hat{m}_{1j22} = 0,$$

et $\hat{m}_{2j22} = \hat{m}_{2j12}\, y_{++22} / y_{2+12}$. Par conséquent, ces estimations du MV tiennent compte à la fois de l'information des répondants et des non-répondants. Nous pouvons aussi obtenir les estimations du MV au moyen de notre algorithme EM décrit à la section 2.1 en posant que $\delta_{ijkl} = 0$ pour tout i, j, k et l. En utilisant ces estimations du MV, nous définissons le troisième type de loi a priori sous la forme

$$\begin{aligned}
\delta_{ij11} &= \nabla_{11} \cdot \left(\frac{\hat{m}_{ij11}}{\hat{m}_{++11}} \right), \delta_{ij12} \\
&= \nabla_{12} \cdot \left(\frac{\hat{m}_{ij12}}{\hat{m}_{++12}} + \frac{1}{I \cdot J} \right) \cdot \frac{1}{2}, \\
\delta_{ij21} &= \nabla_{21} \cdot \left(\frac{\hat{m}_{ij21}}{\hat{m}_{++21}} + \frac{1}{I \cdot J} \right) \cdot \frac{1}{2}, \\
\text{et} \qquad & \qquad (11) \\
\delta_{ij22} &= \nabla_{22} \cdot \left(\frac{\hat{m}_{ij22}}{\hat{m}_{++22}} + \frac{1}{I \cdot J} \right) \cdot \frac{1}{2}
\end{aligned}$$

où $\nabla_{kl} = p \cdot \hat{m}_{++kl} / \hat{m}_{++++}$ pour $k, l = 1, 2,$ et le terme $1/IJ$ est la loi a priori constante de Clogg et coll. (1991) afin d'éviter d'éventuelles solutions limites pour m_{ij12}, m_{ij21}, et m_{ij22} (voir aussi la cinquième loi a priori plus loin). Donc, nous attribuons la troisième loi a priori de δ_{ijkl} pour que l'EMP de m_{ijkl} diminue et se rapproche de l'EMV obtenu sous le modèle de non-réponse non ignorable, tandis que la première loi a priori est obtenue sous le modèle de non-réponse ignorable.

Nous définissons le quatrième type de loi a priori en posant que $\delta_{ij11} = 0$ dans (11) comme nous l'avons fait pour obtenir le deuxième type de loi a priori à partir du premier type. Le dernier type de loi a priori, tiré de Clogg et coll. (1991) est défini comme il suit

$$\begin{aligned}
&\delta_{ij11} = 0, \delta_{ij12} = \frac{p}{3} \cdot \left(\frac{1}{I \cdot J} \right), \delta_{ij21} = \frac{p}{3} \cdot \left(\frac{1}{I \cdot J} \right), \\
\text{et} \qquad & \qquad (12) \\
&\delta_{ij22} = \frac{p}{3} \cdot \left(\frac{1}{I \cdot J} \right).
\end{aligned}$$

Ces cinq types de lois a priori sont résumés au tableau 1 et sont comparés à la section suivante en utilisant des données empiriques et des études par simulation.

Tableau 1
Cinq types de lois a priori δ_{ijkl} ($\hat{m}_{ijkl}$ est l'EMV, I et J sont les nombres de lignes et de colonnes dans un tableau à double entrée, et p est le nombre de paramètres)

	δ_{ij11}	δ_{ij12}	δ_{ij21}	δ_{ij22}	
Type I	$\nabla_{11} \frac{y_{ij11}}{y_{++11}}$	$\nabla_{12} \frac{y_{ij11}}{y_{++11}}$	$\nabla_{21} \frac{y_{ij11}}{y_{++11}}$	$\nabla_{22} \frac{y_{ij11}}{y_{++11}}$,	$\nabla_{kl} = p \cdot \frac{y_{++kl}}{y_{++++}}$
Type II	0	$\nabla_{12} \frac{y_{ij11}}{y_{++11}}$	$\nabla_{21} \frac{y_{ij11}}{y_{++11}}$	$\nabla_{22} \frac{y_{ij11}}{y_{++11}}$,	$\nabla_{kl} = p \cdot \frac{y_{++kl}}{y^*_{++++}}$
Type III	$\nabla_{11} \cdot \left(\frac{\hat{m}_{ij11}}{\hat{m}_{++11}} \right)$	$\frac{\nabla_{12}}{2} \left(\frac{\hat{m}_{ij12}}{\hat{m}_{++12}} + \frac{1}{IJ} \right)$	$\frac{\nabla_{21}}{2} \left(\frac{\hat{m}_{ij21}}{\hat{m}_{++21}} + \frac{1}{IJ} \right)$	$\frac{\nabla_{22}}{2} \left(\frac{\hat{m}_{ij22}}{\hat{m}_{++22}} + \frac{1}{IJ} \right)$,	$\nabla_{kl} = p \cdot \frac{\hat{m}_{++kl}}{\hat{m}_{++++}}$
Type IV	0	$\frac{\nabla_{12}}{2} \left(\frac{\hat{m}_{ij12}}{\hat{m}_{++12}} + \frac{1}{IJ} \right)$	$\frac{\nabla_{21}}{2} \left(\frac{\hat{m}_{ij21}}{\hat{m}_{++21}} + \frac{1}{IJ} \right)$	$\frac{\nabla_{22}}{2} \left(\frac{\hat{m}_{ij22}}{\hat{m}_{++22}} + \frac{1}{IJ} \right)$,	$\nabla_{kl} = p \cdot \frac{\hat{m}_{++kl}}{\hat{m}^*_{++++}}$
Type V	0	$\nabla_{12} \left(\frac{1}{I \cdot J} \right)$	$\nabla_{21} \left(\frac{1}{I \cdot J} \right)$	$\nabla_{22} \left(\frac{1}{I \cdot J} \right)$,	$\nabla_{kl} = \frac{p}{3}$

$y^*_{++++} = y_{++++} - y_{++11}$ et $\hat{m}^*_{++++} = \hat{m}_{++++} - \hat{m}_{++11}$

Jusqu'ici, nous avons présenté des méthodes pour un tableau à double entrée, et y_{ijkl} est défini pour l'effectif de la case (i, j) de la i^{e} ligne et de la j^{e} colonne (c'est-à-dire $X_1 = i, X_2 = j$), et l'indicateur R_1 pour une ligne manquante et R_2 pour une colonne manquante (c'est-à-dire $R_1 = k, R_2 = l$). L'extension à un tableau à triple entrée est facile. Soit y_{ijklmn} l'effectif de la $(i, j, k)^{\text{e}}$ case pour les trois variables de réponse (c'est-à-dire $X_1 = i$, $X_2 = j$ et $X_3 = k$) et les lignes et colonnes manquantes respectives (c'est-à-dire $R_1 = l$, $R_2 = m$ et $R_3 = n$ pour $l, m, n = 1, 2$). Donc, $lmn = 111$ implique que chacune des trois variables est observée, $lmn = 112$ implique que X_1 et X_2 sont observées, mais que X_3 manque ; de même pour $lmn = 121, 122, 211, 212, 221, 222, 1$ indique que la variable est observée et 2, qu'elle manque. Par conséquent, nous pouvons définir l'algorithme EM et les lois a priori pour un tableau de contingence à triple entrée incomplet. À l'étape E, l'espérance conditionnelle pour la $(i, j, k)^{\text{e}}$ case avec l'information inconnue de la marge k est donnée par

$$E_{\text{old}}(y_{ijk112} \mid \pi^{\text{old}}_{ijklmn}, y_{ij+112}) = y_{ij+112} \frac{m^{\text{old}}_{ijk112}}{m^{\text{old}}_{ij+112}}.$$

De même,

$$E_{\text{old}}(y_{ijk122} \mid \pi^{\text{old}}_{ijklmn}, y_{i++122}) = y_{i++122} \frac{m^{\text{old}}_{ijk122}}{m^{\text{old}}_{i++122}}$$

et

$$E_{\text{old}}(y_{ijk222} \mid \pi^{\text{old}}_{ijklmn}, y_{+++222}) = y_{+++222} \frac{m^{\text{old}}_{ijk222}}{m^{\text{old}}_{+++222}}.$$

Nous pouvons définir de la même manière d'autres espérances et cinq types de lois a priori.

Le sondage électoral de l'État de l'Ohio est réalisé par la méthode de composition aléatoire (CA). Si le sondage par CA est strictement autopondéré, aucune modification ne doit être apportée aux méthodes bayésiennes (Lavrakas 1993 ; Potthoff 1994). Cependant, la CA n'est pas toujours exécutée selon un plan autopondéré. Par exemple, un échantillon téléphonique est constitué de ménages et non de personnes. Si une personne est interviewée dans un ménage, la réponse doit être pondérée par le nombre de personnes dans le ménage. Un poids est également nécessaire pour les ménages possédant plus d'un numéro de téléphone. Si l'on dispose d'une estimation exacte du nombre total de ménages, on peut procéder à une stratification par région ou par État et il convient d'envisager une pondération dans une analyse complète. Les sondages électoraux réalisés en Ohio en 1998 ont été effectués par CA. Dans la présente étude, nos méthode et modèles ne comprennent pas de pondération pour tenir compte de la stratification, de la mise en grappes et d'autres facteurs donnant lieu à des probabilités de sélection différentes dans un sondage téléphonique.

Toutefois, il est possible de les étendre afin d'introduire ce genre de pondération. La simple extension qui suit montre comment tenir compte d'une stratification type. Dans un tableau à triple entrée, soit X_3 la troisième variable de réponse marquée de l'indice h ($h = 1, \ldots, H$) que nous supposons être toujours observée. Les H catégories peuvent être des strates dans un échantillonnage stratifié. Puisque X_3 est toujours observée, la variable indicatrice de données manquantes correspondante R_3 est égale à 1 et son observation peut être désignée par y_{ijhlm1}. Alors, pour chaque strate h, nous pouvons écrire la log-vraisemblance suivante :

$$l_h = \sum_{i=1}^{I}\sum_{j=1}^{J} y_{ijh111} \log(\pi_{ijh11}) + \sum_{i=1}^{I} y_{i+h121} \log(\pi_{i+h12}) + \sum_{j=1}^{J} y_{+jh211} \log(\pi_{+jhk21}) + y_{++h221} \log(\pi_{++h22})$$

où $\pi_{ijhlm} = P[X_1 = i, X_2 = j, R_1 = l, R_2 = m \mid X_3 = h]$. Donc, la terminologie X_3 utilisée pour un tableau à triple entrée joue le rôle d'un indicateur pour les strates. Pour chaque strate h, la vraisemblance de (13) est exactement la même que celle d'un tableau à double entrée.

Alors, nous pouvons définir un modèle log-linéaire pour l'espérance par case $m_{ijhlm} = N_h \cdot \pi_{ijhlm}$ de la même façon qu'en (2), où $N_h = \sum_{i, j, l, m} y_{ijhlm}$ pour chaque $h = 1, 2, ..., H$. Un modèle de non-réponse non ignorable est donné par

$$\log(m_{ijhlm}) = \beta_{0h} + \beta^{i}_{X_1 h} + \beta^{j}_{X_2 h} + \beta^{l}_{R_1 h} + \beta^{m}_{R_2 h} + \beta^{ij}_{X_1 X_2 h} + \beta^{il}_{X_1 R_1 h} + \beta^{jm}_{X_2 R_2 h}. \quad (13)$$

Afin d'éviter le problème des solutions limites décrit à la section 2, nous utilisons les lois a priori de Dirichlet pour π_{ijhlm}

$$\prod_i \prod_j \pi^{\delta_{ijh11}}_{ijh11} \cdot \pi^{\delta_{ijh12}}_{ijh12} \cdot \pi^{\delta_{ijh21}}_{ijh21} \cdot \pi^{\delta_{ijh22}}_{ijh22}.$$

Ensuite, nous suivons exactement les mêmes procédures que celles illustrées à la section 2 pour estimer l'espérance par case m_{ijhlm} pour chaque $h = 1, 2\ldots, H$. L'estimation de l'espérance dans la $(i, j)^{\text{e}}$ case est

$$\hat{E}(y_{ij}) = \sum_{h=1}^{H} w_h \sum_{l, m} \hat{m}_{ijhlm}$$

où w_h est le poids connu pour la h^{e} strate et $\hat{m}_{ijhlm}$ est m_{ijhlm} évalué à l'EMP de $\boldsymbol{\beta}$. Par exemple, $w_h = N_h / \sum_h N_h$ est le poids pour un échantillon stratifié où N_h est la taille de la population de la h^{e} strate.

La matrice de variance-covariance d'une approximation de la distribution de $\hat{\mathbf{m}}$ est

$$\frac{\partial \hat{\mathbf{m}}^T}{\partial \boldsymbol{\beta}} \mathrm{Var}(\hat{\boldsymbol{\beta}}_{\mathbf{EMP}}) \frac{\partial \hat{\mathbf{m}}}{\partial \boldsymbol{\beta}} \qquad (14)$$

où $\hat{\mathbf{m}}$ est une expression vectorielle des estimations par case $\hat{m}_{ijhlm}$, $\hat{\boldsymbol{\beta}}_{\mathbf{EMP}}$ est l'EMP de $\boldsymbol{\beta}$ dont la variance $\mathrm{Var}(\hat{\boldsymbol{\beta}}_{\mathbf{EMP}})$ est donnée par l'inverse de (9), et $\partial \mathbf{m} / \partial \boldsymbol{\beta} = N_h \times [\mathrm{diag}(\hat{\boldsymbol{\pi}}) - \hat{\boldsymbol{\pi}}\hat{\boldsymbol{\pi}}^T]\mathbf{Z}$ où $\hat{\boldsymbol{\pi}}$ possède

$$\hat{\pi}_{ijhlm} = \pi_{ijhlm}(\hat{\boldsymbol{\beta}}_{\mathbf{EMP}}) = \frac{\exp(\mathbf{z}_{ijhlm}\hat{\boldsymbol{\beta}}_{\mathbf{EMP}})}{\sum_{k\in(i,j,h,l,m)} \exp(\mathbf{z}_k \hat{\boldsymbol{\beta}}_{\mathbf{EMP}})}$$

comme élément type.

3. Une application à un sondage électoral dans l'État de l'Ohio

Lorsque l'on veut prédire par sondage le gagnant d'une élection, l'exactitude du résultat dépend souvent de la façon dont sont traités les électeurs indécis qui voteront vraisemblablement, mais qui n'ont pas encore décidé quel est leur candidat préféré. Nous comparons les estimations bayésiennes basées sur les cinq types de lois a priori à l'estimation du MV en nous servant des données du Buckeye State Poll (BSP), ou sondage électoral dans l'État de l'Ohio, réalisé en 1998 par le Center for Survey Research de la Ohio State University. Les sondages préélectoraux du BSP ont produit des tableaux de contingence à double entrée incomplets dont une catégorie était le candidat préféré et l'autre, la probabilité de voter aux élections de novembre 1998 en vue d'élire le gouverneur, le procureur général, le maire de Columbus et le trésorier de l'Ohio. Le tableau 2, qui résume ces quatre sondages, révèle un nombre important d'électeurs indécis.

Pour la comparaison, nous considérons le modèle 1 de non-réponse ignorable et les modèles 2 et 3 de non-réponse non ignorable qui suivent.

$$\text{Modèle 1}: \quad \log(m_{ijkl}) = \beta_0 + \beta^i_{X_1} + \beta^j_{X_2} + \beta^k_{R_1} + \beta^l_{R_2} + \beta^{ij}_{X_1X_2} + \beta^{kl}_{R_1R_2},$$

$$\text{Modèle 2}: \quad \log(m_{ijkl}) = \beta_0 + \beta^i_{X_1} + \beta^j_{X_2} + \beta^k_{R_1} + \beta^l_{R_2} + \beta^{ik}_{X_1R_1} + \beta^{jl}_{X_2R_2} + \beta^{ij}_{X_1X_2} + \beta^{kl}_{R_1R_2},$$

$$\text{Modèle 3}: \quad \log(m_{ijkl}) = \beta_0 + \beta^i_{X_1} + \beta^j_{X_2} + \beta^k_{R_1} + \beta^l_{R_2} + \beta^{il}_{X_1R_2} + \beta^{jk}_{X_2R_1} + \beta^{ij}_{X_1X_2} + \beta^{kl}_{R_1R_2}.$$

Dans le modèle 1, les données manquent complètement au hasard et les cas pour lesquels des données manquent peuvent être ignorés dans les inférences de la vraisemblance. Les modèles 2 et 3 sont des modèles de non-réponse non ignorable où la probabilité qu'une variable manque dépend de la variable elle-même dans le modèle 2, tandis qu'elle dépend de l'autre variable dans le modèle 3. Notons que, sous les modèles 1 et 3, les estimations du MV ne se situent pas sur la limite de l'espace des paramètres comme l'ont montré Baker et coll. (1992). En outre, ayant constaté que, sous les modèles 1 et 3, les cinq estimations bayésiennes des effectifs attendus par case sont non seulement assez proches de l'estimation du MV, mais ont aussi presque le même écart-type, nous ne présentons les estimations du MV que pour les modèles 1 et 3.

Nous désignons les estimations du MV sous le modèle 1 de non-réponse ignorable, le modèle 2 de non-réponse non ignorable et le modèle 3 de non-réponse non ignorable par $IG1_{ML}$, $NON2_{ML}$ et $NON3_{ML}$, respectivement. *IG* et *NON* signifient ignorable et non ignorable, respectivement. Soit aussi $NON2_i^{BE}$ l'estimateur bayésien utilisant le i^e type de lois a priori sous le modèle 2. Autrement dit, $NON2_1^{BE}$ utilise les lois a priori axées sur les répondants de (10) et $NON2_2^{BE}$ utilise les mêmes lois a priori que $NON2_1^{BE}$, excepté que $\delta_{ij11} = 0$. De même, $NON2_3^{BE}$ est donné par (11) et $NON2_4^{BE}$ utilise les mêmes lois a priori excepté que $\delta_{ij11} = 0$. $NON2_5^{BE}$ est l'estimation bayésienne obtenue en utilisant les lois a priori constantes de (12). De plus, nous pouvons utiliser la méthode de Stasny (1986, 1988) pour estimer les effectifs prévus par case sous les modèles 1 et 3 qu'elle a supposés implicitement. Cependant, ses estimations semblent être exactement les mêmes que $IG1_{ML}$.

Tableau 2
Données observées pour les sondages préélectoraux du BSP

	Course à l'élection du gouverneur				Course à l'élection du procureur général		
	Fisher	Taft	Autre	Indécis	Montgomery	Cordray	Indécis
Votera probablement	112	140	23	61	197	82	57
Ne votera probablement pas	96	108	21	73	161	65	75
Indécis	7	11	1	4	15	4	0
	Course à l'élection du maire				Course à l'élection du trésorier		
	Coleman	Teater	Espy	Indécis	Deters	Donofrio	Indécis
Votera probablement	40	32	25	30	127	119	90
Ne votera probablement pas	37	47	41	56	127	90	84
Indécis	0	2	1	0	10	7	0

La partie supérieure du tableau 3 donne les résultats prédits des élections obtenus en utilisant uniquement les réponses « votera vraisemblablement » pour les quatre courses, ainsi que les écarts-types entre parenthèses. Les écarts-types sont proches les uns des autres et révèlent des écarts importants entre les premier et deuxième candidats en tête, excepté dans le cas de la course à l'élection du maire. Ce tableau contient aussi les résultats réels des élections et indique si les estimations du MV se situent ou non dans les solutions limites.

La partie inférieure du tableau donne les prédictions des résultats des élections en utilisant les réponses « votera vraisemblablement » et « ne votera vraisemblablement pas » pour voir ce qui se passerait si les personnes ayant déclaré qu'elles ne « voteraient vraisemblablement pas » se rendaient effectivement aux urnes. La comparaison des deux tableaux nous permet de conclure que les gagnants des élections du gouverneur, du procureur général et du trésorier restent les mêmes quelle que soit la probabilité déclarée de voter, tandis que le gagnant de l'élection du maire aurait pu changer si la plupart des personnes ayant déclaré qu'elles « ne voteraient vraisemblablement pas » avaient effectivement voté.

En nous basant sur le tableau 3, nous pouvons classer les sept estimations, sauf $NON2_{ML}$, en deux groupes : $NON2_3^{BE}$, $NON2_4^{BE}$ et $NON2_5^{BE}$ dans le premier groupe et les quatre autres estimations, $NON2_1^{BE}$, $NON2_2^{BE}$, $IG1_{ML}$ et $NON3_{ML}$, dans le deuxième groupe. Comme il fallait s'y attendre, puisque les lois a priori δ_{ijkl} pour $NON2_1^{BE}$ et $NON2_2^{BE}$ sont définis de façon telle que l'estimation de m_{ijkl} diminue et se rapproche du MV sous un modèle de non-réponse ignorable, ces deux estimations bayésienne sont très proches de $IG1_{ML}$ et offrent par conséquent peu d'avantage par rapport à cette dernière estimation. Il est également intéressant de souligner que $NON3_{ML}$ est presque identique à $IG1_{ML}$, bien que leurs modèles log-linéaires soient spécifiés différemment.

Il n'existe aucun critère général pour déterminer s'il convient d'utiliser un modèle de non-réponse ignorable ou un modèle de non-réponse non ignorable. Cependant, comme l'indique Chen et Stasny (2003), l'hypothèse de non-ignorabilité d'une non-réponse peut être raisonnable dans l'étude fondée sur le sondage électoral de l'État de l'Ohio parce que les gens hésitent à indiquer leur appui pour un candidat impopulaire ou que leurs préférences au moment du sondage ne sont pas fermes ou exactes. À cet égard, les estimations $NON2_1^{BE}$, $NON2_2^{BE}$ et $NON3_{ML}$ ne sont peut-être pas appropriées dans ces études de cas particulières, parce qu'elles sont presque les mêmes que les estimations $IG1_{ML}$ du modèle 1.

Tableau 3
Prédiction des résultats des élections basée sur les sondages électoraux de l'État de l'Ohio d'octobre 98 et d'avril 98 (les résultats sont exprimés en pourcentage et les chiffres entre parenthèses sont les écarts-types)

	Gouverneur			Maire			Procureur général		Trésorier	
	Fisher	Taft	Autre	Coleman	Teater	Espy	Mongomery	Cordray	Deters	Donofrio
Utilisation des réponses « votera vraisemblablement » seulement										
$NON2_{ML}$	33,2(2,75)	42,1(3,00)	24,8	31,5(4,65)	25,3(4,23)	43,2	75,6(3,71)	24,4	57,0(3,48)	43,0
$NON2_1^{BE}$	40,6(3,04)	48,5(3,27)	10,9	38,1(5,14)	34,2(4,78)	27,7	72,1(3,61)	27,9	52,7(3,36)	47,3
$NON2_2^{BE}$	40,9(3,01)	50,7(3,20)	8,40	39,9(5,04)	33,6(4,83)	26,5	71,0(3,59)	29,0	52,1(3,34)	47,9
$NON2_3^{BE}$	35,8(2,85)	44,5(3,08)	19,7	35,6(4,87)	29,3(4,51)	35,1	63,0(3,67)	37,0	54,3(3,41)	45,7
$NON2_4^{BE}$	36,3(2,87)	45,2(3,11)	18,6	35,9(4,91)	29,4(4,52)	34,6	63,0(3,64)	37,0	53,9(3,40)	46,1
$NON2_5^{BE}$	38,9(2,99)	47,4(3,20)	13,7	37,7(4,99)	33,6(4,77)	28,7	66,0(3,54)	34,0	51,5(3,32)	48,5
$IG1_{ML}$	40,6(3,03)	51,2(3,28)	8,20	40,8(5,16)	33,4(4,76)	25,8	70,9(3,59)	29,1	51,8(3,32)	48,2
$NON3_{ML}$	40,6(3,03)	51,2(3,28)	8,20	40,9(5,16)	33,3(4,75)	25,8	70,9(3,58)	29,1	51,7(3,32)	48,3
Résultat réel	45	50	5	39	37	24	63	37	57	43
Limite	oui			oui			oui		non	
Utilisation des réponses « votera vraisemblablement » + « ne votera vraisemblablement pas »										
$NON2_{ML}$	32,7(1,83)	39,4(1,91)	27,8	24,8(2,45)	26,2(2,49)	49,0	77,0(1,64)	23,0	60,2(1,93)	39,8
$NON2_1^{BE}$	41,3(1,93)	46,4(1,96)	12,3	30,7(2,68)	37,1(2,75)	32,2	72,8(1,74)	27,2	56,0(1,96)	44,0
$NON2_2^{BE}$	41,9(1,93)	49,2(1,95)	8,90	32,7(2,63)	36,5(2,76)	30,8	71,4(1,77)	28,6	55,3(1,96)	44,7
$NON2_3^{BE}$	35,4(1,87)	41,8(1,93)	22,7	27,8(2,55)	30,5(2,62)	41,7	61,0(1,72)	39,0	57,6(1,95)	42,4
$NON2_4^{BE}$	36,0(1,88)	42,6(1,93)	21,4	28,7(2,57)	30,6(2,62)	40,7	60,9(1,75)	39,1	57,2(1,95)	42,8
$NON2_5^{BE}$	39,1(1,91)	45,1(1,95)	15,8	30,7(2,63)	35,8(2,74)	33,5	64,8(1,88)	35,2	54,8(1,96)	45,2
$IG1_{ML}$	41,5(1,96)	49,8(1,96)	8,70	33,9(2,70)	36,1(2,74)	29,9	71,2(1,78)	28,8	55,0(1,96)	45,0
$NON3_{ML}$	41,5(1,96)	49,8(1,96)	8,70	34,1(2,71)	36,0(2,74)	29,9	71,1(1,78)	28,9	55,0(1,96)	45,0

Comparativement aux résultats réels des élections, $NON2_{ML}$ donne la pire prédiction pour le gouverneur, le maire et le procureur général, parce que les valeurs se situent sur une solution limite ; par contre, elle donne la meilleure prédiction pour le trésorier, parce que les valeurs ne se situent pas sur une solution limite. Dans le cas de l'élection du procureur général, $NON2_3^{BE}$ et $NON2_4^{BE}$ prédisent non seulement le résultat réel exact, mais diffèrent aussi assez bien des autres estimations. Puisque $NON2_3^{BE}$ et $NON2_4^{BE}$ utilisent les lois a priori destinées à refléter les profils de réponse différents des répondants et des électeurs indécis, nous pouvons inférer que la préférence de ces derniers pour un candidat diffère assez fortement de celle des répondants (autrement dit, $NON2_3^{BE}$ et $NON2_4^{BE}$ affectent 19,4 % des électeurs indécis qui voteront vraisemblablement à Montgomery et 80,6 % à Cordray, tandis que dans le tableau 2, les données indiquent que le pourcentage allant à Montgomery par opposition à Cordray est de 29,4 % contre 70,6 % chez les répondants qui voteront vraisemblablement).

Afin de visualiser cette différence entre les répondants et les électeurs indécis en ce qui a trait aux estimations des paramètres et d'examiner l'effet de l'occurrence de la solution limite sur les estimations sous le modèle 2 de non-réponse non ignorable, nous présentons au tableau 4 les estimations du MV et les estimations $NON2_3^{BE}$, ainsi que les écarts-types correspondants pour l'élection du procureur général. Comme il existe une solution limite, toutes les estimations du MV ont un écart-type trop grand comme il fallait s'y attendre. Par ailleurs, $NON2_3^{BE}$ est très stable. Puisque $\beta_{X_1X_2}^{11} = 0,0472$ est la plus petite et que son écart-type est relativement grand, nous négligeons le terme $\beta_{X_1X_2}^{11}$ pour rendre l'interprétation moins complexe. Sous $\beta_{X_1X_2}^{11} = 0$, il n'est pas difficile de montrer qu'en utilisant les estimations de $NON2_3^{BE}$ du tableau 4,

$$\log\frac{m_{1j1l}}{m_{2j1l}} = 2(\beta_{X_1}^{1} + \beta_{X_1R_1}^{11}) = 0,09$$

et

$$\log\frac{m_{1j2l}}{m_{2j2l}} = 2(\beta_{X_1}^{1} - \beta_{X_1R_1}^{11}) = 1,3916$$

pour chaque valeur fixée de j et l, et

$$\log\frac{m_{i1k1}}{m_{i2k1}} = 2(\beta_{X_2}^{1} + \beta_{X_2R_2}^{11}) = 0,8982$$

et

$$\log\frac{m_{i1k2}}{m_{i2k2}} = 2(\beta_{X_2}^{1} - \beta_{X_2R_2}^{11}) = -1,4942$$

pour chaque valeur fixée de i et k. Donc, en vertu de

$$\log\frac{m_{1j1l}}{m_{2j1l}} = 2(\beta_{X_1}^{1} + \beta_{X_1R_1}^{11}) = 0,09,$$

les personnes qui voteront vraisemblablement (c'est-à-dire $i = 1$) sont 1,09 fois (c'est-à-dire $e^{0,09}$) plus nombreuses que celles qui ne voteront vraisemblablement pas (c'est-à-dire $i = 2$) parmi les répondants ($k = 1$), tandis qu'en vertu de

$$\log\frac{m_{1j2l}}{m_{2j2l}} = 2(\beta_{X_1}^{1} - \beta_{X_1R_1}^{11}) = 1,3916,$$

les personnes qui voteront vraisemblablement ($i = 1$) sont 4,02 fois (c'est-à-dire $e^{1,3916}$) plus nombreuses que les personnes qui ne voteront vraisemblablement pas $i = 2$ parmi les électeurs indécis ($k = 2$) ; en vertu de

$$\log\frac{m_{i1k1}}{m_{i2k1}} = 2(\beta_{X_2}^{1} + \beta_{X_2R_2}^{11}) = 0,8982,$$

les personnes qui votent pour Montgomery sont 2,46 fois plus nombreuses que celles qui votent pour Cordray parmi les répondants, tandis qu'en vertu de

$$\log\frac{m_{i1k2}}{m_{i2k2}} = 2(\beta_{X_2}^{1} - \beta_{X_2R_2}^{11}) = -1,4942,$$

les personnes qui ne voteront vraisemblablement pas sont 4,46 fois plus nombreuses que celles qui voteront vraisemblablement parmi les électeurs indécis. Cela implique que le profil de réponse des répondants et des électeurs indécis est fort différent.

Tableau 4
MV et le troisième type d'estimation bayésienne sous le modèle 2 de non-réponse non ignorable pour l'élection du procureur général (les écarts-types sont indiqués entre parenthèses)

	β_0	$\beta_{X_1}^1$	$\beta_{X_2}^1$	$\beta_{R_1}^1$	$\beta_{R_2}^1$	$\beta_{X_1R_1}^{11}$	$\beta_{X_2R_2}^{11}$	$\beta_{X_1X_2}^{11}$	$\beta_{R_1R_2}^{11}$
$NON2_{ML}$	-3,3735	-1,9487	3,2134	4,8496	4,8186	2,0283	-2,7594	-0,0452	-1,5588
		(3,120)	(8,515)	(3,996)	(8,871)	(3,120)	(8,512)	(0,045)	(2,501)
$NON2_3^{BE}$	0,6860	0,3704	-0,1490	3,3024	2,2942	-0,3254	0,5981	0,0472	-1,5450
		(0,118)	(0,052)	(2,501)	(2,501)	(0,117)	(0,052)	(0,041)	(2,501)

La grandeur de cette différence peut être mesurée à l'aide des termes les plus importants, $\beta^{11}_{X_1R_1}$ et $\beta^{11}_{X_2R_2}$, dans le modèle 2 de non-réponse non ignorable. Puisque

$$\beta^{11}_{X_1R_1} = \frac{1}{4}\log\frac{m_{1111}/m_{2111}}{m_{1121}/m_{2121}} = -0,3254$$

et

$$\beta^{11}_{X_2R_2} = \frac{1}{4}\log\frac{m_{1111}/m_{1211}}{m_{1112}/m_{1212}} = 0,5981,\ \beta^{11}_{X_1R_1}$$

est le logarithme du rapport des cotes qui montre la différence logarithmique entre le ratio du nombre de ceux « qui voteront vraisemblablement » à ceux qui « ne voteront vraisemblablement pas » parmi les électeurs décidés pour Montgomery et le même ratio parmi les électeurs indécis qui préfèrent Montgomery, mais qui n'expriment pas leur probabilité de se rendre aux urnes. Par contre, $\beta^{11}_{X_2R_2}$ est le logarithme du rapport des cotes qui montre la log-différence entre le ratio du nombre d'électeurs en faveur de Montgomery au nombre d'électeurs en faveur de Cordray parmi les électeurs décidés qui voteront vraisemblablement et le même ratio parmi les électeurs indécis qui voteront vraisemblablement, mais qui n'indiquent pas quel est leur candidat préféré. Donc, parmi les électeurs en faveur de Montgomery, la possibilité que les électeurs indécis votent par rapport à celle qu'ils ne votent pas est d'environ 3,67 fois

$$\left(\text{c'est-à-dire } \frac{m_{1111}/m_{2111}}{m_{1121}/m_{2121}} = e^{4\times -0,3254} = 3,67^{-1}\right)$$

plus grande que la possibilité pour les électeurs décidés, ce qui veut dire que Montgomery doit mettre en œuvre une stratégie en vue d'accroître la participation des électeurs au scrutin. Par ailleurs, parmi les personnes qui voteront vraisemblablement, le taux de soutien pour Montgomery chez les électeurs décidés est environ 10,94 fois

$$\left(\text{c'est-à-dire } \frac{m_{1111}/m_{1211}}{m_{1112}/m_{1212}} = e^{4\times 0,5981} = 10,94\right)$$

plus grand que celui des électeurs indécis à l'égard de Montgomery, ce qui implique que la plupart des électeurs indécis n'indiquant pas qui est leur candidat préféré voteront vraisemblablement pour Cordray comme procureur général. Cela confirme également la croyance populaire selon laquelle les électeurs ont tendance à demeurer « indécis » dans un sondage s'ils appuient un candidat qui est considéré comme étant inférieur dans une course électorale et ils sont enclins à s'abstenir de voter s'ils appuient le candidat qui domine la course avec certitude.

4. Étude par simulation

Nous considérons un tableau de contingence 2×2 avec marges supplémentaires pour comparer la performance des cinq estimations bayésiennes décrites à la section 2 pour divers pourcentages de données manquantes et divers profils de réponse sous le modèle de non-réponse non ignorable suivant (c'est-à-dire le modèle 2) :

$$\log(m_{ijkl}) = \beta_0 + \beta^i_{X_1} + \beta^j_{X_2} + \beta^k_{R_1} + \beta^l_{R_2} + \beta^{ik}_{X_1R_1} + \beta^{jl}_{X_2R_2} + \beta^{ij}_{X_1X_2} + \beta^{kl}_{R_1R_2}.$$

Donc, nous comparons uniquement $NON2_{ML}$ et $NON2^{BE}_i$ pour $i = 1, ..., 5$ dans cette étude par simulation.

Puisque chacune des variables X_1, X_2, R_1 et R_2 comprend deux niveaux, huit paramètres doivent être déterminés pour l'étude par simulation. D'après les équations de

$$4\beta^{11}_{X_1R_1} = \log\frac{m_{1111}/m_{2111}}{m_{1121}/m_{2121}} \quad \text{et} \quad 4\beta^{11}_{X_2R_2} = \log\frac{m_{1111}/m_{1211}}{m_{1112}/m_{1212}},$$

$$\beta^{11}_{X_1R_1} = \beta^{11}_{X_2R_2} = 0$$

signifie qu'il n'y a aucune différence entre les profils de réponse des répondants et des électeurs indécis. Ces profils de réponse diffèrent d'autant plus que les paramètres $\beta^{11}_{X_1R_1}$ et $\beta^{11}_{X_2R_2}$ sont grands. Nous faisons varier ces deux paramètres de 0,2 à 0,8 par incrément de 0,2. Nous fixons le pourcentage de valeurs manquantes à 20 % et à 30 % en ajustant $\beta^1_{X_1}$ et $\beta^1_{R_1}$ et en posant que

$$\frac{m_{1111}/m_{1211}}{m_{2111}/m_{2211}} = 5,\ \frac{m_{1111}/m_{1112}}{m_{1112}/m_{1122}} = 2,$$

et

$$N = \sum_{ijkl} m_{ijkl} = 1\,000.$$

Cela signifie que la taille et le pourcentage de valeurs manquantes pour la case de $X_1 = 1$ et $X_2 = 1$ sont approximativement égaux à cinq fois et deux fois la taille des trois autres cases, respectivement.

Nous produisons un grand nombre d'échantillons $\{y_{ijkl}, i, j, k, l = 1, 2\}$ dans les conditions susmentionnées jusqu'à ce que nous obtenions 1 000 échantillons aléatoires avec des solutions limites et 1 000 autres sans solution limite. L'occurrence d'une solution limite est déterminée par le critère donné dans Michiels et Molenberghs (1997) (voir aussi Clarke 2002, ainsi que Smith et coll. 1999 pour plus de précisions). En utilisant $\{y_{ij11}, y_{i+12}, y_{+j21}, y_{++22}, i, j, = 1, 2\}$ obtenu d'après les données générées, nous estimons les effectifs attendus par case m_{ijkl} au moyen des

cinq estimations bayésiennes et de l'estimation du MV décrites à la section 2.

Nous calculons les erreurs quadratiques moyennes (EQM) et les biais absolus de $NON2_{ML}$, $NON2_1^{BE}$, ..., $NON2_5^{BE}$ pour $\{\sum_{kl} m_{ijkl}, i, j = 1, 2\}$. Puis, nous prenons la moyenne sur les quatre EQM et sur les quatre biais absolus que nous obtenons à partir de chaque estimation pour voir la performance globale de l'estimation. De même, nous calculons les EQM moyennes et les biais absolus moyens pour $\{m_{ij12} + m_{ij21} + m_{ij22}, i, j = 1, 2\}$ pour voir la performance de chaque estimation en présence d'imputation des non-réponses.

Le tableau 5 donne les ratios des EQM moyennes et des biais absolus moyens des cinq estimations bayésiennes (c'est-à-dire $NON2_1^{BE}$, ..., $NON2_5^{BE}$) par rapport à l'estimation du MV (c'est-à-dire $NON2_{ML}$) quand surviennent des solutions limites, tandis que le tableau 6 donne les mêmes ratios quand aucune solution limite ne se présente. Donc, les valeurs inférieures à 1 impliquent que l'estimation bayésienne correspondante possède une EQM moyenne ou un biais absolu moyen plus faible que l'estimation du MV. Les deux tableaux ne présentent que les cas pour $\beta_{X_1R_1}^{11} < \beta_{X_2R_2}^{11}$ et pour un pourcentage de valeurs manquantes de 20 %, parce que les EQM et les biais sont presque symétriques autour de la coordonnée de $(\beta_{X_1R_1}^{11}, \beta_{X_2R_2}^{11})$. Ils augmentent si nous faisons passer le pourcentage de valeurs manquantes à 30 % en maintenant les mêmes profils d'EQM et de biais que pour le cas où 20 % de données manquent.

Le tableau 5, dans lequel survient une solution limite, montre que $NON2_1^{BE}$, $NON2_3^{BE}$, $NON2_4^{BE}$ ont une EQM plus faible que l'estimation du MV (c'est-à-dire $NON2_{ML}$) pour toutes les valeurs de $\beta_{X_1R_1}^{11}$ et $\beta_{X_2R_2}^{11}$, sauf $(\beta_{X_1R_1}^{11}, \beta_{X_2R_2}^{11}) = (0{,}8, 0{,}8)$. Ici, $NON2_3^{BE}$ a une plus petite EQM que l'estimation du MV. Cela est vrai pour les biais absolus. Par ailleurs, le tableau 6, où ne survient aucune solution limite, montre que seule $NON2_3^{BE}$ est comparable à l'estimation du MV en ce qui concerne l'EQM, mais qu'elle est légèrement biaisée. En particulier, $NON2_3^{BE}$ a une EQM plus faible que l'estimation du MV à condition que $\beta_{X_1R_1}^{11} \neq 0{,}8$ ou $\beta_{X_2R_2}^{11} \neq 0{,}8$ (c'est-à-dire que les profils de réponse des répondants et des non-répondants diffèrent peu).

Tableau 5
Ratios des EQM moyennes et des biais absolus moyens des estimations bayésiennes relativement à l'estimation du MV quand surviennent des solutions limites sous un pourcentage de valeurs manquantes de 20 % (les ratios des biais absolus figurent entre parenthèses)

	$(\beta_{X_1R_1}^{11}, \beta_{X_2R_2}^{11})$	$NON2_1^{BE}$	$NON2_2^{BE}$	$NON2_3^{BE}$	$NON2_4^{BE}$	$NON2_5^{BE}$
Pour $\{m_{ij11} + m_{ij12} + m_{ij21} + m_{ij22}, i, j = 1, 2\}$	(0,2, 0,2)	0,68(0,66)	0,47(0,22)	0,76(0,76)	0,65(0,48)	0,42(0,05)
	(0,2, 0,4)	0,68(0,48)	0,57(0,20)	0,77(0,68)	0,60(0,29)	0,56(0,30)
	(0,2, 0,6)	0,67(0,23)	0,73(0,66)	0,77(0,57)	0,64(0,10)	0,69(0,64)
	(0,2, 0,8)	0,77(0,26)	1,08(1,55)	0,83(0,43)	0,76(0,28)	0,95(1,34)
	(0,4, 0,4)	0,65(0,32)	0,69(0,57)	0,76(0,63)	0,61(0,17)	0,65(0,52)
	(0,4, 0,6)	0,58(0,14)	0,83(0,90)	0,71(0,56)	0,56(0,06)	0,69(0,71)
	(0,4, 0,8)	0,75(0,36)	1,46(2,07)	0,78(0,36)	0,74(0,42)	1,12(1,61)
	(0,6, 0,6)	0,66(0,22)	1,35(1,73)	0,73(0,43)	0,66(0,16)	1,01(1,29)
	(0,6, 0,8)	0,85(0,87)	2,27(3,19)	0,76(0,17)	0,83(0,81)	1,52(2,35)
	(0,8, 0,8)	1,12(1,93)	3,58(5,49)	0,83(0,24)	1,04(1,67)	2,18(3,95)
Pour $\{m_{ij12} + m_{ij21} + m_{ij22}, i, j = 1, 2\}$	(0,2, 0,2)	0,57(0,63)	0,27(0,13)	0,69(0,74)	0,41(0,40)	0,28(0,31)
	(0,2, 0,4)	0,54(0,46)	0,37(0,34)	0,68(0,68)	0,42(0,24)	0,44(0,57)
	(0,2, 0,6)	0,51(0,19)	0,69(0,94)	0,65(0,55)	0,47(0,10)	0,69(0,88)
	(0,2, 0,8)	0,63(0,35)	1,39(2,08)	0,71(0,34)	0,62(0,47)	1,11(1,52)
	(0,4, 0,4)	0,49(0,35)	0,54(0,64)	0,65(0,64)	0,42(0,17)	0,57(0,76)
	(0,4, 0,6)	0,48(0,17)	0,98(1,24)	0,62(0,51)	0,45(0,17)	0,85(1,04)
	(0,4, 0,8)	0,62(0,44)	1,81(2,33)	0,67(0,35)	0,61(0,55)	1,35(1,81)
	(0,6, 0,6)	0,55(0,42)	1,70(1,90)	0,63(0,41)	0,54(0,40)	1,28(1,51)
	(0,6, 0,8)	0,78(0,92)	2,91(3,43)	0,69(0,14)	0,75(0,92)	1,96(2,64)
	(0,8, 0,8)	1,13(1,96)	4,63(5,72)	0,75(0,33)	1,02(1,77)	2,86(4,24)

Tableau 6
Ratios des EQM moyennes et des biais absolus moyens des estimations bayésiennes relativement à l'estimation du MV quand ne survient aucune solution limite sous un pourcentage de valeurs manquantes de 20 % (les ratios des biais absolus figurent entre parenthèses)

	$(\beta^{11}_{X_1R_1}, \beta^{11}_{X_2R_2})$	$NON2_1^{BE}$	$NON2_2^{BE}$	$NON2_3^{BE}$	$NON2_4^{BE}$	$NON2_5^{BE}$
Pour $\{m_{ij11} + m_{ij12} + m_{ij21} + m_{ij22},\ i,\ j = 1,\ 2\}$	(0,2, 0,2)	0,99(3,37)	1,05(7,00)	0,94(2,51)	0,93(4,89)	1,06(8,96)
	(0,2, 0,4)	0,98(2,57)	1,21(5,13)	0,97(1,89)	1,00(3,26)	1,24(5,56)
	(0,2, 0,6)	1,04(2,18)	1,52(3,84)	0,95(1,67)	1,06(2,38)	1,43(3,71)
	(0,2, 0,8)	1,12(2,04)	1,75(3,53)	1,00(1,48)	1,13(2,14)	1,52(3,21)
	(0,4, 0,4)	1,03(2,40)	1,49(4,66)	0,97(1,69)	1,05(2,74)	1,39(4,46)
	(0,4, 0,6)	1,20(2,17)	2,11(3,85)	1,00(1,52)	1,22(2,24)	1,78(3,42)
	(0,4, 0,8)	1,28(2,09)	2,36(3,67)	1,05(1,45)	1,26(2,09)	1,86(3,12)
	(0,6, 0,6)	1,22(2,16)	2,49(3,90)	0,96(1,48)	1,21(2,15)	1,90(3,32)
	(0,6, 0,8)	1,52(1,99)	3,19(3,39)	1,11(1,38)	1,45(1,91)	2,29(2,77)
	(0,8, 0,8)	1,66(1,96)	3,64(3,27)	1,14(1,36)	1,52(1,83)	2,43(2,59)
Pour $\{m_{ij12} + m_{ij21} + m_{ij22},\ i,\ j = 1,\ 2\}$	(0,2, 0,2)	0,88(2,59)	0,89(5,66)	0,87(2,26)	0,89(4,55)	1,21(8,69)
	(0,2, 0,4)	0,93(2,40)	1,27(4,86)	0,93(1,78)	1,00(3,08)	1,50(5,29)
	(0,2, 0,6)	1,09(2,11)	1,93(3,97)	0,98(1,40)	1,15(2,29)	1,85(3,61)
	(0,2, 0,8)	1,24(2,13)	2,36(3,90)	1,02(1,48)	1,27(2,18)	2,06(3,19)
	(0,4, 0,4)	1,03(2,18)	1,81(4,30)	0,96(1,60)	1,12(2,62)	1,85(4,39)
	(0,4, 0,6)	1,23(2,28)	2,62(4,28)	0,99(1,48)	1,29(2,42)	2,28(3,80)
	(0,4, 0,8)	1,42(2,05)	3,26(3,70)	1,07(1,42)	1,44(2,07)	2,53(3,09)
	(0,6, 0,6)	1,33(2,07)	3,22(3,95)	0,99(1,36)	1,36(2,14)	2,54(3,43)
	(0,6, 0,8)	1,65(2,09)	4,14(3,74)	1,13(1,43)	1,61(2,07)	2,98(3,13)
	(0,8, 0,8)	1,91(2,02)	4,48(3,50)	1,16(1,39)	1,66(1,93)	3,03(2,83)

Park et Brown (1994) ont utilisé $NON2_2^{BE}$ pour estimer les effectifs attendus par case dans un tableau à simple entrée incomplet sous un modèle de non-réponse non ignorable. Ils ont montré par des études en simulation que $NON2_2^{BE}$ avait une EQM plus petite que l'estimation du MV bien que son biais soit plus important. Cependant, les valeurs supérieures à 1 pour $NON2_2^{BE}$ dans les tableaux 5 et 6 indiquent qu'il n'en est pas ainsi dans un tableau à double entrée incomplet, indépendamment de la solution limite, et que les méthodes bayésiennes ne donnent pas systématiquement de meilleurs résultats que le MV, même en cas d'une solution limite. L'une des raisons pour lesquelles nos résultats de simulation diffèrent de ceux de Park et Brown (1994) quand a lieu une solution limite tient au choix de $(\beta^{11}_{X_1R_1}, \beta^{11}_{X_2R_2})$, Park et Brown n'ayant exécuté leur simulation que dans les conditions où $\beta^{11}_{X_1R_1} = \beta^{11}_{X_2R_2} = 0,34$. Comme le montre le tableau 5, $NON2_2^{BE}$ est meilleure que le MV quand $\beta^{11}_{X_1R_1} \le 0,4$ et $\beta^{11}_{X_2R_2} \le 0,4$, tandis que $NON2_2^{BE}$ est pire que le MV quand les profils de réponse des répondants et des non-répondants sont fort différents (c'est-à-dire $\beta^{11}_{X_1R_1} \ge 0,6$ ou $\beta^{11}_{X_2R_2} \ge 0,6$).

Le tableau 7 donne la moyenne des écarts-types et des probabilités de couverture à 95 % pour $\beta^{11}_{X_1R_1}$. Ici, nous avons utilisé la formule de variance donnée dans (9) pour calculer les écarts-types, et les probabilités de couverture à 95 % sont les taux de couverture pour les intervalles de confiance à 95 % nominaux. Quand survient une solution limite, même si la probabilité de couverture de l'estimation du MV est celle qui s'approche le plus du niveau de couverture nominal de 95 %, l'estimation du MV possède un écart-type trop grand pour pouvoir être utilisé en pratique. Ces grands écarts-types sont dus au problème des solutions limites de l'estimation du MV. Parmi les estimations bayésiennes, les probabilités de couverture de $NON2_3^{BE}$ sont celles qui s'approchent le plus du niveau de couverture nominal de 95 %, tandis que celles des autres estimations sont généralement inférieures à ce niveau. Cela signifie qu'à part $NON2_3^{BE}$, les estimations bayésiennes sous-estiment les écarts-types.

Quand aucune solution limite ne se présente (deuxième partie du tableau 7), les écarts-types de l'estimation du MV sont nettement plus stables que ceux obtenus dans le cas d'une solution limite. La probabilité de couverture diminue à mesure que $\beta^{11}_{X_1R_1}$ et $\beta^{11}_{X_2R_2}$ augmentent. En particulier, les probabilités de couverture de NON_1^{BE}, NON_2^{BE} et NON_5^{BE} sont considérablement plus faibles que le niveau de couverture nominal de 95 % quand les profils de réponse des répondants et des électeurs indécis sont fort différents (c'est-à-dire $\beta^{11}_{X_1R_1} \ge 0,6$ et $\beta^{11}_{X_2R_2} \ge 0,6$).

Tableau 7
Moyenne des écarts-types et probabilités de couverture à 95 % (entre parenthèses) pour $\beta^{11}_{X_1R_1}$

	$(\beta^{11}_{X_1R_1}, \beta^{11}_{X_2R_2})$	$NON2_{ML}$	$NON2^{BE}_1$	$NON2^{BE}_2$	$NON2^{BE}_3$	$NON2^{BE}_4$	$NON2^{BE}_5$
Solution limite	(0,2, 0,2)	89,5(0,974)	0,082(0,978)	0,064(0,978)	0,093(0,973)	0,071(0,972)	0,060(0,957)
	(0,2, 0,4)	158,3(0,959)	0,072(0,963)	0,096(0,963)	0,079(0,958)	0,066(0,958)	0,058(0,940)
	(0,2, 0,6)	135,3(0,940)	0,065(0,941)	0,057(0,941)	0,071(0,941)	0,062(0,939)	0,056(0,922)
	(0,2, 0,8)	57,4(0,930)	0,070(0,938)	0,061(0,935)	0,076(0,928)	0,066(0,928)	0,060(0,908)
	(0,4, 0,4)	153,4(0,961)	0,079(0,920)	0,061(0,913)	0,096(0,956)	0,072(0,949)	0,060(0,911)
	(0,4, 0,6)	82,2(0,955)	0,072(0,893)	0,059(0,883)	0,086(0,951)	0,069(0,940)	0,058(0,874)
	(0,4, 0,8)	51,2(0,933)	0,071(0,862)	0,059(0,849)	0,084(0,926)	0,068(0,917)	0,059(0,846)
	(0,6, 0,6)	175,5(0,946)	0,077(0,820)	0,060(0,781)	0,101(0,943)	0,074(0,921)	0,061(0,823)
	(0,6, 0,8)	159,6(0,924)	0,071(0,728)	0,057(0,657)	0,089(0,913)	0,069(0,880)	0,058(0,737)
	(0,8, 0,8)	72,8(0,920)	0,070(0,572)	0,056(0,330)	0,093(0,900)	0,070(0,842)	0,058(0,607)
Pas de solution limite	(0,2, 0,2)	0,068(0,949)	0,060(0,959)	0,056(0,959)	0,062(0,937)	0,058(0,935)	0,055(0,922)
	(0,2, 0,4)	0,066(0,960)	0,060(0,970)	0,056(0,970)	0,061(0,935)	0,058(0,931)	0,055(0,951)
	(0,2, 0,6)	0,064(0,940)	0,058(0,945)	0,055(0,945)	0,059(0,959)	0,057(0,919)	0,054(0,909)
	(0,2, 0,8)	0,069(0,933)	0,063(0,944)	0,059(0,941)	0,065(0,926)	0,062(0,925)	0,058(0,920)
	(0,4, 0,4)	0,074(0,910)	0,061(0,836)	0,055(0,828)	0,064(0,899)	0,059(0,884)	0,055(0,824)
	(0,4, 0,6)	0,074(0,915)	0,060(0,815)	0,055(0,806)	0,064(0,922)	0,059(0,879)	0,055(0,792)
	(0,4, 0,8)	0,073(0,891)	0,061(0,786)	0,056(0,771)	0,064(0,873)	0,060(0,852)	0,056(0,763)
	(0,6, 0,6)	0,078(0,859)	0,061(0,567)	0,055(0,470)	0,067(0,853)	0,061(0,795)	0,056(0,572)
	(0,6, 0,8)	0,076(0,843)	0,060(0,515)	0,054(0,402)	0,065(0,817)	0,060(0,767)	0,055(0,556)
	(0,8, 0,8)	0,080(0,755)	0,059(0,110)	0,053(0,017)	0,065(0,728)	0,059(0,607)	0,055(0,158)

5. Conclusion

Nous avons étudié l'application de l'analyse bayésienne à des tableaux de contingence à double entrée incomplets en présence de non-réponse non ignorable. Dans cette situation, les estimations du MV se situent souvent sur la solution limite. Ces solutions limites peuvent donner $G^2 > 0$, même dans le cas d'un modèle saturé (Baker et coll. 1992 ; Park et Brown 1994). Autrement dit, G^2 peut ne pas être appropriée comme statistique pour la spécification des modèles. Pour contourner le problème des solutions limites et obtenir une statistique telle que le facteur de Bayes pour spécifier le modèle indépendamment de l'existence d'une solution limite, nous avons proposé des méthodes d'estimation bayésiennes s'appuyant sur cinq lois a priori différentes. Deux d'entre eux sont nouveaux et les trois autres ont été utilisés antérieurement pour analyser un tableau à simple entrée incomplet. Les deux nouvelles lois a priori tiennent compte des profils de réponse différents entre les répondants et les non-répondants.

L'analyse des données révèle que ces deux nouvelles lois a priori sont plus raisonnables en ce sens qu'ils tiennent mieux compte du mécanisme de non-réponse non ignorable et produisent des estimations proches des résultats réels. En outre, dans le cas des trois lois a priori utilisées antérieurement, notre étude par simulation montre que les estimations bayésiennes peuvent avoir une EQM plus grande que celle des estimations du MV pour un tableau de contingence sans solution limite ainsi qu'avec une solution limite, contrairement aux études antérieures. Cependant, quand a lieu une solution limite, les deux nouvelles lois a priori donnent de meilleures résultats que les trois lois a priori antérieures et que les estimations du MV en ce sens qu'ils sont généralement caractérisés par des EQM plus faibles, des biais plus petits et des probabilités de couverture plus proches du niveau de couverture nominal.

Nous avons discuté brièvement des questions de pondération à la section 2.2. Cependant, elles requièrent une discussion plus rigoureuse que celle présentée dans cette section. Notre discussion pourrait être approfondie de manière à inclure non seulement divers facteurs de pondération, mais aussi des biais de réponse et d'autres sources de biais et de variations. Ces questions pourront être développées avec soin ultérieurement dans un article élargi.

Remerciements

La présente étude a été financée par une subvention de l'Université de la Corée (K0822301).

Bibliographie

Agresti, A. (2002). *Categorical Data Analysis.* 2ième Édition. New York : John Wiley & Sons, Inc.

Baker, S.G., et Laird, N.M. (1988). Regression analysis for categorical variables with outcome subject to nonignorable nonresponse. *Journal of the American Statistical Association*, 83, 62-69.

Baker, S.G., Rosenberger, W.F. et Dersimonian, R. (1992). Closed-form estimates for missing counts in two-way contingency tables. *Statistics in Medicine*, 11, 643-657.

Chen, T. (1972). Mixed-up frequencies and missing data in contingency tables. Mémoire de doctorat non publié, University of Chicago, Dept. of Statistics.

Chen, Q.L., et Stasny, E.A. (2003). Handling undecided voters: Using missing data methods in election forecasting. *Rapport technique*, Department of Statistics, The Ohio State University.

Clarke, P.S. (2002). On boundary solutions and identifiability in categorical regression with non-ignorable non-response. *Biometrical Journal*, 44, 701-717.

Clogg, C.C., Rubin, D.B., Schenker, N. et Schultz, B. (1991). Multiple imputation of industry and occupation codes in Census Public use-samples using Bayesian logistic regression. *Journal of the American Statistical Association*, 86, 68-78.

De Heer W. (1999). International response trends of an international survey. *Journal of Official Statistics*, 15, 129-142.

Dempster, A.P., Laird, N.M. et Rubin, D.M. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, Séries B, 39, 1-38.

Flannelly, K.J., Flannelly, L.T. et McLeod, M.S. Jr. (2000). Reducing undecided voters and other sources of error in election surveys. *International Journal of Market Research*, 42, 231-237.

Fenwick, I, Wiseman, F, Becker, J.F. et Heiman, J.R. (1982). Classifying undecided voters in pre-election polls. *Public Opinion Quarterly*, 46, 383-391.

Forster, J.J., et Smith, R.W.F. (1998). Model-based inference for categorical survey data subject to non-ignorable non-response. *Journal of the Royal Statistical Society B*, 60, 57-70.

Gelman, A., Carlin, J.P., Stern, H.S. et Rubin, D.B. (2004). *Bayesian Data Analysis*. 2ième Édition. New York : Chapman and Hall/CRC.

Groves, R.M., et Couper, M.P. (1998). *Nonresponse in Household Interview Surveys*. New York : John Wiley & Sons, Inc.

Kim, T. (1995). Discriminant analysis as a prediction tool for uncommitted voters in pre-election polls. *International Journal of Public Opinion Research*, 7, 110-127.

Lau, R.R. (1994). An analysis of the accuracy of "trial heat" polls during the 1992 presidential elections. *Public Opinion Quarterly*, 59, 589-605.

Lavrakas, P.J. (1993). *Telephone Survey Method*: *Sampling, selection, and supervision*. 2ième Édition. Newbury Park, Calif. : Sage.

Little, J.A., et Rubin, D.B. (2002). *Statistical analysis with missing data*. 2ième Édition. New York : John Wiley & Sons, Inc.

Martin, E.A., Traugott, M.W. et Kennedy, C. (2005). A review and proposal for a new measure of poll accuracy. *The Public Opinion Quarterly*, 69, 342-369.

Michiels, B., et Molenberghs, G. (1997). Protective estimation of longitudinal categorical data with nonrandom drop-out. *Communications in Statistics*: *Theory and Methods*, 26, 65-94.

Molenberghs, G., Kenward, M.G. et Goetghebeur, E. (2001). Sensitivity analysis for incomplete contingency tables: The Slovenian plebiscite case. *Applied Statistics*, 50, 15-29.

Monterola, C., Lim, M., Garcia, F. et Saloma, C. (2001). Feasibility of a neural network as classifier of undecided respondents in a public opinion survey. *International Journal of Public Opinion Research*, 14, 222-299.

Myers, D.J., et O'Connor, R.E. (1983). The undecided respondents in mandatory voting settings: A Venezuelan exploration. *The Western Political Quarterly*, 36, 420-433.

Park, T., et Brown, M.B. (1994). Models for categorical data with nonignorable nonresponse. *Journal of the American Statistical Association*, 89, 44-52.

Park, T. (1998). An approach to categorical data with nonignorable nonresponse. *Biometrics*, 54, 1579-1690.

Perry, P. (1979). Certain problem in election survey methodology. *Public Opinion Quarterly*, 43, 312-325.

Potthoff, R.F. (1994). Telephone sampling in epidemiologic research: To reap the benefits, avoid the pitfalls. *American Journal of Epidemiology*, 139, 967-978.

Rubin, D.B., Stern, H.S. et Vehovar, V. (1995). Handling "Don't Know" survey responses: the case of the Slovenian plebiscite. *Journal of the American Statistical Association*, 90, 822-828.

Smith, P.W.F., Skinner, C.J. et Clarke, P.S. (1999). Allowing for non-ignorable nonresponse in the analysis of voting intention data. *Applied Statistics*, 48, 563-577.

Smith, T.W. (1984). Non attitudes: A review and evaluation. Dans *Surveying Subjective Phenomena*, (Éds. C.F. Turner et E. Martin), New York : Russell Sage Foundation, 2, 215-255.

Stasny, E.A. (1986). Estimating gross flow using panel data with nonresponse: An example from the Canadian Labor Force survey. *Journal of the American Statistical Association*, 81, 42-47.

Stasny, E.A. (1988). Modelling nonignorable nonresponse in categorical panel data with an example in estimating Gross Labor-Force flows. *Journal of Business and Economic Statistics*, 6, 207-219.