N° 11-522-XIF au catalogue

La série des symposiums internationaux de Statistique Canada - Recueil

Symposium 2003 : Défis reliés à la réalisation d'enquêtes pour la prochaine décennie

2003

Statistique Canada Statistics Canada

Canada

Recueil du Symposium 2003 de Statistique Canada
Défis reliés à la réalisation d'enquêtes pour la prochaine décennie

MÉTHODES DE RÉÉCHANTILLONNAGE POUR L'ESTIMATION EQM AVEC DES MODÈLES RÉGIONAUX NON LINÉAIRES

Sharon L. Lohr et J.N.K. Rao[1]

RÉSUMÉ

Les méthodes d'estimation régionale visant des tailles d'échantillon de sous-populations d'intérêt trop petites pour une estimation directe fiable ont été appliquées à une grande diversité de problèmes d'enquête par sondage et de biostatistique. Nombre de modèles d'estimation régionale livrent la valeur théorique approximative de l'erreur quadratique moyenne (EQM) pour la caractéristique d'intérêt dans chaque petite région. L'estimation de l'EQM est cependant plus complexe, plus particulièrement lorsque le modèle exploité n'est pas linéaire. Nous décrivons une méthode jackknife proposée par Rao (2003) pour l'estimation EQM lorsque des modèles linéaires généralisés ou d'autres modèles non linéaires visent la réponse d'intérêt et nous en démontrons le rendement dans une étude de simulation. À la différence de certaines autres méthodes en usage, la méthode jackknife donne des estimations régionales précises de l'EQM et réduit la masse de calculs à prévoir.

MOTS CLÉS : Bootstrap, erreur quadratique moyenne, jackknife, modèle bêta-binomial.

1. INTRODUCTION

On se sert couramment de méthodes d'estimation régionale pour faire des inférences à propos de sous-populations dont les échantillons sont trop petits pour une estimation directe suffisamment fiable. Comme exemples, on peut mentionner l'estimation des taux de pauvreté à l'âge scolaire dans les divers comtés, l'estimation de la prévalence diabétique dans les sous-groupes démographiques, l'établissement des taux de protection d'assurance-maladie dans les divers États et l'estimation des taux de survie des leucémiques dans les différents hôpitaux. D'ordinaire, on emploie des modèles pour utiliser l'information d'autres petites régions à l'aide de covariables comme les données censitaires et administratives. Le livre récent de Rao (2003) donne un aperçu de la diversité des méthodes et des modèles servant actuellement aux estimations régionales.

Posons l'existence de m sous-populations ou petites régions d'intérêt. Pour la région i, la caractéristique d'intérêt est désignée par θ_i. Ainsi, θ_i peut être le taux de paupérisme dans le comté i, la tension sanguine moyenne dans la sous-population i ou la proportion de diabétiques dans le sous-groupe démographique i.

Dans maintes méthodes répandues d'estimation régionale, on adopte un modèle hiérarchisé à deux degrés. Au premier degré, le vecteur de données de la région i, $\mathbf{y}_i$, se présente sous la forme $\mathbf{y}_i \mid \theta_i \sim f_1(\mathbf{y}_i \mid \theta_i, \lambda)$. Au second degré, le modèle lie la caractéristique d'intérêt θ_i à d'autres régions et à des covariables. Une forme générale de ce modèle est $\theta_i \sim f_2(\mathbf{x}_i, v_i, \phi)$, où $\mathbf{x}_i$ est un vecteur de covariables de la région i, ϕ un vecteur de paramètres inconnus et v_i une variable aléatoire. On pose que les paires $(\mathbf{y}_i, \theta_i)$ sont indépendantes. Dans ce modèle, θ_i est une quantité aléatoire. Le but est de trouver le meilleur prédicteur de θ_i et un estimateur de son erreur quadratique moyenne.

En guise d'illustration, considérons le modèle mixte régional de Fay-Herriot (1979). Dans le cas le plus simple, $\bar{y}_i$ désigne l'estimateur de la moyenne de population dans la région i à partir des données d'enquête. Ainsi, $\bar{y}_i$ pourrait

[1]Sharon L. Lohr, Département de mathématiques et de statistique, Université d'État de l'Arizona, Tempe, Arizona, États-Unis 85287-1804 (sharon.lohr@asu.edu); J. N. K. Rao, École de mathématiques et de statistique, Université Carleton, Ottawa, Ontario, Canada K1S 5B6 (jrao@math.carleton.ca).

être l'estimateur direct du taux de paupérisme dans le cadre de l' enquête sur l'état de la population. Au premier degré, on a $\bar{y}_i \mid \theta_i = \theta_i + e_i$, où $e_i \sim N(0, \psi_i)$ représente l'erreur d'échantillonnage de l'enquête. Au second degré, on pose que $\theta_i = \mathbf{x}_i^T\beta + v_i$, où $v_i \sim N(0, \sigma^2)$ représente l'erreur du modèle par rapport à la moyenne régionale. Dans ce cas, le vecteur de paramètres inconnus est $\phi = (\beta, \sigma^2)$. Dans un tel cadre, la codistribution de $\bar{y}_i$ et θ_i est connue et, par conséquent, le meilleur prédicteur – prédicteur comportant la moindre erreur quadratique moyenne – de θ_i si les valeurs de β et σ^2 sont connues est la moyenne de la distribution postérieure,

$$\hat{\theta}_i^B = E\left[\theta_i \mid \bar{y}_i, \beta, \sigma^2\right] = \gamma_i \bar{y}_i + (1-\gamma_i)\mathbf{x}_i^T\beta, \tag{0.1}$$

où

$\gamma_i = \sigma^2/(\sigma^2 + \psi_i)$. La variance postérieure de θ_i si les valeurs de β et σ^2 sont connues est

$$V\left[\theta_i \mid \bar{y}_i, \beta, \sigma^2\right] = E\left[(\hat{\theta}_i^B - \theta_i)^2 \mid \beta, \sigma^2\right] = \gamma_i\psi_i. \tag{0.2}$$

La variance postérieure en (1.2) est l'EQM de $\hat{\theta}_i^B$ si β et σ^2 sont connus.

Si β et σ^2 sont inconnus, $\hat{\theta}_i^B$ ne peut servir de prédicteur de θ_i. On peut alors plutôt employer le meilleur prédicteur empirique $\hat{\theta}_i^{\mathrm{EB}} = \hat{\gamma}_i\bar{y}_i + (1-\hat{\gamma}_i)\mathbf{x}_i^T\hat{\beta}$, où $\hat{\gamma}_i = \hat{\sigma}^2/(\hat{\sigma}^2 + \psi_i)$ et les statistiques $\hat{\beta}$ et $\hat{\sigma}^2$ se calculent à partir des données. Le surcroît d'incertitude qui tient à l'estimation de β et de σ^2, et non pas à la connaissance de leurs valeurs, confère à $\hat{\theta}_i^{\mathrm{EB}}$ une plus grande EQM qu'à $\hat{\theta}_i^B$, puisque

$$\begin{aligned} \mathrm{EQM}\left[\hat{\theta}_i^{EB}\right] &= E\left[(\hat{\theta}_i^{EB} - \theta_i)^2\right] \\ &= E\left[(\hat{\theta}_i^B - \theta_i)^2\right] + E\left[(\hat{\theta}_i^{EB} - \hat{\theta}_i^B)^2\right] \\ &= \gamma_i\psi_i + O(1/m)\,. \end{aligned} \tag{0.3}$$

Ainsi, si le simple estimateur $\hat{\gamma}_i\psi_i$ servait à l'estimation de l'EQM de $\hat{\theta}_i^{\mathrm{EB}}$, en ignorant l'erreur en estimant β et σ^2, l'EQM serait biaisé négativement.

Des problèmes semblables se posent dans le cas de modèles non linéaires. Ainsi, on applique souvent le modèle bêta-binomial à des données binaires. Dans un tel modèle, la caractéristique d'intérêt θ_i est une proportion souvent désignée par p_i. Les n_i valeurs observées de la région i à partir des données d'enquête, y_{ij}, sont jugées avoir été indépendamment tirées de variables aléatoires de Bernoulli avec le paramètre p_i, si bien que, pour $y_{i+} = \sum_{j=1}^{n_i} y_{ij}$, nous obtenons $y_{i+} \mid p_i \sim B(n_i, p_i)$. Le modèle au second degré reliant les proportions p_i des régions est $p_i \sim Beta(\alpha, \beta)$ pour $\alpha > 0$ et $\beta > 0$. Comme dans le modèle linéaire, on pose que les paires (y_{i+}, p_i) sont indépendantes entre régions. Si les paramètres α et β sont connus, on a $p_i \mid y_{i+}, \alpha, \beta \sim \mathrm{Beta}(y_{i+} + \alpha, n_i - y_{i+} + \beta)$. Dans un tel cas, le meilleur prédicteur de p_i est

$$\hat{p}_i^B = E\left[p_i \mid y_{i+}, \alpha, \beta\right] = \frac{y_{i+} + \alpha}{n_i + \alpha + \beta} \tag{0.4}$$

avec

$$V\left[p_i \mid y_{i+},\alpha,\beta\right] = E\left[(\hat{p}_i^B - p_i)^2 \mid y_{i+},\alpha,\beta\right] = \frac{(y_{i+}+\alpha)(n_i - y_{i+}+\alpha)}{(n_i+\alpha+\beta+1)(n_i+\alpha+\beta)^2}. \quad (0.5)$$

Si α et β sont inconnus, le meilleur prédicteur empirique de p_i obtenu par substitution des valeurs estimées de α et β aux valeurs réelles inconnues est $\hat{p}_i^{EB} = (y_i + \hat{\alpha})/(n_i + \hat{\alpha} + \hat{\beta})$. Comme dans le modèle linéaire, si l'erreur en estimant α et β est ignorée lors de l'estimation du MSE de $\hat{p}_i^{EB}$, l'estimateur du MSE sera biaisé négativement.

Il existe une différence d'expression pour la variance postérieure du modèle linéaire en (1.2) et du modèle non linéaire en (1.5). Dans le premier cas, l'expression $V\left[\theta_i \mid \overline{y}_i, \beta, \sigma^2\right]$ ne dépend pas de $\overline{y}_i$ et, par conséquent, $E\left[(\hat{\theta}_i^B - \theta_i)^2 \mid \beta, \sigma^2\right] = V\left[\theta_i \mid \overline{y}_i, \beta, \sigma^2\right]$. L'expression de variance conditionnelle en (1.5) dépend cependant de la valeur de y_{i+}, si bien que, en général, l'erreur quadratique moyenne conditionnelle EQMC = $E\left[(\hat{p}_i^B - p_i)^2 \mid y_{i+},\alpha,\beta\right]$ ne correspond pas à l'erreur quadratique moyenne inconditionnelle EQMI = $E\left\{E\left[(\hat{p}_i^B - p_i)^2 \mid y_{i+},\alpha,\beta\right]\right\}$. Un estimateur sans biais pour l'EQMI ne le sera pas nécessairement pour l'EQMC.

Dans cet article, nous examinerons la stabilité et les biais conditionnels et inconditionnels de plusieurs estimateurs proposés pour l'estimation de l'EQMI de modèles régionaux. À la section 2, nous passerons en revue les grandes méthodes susceptibles d'être appliquées à l'estimation de l'erreur quadratique moyenne dans des modèles linéaires ou non. À la section 3, nous présenterons un estimateur jackknife EQMC proposé par Rao (2003, p. 199) et en décrirons les propriétés. À la section 4, nous exposerons les résultats d'une petite étude de simulation et, la section 5 contient une discussion.

2. MODES D'ESTIMATION DE L'ERREUR QUADRATIQUE MOYENNE

L'erreur quadratique moyenne (EQM) de $\hat{\theta}_i^{\mathrm{EB}}$ dans le modèle mixte linéaire est donnée en (1.3). On a recouru à plusieurs méthodes d'estimation de l'EQM. L'estimateur le plus simple, que nous appellerons le simple estimateur EQM, se contente de substituer des valeurs estimées aux valeurs réelles inconnues des paramètres en (1.2) en ne tenant pas compte du surcroît de variabilité attribuable à une telle estimation. Dans le modèle linéaire, le simple estimateur EQM est $\hat{\gamma}_i\psi_i$.

Prasad et Rao (1990) ont démontré que ce dernier est biaisé négativement et ils ont conçu un mode analytique d'estimation EQM dans des modèles mixtes linéaires. Ils ont appliqué la formule d'estimation

$$\mathrm{eqm}_{\mathrm{PR}}\left[\hat{\theta}_i^{\mathrm{EB}}\right] = \hat{\gamma}_i\psi_i + g_{2i} + 2g_{3i},$$

où les termes g_{2i} et $2g_{3i}$ rendent compte du surcroît de variabilité attribuable à l'estimation de β et de σ^2. Ils ont prouvé que, sous certaines conditions de régularité, E[eqm($\hat{\theta}_i^{\mathrm{EB}}$)] = EQM($\hat{\theta}_i^{\mathrm{EB}}$)+$o$(1/$m$), de sorte que l'estimateur EQM est approximativement sans biais. Lahiri et Rao (1995) ont montré que l'estimateur EQM de Prasad-Rao est approximativement sans biais même si l'hypothèse de normalité sur v_i ne se vérifie pas, et ce, à condition qu'il existe suffisamment de moments de ν_i..

Les résultats analytiques de Prasad-Rao ont été appliqués à d'autres cas comme celui des modèles mixtes linéaires généralisés (voir l'examen de la question dans Rao, 2003, section 5.6). Dans un traitement analytique cependant, les résultats de chaque cas doivent s'obtenir séparément, et on doit écrire du code pour chaque situation. Les méthodes

de rééchantillonnage permettent d'éviter certains de ces problèmes et divers auteurs ont proposé les méthodes bootstrap et jackknife pour l'estimation de l'EQM d'estimateurs régionaux.

Pfeffermann et Tiller (2001) ont proposé pour leur part un estimateur bootstrap paramétrique de l'EQM[$\hat{\theta}_i^{\mathrm{EB}}$] dans le contexte d'une modélisation d'espace d'états. Ils ont estimé le surcroît de variabilité dû à l'estimation du paramètre inconnu ϕ en produisant des séries bootstrap à partir du modèle tiré des données d'origine pour ainsi démontrer que, dans les conditions de régularité proposées, le biais de la méthode bootstrap était $o(1/m)$.

Jiang, Lahiri et Wan (2002) ont établi les propriétés d'un estimateur jackknife de l'EQM dans un cadre général. Ils ont procédé à la décomposition orthogonale

$$\begin{aligned} \mathrm{EQM}\left[\hat{\theta}_i^{EB}\right] &= E\left[(\hat{\theta}_i^{B}-\theta_i)^2\right] + E\left[(\hat{\theta}_i^{EB}-\hat{\theta}_i^{B})^2\right] \\ &= \quad M_{1i} \quad + \quad M_{2i} \end{aligned} \tag{1.1}$$

et ont employé la méthode jackknife pour estimer M_{1i} et M_{2i} séparément. Cette méthode s'applique de la manière suivante au modèle mixte linéaire de Fay-Herriot. Dans la j^{e} itération jackknife, la région j est supprimée. On se reporte aux données des régions restantes (m–1) pour calculer $\hat{\beta}_{(-j)}$, $\hat{\sigma}^2_{(-j)}$ et $\hat{\gamma}_{i(-j)} = \hat{\sigma}^2_{(-j)}/(\hat{\sigma}^2_{(-j)} + \psi_i)$. Ces quantités servent ensuite au calcul d'un estimateur de suppression j de θ_i sous la forme $\hat{\theta}^{EB}_{i(-j)} = \hat{\gamma}_{i(-j)}\overline{y}_i + (1-\hat{\gamma}_{i(-j)})\mathbf{x}_i^T\hat{\beta}_{(-j)}$. Jiang, Lahiri et Wan (2002) ont estimé M_{1i} par

$$\hat{M}_{1i} = \hat{\gamma}_i\psi_i - \frac{m-1}{m}\sum_{j=1}^{m}\left[\hat{\gamma}_{i(-j)}\psi_i - \hat{\gamma}_i\psi_i\right]. \tag{1.2}$$

La quantité $\hat{M}_{1i}$ en (2.2) correspond au simple estimateur $\hat{\gamma}_i\psi_i$, plus une correction du biais. Le second terme en (2.1) est estimé par

$$\hat{M}_{2i} = \frac{m-1}{m}\sum_{j=1}^{m}\left[\hat{\theta}^{EB}_{i(-j)} - \hat{\theta}^{EB}_i\right]^2. \tag{1.3}$$

Jiang, Lahiri et Wan (2002) démontrent que l'estimateur jackknife EQM, $\mathrm{eqm}_{\mathrm{JLW}} = \hat{M}_{1i} + \hat{M}_{2i}$, est entaché d'un biais de l'ordre $o(1/m)$ si les conditions de régularité sont réunies.

Jiang, Lahiri et Wan (2002) définissent cet estimateur de la même façon pour des modèles non linéaires. En général, si $k_i(\phi) = M_{1i} = E\left[(\hat{\theta}_i^B - \theta_i)^2\right]$,

$$\hat{M}_{1i} = k_i(\hat{\phi}) - \frac{m-1}{m}\sum\nolimits_j\left[k_i(\hat{\phi}_{(-j)}) - k_i(\hat{\phi})\right] \tag{1.4}$$

et $\hat{M}_{2i}$ est comme en (2.3). Pour le modèle bêta-binomial décrit à la section 1,

$$k_i(\alpha,\beta) = E\left[\left(\hat{p}_i^B - p_i\right)^2 \mid \alpha,\beta\right] = \frac{\alpha}{(n_i+\alpha+\beta+1)(n_i+\alpha+\beta)^2}\left\{n_i+\beta+\frac{n_i(n_i-1)\alpha\beta}{(\alpha+\beta)(\alpha+\beta+1)} + \frac{n_i\alpha(\beta-\alpha)}{\alpha+\beta}\right\}.$$

Dans ce cas, $k_i(\phi)$ avec $\phi = (\alpha,\beta)$ peut se calculer analytiquement. Pour la plupart des autres cas non linéaires, ce qui comprend la plupart des modèles mixtes linéaires généralisés, le calcul de $k_i(\phi)$ exige une intégration numérique, laquelle doit se faire (m+1) fois pour l'estimation du terme M_{1i}.

3. MÉTHODE JACKKNIFE D'ESTIMATION RÉGIONALE

Avec les estimateurs EQM de la section 2, on estime l'EQM inconditionnelle EQMI = $E\left[(\hat{\theta}_i^{EB} - \theta_i)^2\right]$.

Dans des modèles linéaires, les EQM inconditionnelle et conditionnelle coïncident. Dans les modèles non linéaires toutefois, l'EQMC diffère de l'EQMI. Parfois, il sera peut-être préférable d'estimer l'EQMC. Booth et Hobert (1998) font valoir que, si les paires ($\mathbf{y}_i$, θ_i) sont indépendantes entre régions et que ϕ est connu, l'inférence au sujet de θ_i devrait être fondée sur la distribution conditionnelle de $\theta_i \mid \mathbf{y}_i, \phi$. Comme ils le disent, seules les observations relatives à la i^e région sont utiles à la prévision de [θ_i] lorsque les paramètres sont connus (p. 265). Fuller (1990) a aussi présenté un estimateur EQMC pour la prévision des valeurs réelles ou vraies à partir du modèle d'erreur de mesure. Dans la présente étude, nous n'utiliserons pas les méthodes hiérarchisées de Bayes, mais il faut noter que, d'un point de vue bayésien, l'EQMC est la variance postérieure si les paramètres ϕ sont connus. Un estimateur EQMC est « régional » pour des modèles mixtes non linéaires, parce qu'il dépend des données observées dans la région *i*.

Dans cette section, nous examinons un estimateur jackknife de l'EQM par lequel on calcule l'EQMC (et donc aussi l'EQMI), ce qui est brièvement proposé par Rao (2003, section 9.4). L'estimateur jackknife régional de l'EQM est

$$\text{eqm}_{AS}(\hat{\theta}_i^{EB}) = \hat{M}_{A1i}(\mathbf{y}_i) + \hat{M}_{2i}, \tag{3.1}$$

où

$$\hat{M}_{A1i}(\mathbf{y}_i) = g_i(\hat{\phi}, \mathbf{y}_i) - \frac{m-1}{m}\sum\nolimits_j [g_i(\hat{\phi}_{(-j)}, \mathbf{y}_i) - g_i(\hat{\phi}, \mathbf{y}_i)] \tag{3.2}$$

et où $g_i(\phi, \mathbf{y}_i) = V(\theta_i \mid \mathbf{y}_i, \phi)$ est la variance postérieure de θ_i compte tenu des données de la région *i* et des valeurs de ϕ. La quantité $\hat{M}_{2i}$ est définie en (2.3). On peut démontrer que, dans des conditions de régularité appropriées, la méthode jackknife régionale en (3.1) est entachée d'un biais conditionnel de l'ordre $o_p(1/m)$ et que le biais de l'estimation EQMI est, lui, de l'ordre $o(1/m)$.

Pour le modèle linéaire de Fay-Herriot, la méthode jackknife décrite en (3.1) est identique à celle de Jiang, Lahiri et Wan (2002), puisque dans ce cas $g_i = k_i$. Dans des modèles non linéaires cependant, la méthode jackknife régionale comporte un faible biais pour l'EQMC et l'EQMI, ce biais étant faible seulement pour l'EQMI avec la méthode de Jiang, Lahiri et Wan. De plus, le calcul jackknife régional est bien plus simple, car on n'a pas à trouver l'espérance de $g_i(\phi, \mathbf{y}_i)$.

La moindre complexité de calcul par la méthode jackknife régionale a pour résultat non seulement moins de temps de calcul, mais aussi peut-être moins d'erreurs numériques. Dans peu d'études consacrées aux estimateurs jackknife, il est question des erreurs numériques des méthodes employées, mais il faut préciser que les effets accumulés de telles erreurs peuvent être importants. Les sommations en (2.3), (2.4) et (3.2) sont normalement petits et sont donc sensibles aux erreurs numériques. On recourt souvent à une estimation de maximum de vraisemblance (EMV) pour calculer $\hat{\phi}$, d'ordinaire à l'aide d'un algorithme itératif comme celui de Newton-Raphson (voir dans Press et coll., 1992, un examen de certaines des inexactitudes numériques qui peuvent se glisser dans les résultats). Si l'estimation $\hat{\phi}$ de l'ensemble des données comporte des erreurs numériques et que celles-ci se propagent dans les quantités jackknife $\hat{\phi}_{(-j)}$, la correction jackknife en (3.2) peut donner une estimation moins fidèle que la simple estimation. Il y a tout particulièrement risque si on se sert de sous-programmes écrits par d'autres chercheurs pour établir les estimations paramétriques, puisque ces programmes spécifient parfois l'itération seulement jusqu'à une tolérance 0,001 ou une valeur plus grande encore. Pour qu'elle puisse s'employer avec la méthode jackknife, l'estimation des paramètres inconnus ϕ par l'ensemble des données devrait être calculée aussi près de la précision-machine que possible.

Avec une estimation $\hat{\phi}$ numériquement exacte, l'algorithme de Newton-Raphson peut servir à accélérer le calcul de $\hat{\phi}_{(-j)}$. Dans les itérations jackknife, les EMV de l'ensemble des données peuvent faire fonction de valeurs initiales et l'algorithme de Newton-Raphson peut être appliqué avec des dérivées numériques dans le calcul de $\hat{\phi}_{(-j)}$ de chaque itération jackknife. On parvient généralement à la convergence dans une des quatre étapes de cette application, et cette convergence se réalise plus rapidement pour $\hat{\phi}_{(-j)}$ lorsque *m* est plus grand.

Dans la pratique, $\hat{M}_{1i}$ en (2.4) ou $\hat{M}_{A1i}(\mathbf{y}_i)$ en (3.2) peut être négatif, auquel cas nous recommandons de substituer $k_i(\hat{\phi})$ à $\hat{M}_{1i}$ et $g_i(\hat{\phi}, \mathbf{y}_i)$ à $\hat{M}_{A1i}(\mathbf{y}_i)$.

4. ÉTUDE DE SIMULATION

Par une petite étude de simulation, nous avons comparé la méthode jackknife régionale au simple estimateur EQM et à l'estimateur jackknife proposé par Jiang, Lahiri et Wan (2002). Le modèle bêta-binomial décrit à la section 1 a servi à produire les données. Dans cette étude de simulation, nous avons appliqué un plan factoriel avec trois valeurs de *m* (10, 30 et 60) et autant de valeurs tant pour α que pour β (0,1, 1,0 et 10,0). Les tailles d'échantillon intrarégionales ont pris des valeurs $n_i \in \{1, 2, 3, 4, 5\}$; chaque valeur a été utilisée un nombre égal de fois pour les *m* petites régions.

Tous les calculs se sont faits en R, version 1.6.1 (obtenue à www.r-project.org). Il y a eu un millier de passages de simulation pour chaque combinaison de facteurs. Pour simplifier les calculs de simulation, nous avons appliqué des estimateurs de la méthode des moments à α et β. Si $\hat{p} = \sum y_{i+} / \sum n_i$ et $\hat{\sigma}^2 = [\sum y_{i+}(y_{i+} - 1)] / [\sum n_i (n_i - 1)] - \hat{p}^2$, $\hat{\alpha} = \hat{p}[\hat{p}(1-\hat{p}) / \hat{\sigma}^2 - 1]$ et $\hat{\beta} = (1-\hat{p})\hat{\alpha} / \hat{p}$ sont les estimateurs de méthode des moments pour α et β lorsque $\hat{\alpha}$ et $\hat{\beta}$ sont définis et positifs. Pour les passages de simulation où $\hat{\alpha}$ ou $\hat{\beta}$ ci-dessus n'était ni positif ni défini, nous avons posé que ces éléments étaient des valeurs importantes avec $\hat{\alpha}/(\hat{\alpha} + \hat{\beta}) = \hat{p}$.

Pour chaque réglage de facteurs expérimentaux, nous avons estimé l'EQMI réelle séparément pour chaque valeur de la taille d'échantillon de petite région n_i sous la forme $EQMI(n_i) = (1/1000)\sum_{k=1}^{1000} \left(\hat{p}_{ik}^{EB} - P_{ik}\right)^2$, où p_{ik} est la valeur réelle ou vraie de p_i issue du k^e passage de simulation et où $\hat{p}_{ik}^{EB}$ est l'estimation de p_i obtenue dans ce passage. Nous avons estimé de même l'EQMC (y_{i+}, n_i) en prenant la moyenne des écarts quadratiques sur les passages de simulation qui comportaient la valeur spécifiée de y_{i+}. Nous avons ensuite calculé le biais relatif inconditionnel d'un estimateur EQM, eqm, pour chaque taille d'échantillon sous la forme $100[epm(n_i) - EQMI(n_i)] / EQMI(n_i)$ où $eqm(n_i)$ est la moyenne des estimations EQM pour cette taille d'échantillon. Le biais relatif conditionnel a fait l'objet d'un calcul semblable en utilisant $100[eqm(y_{i+}, n_i) - EQMC(y_{i+}, n_i)] / EQMC(y_{i+}, n_i)$, où $eqm(y_{i+}, n_i)$ est la moyenne des estimations EQM pour les passages de simulation comportant cette valeur de y_{i+}. Nous avons enfin calculé le coefficient de variation (CV) par (écart-type de l'eqm)/EQM.

Les tableaux 1 à 3 récapitulent les résultats partiels pour $\alpha = \beta = 1$. Les chiffres de ces tableaux sont les moyennes des valeurs absolues de biais relatif (BRA) et des CV pour toutes les tailles d'échantillon. Aux lignes des tableaux livrant des valeurs conditionnelles, nous avons d'abord mis en moyenne – pour faciliter les comparaisons – les valeurs obtenues des différents y_{i+}.

Tableau 1. Résultats de simulation pour *m* = 10

	Simple estimateur	Jiang et coll. (2002)	Méthode jackknife régionale
BRA inconditionnelle	61 (toujours négatif)	73 (le plus souvent positif)	72 (le plus souvent positif)
CV inconditionnel	0,3	2,6	2,7
BRA conditionnelle	60 (toujours négatif)	96 (le plus souvent positif)	88 (le plus souvent positif)
CV conditionnel	0,3	2,3	2,2

Tableau 2. Résultats de simulation pour *m* = 30

	Simple estimateur	Jiang et coll. (2002)	Méthode jackknife régionale
BRA inconditionnelle	31,8 (toujours négatif)	3,8	2,6
CV inconditionnel	0,30	0,51	0,53
BRA conditionnelle	32,4 (toujours négatif)	22,3	7,5
CV conditionnel	0,24	0,53	0,57

Tableau 3. Résultats de simulation pour *m* = 60

	Simple estimateur	Jiang et coll. (2002)	Méthode jackknife régionale
BRA inconditionnelle	17,1 (toujours négatif)	1,9	1,6
CV inconditionnel	0,24	0,17	0,24
BRA conditionnelle	17,8 (toujours négatif)	20,4	5,5
CV conditionnel	0,16	0,16	0,21

Les tableaux 1 à 3 démontrent les résultats théoriques d'une sous-estimation des valeurs réelles ou vraies par le simple estimateur EQM. Celui-ci sous-estimait la valeur réelle de l'EQMI ou de l'EQMC pour toute valeur de n_i (et de y_{i+} pour l'EQMC) et, par conséquent, le biais relatif était toujours d'une valeur négative. Pour $m = 10$ petites régions, les deux méthodes jackknife sont aussi entachées d'un biais important, mais la plupart des valeurs se trouvent à surestimer l'EQMI ou l'EQMC. S'il y a surestimation, c'est en grande partie en raison de l'instabilité de l'estimateur de la méthode des moments pour de petites tailles d'échantillon. Avec 10 régions seulement, la suppression d'une de ces régions avait parfois pour effet de modifier grandement les estimations de α et β, surtout lorsqu'une des estimations de méthode des moments était non positive pour une itération jackknife. Ces changements marqués ont donné lieu à une estimation trop importante de l'EQM.

Pour $m = 30$ ou 60, le tableau est différent. Avec le simple estimateur, on sous-estime toujours la valeur réelle de l'EQMI et de l'EQMC pour chaque cas étudié. Et la méthode de Jiang, Lahiri et Wan (2002) et la méthode régionale comportent un léger biais d'estimation de l'EQMI, comme on pouvait s'y attendre. Les deux méthodes jackknife produisent cependant des résultats différents d'estimation de l'EQMC. La méthode régionale est entachée d'un petit biais conditionnel, qui diminue à mesure qu'augmente *m*. En revanche, la méthode de Jiang, Lahiri et Wan donne de fortes valeurs pour le biais conditionnel, celui-ci étant soit élevé et positif si y_{i+}/n_i est petit ou grand, soit élevé et négatif si y_{i+}/n_i est proche de 0,5. Avec cette méthode, le biais ne diminue toutefois pas à mesure qu'augmente *m*.

L'estimateur jackknife régional comporte un léger biais inconditionnel et conditionnel, mais le biais conditionnel moindre s'obtient au prix d'une certaine instabilité de l'estimateur EQM. La méthode de Jiang, Lahiri et Wan (2002) a un CV quelque peu moindre pour un grand nombre de régions, bien que, même pour $m = 60$, la différence de CV entre les deux méthodes ne soit pas prononcée. La méthode régionale est d'un calcul bien plus simple même pour le cas bêta-binomial lorsque l'intégrale supplémentaire $E[V(\theta_i \mid \mathbf{y}_i, \phi)]$ peut s'évaluer analytiquement.

5. EXAMEN

Dans ce document, nous avons examiné et comparé trois méthodes d'estimation EQM dans des modèles régionaux linéaires ou non, à savoir le simple estimateur, la méthode jackknife proposée par Jiang, Lahiri et Wan (2002) et une méthode jackknife régionale. Les deux méthodes jackknife rendent compte du surcroît de variabilité attribuable à l'ignorance des valeurs de paramètres inconnus comme β et σ^2 dans le modèle linéaire ou α et β dans le modèle bêta-binomial. L'une et l'autre, elles confèrent plus de précision à l'estimation de l'EQM inconditionnelle que le simple estimateur, mais la méthode jackknife régionale présente plusieurs propriétés souhaitables. Elle comporte un faible biais tant conditionnel qu'inconditionnel et nous procure donc un estimateur EQM régional précis. Son calcul se fait aussi en moins d'étapes et s'expose donc moins aux erreurs numériques.

La méthode jackknife régionale est un estimateur EQM prometteur dans le contexte des problèmes d'estimation régionale. Son calcul étant plus simple, elle peut s'adapter à des problèmes comme celui de l'estimation de l'EQM d'autres quantités comme la différence entre deux caractéristiques de sous-populations $\theta_i - \theta_j$.

REMERCIEMENTS

Cette étude a partiellement été financée par la subvention n° 0105852 de la U.S. National Science Foundation et par une subvention du Conseil de recherches en sciences naturelles et en génie (CRSNG) du Canada.

RÉFÉRENCES

Booth, J. et J. Hobert (1998), "Standard Errors of Prediction in Generalized Linear Mixed Models", *Journal of the American Statistical Association,* 93, pp. 262-272.

Fay, R. E. et R. A. Herriot (1979), "Estimates of Income for Small Places: An Empirical Bayes Application of James-Stein Procedures to Census Data", *Journal of the American Statistical Association*, 78, pp. 269-277.

Fuller, W. A. (1990), "Prediction of True Values for the Measurement Error Model", in P. J. Brown and W. A. Fuller (eds.), *Statistical Analysis of Measurement Error Models and Applications, Contemporary Mathematics Vol. 112,* Providence, RI: AMS, pp. 41-57.

Jiang, J., P. Lahiri et S.-M. Wan (2002), "A Unified Jackknife Theory for Empirical Best Prediction with *M*-Estimation", *Annals of Statistics*, 30, pp. 1782-1810.

Lahiri, P. et J. N. K. Rao (1995), "Robust Estimation of Mean Squared Error of Small Area Estimators", *Journal of the American Statistical Association*, 82, pp. 758-766.

Pfeffermann, D. et R. B. Tiller (2001), "Bootstrap Approximation to Prediction MSE for State-Space Models with Estimated Parameters", Technical report, Jerusalem Israel: Hebrew University.

Press, W. H., S. A. Teukolsky, W. T. Vetterling, et B. P. Flannery (1992), *Numerical Recipes in C, 2nd ed.*, Cambridge: Cambridge University Press.

Prasad, N. G. N. et J. N. K. Rao (1990), "The Estimation of the Mean Squared Error of Small-Area Estimators", *Journal of the American Statistical Association,* 85, pp. 163-171.

Rao, J. N. K. (2003), *Small Area Estimation*, New York: Wiley.