

ESTIMATION DANS LES PETITES RÉGIONS: UNE NOUVELLE DÉRIVATION DE L'ERREUR QUADRATIQUE MOYENNE DE PRASAD-RAO

Eve Belmonte¹

ABSTRACT

In sampling theory, serious problems occur when estimating parameters using a small sample size, which is often the case in small areas. Henceforth the direct-survey estimators are not reliable anymore.

By fitting a model for the survey estimates, we are able to improve the efficiency of the direct estimates. The Empirical Bayes method offers an interesting alternative by proposing a combined estimator, i.e. a weighted average of a survey and a synthetic estimator. Such new estimators bring the problem of accurately estimating their mean squared error (MSE). Although MSE estimators already exist in the literature, we suggest a conditional model independent MSE estimator. Finally, we will discuss its relationship with the Prasad-Rao estimator (1990).

KEY WORDS: Empirical Bayes; Mean Squared Error (MSE); Prasad-Rao Estimator; Conditionnal Estimator.

RÉSUMÉ

En échantillonnage, l'estimation de totaux ou de moyennes dans de petites régions est un sérieux problème lorsque la taille de l'échantillon observé est si petite que les estimateurs directs ne conviennent plus. En modélisant les estimations des petites régions, on arrive à contourner ce problème. La méthode bayésienne empirique offre une alternative intéressante en proposant un estimateur combiné, soit une moyenne pondérée d'un estimateur direct et d'un estimateur synthétique.

Un autre problème est de trouver un bon estimateur pour l'erreur quadratique moyenne (EQM) de tels estimateurs. Plusieurs estimateurs de l'EQM existent dans la littérature mais on présente ici un estimateur conditionnel de l'EQM qui mesure la variabilité par rapport au plan d'échantillonnage. On établit ensuite, à l'aide d'une preuve mathématique, la relation qui existe entre l'estimateur conditionnel et l'estimateur de Prasad et Rao (1990).

MOTS-CLÉS: Bayes empirique; erreur quadratique moyenne (EQM); estimateur de Prasad-Rao; estimateur conditionnel.

1 Introduction

Lors d'un sondage, on appelle une petite région: une petite région géographique ou encore une petite sous-population à l'intérieur d'un territoire donné. Cela peut être, par exemple, une région administrative particulière ou un groupe âge/sexe.

Les bureaux de statistique sont souvent appelés à produire des estimations de totaux ou de moyennes pour de petites régions. Comme la précision des estimateurs est déterminée à plus grande échelle (par exemple au niveau national ou provincial), les tailles échantillonnales de ces petites régions sont trop petites ou même inexistantes.

Cette situation cause un problème, car la variance des estimateurs directs est très grande ce qui les rend imprécis.

Il existe plusieurs méthodes pour contourner ce problème. Parmi les plus populaires, on retrouve les estimateurs synthétiques, les estimateurs combinés, les modèles linéaires (EBLUP) et les méthodes de Bayes empirique et hiérarchique. Dans cet article, on s'intéressera plus particulièrement à l'estimateur de Bayes empirique, qui est en réalité un estimateur combiné. Dans la section 2, on présentera le modèle bayésien, on dérivera l'estimateur de Bayes empirique et on discutera de l'estimation des

¹Eve Belmonte, Département de mathématiques et de statistique, Université Laval, Sainte-Foy, Québec, Canada, G1K 7P4, belmonte@mat.ulaval.ca

paramètres.

On rencontre un autre problème lorsqu'il s'agit de trouver une mesure précise de la variabilité de cet estimateur. Dans la section 3, on présentera deux manières d'interpréter l'erreur quadratique moyenne (EQM) avec leur estimateur respectif. Finalement, dans la section 4, on démontrera qu'il existe un lien entre ces deux estimateurs de l'EQM.

2 Modèle de Fay et Herriot (1979)

L'idée de la méthode de Bayes empirique est de modéliser les estimations directes dans le but de produire une série d'estimations plus précises. La construction d'un modèle semblable introduit un biais dans les estimateurs des petites régions, mais du même coup réduit considérablement leur EQM.

Soient $\mathbf{y} = (y_1, \dots, y_n)^t$ le vecteur des estimations directes pour les n petites régions et $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})^t$ un vecteur d'information auxiliaire de dimension p portant sur la petite région i .

Le modèle Fay-Herriot est un modèle linéaire en deux étapes:

$$y_i \sim_{iid} N(\theta_i, \sigma_{ii}) \quad (\text{modèle S})$$

$$\theta_i \sim_{iid} N(\mathbf{x}_i^t \boldsymbol{\beta}, \sigma_v^2) \quad (\text{modèle M})$$

pour $i = 1, \dots, n$.

Donc,

$$y_i = \mathbf{x}_i^t \boldsymbol{\beta} + v_i + e_i$$

où les erreurs v_i et e_i sont indépendantes et normalement distribuées de moyenne 0 et de variance respective σ_v^2 et σ_{ii} . On considère σ_{ii} connue, alors que $\boldsymbol{\beta}$ et σ_v^2 sont des paramètres inconnus qu'on devra estimer à l'aide des données y_1, \dots, y_n . Ici, on note le modèle échantillonnal, modèle S, et le modèle de régression, modèle M. On sait que le modèle échantillonnal S est vérifié au moins approximativement en vertu du théorème limite-central, étant donné que y_i est une somme de variables aléatoires. Par contre, le modèle M est moins évident à justifier, dépendant du contexte.

Pour obtenir l'estimateur de Bayes empirique, on commence par calculer la loi *a posteriori* de θ_i . Il est facile de démontrer que pour $i = 1, \dots, n$

$$\theta_i | y_i, \boldsymbol{\beta}, \sigma_v^2 \sim_{iid} N \left(\frac{\sigma_v^2 y_i + \sigma_{ii} \mathbf{x}_i^t \boldsymbol{\beta}}{\sigma_v^2 + \sigma_{ii}}, \frac{\sigma_{ii} \sigma_v^2}{\sigma_v^2 + \sigma_{ii}} \right).$$

Sous la fonction de perte quadratique, l'estimateur de Bayes est l'espérance *a posteriori* de θ_i qui peut se réécrire comme:

$$E[\theta_i | y_i] = \frac{\sigma_v^2}{\sigma_v^2 + \sigma_{ii}} y_i + \frac{\sigma_{ii}}{\sigma_v^2 + \sigma_{ii}} \mathbf{x}_i^t \boldsymbol{\beta}. \quad (1)$$

L'ennui dans (1), c'est que les paramètres $\boldsymbol{\beta}$ et σ_v^2 sont inconnus. On doit donc les estimer. L'estimateur de Bayes empirique, noté $\hat{\theta}_i^{EB}$, sera obtenu en substituant $\hat{\boldsymbol{\beta}}$ et $\hat{\sigma}_v^2$ à $\boldsymbol{\beta}$ et σ_v^2 dans (1):

$$\hat{\theta}_i^{EB} = \underbrace{\frac{\hat{\sigma}_v^2}{\hat{\sigma}_v^2 + \sigma_{ii}}}_{\lambda_i} y_i + \underbrace{\frac{\sigma_{ii}}{\hat{\sigma}_v^2 + \sigma_{ii}}}_{1-\lambda_i} \mathbf{x}_i^t \hat{\boldsymbol{\beta}}_w \quad (2)$$

où λ_i est le poids associé à l'estimateur direct tel que $0 \leq \lambda_i \leq 1$.

On remarque tout d'abord que $\hat{\theta}_i^{EB}$ est un estimateur combiné, i.e. une moyenne pondérée d'un estimateur direct (non-biaisé à grande variance) et d'un estimateur synthétique (biaisé mais à plus faible EQM). Le poids λ_i a été déterminé par l'approche bayésienne et sert à établir un compromis entre l'introduction d'un biais et la réduction de l'EQM.

Il existe plusieurs alternatives quant à l'estimation des paramètres $\boldsymbol{\beta}$ et σ_v^2 . Comme dans l'article de Ghosh et Rao (1994), on a choisi l'estimateur des moindres carrés pondérés pour estimer $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}}_w = (\mathbf{X}^t \hat{\mathbf{V}}^{-1} \mathbf{X})^{-1} \mathbf{X}^t \hat{\mathbf{V}}^{-1} \mathbf{y} \quad (3)$$

où \mathbf{X} est la matrice de régression: $\mathbf{X} = (\mathbf{x}_1^t, \dots, \mathbf{x}_n^t)^t$ et $\hat{\mathbf{V}} = \text{diag}(\hat{\sigma}_v^2 + \sigma_{ii})$. De plus, un estimateur non-biaisé pour σ_v^2 est:

$$\hat{\sigma}_v^2 = (n-p)^{-1} \left\{ \sum_{i=1}^n (y_i - \mathbf{x}_i^t \hat{\boldsymbol{\beta}})^2 - \sum_{i=1}^n \sigma_{ii} (1 - h_{ii}) \right\} \quad (4)$$

avec $\hat{\boldsymbol{\beta}} = (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t \mathbf{y}$ et $h_{ii} = \mathbf{x}_i^t (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{x}_i$.

3 Définition et évaluation de l'EQM

On distingue deux façons de définir l'erreur quadratique moyenne des estimateurs des petites régions. On appelle la première **EQM inconditionnelle**, notée EQM_1 :

$$EQM_1(\hat{\theta}_i^{EB}) = E_M \left[E_S \left[(\hat{\theta}_i^{EB} - \theta_i)^2 \right] \right] \quad (5)$$

où S et M indiquent les deux étapes du modèle Fay-Herriot.

Sous le modèle Fay-Herriot, une approximation de second ordre de (5) est donnée par (Prasad et Rao, 1990):

$$EQM_{PR}(\hat{\theta}_i^{EB}) = \frac{\sigma_v^2 \sigma_{ii}}{\sigma_v^2 + \sigma_{ii}} + \frac{\sigma_{ii}^2 h_{ii}^*}{(\sigma_v^2 + \sigma_{ii})^2} + \frac{\sigma_{ii}^2 \text{Var}(\hat{\sigma}_v^2)}{(\sigma_v^2 + \sigma_{ii})^3} \quad (6)$$

où $h_{ii}^* = \mathbf{x}_i^t (\mathbf{X}^t \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{x}_i$ et

$$\text{Var}(\hat{\sigma}_v^2) \approx \frac{2}{n} \left\{ (\sigma_v^2)^2 + 2\sigma_v^2 \sum_{i=1}^n \frac{\sigma_{ii}}{n} + \sum_{i=1}^n \frac{\sigma_{ii}^2}{n} \right\}. \quad (7)$$

Une approche naïve d'estimation de (6) serait de remplacer dans (6) les paramètres inconnus par leur estimation. Or, en procédant de la sorte, on sous-estimerait considérablement l'EQM, car on ne tiendrait pas compte de la variabilité additionnelle due à l'estimation de β et σ_v^2 .

Un estimateur de (6) a été dérivé par Prasad et Rao (1990) et est noté $eqm_{PR}(\hat{\theta}_i^{EB})$:

$$eqm_{PR}(\hat{\theta}_i^{EB}) = \frac{\hat{\sigma}_v^2 \sigma_{ii}}{\hat{\sigma}_v^2 + \sigma_{ii}} + \frac{\sigma_{ii}^2 h_{ii}^{**}}{(\hat{\sigma}_v^2 + \sigma_{ii})^2} + \frac{2 \sigma_{ii}^2 v(\hat{\sigma}_v^2)}{(\hat{\sigma}_v^2 + \sigma_{ii})^3} \quad (8)$$

où $h_{ii}^{**} = \mathbf{x}_i^t (\mathbf{X}^t \hat{\mathbf{V}}^{-1} \mathbf{X})^{-1} \mathbf{x}_i$ et $v(\hat{\sigma}_v^2)$ est l'estimateur de $\text{Var}(\hat{\sigma}_v^2)$ obtenu en remplaçant σ_v^2 par $\hat{\sigma}_v^2$.

Cet estimateur de l'EQM présuppose que le modèle de régression (M) est vrai et il est valide pour toutes les réalisations de l'échantillon.

La deuxième définition pour l'EQM est appelée **EQM conditionnelle** et est notée EQM_2 :

$$EQM_2(\hat{\theta}_i^{EB}) = E_S[(\hat{\theta}_i^{EB} - \theta_i)^2] \quad (9)$$

où l'espérance est calculée uniquement sur S , le modèle échantillonnal qu'on sait être approximativement vrai.

Un estimateur de (9) a été proposé par Rivest (1997):

$$eqm_R(\hat{\theta}_i^{EB}) = \max \left(0, \sigma_{ii} + 2\sigma_{ii} \frac{\partial g_i(\mathbf{y})}{\partial y_i} + g_i(\mathbf{y})^2 \right) \quad (10)$$

où la fonction $g_i(\mathbf{y})$ est telle qu'on peut réécrire l'estimateur de Bayes empirique sous la forme:

$$\begin{aligned} \hat{\theta}_i^{EB} &= y_i + g_i(\mathbf{y}) \\ &= y_i + \underbrace{\left(\frac{-\sigma_{ii}}{\hat{\sigma}_v^2 + \sigma_{ii}} \right) (y_i - \mathbf{x}_i^t \hat{\beta}_w)}_{g_i(\mathbf{y})}. \end{aligned} \quad (11)$$

Calculer la dérivée de la fonction g dans l'expression (10) peut être plus ou moins laborieux, tout dépendant de la complexité de cette fonction g . Mais sous le modèle de la section 2, cette dérivée se calcule assez facilement et l'estimateur conditionnel est:

$$\begin{aligned} eqm_R(\hat{\theta}_i^{EB}) &= \sigma_{ii} + 2\sigma_{ii} \left\{ \frac{\sigma_{ii} h_{ii}^{**}}{(\hat{\sigma}_v^2 + \sigma_{ii})^2} - \frac{\sigma_{ii}}{\hat{\sigma}_v^2 + \sigma_{ii}} \right. \\ &\quad + \frac{2}{(n-p)} \frac{\sigma_{ii}}{(\hat{\sigma}_v^2 + \sigma_{ii})^2} r_i^{(1)} r_i^{(2)} \\ &\quad + K_i \sum_{k=1}^n \frac{\mathbf{x}_k \mathbf{x}_k^t}{(\hat{\sigma}_v^2 + \sigma_{kk})^2} \hat{\beta}_w \\ &\quad \left. - K_i \sum_{k=1}^n \frac{\mathbf{x}_k y_k}{(\hat{\sigma}_v^2 + \sigma_{kk})^2} \right\} \\ &\quad + \frac{\sigma_{ii}^2}{(\hat{\sigma}_v^2 + \sigma_{ii})^2} r_i^{(2)2} \end{aligned} \quad (12)$$

où on a simplifié la notation en posant $r_i^{(1)} = (y_i - \mathbf{x}_i^t \hat{\beta})$, $r_i^{(2)} = (y_i - \mathbf{x}_i^t \hat{\beta}_w)$,

$$K_i = \frac{2}{(n-p)} \frac{\sigma_{ii}}{(\hat{\sigma}_v^2 + \sigma_{ii})} r_i^{(1)} \mathbf{x}_i^t \hat{\mathbf{A}}^{-1} \quad (13)$$

et

$$\hat{\mathbf{A}}^{-1} = (\mathbf{X}^t \hat{\mathbf{V}}^{-1} \mathbf{X})^{-1}. \quad (14)$$

L'estimateur conditionnel ne suppose pas que le modèle de régression (M) est vrai. Il est appelé estimateur conditionnel car il est conditionnel à cette réalisation de l'échantillon. De plus, il a l'avantage de pouvoir être calculé pour n'importe quelle technique d'estimation, en autant qu'on puisse trouver une fonction g comparable à celle de l'expression (11).

Les deux estimateurs de l'EQM présentés ci-dessus sont étroitement reliés. On démontrera, dans la prochaine section, le résultat suivant:

Résultat 3.1

$$\begin{aligned} E_M [eqm_R(\hat{\theta}_i^{EB})] &\approx E_M [eqm_{PR}(\hat{\theta}_i^{EB})] \\ \Rightarrow E_M [eqm_R(\hat{\theta}_i^{EB})] &\approx EQM_{PR}(\hat{\theta}_i^{EB}). \end{aligned}$$

4 Espérance de l'estimateur conditionnel

Dans cette section, on calcule l'espérance de l'estimateur de l'EQM conditionnelle de $\hat{\theta}_i^{EB}$, par rapport au modèle de régression M vu à la section 2. On verra qu'en conservant uniquement les termes d'ordre supérieur ou égal à $1/n$, on retrouve l'expression de l'approximation de l'EQM proposée par Prasad et Rao (équation (6)), qui est aussi l'espérance par rapport au modèle M de l'estimateur inconditionnel de Prasad-Rao (équation (8)). On cherche donc à calculer

$$E_M \left[eqm_R(\hat{\theta}_i^{EB}) \right].$$

Tout d'abord, pour calculer l'espérance de cette expression, il sera utile de procéder terme à terme, dans le but de simplifier la compréhension des calculs.

On a essentiellement 6 termes aléatoires plus ou moins complexes qui apparaissent dans l'expression de $eqm_R(\hat{\theta}_i^{EB})$, qu'on appellera respectivement T_1, T_2, \dots, T_6 . On peut alors réécrire l'espérance de $eqm_R(\hat{\theta}_i^{EB})$ sous la forme suivante:

$$\begin{aligned} E_M \left[eqm_R(\hat{\theta}_i^{EB}) \right] = & \\ & \sigma_{ii} + 2\sigma_{ii} \{ E_M [T_1] - E_M [T_2] \\ & + E_M [T_3] + E_M [T_4] - E_M [T_5] \} \\ & + E_M [T_6]. \end{aligned} \quad (15)$$

Notons qu'à partir de maintenant, on remplacera, dans les calculs d'espérance, $(n-p)$ par n , étant donné qu'on cherche un résultat asymptotique.

4.1 Calcul de $E_M [T_1]$

Étant donné que le premier terme est d'ordre $1/n$ (à cause de h_{ii}^{**}), on peut tout simplement calculer une approximation de cette espérance en remplaçant $\hat{\sigma}_v^2$ par σ_v^2 , car cet estimateur est non-biaisé pour σ_v^2 .

Donc, l'espérance devient:

$$\begin{aligned} E_M [T_1] &\approx E_M \left[\frac{\sigma_{ii} h_{ii}^*}{(\sigma_v^2 + \sigma_{ii})^2} \right] \\ &\approx \frac{\sigma_{ii} h_{ii}^*}{(\sigma_v^2 + \sigma_{ii})^2} \end{aligned}$$

car plus rien n'est aléatoire.

4.2 Calcul de $E_M [T_2]$

En ce qui concerne le deuxième terme, il est impossible de remplacer $\hat{\sigma}_v^2$ par σ_v^2 car l'expression n'est pas d'ordre $1/n$. On calculera plutôt une approximation de l'espérance en utilisant la série de Taylor de $\frac{\sigma_{ii}}{(\hat{\sigma}_v^2 + \sigma_{ii})}$:

$$\begin{aligned} E_M [T_2] = & \\ E_M \left[\frac{\sigma_{ii}}{(\sigma_v^2 + \sigma_{ii})} \left(1 - \left(\frac{\hat{\sigma}_v^2 - \sigma_v^2}{\sigma_v^2 + \sigma_{ii}} \right) \right. \right. & \\ \left. \left. + \left(\frac{\hat{\sigma}_v^2 - \sigma_v^2}{\sigma_v^2 + \sigma_{ii}} \right)^2 - \dots \right) \right]. & \quad (16) \end{aligned}$$

Pour l'approximation, on abandonne les termes d'ordre 3 et plus et on calcule l'espérance. Étant donné que $\hat{\sigma}_v^2$ est non-biaisé pour σ_v^2 , on obtient:

$$E_M [T_2] \approx \frac{\sigma_{ii}}{(\sigma_v^2 + \sigma_{ii})} \left(1 + \frac{Var(\hat{\sigma}_v^2)}{(\sigma_v^2 + \sigma_{ii})^2} \right). \quad (17)$$

4.3 Calcul de $E_M [T_3]$

Comme T_3 est aussi d'ordre $1/n$, on remplace tout simplement les estimateurs par les vrais paramètres, puis on calcule l'espérance. On a alors que

$$E_M [T_3] \approx \frac{2\sigma_{ii}}{n(\sigma_v^2 + \sigma_{ii})^2} E_M \left[(y_i - \mathbf{x}_i^t \boldsymbol{\beta})^2 \right].$$

D'après le modèle Fay-Herriot, y_1, \dots, y_n sont indépendantes et identiquement distribuées selon la loi $N(\mathbf{x}_i^t \boldsymbol{\beta}, \sigma_v^2 + \sigma_{ii})$. Conséquemment,

$$E_M \left[(y_i - \mathbf{x}_i^t \boldsymbol{\beta})^2 \right] = \sigma_v^2 + \sigma_{ii}.$$

On obtient finalement

$$E_M [T_3] \approx \frac{2\sigma_{ii}}{n(\sigma_v^2 + \sigma_{ii})}. \quad (18)$$

4.4 Calcul de $E_M [T_4]$

Pour le quatrième terme, on peut aussi substituer σ_v^2 à $\hat{\sigma}_v^2$ et $\boldsymbol{\beta}$ à $\hat{\boldsymbol{\beta}}$ et $\hat{\boldsymbol{\beta}}_w$, car T_4 est aussi d'ordre $1/n$:

$$E_M [T_4] \approx K_i^* \sum_{k=1}^n \frac{\mathbf{x}_k \mathbf{x}_k^t}{(\sigma_v^2 + \sigma_{kk})^2} \boldsymbol{\beta} E_M [y_i - \mathbf{x}_i^t \boldsymbol{\beta}]$$

où

$$K_i^* = \frac{2\sigma_{ii}}{n(\sigma_v^2 + \sigma_{ii})} \mathbf{x}_i^t \mathbf{A}^{-1}.$$

Étant donné que $E_M [y_i - \mathbf{x}_i^t \boldsymbol{\beta}] = 0$, ce terme disparaît dans l'espérance finale.

4.5 Calcul de $E_M [T_5]$

En utilisant la même méthode pour le cinquième terme, on obtient

$$E_M [T_5] \approx K_i^* \sum_{k=1}^n \frac{x_k}{(\sigma_v^2 + \sigma_{kk})^2} E_M [y_k (y_i - x_i^t \beta)].$$

$$\text{Or, } E_M [y_k (y_i - x_i^t \beta)] = \begin{cases} 0 & \text{si } k \neq i \\ \sigma_v^2 + \sigma_{ii} & \text{si } k = i \end{cases}$$

donc,

$$E_M [T_5] \approx \frac{2 \sigma_{ii} h_{ii}^*}{n (\sigma_v^2 + \sigma_{ii})^2}. \quad (19)$$

Mais, h_{ii}^* étant d'ordre $1/n$, $E_M [T_5]$ est alors d'ordre $1/n^2$ et on peut l'abandonner dans le calcul de l'approximation de l'espérance finale.

4.6 Calcul de $E_M [T_6]$

Tout ce qui reste à calculer dans l'expression complète de $E_M [eqm_R(\hat{\theta}_i^{EB})]$ est l'espérance du sixième terme, i.e. $E_M [T_6] = E_M [g_i^2(\mathbf{y})]$.

$$E_M [T_6] = E_M \left[\frac{\sigma_{ii}^2}{(\hat{\sigma}_v^2 + \sigma_{ii})^2} (y_i - x_i^t \hat{\beta}_w)^2 \right].$$

On peut calculer une approximation de $\frac{\sigma_{ii}^2}{(\hat{\sigma}_v^2 + \sigma_{ii})^2}$ en conservant uniquement les termes d'ordre 2 et moins dans la série de Taylor et on remplace dans $E_M [T_6]$:

$$E_M [T_6] \approx \frac{\sigma_{ii}^2}{(\sigma_v^2 + \sigma_{ii})^2} \left\{ E_M [(y_i - x_i^t \hat{\beta}_w)^2] - \frac{2}{(\sigma_v^2 + \sigma_{ii})} E_M [(\hat{\sigma}_v^2 - \sigma_v^2)(y_i - x_i^t \hat{\beta}_w)^2] + \frac{3}{(\sigma_v^2 + \sigma_{ii})^2} E [(\hat{\sigma}_v^2 - \sigma_v^2)^2 (y_i - x_i^t \hat{\beta}_w)^2] \right\}.$$

Calculons ces espérances une à une. Tout d'abord, on peut facilement montrer que

$$E_M [(y_i - x_i^t \hat{\beta}_w)^2] = \sigma_v^2 + \sigma_{ii} - h_{ii}^*.$$

On calcule ensuite la deuxième espérance:

$$E_M [(\hat{\sigma}_v^2 - \sigma_v^2)(y_i - x_i^t \hat{\beta}_w)^2].$$

On remplace d'abord $\hat{\sigma}_v^2$ par son approximation:

$$\hat{\sigma}_v^2 \approx \frac{1}{n} \left\{ \sum_{k=1}^n (y_i - x_i^t \beta)^2 - \sum_{k=1}^n \sigma_{ii} \right\} \quad (20)$$

et on substitue ensuite β à $\hat{\beta}_w$. L'espérance devient alors:

$$E_M [((\hat{\sigma}_v^2)^* - \sigma_v^2)(y_i - x_i^t \beta)^2]$$

où $(\hat{\sigma}_v^2)^*$ est l'approximation de $\hat{\sigma}_v^2$ (expression (20)) donnée plus haut.

Finalement, on obtient

$$E_M [(\hat{\sigma}_v^2 - \sigma_v^2)(y_i - x_i^t \hat{\beta}_w)^2] \approx \frac{2}{n} (\sigma_v^2 + \sigma_{ii})^2.$$

En procédant de la même façon pour la dernière espérance, on obtient

$$E_M [(\hat{\sigma}_v^2 - \sigma_v^2)^2 (y_i - x_i^t \hat{\beta}_w)^2] \approx \frac{2}{n^2} (\sigma_v^2 + \sigma_{ii}) \sum_{k=1}^n (\sigma_v^2 + \sigma_{kk})^2.$$

On peut maintenant remplacer ces trois espérances dans $E_M [T_6]$:

$$E_M [T_6] \approx \frac{\sigma_{ii}^2}{(\sigma_v^2 + \sigma_{ii})} - \frac{\sigma_{ii}^2 h_{ii}^*}{(\sigma_v^2 + \sigma_{ii})^2} - \frac{4 \sigma_{ii}^2}{n (\sigma_v^2 + \sigma_{ii})} + \frac{6 \sigma_{ii}^2}{n^2 (\sigma_v^2 + \sigma_{ii})^3} \sum_{k=1}^n (\sigma_v^2 + \sigma_{kk})^2. \quad (21)$$

4.7 Calcul de $E_M [eqm_R(\hat{\theta}_i^{EB})]$

En rassemblant toutes les espérances calculées précédemment, on peut facilement obtenir, en réarrangeant les termes:

$$E_M [eqm_R(\hat{\theta}_i^{EB})] \approx \frac{\sigma_v^2 \sigma_{ii}}{(\sigma_v^2 + \sigma_{ii})} + \frac{\sigma_{ii}^2 h_{ii}^*}{(\sigma_v^2 + \sigma_{ii})^2} - \frac{2 \sigma_{ii}^2 \text{Var}(\hat{\sigma}_v^2)}{(\sigma_v^2 + \sigma_{ii})^3} + \frac{6 \sigma_{ii}^2}{n^2 (\sigma_v^2 + \sigma_{ii})^3} \sum_{k=1}^n (\sigma_v^2 + \sigma_{kk})^2. \quad (22)$$

Comme on a déjà vu que

$$\text{Var}(\hat{\sigma}_v^2) \approx \frac{2}{n^2} \sum_{k=1}^n (\sigma_v^2 + \sigma_{kk})^2, \quad (23)$$

en remplaçant (23) dans (22), on obtient l'expression suivante:

$$E_M [eqm_R(\hat{\theta}_i^{EB})] \approx \frac{\sigma_v^2 \sigma_{ii}}{(\sigma_v^2 + \sigma_{ii})} + \frac{\sigma_{ii}^2 h_{ii}^*}{(\sigma_v^2 + \sigma_{ii})^2} + \frac{\sigma_{ii}^2 \text{Var}(\hat{\sigma}_v^2)}{(\sigma_v^2 + \sigma_{ii})^3}$$

qui est en tout point identique à l'expression (6). On vient donc de prouver le résultat (3.1):

$$E_M [eqm_R(\hat{\theta}_i^{EB})] \approx E_M [eqm_{PR}(\hat{\theta}_i^{EB})].$$

5 Conclusion

Dériver un estimateur de l'EQM par l'approche conditionnelle offre plusieurs avantages. Entre autres, l'intérêt de cette méthode est de fournir une dérivation directe de l'EQM de Prasad-Rao, dont la preuve a été redémontrée par Singh, Stukel et Pfeffermann (1998). De plus, l'estimateur conditionnel permet d'envisager des généralisations de l'estimateur de Prasad-Rao à des cas où le calage aux marges est utilisé.

Références

- [1] DICK, P. (1995). Modelling net undercoverage in the 1991 Canadian census. *Survey Methodology*, **21**, 45-54.
- [2] FAY, R. E. ET HERRIOT, R. A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association*, **74**, 269-277.
- [3] GHOSH, M. ET RAO, J. N. K. (1994). Small area estimation: an appraisal (with discussion). *Statistical Science*, **9**, 55-93.
- [4] PRASAD, N. G. N. ET RAO, J. N. K. (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association*, **85**, 163-171.
- [5] RIVEST, L.P. (1997). An estimator for the mean squared error of small area estimates. Prépublication, Département de mathématiques et statistique, Université Laval.
- [6] SINGH, A.C., STUKEL, D.M. ET PFEFFERMANN, D. (1998). Bayesian versus frequentist measures of error in small area estimation. *Journal. Royal Statistical Society. Series B*, **60**, 377-396.