

LI ZHU

APPLICATION DES MODÈLES DE RÉGRESSION SUR DES
DONNÉES D'ENQUÊTE

Essai
présenté
à la Faculté des études supérieures
de l'Université Laval
pour l'obtention
du grade de maître ès sciences (M. Sc.)

Département de mathématiques et de statistique
FACULTÉ DES SCIENCES ET DE GÉNIE
UNIVERSITÉ LAVAL

Décembre 2005

© Li Zhu, 2005

Résumé

Dans ce document, on étudie deux approches pour traiter les données d'enquête avec un plan de sondage à deux degrés dans le cadre de l'analyse de régression. Une approche est basé sur le modèle de régression multiple où un poids d'échantillonnage final est utilisé pour tenir compte du plan de sondage. L'autre approche est fondée sur le modèle de régression multiniveau où les poids d'échantillonnage sont utilisés à deux degrés du plan. Ce document se concentre sur les estimations des paramètres de recensement et les estimations des paramètres d'enquête. Les estimations des variances de ces estimateurs sont aussi un thème abordé. Les procédures sont illustrés en utilisant les données de l'Enquête auprès des jeunes en maison d'accueil.

Avant-propos

Je tiens à remercier en tout premier lieu M. Louis-Paul Rivest, mon directeur de recherche pour l'aide compétente qu'il m'a apportée, pour sa disponibilité et son encouragement à finir mon premier travail de recherche en statistique. J'aimerais aussi le remercier pour son soutien financier qui a été particulièrement important tout au long de mes études.

Je tiens à remercier tous les professeurs du Département de mathématiques et statistique sans oublier mon directeur de recherche qui me donnent un opportunité de me réorienter en statistique et qui m'aident à réussir.

Je tiens encore à remercier mes amis et collègues au Département de mathématiques et statistique. J'apprécie beaucoup leur aide au cours de mes études en statistique.

Finalement, j'exprime mes profonds remerciements à ma famille pour leurs soutien et leurs compréhension sur la réorientation de mes études. Je les remercie vivement.

Table des matières

Résumé	1
Avant-Propos	4
1 Introduction	9
2 Modèle de régression multiple et application du modèle sur des données d'enquête	11
2.1 Notations et hypothèses	11
2.2 Estimation des paramètres du modèle	12
2.3 Estimation des paramètres-recensement	13
2.4 Estimation de la variance de l'estimateur	15
3 Modèle de régression linéaire multiniveau	17
3.1 Notations et hypothèses	17
3.2 Estimation des paramètres du modèle	19
3.2.1 Calcul de la fonction de vraisemblance et du maximum de la fonction	19
3.2.2 Algorithme pour maximiser la fonction de vraisemblance	21

3.3	Mise en œuvre de l'algorithme	22
3.4	Cas particulier	28
3.5	Annexe : calculs détaillés de certaines expressions	34
4	Utilisation du modèle de régression multiniveau sur des données d'enquête complexe	39
4.1	Définition des paramètres-recensement	39
4.2	Estimation des paramètres-recensement	40
4.2.1	Plan du sondage	40
4.2.2	Estimation des paramètres-recensement	40
4.3	Implantation	43
4.4	Estimation de variances des estimateurs	45
4.4.1	Estimation de la variance de \hat{B}	45
4.4.2	Estimation de la variance de $\hat{\Theta}$	48
5	Application	49
5.1	Plan de sondage	49
5.2	Modèles de régression	50
5.3	Résultats	51
6	Conclusion	55
	Bibliographie	56
	Annexe A :Programmation en SAS	57
	Annexe B :Implantation du modèle multiniveau pour des données d'enquête	61

Chapitre 1

Introduction

Certaines études associées à des enquêtes de grande envergure nécessitent l'analyse statistique de populations présentant une structure hiérarchique complexe. Due à la nature de la population, à l'étape de planification d'enquête, les méthodologistes utilisent souvent des plans d'échantillonnage multiniveaux. En outre, à l'étape d'analyse, habituellement, les analystes ne tiennent pas compte de la complexité du plan de sondage et produisent souvent des estimateurs biaisés des paramètres d'intérêt.

En fait, dans le cadre de l'analyse de régression linéaire multiple, une méthode basée sur le plan d'enquête est déjà disponible. Cette méthode est une mise en œuvre de la régression pondérée en utilisant les poids d'échantillonnage. Elle permet d'obtenir des estimateurs des paramètres asymptotiquement non-biaisés, lorsque les unités d'échantillonnage sont indépendantes les unes des autres. Elle a cependant des limites : Si les données ne sont pas indépendantes, elle n'est pas suffisante pour représenter convenablement la structure particulière de la population. On a donc besoin de nouvelles méthodes statistiques pour analyser les données hiérarchiques dans un plan complexe.

Récemment, certaines études se sont concentrées sur les méthodes permettant d'utiliser l'information sur le plan de sondage lors de la modélisation des données hiérarchiques : Graubard et Korn (1996) ont proposé une méthode des moments, Pfefferman et al. (1998) ont proposé une méthode itérative généralisée des moindres carrés pondérées et Kovacevic et Rai (2003) ont proposé une méthode de pseudo-vraisemblance pour l'estimation des variances associées à chaque niveau.

Dans ce document, on essaie de présenter deux approches pour traiter les données d'enquête complexe dans le cadre de l'analyse de régression linéaire. Dans le chapitre 2, la régression multiple avec un plan de sondage sera illustrée. Le chapitre 3 et le chapitre 4 présentent une méthode d'estimation des paramètres basée sur le plan de sondage dans un modèle de régression multiniveau. Le chapitre 3 se penchera sur l'introduction du modèle de régression multiniveau, alors que le chapitre 4 présentera la méthode proposée par Pfefferman et al. (1998) : la méthode itérative généralisée des moindres carrés pondérées, qui permet de faire des inférences basées sur le plan de sondage dans un modèle de régression multiniveau pour les données d'enquête. Finalement, dans le chapitre 5, on donnera un exemple de l'application des méthodes aux données de l'Enquête auprès des jeunes en maison d'accueil réalisée en 1987 aux États-Unis. Le jeu de données de cette enquête est extrait du livre de Lohr (2001).

Chapitre 2

Modèle de régression multiple et application du modèle sur des données d'enquête

Ce chapitre se concentre sur l'estimation des paramètres d'enquête à l'aide d'un modèle de régression multiple. Dans la section 2.1, on introduit le modèle standard de régression multiple. La section 2.2 présente la méthode d'estimation des paramètres du modèle : la méthode des moindres carrés. Enfin, les sections 2.3 et 2.4 portent sur l'estimation du paramètre d'enquête et la variance associée.

2.1 Notations et hypothèses

Si on observe N paires $(y_1, \mathbf{x}_1), \dots, (y_N, \mathbf{x}_N)$ et que, pour $i = 1 \dots N$, y_i et \mathbf{x}_i ont une relation linéaire, le modèle de régression usuel est donné par

$$y_i = \mathbf{x}'_i \beta + \varepsilon_i, \quad (2.1.1)$$

où $\mathbf{x}'_i = (1 \ x_{1i} \ \dots \ x_{(p-1)i})$.

Notons que, dans le modèle (2.1.1),

y_1, \dots, y_N sont N observations de la variable réponse ;

$\mathbf{x}_1, \dots, \mathbf{x}_N$ sont N observations des variables explicatives ;

$\varepsilon_1, \dots, \varepsilon_N$ sont les termes d'erreur ;

β est le vecteur des paramètres à estimer.

Afin de faciliter les présentations des calculs effectuées dans les sections suivantes, on ré-écrit (2.1.1) en utilisant une notion matricielle :

$$Y = X\beta + \varepsilon,$$

où

$$Y = (y_1, \dots, y_N)'$$

$$X = \begin{pmatrix} 1 & x_{11} & \dots & x_{(p-1)1} \\ 1 & x_{12} & \dots & x_{(p-1)2} \\ \vdots & \vdots & & \vdots \\ 1 & x_{1N} & \dots & x_{(p-1)N} \end{pmatrix}$$

$$\beta = (\beta_0, \beta_1, \dots, \beta_{(p-1)})$$

$$\varepsilon = (\varepsilon_1, \dots, \varepsilon_N).$$

On postule trois hypothèses pour ce modèle : ce sont respectivement les linéarité, non-corrélation et normalité. En terme mathématique, $\varepsilon \sim \mathcal{N}(0, \text{diag}(\sigma_i^2))$, où $\sigma_i^2 = \text{Var}(\varepsilon_i)$, $i = 1, \dots, N$. Notons que le modèle nous permet de tenir compte de l'hétéroscédasticité.

2.2 Estimation des paramètres du modèle

Pour le modèle de régression linéaire multiple présenté dans la section précédente, la méthode des moindres carrés pondérées est une approche standard pour estimer

le vecteur des paramètres β . Cette méthode met l'accent sur la minimisation de la distance entre les observations et le modèle de régression pondéré en tenant compte de la propriété de l'hétéroscédaticité du modèle. Autrement dit, la méthode cherche les valeurs de β qui minimisent $(Y - X\beta)'W(Y - X\beta)$, où $W = 1/\text{diag}(\sigma_i^2)$.

En fait, $(Y - X\beta)'W(Y - X\beta)$ est une fonction continue et convexe. La minimisation de cette fonction s'effectue en dérivant $(Y - X\beta)'W(Y - X\beta)$ par rapport à β , et en posant la dérivée égale à zéro :

$$\begin{aligned}\frac{\partial}{\partial \beta}(Y - X\beta)'W(Y - X\beta) &= \frac{\partial}{\partial \beta}(Y'WY - 2\beta'X'WY + \beta'X'WX\beta) \\ &= 0 - 2X'WY + 2X'WX\beta,\end{aligned}$$

$$0 - 2X'WY + 2X'WX\hat{\beta} = 0 .$$

Finalement, on obtient un estimateur pour β :

$$\hat{\beta} = (X'WX)^{-1}X'WY. \quad (2.2.2)$$

Cas Particulier

Supposons que les variance de ε_i sont homogènes, c-à-d, $\sigma_i^2 = \sigma^2, \forall i$. On a alors $W = 1/\sigma^2\mathbf{I}$. L'estimateur de β devient donc

$$\hat{\beta} = (X'X)^{-1}X'Y,$$

une forme classique de $\hat{\beta}$ dans l'étude de régression.

2.3 Estimation des paramètres-recensement

Dans cette section, on présente une méthode d'estimation des paramètres dans un modèle de régression multiple en tenant compte du plan de sondage.

Définition des paramètres-recensement

Tout d'abord, considérons une population finie U avec N unités indépendantes. On reprend (2.1.1) pour décrire la relation entre Y et X dans la population. Puisque toutes les données sont dans la même population, il est raisonnable de penser que les variances des termes d'erreur sont homogènes. Un estimateur du paramètre-recensement pour β est donc donné par

$$B = (X'_U X_U)^{-1} X'_U Y_U,$$

où

X_U = matrice X pour toutes les données de la population,

Y_U = vecteur Y pour toutes les données de la population.

Estimation des paramètres-recensement

Maintenant, on applique un plan de sondage complexe à cette population : On tire un échantillon S avec n unités d'échantillonnage. Si la probabilité d'inclusion pour une telle unité d'échantillonnage est π_i , où $i \in S$, le poids d'échantillonnage est bien l'inverse de cette probabilité $\{w_i | w_i = 1/\pi_i, i \in S\}$.

Pour estimer les paramètres d'enquête, on fait appel à la méthode de moindres carrées pondérées qu'on a présentée dans la section précédente, et on redéfinit la matrice W en se servant du poids d'échantillonnage. Un estimateur du paramètre d'enquête est donc

$$\hat{B} = (X'_S W_S X_S)^{-1} X'_S W_S Y_S,$$

où

X_S est la matrice X pour toutes les données dans l'échantillon, dont la dimension est $n \times p$,

Y_S est le vecteur Y pour toutes les données dans l'échantillon, dont la dimension est $n \times 1$,

W_S est une matrice diagonale $n \times n$ avec $W_S = \text{diag}(w_1, \dots, w_n)$.

L'étude des propriétés de \hat{B} par rapport au plan de sondage est présentée dans Lohr (1999, Chapitre 11).

2.4 Estimation de la variance de l'estimateur

Cette section se penche sur l'estimation de la variance de \hat{B} par la méthode de linéarisation.

Puisque \hat{B} est une statistique non-linéaire, premièrement, on devrait la linéariser pour que l'on puisse calculer sa variance. En linéarisant \hat{B} , on obtient

$$\begin{aligned}\hat{B} &= (X'_S W_S X_S)^{-1} X'_S W_S Y_S \\ &= (X'_S W_S X_S)^{-1} X'_S W_S (Y_S - X_S B + X_S B) \\ &= B + (X'_S W_S X_S)^{-1} X'_S W_S (Y_S - X_S B).\end{aligned}$$

Notons que

- 1) $(X'_S W_S X_S)^{-1}$ converge en probabilité vers $(X'_U X_U)^{-1}$
- 2) $X'_S W_S (Y_S - X_S B)$ converge en loi vers une loi normale dont l'espérance est nulle.

D'après le théorème de Slutsky, on a $(X'_S W_S X_S)^{-1} X'_S W_S (Y_S - X_S B)$ converge vers une loi normale dont l'espérance est 0 et la variance $Var((X'_U X_U)^{-1} X'_S W_S (Y_S - X_S B))$.

Alors, la variance de \hat{B} est donnée par

$$\begin{aligned}Var(\hat{B}) &\approx Var[B + (X'_U X_U)^{-1} X'_S W_S (Y_S - X_S B)] \\ &\approx (X'_U X_U)^{-1} Var[X'_S W_S (Y_S - X_S B)] (X'_U X_U)^{-1}.\end{aligned}$$

En approximant B par \hat{B} et $(X'_U X_U)^{-1}$ par $(X'_S W_S X_S)^{-1}$, on obtient un estimateur de $Var(B)$:

$$\begin{aligned}v(\hat{B}) &\approx (X'_S W_S X_S)^{-1} v(X'_S W_S (Y_S - X_S \hat{B})) (X'_S W_S X_S)^{-1} \\ &\approx \left(\sum_{i \in S} w_i x_i x'_i \right)^{-1} v \left(\sum_{i \in S} w_i \hat{\varepsilon}_i x_i \right) \left(\sum_{i \in S} w_i x_i x'_i \right)^{-1},\end{aligned}$$

où $\hat{\varepsilon}_i = y_i - x_i \hat{B}$.

Dans un plan complexe à deux niveaux (un échantillon avec m unités primaires et n_j unités secondaires dans chaque unité primaire),

$$v\left(\sum_{i \in S} w_i \hat{\varepsilon}_i x_i\right) = \sum_{j=1}^m \frac{n_j - 1}{n_j} \sum_{i=1}^{n_j} (a_{ij} - \bar{a}_{ij})(a_{ij} - \bar{a}_{ij})',$$

avec $a_{ij} = \sum_{i=1}^{n_j} w_{ij} \hat{\varepsilon}_{ij} x_{ij}$ et $\bar{a}_{ij} = \frac{1}{n_j} \sum_{i=1}^{n_j} a_{ij}$.

Chapitre 3

Modèle de régression linéaire multiniveau

À partir de ce chapitre, on s'intéresse à analyser des données d'enquête à l'aide du modèle de régression linéaire multiniveau. Dans un premier temps, il faut donc préciser le modèle de régression multiniveau qu'on utilisera pour notre étude. Dans la section 3.1, on introduit d'abord un modèle de régression à deux niveaux. Puis, dans la section 3.2, on présente la méthode du maximum de vraisemblance, une méthode d'estimation des paramètres du modèle, et on développe un algorithme d'estimation pour cette méthode. La section 3.3 met en œuvre cet algorithme pour notre modèle. Enfin, dans la section 3.4, un cas spécial où on peut obtenir les estimateurs exacts des paramètres sans itération sera présenté.

3.1 Notations et hypothèses

Soit une structure de données hiérarchiques avec M unités au niveau 2 et N_j unités au niveau 1 emboîtées dans la $j^{\text{ième}}$ unité du niveau 2. Supposons que toutes les unités au deuxième niveau sont indépendantes les unes des autres, et que les unités au niveau 1 emboîtées dans une telle unité du niveau 2 sont corrélées.

On peut donc construire un modèle linéaire mixte à deux niveaux à partir de cette structure de données. À ce chapitre, on ne présente qu'un modèle mixte avec l'ordonne à l'origine aléatoire.

Le modèle est défini par :

$$y_{ij} = \mathbf{x}'_{ij}\beta + z_{ij}u_j + z_{0ij}\varepsilon_{ij}, \quad (3.1.1)$$

où $i = 1, \dots, N_j$, $j = 1, \dots, M$, $\mathbf{x}'_{ij} = (1 \ x_{1ij} \ \dots \ x_{(p-1)ij})$.

Les hypothèses du modèle sont les suivantes :

i) $u_j \sim \mathcal{N}(0, \sigma_u^2)$

ii) $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$

iii) $Cov(u_j, \varepsilon_{ij}) = 0 \quad \forall i, j$

iv) $z_{ij} = 1 \quad \forall i, j$

v) $z_{0ij} = 1 \quad \forall i, j$

Sous la forme matricielle :

$$Y = X\beta + Zu + Z_0\epsilon \quad (3.1.2)$$

où

Y , le vecteur de la variable réponse, est un vecteur de $\mathfrak{R}^{\sum_j N_j}$.

X , la matrice des variables explicatives, est une matrice de dimension $\sum_j N_j \times p$.

β , le vecteur des paramètres des effets fixes, est un vecteur de \mathfrak{R}^p .

Z , la matrice des coefficients des effets aléatoires au niveau 2 de la population, est une matrice de dimension $\sum_j N_j \times M$.

u , le vecteur des effets aléatoires au niveau 2 de la population, est un vecteur de \mathfrak{R}^M .

ϵ , le vecteur des termes d'erreurs, est un vecteur de $\mathfrak{R}^{\sum_j N_j}$.

Z_0 , la matrice des coefficients des effets aléatoires qui permet de tenir compte de l'hétéroscédasticité au niveau 1 de la population, est une matrice diagonale de dimension $\sum_j N_j \times \sum_j N_j$.

Ainsi, la matrice de covariance du vecteur des y pour la $j^{\text{ième}}$ grappe est

$$V_j = Z_j \sigma_u^2 Z_j' + \sigma^2 D_j,$$

où

$$Z_j = (z_{1j} \dots z_{N_j j})', \quad (3.1.3)$$

$$D_j = \text{diag}(z_{01j}^2 \dots z_{0N_j j}^2). \quad (3.1.4)$$

Il est également intéressant d'écrire la matrice de covariance

$$V_j(\theta) = \sum_{k=1}^2 \theta_k G_{kj} \quad (3.1.5)$$

comme une combinaison linéaire des matrices symétriques G_{kj} , où $\theta_1 = \sigma_u^2$, $\theta_2 = \sigma^2$,

$$G_{kj} = \delta_k D_j + (1 - \delta_k) Z_j Z_j' \quad \text{et} \quad \delta_k = \begin{cases} 0 & \text{si } k = 1 \\ 1 & \text{si } k = 2 \end{cases}.$$

3.2 Estimation des paramètres du modèle

Plusieurs méthodes ont déjà été développées pour estimer les composantes de la variance et les paramètres dans un modèle mixte. Ce sont la méthode du maximum de vraisemblance restreint (REML), la méthode du maximum de vraisemblance (ML) et la méthode MINQUE0. Parmi eux, la méthode du maximum de vraisemblance restreint (REML) est la méthode qu'on utilise le plus souvent.

Ce document se penche sur la méthode du maximum de vraisemblance. Puisque la méthode du maximum de vraisemblance (ML) est équivalente à la méthode itérative généralisée des moindres carrés (IGLS) dans le cas de la normalité (voir Goldstein (1995)), on la présente sous cette forme.

3.2.1 Calcul de la fonction de vraisemblance et du maximum de la fonction

Vu que les unités primaires de la population présentée dans la section précédente sont indépendantes et que les unités secondaires ne le sont pas, on donne la densité

conjointe des données sous la forme suivante :

$$L(\beta, V_j) = \prod_j \{(2\pi)^{-n/2} |V_j|^{-1/2} \exp[-(Y_j - X_j\beta)' V_j^{-1} (Y_j - X_j\beta)/2]\}.$$

La méthode du maximum de vraisemblance consiste à chercher la valeur de β et V_j qui rend la valeur de la fonction de vraisemblance (L) maximale. Puisque V_j est en fonction de θ , la démarche est équivalente à chercher la valeur de β et θ pour maximiser la fonction L .

Pour commencer, on calcule la fonction de log-vraisemblance :

$$l(\beta, V_j) = \log(L) = \sum_j \left[-\frac{n}{2} \log(2\pi) - \frac{1}{2} \log |V_j| - (Y_j - X_j\beta)' V_j^{-1} (Y_j - X_j\beta)/2 \right].$$

Ensuite, on calcule les dérivées partielles par rapport à tous les paramètres. À cette étape, on peut facilement avoir le vecteur des dérivées partielles par rapport à β :

$$\frac{\partial l}{\partial \beta} = -\sum_j X_j' V_j^{-1} X_j \beta + \sum_j X_j' V_j^{-1} Y_j. \quad (3.2.6)$$

Le calcul des dérivées partielles par rapport à θ fait appel aux résultats suivants (voir Christensen (1987), p. 231). Soit une matrice A en fonction d'un scalaire s , alors

i)

$$\frac{\partial A^{-1}}{\partial s} = -A^{-1} \frac{\partial A}{\partial s} A^{-1},$$

ii)

$$\frac{\partial}{\partial s} \log |A| = \text{tr} \left(A^{-1} \frac{\partial A}{\partial s} \right).$$

Ces résultats permettent d'obtenir les dérivées partielles de la log-vraisemblance par rapport à θ : Avec $k = 1, 2$,

$$\frac{\partial l}{\partial \theta_k} = \sum_j \left[-\frac{1}{2} \text{tr} \left(V_j^{-1} \frac{\partial V_j}{\partial \theta_k} \right) + \frac{1}{2} (Y_j - X_j\beta)' V_j^{-1} \frac{\partial V_j}{\partial \theta_k} V_j^{-1} (Y_j - X_j\beta) \right] \quad (3.2.7)$$

$$\begin{aligned}
&= \sum_j \left[-\frac{1}{2} \text{tr} \left(V_j^{-1} \frac{\partial V_j}{\partial \theta_k} V_j^{-1} V_j \right) + \frac{1}{2} (Y_j - X_j \beta)' V_j^{-1} \frac{\partial V_j}{\partial \theta_k} V_j^{-1} (Y_j - X_j \beta) \right] \\
&= \sum_j \left[-\frac{1}{2} \text{tr} \left(V_j^{-1} \frac{\partial V_j}{\partial \theta_k} V_j^{-1} \sum_{l=1}^2 \theta_l G_{lj} \right) + \frac{1}{2} (Y_j - X_j \beta)' V_j^{-1} \frac{\partial V_j}{\partial \theta_k} V_j^{-1} (Y_j - X_j \beta) \right].
\end{aligned}$$

Afin de trouver les valeurs de β et θ qui maximisent la fonction L , on résout les équations (3.2.6) = 0 et (3.2.7) = 0 .

Malheureusement, on ne peut pas toujours obtenir des formes explicites de β et θ en résolvant ces équations. On adopte ainsi un processus itératif pour faire face à ce problème.

3.2.2 Algorithme pour maximiser la fonction de vraisemblance

L'algorithme pour ce processus se présente en quatre étapes :

- Première étape : initialiser $\theta^{(0)}$ avec des valeurs quelconques ;
- Deuxième étape : sachant $\theta^{(r-1)}$, calculer $\beta^{(r)}$ en résolvant l'équation

$$P^{(r)} \beta^{(r)} = Q^{(r)},$$

où P est une matrice $p \times p$

$$P^{(r)} = \sum_j X_j' V_j^{(r-1)-1} X_j, \quad (3.2.8)$$

Q est un vecteur de \mathfrak{R}^p

$$Q^{(r)} = \sum_j X_j' V_j^{(r-1)-1} Y_j; \quad (3.2.9)$$

- Troisième étape : sachant $\beta^{(r)}$, calculer $\theta^{(r)}$ en résolvant l'équation

$$R^{(r)} \theta^{(r)} = S^{(r)},$$

où $R^{(r)}$ est une matrice 2×2 dont l'élément (k, l) est donné par

$$\sum_j \text{tr} \left(V_j^{(r-1)-1} G_{kj} V_j^{(r-1)-1} G_{lj} \right), \quad (3.2.10)$$

$S^{(r)}$ est un vecteur de \Re^2 dont le $k^{ième}$ élément est donné par

$$\sum_j tr\left((Y_j - X_j\beta^{(r)})'V_j^{(r-1)-1}G_{kj}V_j^{(r-1)-1}(Y_j - X_j\beta^{(r)})\right); \quad (3.2.11)$$

- Quatrième étape : répéter la deuxième étape et la troisième étape jusqu'à convergence, c-à-d, la différence entre $\beta^{(r-1)}$ et $\beta^{(r)}$ et celle entre $\theta^{(r-1)}$ et $\theta^{(r)}$ sont très petites.

3.3 Mise en œuvre de l'algorithme

Dans cette section, on fait des efforts pour préciser $P^{(r)}$, $Q^{(r)}$, $R^{(r)}$ et $S^{(r)}$ dans notre modèle. Ainsi, on essaie de tous les exprimer sous forme de totaux. Le travail de cette section est long, mais les résultats qu'on attend sont moins volumineux.

L'avantage de cette mise en œuvre tient du fait qu'elle élabore un lien entre la méthode itérative généralisée des moindres carrés (IGLS) et la méthode itérative généralisée des moindres carrés pondérées (PWIGLS) qu'on présentera au chapitre suivant. Plus tard, dans la section 4.1, on verra comment les résultats obtenus dans cette section peuvent faciliter l'application du modèle à des données d'enquête recueillées avec un plan complexe.

Pour simplifier la présentation dans cette section, on note $V_j^{(r-1)}$ par V_j , $D_j^{(r-1)}$ par D_j , $Z_j^{(r-1)}$ par Z_j , $\theta_1^{(r-1)}$ par θ_1 et $\theta_2^{(r-1)}$ par θ_2 .

Récrire $P^{(r)}$, $Q^{(r)}$, $R^{(r)}$ et $S^{(r)}$ sous formes de totaux

Le processus comporte deux étapes principales, dont la première est de remplacer V_j^{-1} par sa forme matricielle dans (3.2.8), (3.2.9), (3.2.10) et (3.2.11).

Pour démarrer, on calcule l'inverse de la matrice V_j . Une forme matricielle de V_j^{-1} est donnée par

$$V_j^{-1} = \frac{1}{\theta_2} \left(D_j^{-1} - D_j^{-1} Z_j A_j Z_j' D_j^{-1} \right), \quad (3.3.12)$$

où

$$A_j = \left(Z_j' D_j^{-1} Z_j + \frac{\theta_2}{\theta_1} \right)^{-1}.$$

Ensuite, on remplace V_j^{-1} par sa forme matricielle déduite de l'équation (3.3.12) dans les équations (3.2.8) et (3.2.9), et on trouve

$$P^{(r)} = \frac{1}{\theta_2} \sum_j (X_j' D_j^{-1} X_j - X_j' D_j^{-1} Z_j A_j Z_j' D_j^{-1} X_j), \quad (3.3.13)$$

$$Q^{(r)} = \frac{1}{\theta_2} \sum_j (X_j' D_j^{-1} Y_j - X_j' D_j^{-1} Z_j A_j Z_j' D_j^{-1} Y_j). \quad (3.3.14)$$

Quant à l'élément (k, l) de la matrice R et l'élément $[k]$ du vecteur S , non seulement V_j^{-1} est remplacé par sa forme explicite, mais G_{kj} et G_{lj} sont aussi remplacés par leurs formes matricielles montrées dans (3.1.5). D'ailleurs, beaucoup d'efforts sont faits pour réduire les formes de résultats. Finalement, les expressions deviennent

$$\begin{aligned} R^{(r)}[k, l] &= \frac{1}{\theta_2^2} \sum_j [\delta_k \delta_l N_j + \delta_l Z_j' D_j^{-1} Z_j C_{kj} + \delta_k (1 - \delta_l) Z_j' D_j^{-1} Z_j \\ &\quad + (1 - \delta_l) (Z_j' D_j^{-1} Z_j)^2 C_{kj}] \end{aligned} \quad (3.3.15)$$

$$S^{(r)}[k] = \frac{1}{\theta_2^2} \sum_j (\delta_k E_j' D_j^{-1} E_j + E_j' D_j^{-1} Z_j C_{kj} Z_j' D_j^{-1} E_j), \quad (3.3.16)$$

où

$$\begin{aligned} C_{kj} &= B_{kj} - \delta_k A_j - B_{kj} Z_j' D_j^{-1} Z_j A_j \\ B_{kj} &= \frac{\theta_2}{\theta_1} (1 - \delta_k) A_j - \delta_k A_j \\ E_j &= Y_j - X_j \beta. \end{aligned}$$

(Les détails des calculs sont présentés à la section 3.5.)

En second temps, on se concentre sur l'expression des formes matricielles à l'aide de totaux. C'est le moment de nous rappeler les équations (3.1.3) et (3.1.4). Ce sont ces deux égalités qu'on utilisera tout au long des calculs suivants dans cette section pour préciser le vecteur Z_j et la matrice D_j de notre modèle.

En fait, pour obtenir des formes de totaux pour la matrice $P^{(r)}$ et le vecteur $Q^{(r)}$ à partir de (3.1.3), (3.1.4), (3.3.13) et (3.3.14), ce n'est qu'une simple transformation.

On a éventuellement

Matrice $P^{(r)}$

$$P^{(r)} = \frac{1}{\theta_2} \sum_j (T_{1j} - a_j T_{2j} T'_{2j}), \quad (3.3.17)$$

où

$$T_{1j} = \sum_i \frac{\mathbf{x}_{ij} \mathbf{x}'_{ij}}{z_{0ij}^2}$$

$$T_{2j} = \sum_i \frac{\mathbf{x}_{ij} z_{ij}}{z_{0ij}^2}$$

$$a_j = \left(T_{5j} + \frac{\theta_2}{\theta_1} \right)^{-1}$$

$$T_{5j} = \sum_i \frac{z_{ij}^2}{z_{0ij}^2}.$$

Vecteur $Q^{(r)}$

$$Q^{(r)} = \frac{1}{\theta_2} \sum_j (T_{3j} - a_j T_{2j} T_{4j}), \quad (3.3.18)$$

où

$$T_{3j} = \sum_i \frac{\mathbf{x}_{ij} y_{ij}}{z_{0ij}^2}$$

$$T_{4j} = \sum_i \frac{y_{ij} z_{ij}}{z_{0ij}^2}.$$

On remarque que $T_{1j}, T_{2j}, T_{3j}, T_{4j}$ et T_{5j} sont tous les totaux au niveau 1 des données, et que $P^{(r)}$ et $Q^{(r)}$ sont tous sous des formes de totaux au niveau 2.

Les calculs sur $R^{(r)}$ et $S^{(r)}$ sont beaucoup plus complexes et ils se font élément par élément.

Matrice $R^{(r)}$

1) $R^{(r)}(1, 1)$

Avec $k = 1$ et $l = 1$, on a $\delta_k = 0$ et $\delta_l = 0$. Par conséquence, les trois premiers termes de (3.3.15) sont nuls. On a alors

$$\begin{aligned}
 R^{(r)}(1, 1) &= \frac{1}{\theta_2^2} \sum_j [0 + 0 + 0 + (Z'_j D_j^{-1} Z_j)^2 C_{1j}] \\
 &= \frac{1}{\theta_2^2} \sum_j \left(\sum_i \left(\frac{z_{ij}^2}{z_{0ij}^2} \right)^2 C_{1j} \right) \\
 &= \frac{1}{\theta_2^2} \sum_j \left(\sum_i \frac{z_{ij}^2}{z_{0ij}^2} \right)^2 \left(\frac{\theta_2}{\sum_i (z_{ij}^2 / z_{0ij}^2) \theta_1 + \theta_2} \right)^2 \\
 &= \sum_j \left(\frac{1}{(\sum_i z_{ij}^2 / z_{0ij}^2)^{-1} \theta_2 + \theta_1} \right)^2 \\
 &= \sum_j b_j^2,
 \end{aligned}$$

où

$$b_j = \left(\theta_1 + \frac{\theta_2}{T_{5j}} \right)^{-1}.$$

2) $R^{(r)}(1, 2)$

Cette fois-ci, on a $\delta_k = 1$ et $\delta_l = 0$. En utilisant ces égalités dans (3.3.15), on obtient

$$R^{(r)}(1, 2) = \sum_j \frac{1}{\theta_2^2} (0 + Z'_j D_j^{-1} Z_j C_{1j} + 0 + 0)$$

$$\begin{aligned}
 &= \sum_j \left(\sum_i \frac{z_{ij}^2}{z_{0ij}^2} \right) \left(\frac{\theta_2}{\sum_i (z_{ij}^2/z_{0ij}^2)\theta_1 + \theta_2} \right)^2 \\
 &= \sum_j \frac{b_j^2}{T_{5j}} .
 \end{aligned}$$

3) $R^{(r)}(2, 1)$

Pour cet élément, il suffit de faire appel à un théorème de trace : Soit deux matrices carrées A et B . Alors, $tr(AB) = tr(BA)$. (Voir Graybill (1983), p.302)

Dans notre cas, $V_j^{-1}G_{kj}$ est une matrice carée. On a donc $tr(V_j^{-1}G_{2j}V_j^{-1}G_{1j}) = tr(V_j^{-1}G_{1j}V_j^{-1}G_{2j})$, plus précisément $R^{(r)}(2, 1) = R^{(r)}(1, 2)$.

4) $R^{(r)}(2, 2)$

Dans ce cas-ci, $\delta_k = 0$, $\delta_l = 0$ et deux termes de (3.3.15) sont nuls. Il devient

$$\begin{aligned}
 R^{(r)}(2, 2) &= \frac{1}{\theta_2^2} \sum_j (N_j + Z_j' D_j^{-1} Z_j C_{2j} + 0 + 0) \\
 &= \frac{1}{\theta_2^2} \sum_j \left[N_j + \left(\sum_i \frac{z_{ij}^2}{z_{0ij}^2} \right) \left(\sum_i \frac{z_{ij}^2}{z_{0ij}^2} A_j^2 - 2A_j \right) \right] \\
 &= \sum_j \left[\frac{1}{\theta_2^2} (N_j - 1) + \frac{1}{\theta_2^2} \left(\sum_i \frac{z_{ij}^2}{z_{0ij}^2} A_j - 1 \right) \right] \\
 &= \sum_j \left[\frac{1}{\theta_2^2} (N_j - 1) + \frac{1}{\theta_2^2} \left(\frac{1}{(\sum_i z_{ij}^2/z_{0ij}^2)^{-1}\theta_2 + \theta_1} \right)^2 \left(\frac{1}{\sum_i z_{ij}^2/z_{0ij}^2} \right)^2 \right] \\
 &= \sum_j \left[\frac{1}{\theta_2^2} (N_j - 1) + \frac{b_j^2}{(\sum_i z_{ij}^2/z_{0ij}^2)^2} \right] \\
 &= \sum_j \left(\frac{1}{\theta_2^2} (N_j - 1) + \frac{b_j^2}{T_{5j}^2} \right) .
 \end{aligned}$$

En résumé,

$$R^{(r)} = \begin{pmatrix} \sum_j b_j^2 & \sum_j b_j^2/T_{5j} \\ \sum_j b_j^2/T_{5j} & \sum_j \left(\theta_2^{-2} (N_j - 1) + b_j^2/T_{5j}^2 \right) \end{pmatrix} . \quad (3.3.19)$$

Vecteur $S^{(r)}$

D'une même manière, on retravaille sur la forme de $S^{(r)}[k]$ présentée dans (3.3.16) pour l'obtenir sous forme de totaux.

1) $S^{(r)}[1]$

$$\begin{aligned} S^{(r)}[1] &= \frac{1}{\theta_2^2} \sum_j \left(0 + (E'_j D_j^{-1} Z_j)^2 C_{1j} \right) \\ &= \frac{1}{\theta_2^2} \sum_j \left(\left(\sum_i \frac{e_{ij} z_{ij}}{z_{0ij}^2} \right)^2 \left(\frac{\theta_2}{(\sum_i z_{ij}^2 / z_{0ij}^2) \theta_1 + \theta_2} \right)^2 \right) \\ &= \sum_j b_j^2 \mu_j^2, \end{aligned}$$

où

$$\mu_j = \frac{\sum_i e_{ij} z_{ij} / z_{0ij}^2}{T_{5j}}$$

$$e_{ij} = y_{ij} - \mathbf{x}_{ij} \beta. \quad (3.3.20)$$

2) $S^{(r)}[2]$

$$\begin{aligned} S^{(r)}[2] &= \frac{1}{\theta_2^2} [E'_j D_j^{-1} E_j + (E'_j D_j^{-1} Z_j)^2 C_{2j}] \\ &= \frac{1}{\theta_2^2} \sum_j \left[\sum_i \frac{e_{ij}^2}{z_{0ij}^2} + \left(\sum_i \frac{e_{ij} z_{ij}}{z_{0ij}^2} \right)^2 \left(\frac{b_j^2 \theta_2^2 - T_{5j}^2}{T_{5j}^3} \right) \right] \\ &= \frac{1}{\theta_2^2} \sum_j \left[\sum_i \frac{e_{ij}^2}{z_{0ij}^2} + \mu_j^2 \left(-T_{5j} + \frac{b_j^2 \theta_2^2}{T_{5j}} \right) \right] \\ &= \frac{1}{\theta_2^2} \sum_j \left[\sum_i \frac{e_{ij}^2}{z_{0ij}^2} + \mu_j^2 T_{5j} - 2\mu_j^2 T_{5j} + \frac{b_j^2 \mu_j^2 \theta_2^2}{T_{5j}} \right] \\ &= \frac{1}{\theta_2^2} \sum_j \left[\sum_i \frac{e_{ij}^2}{z_{0ij}^2} + \mu_j^2 T_{5j} - 2 \sum_i \frac{e_{ij} z_{ij}}{z_{0ij}^2} \mu_j + \frac{b_j^2 \mu_j^2 \theta_2^2}{T_{5j}} \right] \end{aligned}$$

$$\begin{aligned}
&= \frac{1}{\theta_2^2} \sum_j \left(\sum_i \frac{e_{ij}^2}{z_{0ij}^2} + \sum_i \frac{z_{ij}^2 u_j^2 - 2e_{ij} z_{ij} \mu_j}{z_{0ij}^2} + \frac{b_j^2 \mu_j^2 \theta_2^2}{T_{5j}} \right) \\
&= \sum_j \left[\frac{1}{\theta_2^2} \sum_i \frac{(e_{ij} - z_{ij} \mu_j)^2}{z_{0ij}^2} + \frac{b_j^2 \mu_j^2}{T_{5j}} \right] \\
&= \sum_j \left(\frac{1}{\theta_2^2} \sum_i v_{ij}^2 + \frac{b_j^2 \mu_j^2}{T_{5j}} \right) \\
&= \sum_j \left(\frac{1}{\theta_2^2} T_{6j} + \frac{b_j^2 \mu_j^2}{T_{5j}} \right),
\end{aligned}$$

où

$$T_{6j} = \sum_i v_{ij}^2$$

$$v_{ij} = \frac{(e_{ij} - z_{ij} \mu_j)}{z_{0ij}}.$$

Notons que v_{ij} est le résidu où on enlève les effets aléatoires.

En bref,

$$S^{(r)} = \begin{pmatrix} \sum_j b_j^2 \mu_j^2 \\ \sum_j (\theta_2^{-2} T_{6j} + b_j^2 \mu_j^2 / T_{5j}) \end{pmatrix}. \quad (3.3.21)$$

(Note : Pour les détails des calculs de A_j , B_{kj} et C_{kj} , cf Annexe)

Malgré des longs calculs, on a terminé avec les résultats moins volumineux. Comme on le voit, T_{5j} et T_{6j} sont les totaux au niveau 1 et tous les éléments de $R^{(r)}$ et $S^{(r)}$ s'écrivent sous des formes de totaux au niveau 2.

3.4 Cas particulier

Dans cette section, on traite un cas où les estimateurs du maximum de vraisemblance ont une forme explicite.

Hypothèse

Prenant la même structure de la population définie dans la section 3.1, on postule que le nombre d'unités secondaires est égal pour toutes les unités primaires $N_j = N$. D'ailleurs, le modèle ne contient qu'un vecteur de $x = (1, x_{1ij})'$, avec $\sum_j x_{1ij} = 0$ et $x_{1ij} = x_{1ij'} = x_{1i}$ où $j \neq j'$.

Alors, le modèle devient

$$y_{ij} = \beta_0 + u_j + \beta_1 x_{1i} + \varepsilon_{ij}$$

où $j = 1, \dots, M$, $i = 1, \dots, N$

Calcul des paramètres

Premièrement, on initialise $\tilde{\theta}_1 = 1, \tilde{\theta}_2 = 1$.

Deuxièmement, on calcule $\hat{\beta}$ en résolvant l'équation

$$P\beta = Q . \tag{3.4.22}$$

Prenant $\sum_j x_{1i} = 0$, on a

$$T_{1j} = \sum_i \begin{pmatrix} 1 & x_{1i} \\ x_{1i} & x_{1i}^2 \end{pmatrix} = \begin{pmatrix} N & 0 \\ 0 & \sum_i x_{1i}^2 \end{pmatrix},$$

$$T_{2j} = \sum_i \begin{pmatrix} 1 \\ x_{1i} \end{pmatrix} = \begin{pmatrix} N \\ 0 \end{pmatrix},$$

$$T_{3j} = \sum_i \begin{pmatrix} 1 \\ x_{1i} \end{pmatrix} y_{ij} = \begin{pmatrix} \sum_i y_{ij} \\ \sum_i x_{1i} y_{ij} \end{pmatrix},$$

$$T_{4j} = \sum_i y_{ij},$$

$$a_j = \frac{1}{1 + N}.$$

Selon (3.3.17) et (3.3.18), la matrice P et le vecteur Q deviennent

$$\begin{aligned}
P &= \sum_j (T_{1j} - a_j T_{2j} T'_{2j}) \\
&= \sum_j \left[\begin{pmatrix} N & 0 \\ 0 & \sum_i x_{1i}^2 \end{pmatrix} - \begin{pmatrix} a_j N^2 & 0 \\ 0 & 0 \end{pmatrix} \right] \\
&= \sum_j \begin{pmatrix} N - a_j N^2 & 0 \\ 0 & \sum_i x_{1i}^2 \end{pmatrix} \\
&= \begin{pmatrix} \frac{MN}{1+N} & 0 \\ 0 & \sum_j \sum_i x_{1i}^2 \end{pmatrix};
\end{aligned}$$

$$\begin{aligned}
Q &= \sum_j (T_{3j} - a_j T_{2j} T_{4j}) \\
&= \sum_j \left[\begin{pmatrix} \sum_i y_{ij} \\ \sum_i x_{1i} y_{ij} \end{pmatrix} - \begin{pmatrix} a_j N \sum_i y_{ij} \\ 0 \end{pmatrix} \right] \\
&= \sum_j \begin{pmatrix} \sum_i y_{ij} (1 - a_j N) \\ \sum_i x_{1i} y_{ij} \end{pmatrix} \\
&= \begin{pmatrix} \frac{1}{1+N} \sum_j \sum_i y_{ij} \\ \sum_j \sum_i x_{1i} y_{ij} \end{pmatrix}.
\end{aligned}$$

Ces résultats nous permettent de récrire l'équation (3.4.22) sous la forme suivante :

$$\begin{pmatrix} \frac{MN}{1+N} & 0 \\ 0 & \sum_j \sum_i x_{1i}^2 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} \frac{1}{1+N} \sum_j \sum_i y_{ij} \\ \sum_j \sum_i x_{1i} y_{ij} \end{pmatrix}.$$

Un résultat exact pour $\hat{\beta}$ est donc donné par

$$\begin{aligned}
\hat{\beta}_0 &= \frac{1}{MN} \sum_j \sum_i y_{ij} = \bar{y}_{..} \\
\hat{\beta}_1 &= \frac{1}{M} \frac{\sum_j \sum_i x_{1i} y_{ij}}{\sum_i x_{1i}^2}.
\end{aligned}$$

Notons que, tout au long du calcul, les valeurs de β ne dépendent pas de valeurs de $\tilde{\theta}$. Autrement dit, le vecteur β reste toujours le même quelque soit le nombre d'itérations.

Troisièmement, on calcule $\hat{\theta}$ à partir de l'équation

$$R\theta = S . \quad (3.4.23)$$

Quelques calculs préliminaires nous donnent

$$T_{5j} = N , T_{6j} = \sum_i (e_{ij} - \frac{\sum_i e_{ij}}{N})^2 ,$$

$$b_j = \frac{N}{1+N} , \mu_j = \frac{\sum_i e_{ij}}{N} , v_{ij} = e_{ij} - \frac{\sum_i e_{ij}}{N} ,$$

où $e_{ij} = y_{ij} - \mathbf{x}_{ij}\beta$.

Avant de résoudre l'équation 3.4.23, on simplifie l'élément $S[2]$:

$$\begin{aligned} S[2] &= \sum_j \left(\frac{1}{\theta_2^2} T_{6j} + \frac{b_j^2 \mu_j^2}{T_{5j}} \right) \\ &= \sum_j \left[\sum_i \left(e_{ij} - \frac{\sum_i e_{ij}}{N} \right)^2 + \frac{(\sum_i e_{ij})^2}{N(1+N)^2} \right] \\ &= \sum_j \left[\sum_i e_{ij}^2 - \frac{2}{N} (\sum_i e_{ij})^2 + \frac{(\sum_i e_{ij})^2}{N} + \frac{(\sum_i e_{ij})^2}{N(1+N)^2} \right] \\ &= \sum_j \left[\sum_i e_{ij}^2 - \left(\sum_i e_{ij} \right)^2 \left(\frac{2}{N} - \frac{1}{N} \right) - \frac{1}{N(1+N)^2} \right] \\ &= \sum_j \sum_i e_{ij}^2 - \frac{2+N}{(1+N)^2} \sum_j \left(\sum_i e_{ij} \right)^2 \end{aligned}$$

Selon (3.3.19) et (3.3.21), la matrice R et le vecteur S sont donnés par

$$R = \begin{pmatrix} M \left(\frac{N}{1+N} \right)^2 & M \frac{N}{(1+N)^2} \\ M \frac{N}{(1+N)^2} & M \left[(N-1) + \frac{1}{(1+N)^2} \right] \end{pmatrix}$$

et

$$S = \begin{pmatrix} \frac{1}{(1+N)^2} \sum_j (\sum_i e_{ij})^2 \\ \sum_j \sum_i e_{ij}^2 - \frac{2+N}{(1+N)^2} \sum_j (\sum_i e_{ij})^2 \end{pmatrix}$$

L'équation (3.4.23) devient alors

$$\begin{pmatrix} M\left(\frac{N}{1+N}\right)^2 & M\frac{N}{(1+N)^2} \\ M\frac{N}{(1+N)^2} & M\left[(N-1) + \frac{1}{(1+N)^2}\right] \end{pmatrix} \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} = \begin{pmatrix} \frac{1}{(1+N)^2} \sum_j (\sum_i e_{ij})^2 \\ \sum_j \sum_i e_{ij}^2 - \frac{2+N}{(1+N)^2} \sum_j (\sum_i e_{ij})^2 \end{pmatrix}$$

De ce fait, on obtient deux équations avec deux inconnues.

$$M\left(\frac{N}{1+N}\right)^2 \theta_1 + M\frac{N}{(1+N)^2} \theta_2 = \frac{1}{(1+N)^2} \sum_j (\sum_i e_{ij})^2 \quad (3.4.24)$$

et

$$M\frac{N}{(1+N)^2} \theta_1 + M\left[(N-1) + \frac{1}{(1+N)^2}\right] \theta_2 = \sum_j \sum_i e_{ij}^2 - \frac{2+N}{(1+N)^2} \sum_j (\sum_i e_{ij})^2. \quad (3.4.25)$$

Prenons (3.4.25)*N -(3.4.24). Il nous reste seulement θ_2 dans l'équation :

$$MN(N-1)\theta_2 = N \sum_j \sum_i e_{ij}^2 - \frac{N(2+N)}{(1+N)^2} \sum_j (\sum_i e_{ij})^2 - \frac{1}{(1+N)^2} \sum_j (\sum_i e_{ij})^2.$$

En résolvant cette équation, on obtient enfin

$$\begin{aligned} \hat{\theta}_2 &= \frac{1}{M(N-1)} \left[\sum_j \sum_i e_{ij}^2 - \frac{(2+N)}{(1+N)^2} \sum_j (\sum_i e_{ij})^2 - \frac{1}{N(1+N)^2} \sum_j (\sum_i e_{ij})^2 \right] \\ &= \frac{1}{M(N-1)} \left[\sum_j \sum_i e_{ij}^2 - \frac{1}{N} \sum_j (\sum_i e_{ij})^2 \right] \\ &= \frac{1}{M(N-1)} \left[\sum_j \sum_i \left(e_{ij} - \frac{\sum_i e_{ij}}{N} \right)^2 \right] \\ &= \frac{1}{M(N-1)} \sum_j \sum_i (y_{ij} - \bar{y}_{..} - \beta_1 x_{1i} - \bar{y}_{.j} + \bar{y}_{..} + 0)^2 \\ &= \frac{1}{M(N-1)} \sum_j \sum_i (y_{ij} - \bar{y}_{.j} - \beta_1 x_{1i})^2. \end{aligned}$$

Ensuite, on calcule $\hat{\theta}_1$ à partir de l'équation (3.4.24). En enlevant le facteur commun aux deux cotés de cette équation, on a

$$MN^2\theta_1 + MN\theta_2 = \sum_j (\sum_i e_{ij})^2 .$$

Alors,

$$\begin{aligned} \hat{\theta}_1 + \frac{1}{N}\hat{\theta}_2 &= \frac{1}{MN^2} \sum_j (\sum_i y_{ij} - N\bar{y}_{..} - N\beta_1 x_{1i})^2 \\ &= \frac{1}{MN^2} \sum_j (\sum_i y_{ij} - N\bar{y}_{..})^2 \\ &= \frac{1}{M} \sum_j (\bar{y}_{..} - \bar{y}_{.j})^2 . \end{aligned}$$

Finalement, on remplace θ_2 par sa forme explicite et on obtient

$$\hat{\theta}_1 = \frac{1}{M} \sum_j (\bar{y}_{..} - \bar{y}_{.j})^2 - \frac{1}{MN(N-1)} \sum_j \sum_i (y_{ij} - \bar{y}_{.j} - \beta_1 x_{1i})^2 .$$

3.5 Annexe : calculs détaillés de certaines expressions

I. Les détails des calculs pour (3.3.15) et (3.3.16)

$$V_j^{-1} = \frac{1}{\theta_2} \left(D_j^{-1} - D_j^{-1} Z_j A_j Z_j' D_j^{-1} \right),$$

où $A_j = (Z_j' D_j^{-1} Z_j + \frac{\theta_2}{\theta_1})^{-1}$,

$$\begin{aligned} V_j^{-1} G_{kj} &= \frac{1}{\theta_2} \left(\delta_k I + (1 - \delta_k) D_j^{-1} Z_j Z_j' - (1 - \delta_k) D_j^{-1} Z_j A_j Z_j' D_j^{-1} Z_j Z_j' \right. \\ &\quad \left. - \delta_k D_j^{-1} Z_j A_j Z_j' \right) \\ &= \frac{1}{\theta_2} \left[\delta_k I + D_j^{-1} Z_j \left((1 - \delta_k) (I - A_j Z_j' D_j^{-1} Z_j) - \delta_k A_j \right) Z_j' \right] \\ &= \frac{1}{\theta_2} (\delta_k I + D_j^{-1} Z_j B_{kj} Z_j') \end{aligned}$$

où

$$\begin{aligned} B_{kj} &= (1 - \delta_k) (I - A_j Z_j' D_j^{-1} Z_j) - \delta_k A_j \\ &= (1 - \delta_k) (A_j A_j^{-1} - A_j Z_j' D_j^{-1} Z_j) - \delta_k A_j \\ &= (1 - \delta_k) A_j (A_j^{-1} - Z_j' D_j^{-1} Z_j) - \delta_k A_j \\ &= \frac{\theta_2}{\theta_1} (1 - \delta_k) A_j - \delta_k A_j \end{aligned}$$

$$\begin{aligned} V_j^{-1} G_{kj} V_j^{-1} G_{lj} &= \frac{1}{\theta_2^2} (\delta_k I + D_j^{-1} Z_j B_{kj} Z_j') (\delta_l I + D_j^{-1} Z_j B_{lj} Z_j') \\ &= \frac{1}{\theta_2^2} (\delta_k \delta_l I + \delta_k D_j^{-1} Z_j B_{lj} Z_j' \\ &\quad + \delta_l D_j^{-1} Z_j B_{kj} Z_j' + D_j^{-1} Z_j B_{kj} Z_j' D_j^{-1} Z_j B_{lj} Z_j') \end{aligned}$$

$$\begin{aligned}
 &= \frac{1}{\theta_2^2} [\delta_k \delta_l I + \delta_k D_j^{-1} Z_j (\frac{\theta_2}{\theta_1} (1 - \delta_l) A_j - \delta_l A_j) Z_j' \\
 &\quad + \delta_l D_j^{-1} Z_j B_{kj} Z_j' \\
 &\quad + D_j^{-1} Z_j B_{kj} Z_j' D_j^{-1} Z_j (\frac{\theta_2}{\theta_1} (1 - \delta_l) A_j - \delta_l A_j) Z_j'] \\
 &= \frac{1}{\theta_2^2} [\delta_k \delta_l I + \delta_l (D_j^{-1} Z_j B_{kj} Z_j' - \delta_k D_j^{-1} Z_j A_j Z_j' \\
 &\quad - D_j^{-1} Z_j B_{kj} Z_j' D_j^{-1} Z_j A_j Z_j') + \delta_k (1 - \delta_l) (\frac{\theta_2}{\theta_1} D_j^{-1} Z_j A_j Z_j') \\
 &\quad + (1 - \delta_l) (\frac{\theta_2}{\theta_1} D_j^{-1} Z_j B_{kj} Z_j' D_j^{-1} Z_j A_j Z_j')] \\
 &= \frac{1}{\theta_2^2} [\delta_k \delta_l I + \delta_l D_j^{-1} Z_j (B_{kj} - \delta_k A_j - B_{kj} Z_j' D_j^{-1} Z_j A_j) Z_j' \\
 &\quad + \delta_k (1 - \delta_l) D_j^{-1} Z_j A_j (A_j^{-1} - Z_j' D_j^{-1} Z_j) Z_j' \\
 &\quad + (1 - \delta_l) D_j^{-1} Z_j B_{kj} Z_j' D_j^{-1} Z_j A_j (A_j^{-1} - Z_j' D_j^{-1} Z_j) Z_j'] \\
 &= \frac{1}{\theta_2^2} [\delta_k \delta_l I + \delta_l D_j^{-1} Z_j C_{kj} Z_j' \\
 &\quad + \delta_k (1 - \delta_l) D_j^{-1} Z_j Z_j' + (1 - \delta_l) Z_j' D_j^{-1} Z_j C_{kj} D_j^{-1} Z_j Z_j']
 \end{aligned}$$

où $C_{kj} = B_{kj} - \delta_k A_j - B_{kj} Z_j' D_j^{-1} Z_j A_j$

$$\begin{aligned}
 E_j' V_j^{-1} G_{kj} V_j^{-1} E_j &= \left(E_j' \frac{1}{\theta_2^2} (\delta_k I + D_j^{-1} Z_j B_{kj} Z_j') (D_j^{-1} - D_j^{-1} Z_j A_j Z_j' D_j^{-1}) E_j \right) \\
 &= \frac{1}{\theta_2^2} \left(E_j' (\delta_k D_j^{-1} + D_j^{-1} Z_j B_{kj} Z_j' D_j^{-1} \right. \\
 &\quad \left. - D_j^{-1} Z_j B_{kj} Z_j' D_j^{-1} Z_j A_j Z_j' D_j^{-1} \right. \\
 &\quad \left. - \delta_k D_j^{-1} Z_j A_j Z_j' D_j^{-1}) E_j \right) \\
 &= \frac{1}{\theta_2^2} \left(E_j' (\delta_k D_j^{-1} + D_j^{-1} Z_j C_{kj} Z_j' D_j^{-1}) E_j \right)
 \end{aligned}$$

où $E_j = Y_j - X_j \beta$

Selon (3.2.10) ,

$$\begin{aligned}
 R^{(r)}(k, l) &= \frac{1}{\theta_2^2} \sum_j [\delta_k \delta_l N_j + \delta_l \text{tr}(Z_j' D_j^{-1} Z_j C_{kj}) + \delta_k (1 - \delta_l) \text{tr}(Z_j' D_j^{-1} Z_j) \\
 &\quad + (1 - \delta_l) \text{tr}(Z_j' D_j^{-1} Z_j C_{kj} Z_j' D_j^{-1} Z_j)] \\
 &= \frac{1}{\theta_2^2} \sum_j [\delta_k \delta_l N_j + \delta_l Z_j' D_j^{-1} Z_j C_{kj} + \delta_k (1 - \delta_l) Z_j' D_j^{-1} Z_j \\
 &\quad + (1 - \delta_l) (Z_j' D_j^{-1} Z_j)^2 C_{kj}]
 \end{aligned}$$

Selon (3.2.11) ,

$$\begin{aligned}
 S^{(r)}[k] &= \frac{1}{\theta_2^2} \sum_j \text{tr}(\delta_k E_j' D_j^{-1} E_j + E_j' D_j^{-1} Z_j C_{kj} Z_j' D_j^{-1} E_j) \\
 &= \frac{1}{\theta_2^2} \sum_j (\delta_k E_j' D_j^{-1} E_j + E_j' D_j^{-1} Z_j C_{kj} Z_j' D_j^{-1} E_j)
 \end{aligned}$$

II. Les détails des calculs pour de A_j , B_{kj} et C_{kj}

$$\begin{aligned}
 A_j &= \frac{\theta_1}{\left(\sum_i z_{ij}^2 / z_{0ij}^2 \right) \theta_1 + \theta_2} \\
 B_{kj} &= \frac{\theta_2}{\theta_1} (1 - \delta_k) A_j - \delta_k A_j \\
 C_{kj} &= B_{kj} - \delta_k A_j - B_{kj} Z_j' D_j^{-1} Z_j A_j
 \end{aligned}$$

i) $k = 1$

$$\begin{aligned}
 B_{1j} &= \frac{\theta_2}{\theta_1} \frac{\theta_1}{\left(\sum_i z_{ij}^2 / z_{0ij}^2 \right)^2 + \theta_2} \\
 &= \frac{\theta_2}{\left(\sum_i z_{ij}^2 / z_{0ij}^2 \right) \theta_1 + \theta_2}
 \end{aligned}$$

$$\begin{aligned}
C_{1j} &= B_{1j} - B_{1j} \sum_i \frac{z_{ij}^2}{z_{0ij}^2} \frac{\theta_1}{\left(\sum_i z_{ij}^2 / z_{0ij}^2\right) \theta_1 + \theta_2} \\
&= B_{1j} \frac{\theta_2}{\left(\sum_i z_{ij}^2 / z_{0ij}^2\right) \theta_1 + \theta_2} \\
&= \left(\frac{\theta_2}{\left(\sum_i z_{ij}^2 / z_{0ij}^2\right) \theta_1 + \theta_2} \right)^2
\end{aligned}$$

ii) $k = 2$

$$B_{2j} = -A_j \quad (3.5.26)$$

$$C_{2j} = -2A_j + A_j^2 \sum_i \frac{z_{ij}^2}{z_{0ij}^2} \quad (3.5.27)$$

Chapitre 4

Utilisation du modèle de régression multiniveau sur des données d'enquête complexe

Le modèle mixte est un module spécialisé permettant de traiter des données avec une structure hiérarchique. Souvent, des populations à étudier à l'aide d'une enquête ont cette structure, et cette similarité nous permet d'utiliser le modèle mixte à l'étude de sondage.

Le but de ce chapitre est de présenter une méthode d'estimation des paramètres d'enquête à l'aide d'un modèle linéaire mixte à deux niveaux. Pour atteindre cet objectif, on définit les paramètres-recensement dans la section 4.1 et on présente une méthode d'estimation des paramètres-recensement dans la section 4.2. Ensuite, dans la section 4.3, on introduit une technique pour l'implantation de cette méthode. Finalement, on calcule des estimateurs de variances dans la section 4.4.

4.1 Définition des paramètres-recensement

Les paramètres-recensement sont des paramètres calculés à partir des données de la population. Supposons qu'on a une population et que l'on postule que le modèle de régression multiniveau s'applique à la population complète. La description de la population est la suivante :

Soit une population U avec M grappes. La taille de la grappe j ($j = 1, \dots, M$) est N_j . Supposons que les grappes sont indépendantes les unes des autres, et que les unités dans une grappe sont corrélées. Notons que les grappes et les unités dans les grappes correspondent respectivement aux unités au niveau 2 et aux unités au niveau 1 dans le modèle de la section 3.1.

Quant aux calculs des paramètres-recensement, ce n'est qu'une application directe des résultats du chapitre précédent. En se servant des données de la population, on obtient facilement $P_U^{(r)}$, $Q_U^{(r)}$, $R_U^{(r)}$ et $S_U^{(r)}$ avec les résultats qu'on a eu dans la section 3.3. Ensuite, en suivant l'algorithme présenté dans la section 3.2, on obtient les paramètres-recensement : B et Θ .

4.2 Estimation des paramètres-recensement

4.2.1 Plan du sondage

Prenons un plan de sondage à deux degrés avec lequel on tire un échantillon S en effectuant un tirage aléatoire à deux niveaux de la population U :

1. Sélection de m unités primaires d'échantillonnage parmi M unités au niveau 2 ;
2. Pour chaque unité primaire tirée au niveau 2, tirer un échantillon aléatoire de n_j unités secondaires parmi les N_j unités que contient cette unité primaire.

Soit la probabilité d'inclusion d'une unité primaire donnée est π_j , et que la probabilité conditionnelle d'une unité secondaire ij étant donnée que la unité primaire j est tirée dans l'échantillon est $\pi_{i|j}$. Les poids d'échantillonnage sont données par l'inversion de la probabilité d'inclusion, avec $w_j = 1/\pi_j$ au niveau 2 du plan et $w_{i|j} = 1/\pi_{i|j}$ au niveau 1 du plan.

4.2.2 Estimation des paramètres-recensement

Pour estimer des paramètres-recensement, on travaille sur des méthodes pour incorporer les poids d'échantillonnage. Due à la complexité du plan de sondage (on a

effectivement deux poids d'échantillonnage), la méthode pondérée utilisée au chapitre 2 n'est plus appropriée. On cherche donc d'autres méthodes pour faire face à ce problème.

Dans cette section, on présente une méthode proposée par Pfeffermann et coll. (1998), notamment la méthode itérative généralisée des moindres carrés pondérés (PWIGLS). C'est une méthode fondée sur la méthode du maximum de vraisemblance (ML) et la méthode itérative généralisée des moindres carrés (IGLS). Au début, il est utile de jeter un coup d'œil sur la méthode d'estimation des totaux de population finie.

Soit un échantillon S_0 tirée d'une population U_0 . Si on est intéressé à estimer le total d'une variable y dans cette population, on ne fait appel qu'à la formule suivante :

$$\hat{t}_y = \sum_{i \in S_0} w_i y_i,$$

où \hat{t}_y est l'estimation du total de la variable y dans une population U_0 ; w_i est le poids d'échantillonnage associé au $i^{\text{ième}}$ individu dans l'échantillon S_0 ; y_i est la variable d'intérêt mesurée pour le $i^{\text{ième}}$ individu dans S_0 .

Maintenant, c'est le moment de profiter des résultats calculés dans la section 3.3 pour élaborer la méthode de prévision des totaux et la méthode itérative généralisée des moindres carrés (IGLS).

Avec des données d'enquête, les estimations de T_{1j}, \dots, T_{7j} peuvent être données par

$$\hat{T}_{1j} = \sum_{i \in S} w_{i|j} \frac{\mathbf{x}_{ij} \mathbf{x}'_{ij}}{z_{0ij}^2}$$

$$\hat{T}_{2j} = \sum_{i \in S} w_{i|j} \frac{\mathbf{x}_{ij} z_{ij}}{z_{0ij}^2}$$

$$\hat{T}_{3j} = \sum_{i \in S} w_{i|j} \frac{\mathbf{x}_{ij} y_{ij}}{z_{0ij}^2}$$

$$\hat{T}_{4j} = \sum_{i \in S} w_{i|j} \frac{y_{ij} z_{ij}}{z_{0ij}^2}$$

$$\hat{T}_{5j} = \sum_{i \in S} w_{i|j} \frac{z_{ij}^2}{z_{0ij}^2}$$

$$\hat{T}_{6j} = \sum_{i \in S} w_{i|j} \hat{v}_{ij}^2$$

Notons que T_{1j}, \dots, T_{6j} sont les totaux au niveau 1 du plan. Il est donc propre d'utiliser $w_{i|j}$, le poids au niveau 1 du plan.

Les formes des estimateurs pour a_j, b_j, μ_j et v_{ij} deviennent

$$\hat{a}_j = \left(\hat{T}_{5j} + \frac{\hat{\theta}_2}{\hat{\theta}_1} \right)^{-1}$$

$$\hat{b}_j = \left(\hat{\theta}_1 + \frac{\hat{\theta}_2}{\hat{T}_{5j}} \right)^{-1}$$

$$\hat{\mu}_j = \left(\sum_{i \in S} w_{i|j} e_{ij} z_{ij} / z_{0ij}^2 \right) / \hat{T}_{5j}$$

$$\hat{v}_{ij} = (e_{ij} - z_{ij} \hat{\mu}_j) / z_{0ij}$$

où $\hat{\theta}_1 = \theta_1^{(r-1)}$ et $\hat{\theta}_2 = \theta_2^{(r-1)}$. Et pour l'expression de e_{ij} , on fait appel à (3.3.20).

Finalement, les sommes au niveau 2 du plan font intervenir les poids w_j . Les estimateurs de $P_U^{(r)}, Q_U^{(r)}, R_U^{(r)}$ et $S_U^{(r)}$ peuvent donc s'écrire sous les formes

$$\hat{P}_S^{(r)} = \frac{1}{\hat{\theta}_2} \sum_j w_j (\hat{T}_{1j} - \hat{a}_j \hat{T}_{2j} \hat{T}'_{2j}) \quad (4.2.1)$$

$$\hat{Q}_S^{(r)} = \frac{1}{\hat{\theta}_2} \sum_j w_j (\hat{T}_{3j} - \hat{a}_j \hat{T}_{2j} \hat{T}_{4j}) \quad (4.2.2)$$

$$\hat{R}_S^{(r)} = \begin{pmatrix} \sum_j w_j \hat{b}_j^2 & \sum_j w_j \hat{b}_j^2 / \hat{T}_{5j} \\ \sum_j w_j \hat{b}_j^2 / \hat{T}_{5j} & \sum_j w_j (\hat{\theta}_2^{-2} (\hat{N}_j - 1) + \hat{b}_j^2 / \hat{T}_{5j}^2) \end{pmatrix} \quad (4.2.3)$$

$$\hat{S}^{(r)} = \begin{pmatrix} \sum_j w_j \hat{b}_j^2 \hat{\mu}_j^2 \\ \sum_j w_j (\hat{\theta}_2^{-2} \hat{T}_{6j} + \hat{b}_j^2 \hat{\mu}_j^2 / \hat{T}_{5j}) \end{pmatrix}. \quad (4.2.4)$$

À l'aide de la procédure itérative de la section 3.2, on peut éventuellement obtenir des estimateurs pour \hat{B} et $\hat{\Theta}$.

4.3 Implantation

La réutilisation est au cœur de l'implantation des méthodes de pondération. Vu que la méthode itérative généralisée des moindres carrés pondérées est une extension de la méthode itérative généralisée des moindres carrés dans un contexte du sondage, une réutilisation de l'implantation de la première méthode dans le deuxième cas est une approche attrayante.

L'implantation se déroule en deux étapes :

Étape 1 : Remplacer z_{ij} par $w_j^{-1/2} z_{ij}$ et z_{0ij} par $w_j^{-1/2} w_{ij}^{-1/2} z_{0ij}$.

À cette étape, les totaux deviennent

$$\begin{aligned} \tilde{T}_{1j} &= \sum_{i \in S} w_j w_{ij} \frac{\mathbf{x}_{ij} \mathbf{x}'_{ij}}{z_{0ij}^2} = w_j \hat{T}_{1j} \\ \tilde{T}_{2j} &= \sum_{i \in S} w_j^{\frac{1}{2}} w_{ij} \frac{\mathbf{x}_{ij} z_{ij}}{z_{0ij}^2} = w_j^{\frac{1}{2}} \hat{T}_{2j} \end{aligned} \quad (4.3.5)$$

$$\begin{aligned}\tilde{T}_{3j} &= \sum_{i \in S} w_j w_{ij} \frac{\mathbf{x}_{ij} y_{ij}}{z_{0ij}^2} = w_j \hat{T}_{3j} \\ \tilde{T}_{4j} &= \sum_{i \in S} w_j^{\frac{1}{2}} w_{ij} \frac{y_{ij} z_{ij}}{z_{0ij}^2} = w_j^{\frac{1}{2}} \hat{T}_{4j} \\ \tilde{T}_{5j} &= \sum_{i \in S} w_{ij} \frac{z_{ij}^2}{z_{0ij}^2} = \hat{T}_{5j} \\ \tilde{T}_{6j} &= \sum_{i \in S} \tilde{v}_{ij}^2 = w_j \hat{T}_{6j}\end{aligned}$$

et

$$\tilde{a}_j = (\tilde{T}_{5j} + \frac{\hat{\theta}_2}{\hat{\theta}_1})^{-1} = \hat{a} \quad (4.3.6)$$

$$\tilde{b}_j = (\hat{\theta}_1 + \frac{\hat{\theta}_2}{\tilde{T}_{5j}})^{-1} = \hat{b}$$

$$\tilde{\mu}_j = \sum_{i \in S} (w_j^{\frac{1}{2}} w_{ij} \frac{e_{ij} z_{ij}}{z_{0ij}^2}) / \tilde{T}_{5j} = w_j^{\frac{1}{2}} \hat{\mu}$$

$$\tilde{v}_{ij} = (e_{ij} - w_j^{-\frac{1}{2}} z_{ij} \tilde{\mu}_j) / (w_j^{-\frac{1}{2}} w_{ij}^{-\frac{1}{2}} z_{0ij}) = w_j^{\frac{1}{2}} w_{ij}^{\frac{1}{2}} \hat{v}_{ij} ,$$

où $\hat{\theta}_1 = \theta_1^{(r-1)}$ et $\hat{\theta}_2 = \theta_2^{(r-1)}$.

Afin de comparer les résultats obtenus à cette étape avec ceux de la méthode itérative généralisée des moindres carrés pondérées, on calcule les valeurs de $\tilde{P}^{(r)}$, $\tilde{Q}^{(r)}$, $\tilde{R}^{(r)}$ et $\tilde{S}^{(r)}$ à l'aide de (3.3.17), (3.3.18), (3.3.19) et (3.3.21) avec les changements effectués à l'étape 1 et on les présente sous forme de \hat{T}_j .

$$\tilde{P}^{(r)} = \frac{1}{\hat{\theta}_2} \sum_j (\tilde{T}_{1j} - \tilde{a}_j \tilde{T}_{2j} \tilde{T}'_{2j}) = \frac{1}{\hat{\theta}_2} \sum_j w_j (\hat{T}_{1j} - \hat{a}_j \hat{T}_{2j} \hat{T}'_{2j})$$

$$\tilde{Q}^{(r)} = \frac{1}{\hat{\theta}_2} \sum_j (\tilde{T}_{3j} - \tilde{a}_j \tilde{T}_{2j} \tilde{T}_{4j}) = \frac{1}{\hat{\theta}_2} \sum_j w_j (\hat{T}_{3j} - \hat{a}_j \hat{T}_{2j} \hat{T}_{4j})$$

$$\begin{aligned}\tilde{R}^{(r)} &= \begin{pmatrix} \sum_j \tilde{b}_j^2 & \sum_j \tilde{b}_j^2 / \tilde{T}_{5j} \\ \sum_j \tilde{b}_j^2 / \tilde{T}_{5j} & \sum_j \left(\hat{\theta}_2^{-2} (n_j - 1) + \tilde{b}_j^2 / \tilde{T}_{5j}^2 \right) \end{pmatrix} \\ &= \begin{pmatrix} \sum_j \hat{b}_j^2 & \sum_j \hat{b}_j^2 / \hat{T}_{5j} \\ \sum_j \hat{b}_j^2 / \hat{T}_{5j} & \sum_j \left(\hat{\theta}_2^{-2} (n_j - 1) + \hat{b}_j^2 / \hat{T}_{5j}^2 \right) \end{pmatrix}\end{aligned}$$

$$\begin{aligned}\tilde{S}^{(r)} &= \begin{pmatrix} \sum_j \tilde{b}_j^2 \tilde{\mu}_j^2 \\ \sum_j (\hat{\theta}_2^{-2} \tilde{T}_{6j} + \tilde{b}_j^2 \tilde{\mu}_j^2 / \tilde{T}_{5j}) \end{pmatrix} \\ &= \begin{pmatrix} \sum_j w_j \hat{b}_j^2 \hat{\mu}_j^2 \\ \sum_j w_j (\hat{\theta}_2^{-2} \hat{T}_{6j} + \hat{b}_j^2 \hat{\mu}_j^2 / \hat{T}_{5j}) \end{pmatrix}.\end{aligned}$$

Notons que la matrice $\tilde{P}^{(r)}$ et les vecteurs $\tilde{Q}^{(r)}$ et $\tilde{S}^{(r)}$ sont identiques à ceux qu'on a obtenus dans la section 4.1. Par contre, la matrice $\tilde{R}^{(r)}$ obtenue à cette étape n'est pas identique à celle qu'on a obtenue dans (4.2.3).

Étape 2 : Modifier la matrice $\tilde{R}^{(r)}$

- multiplier les termes sommés dans les quatre sommes en j par w_j ,
- pour $\tilde{R}^{(r)}(2, 2)$, remplacer n_j par $\hat{N}_j = \sum_{i \in s} w_{ij}$.

4.4 Estimation de variances des estimateurs

Dans cette section, on travaille sur les estimations des variances des estimateurs \hat{B} et $\hat{\Theta}$ en se servant de la méthode de linéarisation.

4.4.1 Estimation de la variance de \hat{B}

En premier lieu, on linéarise la statistique \hat{B} . Prenons $\hat{P}_S = \lim_{r \rightarrow \infty} \hat{P}_S^{(r)}$ et $\hat{Q}_S = \lim_{r \rightarrow \infty} \hat{Q}_S^{(r)}$:

$$\begin{aligned}\hat{B} &= \hat{P}_S^{-1} \hat{Q}_S \\ &= \hat{P}_S^{-1} \left[\frac{1}{\hat{\theta}_2} \sum_j w_j (\hat{T}_{3j} - \hat{a}_j \hat{T}_{2j} \hat{T}_{4j}) \right] \\ &= \hat{P}_S^{-1} \left\{ \frac{1}{\hat{\theta}_2} \sum_j w_j \left[\sum_{i \in S} w_{ij} \frac{\mathbf{x}_{ij}(y_{ij} - \mathbf{x}_{ij}'B + \mathbf{x}_{ij}'B)}{z_{0ij}^2} \right] \right\}\end{aligned}$$

$$\begin{aligned}
 & -\hat{a}_j \hat{T}_{2j} \sum_{i \in S} w_{ij} \frac{(y_{ij} - \mathbf{x}_{ij}'B + \mathbf{x}_{ij}'B)z_{ij}}{z_{0ij}^2} \Big] \Big\} \\
 = & \hat{P}_S^{-1} \left\{ \frac{1}{\hat{\theta}_2} \sum_j w_j \left[\sum_{i \in S} w_{ij} \frac{\mathbf{x}_{ij}(y_{ij} - \mathbf{x}_{ij}'B)}{z_{0ij}^2} \right. \right. \\
 & \left. \left. - \hat{a}_j \hat{T}_{2j} \sum_{i \in S} w_{ij} \frac{(y_{ij} - \mathbf{x}_{ij}'B)z_{ij}}{z_{0ij}^2} \right] \right\} + \hat{P}^{-1} \hat{P}B \\
 = & B + \hat{P}_S^{-1} \left\{ \frac{1}{\hat{\theta}_2} \sum_j w_j \left[\sum_{i \in S} w_{ij} \frac{\mathbf{x}_{ij}(y_{ij} - \mathbf{x}_{ij}'B)}{z_{0ij}^2} \right. \right. \\
 & \left. \left. - \hat{a}_j \hat{T}_{2j} \sum_{i \in S} w_{ij} \frac{(y_{ij} - \mathbf{x}_{ij}'B)z_{ij}}{z_{0ij}^2} \right] \right\}.
 \end{aligned}$$

Notons que

- 1) \hat{P}_S converge en probabilité vers \hat{P}_U , où $\hat{P}_U = \lim_{r \rightarrow \infty} \hat{P}_U^{(r)}$;
- 2) Étant donnée $\hat{\Theta}$,

$$\frac{1}{\hat{\theta}_2} \sum_j w_j \left[\sum_{i \in S} w_{ij} \frac{\mathbf{x}_{ij}(y_{ij} - \mathbf{x}_{ij}'B)}{z_{0ij}^2} - \hat{a}_j \hat{T}_{2j} \sum_{i \in S} w_{ij} \frac{(y_{ij} - \mathbf{x}_{ij}'B)z_{ij}}{z_{0ij}^2} \right]$$

converge vers une loi normale quand m est grand.

Selon le théorème de Slutsky, étant donnée $\hat{\Theta}$,

$$\hat{P}_S^{-1} \left\{ \frac{1}{\hat{\theta}_2} \sum_j w_j \left[\sum_{i \in S} w_{ij} \frac{\mathbf{x}_{ij}(y_{ij} - \mathbf{x}_{ij}'B)}{z_{0ij}^2} - \hat{a}_j \hat{T}_{2j} \sum_{i \in S} w_{ij} \frac{(y_{ij} - \mathbf{x}_{ij}'B)z_{ij}}{z_{0ij}^2} \right] \right\}$$

converge vers une loi normale tant que m est grand (voir Pfefferman et coll.(1998)).

Alors, la variance de \hat{B} est donnée par

$$\begin{aligned}
 Var(\hat{B}) = & Var \left\{ \hat{P}_U^{-1} \frac{1}{\hat{\theta}_2} \sum_j w_j \left[\sum_{i \in S} w_{ij} \frac{\mathbf{x}_{ij}(y_{ij} - \mathbf{x}_{ij}'B)}{z_{0ij}^2} \right. \right. \\
 & \left. \left. - \hat{a}_j \hat{T}_{2j} \sum_{i \in S} w_{ij} \frac{(y_{ij} - \mathbf{x}_{ij}'B)z_{ij}}{z_{0ij}^2} \right] \right\}
 \end{aligned}$$

En approximant B par \hat{B} et \hat{P}_U par \hat{P}_S , on obtient un estimateur de cette variance :

$$\begin{aligned} v(\hat{B}) &\approx \frac{1}{\hat{\theta}_2^2} \hat{P}_S^{-1} v \left[\sum_j w_j \left(\sum_{i \in S} w_{i|j} \frac{\mathbf{x}_{ij} e_{ij}}{z_{0ij}^2} - \hat{a}_j \hat{T}_{2j} \sum_{i \in S} w_{i|j} \frac{e_{ij} z_{ij}}{z_{0ij}^2} \right) \right] \hat{P}_S^{-1} \\ &= \frac{1}{\hat{\theta}_2^2} \hat{P}_S^{-1} \frac{m}{m-1} \left(\sum_{j \in S} w_j^2 c_j c_j' \right) \hat{P}_S^{-1}, \end{aligned}$$

où

$$\begin{aligned} c_j &= \hat{T}_{7j} - \hat{a}_j \hat{T}_{2j} \hat{T}_{8j}, \\ \hat{T}_{7j} &= \sum_{i \in S} w_{i|j} \frac{\mathbf{x}_{ij} e_{ij}}{z_{0ij}^2}, \\ \hat{T}_{8j} &= \sum_{i \in S} w_{i|j} \frac{e_{ij} z_{ij}}{z_{0ij}^2}, \\ e_{ij} &= y_{ij} - x_{ij}' \hat{B}. \end{aligned}$$

En suivant la même implantation introduite dans la section 4.3, on obtient

$$\begin{aligned} \tilde{T}_{7j} &= \sum_{i \in S} w_j w_{i|j} \frac{\mathbf{x}_{ij} e_{ij}}{z_{0ij}^2} = w_j \hat{T}_{7j}, \\ \tilde{T}_{8j} &= \sum_{i \in S} w_j^{\frac{1}{2}} w_{i|j} \frac{e_{ij} z_{ij}}{z_{0ij}^2} = w_j^{\frac{1}{2}} \hat{T}_{8j}, \end{aligned}$$

$$\tilde{c}_j = \tilde{T}_{7j} - \tilde{a}_j \tilde{T}_{2j} \tilde{T}_{8j} = w_j [\hat{T}_{7j} - \hat{a}_j \hat{T}_{2j} \hat{T}_{8j}].$$

Les calculs de \tilde{a}_j et \tilde{T}_{2j} font références aux équations (4.3.5) et (4.3.6).

Finalement, on évalue $\tilde{v}(\hat{B})$ et on la compare à avec $\hat{v}(\hat{B})$. On obtient

$$\tilde{v}(\hat{B}) = \frac{1}{\hat{\theta}_2^2} \tilde{P}^{-1} \frac{m}{m-1} \left(\sum_{j \in S} \tilde{c}_j \tilde{c}_j' \right) \tilde{P}^{-1} = \frac{1}{\hat{\theta}_2^2} \hat{P}_S^{-1} \frac{m}{m-1} \left(\sum_{j \in S} w_j^2 c_j c_j' \right) \hat{P}_S^{-1} = \hat{v}(\hat{B}).$$

Notons que $\tilde{v}(\hat{B})$ obtenue est identique à $\hat{v}(\hat{B})$.

4.4.2 Estimation de la variance de $\hat{\Theta}$

En suivant la même démarche, premièrement, on exprime $\hat{\Theta}$ par une fonction linéaire :

$$\begin{aligned}\hat{\Theta} &= \hat{R}_S^{-1} \hat{S}_S \\ &= \hat{R}_S^{-1} \hat{S}_S - \hat{R}_S^{-1} \hat{R}_S \Theta + \Theta \\ &= \Theta + \hat{R}_S^{-1} (\hat{S}_S - \hat{R}_S \Theta).\end{aligned}$$

Notons que $\hat{R}_S = \lim_{r \rightarrow \infty} \hat{R}_S^{(r)}$ et $\hat{S}_S = \lim_{r \rightarrow \infty} \hat{S}_S^{(r)}$.

De même,

1) \hat{R}_S converge en probabilité vers \hat{R}_U , où $\hat{R}_U = \lim_{r \rightarrow \infty} \hat{R}_U^{(r)}$;

2) Sous certaines conditions, $\hat{R}_S^{-1} (\hat{S}_S - \hat{R}_S \Theta)$ converge vers une loi normale (voir Pfefferman et coll.(1998) pour la discussion des conditions de convergence).

Par conséquent, $\hat{R}_S^{-1} (\hat{S}_S - \hat{R}_S \Theta)$ converge vers une loi normale. On obtient donc la variance de $\hat{\Theta}$:

$$Var(\hat{\Theta}) = Var[\Theta + \hat{R}_S^{-1} (\hat{S}_S - \hat{R}_S \Theta)].$$

Un estimateur de cette variance est donné par

$$\begin{aligned}v(\hat{\Theta}) &= \hat{R}_S^{-1} v(\hat{S} - \hat{R}\Theta) \hat{R}_S^{-1} \\ &= \hat{R}_S^{-1} \frac{m}{m-1} \left(\sum_{j \in S} w_j^2 d_j d_j' \right) \hat{R}_S^{-1},\end{aligned}$$

où

$$d_j = \begin{pmatrix} \hat{b}_j^2 (\hat{\mu}_j^2 - \hat{\theta}_1 - \hat{\theta}_2 / \hat{T}_{5j}) \\ \hat{\theta}_2^{-2} [\hat{T}_{6j} - (\hat{N}_j - 1) \hat{\theta}_2] + [\hat{b}_j^2 (\hat{\mu}_j^2 - \hat{\theta}_1 - \hat{\theta}_2 / \hat{T}_{5j})] / \hat{T}_{5j} \end{pmatrix}.$$

Chapitre 5

Application

Dans les chapitres précédents, on a étudié quatre méthodes d'estimation ; dans ce chapitre, on applique ces méthodes aux données d'enquête. L'enquête qu'on a choisi est l'Enquête auprès des jeunes en maison d'accueil qui a été réalisée en 1987 aux États-Unis par le Ministère de la justice américaine.

5.1 Plan de sondage

La base d'échantillonnage de cette enquête est deux cent six institutions aux États-Unis. Les institutions sont réparties en 16 strates selon leurs taille. En effet, les cinq premières strates sont des regroupements des 195 petites institutions et les neuf plus grandes institutions sont chacune une strate .

Dans les cinq premières strates, on a un plan d'échantillon à deux degrés : le premier degré (niveau 2) sont des institutions sélectionnées dans chaque strate et le deuxième degré (niveau 1), des jeunes sélectionnés dans chaque institution.

Puisque, dans ce document, on n'étudie que le plan de sondage à deux degrés, on ne saisit pas d'information sur la stratification et on considère seulement les données dans les cinq premières strates. Finalement, notre base de données contient 39 institutions (unités primaires) et 1799 jeunes contrevenants (unités secondaires).

Les variables qui servent à l'analyse sont les suivantes :

- PSU : identificateur de l'unité primaire
- PSUSIZE : nombre de jeunes contrevenants dans l'unité primaire en 1987
- FINALWT : poids final comprend un ajustement pour la non-réponse et une calibration au total des jeunes contrevenants obtenu dans un recensement de 1987
- AGE : âge du jeune contrevenant ou de la jeune contrevenante (en année)
- AGEFIRST : âge hors de la première arrestation (en année)
- NUMARR : nombre d'arrestations

À partir de la variable AGE et AGEFIRST, on a créé une variable YEARS qui mesure le nombre d'années après la première arrestation ($YEARS = AGE - AGEFIRST$).

Dans la base de données, certaines valeurs pour la variable PSUSIZE sont absentes. Les valeurs manquantes sont remplacées par la valeur imputée qui est égale à la taille moyenne de toutes les institutions dans la strate à laquelle l'observation appartient.

5.2 Modèles de régression

Les modèles de régression à étudier ont tous une variable de réponse et une variable explicative. La variable de réponse et la variable explicative sont respectivement NUMARR (nombre d'arrestation) et YEARS (nombre d'années après la première arrestation).

Lors de l'ajustement du modèle de régression simple, les hypothèses de linéarité, de normalité et d'homoscédasticité ne sont pas respectées, qui nécessite une transformation sur les données. Alors, on prend une transformation pour la variable de réponse $\log NUMARR = \log(NUMARR)$, qui nous permet d'obtenir le modèle dont les hypothèses de base sont bien respectées.

Basé sur les données transformées, on écrit les deux modèles de régression suivants :

I. Le modèle de régression simple :

$$\log NUMARR_i = \beta_0 + \beta_1 * YEARS_i + \varepsilon_i ,$$

avec $\varepsilon \sim \mathcal{N}(0, \sigma_i^2)$.

Lors de l'application du modèle sur les données de l'enquête, on utilise le poids FINALWT.

II. Le modèle de régression multinationnel :

$$\text{LogNUMARR}_{ij} = \beta_0 + u_j + \beta_1 * \text{YEARS}_{ij} + \varepsilon_{ij}$$

avec $u \sim \mathcal{N}(0, \theta_1)$ et $\varepsilon \sim \mathcal{N}(0, \theta_2)$.

Lors de l'application du modèle sur les données de l'enquête, on applique les poids à chaque niveau. Ainsi :

au niveau 1,

$$w_{i|j} = \frac{1}{\pi_{i|j}} = \frac{\text{FINALWT}_{ij} * \text{PSUSIZE}_j}{\sum_{i=1}^{\text{PSU}} \text{FINALWT}_{ij}},$$

et au niveau 2,

$$w_j = \frac{1}{\pi_j} = \frac{\sum_{i=1}^{\text{PSU}} \text{FINALWT}_{ij}}{\text{PSUSIZE}_j}.$$

En fait, les estimateurs de B et Θ sont invariants d'échelle de w_j . Pourtant, il conviendrait de réduire le biais de l'estimation de Θ dès que la normalisation des poids $w_{i|j}$ sera entrée en vigueur. Pfeiffermann et al. (1998) ont donc proposé deux méthodes de normalisation en $w_{i|j}$:

Méthode 1 : Remplacer $w_{i|j}$ par $w_{i|j}^* = \lambda_{1j} w_{i|j}$, où $\lambda_{1j} = \sum_i w_{i|j} / \sum_i w_{i|j}^2$;

Méthode 2 : Remplacer $w_{i|j}$ par $w_{i|j}^* = \lambda_{2j} w_{i|j}$, où $\lambda_{2j} = n_j / \sum_i w_{i|j}$.

Notons que, après la normalisation, le somme des poids comporte certaines caractéristiques de l'échantillon. Le somme des poids $((\sum_i w_{i|j})^2 / \sum_i w_{i|j}^2)$, dans le premier cas, peut être interprété comme le taille d'échantillon effective, tandis que le somme des poids (n_j) est la taille d'échantillon actuelle dans le deuxième cas.

5.3 Résultats

Les calculs des estimateurs sont faits par les logiciels SAS 9.1 et Splus. Dans le modèle de régression simple, on utilise PROC REG (SAS 9.1) pour estimer les

paramètres non-pondérés et PROC SURVEYREG (SAS 9.1) pour estimer les paramètres pondérés. Dans le modèle multiniveau, on calcule les estimateurs non-pondérés en se servant de PROC MIXED (SAS 9.1). Quant aux estimations des estimateurs pondérés dans le modèle multiniveau, il n'y a pas de procédures en SAS destinés à ce genre du problème, on le programme donc en Splus. (Voir les annexes A et B pour toutes les programmations.)

Estimateurs	Régression simple		Régression multi-niveaux			
	non pond.	pondérée	non pond.	pondérée	méthode1	méthode2
constante (β_0)	0,77165 (0,03502)	0,74947 (0,06579)	0,7863 (0,04997)	0,79774 (0,06289)	0,77058 (0,06211)	0,77094 (0,06218)
YEARS (β_1)	0,28245 (0,00911)	0,28693 (0,01455)	0,2793 (0,00927)	0,27977 (0,01293)	0,28475 (0,01260)	0,28465 (0,01262)
Institution (θ_1)			0,04268 (0,01426)	0,07297 (0,03674)	0,06416 (0,00133)	0,06425 (0,00133)
Erreur (θ_2)	0,74417	0,73451	0,6996 (0,02393)	0,21027 (0,13639)	0,35521 (0,00353)	0,35303 (0,00346)

Note : les erreurs types sont données dans les parenthèses.

Interprétation du modèle

Prenant l'exemple du modèle de régression multiniveau, on interprète les estimateurs comme suivants : Durant la première année après la première arrestation (YEARS=0), les nombres d'arrestation de jeunes contrevenants aux États-unis varient d'une institution à l'autre avec une moyenne de 2, 2205 ($=e^{0,79774}$). Si le nombre d'années après la première arrestation augmente d'une unité, le nombre d'arrestation va être multiplié par 32% ($e^{0,27977} = 1,3228$). La variance inter-institutions est 0,07297 et la variance intra-institutions est 0,21027.

Comparaisons des résultats

- Régression simple vs. régression multiniveau

On constate que les estimations pour les β ne varient pas beaucoup d'un modèle à l'autre. Quant aux estimateurs θ , on remarque que les variations due aux erreurs (θ_2) dans le modèle de régression multiniveau sont moins élevées que celles dans le modèle de régression simple. En effet, dans le modèle multiniveau, on considère qu'une proportion de variation vient de la variation inter-institutions.

- Résultats pondérés vs. résultats non pondérés

Dans le cadre du modèle multiniveau, on observe que tous les estimateurs pondérés de la variation inter-institutions (θ_1) sont plus grands que l'estimateur non pondéré. Par contre, les estimateurs pondérés de la variation intra-institution (θ_2) est beaucoup plus petits que l'estimateur non pondéré.

- Poids final vs. poids normalisés

À partir de la figure 5.1, on observe que la moyenne et la variance de poids au deuxième degré baissent après la normalisation. En effet, les poids normalisés ($w_{i|j}^*$) sont presque égaux à une constante dans deux cas.

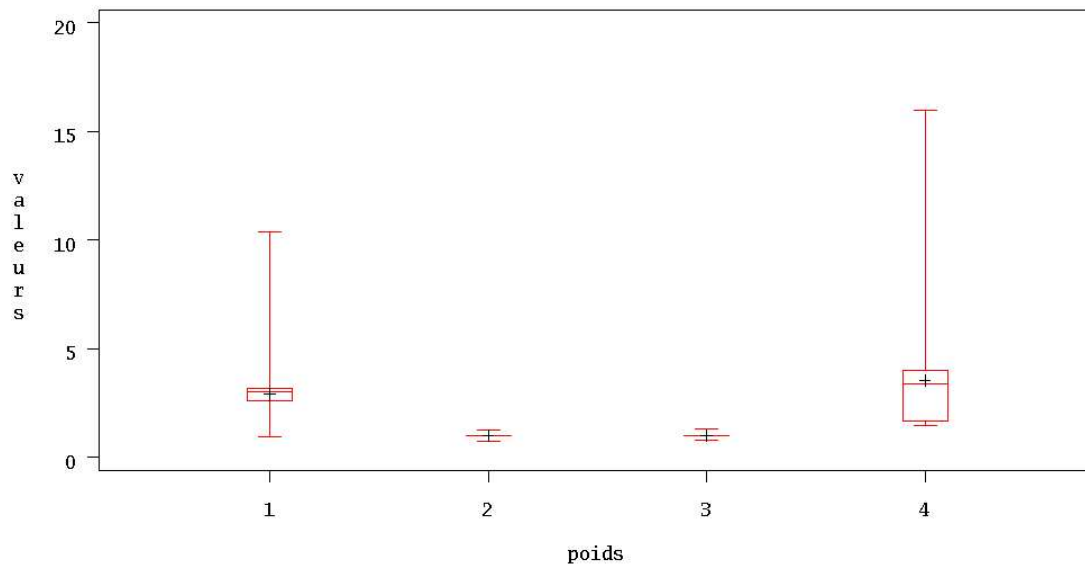


FIG. 5.1 – Boxplot : comparaison des poids. 1 : $w_{i|j}$, 2 : $\lambda_{1j}w_{i|j}$, 3 : $\lambda_{2j}w_{i|j}$, 4 : w_j .

En plus, l'effet de la normalisation est très important sur l'estimation des paramètres. On constate que les estimateurs de la variation inter-institution (θ_1) calculé à propos des poids normalisés sont plus petits que celui obtenu à partir du poids final. Ainsi, les estimateurs de la variation intra-institution (θ_2) calculés à partir des poids normalisés sont plus grands que celui obtenu à partir du poids final.

Chapitre 6

Conclusion

Dans ce document, on a présenté l'utilisation de deux modèles de régression sur des données d'enquête complexe. Pour les données dont la population possède une structure hiérarchique, la préférence est donnée à la méthode de pondération basée sur le modèle de régression multiniveau.

Les avantages de l'usage de cette dernière méthode sont :

- Le modèle correspond bien à la structure du plan de sondage et à la nature de la population ;
- La méthode permet de tenir compte précisément d'informations du plan de sondage : un usage des poids d'échantillonnage est fait à tous les degrés du plan de sondage ;
- L'implantation de cette méthode de pondération est simple.

Quelques inconvénients à l'usage :

- Souvent des fichiers publics de données d'enquête ne contiennent qu'un seul poids final. Il n'y donc pas suffisamment d'information pour utiliser cette méthode. Dans ce cas-ci, un modèle de régression multiple à un niveau s'applique ;
- Afin d'atteindre une convergence souhaitée, une certaine normalisation sur les données d'enquête sera probablement nécessaire.

D'ailleurs, dans l'exemple donné, on ne fait qu'un petit essai sur des données d'enquête. Dans le futur, des études approfondies de l'aspect pratique de cette

méthode seront effectuées.

Bibliographie

- [1] Christensen, R. (1987). *Plane answers to complex questions : the theory of linear models*. Springer-Verlag, New York.
- [2] Casella, G., Berger, R. L. (1990). *Statistical inference*. Brooks/Cole Pub. Co., Pacific Grove, Calif.
- [3] Duchesne, T. (2004). *Notes de cours de Théorie et applications des méthodes de régression*.
- [4] Goldstein, H. (1995). *Multilevel Statistical Models*. London, Arnold.
- [5] Graybill, F. A. (1983). *Matrices with applications in statistics*. Wadsworth International Group, Belmont, Calif.
- [6] Graubard, B. and Korn, E. (1996). Modeling the Sampling Design in the Analysis of Health Surveys. *Statistical Methods in Medical Research*, vol. 5, 263-281.
- [7] Kovacevic, M. S. and Rai, S. N. (2003). A pseudo maximum likelihood approach to multilevel modeling of survey data. *Communications in Statistics, Theory and Methods*, vol. 32, 103-121.
- [8] Lohr, S. (1999). *Sampling : Design and Analysis*. Duxbury Press.
- [9] Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H. and Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models. *J.R.Statist.Soc. B*, vol. 60, 22-34.
- [10] Rivest, L.P. (2005). *Notes de cours de Sondages : Modèles et techniques*.

Annexe A :Programmation en SAS

```
data syc;
infile "Devoir4_donnees.txt" firstobs=2;
input
    STRATUM PSU PSUSIZE INITWT FINALWT RANDGRP AGE RACE ETHNICTY
    EDUC SEX LIVewith FAMTIME CRIMTYPE EVERVIOL NUMARR PROBTN
    CORRINST EVERTIME PRVIOL PRPROP PRDRUG PRPUB PRJUV AGEFIRST
    USEWEPN ALCUSE EVERDRUG ;
run;
```

```
data mod;
    set syc;
    if age=99 then age=.;
    if numarr=99 then numarr=. ;
    if agefirst=99 then agefirst=. ;
run;
```

```
*****;
*                               data mod2                               *;
*****;
```

*****Base de donnees contient seulement strate1-5;

*****On impute des valeurs de PSUSIZE pour remplacer les ;
*****valeurs manquantes; On l'impute par la taille moyenne;
*****de toutes les institution dans le strate qu'elle ;

```
*****appartient
```

```
data mod2;
set mod;
if STRATUM > 5 then delete;
if stratum = 1 & psusize = 999 then psusize = 29;
run;
```

```
*****;
*      data trans:transformation de donnees      *;
*****;
```

```
data trans;
  set mod2;
  lognumarr = log(numarr);
  years = age - agefirst;
run;
```

```
*****;
*      regression simple      *;
*****;
```

```
***** non-ponderee;
proc reg data=trans;
model lognumarr = years;
title 'regression multiple sans plan';
run;
quit;
```

```
***** ponderee;
proc surveyreg data=trans;
  model lognumarr = years / anova;
  cluster psu;
  weight finalwt;
  title 'regression multiple avec plan ';
run;
```

```
*****;
*           regression multiniveau           *;
*****;

***** non-ponderee;
proc mixed data=trans method=ML covtest ;
class psu;
model lognumarr=years /S  XPVIX XPVIXC XPVIXI ;
random intercept / subject=psu;
title'regression multi-niveaux sans plan';
run;
```


Annexe B :Implantation du modèle multiniveau

```
IGLSvar<-function(donnees.matrice,precision,maxIteration)
#####
#Entree:
#i)donnees.matricecontienne[sumj(Nj)x(p+2)]
# unite primaire
# le vecteur Y [sumj(Nj)x1],
# le matrice X [sumj(Nj)xp],(x_1,...,x_(p-1))
# le poids wj (w2)
# le poids wi/j(w1)
#ii)precision:precision pour beta et theta demandee
# par l'utilisateur(ex:0.001)
#iii)maxIteration:le plus grand nombre d'iteration demande
# par l'utilisateur
#Algorithme:
# Methode d'implantation introduite dans la section 4.3
# Algorithme introduit dans la section 3.2.2
# (formules utilisees dans la section 2.3)
# Estimation de la variance dans la section 4.4
#####
{

#=====  
# remplir z_ij par w_j^(-1/2)z_ij #
```

```

#          z_oij par  $w_j^{-1/2} * w_{i/j}^{-1/2} z_{0ij}$           #
#=====#

#w_j^{-1/2} z_ij
#w_j^{-1/2} * w_{i/j}^{-1/2} z_{0ij}

#initialiser  $w_j^{-1/2} z_{ij}$  et  $w_j^{-1/2} * w_{i/j}^{-1/2} z_{0ij}$ 
  all.matrice<-donnees.matrice
  all.matrice[, (ncol(donnees.matrice)-1)]<-
  donnees.matrice[, (ncol(donnees.matrice)-1)]^{-1/2}
  all.matrice[, ncol(donnees.matrice)]<-
diag(donnees.matrice[, (ncol(donnees.matrice)-1)]^{-1/2},
nrow(donnees.matrice))
%*%donnees.matrice[, ncol(donnees.matrice)]^{-1/2}

#ajouter une colonne de 1 pour X
  all.matrice<-cbind(all.matrice[, 1:2], rep(1, nrow(donnees.matrice)),
  all.matrice[, 3:ncol(all.matrice)])

#all.matrice[sumj(Nj)x(p+4)]
#  unite primaire
#  le vecteur Y [sumj(Nj)x1],
#  le matrice X [sumj(Nj)xp], (1, ..., x_(p-1))
#   $w_j^{-1/2} z_{ij}$ 
#   $w_j^{-1/2} * w_{i/j}^{-1/2} z_{0ij}$ 

#=====AlgorithmeEtape1=====#
#          Initialisation du vecteur theta          #
#=====#

theta.vecteur<-c(1,1)
thetaAncien.vecteur<-c(1,1)
betaAncien.vecteur<-c(rep(-9999, ncol(donnees.matrice)-4))

beta.vecteur<-c(rep(0, ncol(donnees.matrice)-4))
nbIteration=0

```

```

#=====#
#           Condition d'arret           #
#=====#
#condition d'arret:
#1)la difference entre beta.vecteur et betaAncien.vecteur
#et celle entre theta.vecteur et thetaAncien.vecteur sont tres petites.
#Note:Ancien.vecteur est un vecteur qui est calcule dans l'iteration
#precedente
#2)nb d'iteration depasse la valeur demande par l'utilisateur

while(sum(abs(thetaAncien.vecteur-theta.vecteur)>
c(rep(precision,length(theta.vecteur))))>0
||
sum(abs(betaAncien.vecteur-beta.vecteur)>
c(rep(precision,length(beta.vecteur))))>0
&&nbIteration<maxIteration)

#abs(thetaAncien.vecteur-theta.vecteur)>
#c(rep(precision,length(theta.vecteur)))
#sort un vecteur de boolean.

#Lorsque sum(>0),les valeurs de theta ne sont pas
#toutes atteintes leurs prevision.Il est donc necessaire
#de continuer l'iteration.

{

#mettre a jour les informations de l'iteration
nbIteration=nbIteration+1
thetaAncien.vecteur<-theta.vecteur
betaAncien.vecteur<-beta.vecteur

#=====AlgorithmeEtape2=====#
#           Calculer beta(r)           #
#=====#
P.matrice<-
matrix(0,nrow=ncol(all.matrice)-4,ncol=ncol(all.matrice)-4,byrow=T)

```

```

Q.vecteur<-c(rep(0,ncol(all.matrice)-4))

for(i in 1:length(unique(all.matrice[,1])))
{
  valeurj=unique(all.matrice[,1])[i]

#calculerP

  Xj.matrice<-
all.matrice[all.matrice[,1]==valeurj,(3:(ncol(all.matrice)-2))]
  Dj.matrice<-
diag(all.matrice[all.matrice[,1]==valeurj,ncol(all.matrice)])
  Zj.vecteur<-
all.matrice[all.matrice[,1]==valeurj,(ncol(all.matrice)-1)]

  T1j.vecteur<-t(Xj.matrice)%%ginverse(Dj.matrice)%%Xj.matrice
  T2j.vecteur<-t(Xj.matrice)%%ginverse(Dj.matrice)%%Zj.vecteur
  T5j.vecteur=t(Zj.vecteur)%%ginverse(Dj.matrice)%%Zj.vecteur

  aj.vecteur=1/(T5j.vecteur+theta.vecteur[2]/theta.vecteur[1])

  P.matrice<-
1/theta.vecteur[2]*(T1j.vecteur-aj.vecteur[1]*(T2j.vecteur)
%%t(T2j.vecteur))+P.matrice

#calculerQ

  Yj.vecteur<-all.matrice[all.matrice[,1]==valeurj,2]
  T3j.vecteur<-t(Xj.matrice)%%ginverse(Dj.matrice)%%Yj.vecteur
  T4j.vecteur<-t(Zj.vecteur)%%ginverse(Dj.matrice)%%Yj.vecteur

  Q.vecteur<-1/theta.vecteur[2]
*(T3j.vecteur-aj.vecteur[1]*T2j.vecteur%%T4j.vecteur)+Q.vecteur

}#fin du bloc for()

#Afficher le message,si la matrice n'est pas inversible

```

```

erreur<-try(solve(P.matrice,Q.vecteur))
if(inherits(erreur,"Error"))
{
print('La matrice nest pas inversible,hors du calcul de B=PQ')
}

beta.vecteur<-solve(P.matrice,Q.vecteur)

#####AlgorithmeEtape3#####
#####Implantation Etape2#####
#           Calculer theta(r)           #
#####

R11=0
R12=0
R22=0
S1=0
S2=0

for(i in 1:length(unique(all.matrice[,1])))
{
    valeurj=unique(all.matrice[,1])[i]

    Xj.matrice<-
all.matrice[all.matrice[,1]==valeurj,(3:(ncol(all.matrice)-2))]
    Dj.matrice<-
diag(all.matrice[all.matrice[,1]==valeurj,ncol(all.matrice)])
    Zj.vecteur<-
all.matrice[all.matrice[,1]==valeurj,(ncol(all.matrice)-1)]

    T5j.vecteur=t(Zj.vecteur)%*%ginverse(Dj.matrice)%*%Zj.vecteur
    aj.vecteur=1/(T5j.vecteur+theta.vecteur[2]/theta.vecteur[1])
    Yj.vecteur<-all.matrice[all.matrice[,1]==valeurj,2]
    Ej.vecteur<-Yj.vecteur-Xj.matrice%*%beta.vecteur

    bj.vecteur=1/(theta.vecteur[1]+theta.vecteur[2]/T5j.vecteur)

```

```

    uj.vecteur=
t(Ej.vecteur)%*%solve(Dj.matrice)%*%Zj.vecteur/T5j.vecteur
    T6j.vecteur=t(Ej.vecteur-uj.vecteur[1]*Zj.vecteur)%*%
solve(Dj.matrice)%*%(Ej.vecteur-uj.vecteur[1]*Zj.vecteur)

    Nj=sum(donnees.matrice[donnees.matrice[,1]==valeurj,
ncol(donnees.matrice)])

#calculerR

    R11=(donnees.matrice
[donnees.matrice[,1]==valeurj,(ncol(donnees.matrice)-1)][1])
*(bj.vecteur[1]^2)+R11

    R12=(donnees.matrice
[donnees.matrice[,1]==valeurj,(ncol(donnees.matrice)-1)][1])
*(bj.vecteur[1]^2)/T5j.vecteur[1]+R12

    R22=(donnees.matrice
[donnees.matrice[,1]==valeurj,(ncol(donnees.matrice)-1)][1])
*(1/theta.vecteur[2]^2*(Nj-1)+(bj.vecteur[1]^2)/T5j.vecteur[1]^2)
+R22

#calculerS

    S1=bj.vecteur[1]^2*uj.vecteur[1]^2+S1
    S2=(1/theta.vecteur[2]^2*T6j.vecteur[1]
+bj.vecteur[1]^2*uj.vecteur[1]^2/T5j.vecteur[1])+S2

}#fin du bloc for()

R.matrice<-matrix(c(R11,R12,R12,R22),nrow=2,ncol=2,byrow=T)

S.vecteur<-matrix(c(S1,S2),nrow=2,ncol=1,byrow=T)

#Afficher le message,si la matrice n'est pas inversible

```

```

erreur<-try(solve(R.matrix,S.vecteur))
if(inherits(erreur,"Error"))
{
print('Matrice nest pas inversible hors du calcul de Theta=RS')
}

theta.vecteur<-solve(R.matrice,S.vecteur)

#modification si des variances sont negatives

if(theta.vecteur[1]<0)
{
theta.vecteur[1]=0
}

if(theta.vecteur[2]<0)
{
theta.vecteur[2]=0
}

}#fin de while

#=====#
#          Calculer les variances de Beta          #
#=====#

w2cc.matrice<-
matrix(0,nrow=ncol(all.matrice)-4,ncol=ncol(all.matrice)-4,byrow=T)

for(i in 1:length(unique(all.matrice[,1])))
{
    valeurj=unique(all.matrice[,1])[i]

    Xj.matrice<-
all.matrice[all.matrice[,1]==valeurj,(3:(ncol(all.matrice)-2))]
    Dj.matrice<-
diag(all.matrice[all.matrice[,1]==valeurj,ncol(all.matrice)])

```

```

Zj.vecteur<-
all.matrice[all.matrice[,1]==valeurj,(ncol(all.matrice)-1)]
T5j.vecteur=t(Zj.vecteur)%*%ginverse(Dj.matrice)%*%Zj.vecteur
aj.vecteur=1/(T5j.vecteur+theta.vecteur[2]/theta.vecteur[1])
Yj.vecteur<-all.matrice[all.matrice[,1]==valeurj,2]
Ej.vecteur<-Yj.vecteur-Xj.matrice%*%beta.vecteur

T2j.vecteur<-t(Xj.matrice)%*%ginverse(Dj.matrice)%*%Zj.vecteur
T7j.vecteur<-t(Xj.matrice)%*%ginverse(Dj.matrice)%*%Ej.vecteur
T8j.vecteur<-t(Ej.vecteur)%*%ginverse(Dj.matrice)%*%Zj.vecteur

cj.vecteur<-
T7j.vecteur-aj.vecteur[1]*(T2j.vecteur)%*%t(T8j.vecteur)
w2cc.matrice<-cj.vecteur%*%t(cj.vecteur)+w2cc.matrice

}#fin du bloc for()

m=length(unique(all.matrice[,1]))

varBeta.vecteur<-
m/(m-1)*ginverse(P.matrice)%*%w2cc.matrice
%*%ginverse(P.matrice)/theta.vecteur[2]^2

#=====
#          Calculer les variances de Theta          #
#=====

w2dd.matrice<-matrix(0,nrow=2,ncol=2,byrow=T)

for(i in 1:length(unique(all.matrice[,1])))
{
  valeurj=unique(all.matrice[,1])[i]

  Xj.matrice<-
all.matrice[all.matrice[,1]==valeurj,(3:(ncol(all.matrice)-2))]
  Dj.matrice<-
diag(all.matrice[all.matrice[,1]==valeurj,ncol(all.matrice)])
  Zj.vecteur<-

```

```

all.matrice[all.matrice[,1]==valeurj, (ncol(all.matrice)-1)]

T5j.vecteur=t(Zj.vecteur)%*%ginverse(Dj.matrice)%*%Zj.vecteur
aj.vecteur=1/(T5j.vecteur+theta.vecteur[2]/theta.vecteur[1])
Yj.vecteur<-all.matrice[all.matrice[,1]==valeurj,2]
Ej.vecteur<-Yj.vecteur-Xj.matrice%*%beta.vecteur

bj.vecteur=1/(theta.vecteur[1]+theta.vecteur[2]/T5j.vecteur)
uj.vecteur=
t(Ej.vecteur)%*%solve(Dj.matrice)%*%Zj.vecteur/T5j.vecteur
T6j.vecteur=
t(Ej.vecteur-uj.vecteur[1]*Zj.vecteur)%*%solve(Dj.matrice)
%*%(Ej.vecteur-uj.vecteur[1]*Zj.vecteur)

Nj=sum(donnees.matrice[donnees.matrice[,1]==valeurj,
ncol(donnees.matrice)])

#calculerR

R11=(donnees.matrice
[donnees.matrice[,1]==valeurj, (ncol(donnees.matrice)-1)][1])
*(bj.vecteur[1]^2)

R12=(donnees.matrice
[donnees.matrice[,1]==valeurj, (ncol(donnees.matrice)-1)][1])
*(bj.vecteur[1]^2)/T5j.vecteur[1]

R22=donnees.matrice
[donnees.matrice[,1]==valeurj, (ncol(donnees.matrice)-1)][1]*
(1/theta.vecteur[2]^2*(Nj-1)+(bj.vecteur[1]^2)/T5j.vecteur[1]^2)

#calculerS

S1=bj.vecteur[1]^2*uj.vecteur[1]^2
S2=(1/theta.vecteur[2]^2*T6j.vecteur[1]
+bj.vecteur[1]^2*uj.vecteur[1]^2/T5j.vecteur[1])

Rj.matrice<-matrix(c(R11,R12,R12,R22),nrow=2,ncol=2,byrow=T)

```

```
Sj.vecteur<-matrix(c(S1,S2),nrow=2,ncol=1,byrow=T)

wdj.vecteur<-Sj.vecteur-Rj.matrice%%theta.vecteur
w2dd.matrice<-wdj.vecteur%%t(wdj.vecteur)

}#fin du bloc for()

varTheta.vecteur<-
m/(m-1)*ginverse(R.matrice)%*%w2dd.matrice%%ginverse(R.matrice)

ecartBeta<-c((varBeta.vecteur[1,1])^0.5,(varBeta.vecteur[2,2])^0.5)
ecartTheta<-
c((varTheta.vecteur[1,1])^0.5,(varTheta.vecteur[2,2])^0.5)

#Affichage
print('Beta:')
print(beta.vecteur)
print('Theta:')
print(theta.vecteur)
print('#Iteration:')
print(nbIteration)

print('BetaEcart')
print(ecartBeta)

print('ThetaEcart')
print(ecartTheta)

return(NULL)

}#findefonction
```

