Recueil du Symposium 2002 de Statistique Canada
Modélisation des données d'enquête pour la recherche économique et sociale

RÉDUCTION DU BIAIS DES ERREURS-TYPES POUR LES MODÈLES LINÉAIRES ET LINÉAIRES GÉNÉRALISÉS DANS LE CAS D'ÉCHANTILLONS À PLUSIEURS DEGRÉS

Daniel F. McCaffrey[1] et Robert M. Bell[2]

RÉSUMÉ

La linéarisation et le jackknife sont des méthodes utilisées très fréquemment pour estimer les erreurs-types des coefficients des modèles de régression linéaire ajustés à des échantillons à plusieurs degrés. Dans certains plans d'échantillonnage, les estimateurs par linéarisation peuvent présenter un biais négatif important et le jackknife, de façon correspondante, un biais positif important. Nous proposons un estimateur de rechange, que nous appelons estimateur par linéarisation à biais réduit (LBR), fondé sur des résidus ajustés afin de mieux approximer la covariance des erreurs réelles. Si les erreurs sont indépendantes et identiquement distribuées (i.i.d.), l'estimateur LBR est sans biais. La méthode LBR s'applique à des échantillons dont les poids de sélection ne sont pas constants et à des modèles linéaires généralisés tels que la régression logistique. Nous examinons aussi les estimateurs LBR de l'erreur-type pour des modèles à équation d'estimation généralisée qui modélisent explicitement l'interdépendance des observations provenant de la même UPÉ. Les résultats d'une étude en simulation montrent que les erreurs-types calculées par la méthode LBR combinées à l'approximation de Satterthwaite pour déterminer la distribution de référence produisent des tests avec taux d'erreurs de première espèce (type I) proches des valeurs nominales.

MOTS-CLÉS : échantillons complexes, linéarisation, jackknife, approximation de Satterthwaite, degrés de liberté, équations d'estimation généralisées (EEG)

1. INTRODUCTION

La linéarisation est une méthode non paramétrique très souvent utilisée pour estimer les erreurs-types des coefficients des modèles de régression linéaire et de régression linéaire généralisée (Binder, 1983; Skinner, 1989). Si l'estimateur par linéarisation classique des erreurs-types donne de bons résultats dans le cas d'échantillons comportant un grand nombre d'unités primaires d'échantillonnage (UPÉ), l'estimateur peut être biaisé, surtout à la baisse, lorsque le nombre d'UPÉ est restreint ou que les variables prédictives ne sont pas équilibrées d'une UPÉ à l'autre (Bell et McCaffrey, 2002; Kott, 1994; Mancl et DeRouen, 2001). Par exemple, Bell et McCaffrey (2002) montrent que l'estimateur par linéarisation classique des erreurs-types pour les moindres carrés ordinaires est biaisé à la baisse, sauf dans le cas d'une hypothèse très restrictive concernant la distribution des variables explicatives. Bell et McCaffrey montrent que dans des conditions semblables, l'estimateur par le jackknife est biaisé à la hausse.

Kott (1996) a proposé une méthode pour réduire le biais de l'estimateur par linéarisation dans le cas de la régression par moindres carrés linéaires. Mancl et DeRouen (2001) ont élaboré une solution de rechange différente dans le contexte des équations d'estimation généralisées (EEG). Les deux approches consistent à modifier les vecteurs résiduels utilisés dans l'estimateur par linéarisation classique. Nous allons y revenir en détail. Dans Bell et McCaffrey (2002), nous proposons une méthode de rechange pour rajuster les résidus, appelée linéarisation à biais réduit (LBR).

Dans la présente communication, nous passons en revue nos résultats pour la méthode des moindres carrés ordinaires et présentons des généralisations de la méthode LBR pour 1) les moindres carrés pondérés, 2) les moindres carrés généralisés, 3) les modèles linéaires généralisés, et 4) les équations d'estimation généralisées. Nous terminons en présentant une application de la régression logistique utilisée pour estimer l'effet d'un traitement dans le cadre d'une expérience d'échantillonnage aléatoire par grappes.

[1] RAND, 201 North Craig Street, Suite 102, Pittsburgh, PA, USA, 15090
[2] AT&T Labs-Research, Room C211, 180 Park Avenue, Florham Park, NJ, USA, 07932

2. MOINDRES CARRÉS ORDINAIRES

Nous utilisons la méthode des moindres carrés ordinaires sur un échantillon à deux degrés pour élaborer l'estimateur LBR. Nous présentons les étapes essentielles servant à trouver l'estimateur pour les moindres carrés, et ces étapes renvoient naturellement à des généralisations pour les modèles linéaires généralisés, les EEG et les analyses pondérées.

2.1 Linéarisation et jackknife

Soit n, le nombre d'UPÉ et m_i, le nombre d'unités finales d'échantillonnage provenant de la i^e UPÉ, pour $i = 1,\ldots, n$. La taille globale de l'échantillon est $M = \sum_i m_i$. Nous supposons que $y_{ij} = \beta' x_{ij+} \varepsilon_{ij}$, où $\boldsymbol{\varepsilon}$ est caractérisé par une moyenne nulle et une matrice de covariance $\mathbf{V}$, et où les y_{ij}, x_{ij} et ε_{ij} se rapportent tous à la j^e observation provenant de la i^e UPÉ. Nous laissons tomber l'hypothèse type des MCO voulant que les erreurs soient *i.i.d.* et supposons uniquement que les erreurs provenant d'UPÉ distinctes ne sont pas corrélées. Plus précisément, nous supposons que $\mathbf{V}$ est une matrice diagonale par blocs comptant $m_i \times m_i$ blocs $\mathbf{V}_i$ pour $i = 1,\ldots, n$. En plus de la notation de ce modèle, dans la présente communication, $\mathbf{I}$ représente une matrice d'identité $M \times M$ et $\mathbf{I}_i$, une matrice d'identité $m_i \times m_i$.

Représentons par $\hat{\beta}$ les coefficients estimés du modèle de régression linéaire. Pour simplifier la présentation, nous discutons généralement d'une combinaison linéaire de coefficients de régression, $l'\hat{\beta}$, pour un vecteur colonne arbitraire l. Dans le cas particulier où un élément de $l = 1$ et les autres sont nuls, $l'\hat{\beta}$ représente un coefficient estimé unique. Si les erreurs ne sont pas corrélées entre les UPÉ, la variance de $l'\hat{\beta}$ est

$$Var(l'\hat{\beta}) = l'(\mathbf{X}'\mathbf{X})^{-1}\left(\sum_{i=1}^{n}\mathbf{X}_i'\mathbf{V}_i\mathbf{X}_i\right)(\mathbf{X}'\mathbf{X})^{-1}l\,, \qquad (2.1)$$

où $\mathbf{X}$ et $\mathbf{X}_i$ sont les matrices de plan d'échantillonnage pour l'échantillon complet et pour l'UPÉ i, respectivement.

L'estimateur par linéarisation standard de la variance de $l'\hat{\beta}$ est donné par :

$$v_L = l'(\mathbf{X}'\mathbf{X})^{-1}\left(c\sum_{i=1}^{n}\mathbf{X}_i'\mathbf{r}_i\mathbf{r}_i'\mathbf{X}_i\right)(\mathbf{X}'\mathbf{X})^{-1}l \qquad (2.2)$$

où $\mathbf{r}_i$ est le vecteur des résidus pour la i^e UPÉ. Les matrices inconnues $\mathbf{V}_i$ sont estimées par $c\mathbf{r}_i\mathbf{r}_i'$ et c égale habituellement $n/(n-1)$.

On utilise parfois le jackknife comme solution de rechange à la linéarisation (Rust et Rao, 1996). Soit $\{\tilde{\beta}_{[i]}\}$ un ensemble de pseudo-valeurs ou d'estimations de β d'après des données n'incluant pas la i^e UPÉ. Pour un échantillon à plusieurs degrés, l'estimateur par le jackknife se présente comme suit :

$$v_{JK} = [(n-1)/n]\ \sum_i l'\left(\tilde{\beta}_{[i]} - \hat{\beta}\right)\left(\tilde{\beta}_{[i]} - \hat{\beta}\right)' l \qquad (2.3)$$

Dans Bell et McCaffrey (2002), nous montrons qu'on peut formuler l'estimateur par le jackknife comme suit : $v_{JK} = c\, l'(\mathbf{X}'\mathbf{X})^{-1}\ \{\sum_{i=1}^{n}\ \mathbf{X}_i'\ (\mathbf{I}_i - \mathbf{H}_{ii})^{-1}\ \mathbf{r}_i\mathbf{r}_i'(\mathbf{I}_i - \mathbf{H}_{ii})^{-1}\mathbf{X}_i\}l$, où $c=(n\text{-}1)/n$, où $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})\mathbf{X}'$ et $\mathbf{X}_i$ et $\mathbf{H}_{ii}$ représentent les sous-matrices de $\mathbf{X}$ et de $\mathbf{H}$ correspondant à la i^e UPÉ. L'estimateur par le jackknife est donc semblable à l'estimateur par linéarisation, mais le produit externe des résidus bruts y est remplacé par le produit externe du résidu rajusté. Les théorèmes 1 et 2 de Bell et McCaffrey (2002) montrent que lorsque $\mathbf{V} = \mathbf{I}$, l'estimateur par linéarisation est légèrement biaisé, sauf dans des conditions restrictives, tandis que l'estimateur par le jackknife corrige de façon excessive et devient biaisé à la hausse.

D'autres auteurs ont proposé des rajustements pour réduire le biais de l'estimateur par linéarisation. Kott (1996) propose de calculer le ratio de $Var(l'\hat{\beta})$ to $E(v_L)$ sous l'hypothèse que $\mathbf{V} = \mathbf{I}$ et de rajuster v_L par le ratio. Si $\mathbf{V} = \sigma^2\mathbf{I}$, alors l'estimateur résultant sera sans biais. Dans le contexte des équations d'estimation généralisées, Mancel et DeRouen (2001) proposent d'ajuster les résidus provenant de chaque UPÉ pour réduire le biais de $\mathbf{r}_i\mathbf{r}_i'$ en tant

qu'estimateur de $\mathbf{V}_i$. Pour un modèle linéaire non pondéré, leur méthode consiste à approximer $E(\mathbf{r}_i\mathbf{r}_i')$ par $(\mathbf{I}_i - \mathbf{H}_{ii})\mathbf{V}_i(\mathbf{I}_i - \mathbf{H}_{ii})$ et à remplacer $\mathbf{r}_i$ par $(\mathbf{I}_i - \mathbf{H}_{ii})^{-1}\mathbf{r}_i$ dans l'équation (2). Donc, comme le montrent Bell et McCaffrey (2002), l'estimateur de Mancl et DeRouen est égal à $[n/(n-1)]v_{JK}$ pour les modèles linéaires non pondérés.

2.2 Erreurs-types liées à la linéarisation à biais réduit

Dans Bell et McCaffrey, nous proposons un compromis entre les estimateurs par linéarisation et par le jackknife, que nous appelons estimateur par linéarisation à biais réduit. À l'instar du jackknife, la LBR utilise le produit externe des résidus rajustés. Toutefois, le rajustement découle de l'équation $E(\mathbf{r}_i\mathbf{r}_i') = (\mathbf{I} - \mathbf{H})_i\mathbf{V}(\mathbf{I} - \mathbf{H})_i'$. Si nous connaissions $\mathbf{V}$, nous pourrions alors déterminer les matrices $\mathbf{A}_i$ de sorte que $\mathbf{A}_i[(\mathbf{I} - \mathbf{H})_i\mathbf{V}(\mathbf{I} - \mathbf{H})_i']\mathbf{A}_i' = \mathbf{V}_i$. Comme nous ne connaissons pas $\mathbf{V}$, nous utilisons plutôt une matrice des covariances de travail pour calculer notre estimateur. Plus précisément, nous proposons d'utiliser une matrice des covariances de travail de la forme $\mathbf{U} = \sigma^2\mathbf{I}$, qui simplifie la condition sur $\mathbf{A}_i$ en $\mathbf{A}_i(\mathbf{I}_i - \mathbf{H}_{ii})\mathbf{A}_i' = \mathbf{I}$ ou $\mathbf{A}_i = (\mathbf{I}_i - \mathbf{H}_{ii})^{-1/2}$. Le théorème 3 de Bell et McCaffrey montre que si $\mathbf{V} = \sigma^2\mathbf{I}$, alors v_{LBR} est sans biais. Dans la section 3, nous envisageons d'autres matrices des covariances de travail.

Si $m_i > 1$, $\mathbf{A}_i$ n'est pas unique. Si $\mathbf{V} = \sigma^2\mathbf{I}$, le choix de $\mathbf{A}_i$ est sans importance, car toute solution de (2.4) produira un estimateur sans biais de la variance. Cependant, les estimateurs résultants sont biaisés lorsque $\mathbf{V} \neq \sigma^2\mathbf{I}$, et le biais peut varier fortement selon le choix de $\mathbf{A}_i$. Parmi les solutions de rechange que nous avons essayées, nous avons trouvé (Bell et McCaffrey, 2002) que la racine carrée symétrique de $(\mathbf{I}_i - \mathbf{H}_{ii})^{-1}$ donnait le meilleur résultat, et nous avons donné à l'estimateur qui utilise cette racine le nom d'estimateur par linéarisation à biais réduit $v_{LBR} = l'(\mathbf{X}'\mathbf{X})^{-1}\sum_i\mathbf{X}_i'\mathbf{A}_i\mathbf{r}_i\mathbf{r}_i'\mathbf{A}_i\mathbf{X}_i\,(\mathbf{X}'\mathbf{X})^{-1}l$.

2.2 Variation des estimateurs de la variance

Bell et McCaffrey (2002) montrent que v_{LBR} est égal à la somme pondérée de n variables aléatoires indépendantes χ_1^2 où les poids sont les valeurs propres de la matrice $n \times n$ $\mathbf{G} = \{\mathbf{g}_i'\mathbf{V}\mathbf{g}_j\}$, pour $\mathbf{g}_i = (\mathbf{I} - \mathbf{H})_i'\mathbf{A}_i\mathbf{X}_i(\mathbf{X}'\mathbf{X})^{-1}l$. Dans cet article, nous montrons également que v_L et v_{JK} présentent des distributions semblables où $\mathbf{G}$ est défini respectivement par $\mathbf{g}_{Li} = [n/(n-1)]^{1/2}(\mathbf{I} - \mathbf{H})_i'\mathbf{X}_i(\mathbf{X}'\mathbf{X})^{-1}l$ et $\mathbf{g}_{JKi} = [(n-1)/n]^{1/2}(\mathbf{I} - \mathbf{H})_i'(\mathbf{I} - \mathbf{H}_{ii})^{-1}\mathbf{X}_i(\mathbf{X}'\mathbf{X})^{-1}l$. Si $\mathbf{V} = \sigma^2\mathbf{I}$ et $\mathbf{X}_i'\mathbf{X}_i(\mathbf{X}'\mathbf{X})^{-1}l$ pour $i = 1, ..., n$ sont constants, alors av_L, av_{JK} et av_{LBR} suivent tous la loi de distribution χ_{n-1}^2 pour $a = (n-1)/Var(l'\hat{\beta})$ (Bell et McCaffrey, 2002). Cependant, en général, $\mathbf{X}_i'\mathbf{X}_i(\mathbf{X}'\mathbf{X})^{-1}l$ n'est pas constante et le carré du coefficient de variation est supérieur à $2/(n-1)$, la statistique correspondante pour la variable aléatoire χ_{n-1}^2. Pour n'importe quel estimateur non paramétrique de la variance, le coefficient de variation peut être très grand pour certains plans d'échantillonnage. On observe une forte variabilité dans les mêmes conditions que celles où v_L et v_{JK} sont le plus fortement biaisés, c.-à-d. lorsque les résidus de quelques UPÉ seulement déterminent effectivement l'estimation finale de la variance.

Il convient notamment de se préoccuper de cette variabilité excédentaire lorsqu'on envisage, pour le test de vérification de l'hypothèse nulle $l'\beta = 0$, des distributions de référence dont le critère de test est de la forme $t = l'\hat{\beta} / \sqrt{v^*}$. Pour v_L, Shah, Holt et Folsom (1977) proposent de comparer t à une distribution t de référence à $n-1$ degrés de liberté. Cependant, comme la variance de $(n-1)v_L/E(v_L)$ a tendance à être plus grande que $2(n-1)$, les tests basés sur une distribution t à $n-1$ degrés de liberté auront tendance à produire un taux d'erreurs de première espèce supérieur à la valeur nominale, même si v_L est sans biais. Satterthwaite (1946) propose une autre méthode d'approximation de la distribution des estimateurs de la variance. En appariant les deux premiers moments à celui d'une variable aléatoire χ^2, nous approximons, jusqu'à une constante d'échelle, la distribution de v_L, v_{LBR} ou v_{JK} par une χ_f^2 où $f = 2/cv^2 = \left(\sum_{i=1}^n \lambda_i\right)^2 \Big/ \sum_{i=1}^n \lambda_i^2$ et où les λ_i sont les valeurs propres de la matrice $\mathbf{G}$ correspondante. Les tests basés sur les distributions t de référence ayant f degrés de liberté devraient, en principe, donner un meilleur taux d'erreurs de première espèce que ceux fondés sur $n-1$ degrés de liberté. Pan et Wall (2001) et Kott (1994, 1996) proposent d'utiliser l'approximation de Satterthwaite pour estimer le nombre de degrés de liberté pour des tests fondés sur la linéarisation standard ou sur les solutions de rechange à linéarisation proposées par Kott. Le nombre f de degrés de liberté de Satterthwaite oblige à spécifier la matrice inconnue $\mathbf{V}$. Nous supposons que $\mathbf{V}$ est

identiquement égale à la matrice d'identité—c.-à-d. que nous supposons que les erreurs sont indépendantes et homoscédastiques aux fins de la détermination du nombre de degrés de liberté.

La distribution de v_{LBR} (et des autres estimateurs de la variance) a tendance à être moins asymétrique et à avoir une queue inférieure moins lourde que la distribution d'une variable χ^2_f où f est égal au nombre de degrés de liberté de Satterthwaite. Donc, les distributions t de référence fondées sur l'approximation de Satterthwaite ont tendance à surestimer les probabilités de queue de distribution. Par exemple, si les données observées pour quelque UPÉ déterminent pour ainsi dire la valeur d'un coefficient, le nombre de degrés de liberté de Satterthwaite peut être inférieur à deux, ce qui implique incorrectement une densité du chi carré infinie à zéro. Conséquemment, la probabilité d'une statistique t très grande pourrait ne pas être aussi forte que le laisse entendre l'approximation de Satterthwaite, particulièrement lorsque le nombre de degrés de liberté de Satterthwaite est inférieur à 4 ou à 5. Dans ces conditions, les approximations par la méthode du col (Huzurbazar, 1999) offrent une solution de rechange prometteuse.

3. GÉNÉRALISATIONS

Le calcul de l'estimateur LBR pour la méthode MCO comportait quatre étapes :

1. calculer le $Var(l'\hat{\beta})$ en additionnant les termes $\mathbf{b}_i'\mathbf{V}_i\mathbf{b}_i$;
2. calculer le $E(\mathbf{r}_i\mathbf{r}_i') = \mathbf{Q}_i$ en utilisant une matrice des variances/covariances de travail, $\mathbf{U}$, pour l'inconnue $\mathbf{V}_i$;
3. trouver les solutions symétriques à $\mathbf{A}_i\mathbf{Q}_i\mathbf{A}_i = \mathbf{U}_i$;
4. v_{LBR} est égal à la somme des termes $\mathbf{b}_i'\mathbf{A}_i\mathbf{r}_i\mathbf{r}_i'\mathbf{A}_i\mathbf{b}_i$.

Nous considérons des modèles de rechange et nous leur étendons l'estimateur LBR en utilisant le modèle MCO et en calculant des formules pour les $\mathbf{b}_i$, $\mathbf{Q}_i$ et $\mathbf{A}_i$. Pour toutes les généralisations, $\mathbf{Q}_i$ **et** $\mathbf{A}_i$ seront de forme semblable. $\mathbf{Q}_i$ est égal à la matrice des covariances de travail avant et après multiplication par les rangées de la matrice de projection qui définit les résidus et leur transposée. Les valeurs de $\mathbf{A}_i$' sont égales aux racines des matrices associées aux produits des racines de la matrice des variances/covariances de travail et de $\mathbf{Q}_i$.

Les propriétés des généralisations aux modèles linéaires découlent directement des résultats pour la méthode MCO. Plus précisément, les estimateurs sont sans biais lorsque la matrice des covariances de travail est proportionnelle à la matrice de covariance réelle. Comme les résultats pour la méthode MCO ne s'appliquent pas aux estimateurs pour les modèles linéaires généralisés ni pour les EEG, les propriétés de ces estimateurs pour un petit échantillon doivent être étudiées au moyen d'une simulation.

3.1 LBR par les moindres carrés ordinaires pour une covariance de travail qui n'est pas l'identité

Dans l'équation (2.4), nous définissons les matrices de rajustement pour l'estimateur LBR en supposant une matrice des variances/covariances, $\mathbf{V} = k\mathbf{I}$, pour une constante k non précisée. Dans certains cas, nous utiliserions peut-être une autre matrice diagonale par blocs comme matrice des covariances de travail, $\mathbf{U}$, pour estimer nos erreurs-types. Dans ce cas, nous continuerions à estimer la variance de $l'\hat{\beta}_{MCO}$ donnée par l'équation (2.1), mais maintenant, $\mathbf{Q}_i = (\mathbf{I} - \mathbf{H})_i\mathbf{U}(\mathbf{I} - \mathbf{H})_i$ et les matrices de rajustement résolvent l'équation

$$\mathbf{A}_i(\mathbf{I} - \mathbf{H})_i\mathbf{U}(\mathbf{I} - \mathbf{H})_i'\mathbf{A}_i' = \mathbf{U}_i. \qquad (3.1)$$

Soit $\mathbf{Q}^{1/2}$ qui représente n'importe quelle matrice $\mathbf{Q}^{1/2\prime}\mathbf{Q}^{1/2} = \mathbf{Q}$ et $\mathbf{Q}^*$ qui représente la racine symétrique de $\mathbf{Q}^{-1}$ pourvu qu'elle existe; ainsi, $\mathbf{Q}^*\mathbf{Q}^* = \mathbf{Q}^{-1}$ et $\mathbf{Q}^*\mathbf{Q}\mathbf{Q}^* = \mathbf{I}$. La solution symétrique à l'équation (3.1) est la suivante :

$$\mathbf{A}_i = \mathbf{U}_i^{1/2}(\mathbf{U}_i^{1/2}\mathbf{Q}_i\mathbf{U}_i^{1/2'})^*\mathbf{U}_i^{1/2}. \qquad (3.2)$$

3.2 Moindres carrés pondérés

Prenons le cas où l'on assigne à chaque observation un poids w_{ij} et où $\mathbf{W} = \text{diag}\{w_{ij}\}$. Les estimateurs par les moindres carrés pondérés des coefficients de régression sont $\hat{\beta}_W = (\mathbf{X'WX})^{-1}\mathbf{X'Wy}$ et $var(l'\hat{\beta}_W) = l'(\mathbf{X'WX})^{-1}[\sum_i \mathbf{X}_i'\mathbf{W}_i\mathbf{V}_i\mathbf{W}_i\mathbf{X}_i](\mathbf{X'WX})^{-1}l$. Comme $\mathbf{r}_i = \mathbf{X}(\mathbf{X'WX})^{-1}\mathbf{X'Wy} = \mathbf{G}_W\mathbf{y}$, nous avons $\mathbf{Q}_i = (\mathbf{I} - \mathbf{G}_W)_i\mathbf{U}(\mathbf{I} - \mathbf{G}_W)_i'$ et $\mathbf{A}_i = \mathbf{U}_i^{1/2}(\mathbf{U}_i^{1/2}\mathbf{Q}_i\mathbf{U}_i^{1/2})^*\mathbf{U}_i^{1/2}$.

3.3 Moindres carrés généralisés linéaires

Nous considérons l'estimation par les moindres carrés généralisés des coefficients en utilisant la matrice des covariances de travail **U**. Toutefois, au lieu d'utiliser l'erreur-type fondée sur le modèle, nous utilisons l'estimateur de l'erreur-type par linéarisation pour protéger l'inférence contre une erreur de spécification dans la matrice des covariances de travail. On utilise couramment cette pratique dans l'analyse de données longitudinales (voir, par exemple, Liang et Zeger, 1986).

Les estimateurs par les moindres carrés pondérés des coefficients de régression sont $\hat{\beta}_{MCG} = (\mathbf{X'U}^{-1}\mathbf{X})^{-1}\mathbf{X'U}^{-1}\mathbf{y}$ et $var(l'\hat{\beta}_{MCG}) = l'(\mathbf{X'U}^{-1}\mathbf{X})^{-1}[\sum_i \mathbf{X}_i'\mathbf{U}_i^{-1}\mathbf{V}_i\mathbf{U}_i^{-1}\mathbf{X}_i](\mathbf{X'U}^{-1}\mathbf{X})^{-1}l$. La matrice de projection est $\mathbf{G}_{MCG} = \mathbf{X}(\mathbf{X'U}^{-1}\mathbf{X})^{-1}\mathbf{X'U}^{-1}$ et $\mathbf{Q}_i = (\mathbf{I} - \mathbf{G}_{MCG})_i\mathbf{U}(\mathbf{I} - \mathbf{G}_{MCG})_i'$. $\mathbf{A}_i$ est égal à $\mathbf{U}_i^{1/2}(\mathbf{U}_i^{1/2}\mathbf{Q}_i\mathbf{U}_i^{1/2})^*\mathbf{U}_i^{1/2}$. Toutefois, $\mathbf{Q}_i = \mathbf{U}_i^{1/2}(\mathbf{I} - \mathbf{H}_{MCG})_i\mathbf{U}^{-1/2}\mathbf{U}\mathbf{U}^{-1/2}(\mathbf{I} - \mathbf{H}_{MCG})_i'\mathbf{U}i^{1/2} = \mathbf{U}_i^{1/2}(\mathbf{I} - \mathbf{H}_{MCG,ii})\mathbf{U}i^{1/2}$, de sorte que

$$\mathbf{A}_i = \mathbf{U}_i^{1/2}[\mathbf{U}_i(\mathbf{I}_i - \mathbf{H}_{MCG,ii})\mathbf{U}_i]^*\mathbf{U}_i^{1/2} \tag{3.7}$$

où $\mathbf{H}_{MCG} = \mathbf{U}^{-1/2}\mathbf{X}(\mathbf{X'U}^{-1}\mathbf{X})^{-1}\mathbf{X'U}^{-1/2}$.

3.4 Modèles linéaires généralisés

Nous considérons le modèle linéaire généralisé où l'on suppose que la fonction de densité pour la réponse individuelle Y_{ij} est la suivante :

$$f_Y(y_{ij}) = exp\{(y_{ij}\theta_{ij} - b(\theta_{ij}))/a(\phi) + c(y_{ij}, \phi)\} \tag{3.8}$$

où $\theta_{ij} = h(\eta_{ij})$ et $\eta_{ij} = \boldsymbol{x}_{ij}'\beta$. La moyenne et la variance sont données par $\mu_{ij} = E(y_{ij}) = \dot{b}(\theta_{ij})$ et $v = \text{E}(y) = \ddot{b}(\theta_{ij})a(\phi)$. Pour estimer les coefficients, on suppose que les observations sont indépendantes de sorte que les estimations par le maximum de vraisemblance des coefficients apparaissent comme la solution à l'équation d'estimation suivante :

$$\sum_i \mathbf{X}_i'\Delta_i(\mathbf{y}_i - \quad_i) = 0 \tag{3.9}$$

où $\Delta_i = \text{diag}\{d\theta_{ij}/d\eta_{ij}\}$. On trouve les solutions à l'équation (3.9) par la méthode des moindres carrés repondérés itérativement où, à la dernière itération,

$$\hat{\beta}_{EMV} = (\mathbf{X'U}^{-1}\mathbf{X})^{-1}\mathbf{X'U}^{-1}\mathbf{z} \tag{3.10}$$

où $z_{ij} = \boldsymbol{x}_{ij}'\hat{\beta}_{EMV} + (y_{ij} - \hat{\mu}_{ij}) / \{\ddot{b}(\hat{\theta}_{ij})\dot{h}(\hat{\eta}_{ij})\}$, $\mathbf{U}^{-1} = \text{diag}\{\ddot{b}(\hat{\theta}_{ij})\dot{h}(\hat{\eta}_{ij})\}$, $\hat{\eta}_{ij} = \boldsymbol{x}_{ij}'\hat{\beta}_{EMV}$, $\hat{\theta}_{ij} = h(\hat{\eta}_{ij})$ et $\hat{\mu}_{ij} = \dot{b}(\hat{\theta}_{ij})$. Selon l'hypothèse de travail des observations indépendantes, la variance de $\mathbf{z}_i$ est approximativement $\mathbf{U}_i$, jusqu'à un terme d'échelle. Les modèles linéaires généralisés sont donc analogues aux moindres carrés généralisés pour les modèles linéaires et nous pouvons calculer un estimateur LBR pour les MLG en utilisant les formules pour les MCG.

Nous avons d'abord besoin d'une estimation de *var*($l'\hat{\beta}_{EMV}$) selon l'hypothèse moins restrictive que *var*(**z**) est une matrice diagonale par blocs, $\mathbf{V}_i$. $l'\hat{\beta}_{EMV}$ présente une distribution normale approximative avec $l'\left(\mathbf{X}'\mathbf{U}^{-1}\mathbf{X}\right)^{-1}\left[\sum_i \mathbf{X}_i'\mathbf{U}_i^{-1}\mathbf{V}_i\mathbf{U}_i^{-1}\mathbf{X}_i\right]\left(\mathbf{X}'\mathbf{U}^{-1}\mathbf{X}\right)^{-1}l$.

Puis, nous devons calculer $\mathbf{Q}_i$. Soit $\mathbf{G}_{MLG} = \mathbf{X}(\mathbf{X}'\mathbf{U}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{U}^{-1}$ et $\mathbf{r}_i = (\mathbf{z}_i - \hat{\mathbf{z}}_i)$, et les approximations du premier ordre donnent

$$E(\mathbf{r}_i\mathbf{r}_i') = (\mathbf{I} - \mathbf{G}_{MLG})_i\mathbf{V}_i(\mathbf{I} - \mathbf{G}_{MLG})_i'$$

Ainsi, tout comme dans le cas de MCG, $\mathbf{Q}_i = (\mathbf{I} - \mathbf{G}_{MLG})_i\mathbf{U}_i(\mathbf{I} - \mathbf{G}_{MLG})_i'$. Enfin, nous calculons des matrices de rajustement pour résoudre

$$\mathbf{A}_i(\mathbf{I} - \mathrm{G}_{MLG})_i\mathrm{U}(\mathbf{I} - \mathrm{G}_{MLG})_i\mathbf{A}_i' = \mathbf{U}_i,$$

qui, d'après les calculs MCG, sont donnés par

$$\mathbf{A}_i = \mathbf{U}_i^{1/2}(\mathbf{U}_i^{1/2}\mathbf{Q}_i\mathbf{U}_i^{1/2})^*\mathbf{U}_i^{1/2} = \mathbf{U}_i^{1/2}[\mathbf{U}_i(\mathbf{I}_i - \mathrm{H}_{MLG,ii})\mathbf{U}_i]^*\mathbf{U}_i^{1/2}$$

où $\mathrm{H}_{MLG} = \mathbf{U}^{-1/2}\mathbf{X}(\mathbf{X}'\mathbf{U}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{U}^{-1/2}$. L'estimateur LBR pour la méthode MLG est le suivant :

$$v_{LBR,MLG} = l'\left(\mathbf{X}'\mathbf{U}^{-1}\mathbf{X}\right)^{-1}\left[\sum_i \mathbf{X}_i'\mathbf{U}_i^{-1}\mathbf{A}_i\mathbf{r}_i\mathbf{r}_i'\mathbf{A}_i'\mathbf{U}_i^{-1}\mathbf{X}_i\right]\left(\mathbf{X}'\mathbf{U}^{-1}\mathbf{X}\right)^{-1}l$$

3.5 Équations d'estimation généralisées

Les équations d'estimation généralisées étendent les modèles linéaires généralisés pour permettre la corrélation parmi les observations provenant de la même unité. La matrice des covariances de travail est donnée par une matrice diagonale par blocs **U** de sorte que $a(\phi)var(\mathbf{y}_{ij}) = \mathbf{U}_i = \mathbf{\Omega}_i^{1/2}\mathbf{R}_i\mathbf{\Omega}_i^{1/2}$. On estime les coefficients de régression pour résoudre les équations d'estimation

$$\sum_i \mathbf{D}_i'\mathbf{U}_i^{-1}(\mathbf{y}_i - \quad_i) = 0\,, \qquad (3.11)$$

où $\mathbf{D}_i = d\mathbf{\mu}_i/d\beta = \mathbf{\Omega}_i\mathbf{\Delta}_i\mathbf{X}_i$. Comme l'expliquent Liang et Zeger (1986), on trouve les estimations de coefficient par itération entre une notation de Fisher modifiée pour estimer β sous réserve des valeurs courantes des paramètres de **R** et ϕ et de la méthode d'estimation du moment de ces paramètres de corrélation et d'échelle. L'algorithme à notation de Fisher modifiée équivaut à la méthode des moindres carrés repondérés itérativement. Soit z_{ij} défini comme ci-dessus et $\tilde{\mathbf{\Omega}}$ et $\tilde{\mathbf{R}}$ qui représentent les matrices **Ω** et **R** évaluées aux valeurs estimées courantes de leurs paramètres, de sorte que la matrice des covariances de travail pour $\mathbf{z}_i$ est $\mathbf{U}_i = \tilde{\mathbf{\Omega}}_i^{-1/2}\tilde{\mathbf{R}}_i\tilde{\mathbf{\Omega}}_i^{-1/2}$. L'estimateur EEG des coefficients est alors donné par l'équation suivante :

$$\hat{\beta}_{EEG} = (\mathbf{X}'\tilde{\mathbf{\Omega}}^{1/2}\tilde{\mathbf{R}}^{-1}\tilde{\mathbf{\Omega}}^{1/2}\mathbf{X})^{-1}\mathbf{X}'\tilde{\mathbf{\Omega}}^{1/2}\tilde{\mathbf{R}}^{-1}\tilde{\mathbf{\Omega}}^{1/2}\mathbf{z} = (\mathbf{X}'\mathbf{U}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{U}^{-1}\mathbf{z}$$

Toujours à l'instar de la méthode MCG, $(\mathbf{X}'\mathbf{U}^{-1}\mathbf{X})^{-1}\left\{\sum_i \mathbf{X}_i'\mathbf{U}_i^{-1}\mathbf{V}_i\mathbf{U}_i^{-1}\mathbf{X}_i\right\}(\mathbf{X}'\mathbf{U}^{-1}\mathbf{X})^{-1}$ approxime la variance de la distribution normale asymptotique de $l'\hat{\beta}_{EEG}$ et $E(\mathbf{r}_i\mathbf{r}_i') \approx (\mathbf{I} - \mathbf{G}_{EEG})_i\mathbf{V}_i(\mathbf{I} - \mathbf{G}_{EEG})_i'$, $\mathbf{G}_{EEG} = \mathbf{G}_{EEG} = \mathbf{X}(\mathbf{X}'\mathbf{U}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{U}^{-1} = \mathbf{X}(\mathbf{X}'\tilde{\mathbf{\Omega}}^{1/2}\tilde{\mathbf{R}}^{-1}\tilde{\mathbf{\Omega}}^{1/2}\mathbf{X})^{-1}\mathbf{X}'\tilde{\mathbf{\Omega}}^{1/2}\tilde{\mathbf{R}}^{-1}\tilde{\mathbf{\Omega}}^{1/2}$. Soit $\mathbf{Q}_i = (\mathbf{I} - \mathbf{G}_{EEG})_i\mathbf{U}_i(\mathbf{I} - \mathbf{G}_{EEG})_i' = (\mathbf{I} - \mathbf{G}_{EEG})_i'\ (\tilde{\mathbf{\Omega}}_i^{-1/2}\tilde{\mathbf{R}}_i\tilde{\mathbf{\Omega}}_i^{-1/2})(\mathbf{I} - \mathbf{G}_{EEG})_i'$ et

$$\mathbf{A}_i = \mathbf{U}_i^{1/2}(\mathbf{U}_i^{1/2}\mathbf{Q}_i\mathbf{U}_i^{1/2})^*\mathbf{U}_i^{1/2}$$

$$= \tilde{\mathbf{\Omega}}_i^{-1/2}\tilde{\mathbf{R}}_i^{1/2}\{\tilde{\mathbf{R}}^{1/2}\tilde{\mathbf{\Omega}}^{-1}\tilde{\mathbf{R}}^{1/2}(\mathbf{I}_i - \mathbf{H}_{EEGii})\tilde{\mathbf{R}}^{1/2}\tilde{\mathbf{\Omega}}^{-1}\tilde{\mathbf{R}}^{1/2}\}\tilde{\mathbf{R}}_i^{1/2}\tilde{\mathbf{\Omega}}_i^{-1/2} \qquad (3.12)$$

où $\mathbf{H}_{EEG} = \tilde{\mathbf{R}}^{-1/2}\tilde{\boldsymbol{\Omega}}^{1/2}\mathbf{X}(\mathbf{X}'\tilde{\boldsymbol{\Omega}}^{1/2}\tilde{\mathbf{R}}^{-1}\tilde{\boldsymbol{\Omega}}^{1/2}\mathbf{X})^{-1}\mathbf{X}'\tilde{\boldsymbol{\Omega}}^{1/2}\tilde{\mathbf{R}}^{-1/2}$. L'équation (3.12) est valable parce que $(\mathbf{I} - \mathbf{G}_{EEG}) = \tilde{\boldsymbol{\Omega}}_i^{-1/2}\tilde{\mathbf{R}}_i^{1/2}(\mathbf{I} - \mathbf{H}_{EEG})\tilde{\mathbf{R}}_i^{-1/2}\tilde{\boldsymbol{\Omega}}_i^{1/2}$.

L'estimateur LBR de la variance de $l'\hat{\beta}_{EEG}$ est

$$V_{LBR,EEG} = (\mathbf{X}'\tilde{\boldsymbol{\Omega}}^{1/2}\tilde{\mathbf{R}}^{-1}\tilde{\boldsymbol{\Omega}}^{1/2}\mathbf{X})^{-1}\left\{\sum_i \mathbf{X}_i'\tilde{\boldsymbol{\Omega}}_i\tilde{\mathbf{R}}_i^{-1}\tilde{\boldsymbol{\Omega}}_i\mathbf{A}_i\mathbf{r}_i\mathbf{r}_i'\mathbf{A}_i'\tilde{\boldsymbol{\Omega}}_i\tilde{\mathbf{R}}_i^{-1}\tilde{\boldsymbol{\Omega}}_i\mathbf{X}_i\right\}(\mathbf{X}'\tilde{\boldsymbol{\Omega}}^{1/2}\tilde{\mathbf{R}}^{-1}\tilde{\boldsymbol{\Omega}}^{1/2}\mathbf{X})^{-1}.$$

4. RÉSULTATS EMPIRIQUES

4.1 Étude de Monte Carlo pour les moindres carrés ordinaires

Dans Bell et McCaffrey (2002), nous présentons les résultats d'une simulation de Monte Carlo pour étudier les propriétés d'estimateurs de la variance et de tests de rechange pour MCO et un échantillon en grappes à deux degrés équilibré de n = 20 UPÉ où le nombre d'observations m = 10 est le même pour chaque UPÉ. Voici un résumé de ces résultats.

4.1.1 Conception de l'étude en simulation

Dans cette étude, pour toutes les répétitions de la simulation, nous utilisons une même matrice de plan d'échantillonnage **X** en choisissant quatre variables explicatives représentant une gamme de difficultés pour les estimateurs non paramétriques de la variance. Les deux premières variables explicatives, x_1 et x_2, sont dichotomiques (0 ou 1) et constantes à l'intérieur des UPÉ. La valeur de la variable x_1 est 1 dans la moitié des grappes : 1, 3,...,19, tandis que celle de x_2 est 1 dans trois grappes seulement : 9, 10 et 11. Les variables x_3 et x_4 sont toutes deux générées d'après une distribution normale standard. Elles diffèrent en ce sens que x_3 est générée d'après une loi normale multivariée dont la corrélation intra-grappe est de 0,5 dans les UPÉ, tandis que x_4 est générée à partir de distributions normales indépendantes.

La variable dépendante $y_{ij} = \beta' x_{ij} + \varepsilon_{ij}$, où $\beta = 0$ et les $\boldsymbol{\varepsilon}_i$ sont des variables aléatoires normales multivariées standard dont la corrélation intra-grappe est ρ. Nous utilisons deux autres valeurs de $\rho = 0$ et 1/3, qui correspondent, pour la moyenne d'échantillon, aux effets de plan $DEFF$ = 1 et 4, respectivement ($DEFF$=1+(m-1)ρ). Les résultats de la simulation de Monte Carlo sont basés sur 100 000 répétitions de **y** pour notre matrice fixe **X**. Les résultats pour ρ = 1/9 sont présentés dans Bell et McCaffrey (2002).

Nous avons évalué l'estimateur de la variance par les moindres carrés ordinaires (MCO), $s^2 l'(\mathbf{X}'\mathbf{X})^{-1} l$ et quatre estimateurs non paramétriques de la variance : l'estimateur par linéarisation standard donné par l'équation (2.2) pour $c = n/(n-1)$; l'estimateur par le jackknife donné par (2.3); l'estimateur par linéarisation à biais réduit; et la méthode de Kott (1996). La LBR et les ajustements de Kott sont fondés sur des corrélations intra-grappe de travail de $\rho = 0$.

Nous avons estimé les taux d'erreurs de première espèce pour huit variantes, basées sur 100 000 répétitions, du test de répétition de l'hypothèse nulle $\beta_k = 0$, pour k = 0 à 4. Chaque variante consiste à comparer une « statistique t » à une distribution t de référence. Pour les statistiques t fondées sur la linéarisation, le jackknife et la LBR, nous utilisons les valeurs critiques des distributions t pour un nombre de degrés de liberté égal à $(n-1) = 19$, ainsi qu'à l'approximation de Satterthwaite correspondante. Pour la méthode de Kott, nous utilisons le nombre de degrés de liberté proposé par ce dernier. Tous les calculs ont été réalisés en SAS.

4.1.2 Résultats des simulations

Le tableau 1 montre le biais de plusieurs estimateurs de la variance pour les cinq coefficients de régression (y compris l'ordonnée à l'origine) pour $\rho = 0$ et 1/3. Les variances MCO sont sans biais pour $\rho = 0$, mais fortement

biaisées pour ρ = 1/3. Pour les variables au niveau de l'UPÉ (y compris l'ordonnée à l'origine), les variances MCO sont sous-estimées d'environ un facteur 1/*DEFF*. De la même façon, le biais est plus faible, mais reste considérable pour x_3, la variable de niveau individuel pour laquelle la corrélation intra-grappe est grande. Le biais positif de la variance MCO de $\hat{\beta}_4$ résulte de la corrélation intra-grappe légèrement négative pour x_4.

Tableau 1. Biais des estimateurs de la variance (en pourcentage de la variance réelle).

Estimateur	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
	$\rho = 0$				
MCO	0,0	0,0	0,0	0,0	0,0
Linéarisation	-9,6	-13,2	-32,5	-13,3	-1,8
Jackknife	11,7	17,2	51,2	17,6	2,1
Kott (1996)	0,0	0,0	0,0	0,0	0,0
LBR	0,0	0,0	0,0	0,0	0,0
	ρ =1/3				
MCO	-75,8	-75,5	-76,2	-65,3	13,8
Linéarisation	-10,7	-14,8	-33,5	-19,9	-4,1
Jackknife	10,7	15,9	49,5	21,4	5,9
Kott (1996)	-1,2	-1,9	-1,5	-7,7	-2,3
LBR	-1,0	-1,5	-1,3	-2,1	0,4

Source : Bell et McCaffrey (2002)

Les estimateurs par linéarisation et par le jackknife présentent tous deux un biais important; ces biais sont relativement indépendants de ρ, mais de signes opposés. Pour chaque estimateur, la grandeur du biais varie fortement d'un coefficient à l'autre. Les biais les plus importants (en valeur absolue) sont observés pour $\hat{\beta}_2$, dont la valeur dépend principalement des données provenant de trois UPÉ. Vient ensuite, par ordre décroissant, le biais observé pour $\hat{\beta}_3$, suivi de près par ceux de $\hat{\beta}_1$ et $\hat{\beta}_0$.

La méthode de Kott (1996) et la LBR éliminent à dessein le biais pour ρ = 0. Pour ρ = 1/9 et 1/3, les deux méthodes réduisent spectaculairement la grandeur du biais comparativement à la linéarisation. S'il est pratiquement impossible de distinguer les deux méthodes pour les variables au niveau des UPÉ, celle de Kott (1996) donne des résultats nettement moins bons pour $\hat{\beta}_3$ et $\hat{\beta}_4$ avec des biais relatifs de –7,7 % et –2,3 %, contre –2,1 et 0,4 dans le cas de la LBR.

Le tableau 2 montre que le taux d'erreurs de première espèce pour la méthode de linéarisation standard avec (n-1) degrés de liberté excède systématiquement 5 % pour les deux valeurs de ρ. C'est pour $\hat{\beta}_2$ que les erreurs de première espèce sont les plus courantes, leur taux pouvant atteindre 16 %, mais elles sont également beaucoup trop fréquentes pour $\hat{\beta}_0$, $\hat{\beta}_1$ et $\hat{\beta}_3$, leur taux variant de 7,0 % à 8,8 %. L'importance de ce problème est fortement corrélée à la taille du biais de l'estimateur par linéarisation (voir le tableau 1). Le taux d'erreurs de première espèce est nettement plus faible, de 5,7 % à 6,4 %, pour les tests où le nombre de degrés de liberté correspond à l'approximation de Satterthwaite. Donc, l'utilisation de l'autre nombre de degrés de liberté améliore d'environ 30 % à 88 % le taux d'erreurs de première espèce.

Tableau 2. Taux d'erreurs de première espèce pour les tests de vérification de l'hypothèse nulle $\beta = 0$

Estimateur	Df	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
		$\rho = 0$				
Linéarisation	*n*-1	7,54	7,00	15,99	7,35	5,38
Linéarisation	Satt	5,75	6,45	6,33	6,28	5,18
Jackknife	*n*-1	5,01	3,92	7,58	4,52	5,02
Jackknife	Satt	3,80	3,43	1,41	3,26	4,77
Kott (1996)	Kott	5,11	5,08	4,85	4,76	5,07
LBR	*n*-1	6,28	5,37	11,25	5,90	5,21
LBR	Satt	4,73	4,86	3,12	4,72	5,00
		$\rho = 1/3$				
Linéarisation	*n*-1	8,10	7,28	16,39	8,79	5,66
Linéarisation	Satt	6,30	6,78	6,62	7,53	5,44
Jackknife	*n*-1	5,45	4,11	7,76	4,56	4,67
Jackknife	Satt	4,13	3,61	1,51	3,35	4,46
Kott (1996)	Kott	5,59	5,44	5,14	5,88	5,31
LBR	*n*-1	6,76	5,63	11,55	6,45	5,19
LBR	Satt	5,18	5,14	3,30	5,26	4,98

NOTA : Les entrées dont la valeur réelle est de 5,00 % ont une erreur-type de 0,07 %.
Source : Bell et McCaffrey (2002)

Le profil des probabilités d'erreur de première espèce est moins uniforme pour la méthode du jackknife. Avec (*n*-1) degrés de liberté, cette méthode a tendance à produire, pour $\hat{\beta}_1$ et $\hat{\beta}_3$, des résultats prudents en harmonie avec le biais positif de l'estimation de la variance par le jackknife. Par contre, la probabilité d'une erreur de première espèce est beaucoup trop élevée pour $\hat{\beta}_2$, et un peu trop élevée pour l'ordonnée à l'origine $\hat{\beta}_0$ lorsque $\rho = 1/3$. Cette situation semble due au fait que choisir (*n*-1) degrés de liberté pour la distribution *t* de référence compense parfois le biais de l'estimation de la variance par le jackknife. Le taux très faible d'erreurs de première espèce observé pour la méthode du jackknife avec le nombre de degrés de liberté de Satterthwaite appuie cette conclusion; un plus petit nombre de degrés de liberté combiné à un biais positif important produit des valeurs de test très prudentes.

La LBR avec (*n*-1) degrés de liberté donne des résultats nettement meilleurs que la linéarisation avec le même nombre de degrés de liberté. Comme la LBR est sans biais quand $\rho = 0$, la comparaison de la cinquième à la première ligne du tableau montre la réduction du taux d'erreurs de première espèce qui résulte de l'élimination du biais de linéarisation. Sauf pour $\hat{\beta}_4$, la LBR réduit le taux d'erreurs de première espèce dans une proportion allant de 45 % à 88 %. Cependant, la LBR avec (*n*-1) degrés de liberté continue de produire des résultats systématiquement libéraux, particulièrement pour $\hat{\beta}_2$. La comparaison des lignes 2 et 6 de chaque section du tableau montre l'effet relatif de la réduction du biais et de l'ajustement de Satterthwaite. Pour $\hat{\beta}_0$ et $\hat{\beta}_2$, le nombre de degrés de liberté est le facteur le plus important, tandis que pour $\hat{\beta}_1$ et $\hat{\beta}_3$, c'est le biais qui importe. Pour la LBR avec l'approximation de Satterthwaite, les résultats sont très bons, sauf pour $\hat{\beta}_2$, pour lequel le taux d'erreurs de première espèce tombe à environ 3 %.

Les tests fondés sur l'estimateur de Kott de 1996 donnent aussi de bons résultats. Pour presque tous les coefficients et pour les deux valeurs de ρ, le taux d'erreurs de première espèce s'approche de 5 %. Fait exception le test pour $\hat{\beta}_3$, lorsque $\rho = 1/3$, pour lequel le taux d'erreurs est de 5,88 % à cause du biais modéré de l'estimateur de la variance.

Nous avons également mené des études en simulation pour généraliser la LBR, et les propriétés souhaitables obtenues par les MCO semblent se retrouver dans ces autres modèles. Dans McCaffrey, Bell et Botts (2001), nous utilisons la même matrice de plan d'échantillonnage que Bell et McCaffrey (2002) pour étudier les propriétés des

estimateurs LBR pour les moindres carrés pondérés et généralisés. Pour les méthodes MCP et MCG, les erreurs-types de la LBR sont très légèrement biaisées lorsque la matrice des covariances de travail s'écarte de la matrice des covariances réelles des termes d'erreur. Pour les modèles linéaires généralisés (régression logistique), les résultats provisoires des études en simulation donnent à penser que l'inférence fondée sur les estimations LBR et les distributions de référence utilisant le nombre approximatif de degrés de liberté à la manière de Satterthwaite ont tendance à présenter des erreurs de première espèce proches de la valeur nominale sur une gamme de valeurs réelles pour les coefficients de régression et la corrélation intra-grappe. Toutefois, il faudrait poursuivre les études en simulation pour vérifier le caractère général de ces constatations provisoires.

4.3 Application : régression logistique pour l'intervention dans l'exemple de Partners in Care

Nous illustrons les méthodes décrites dans la présente communication à l'aide de données provenant de Partners in Care, une expérience longitudinale réalisée en vue d'évaluer l'effet des programmes d'« amélioration de la qualité » des soins prodigués aux personnes déprimées par les organismes de gestion intégrée des soins de santé (OGISS) (Wells *et coll.*, 2000). L'expérience consistait à suivre 1 356 patients, provenant de 43 cliniques de sept OGISS, chez qui on avait diagnostiqué une dépression en 1996-1997. Dans chacun des neuf blocs constitués, des ensembles d'une à quatre cliniques ont été répartis au hasard entre trois cellules expérimentales : soins habituels, programme d'amélioration de la qualité complété par des ressources pour le suivi du traitement médicamenteux et programme d'amélioration de la qualité complété par des ressources pour l'accès à la psychothérapie. Six OGISS représentent, chacun, un bloc unique et un OGISS a été scindé en trois blocs d'après la composition ethnique des cliniques. Dans les blocs comptant plus de trois cliniques, les ensembles de cliniques ont été combinés de façon à obtenir la meilleure concordance possible avec la taille prévue d'échantillon et les caractéristiques des patients. Pour des précisions supplémentaires, voir Wells *et coll.* (2000).

La réception de soins indiqués au cours des six mois précédant le premier suivi constitue un résultat particulièrement intéressant. La réception de soins indiqués était codée comme une variable dichotomique égale à un si le patient recevait le traitement thérapeutique ou médicamenteux pertinent, et à zéro dans le cas contraire (Wells *et coll.*, 2000). Nous présentons les résultats d'un modèle de régression logistique pour les soins indiqués pour 1 143 patients lors du suivi après six mois. Comme dans Wells *et coll.* (2000), la variable indépendante présentant le plus grand intérêt est un indicateur d'intervention qui estime l'effet combiné du traitement médicamenteux et de la thérapie par opposition aux soins habituels. Notre régression diffère de la leur parce que nous ne corrigeons pas la non-réponse par pondération et n'imputons pas de données pour remplacer les valeurs manquantes de la variable étudiée, mais les résultats pour l'effet d'intervention concordent raisonnablement.

Comme les patients provenant d'une même clinique pourraient présenter des résultats comparables, les erreurs-types par la régression logistique pourraient facilement être trop faibles—particulièrement pour les variables au niveau de l'UPÉ, comme l'intervention. Nous comparons l'estimateur par linéarisation à l'estimateur LBR donné dans la section 3, $v_{LBR,MLG}$, en utilisant les matrices de rajustement données dans l'équation 3.7.

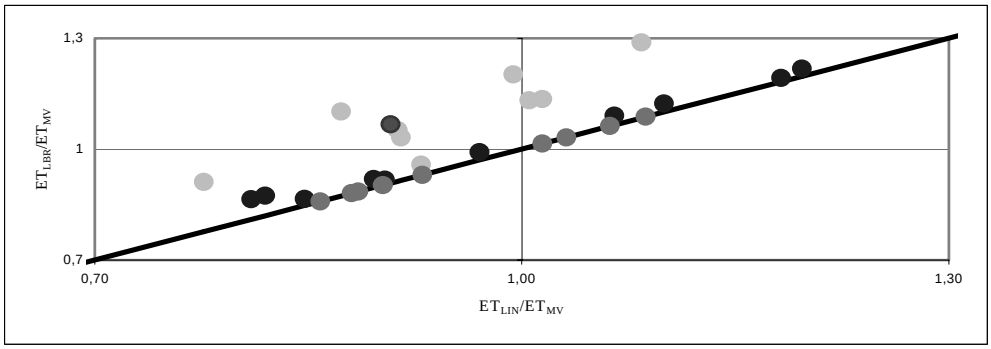

Figure 1. Ratio des ET_{LBR} aux ET_{MV} par opposition à celui des ET_{LIN} aux ET_{MV} pour les coefficients de modèle suivants : soins indiqués, intervention (rouge), autres variables au niveau des grappes (rose), caractéristiques démographiques (bleu) et niveau de santé de référence (brun).

Pour les patients provenant d'une même clinique, la réception de soins indiqués était essentiellement non corrélée. Par la méthode EEG de Liang et Zeger (1986), nous calculons que la corrélation intra-clinique des erreurs est égale à –0,0014, valeur qui concorde facilement avec une valeur réelle de 0. Donc, les erreurs-types fondées sur le maximum de vraisemblance (MV), qui sont précises pour un échantillon de cette taille, devraient aussi être exactes, et il n'y a aucune raison de s'attendre à ce que toute erreur-type correcte soit nettement plus faible que celle obtenue par la régression logistique. Toutefois, les erreurs-types par linéarisation sont inférieures aux erreurs-types par MV pour 18 des 29 coefficients et pour 7 des 10 coefficients dans le cas des variables au niveau des grappes. On le constate dans la figure 1. L'axe horizontal représente graphiquement le ratio de la linéarisation aux erreurs-types (ET) par MV, et bon nombre des points se trouvent à gauche de l'axe vertical situé à 1,00, où les estimations MV sont égales à celles de la linéarisation. La figure montre également la variabilité considérable des estimateurs par linéarisation, qui est manifeste dans la fourchette des ratios, qui va d'environ 0,8 à 1,2.

L'estimateur LBR donne des résultats très supérieurs à ceux de l'estimateur par linéarisation classique. Le ratio des estimateurs LBR aux estimateurs MV est représenté graphiquement sur l'axe vertical de la figure 1, et 8 des 10 points roses se situent au-dessus de l'axe horizontal à 1,00. Tous les points se situent au-dessus de 45°, ce qui indique que les estimations LBR sont supérieures aux estimations par linéarisation pour chaque coefficient.

RÉFÉRENCES

Bell R.M. et McCaffrey, D.F., (2002) "Bias Reduction in Standard Errors for Linear Regression with Multi-Stage Samples," *Survey Methodology*, forthcoming.

Binder, D. (1983), "On the Variance of Asymptotically Normal Estimators from Complex Surveys," *International Statistical Review*, 51, pp. 279-292.

Huzurbazar, S. (1999), "Practical Saddlepoint Approximations," *American Statistician*, 53, pp. 225-232.

Kott, P. S. (1994), "A Hypothesis Test of Linear Regression Coefficients with Survey Data," *Survey Methodology*, 20, pp. 159-64.

Kott, P. S. (1996), "Linear Regression in the Face of Specification Error: Model-Based Exploration of Randomization-Based Techniques," *Statistical Society of Canada Proceedings of the Survey Methods Section*, pp. 39-47.

Liang, K-,Y., et Zeger, S. L. (1986) "Longitudinal Data Analysis Using Generalized Linear Model," *Biometrika*, 73, pp. 13-22.

Mancl, L. A. et DeRouen T. A. (2001) "A Covariance Estimator for GEE with Improved Small-Sample Properties," *Biometrics*, 57, pp. 126-134.

McCaffrey, D. F., Bell R. M. et Botts, C. H. (2001), "Generalizations of Bias Reduced Linearization," *Proceeding of the Annual Meeting of the American Statistical Association, August 5-9, 2001*.

Pan, W. et Wall, M. (2001) "Small-sample Adjustments in Using the Sandwich Variance Estimator in Generalized Estimating Equations," *Statistics in Medicine*, 21, pp. 1429-1441.

Rust, K. F., et Rao, J. N. K. (1996), "Variance Estimation for Complex Surveys Using Replication Techniques," *Statistical Methods in Medical Research*, 5, pp. 283-310.

Satterthwaite, F. (1946), "An Approximate Distribution of Estimates of Variance Components," *Biometrics*, 2, pp. 110-114.

Shah, B. V., Holt, M. M., et Folsom, R. E., (1977), "Inference About Regression Models from Survey Data," *Bulletin of the International Statistical Institute*, 41, pp. 43-57.

Skinner, C. J., (1989), "Domain Means, Regression and Multivariate Analyses," in *Analysis of Complex Surveys*, eds. C. J. Skinner, D. Holt, and T. M. F. Smith, New York: Wiley, pp. 59-88.

Wells, K. B, Sherbourne, C., Schoenbaum, M., Duan, N., Meredith, L., Unutzer, J., Miranda, J., Carney, M., et Rubenstein, L. V., (2000), "Impact of Disseminating Quality Improvement Programs for Depression in Managed Primary Care: A Randomized Controlled Trial," *JAMA*, 283, pp. 212-220.