

**REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR
ET DE LA RECHERCHE SCIENTIFIQUE**

UNIVERSITE MENTOURI CONSTANTINE

FACULTE DES SCIENCES EXACTES

DEPARTEMENT DE MATHEMATIQUES

**N° d'ordre :...../ TE / 2006
série : / MAT / 2006**

**THESE PRESENTEE POUR L'OBTENTION DU DIPLOME DE DOCTORAT
D'ETAT EN MATHEMATIQUES.**

**«ETUDE DE DONNEES LONGITUDINALES ET DE MESURES REPETEES
DESEQUILIBREES DANS UN MODELE MIXTE : DIFFERENTES APPROCHES»**

OPTION : STATISTIQUES APPLIQUEES

Par

Ahsène Lanani

**Soutenue le : 19/12/2006
composé de :**

devant le jury

**Mr Z. MOHDEB Pr U. Mentouri Constantine Président
Mme F. MESSACI M.C U. Mentouri Constantine Rapporteur
Mr L. ABBAOUI Pr U. de Sétif Examineur
Mme N. SEDDIK AMEUR M.C. U. de Annaba Examineur
Mr F. L. RAHMANI M.C. U. Mentouri Constantine Examineur
Mr M. REGHIOUA M.C. E.N.S de Constantine Examineur**

Remerciements

Je voudrais tout d'abord exprimer ma profonde reconnaissance et remercier mon directeur de thèse le Professeur James K.Lindsey, Professeur à Liège et au L.U.C en Belgique, pour avoir accepté de m'encadrer malgré les difficultés de l'éloignement.

Je voudrais remercier le Professeur Geert Molenberghs, Directeur du laboratoire de Biostatistiques au L.U.C en Belgique, qui m'a fait un grand honneur en acceptant de consacrer une partie de son temps précieux pour juger cette thèse.

Je tiens à remercier Madame Fatiha Messaci Maître de conférences à l'université de Constantine pour avoir accepté d'être mon Co-encadreur puis mon encadreur ainsi que pour ses encouragements et sa permanente disponibilité.

Mes remerciements vont également :

A Monsieur Zoheir Mohdeb Professeur à l'université de Constantine pour avoir accepté de présider ce jury.

A Monsieur Lyazid Abbaoui Professeur à l'université de Sétif pour avoir accepté d'être membre du jury.

A Madame Nacéra Seddik Ameer Maître de conférences à l'université de Annaba pour avoir accepté d'être membre du jury.

A Monsieur Fouad Lazhar Rahmani Maître de conférences à l'université de Constantine pour avoir accepté d'être membre du jury.

A Monsieur Mohammed Reghioua Maître de conférences à LE.N.S de Constantine pour avoir accepté d'être membre du jury.

Je tiens aussi à remercier :

Madame Vigeannel Odile Larive, responsable de la bibliothèque de Jussieu-Chevaleret à l'université de Paris5, pour sa disponibilité permanente et son accueil.

Monsieur le Professeur Pierre Cartigny, Directeur du Laboratoire LASB, unité de biométrie à l'ENSAM-INRA Montpellier pour son chaleureux accueil et pour sa disponibilité, ainsi qu'à tous les membres de son équipe, qui ont essayé de près ou de loin à rendre mon séjour agréable.

Le Professeur Hélène Jacqmin-Gadda de l'université de Bordeaux pour sa documentation.

Le professeur Deborah Crane tufts-nemc (New England Medical Center USA) pour le temps consacré à me procurer et à m'envoyer la documentation demandée.

Enfin je tiens à remercier Monsieur A.Ouakas, qui était doctorant à l'ULB en Belgique pour le traitement des données.

Résumé :

Les modèles linéaires mixtes sont largement utilisés dans plusieurs domaines (surtout en génétique quantitative; en médecine; en biologie ; ou en écologie). Ces modèles s'avèrent s'adapter convenablement aux données longitudinales et aux mesures répétées équilibrées ou déséquilibrées, également en présence de données manquantes.

Cette thèse présente l'estimation des paramètres d'un modèle linéaire mixte, ainsi que ceux générant sa matrice de variance-covariance, d'une part; d'autre part, la recherche d'un choix parcimonieux de la structure de cette matrice a été concrétisée par l'application d'une première approche à un exemple de données dentaires. L'approche considérée est celle du maximum de vraisemblance (ML) ou du maximum de vraisemblance restreinte (REML). Les équations normales obtenues sont souvent non linéaires, d'où l'utilisation pour leur résolution de procédés itératifs ou algorithmes, tels que, Newton-Raphson; l'algorithme EM ou celui des scores de Fisher.

Nous introduisons ensuite une deuxième approche; c'est celle des modèles de quasi-vraisemblance et de quasi-vraisemblance étendue, d'où l'utilisation des modèles linéaires généralisés (GLM). Une extension de ces modèles, nous a permis d'introduire les modèles marginaux. Ces derniers, utilisent les équations d'estimation généralisées (GEE).

Enfin, nous terminons ce travail par la confrontation de cette approche à la première, celle des modèles à effets aléatoires utilisant le maximum de vraisemblance (ML) ou le maximum de vraisemblance restreinte (RE ML).

Mots clés : Algorithme des scores de Fisher; EM algorithme; Estimation des composantes de la variance; GEE; Modèles linéaires mixtes; Newton-Raphson; Structure de Variance-Covariance.

Summary :

The mixed linear models are largely used in several fields (especially in quantitative genetics; medicine; biology; ecology...). These models prove to adapt suitably to the longitudinal data and repeated balanced or unbalanced measurements, also in the presence of missing data.

This thesis presents, on the one hand, the estimate of the parameters of a mixed linear model, like those generating its variance-covariance matrix; on the other hand, the search for a parsimonious choice of the structure of this matrix was concretized by the application of a first approach to an example of dental data. The considered approach is the maximum likelihood (ML) or restricted maximum likelihood (REML) applied to mixed linear models. The obtained normal equations are often nonlinear, from where the use of iterative processes or algorithms for their resolution, such as, Newton-Raphson; the EM algorithm or the Fisher scoring ones.

We introduce then a second approach; it is that of the quasi-likelihood models and extended quasi-likelihood, from where the use of the generalized linear models (GLM). An extension of these models, enabled us to introduce the marginal models. The latter, use the generalized estimating equations (GEE).

Lastly, we complete this work by the confrontation of this approach to the first, that of the random effects models using the maximum likelihood (ML) or restricted maximum likelihood (RE ML).

key Words : EM algorithm; Estimate of the variance components; Fisher Scoring Algorithm; GEE; Mixed Linear Models; Newton-Raphson; Variance-Covariance Structure.

Table des matières

| | |
|---|-----------|
| Introduction | 5 |
| 1 Estimation des composantes de la variance dans un modèle à effets mixtes. | 9 |
| 1.1 Notions préliminaires | 9 |
| 1.2 Modèles | 11 |
| 1.2.1 Courbes de croissance | 11 |
| 1.2.2 Modèles linéaires mixtes | 13 |
| 1.2.3 Estimation des paramètres du modèle et ceux de sa matrice de variance-covariance. | 15 |
| 1.2.4 Estimation | 18 |
| 1.2.5 Algorithme EM | 21 |
| 2 Modélisation de la matrice de variance-covariance en modèles linéaires mixtes. | 31 |
| 2.1 Introduction | 31 |
| 2.2 Modèle | 32 |

| | | |
|------------|---|-----------|
| 2.3 | Algorithmes pour le maximum de vraisemblance. | 33 |
| 2.3.1 | La fonction vraisemblance et les équations d'estimation. | 33 |
| 2.3.2 | Algorithmes de Newton-Raphson et 'Fisher-scoring' : | 34 |
| 2.3.3 | Fonction de vraisemblance restreinte (résiduelle) | 36 |
| 2.4 | Autres Modèles | 36 |
| 2.4.1 | Modèle AR(1)RE | 36 |
| 2.4.2 | Modèle à amortissement exponentiel | 38 |
| 2.5 | modèles linéaires mixtes | 41 |
| 2.5.1 | Introduction | 41 |
| 2.5.2 | Modèle | 43 |
| 2.5.3 | Structures de la covariance | 45 |
| 2.5.4 | Exemple | 49 |
| 2.5.5 | Ajustement | 51 |
| 2.5.6 | Comparaison des différentes structures en utilisant les résultats des ajustements. | 53 |
| 3 | Etude de la fonction variance pour des données longitudinales utilisant une extension des modèles linéaires généralisés ('GLM'). | 55 |
| 3.1 | Définition d'un modèle linéaire généralisé | 55 |
| 3.2 | Extensions des modèles linéaires généralisés. | 56 |
| 3.2.1 | Introduction | 56 |
| 3.2.2 | Quasi-vraisemblance; Quasi-vraisemblance étendue; fonction variance et son estimation. | 56 |

| | |
|--|-----------|
| | 3 |
| 4 Modèles marginaux - équations d'estimation généralisées | 63 |
| 4.1 Introduction | 63 |
| 4.2 Analyse de données longitudinales utilisant les 'GLM' | 65 |
| 4.2.1 Introduction | 65 |
| 4.2.2 Notations et modèle | 65 |
| 4.3 Equations d'estimation | 66 |
| 4.3.1 Equation d'estimation et indépendance | 66 |
| 4.3.2 Equations d'estimation généralisées | 67 |
| 5 Les Modèles à effets aléatoires versus les modèles marginaux. | 73 |
| 5.1 Introduction | 73 |
| 5.2 Modèles SS et PA | 74 |
| 5.2.1 Modèle SS | 74 |
| 5.2.2 Modèle PA | 75 |
| 5.3 Equations d'estimation généralisées pour les modèles PA et SS | 76 |
| 5.3.1 Estimation des coefficients de regression | 76 |
| 5.3.2 Estimation de D et de ϕ | 80 |
| Conclusion et Perspectives | 86 |
| Références | 88 |
| Annexe A | 92 |
| Annexe B | 96 |

Introduction

La fréquence des données groupées en biologie, en épidémiologie, dans les problèmes de santé ou ceux du vivant en général est à l'origine de l'intérêt croissant des biostatisticiens pour les méthodes d'analyses statistiques adaptées surtout aux données corrélées. Le choix d'un modèle statistique parmi une famille de modèles et d'une méthode d'analyse parmi tant d'autres, n'est pas une tâche facile. Ceci dépend du domaine d'application, de l'objectif visé, de la structure de l'échantillon ou du degré de dépendance à l'intérieur des groupes d'individus.

L'objectif de cette thèse est l'estimation des paramètres d'un modèle linéaire mixte ainsi que ceux générant sa matrice de variance-covariance d'une part en utilisant différentes méthodes ou différents procédés itératifs; d'autre part, mettre en évidence l'importance de cette matrice, le choix de sa structure qui donnera le " meilleur " modèle; son effet sur les paramètres de régression dans le contexte des tests d'hypothèses. Finalement, chercher son importance dans le cadre des équations d'estimation généralisées (GEE) par rapport à la matrice de corrélation de travail dans le cadre des modèles marginaux; autrement dit, nous sommes en présence de la " confrontation " ou de la " compétition " entre deux grandes familles de modèles; les modèles linéaires mixtes et les modèles marginaux.

Dans la première partie de cette thèse, nous rappelons la méthode d'analyse multivariée (" Multivariate analysis "), nous décrirons les modèles linéaires mixtes, puis on passe à l'estimation des paramètres d'un modèle linéaire mixte général ainsi que

ceux générant sa matrice de variance-covariance (qui n'est pas forcément une combinaison linéaire de matrices connues). La méthode d'estimation utilisée est le maximum de vraisemblance (ML) ou le maximum de vraisemblance restreinte (REML), en utilisant le EM algorithme.

Dans le deuxième chapitre, nous nous intéressons à la matrice de variance-covariance d'un modèle linéaire mixte ; aux différentes structures possibles pour celle-ci, afin d'en choisir " le meilleur modèle " ; ainsi qu'à l'influence de cette matrice sur les effets fixes. La méthode d'estimation utilisée est le ML ou le REML utilisant les deux algorithmes Newton-Raphson et Fisher-scoring. Un exemple classique de données de croissance a été étudié afin d'atteindre notre but.

Le troisième chapitre est consacré aux rappels sur les modèles linéaires généralisés, sur des extensions de ceux-ci ; cela a permis d'introduire les modèles de quasi-vraisemblance (QL) et de quasi-vraisemblance étendue. Nous nous sommes intéressés par la suite à la fonction variance de ces modèles et à son estimation.

Le chapitre quatre, est une extension ou prolongement du précédent. Dans cette partie, nous introduisons les modèles marginaux. Ceci fait l'objet de la deuxième approche d'estimation qui est celle de la résolution d'équations, appelées équations d'estimation généralisées (GEE). Cette méthode se subdivise en deux autres méthodes (GEE1) et (GEE2). La première est appliquée lorsque l'on s'intéresse uniquement à l'estimation de l'effet fixe et la corrélation est vue comme un paramètre de nuisance. La seconde est pour sa part appliquée lorsque l'on s'intéresse aux deux quantités.

Enfin le chapitre cinq concerne la comparaison entre les deux familles de modèles (les modèles linéaires mixtes versus les modèles marginaux).

Chapitre 1

Estimation des composantes de la variance dans un modèle à effets mixtes.

1.1 Notions préliminaires

Les études longitudinales sont des études où chaque individu est observé en deux occasions ou plus sur des périodes étendues (en tenant compte du temps). Pour ce qui est des mesures répétées, ce sont des mesures qui sont prises durant une période d'étude qui est très courte (en tenant compte des conditions expérimentales), (Ware 1985).

Le livre de Diggle ; Liang et Zeger 1994 (1ière édition),2002 (seconde édition), est un ouvrage complet traitant de l'analyse de données longitudinales.

Pour ce qui est des mesures répétées un ouvrage complet est celui de Lindsey 1993.

Dans les dispositifs longitudinaux, on s'intéresse à l'étude du comportement (changement d'un caractère) d'un individu à travers le temps, ainsi qu'aux effets des covariables sur les plans des réponses de l'individu. Les mesures faites peuvent être par exemple, la pression du sang; le niveau du cholestérol; le volume respiratoire; le glucose du serum;...etc.

Notons d'autre part, que la plupart des modèles stochastiques de mesures en séries peuvent être classés comme des modèles multivariés et complets (il n'y a pas de données manquantes) ou comme des modèles à effets aléatoires à plusieurs niveaux.

Dans les premiers, on suppose que chaque vecteur des observations suit une loi gaussienne et la matrice de variance-covariance de ce vecteur a une forme générale ou sans structure (" Unstructured ").

Cette approche que l'on peut qualifier par approche multivariée (" Multivariate analysis ") devient compliquée surtout quand la dimension de cette matrice de variance-covariance est grande ou quand les données sont fortement déséquilibrées (nombres d'observations différents chez les individus, par exemple), aussi quand on a des données manquantes. Il faut ajouter à cela que les modèles complets multivariés ne permettent pas l'estimation des effets aléatoires des individus.

Pour ce qui est des modèles aléatoires à deux niveaux, la plupart d'entre eux peuvent être décrits soit comme des modèles de croissance (étude de la variation intra-individus et estimation des caractéristiques des différentes populations), ou

comme des modèles de mesures répétées (effets individus supposés constants mais il y a une variation des effets expérimentaux).

Notons que les modèles à effets aléatoires à deux niveaux sont utilisés pour l'identification des paramètres des individus et celle des caractéristiques de population et fréquemment leur forme aboutit à une situation déséquilibrée. Ce genre de modèles, sont les modèles linéaires mixtes. Nous allons insister dans notre travail sur ces modèles. Ces modèles linéaires mixtes sont intéressants, car d'une part les données traitées ne doivent pas être forcement équilibrées ; d'autre part, on a une modélisation explicite des effets intra et inter individus.

Dans cette première partie de notre travail, nous allons considérer un modèle linéaire mixte et essayer d'estimer les paramètres de ce modèle ; ainsi que ceux générant sa matrice de variance-covariance.

1.2 Modèles

1.2.1 Courbes de croissance

Formulation de Laird et al. 1987. Soit le modèle linéaire mixte :

$$y_i = X_i\alpha + Z_ib_i + e_i, \quad i = 1, \dots, m \quad (1.1)$$

y_i : observations faites sur le ième individu.

α : vecteur inconnu des paramètres de population, de dimension $(p \times 1)$.

X_i : matrice du dispositif liant y_i à α , connue, de dimension $(n_i \times p)$ avec n_i le nombre d'observations de l'individu i .

b_i : vecteur inconnu des effets individus, aléatoire de dimension $(k \times 1)$; variables supposées indépendantes et suivent $N(0, D_{k \times k})$.

Z_i : matrice du dispositif liant y_i à b_i , connue, de dimension $(n_i \times k)$.

Les erreurs e_i sont des variables indépendantes et suivent $N(0, \sigma^2 I'_{n_i \times n_i})$.

Ce modèle mixte peut être appliqué pour des courbes de croissance. Pour cela, il faut spécifier les caractéristiques de la courbe de croissance de chaque individu puis modéliser les paramètres des courbes de croissance comme des fonctions linéaires des caractéristiques de l'individu, comme nous allons le faire.

Spécifiquement pour les courbes de croissance, on suppose que :

$$y_i = Z_i \beta_i + e_i, \quad i = 1, \dots, m \quad (1.2)$$

y_i , Z_i et e_i définis précédemment.

La composante $Z_i \beta_i$ définit la courbe de croissance de l'individu. Les paramètres aléatoires β_i sont uniques pour chaque individu. Ces variables β_i sont indépendantes et suivent $N(A_i \alpha, D)$.

α et D définis précédemment ; A_i de dimension $(q \times p)$, c'est la matrice du dispositif.

Donc :

$$E(y_i) = Z_i A_i \alpha = X_i \alpha$$

et

$$\text{Var}(y_i) = \sigma^2 I'_{ni} + Z_i D Z_i^T$$

où

$$X_i = Z_i A_i.$$

Il s'ensuit que $\beta_i = A_i \alpha + b_i$. Donc b_i peut être vu comme un vecteur des résidus représentant la déviation des paramètres de croissance des individus par rapport aux paramètres de croissance de la population, contrairement aux résidus ordinaires définis par $r_i = y_i - X_i \alpha$.

Des exemples d'études de courbes de croissance ont été traités par Potthoff et Roy 1964 ; Rao 1965 ; Grizzle et Allen 1969 ; un autre exemple a été traité par Rochon et Helms 1989, dans le contexte des modèles ARMA.

1.2.2 Modèles linéaires mixtes

Les modèles que nous allons utiliser seront du type :

$$y_i = X_i \alpha + \varepsilon_i, \quad i = 1, \dots, m \quad (1.3)$$

Qui peuvent aussi s'écrire :

$$y_i = X_i \alpha + Z_i b_i + e_i, \quad i = 1, \dots, m \quad (1.4)$$

En prenant :

$\varepsilon_i = Z_i b_i + e_i, \quad i = 1, \dots, m.$ dans le modèle donné par la formule (1.3) :

y_i : observations faites sur le i ème individu.

α : vecteur inconnu des paramètres de population, de dimension $(p \times 1)$.

X_i : matrice du dispositif liant y_i à α , connue, de dimension $(n_i \times p)$ avec n_i le nombre d'observations de l'individu i .

b_i : vecteur inconnu des effets individus, aléatoire de dimension $(k \times 1)$.

Z_i : matrice du dispositif liant y_i à b_i , connue, de dimension $(n_i \times k)$.

Ces modèles qui sont des modèles linéaires mixtes sont souvent appelés modèles à deux niveaux (" two-stage models ").

Un exemple est l'étude appliquée à des données longitudinales, faite par Laird et Ware 1982.

Etape 1 :

$$\forall i, y_i = X_i \alpha + Z_i b_i + e_i, \quad i = 1, \dots, m$$

Les variables e_i sont indépendantes et suivent $N(0, R_i)$, α et b_i fixes.

Dans cette étape, on est intéressé par la variation intra-individus (" within individual variation ") et qui est formulée par R_i .

Etape 2 :

Les b_i sont supposées indépendantes et suivent $N(0, D_{k \times k})$, α fixes.

Dans cette étape, on est intéressé par la variation inter-individus (" between individual variation ") et qui est formulée par D .

Les y_i sont indépendantes et suivent $N(X_i \alpha, Z_i D Z_i^T + R_i)$.

Pour simplifier les calculs, on peut prendre dans un premier temps :

$$R_i = \sigma^2 I'_{n_i \times n_i} \dots \dots \dots \text{ (erreurs homogènes)}.$$

Par la suite, on peut généraliser en prenant R_i et D quelconques, générées par un paramètre inconnu θ

Soit donc $\theta = (\theta)_{q \times 1}$ paramètre tel que : $R_i = R_i(\theta)$ et $D = D(\theta)$.

Une généralisation de ce modèle linéaire mixte a été faite par Jones et Boadi-Boateng 1991, où R_i n'est plus nécessairement diagonale ; e_i est décomposée en une composante autoregressive et une erreur de mesure.

1.2.3 Estimation des paramètres du modèle et ceux de sa matrice de variance-covariance.

Pour estimer α , b_i et θ , on a deux méthodes

lière Méthode :

Pour estimer α et θ , on a une approche basée sur le maximum de vraisemblance (ML), (Foulley 1991 p75).

Ayant $f(y/b; \alpha, \theta)$ dans la première étape et $f(b; \theta)$ dans la deuxième, d'où $f(y; \theta, \alpha)$; alors on maximise la loi marginale de y et qui est donnée par :

$$f(y; \theta, \alpha) = \int f(y/b; \alpha, \theta) f(b) db$$

Pour estimer b_i , on applique une méthode qui consiste à trouver un estimateur dont l'erreur du carré moyen (" MSE ") est la plus petite de la combinaison linéaire $\lambda'_1 \alpha + \lambda'_2 b$, avec $\lambda'_1 \alpha$ supposé estimable (version du théorème de Gauss- Markov étendue aux effets aléatoires, Harville1976 ; approche revue par Harville 1977).

2ième Méthode :

Vu que les estimateurs des composantes de la variance sont biaisés dans la méthode du ML, on propose une alternative basée sur une formulation bayésienne. On estime θ en maximisant la vraisemblance marginale de y après avoir intégré sur α et b_i . On se donne en plus une loi a priori pour α (typiquement, on n'a aucune information sur α , paramètre modélisant la population). on maximise donc :

$$f(y; \theta) = \int_{R^p} \int_{R^k} f(y/b; \alpha; \theta) f(b; \theta) f(\alpha) db d\alpha \quad (1.5)$$

Approche considérée par Harville (1974 ;1976) et Dempster et al.1981. Cette méthode donne un estimateur du maximum de vraisemblance restreinte (" REML"). On rappelle l'équivalence qu'on a entre l'estimateur REML et celui de Bayes (Foulley 1991 p90).

Les estimateurs de Bayes empiriques de α et de b_i sont les moyennes estimées des distributions conditionnelles. Ces estimateurs sont données par les formules (1.20) et (1.21).

L'estimateur REML est aussi obtenu (Patterson et Thompson 1971) en maximisant la vraisemblance de θ basée non pas sur y mais sur $n - p$ contrastes d'erreur (Harville 1977), choisis tels que :

$$E(u^T y) = 0 \iff u^T y = 0$$

Exemple : choix de contrastes d'erreur pour trouver l'estimateur REML

Cette approche est basée sur l'écriture de la vraisemblance d'une combinaison linéaire des données Ly libre des effets fixes; qui est donc telle que : $LX = 0$. On obtient une telle combinaison linéaire en travaillant sur un vecteur r des résidus des données ajustées par leur estimateur des moindres carrés, soit

$$r = y - X\alpha^0$$

où :

α^0 est l'estimateur des moindres carrés ordinaires de α et qui est solution de :

$$X^T X \alpha^0 = X^T y$$

Si l'on pose $P_X = X(X^T X)^- X^T$; A^- étant un inverse généralisé de A et $M_X = I - X(X^T X)^- X^T = I - P_X$. On a conformément aux propriétés des projecteurs P_X et M_X :

$$X\alpha^0 = P_X y$$

$$r = M_X y$$

ou encore :

$$P_X X = X$$

$$M_X = 0$$

On peut donc choisir comme combinaison linéaire L la matrice M_X définie par $(I - P_X)$ appelée aussi projecteur sur l'espace orthogonal à l'espace $C(X)$ engendré par les colonnes de X . Comme $\text{rang}M_X = N - \text{rang}X$, avec $N = \dim y$, il suffit de considérer $N - \text{rang}X$ éléments du vecteur r ou combinaisons indépendantes de celui-ci. Ce sous vecteur r^* des résidus ou contrastes d'erreur (Harville, 1977) étant défini, il reste à exprimer sa vraisemblance (elle est donnée dans Searle, 1979, et qu'on retrouve plus loin, Chapitre 2).

1.2.4 Estimation

1.2.4.1 Estimation de α et de b_i en supposant la variance connue :

$$\text{Var}(y_i) = V_i = Z_i D Z_i^T + R_i, \quad i = 1, \dots, m \quad (1.6)$$

soit :

$$W_i = [V_i]^{-1}$$

$$\hat{\alpha} = (\sum_i X_i^T W_i X_i)^{-1} (\sum_i X_i^T W_i y_i) \quad (1.7)$$

L'estimateur de α est celui maximisant la vraisemblance basée sur les distributions marginales des données et il est à variance minimum (souvent appelé "BLUE").

Pour sa part, l'estimateur de b_i est donné par :

$$\hat{b}_i = DZ_i^T W_i (y_i - X_i \hat{\alpha}) \quad (1.8)$$

\hat{b}_i n'est pas l'estimateur du maximum de vraisemblance mais l'estimateur de Bayes empirique

(souvent appelé "BLUP").

$$\hat{b}_i = E(b_i / y_i, \hat{\alpha}, \theta) \quad (1.9)$$

Puisque $\hat{\alpha}$ et \hat{b}_i sont des fonctions linéaires de y , les expressions de leurs erreurs standards peuvent être facilement calculées et sont données par :

$$Var(\hat{\alpha}) = (\sum_i X_i^T W_i X_i)^{-1}$$

et

$$Var(\hat{b}_i) = DZ_i^T \left\{ W_i - W_i X_i (\sum_i X_i^T W_i X_i)^{-1} X_i^T W_i \right\} Z_i D$$

Remarque :

On peut aussi avoir $\hat{\alpha}$ et \hat{b}_i en résolvant le système de Henderson (1959), qui est

repris dans Searle (1971) :

$$\begin{bmatrix} X_i^T R_i^{-1} X_i & X_i^T R_i^{-1} Z_i \\ Z_i^T R_i^{-1} X_i & Z_i^T R_i^{-1} Z_i + D^{-1} \end{bmatrix} \begin{bmatrix} \hat{\alpha} \\ \hat{b}_i \end{bmatrix} = \begin{bmatrix} X_i^T R_i^{-1} y_i \\ Z_i^T R_i^{-1} y_i \end{bmatrix} \quad (1.10)$$

Ce système est obtenu en dérivant par rapport à α et à b_i la densité conjointe de y_i et b_i et qui est donnée par :

$$f(y_i, b_i) = g(y_i/b_i)h(b_i) \quad (1.11)$$

$$f(y_i, b_i) =$$

$$C \exp \left\{ -1/2(y_i - X_i\alpha - Z_ib_i)^T R_i^{-1}(y_i - X_i\alpha - Z_ib_i) \right\} \exp -1/2b_i^T D_i^{-1}b_i$$

où C est une constante de normalisation contenant notamment les déterminants de R_i^{-1} et de D .

Récemment, un algorithme résolvant ce système ainsi que l'estimation des composantes de la matrice de variance-covariance a été donné par Witkovsky (2002), utilisant le logiciel Matlab.

1.2.4.2 Estimation de α et de b_i avec la variance inconnue :

On estime la variance ou le paramètre θ qui la génère. On a deux estimateurs pour θ , l'estimateur du maximum de vraisemblance (ML) noté par θ_M et l'estimateur du maximum de vraisemblance restreinte (ou résiduelle) noté par θ_R (ou θ_{REML}). Pour calculer ces deux estimateurs, on applique le EM algorithme (Dempster et al.1977; Dempster et al.1981).

Remarque :

Dans cette partie, le EM algorithme n'est pas appliqué à des données incomplètes mais plutôt pour estimer des paramètres aléatoires non observables.

1.2.5 Algorithme EM

Soit X une variable aléatoire de densité $f(x/\theta)$ où θ est un paramètre inconnu. Supposons que X ne soit pas complètement observée; c'est-à-dire que l'on observe une partie Y de X . Soit $Y = Y(X)$, une variable aléatoire de densité $g(y/\theta)$. Soit $t(x)$ un vecteur de statistiques exhaustives pour θ . L'algorithme EM (E pour espérance et M pour maximisation) a pour but de trouver la valeur de θ qui maximise la vraisemblance $g(y/\theta)$ étant donnée une valeur de y . Cette maximisation (équations normales) donne l'équation suivante :

$$E(t(x)/\theta) = E(t(x)/y, \theta) \quad (1.12)$$

L'algorithme EM utilise deux étapes pour résoudre cette équation en θ .

1ère étape :

E-Step : On calcule la quantité :

$$t(x) = E(t(x)/y, \theta). \quad (1.13)$$

2ième étape :

M-Step : On résout l'équation en θ :

$$E(t(x)/\theta) = t(x). \quad (1.14)$$

En d'autres termes, dans le E-Step, étant donnée une valeur initiale pour θ ; $\theta^{(p)}$ sa valeur à l'étape (p) . A l'étape $(p+1)$, on calcule la valeur de $t(x)$ notée par $t^{(p)}$ et qui est donnée par :

$$t^{(p)} = E(t(x/y, \theta^{(p)}))$$

Dans le M-Step, étant donnée la valeur de $t^{(p)}$ calculée dans le E-Step, on résout l'équation en $\theta^{(p+1)}$ et qui est donnée par :

$$E(t(x)/\theta^{(p+1)}) = t^{(p)}$$

1.2.5.1 Utilisation du EM algorithme pour calculer l'estimateur du maximum de vraisemblance.

Pour mettre le problème des données longitudinales de cette étude dans le contexte des données incomplètes, on considère que e_i et b_i sont des observations (en plus de y_i).

Les estimateurs des composantes de θ sont basés sur des formules quadratiques de b_i et de e_i .

Par exemple, si $R_i = \sigma^2 I'_{n_i \times n_i}$ et D matrice arbitraire de dimension $(k \times k)$ définie non négative, on

utilisera la statistique (exhaustive pour le paramètre θ) $\sum_i e_i' e_i$ pour estimer σ^2 et la statistique $\sum_i b_i' b_i$ pour estimer D .

résultats : (cf calculs dans la partie 1.2.5.3, çï-dessous).

Les estimateurs de σ^2 et de D sont donnés par :

$$\hat{\sigma}^2 = (\sum_i e_i' e_i / \sum_i n_i) = t / \sum_i n_i \quad (1.15)$$

$$\hat{D} = \sum_i b_i b_i' / m \quad (1.16)$$

Soit $\hat{\theta}$ un estimateur de θ , on calcule :

$$\hat{t}_1 = E \left\{ \sum_i e_i' e_i / y_i, \hat{\alpha}(\hat{\theta}), \hat{\theta} \right\} = \sum_i \left[\hat{e}_i(\hat{\theta})' \hat{e}_i(\hat{\theta}) + \text{tracevar}(e_i / y_i; \hat{\alpha}(\hat{\theta}), \hat{\theta}) \right] \quad (1.17)$$

$$\hat{t}_2 = E \left\{ \sum_i b_i b_i' / y_i, \hat{\alpha}(\hat{\theta}), \hat{\theta} \right\} = \sum_i \left[\hat{b}_i(\hat{\theta}) \hat{b}_i(\hat{\theta})' + \text{var}(b_i / y_i; \hat{\alpha}(\hat{\theta}), \hat{\theta}) \right] \quad (1.18)$$

où :

$$\hat{e}_i(\hat{\theta}) = E(e_i / y_i, \hat{\alpha}(\hat{\theta}), \hat{\theta}) = y_i - X_i \hat{\alpha}(\hat{\theta}) - Z_i \hat{b}_i(\hat{\theta})$$

Pour obtenir l'estimateur du maximum de vraisemblance, $\hat{\theta}_M$, on commence par une valeur initiale convenable de $\hat{\theta}$ (estimation préalable) ensuite , on fait des itérations entre (1.17) et (1.18) définissant le E-step et (1.15) et (1.16) définissant le M-step. A la convergence, on a non seulement $\hat{\theta}_M$ mais aussi $\hat{\alpha}(\hat{\theta}_M)$ et $\hat{b}(\hat{\theta}_M)$ résultant du calcul du dernier E-Step (en utilisant toujours (1.17) et (1.18) avec $\hat{\theta}_M$ à la place de θ).

1.2.5.2 Utilisation du EM algorithme pour calculer l'estimateur du maximum de vraisemblance restreinte REML :

Etape 1 :

$$\forall i, y_i = X_i \alpha + Z_i b_i + e_i, \quad i = 1, \dots, m$$

α et b_i fixes.

Etape 2 :

$$\alpha \sim N(0, \Gamma), b_i \sim N(0, D), D = D(\theta), \text{cov}(\alpha, b_i) = 0$$

$$y_i \sim N(0, X_i \Gamma X_i' + Z_i D Z_i' + R_i)$$

Puisque généralement Γ ne peut être estimé car on n'a aucune information sur α ; prenons $\Gamma^{-1} = 0$. Un estimateur de θ est obtenu en maximisant la vraisemblance marginale limite de θ quand $\Gamma^{-1} \rightarrow 0$ sachant y , c'est à dire on cherche θ maximisant :

$$\begin{aligned} f(y; \theta) &= \lim_{\Gamma^{-1} \rightarrow 0} \int_{R^P} \int_{R^k} f(y/b; \alpha; \theta) f(b; \theta) f(\alpha) db d\alpha \\ &= \lim_{\Gamma^{-1} \rightarrow 0} \int_{R^P} f(y; \theta, \alpha) f(\alpha) d\alpha \\ &= \lim_{\Gamma^{-1} \rightarrow 0} \int_{R^P} f(y; \theta, \alpha) \exp -(1/2) \alpha' \Gamma^{-1} \alpha |\Gamma|^{-1/2} d\alpha \end{aligned} \quad (1.19)$$

Harville (1976) a montré que cette vraisemblance limite est équivalente à la vraisemblance restreinte.

Les estimateurs de Bayes empiriques sont :

$$E(\alpha/y, \Gamma^{-1} = 0, \hat{\theta}_R) = \hat{\alpha}(\hat{\theta}_R) \quad (1.20)$$

$$E(b_i/y, \Gamma^{-1} = 0, \hat{\theta}_R) = E(b_i/y_i, \hat{\alpha}(\hat{\theta}_R)\hat{\theta}_R) = \hat{b}_i(\hat{\theta}_R) \quad (1.21)$$

Calcul de $\hat{\theta}_R$: application du EM algorithme.

M-Step : le même que dans a).

E-Step : analogue à a) avec α éliminé du calcul par intégration.

$$\hat{t}_1 = E(\Sigma e_i' e_i / y_i, \hat{\theta}) = \sum_i \left[\hat{e}_i(\hat{\theta})' \hat{e}_i(\hat{\theta}) + \text{tracevar}(e_i / y_i; \hat{\theta}) \right] \quad (1.22)$$

$$\hat{t}_2 = E \left\{ \sum_i b_i b_i' / y_i, \hat{\theta} \right\} = \sum_i \left[\hat{b}_i(\hat{\theta}) \hat{b}_i(\hat{\theta})' + \text{var}(b_i / y_i; \hat{\theta}) \right] \quad (1.23)$$

où :

$$\hat{e}_i(\hat{\theta}) = y_i - X_i \hat{\alpha}(\hat{\theta}) - Z_i \hat{b}_i(\hat{\theta})$$

On peut se référer à Stiratelli et al.(1984).

1.2.5.3 Calculs :

En utilisant les notations de Laird et al.(1987), on a :

$$\alpha^{(W)} = (\sum_i X_i' W_i^{(W)} X_i)^{-1} (\sum_i X_i' W_i^{(W)} y_i), \quad (1.24)$$

$$\begin{aligned} W_i^{(W)} &= [\Sigma_i^{(W)}]^{-1} \\ \Sigma_i^{(W)} &= \sigma_i^2 . I + Z_i D^{(W)} Z_i' \end{aligned} \quad (1.25)$$

$$b_i^{(W)} = D^{(W)} Z_i' W_i^{(W)} r_i^{(W)} \quad (1.26)$$

où :

$$r_i^{(W)} = y_i - X_i \alpha^{(W)}.$$

E-Step :

$$\begin{aligned} \hat{t}_1^{(W)} &= E \left\{ \sum_i e_i' e_i / y_i, \hat{\alpha}^{(W)}, \theta^{(W)} \right\} \\ &= \sum_i \left[\hat{e}_i(\theta^{(W)})' \hat{e}_i(\theta^{(W)}) + \text{tracevar}(e_i / y_i; \hat{\alpha}^{(W)}, \theta^{(W)}) \right] \end{aligned} \quad (1.27)$$

$$\begin{aligned} \hat{t}_2^{(W)} &= E \left\{ \sum_i b_i b_i' / y_i, \hat{\alpha}^{(W)}, \theta^{(W)} \right\} \\ &= \sum_i \left[\hat{b}_i(\theta^{(W)}) \hat{b}_i(\theta^{(W)})' + \text{var}(b_i / y_i; \hat{\alpha}^{(W)}, \theta^{(W)}) \right] \end{aligned} \quad (1.28)$$

M-Step :

$$\begin{aligned} &\sum_i E \left\{ e_i' e_i / \hat{\alpha}(\theta^{(W+1)}), \theta^{(W+1)} \right\} = \hat{t}_1^{(W)} \\ &= \sum_i \left[\hat{e}_i(\theta^{(W+1)})' \hat{e}_i(\theta^{(W+1)}) + \text{trvar}(e_i / \hat{\alpha}(\theta^{(W+1)}), \theta^{(W+1)}) \right] \\ &= \sum_i n_i (\sigma^2)^{(W+1)} = \hat{t}_1^{(W)} \Rightarrow (\sigma^2)^{(W+1)} = \hat{t}_1^{(W)} / \sum_i n_i = \hat{t}_1^{(W)} / N. \end{aligned} \quad (1.29)$$

où :

$$N = \sum_i n_i, tr = trace.$$

D'où :

$$(\sigma^2)^{(W+1)} =$$

$$\left\{ \sum_i \left\{ \left[r_i^{(W)} - Z_i b_i^{(W)} \right]' \left[r_i^{(W)} - Z_i b_i^{(W)} \right] + (\sigma^2)^{(W)} tr(I - (\sigma^2)^{(W)} W_i^{(W)}) \right\} \right\} / N \quad (1.30)$$

et :

$$D^{(W+1)} = \sum_i \left[b_i^{(W)} b_i^{(W)'} + var(b_i/y) \right] / m.$$

$$D^{(W+1)} = \sum_i \left[b_i^{(W)} b_i^{(W)'} + D^{(W)} (I - Z_i' W_i^{(W)} Z_i D^{(W)}) \right] / m. \quad (1.31)$$

Pour l'estimateur REML, on a :

$$(\sigma^2)^{(W+1)} =$$

$$\left\{ \sum_i \left\{ \left[r_i^{(W)} - Z_i b_i^{(W)} \right]' \left[r_i^{(W)} - Z_i b_i^{(W)} \right] + (\sigma^2)^{(W)} tr(I - (\sigma^2)^{(W)} p_i^{(W)}) \right\} \right\} / N \quad (1.32)$$

et :

$$D^{(W+1)} = \sum_i \left[b_i^{(W)} b_i^{(W)'} + D^{(W)} (I - Z_i' p_i^{(W)} Z_i D^{(W)}) \right] / m. \quad (1.33)$$

Où :

$$p^{(W)} = W_i^{(W)} \left[I - X_i (\sum_i X_i' W_i^{(W)} X_i)^{-1} X_i' W_i^{(W)} \right].$$

A la convergence, on a en plus :

$$\hat{Var}(\hat{\alpha}) = (\sum_i X_i' W_i^{(\infty)} X_i)^{-1} \quad \infty : \text{"Indice" de la convergence.}$$

$$\hat{Var}(\hat{b}_i - b_i) = D^{(\infty)}(I - Z_i' p_i^{(\infty)} Z_i D^{(\infty)}).$$

Des calculs analogues sont donnés dans Foulley (1991 p92).

Remarques :

1) Le choix des valeurs initiales est déterminé par la forme du modèle et les caractéristiques des données.

Suivant Cook (1982b), des valeurs initiales acceptables sont les estimateurs des moindres carrés ordinaires, qui sont donnés par :

$$\hat{\alpha}_0 = (\sum_i X_i' X_i)^{-1} \sum_i X_i' y_i$$

$$\hat{b}_i = (Z_i' Z_i)^{-1} Z_i' (y_i - X_i \hat{\alpha}_0)$$

$$\hat{\sigma}_0^2 = (\sum_i y_i' y_i - \hat{\alpha}_0' \sum_i X_i' y_i - \sum_i \hat{b}_i Z_i' (y_i - X_i \hat{\alpha}_0)) / (N - (m - 1)q - p)$$

Où :

$$p = \dim \alpha; \quad q = \dim \theta \text{ et } \hat{D}_0 = (\sum_i b_i b_i' / m) - \hat{\sigma}_0^2 \sum_i (Z_i' Z_i)^{-1} / m.$$

2) Pour inverser \sum_i qui sert à calculer W_i ; de même que pour avoir $\sum_i X_i' W_i X_i$, on simplifie les calculs en utilisant l'identité :

$$W_i = [I - Z_i(\sigma^2 D^{-1} + Z_i' Z_i)^{-1} Z_i'] / \sigma^2.$$

La dimension de $(\sigma^2 D^{-1} + Z_i' Z_i)$ qui est égale à $k_i \times k_i$ est généralement plus petite que celle de Σ_i et qui est égale à $n_i \times n_i$; dans le cas contraire, on a la troisième remarque ci-dessous.

3) Quand la dimension de D est grande par rapport à n_i (on a moins d'observations que de paramètres) , il est plus efficace du point de vue des calculs d'utiliser les décompositions matricielles de Cholesky que d'inverser Σ_i (Laird et al.1987, p 100 ; Lindstrom & Bates 1988, p 1015).

Notons que cette décomposition n'est pas la factorisation de Cholesky.

Chapitre 2

Modélisation de la matrice de variance-covariance en modèles linéaires mixtes.

2.1 Introduction

Le modèle adéquat pour analyser des données longitudinales ou des données incomplètes et / ou déséquilibrées correspondant à des mesures répétées est le modèle linéaire mixte. Plusieurs travaux ont été faits dans ce domaine (Harville1977 ; Laird et Ware1982 ; Jennrich et Schluchter 1986 ; Lindstrom et Bates 1988 ; Chi et Reinsel 1989 ; Littell et al.2000 ; Park et al. 2001). Un ouvrage traitant ces modèles, est celui de Verbeke et Molenberghs 2000.

Pour l'estimation des paramètres fixes et aléatoires, nous avons la méthode du

maximum de vraisemblance (ML) ou celle du maximum de vraisemblance restreinte (REML). D'autre part, parmi ces auteurs, certains (Jennrich et Schluchter 1986 ; Park et al. 2001 p2453, entre autres) ont montré qu'une modélisation parcimonieuse de la matrice de variance-covariance du modèle améliore l'efficacité des estimateurs des paramètres de régression (les effets fixes). Pour cette raison, nous allons considérer dans cette partie plusieurs structures de la matrice de variance-covariance des observations afin d'en choisir la "meilleure".

On considère les modèles les plus répandus de mesures répétées déséquilibrées avec des matrices de variance-covariance structurées. Pour calculer les estimateurs du maximum de vraisemblance (ML) des paramètres de régression et ceux de la covariance, on utilise l'algorithme de Newton-Raphson ou celui des scores de Fisher.

2.2 Modèle

Le modèle de base est un modèle général du type :

$$y_i = X_i\beta + \varepsilon_i, \quad i = 1, \dots, n \quad (2.1)$$

Jennrich et Schluchter 1986 ont dressé une table comprenant différentes structures pour la matrice de variance-covariance des observations où la plupart des structures étaient linéaires du type :

$$\Sigma = \sum_{i=1}^q \theta_i G_i$$

où :

Σ est la matrice de variance-covariance du modèle.

G_1, \dots, G_q matrices connues.

$\theta_1, \dots, \theta_q$ paramètres à estimer.

2.3 Algorithmes pour le maximum de vraisemblance.

2.3.1 La fonction vraisemblance et les équations d'estimation.

Le logarithme noté par λ de la vraisemblance des données y_1, \dots, y_n est :

$$\lambda = \text{const..} - (1/2) \sum_i \log |\Sigma_i| - (1/2) \sum_i (y_i - X_i \beta)' \Sigma_i^{-1} (y_i - X_i \beta) \quad (2.2)$$

le vecteur S des scores et la matrice hessienne H sont :

$$S = \begin{bmatrix} S_\beta \\ S_\theta \end{bmatrix} = \begin{bmatrix} \partial \lambda / \partial \beta \\ \partial \lambda / \partial \theta \end{bmatrix} \quad (2.3)$$

$$H = \begin{bmatrix} H_{\beta\beta} & H_{\beta\theta} \\ H_{\theta\beta} & H_{\theta\theta} \end{bmatrix} = \begin{bmatrix} \partial^2 \lambda / \partial \beta \partial \beta & \partial^2 \lambda / \partial \beta \partial \theta \\ \partial^2 \lambda / \partial \beta \partial \theta & \partial^2 \lambda / \partial \theta \partial \theta \end{bmatrix} \quad (2.4)$$

Les expressions des éléments de S et de H sont :

$$S_\beta = \sum_i X'_i \Sigma_i^{-1} e_i \quad (2.5a)$$

$$[S_\theta]_r = (1/2) \sum_i \text{trace} \Sigma_i^{-1} (e_i e'_i - \Sigma_i) \Sigma_i^{-1} \Sigma_{ir}^0; r = 1, \dots, q \quad (2.5b)$$

$$H_{\beta\beta} = -\sum_i X'_i \Sigma_i^{-1} X_i \quad (2.6a)$$

$$[H_{\beta\theta}]_{jr} = -\sum_i X'_{ij} \Sigma_i^{-1} \Sigma_{ir}^0 \Sigma_i^{-1} e_i; j = 1, \dots, p; r = 1, \dots, q \quad (2.6b)$$

$$\begin{aligned} [H_{\theta\theta}]_{rs} = & -(1/2) \sum_i \text{trace} \Sigma_i^{-1} \Sigma_{ir}^0 \Sigma_i^{-1} (2e_i e'_i - \Sigma_i) \Sigma_i^{-1} \Sigma_{is}^0 \\ & + (1/2) \sum_i (e_i e'_i - \Sigma_i) \Sigma_i^{-1} \Sigma_{i,r,s}^{00}; r, s = 1, \dots, q \end{aligned} \quad (2.6c)$$

où :

$$e_i = y_i - X_i \beta; \Sigma_{ir}^0 = \partial \Sigma_i / \partial \theta_r; \Sigma_{i,r,s}^{00} = \partial^2 \Sigma_i / \partial \theta_r \partial \theta_s$$

Les éléments de Σ_{ir}^0 et de $\Sigma_{i,r,s}^{00}$ sont respectivement les dérivées premières et secondes des éléments de Σ_i par rapport à $\theta_1, \dots, \theta_q$.

2.3.2 Algorithmes de Newton-Raphson et 'Fisher-scoring' :

Principes : L'algorithme de Newton-Raphson utilise la matrice hessienne H ; tandis que l'algorithme 'Fisher-scoring' utilise à la place de H son espérance.

Newton-Raphson :

$$\begin{bmatrix} \hat{\beta} \\ \hat{\theta} \end{bmatrix} = \begin{bmatrix} \beta \\ \theta \end{bmatrix} - \begin{bmatrix} H_{\beta\beta} & H_{\beta\theta} \\ H_{\theta\beta} & H_{\theta\theta} \end{bmatrix}^{-1} \begin{bmatrix} S_{\beta} \\ S_{\theta} \end{bmatrix} \quad (2.7)$$

β et θ sont des valeurs actuelles.

'Fisher-scoring' :

$$EH_{\beta\beta} = -\sum_i X_i' \Sigma_i^{-1} X_i \quad (2.8a)$$

$$EH_{\beta\theta} = 0. \quad (2.8b)$$

$$E[H_{\theta\theta}]_{rs} = -(1/2) \sum_i \text{trace} \Sigma_i^{-1} \Sigma_{ir} \Sigma_i^{-1} \Sigma_{is}; r, s = 1, \dots, q \quad (2.8c)$$

Puisque $EH_{\beta\theta} = 0$, les nouvelles valeurs $\tilde{\beta}$ et $\tilde{\theta}$ sont obtenues en résolvant l'équation (2.7) sous forme de deux équations séparées.

Les nouvelles valeurs $\tilde{\beta}$ sont aussi obtenues par la méthode des moindres carrés généralisés ('GLS') avec les poids Σ_i^{-1} .

$$\tilde{\beta} = (\sum_i X_i' \Sigma_i^{-1} X_i)^{-1} (\sum_i X_i' \Sigma_i^{-1} y_i) \quad (2.9)$$

Les nouveaux estimateurs $\tilde{\theta}$ sont alors calculés (équation 2.7) par :

$$\tilde{\theta} = \theta + I_{\theta\theta}^{-1} \cdot S_{\theta} \quad (2.10)$$

avec $I_{\theta\theta}$ l'opposée de la matrice hessienne $E[H_{\theta\theta}]$ définie par (2.8c) et S_{θ} calculé à partir de (2.5b) en utilisant les nouveaux résidus $e_i = y_i - X_i \tilde{\beta}$. Ceci présente une

légère modification de la définition de l'algorithme des scores dans lequel S_θ est calculé en utilisant la nouvelle valeur $\tilde{\beta}$ au lieu de la valeur actuelle β .

2.3.3 Fonction de vraisemblance restreinte (résiduelle)

Il est bien connu que les estimateurs du maximum de vraisemblance des composantes de la variance sont biaisés. En effet, dans l'estimation du maximum de vraisemblance on ne tient pas compte que l'effet fixe β est aussi estimé. La méthode introduite par Patterson et Thompson 1971, consiste à modifier la vraisemblance en une vraisemblance restreinte qui tient compte de l'estimation de l'effet fixe et qui donne des estimateurs sans biais des composantes de la variance.

Le logarithme noté par λ_{REML} de la vraisemblance restreinte des données y_1, \dots, y_n est :

$$\begin{aligned} \lambda_{REML} = & \text{const..} - (1/2) \sum_i \log |\Sigma_i| - (1/2) \sum_i (y_i - X_i \tilde{\beta})' \Sigma_i^{-1} (y_i - X_i \tilde{\beta}) \\ & - (1/2) \sum_i \log |X_i' \Sigma_i^{-1} X_i| \end{aligned} \quad (2.11)$$

2.4 Autres Modèles

2.4.1 Modèle AR(1)RE

Dans leur étude longitudinale, Laird et Ware 1982 considéraient un modèle linéaire mixte avec des erreurs indépendantes. Jennrich et Schluchter 1986 ont dressé une table où les erreurs n'étaient pas indépendantes mais admettaient différentes structures

pour leurs covariances. Il s'est avéré à travers des exemples que combiner des effets aléatoires et des erreurs suivant un AR(1) (ou d'autres séries chronologiques, AR(p) ; MA(q) ; ARMA(p,q) ;...) donnait une structure plus ou moins parcimonieuse pour la matrice de variance-covariance (Chi et Reinsel 1989). Le modèle issu de cette combinaison sera intitulé modèle AR(1)RE.

2.4.1.1 Modèle

$$y_k = X_k\beta + C_k\tau_k + u_k, \quad k = 1, \dots, N \quad (2.12)$$

y_k : observations faites sur le *kième* individu ; vecteur de dimension $(T_k \times 1)$

β : vecteur inconnu des paramètres de population, de dimension $(n \times 1)$.

X_k : matrice du dispositif liant y_k à β , connue, de dimension $(T_k \times n)$.

τ_k : vecteur inconnu des effets individus, aléatoire de dimension $(m \times 1)$. $\tau_k \sim N(0, \Gamma_{m \times m})$

C_k : matrice du dispositif liant y_k à β , connue, de dimension $(T_k \times m)$.

u_k : vecteur inconnu des erreurs intra-individus, aléatoire de dimension $(T_k \times 1)$.

$u_k \sim AR(1)$.

$$u_{k,t} = \phi u_{k,t-1} + \varepsilon_{k,t-1}; \varepsilon_{k,t-1} \sim N(0, \sigma^2) \quad (2.13)$$

La matrice de variance-covariance du modèle est :

$$\Sigma_k = cov(y_k) = C_k \Gamma C_k' + \sigma^2 \Omega_k. \quad (2.14)$$

avec :

$\sigma^2\Omega_k$ matrice de covariance de u_k .

2.4.1.2 Estimation

Pour estimer les paramètres du modèle on applique l'algorithme des scores de Fisher.

Remarque1 :

Pour tester la présence de l'autocorrélation, un test des scores a été proposé par chi & Reinsel 1989.

Remarque2 :

Dans un contexte analogue, d'autres auteurs se sont intéressés aux tests en modèles linéaires mixtes; on peut citer Jacqmin 1994; Park et al.2001; Robert-Granié et Foulley 2002.,...entres autres.

2.4.2 Modèle à amortissement exponentiel

Ce modèle a été ramené par Munoz et al.1992. Il englobe comme cas spécial le modèle C.S ("Compound Symmetry") considéré par Jennrich et Schluchter 1986; le AR(1) et le MA(1) mais pas le AR(1)RE. Ce modèle est intitulé modèle à amortissement exponentiel et il est noté par DE ("Damped Exponential").

2.4.2.1 Modèle

$$y_i \sim MVN(X_i\beta; \sigma^2 V_i(\gamma, \theta; s_i)), 0 \leq \gamma \leq 1, 0 \leq \theta. \quad (2.15)$$

Le j, k ($j < k$) élément de V_i est :

$$Corr(y_{ij}, y_{ik}) = [V_i(\gamma, \theta; s_i)]_{jk} = \gamma^{(s_{ik} - s_{ij})^\theta} \quad (2.16)$$

La corrélation entre deux observations séparées par s unités de temps est γ^{s^θ} ; γ est la corrélation entre deux éléments séparés par 1 -unité de temps ; θ est le paramètre d'amortissement.

Le cas $\theta = 0$ correspond au modèle C.S dont la matrice de corrélation est :

$$Corr(y_{it}, y_{i,t+s}) = I_{\{s=0\}} + \gamma I_{\{s>0\}} \quad (2.17)$$

Le cas $\theta = 1$ correspond au modèle AR(1) dont la matrice de corrélation est :

$$Corr(y_{it}, y_{i,t+s}) = \gamma^{|s|} \quad (2.18)$$

Le cas $\theta \rightarrow \infty$ correspond au modèle MA(1) dont la matrice de corrélation est :

$$Corr(y_{it}, y_{i,t+s}) = I_{\{s=0\}} + \gamma I_{\{s=1\}} \quad (2.19)$$

I_A fonction indicatrice de l'ensemble A .

2.4.2.2 Estimation

Pour l'estimation des paramètres du modèle, on utilise l'algorithme de Newton Raphson (Munoz et al. 1992).

Remarque1 :

Si $0 < \theta < 1$, on obtient une famille de structure de corrélation dont la vitesse de décroissance de la corrélation est entre celle du "compound symmetry" et celle du AR(1).

Remarque2 :

Autres modèles alternatifs à celui-ci :

a) Modèle AR(2) caractérisé par :

$$y_t - \phi_1 y_{t-1} - \phi_2 y_{t-2}$$

b) Modèle AR(1)RE vu précédemment dans 2.4.1

Le travail suivant, qui a fait l'objet d'une publication, résume la plupart des modèles que l'on rencontre fréquemment dans les analyses longitudinales ou celles des mesures répétées déséquilibrées.

2.5 modèles linéaires mixtes

Résumé

Les mesures répétées faites sur une même unité expérimentale sont souvent corrélées qu'elles soient prises à travers le temps ou à travers l'espace. La plupart des statisticiens, surtout ceux utilisant les modèles marginaux (Liang et Zeger 1986), négligent cette corrélation ou la covariance entre les observations et la considèrent comme un paramètre de nuisance et ne s'intéressent qu'à la moyenne des observations. Dans ce travail, on s'est intéressé à des données correspondant à des mesures répétées. Pour mettre en évidence l'importance de la matrice de variance-covariance, on a considéré un modèle linéaire général avec différentes structures de cette matrice. Enfin, en comparant les différents modèles correspondant aux différentes structures à travers un exemple, on conclut au choix du " meilleur " modèle.

Mots clés : Mesures répétées déséquilibrées ; modèle linéaire mixte ; matrice de variance-covariance.

2.5.1 Introduction

Dans le domaine biomédical, la répétition des mesures dans un même groupe ou chez un même sujet entraîne une corrélation entre les observations dont le traitement doit être spécifique. Le modèle multivarié, ainsi que les techniques des séries chronologiques sont en particulier inadaptées à cause des problèmes calculatoires d'une part ; d'autre part l'effet aléatoire de l'individu ainsi que la corrélation inter-

individu n'ont pas de formes explicites dans ces contextes, d'où le passage à d'autres modèles.

Quand les observations sont équilibrées, c'est-à-dire en nombres égaux, il est facile de modéliser directement la matrice de variance-covariance (en utilisant par exemple une forme générale, "Unstructured"). Cependant, souvent les nombres des observations sont inégaux ; dans cette situation il est plus commode d'utiliser des modèles mixtes dont les paramètres aléatoires produisent naturellement une structure de corrélation entre les observations. Cette structure est importante, car dans une analyse de mesures répétées, généralement on veut savoir s'il y a une signification entre différents traitements par exemple ; ou, on veut faire des tests pour savoir si certaines variables ont des effets significatifs sur les résultats. La signification des effets des covariables dépend beaucoup du choix de la structure de la matrice de variance-covariance. Donc il est important de prendre un choix parcimonieux pour la structure de cette matrice. Une alternative au modèle 'multivarié' dont la matrice de variance-covariance admet une forme générale ('Unstructured') consiste à utiliser des modèles linéaires mixtes qui paraissent bien s'adapter à cette situation. Ces modèles sont de plus en plus utilisés pour l'analyse de données incomplètes et / ou déséquilibrées correspondant à des mesures répétées. Le choix de la matrice de variance-covariance ainsi que les problèmes calculatoires sont en partie résolus ces dernières années avec l'introduction et la mise en oeuvre par la procédure PROC MIXED de SAS ou par BMDP5V. Plusieurs travaux ont été faits dans ce domaine (Harville 1977 ; Laird & Ware 1982 ; Jennrich & Schluchter 1986 ; Lindstrom & Bates 1988 ; Chi & Reinsel 1989) et récemment Park & Lee 2002.

Dans ce papier, nous allons analyser des données de mesures répétées (données de Potthof & Roy 1964). Notre objectif est de choisir une structure parcimonieuse pour la matrice de variance-covariance. Dans la section 2, on introduit un modèle mixte simple ; par la suite, on utilise un modèle linéaire général mixte qui s'applique pour des mesures répétées équilibrées ainsi que déséquilibrées. Dans la section 3, on considère différentes structures de la matrice de variance-covariance. Un exemple est utilisé dans la section 4. Dans la section 5, des ajustements de plusieurs modèles y sont donnés en utilisant la procédure PROC MIXED de SAS (Littell et al.1996). Dans la section 6 , des comparaisons des ajustements des différentes structures de la matrice de variance-covariance sont faites afin d'en choisir le "meilleur" modèle.

2.5.2 Modèle

Dans cette partie, nous allons considérer un modèle mixte simple. Soit g le nombre de groupes de sujets ; n_i : nombre de sujets correspondant au groupe i .

Chaque sujet est mesuré t fois. Soit Y_{ijk} la valeur de la réponse mesurée à l'instant k sur le sujet j dans le groupe i , $i = 1, \dots, g, j = 1, \dots, n_i$ et $k = 1, \dots, t$. Notons que l'effet fixe du modèle spécifie la valeur moyenne des Y_{ijk} et est donné par $E(Y_{ijk}) = \mu_{ijk}$. L'effet aléatoire du modèle spécifie la structure de la covariance des observations.

On suppose que les observations des différents sujets sont indépendantes, donc :

$$Cov(Y_{ijk}, Y_{i'j'l}) = 0 \text{ pour } i \neq i' \text{ ou } j \neq j' .$$

$Cov(Y_{ijk}, Y_{ijl}) = \sigma_{k,l}$, covariance entre les mesures aux instants k et l sur le même individu.

$\sigma_{k,k} = \sigma_k^2$ est la variance à l'instant k .

Soit $Y_{ij} = (Y_{ij1}, Y_{ij2}, \dots, Y_{ijt})'$ vecteur des données aux instants $1, \dots, t$ de l'individu j dans le groupe i .

Le modèle en notation matricielle est donné par :

$$Y_{ij} = \mu_{ij} + \epsilon_{ij}$$

$$\mu_{ij} = (\mu_{ij1}, \mu_{ij2}, \dots, \mu_{ijt})'; \quad \epsilon_{ij} = (\epsilon_{ij1}, \epsilon_{ij2}, \dots, \epsilon_{ijt})'$$

La moyenne et la variance de Y_{ij} sont : $E(Y_{ij}) = \mu_{ij}$ et $V(Y_{ij}) = V_{ij}$; V_{ij} est la matrice de dimension (txt) avec l'élément $\sigma_{k,l}$ figurant dans la ligne k et la colonne l . Le vecteur des données pour tous les sujets est :

$$Y = (Y'_{11}, \dots, Y'_{1n}, \dots, Y'_{21}, \dots, Y'_{2n}, \dots, Y'_{g1}, \dots, Y'_{gn})'$$

De même

$$E(Y) = \mu = (\mu'_{11}, \dots, \mu'_{1n}, \dots, \mu'_{21}, \dots, \mu'_{2n}, \dots, \mu'_{g1}, \dots, \mu'_{gn})'$$

$$\epsilon = (\epsilon'_{11}, \dots, \epsilon'_{1n}, \dots, \epsilon'_{21}, \dots, \epsilon'_{2n}, \dots, \epsilon'_{g1}, \dots, \epsilon'_{gn})'$$

D'où le modèle (1) :

$$Y = \mu + \epsilon \tag{1}$$

Et $V(Y) = V = \text{diag}\langle V_{ij} \rangle$

Où $diag\langle V_{ij} \rangle$ est la matrice block diagonale avec V_{ij} dans chaque block.

En prenant :

$$\mu = X\beta \text{ et } \varepsilon = ZU + e,$$

on obtient le modèle (2) :

$$Y = X\beta + ZU + e \quad (2)$$

X : matrice du dispositif liant Y à β .

β : vecteur des effets fixes.

Z : matrice du dispositif liant Y à U .

U : vecteur des effets aléatoires.

e : vecteur des erreurs.

Soit $G = V(U)$ et $R = V(e)$, alors si U et e sont indépendantes, la matrice de variance-covariance des données est donnée par l'expression (3) :

$$V(Y) = ZGZ' + R \quad (3)$$

Dans la plupart des applications de mesures répétées, ZGZ' représente la partie inter-sujets et R représente la partie intra-sujets.

2.5.3 Structures de la covariance

Dans cette partie, nous présentons plusieurs structures pour la matrice de variance-covariance caractérisées par l'expression (3). Cependant, nous allons consid-

érer la covariance et la corrélation entre deux observations aux instants k et l sur le même sujet et non pas la covariance ou la corrélation totale du modèle (2).

2.5.3.1 Simple Structure (SIM) :

$$Cov(Y_{ijk}, Y_{ijl}) = 0 \text{ si } k \neq l, V(Y_{ijk}) = \sigma_{SIM}^2 \quad (4)$$

Cette structure suppose les observations indépendantes(même sur le même sujet) et ont des variances homogènes données par : $V(Y_{ijk}) = \sigma_{SIM}^2$.

La corrélation est :

$$Corr(lag) = 0. \quad (5)$$

2.5.3.2 Structure du " Compound symmetry " (C.S)

$$Cov(Y_{ijk}, Y_{ijl}) = \sigma_{CS,b}^2 \text{ si } k \neq l, V(Y_{ijk}) = \sigma_{CS,b}^2 + \sigma_{CS,w}^2 \quad (6)$$

Dans cette situation, les observations sur le même sujet ont des covariances homogènes ainsi que les variances.

La corrélation est donnée par :

$$Corr_{CS}(lag) = \sigma_{CS,b}^2 / (\sigma_{CS,b}^2 + \sigma_{CS,w}^2) \quad (7)$$

On remarque que la corrélation ne dépend pas du lag.

2.5.3.3 Structure AR(1)

$$Cov(Y_{ijk}, Y_{ijl}) = \sigma_{AR(1)}^2 \rho_{AR(1)}^{|k-l|} \quad (8)$$

Pour cette structure, les variances sont homogènes et sont données par :

$$V(Y_{ijk}) = \sigma_{AR(1)}^2. \quad (9)$$

On remarque que les covariances ne sont pas égales et qu'elles décroissent vers zéro quand le lag croit.

Les corrélations sont données par :

$$Corr_{AR(1)}(lag) = \rho_{AR(1)}^{lag} \quad (10)$$

2.5.3.4 Structure AR(1)RE

Cette structure est une structure autorégressive avec effet aléatoire pour le sujet.

$$Cov(Y_{ijk}, Y_{ijl}) = \sigma_{AR(1)RE,b}^2 + \sigma_{AR(1)RE,W}^2 \rho_{AR(1)RE}^{|k-l|} \quad (11)$$

Les variances sont homogènes et sont données par :

$$V(Y_{ijk}) = (\sigma_{AR(1)RE,b}^2 + \sigma_{AR(1)RE,W}^2) \quad (12)$$

La corrélation est donnée par :

$$Corr_{AR(1)RE}(lag) =$$

$$(\sigma_{AR(1)RE,b}^2 + \sigma_{AR(1)RE,W}^2 \rho_{AR(1)RE}^{|k-l|}) / (\sigma_{AR(1)RE,b}^2 + \sigma_{AR(1)RE,W}^2) \quad (13)$$

2.5.3.5 Structure Toeplitz (TOEP)

$$Cov(Y_{ijk}, Y_{ijl}) = \sigma_{TOEP,|k-l|} \quad (14)$$

La covariance dépend du lag.

La variance est :

$$V(Y_{ijk}) = \sigma_{TOEP}^2. \quad (15)$$

La corrélation est donnée par :

$$Corr_{TOEP}(lag) = \sigma_{TOEP,lag} / \sigma_{TOEP}^2 \quad (16)$$

Tous les éléments en dessous de la diagonale $|k - l| = lag$ sont $\sigma_{TOEP,|k-l|}$

Où k est la ligne et l est la colonne.

2.5.3.6 Covariance Sans structure (Unstructured "UN")

$$Cov(Y_{ijk}, Y_{ijl}) = \sigma_{UN,kl} \quad (17)$$

La structure "Unstructured" ne présente aucune forme pour la matrice de covariance; c'est une forme générale.

2.5.4 Exemple

Le premier outil d'analyse de mesures répétées ou de données longitudinales est la représentation graphique des observations en fonction du temps. Cette première analyse est une analyse exploratoire. Elle est basée sur de simples graphes qui peuvent donner des renseignements ou des informations importantes. Cela permet par exemple d'identifier les individus anormaux c'est à dire ceux dont les observations sont des valeurs aberrantes. Nous allons prendre un exemple classique celui de Potthoff et Roy 1964, repris par Jennrich et Schluchter 1986. On s'intéresse à un ensemble de données de croissance concernant 11 filles et 16 garçons. Pour chaque sujet, on mesure la distance (en mm) séparant le centre pituitaire et la fissure pterygomaxillaire aux différents âges : 8;10;12 et 14 ans (problème de chirurgie dentaire). Les données sont analysées en utilisant la procédure PROC MIXED du système SAS. Les données des filles ainsi que celles des garçons sont représentées graphiquement en fonction de l'âge dans les figures 1 et 2. Les moyennes de ces données sont représentées graphiquement en fonction de l'âge dans la figure 3.

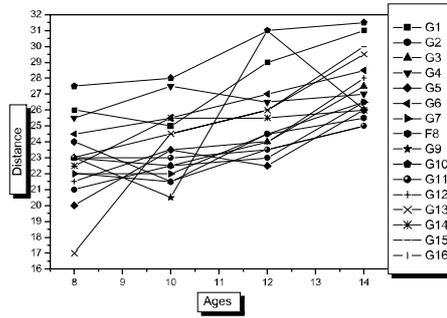


Fig1. Distance vs age (boys)

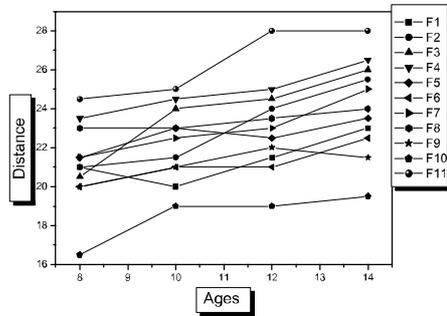


Fig2. Distance vs age (girls)

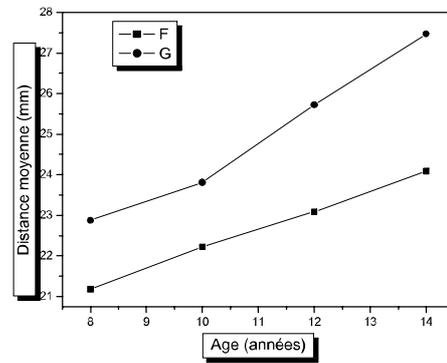


Fig3. mean distance (boys and girls)

On remarque qu'il n'y a pas d'anomalies au sujet de ces données, sauf peut être pour la figure représentant les distances des garçons, où le neuvième garçon a la plus petite valeur à l'âge de 10 ans et la plus grande à l'âge qui suit ; on peut penser que ce sont ou du moins l'une des deux valeurs (probablement la deuxième) est une valeur aberrante. D'autre part, la figure 3 montre que les distances moyennes des garçons sont toujours plus grandes que celles des filles.

2.5.5 Ajustement

La matrice de variance covariance de la structure "Unstructured" est représentée dans la table I. Celles des autres modèles, sont représentées dans la table II. Dans la table II, nous avons représenté uniquement la première ligne de la matrice de variance-covariance de chaque modèle ; ainsi que pour la matrice de corrélation.

TableI. Estimateurs de la covariance en utilisant le REML

| | AG1 | AG2 | AG3 | AG4 |
|--|----------|--------|--------|--------|
| | 1.499E-7 | | | |
| | -0.00043 | 2.8059 | | |
| | -0.00035 | 1.2017 | 3.6886 | |
| | -0.00049 | 1.8411 | 2.4203 | 3.7174 |

Table II. Estimateurs de la covariance et de la corrélation en utilisant le REML

| AG1 | AG2 | AG3 | AG4 |
|------------|--------|---------|---------|
| 1.Simple | | | |
| 2.7764 | 0.0 | 0.0 | 0.0 |
| 1.0 | 0.0 | 0.0 | 0.0 |
| 2.C.S | | | |
| 2.8014 | 0.8264 | 0.8264 | 0.8264 |
| 1.0 | 0.2950 | 0.2950 | 0.2950 |
| 3.AR(1) | | | |
| 2.7639 | 0.8399 | 0.2552 | 0.07755 |
| 1.0 | 0.3039 | 0.09234 | 0.02806 |
| 4.AR(1)RE | | | |
| 2.7964 | 0.8427 | 0.7928 | 0.7915 |
| 1.0 | 0.3013 | 0.2835 | 0.2830 |
| 5.Toeplitz | | | |
| 2.8231 | 0.8068 | 1.0517 | -0.9162 |
| 1.0 | 0.2858 | 0.3725 | -0.3245 |

Les estimateurs des paramètres de la matrice de variance- covariance et ceux de la corrélation sont :

$$1) \text{ SIM} : \hat{\sigma}_{SIM}^2 = 2.7764$$

$$2) \text{ C.S} : \hat{\sigma}_{CS,b}^2 = 0.8264; \quad \hat{\sigma}_{CS,w}^2 = 1.9750$$

$$3) \text{ AR}(1) : \hat{\sigma}_{AR(1)}^2 = 2.7639; \quad \hat{\rho}_{AR(1)} = 0.3039$$

$$4) \text{ AR}(1)\text{RE} : \hat{\sigma}_{AR(1)RE,b}^2 = 0.7915; \quad \hat{\sigma}_{AR(1)RE,w}^2 = 2.0049; \quad \hat{\rho}_{AR(1)RE} = 0.02553$$

5) Toeplitz : $\hat{\sigma}_{TOEP}^2=2.8231$; $\hat{\sigma}_{TOEP2}=0.8068$; $\hat{\sigma}_{TOEP3}=1.0517$; $\hat{\sigma}_{TOEP4}=-0.9162$

6) Unstructured : les valeurs des estimateurs sont données dans la table I.

2.5.6 Comparaison des différentes structures en utilisant les résultats des ajustements.

Table III. AIC (Akaike's information criterion) et BIC (Baye's information criterion)

| Covariance structure | AIC | BIC |
|--|-------|-------|
| 1.Simple | 411 | 412.3 |
| 2.Compound symmetric | 402.9 | 405.5 |
| 3.Autoregressive(1) | 404.9 | 407.5 |
| 4.Autoregressive(1) with random effect for patient | 404.9 | 408.8 |
| 5.Toeplitz | 396.0 | 401.2 |

La comparaison des différentes structures de la matrice de variance-covariance ou des différents modèles est basée sur les résultats figurant dans la table III. Généralement, le "meilleur" modèle est celui qui admet le plus grand AIC ou le plus grand BIC. Suivant ce critère notre préférence va au modèle dont la structure de la covariance est celle du modèle AR(1)RE. Il est bien évident qu'on n'a pas tenu compte du premier modèle (Simple) quoiqu' il admette les plus grandes valeurs pour le AIC et le BIC. Ce modèle ne reflète pas une situation réelle car il suppose que les données sont indépendantes sur le même sujet.

REFERENCES

- Chi, E.M. & Reinsel, G.C. Models for longitudinal data with random effects and AR(1) errors. *JASA*, June 1989, Vol.84, No.406, Theory and Methods.
- Harville, D.A. Maximum likelihood approaches to variance components estimation and to related problem. *JASA*, June 1977, Vol.72, No.358, pp320-329.
- Jennrich, R.I. & Schluchter M.D. Unbalanced Repeated-Measures Models with Structured Covariance Matrices. *Biometrics* 42, 805-820 December 1986.
- Laird, N.M & Ware, J.H. Random effects models for longitudinal data. *Biometrics*, 38, 963-974. December 1982.
- Liang, K.Y. & Zeger, S. L. Longitudinal data analysis using GLM. *Biometrika* 1986, 73, 1, pp13-22.
- Lindstrom, M.J. & Bates, D.M. Newton-Raphson and E-M algorithms for linear mixed effect models for repeated -measures data. *JASA*, Dec. 1988, V83, N404, pp1014-1022.
- Littell, R.C. Milliken, G.A., Stroup, W.W., Wolfinger, R.D. System for mixed models. SAS institute inc : Cary, N.C., 1996.
- Park, T. & Lee, Y.J. Covariance models for nested measures data : analysis of ovarian steroid secretion data. *Statist. Med* 2002; 21 :143-164.
- Potthof, R.F. & Roy, S.N. A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika*, 51, 313-336. 1964.

Chapitre 3

Etude de la fonction variance pour des données longitudinales utilisant une extension des modèles linéaires généralisés ('GLM').

3.1 Définition d'un modèle linéaire généralisé

Etant donné y_1, y_2, \dots, y_n , des variables aléatoires indépendantes telles que :

$$y_i = \mu_i + \varepsilon_i \quad (3.1)$$

Un modèle linéaire généralisé est la donnée :

a) D'une loi de distribution des ε_i

b) D'une matrice X de dimensions $(n \times m)$ de variables explicatives ; ceci définit un prédicteur linéaire $\eta = X\beta$.

c) D'une fonction de lien g , inversible telle que $g(\mu) = X\beta$.

3.2 Extensions des modèles linéaires généralisés.

3.2.1 Introduction

Wedderburn 1974, en introduisant la quasi-vraisemblance ('QL'), définie ci-dessous, élargissait le champ des modèles linéaires généralisés en remplaçant les hypothèses sur les distributions au sujet des composantes aléatoires dans le modèle par de faibles hypothèses dans lesquelles, seuls les moments d'ordre un et deux existent et sont suffisants pour cette quasi-vraisemblance pour faire les différentes estimations et inférences comme pour une vraisemblance.

Nelder et Pregibon 1987, introduisaient la quasi-vraisemblance étendue qui permet de comparer les différentes forme de la fonction variance.

3.2.2 Quasi-vraisemblance ; Quasi-vraisemblance étendue ; fonction variance et son estimation.

3.2.2.1 Quasi-vraisemblance

Supposons les observations y_1, y_2, \dots, y_n , indépendantes, de moyennes $\mu_i = \mu$ et de variances $Var(y_i) = V(\mu_i) = V(\mu)$; V étant une fonction connue.

On définit (Wedderburn 1974) le log de la quasi-vraisemblance pour une observation de moyenne μ et de variance $V(\mu)$ par :

$$Q(y; \mu) = \int_y^{\mu} (y - u)/V(u) du \quad (3.2)$$

Cette formule est équivalente à la suivante :

$$(\partial Q/\partial \mu)(y; \mu) = (y - \mu)/V(\mu) \quad (3.3)$$

3.2.2.2 Déviance

On définit (Nelder et Pregibon 1987) la déviance par :

$$D(y; \mu) = -2 \{Q(y; \mu) - Q(y; y)\} = -2 \int_y^{\mu} (y - u)/V(u) du \quad (3.4)$$

On suppose de plus que :

$$g(\mu_i) = x_i^T \beta \quad \text{et} \quad \text{Var}(y_i) = \phi V(\mu_i) \quad (3.5)$$

g : fonction de lien, connue et strictement monotone et différentiable.

x_i : Vecteur associé de dimension q .

β : paramètre inconnu.

ϕ : paramètre de taille ou de dispersion, inconnu.

Remarques :

- 1) ϕ est introduit pour tenir compte d'une éventuelle grande dispersion.
- 2) Ce paramètre peut ne pas être estimé.

3.2.2.3 Quasi-vraisemblance étendue

On définit (Nelder et Pregibon 1987) la quasi-vraisemblance étendue pour une observation par :

$$Q^+(y; \mu) = (-1/2) \log \{2\pi\phi V(y)\} - (1/2)D(y; \mu)/\phi \quad (3.6)$$

Un estimateur de ϕ , obtenu en maximisant la quasi-vraisemblance étendue totale, est :

$$\hat{\phi} = D(y; \hat{\mu})/n \quad (3.7)$$

Cette expression représente la déviance moyenne.

Les estimateurs de β obtenus en maximisant Q^+ sont les mêmes que ceux obtenus en maximisant Q , ce sont donc les estimateurs du maximum de quasi-vraisemblance (Wedderburn 1974; Mc Cullagh 1983). Ceci est dû au fait que Q^+ est une fonction linéaire de Q (Nelder et Pregibon 1987).

En général, la fonction $V(\mu)$ dépend d'un paramètre inconnu θ , d'où :

$$Var(y) = \phi V_{\theta}(\mu) \quad (3.8)$$

En plus, on propose généralement $V_{\theta}(\mu) = \mu^{\theta}$ (Modèle de Pregibon 1984).

Pour $\theta = 0, 1, 2, 3$, on a les fonctions variances associées aux lois normale, poisson, gamma et inverse de Gauss respectivement.

Pour la fonction variance de la famille $V_{\theta}(\mu) = \mu^{\theta}$, la déviance s'écrit :

$$D_{\theta}(y; \mu) = 2 \{y \log(y/\mu) - (y - \mu)\}, \quad \theta = 1$$

$$D_{\theta}(y; \mu) = 2 \{(y/\mu) - \log(y/\mu) - 1\}, \quad \theta = 2$$

$$D_{\theta}(y; \mu) = 2 \{y^{2-\theta} - (2 - \theta)y\mu^{1-\theta} + (1 - \theta)\mu^{2-\theta}\} / (1 - \theta)(2 - \theta), \text{ ailleurs}$$

On peut aussi considérer le modèle général de Quasi-vraisemblance (Pregibon 1984) :

$$g(\mu) = X\beta, \quad h(\phi) = z\theta, \quad Var(y) = \phi V(\mu) \quad (3.9)$$

Remarque 1 :

Il faut utiliser la Quasi-vraisemblance étendue pour déterminer ϕ , car la Quasi-vraisemblance admet une solution triviale avec ϕ infini, cette solution est pénalisée par la présence de $\log \phi$ dans la Quasi-vraisemblance étendue.

Remarque 2 :

Pour la loi de Poisson, Binomiale, Binomiale négative, (dans les cas discrets), on utilise l'approximation de Stirling $k! = (2\pi)^{1/2} k^k e^{-k}$ dans le logarithme de la vraisemblance pour avoir la Quasi-vraisemblance étendue.

Exemple :

Loi de Poisson de paramètre μ

$$f(y; \mu) = (e^{-\mu} \mu^y) / y!$$

$$\text{Log} f(\mu; y) = -\mu + y \log \mu - \log y! \simeq -\mu + y \log \mu - \log [(2\pi y)^{1/2} y^y e^{-y}]$$

$$\simeq -\mu + y \log \mu - (1/2) \log(2\pi y) - y \log y + y$$

Calculons la Quasi-vraisemblance étendue, notée par Q^+ .

Par définition :

$$Q^+(y; \mu) = (-1/2) \log \{2\pi\phi V(y)\} - (1/2)D(y; \mu)/\phi$$

Pour une loi de Poisson : $\phi = 1$, $V(y) = y$

$$D(y; \mu) = -2 \int_y^\mu \{(y-u)/u\} du = 2 \{y \log(y/\mu) - (y-\mu)\}$$

D'où :

$$Q^+(y; \mu) = (-1/2) \log \{2\pi y\} - (1/2) \{2y \log(y/\mu) - 2(y-\mu)\}$$

$$= (-1/2) \log \{2\pi y\} - y \log y - y \log \mu + y - \mu$$

On voit bien que cette Quasi-vraisemblance étendue est équivalente à la vraisemblance.

3.2.2.4 Inférence concernant les paramètres non linéaires.

Pour estimer θ , on utilise les intervalles de quasi-vraisemblance profil (des calculs sont donnés dans Nelder et Pregibon 1987).

Chapitre 4

Modèles marginaux - équations d'estimation généralisées

4.1 Introduction

Dans beaucoup d'applications, l'intérêt essentiel réside dans la moyenne ; les variance-covariances sont vues comme des paramètres de nuisance. Cependant, nous avons deux conséquences en ignorant cette matrice de variance-covariance :

- i) Les inférences concernant le paramètre de régression peuvent être incorrectes.
- ii) Les estimateurs des coefficients de régression seront inefficaces.

Aussi, pour mettre en évidence l'importance de cette matrice de variance-covariance, nous allons considérer deux approches. La première, la plus traditionnelle est basée sur la fonction vraisemblance (ML) mais elle s'est avérée insuffisante dans le sens où elle aboutit à des estimateurs biaisés asymptotiquement quand la matrice de variance-covariance n'est pas complètement spécifiée. Une 'amélioration' peut être considérée

en utilisant la fonction vraisemblance restreinte (ou résiduelle) (REML). Cependant, ces deux méthodes sont basées sur la connaissances des distributions qui sont fréquemment gaussiennes (ou à la limite exponentielles). Donc cette approche du maximum de vraisemblance est appliquée lorsque les données sont approximativement normales et beaucoup d'études ont été faites dans ce contexte en utilisant des modèles linéaires mixtes pour analyser des données longitudinales ou des mesures répétées déséquilibrées. Parmi les auteurs ayant utilisé cette approche, on peut citer : Harville1977; Laird et Ware 1982; Jennrich et Schluchter1986; Lindstrom et Bates1988, Foulley et al.2000; Littel et al. 2000; Park et Lee 2002;... entre autres. La plupart de ces travaux ont été faits dans le contexte du modèle linéaire mixte.

La deuxième approche est appliquée pour des données normales ou non normales mais elle est souvent appliquée pour le second cas. Cette méthodologie plus récente basée sur les modèles linéaires généralisés ('GLM') qui sont développés dans le livre de 'référence' de Mc Cullagh et Nelder1989, et l'estimation de quasi-vraisemblance ('QL') développée par Wedderburn 1974, en a été une grande alternative et qui a aboutit par la suite aux équations d'estimations généralisées ('GEE') développées par Liang et Zeger 1986. Beaucoup de travaux ont été publiés par la suite dans ce contexte, qui est celui du modèle marginal. On peut citer par exemple, Liang et Zeger 1986; Zeger et Liang 1986; Zeger, Liang et Albert 1988; Zhao et Prentice 1990; Prentice et Zhao 1991; Liang et al.1992; Fitzmaurice et Laird 1993; Park 1993; Crowder 1995; Crowder 2001;...entres autres.

Dans le chapitre 2, nous avons utilisé la première approche à des modèles linéaires

mixtes. Dans cette partie, nous allons utiliser la deuxième approche à des modèles marginaux.

4.2 Analyse de données longitudinales utilisant les 'GLM'

4.2.1 Introduction

Dans cette partie, on propose une autre extension des GLM pour analyser des données longitudinales ; ces données ne sont pas supposées suivre une distribution de probabilité gaussienne, on suppose seulement une forme de la distribution marginale à chaque instant (Liang et Zeger 1986).

4.2.2 Notations et modèle

Soit :

y_{it} : vecteur des résultats du i ème individu à l'instant t , de dimension $(p \times 1)$;

$t = 1, \dots, n_i ; i = 1, \dots, K$.

x_{it} : vecteur des covariables de dimension $(p \times 1)$.

$Y_i = (y_{i1}, \dots, y_{in_i})^T$: vecteur des résultats de dimension $(n_i \times 1)$ et $X_i = (x_{i1}, \dots, x_{in_i})^T$:

matrice des valeurs des covariables de dimension $(n_i \times p)$ pour le i ème sujet.

Supposons que la densité marginale de y_{it} est donnée par :

$$f(y_{it}) = \exp [\{y_{it}\theta_{it} - a(\theta_{it}) + b(y_{it})\} \phi] \quad (4.1)$$

Où :

$\theta_{it} = h(\eta_{it})$, $\eta_{it} = x_{it}\beta$, la fonction h est monotone et différentiable. Avec cette formulation, on a (Mc Cullagh et Nelder 1983) :

$$E(y_{it}) = a'(\theta_{it}); \text{Var}(y_{it}) = a''(\theta_{it})/\phi \quad (4.2)$$

avec a' et a'' dérivées première et seconde de la fonction a , qui est supposée connue.

ϕ : paramètre de taille.

Des calculs pour démontrer (4.2) sont donnés dans Annexe A.

4.3 Equations d'estimation

4.3.1 Equation d'estimation et indépendance

Pour estimer β , on introduit les équations d'estimation. On suppose que les observations répétées pour un sujet sont indépendantes. Sous cette hypothèse d'indépendance, les équations scores de l'analyse de vraisemblance ont la forme suivante :

$$U_I(\beta) = \sum_i X_i^T \Delta_i S_i \quad (4.3)$$

$\Delta_i = \text{diag}(d\theta_{it}/d\eta_{it})$ et $S_i = Y_i - a'(\theta_{it})$, (on suppose que $n_i = n$, données équilibrées).

L'estimateur de β noté par $\hat{\beta}_I$ est solution de l'équation (4.3) ci-dessus.

(Une méthode de résolution est donnée dans Liang et Zeger 1986 ; peut aussi être calculé en utilisant le logiciel GLIM, Baker & Nelder 1978).

Soit : $A_i = \text{diag}(a''(\theta_{it}))$, matrice de dimension $(n \times n)$. Sous de faibles conditions de régularité, on a le théorème suivant :

Théorème 1 :

L'estimateur $\hat{\beta}_I$ est consistant (Rao 1973, p364) et $K^{1/2}(\hat{\beta}_I - \beta)$ est asymptotiquement gaussien quand $K \rightarrow \infty$ avec une moyenne nulle et une matrice de covariance V_I donnée par :

$$\begin{aligned} V_I &= \lim_{K \rightarrow \infty} K (\sum_i X_i^T \Delta_i A_i \Delta_i X_i)^{-1} (\sum_i X_i^T \Delta_i \text{cov}(Y_i) \Delta_i X_i) (\sum_i X_i^T \Delta_i A_i \Delta_i X_i)^{-1} \\ &= \lim_{K \rightarrow \infty} K \{H_1(\beta)\}^{-1} . H_2(\beta) . \{H_1(\beta)\}^{-1} \end{aligned} \quad (4.4)$$

Preuve : Cas particulier du théorème 2, ci-dessous, en posant $R(\alpha) = I$

4.3.2 Equations d'estimation généralisées

Dans cette partie, on tient compte de la corrélation (éventuellement de façon grossière) des observations répétées sur le même sujet pour augmenter l'efficacité.

Soit :

$$V_I = A_i^{1/2} R(\alpha) A_i^{1/2} / \phi \quad (4.5)$$

La matrice $R(\alpha)$ qui est de dimension $(n \times n)$ est appelée matrice de corrélation de travail ('Working correlation matrix') ; on a supposé les données équilibrées ($n_i = n$).

Les équations d'estimation généralisées sont données par :

$$\sum_i D_i^T V_i^{-1} S_i = 0 \quad (4.6)$$

$D_i = d\{a'(\theta_{it})\}/d\beta = A_i \Delta_i X_i$ où $A_i = \text{diag}(a''(\theta_{it}))$, matrice de dimension $(n \times n)$.

L'estimateur $\hat{\beta}_G$ de β est solution de l'équation (4.6) ci-dessus.

Une méthode de résolution est donnée dans annexe C; c'est celle des moindres carrés pondérés itérativement ('IRLS').

Remarques :

1) Si $R(\alpha) = I$, (4.6) se réduit à (4.3).

2) $\forall i, U_i(\beta, \alpha) = D_i^T V_i^{-1} S_i$ est la même fonction que celle obtenue de la quasi-vraisemblance de McCullagh 1983, sauf que dans cette partie, V_i est non seulement une fonction de β mais aussi de α .

L'équation (4.6) peut être réexprimée comme une fonction de β seulement, en remplaçant α dans (4.5) et (4.6) par $\hat{\alpha}(Y, \beta, \phi)$, un estimateur de α , qui est $K^{1/2}$ consistant quand β et ϕ sont connus; c'est-à-dire qui est tel que : $K^{1/2}(\hat{\alpha} - \alpha) = O_P(1)$.

Excepté pour des choix particuliers de R et de $\hat{\alpha}$, le paramètre de taille ϕ reste généralement présent dans (4.6). Pour compléter le processus, on remplace ϕ par $\hat{\phi}(Y, \beta)$, qui est un estimateur $K^{1/2}$ consistant quand β est connu.

Par conséquent (4.6) devient :

$$\sum_i \left[U_i(\beta, \hat{\alpha}\{\beta, \phi(\beta)\}) \right] = 0 \quad (4.7)$$

et $\hat{\beta}_G$ est défini comme solution de cette équation, (4.7).

Théorème 2 :

Sous de faibles conditions de régularité et sachant que :

- i) $\hat{\alpha}$ estimateur $K^{1/2}$ consistant connaissant β et ϕ .
- ii) $\hat{\phi}$ estimateur $K^{1/2}$ consistant connaissant β .
- iii) $\left| \partial \hat{\alpha}(\beta, \phi) / \partial \phi \right| \leq H(Y, \beta)$ qui est $O_P(1)$.

Alors :

$K^{1/2}(\hat{\beta}_G - \beta)$ est asymptotiquement gaussien quand $K \rightarrow \infty$ avec une moyenne nulle et une matrice de covariance V_G donnée par :

$$V_G = \lim_{K \rightarrow \infty} K (\sum_i D_i^T V_i^{-1} D_i)^{-1} (\sum_i D_i^T V_i^{-1} \text{cov}(Y_i) V_i^{-1} D_i) (\sum_i D_i^T V_i^{-1} D_i)^{-1} \quad (4.8)$$

Preuve :

Soit $\alpha^*(\beta) = \hat{\alpha} \left\{ \beta, \hat{\phi}(\beta) \right\}$, et sous certaines conditions de régularité, un développement de Taylor de $U_i(\hat{\beta}_G, \alpha^*)$ au voisinage de β nous donne :

$$K^{1/2}(\hat{\beta}_G - \beta) \simeq \left[\sum_i - \frac{\delta}{\delta \beta} U_i \{ \beta, \alpha^*(\beta) \} / K \right]^{-1} \cdot \left[\sum_i U_i \{ \beta, \alpha^*(\beta) \} / K^{1/2} \right] \quad (4.9)$$

où :

$$\frac{\delta}{\delta \beta} U_i \{ \beta, \alpha^*(\beta) \} = \frac{\partial}{\partial \beta} U_i \{ \beta, \alpha^*(\beta) \} + \left[\frac{\partial}{\partial \alpha^*} U_i \{ \beta, \alpha^*(\beta) \} \right] \cdot \frac{\partial \alpha^*(\beta)}{\partial \beta}$$

$$= A_i + B_i.C$$

Soit β fixé, un développement de Taylor de $U_i \{\beta, \alpha^*(\beta)\}$ au voisinage de α nous donne :

$$\begin{aligned} \frac{\sum_i^K U_i \{\beta, \alpha^*(\beta)\}}{K^{1/2}} &= \frac{\sum_i^K U_i \{\beta, \alpha\}}{K^{1/2}} + \frac{\sum_i^K \partial U_i \{\beta, \alpha\}}{\partial \alpha} \cdot \frac{1}{K} \cdot K^{1/2} (\alpha^* - \alpha) + o_P(1) \\ &= A^* + B^* C^* + o_P(1) \end{aligned} \quad (4.10)$$

On a : $B^* = o_P(1)$, car les $\frac{\partial U_i \{\beta, \alpha\}}{\partial \alpha}$ sont linéaires en S_i dont les moyennes sont nulles, et avec les 3 conditions i) - iii), on a :

$$\begin{aligned} C^* &= K^{1/2} \left[\hat{\alpha} \left\{ \beta, \hat{\phi}(\beta) \right\} - \hat{\alpha} \{\beta, \phi\} + \hat{\alpha} \{\beta, \phi\} - \alpha \right] \\ &= K^{1/2} \left[\frac{\partial \hat{\alpha}}{\partial \phi}(\beta, \phi) \cdot (\phi^* - \phi) + \hat{\alpha} \{\beta, \phi\} - \alpha \right] = O_P(1). \end{aligned}$$

Par conséquent, $\sum_i U_i \{\beta, \alpha^*(\beta)\} / K^{1/2}$ est asymptotiquement équivalent à A^* dont la distribution asymptotique est une multigaussienne de moyenne nulle et de matrice de covariance :

$$\lim_{K \rightarrow \infty} \left\{ \frac{\sum_i D_i^T V_i^{-1} \text{cov}(Y_i) V_i^{-1} D_i}{K} \right\} \quad (4.11)$$

Finalement, en remarquant que $\sum_i B_i = o_P(K)$, $C = O_P(1)$ et que $\sum_i A_i / K$ converge quand $K \rightarrow \infty$ vers $-\sum_i D_i^T V_i^{-1} D_i / K$; on a le résultat.

Remarque : α et ϕ sont estimés de différentes manières (Liang et Zeger 1986).

Estimation de α et de ϕ : à une itération donnée, α et ϕ peuvent être estimés à partir des résidus de Pearson définis par :

$$\hat{r}_{it} = \left\{ y_{it} - a'(\hat{\theta}_{it}) \right\} / \left\{ a''(\hat{\theta}_{it}) \right\}^{1/2} \quad (4.12)$$

où $\hat{\theta}_{it}$ dépend de la valeur actuelle de β .

ϕ peut être estimé par :

$$\hat{\phi}^{-1} = \frac{\sum_i^K \sum_t n_i \hat{r}_{it}^2}{(N - p)}, \quad \text{où } N = \sum_i n_i \quad (4.13)$$

Un estimateur (consistant) de α peut être obtenu en fonction de l'expression suivante :

$$\hat{R}_{uv} = \frac{\sum_i \hat{r}_{iu} \hat{r}_{iv}}{(N - p)} \quad (4.14)$$

Différents estimateurs sont donnés suivant différents exemples (Liang et Zeger 1986). Estimations données dans annexe B.

Des estimateurs pour ϕ sont donnés dans Mc Cullagh et Nelder 1983, p 83.

Chapitre 5

Les Modèles à effets aléatoires versus les modèles marginaux.

5.1 Introduction

Les données longitudinales sont caractérisées par le fait que pratiquement les observations répétées dans un groupe de sujets ou chez le même sujet tendent à être corrélées; contrairement aux cas des modèles linéaires généralisés ('GLM'), (Mc Cullagh & Nelder 1983) et ceux de quasi-vraisemblance (Wedderburn 1974; Mc Cullagh 1983), où les observations sont supposées indépendantes. L'objectif dans cette partie est de décrire la moyenne marginale d'une réponse (y_{it}) pour le sujet i à l'instant t en fonction des variables (x_{it}), d'une part; d'autre part, comparer les modèles mixtes à effets aléatoires avec les modèles marginaux.

On considère deux approches (Zeger, Liang & Albert 1988).

1ère Approche :

L'hétérogénéité dans les paramètres de régression peut être modélisée explicitement ; on se référera pour cela à l'approche SS (Spécifique au sujet), ('Subject Specific'). Cette approche est surtout utile quand l'objectif est la réponse d'un individu plutôt que celle d'une population ; par exemple dans les études de courbes de croissance (chapitre 1).

2ième Approche :

La réponse d'un modèle moyennant sur une population peut être modélisée comme une fonction des covariables sans tenir compte de l'hétérogénéité, explicitement d'un sujet à l'autre. Les coefficients de régression ont une interprétation de la population plutôt que pour un individu d'où l'utilisation de l'expression moyennant sur une population (' PA : Population-Averaged ').

5.2 Modèles SS et PA

5.2.1 Modèle SS

Les modèles à effets mixtes permettent de définir l'espérance conditionnelle de la variable réponse.

Soit (y_{it}) la réponse à l'instant t pour l'individu i , (x_{it}) vecteur des variables fixes de dimension $(p \times 1)$, où $t = 1, \dots, n_i$ et $i = 1, \dots, k$, (z_{it}) vecteur de variables (généralement un sous ensemble de (x_{it})) de dimension $(q \times 1)$, associé au vecteur à effets aléatoires b_i et de dimension $(q \times 1)$, et soit $\mu_{it} = E(y_{it}/b_i)$. Sous le GLM mixte

('Generalized linear mixed model' noté par certains auteurs par GL2M), on a :

$$h(u_{it}) = x'_{it}\beta + z'_{it}b_i, \quad Var(y_{it}/b_i) = g(u_{it}).\phi \quad (5.1)$$

h est la fonction de lien et g est la fonction variance.

Les b_i sont supposées indépendantes et suivent une distribution mixte (notée par F).

Objectifs : estimer β ; les paramètres de F et éventuellement ϕ .

5.2.2 Modèle PA

Comme leur nom l'indique, les modèles marginaux ont pour objectif de modéliser l'espérance marginale de la variable réponse.

On s'intéresse à :

$$\mu_{it} = E(y_{it}). \quad (5.2)$$

Supposons donc :

$$h^*(\mu_{it}) = x'_{it}\beta^*, \quad Var(y_{it}) = g^*(\mu_{it}).\phi \quad (5.3)$$

β^* décrit comment la réponse moyenne, plutôt que la réponse d'un sujet, dépend des covariables. Seuls les modèles PA décrivent la covariance des observations répétées pour un sujet ; les modèles SS expliquent la source de cette covariance.

Un avantage des modèles PA est que la réponse moyenne sur une population pour

une valeur donnée (x_{it}), est directement estimée à partir des observations sans tenir compte de la distribution des b_i .

Sous le modèle GLM mixte dans 5.2.1, formule (5.1), on a :

$$\mu_{it} = E(y_{it}) = E(E(y_{it}/b_i)) = \int_{R^q} h^{-1}(x'_{it}\beta + z'_{it}b_i)dF(b_i) \quad (5.4)$$

Cette formule nous permet de passer du modèle SS au modèle PA.

5.3 Equations d'estimation généralisées pour les modèles PA et SS

Dans cette partie nous allons décrire des équations d'estimation généralisées qui ajustent soit le modèle SS soit le modèle PA pour des données longitudinales.

Pour plus de détail, on se référera à Liang et Zeger 1986 ; Zeger et Liang 1986. Rappelons que cette méthode d'estimation a été introduite pour la première fois par Gilmour et al. 1985 dans le contexte des modèles linéaires généralisés mixtes et développée par la suite par Liang et Zeger en 1986).

5.3.1 Estimation des coefficients de regression

Pour modéliser la moyenne marginale μ_{it} , on suppose que :

$$h^*(\mu_{it}) = x'_{it}\beta^*, \quad Var(y_{it}) = g^*(\mu_{it}) \cdot \phi$$

Soit :

$$\mu_i = E(y_i) = (h^{*-1}(x'_{i1}\beta^*), \dots, h^{*-1}(x'_{in_i}\beta^*))$$

et

$$A_i = \text{diag}(g^*(\mu_{i1}), \dots, g^*(\mu_{in_i}))$$

Un estimateur de β^* , noté par $\hat{\beta}^*$ est solution des équations d'estimation généralisées ('GEE') et qui sont données par :

$$U(\beta^*) = \sum_i \frac{\partial \mu'_i}{\partial \beta^*} V_i^{-1}(\alpha)(y_i - \mu_i) \quad (5.5)$$

Avec μ'_i transposée de μ_i et $V_i(\alpha) = A_i^{1/2} R(\alpha) A_i^{1/2}$.

Liang et Zeger 1986, ont montré que $\hat{\beta}^*$ est consistant et est asymptotiquement (lorsque $k \rightarrow \infty$) gaussien ayant uniquement une spécification correcte de la moyenne et les conditions de régularité usuelles.

Un estimateur dit estimateur robuste ou estimateur 'Sandwich' de la variance de l'estimateur $\hat{\beta}^*$ est donné par :

$$V_{\hat{\beta}^*} = M_0^{-1} M_1 M_0^{-1} \quad (5.6)$$

où :

$$M_0 = \sum_i \frac{\partial \mu'_i}{\partial \beta^*} V_i^{-1} \frac{\partial \mu_i}{\partial \beta^*}$$

et

$$M_1 = \sum_i \frac{\partial \mu'_i}{\partial \beta^*} V_i^{-1} (y_i - \hat{\mu}_i)(y_i - \hat{\mu}_i)' V_i^{-1} \frac{\partial \mu_i}{\partial \beta^*}$$

Cet estimateur est aussi consistant même quand $cov(y_i) \neq V_i$ (Zeger, Liang & Albert 1988).

Remarque1 :

Pour le modèle SS, β est aussi solution des mêmes équations (les équations GEE (5.5)); seulement on a donné comme exemple dans le modèle SS, le modèle GLM mixte, donc pour trouver $\hat{\beta}$, nous avons la démonstration qui suit :

Pour utiliser l'approche (GEE) pour le modèle GLM mixte, calculons μ_i et V_i .

Ayant les moments conditionnels dans (5.1) et la distribution F des effets aléatoires, l'espérance marginale de μ_i a la forme de (5.2).

La matrice de covariance marginale notée par V_i , est donnée par :

$$V_i = Cov \left[E(y_i/b_i) \right] + E \left[Cov(y_i/b_i) \right] \quad (5.7)$$

ses éléments sont définis comme suit :

Exemple la covariance entre le *sième* et le *tième* élément est donnée par :

$$[V_i]_{st} = \int (u_{is} - \mu_{is})(u_{it} - \mu_{it})dF(b_i) + \phi I(s = t) \int g(u_{is})dF(b_i)$$

où :

$u_{it} = h^{-1}(x'_{it}\beta + z'_{it}b_i)$ et I_A fonction indicatrice de l'ensemble A .

Ayant calculé μ_i et V_i pour chaque sujet, on calcule β dans l'équation (5.5).

Remarque2 :

Quand k est grand relativement à n_i et F est donnée (par exemple une distribution gaussienne avec une moyenne nulle et une matrice de variance- covariance notée par D), un développement de Taylor au voisinage de $(b_i = 0)$, de la fonction réciproque de la fonction de lien, donne (à partir des équations (5.1) et (5.7)) l'approximation suivante :

$$\begin{aligned} Cov(y_i) &\simeq Cov \left[h^{-1}(x'_{it}\beta) + \frac{\partial h^{-1}}{\partial b_i}(x'_{it}\beta)b_i \right] + \\ &\quad \phi E \left\{ g \left[h^{-1}(x'_{it}\beta) + \frac{\partial h^{-1}}{\partial b_i}(x'_{it}\beta)b_i \right] \right\} \\ &\simeq L_i Z_i D Z'_i L_i + \phi A_i \simeq \tilde{V}_i \end{aligned} \quad (5.8)$$

où :

$$L_i = diag \left\{ \frac{\partial h^{-1}}{\partial u}(u), u = x'_{it}\beta; t = 1, \dots, n_i \right\}$$

En utilisant \tilde{V}_i comme approximation de V_i , on a conditionnellement à F l'approximation asymptotique suivante :

$$\sqrt{k}(\hat{\beta} - \beta) \xrightarrow[k \rightarrow \infty]{} N(0, V_{\beta})$$

$$\hat{V}_{\beta} = \tilde{M}_0^{-1} \tilde{M}_1^{-1} \tilde{M}_0^{-1}$$

où :

$$\tilde{M}_0 = \sum_i \frac{\partial \mu'_i}{\partial \beta} \tilde{V}_i^{-1} \frac{\partial \mu_i}{\partial \beta}$$

et

$$\tilde{M}_1 = \sum_i \frac{\partial \mu_i'}{\partial \beta} \tilde{V}_i^{-1} (y_i - \hat{\mu}_i)(y_i - \hat{\mu}_i)' \tilde{V}_i^{-1} \frac{\partial \mu_i}{\partial \beta}$$

5.3.2 Estimation de D et de ϕ

L'approximation (5.8) donne des estimateurs " peu fiables " ou grossiers de D , on les estime par :

$$V_i \simeq L_i Z_i D Z_i' L_i + \phi A_i$$

D'où :

$$D \simeq (Z_i' Z_i)^{-1} Z_i' L_i^{-1} (V_i - \phi A_i) L_i^{-1} Z_i (Z_i' Z_i)^{-1}$$

et

$$\hat{D} = \frac{1}{k} \sum_i (Z_i' Z_i)^{-1} Z_i' \hat{L}_i^{-1} \left[(y_i - \hat{\mu}_i)(y_i - \hat{\mu}_i)' - \hat{\phi} A_i \right] \hat{L}_i^{-1} Z_i (Z_i' Z_i)^{-1} \quad (5.9)$$

Le paramètre de taille ϕ est estimé à partir des termes diagonaux de la matrice de covariance.

Notons que :

$$E(y_{it} - \mu_{it})^2 \simeq \phi g(\mu_{it}) + (L_{itt})^2 Z_{it}' D Z_{it}$$

ce qui donne :

$$\hat{\phi} = \frac{\sum_i \sum_t (y_{it} - \hat{\mu}_{it})^2 - (L_{itt})^2 Z'_{it} D Z_{it}}{g(\hat{\mu}_{it})}$$

Pour estimer $\hat{\beta}$ et $(\hat{D}, \hat{\phi})$ simultanément, on résout les équations itératives (5.5) et (5.9).

Conclusion :

Pour les modèles PA, seule la fonction de lien doit être correctement spécifiée pour avoir des estimateurs consistants des paramètres de régression ; par contre, dans les modèles SS il faut ajouter à cette condition la distribution paramétrique F des effets aléatoires (en pratique F est supposée gaussienne).

Notes et discussion :

Des avantages et des inconvénients des modèles marginaux et des équations d'estimation généralisées sont évoqués dans plusieurs travaux , on peut citer ceux de Zhao et Prentice 1990 ; Prentice et Zhao 1991 ; Liang et al.1992 ; Fitzmaurice et Laird 1993 ; Park 1993 ; Crowder 1995 ; Lindsey et Lambert 1998 ; Crowder 2001 ;...entres autres.

Rappelons que Liang et Zeger 1986 ont introduit leur approche pour l'analyse de données corrélées. Leur idée était de modéliser l'espérance marginale de la variable réponse et d'en estimer les paramètres de régression par la résolution des équations d'estimation généralisées. Ces équations font intervenir une matrice de covariance (ou de corrélation) de travail dépendant d'un paramètre α , cette matrice est arbitraire et peut ne pas être correctement spécifiée. Ils proposent par la suite un estimateur de la variance des paramètres de régression (β) dit estimateur robuste ou estimateur 'sandwich' et démontrent que les estimateurs de β et de leurs variances sont convergents même si la matrice de covariance ou de corrélation de travail est mal spécifiée.

Prentice 1988 a étendu cette idée dans le contexte de réponses binaires en introduisant des équations d'estimation pour le paramètre de corrélation noté par α . L'objectif était d'estimer conjointement les paramètres de régression et de corrélation.

Prentice et Zhao 1991 et Zeger et Liang 1992 ont généralisé cette méthode pour une réponse quelconque.

Crowder 1995, à travers des exemples pris pour la matrice de corrélation de travail

et pour la vraie matrice de corrélation a montré que l'estimateur de α peut ne pas être consistant (voire ne pas exister du tout); ceci relève un problème dans la première hypothèse du théorème 2 de Liang et Zeger 1986; c'est-à-dire, pour satisfaire cette hypothèse, il faut trouver les situations où l'estimateur de α est $K^{1/2}$ consistant, c'est-à-dire les situations où $K^{1/2}(\hat{\alpha} - \alpha) = O_p(1)$.

Park et Shin 1999, ont critiqué le travail de Crowder 1995 et ont contredit les résultats de l'auteur par des simulations; cependant, ces simulations étaient faites sur des échantillons de petites tailles ($n = 25$ et $n = 100$). Qu'en est-il donc pour les grands échantillons; car, on rappelle que le travail de Crowder 1995 a porté sur de grands échantillons ayant remis en cause ceux, asymptotiques, de Liang et Zeger 1986.

Crowder 2001, pour y remédier aux inconvénients des équations d'estimation généralisées de Liang et Zeger 1986, a proposé des améliorations de celles-ci en combinant une approche notée GE ('Gaussian Estimation' basée sur le maximum de vraisemblance) avec celle des équations GEE. Cette méthode est beaucoup plus basée sur la méthode GE, où l'auteur concluait par le remarque suivante, est qu'il est plus avantageux et plus facile de maximiser une fonction (exemple la vraisemblance) et qu'un maximum existe presque toujours, même s'il est local; que de résoudre des équations (exemple les GEE) qui peuvent parfois ne pas admettre de solutions.

D'autres auteurs ont essayé d'apporter des améliorations concernant les équations GEE.

Notamment, Lipsitz, Laird et Harrington 1991, ont proposé l'odds ratio (OR)

par paire comme mesure de l'association intragroupe à la place de la corrélation ou de la covariance.

Liang ; Zeger et Qaqish 1992, ainsi que Fitzmaurice et Laird 1993, ont aussi utilisé l'odds ratio.

Des comparaisons entre l'approche du Maximum de vraisemblance et celle des équations GEE ont été faites par Park 1993 qui concluait au choix de la première.

Lindsey et Lambert 1998, ont souligné des avantages et surtout des désavantages des modèles marginaux (par exemple un traitement peut être très bon en moyenne alors qu'il est mauvais pour chaque sujet). Par contre ont souligné les auteurs, ces modèles peuvent être appropriés pour des études descriptives, telles que les études épidémiologiques. Cependant, ce sont des modèles qui ne peuvent être appliqués qu'avec une grande précaution dans les études expérimentales, tels que les essais cliniques. Des exemples sont donnés par les auteurs pour comparer les modèles marginaux versus les modèles conditionnels.

Hall et Severini 1998 , ont proposé une extension des GEE afin d'améliorer l'efficacité des estimateurs des paramètres d'association α . Leur méthode est intitulée méthode EGEE ('extended generalized estimating equations').

Enfin, notons que Hu et Lachin 2001, ont insisté sur le fait que différentes matrices de corrélation de travail aboutissent à différentes conclusions par le biais d'une étude sur le traitement du diabète.

En se basant sur les résultats de la littérature conclus par ces différents auteurs, on

peut dire que le choix de la matrice de corrélation de travail , voire le choix même de la méthode GEE est assez délicat et que cette méthode reste très discutable ; surtout, vis-à-vis de celle du maximum de vraisemblance.

Conclusion et Perspectives

Le premier objectif de ce travail de recherche était basé sur le choix de la matrice de variance-covariance dans un modèle linéaire à effets mixtes, d'une part ; d'autre part, vu que le choix de cette matrice est assez délicat et vu son influence sur les paramètres (fixes) de régression, ceci nous a permis de passer à d'autres modèles.

Dans un premier temps, on a commencé par introduire les modèles de quasi-vraisemblance puis leur extension a donné naissance aux équations d'estimation généralisées. La comparaison des deux approches, la première, celle du maximum de vraisemblance appliquée aux modèles linéaires mixtes avec la seconde, qui est celle de la résolution des équations d'estimation généralisées dans le cadre des modèles marginaux ; est assez difficile car chacune des deux méthodes appliquée relativement aux deux familles de modèles, a ses avantages et ses inconvénients.

Nous sommes en présence de la confrontation ou de la compétition de deux familles de modèles. Les modèles linéaires à effets mixtes et les modèles marginaux.

Les premiers ont été largement utilisés dans plusieurs domaines (surtout en génétique quantitative ; en médecine ; en biologie ; ou en écologie). Ces modèles s'avèrent s'adapter convenablement aux données longitudinales et aux mesures répétées équilibrées ou déséquilibrées, également en présence de données manquantes.

Pour les seconds, ils sont d'actualité. Beaucoup d'auteurs les ont comparés à d'autres approches classiques et ont essayé d'apporter des améliorations à ces modèles (ou à ces équations GEE). Par exemple approche du maximum de vraisemblance

et GEE, Park 1993; méthode EGEE ('Extended Generalized Estimating equation') Hall et Severini 1998 ;Pseudo-vraisemblance restreinte pour données discrètes, Evans et al.2001 ; Combinaison GE et GEE, Crowder 2001 ;...etc). Cette deuxième approche, reste donc un champ assez ouvert et très discutable.

REFERENCES

- [1] Baker, R.J & Nelder, J.A. 1978. The GLIM System. Release 3. Generalized Linear Interactive Modelling. Oxford : Numerical Algorithm Group.
- [2] Chi, E.M. & Reinsel, G.C. Models for longitudinal data with random effects and AR(1) errors. JASA June 1989, Vol.84, No.406, pp.452-459. Theory and Methods.
- [3] Cook N. 1982*b*, A fortran program for random-effects models. Technical report Harvard school of public health, dpt of biostat.
- [4] Crowder M. 1995. On the use of a working matrix correlation in using GLM for repeated measures. Biometrika, 82,2, pp407-10.
- [5] Crowder M. 2001. On repeated measures analysis with misspecified covariance structure. JRSS B, 63, part1, pp.55-62.
- [6] Dempster A.P.; Laird N.M. & Rubin D.B. 1977. Maximum likelihood with incomplete data via the E-M algorithm. Jrss B39, 1-38.
- [7] Dempster A.P.; Rubin D.B. & Tsutakawa R.K. 1981. Estimation covariance component models. Jasa 76, 341-353.
- [8] Diggle P.J.; Liang K.Y. & Zeger S.L. 1994. Analysis of longitudinal data. Oxford science publications.
- [9] Diggle P.J.; Heagerty P.; Liang K.Y. & Zeger S.L. 2002. Analysis of longitudinal data. 2nd edn. Oxford : Oxford university press.
- [10] Dixon W.J. (ED.) BMDP statistical software manual volume 2. University of California press : Berkley, CA, 1988.
- [11] Fitzmaurice G.M. & Laird N.M. 1993. A likelihood based method for analysing longitudinal binary responses. Biometrika 80,1, pp141-51.
- [12] Foulley J.L. 1991. Estimation des composantes de la variance en modèles linéaires. DEA de Stat. 1991/1992. Paris INSERM. INRA 78252 Jouy en Josas.
- [13] Foulley J.L., Jaffrézic F & Granié C.R. 2000. EM-REML estimation of covariance parameters in gaussian mixed models for longitudinal data analysis. Genet. Sel. Evol. 32, 129-141.
- [14] Gilmour A.R., Anderson R.D. & Rae A.L. 1985. The analysis of binomial data by a generalized linear mixed model. Biometrika 72, 593-599.
- [15] Grizzle J. & Allen D. 1969, Analysis of growth and dose-response curves. Biometrics 25, 357-381.
- [16] Hall B. & Severini T. 1998. Extended Generalized Estimating Equations for clustered data. JASA V93, N444, 1365-1374.
- [17] Harville, D.A. 1974. Bayesian inference of variance components using only errors contrasts. Biometrika 61, 393-385.
- [18] Harville, D.A. 1976, Extension of the Gauss Markov theorem to include the estimation of random effects. Ann. Stat., V4, 384-395.

- [19] Harville, D.A. 1977. Maximum likelihood approaches to variance components estimation and to related problem. *JASA*, V72, No358, pp320-329.
- [20] Henderson C.R. ; Kempthorne O. ; Searle S.R. & Vonkrosigk C.N. 1959. Estimation of environmental and genetic trends from records subject to culling. *Biometrics* 15, 192-218.
- [21] Hu M. & Lachin J.M. 2001. Application robust estimating equations to the analysis of quantitative longitudinal data. *Stat. Med.* V20, I22, 3411-3428.
- [22] Jacquemin-Gadda H. 1994. Tests d'homogénéité pour les modèles linéaires généralisés. Applications à l'étude des variations géographiques du déficit cognitif chez les personnes âgées. Thèse pour le Doctorat de l'université de Paris VII.
- [23] Jennrich, R.I. & Schluchter M.D. 1986. Unbalanced Repeated-Measures Models with Structured Covariance Matrices. *Biometrics* 42, 805-820.
- [24] Jones, H.R. & Boadi-Boateng. Unequally spaced longitudinal data with AR(1) serial correlation. *Biometrics* 47, 161-175 Mar. 1991.
- [25] Laird, N.M & Ware, J.H. Random effects models for longitudinal data. *Biometrics*, 38, 963-974. December 1982.
- [26] Laird, N.M. ; Lange N. & Stram D. 1987. Maximum likelihood computations with repeated measures : application of the E-M algorithm. *JASA* Mar. 1987, V82, N397, pp97-105.
- [27] Liang, K.Y. & Zeger, S. L. 1986. Longitudinal data analysis using GLM. *Biometrika* 1986 73, 1, pp13-22.
- [28] Liang, K.Y. ; Zeger, S. L. & Quaquez B. 1992. Multivariate regression analysis for categorical data. *JRSS B*, 54, N1, PP3-40.
- [29] Lindsey, J.K. 1993. Models for repeated measurements. New York : Oxford University Press.
- [30] Lindsey, J.K. & Lambert, P. 1998. On the appropriateness of marginal models for Repeated Measurements in Clinical Trials. *Stat. In Med.* 1998, 17, 447-469.
- [31] Lindstrom M.J. & Bates D.M. 1988. Newton-Raphson and E-M algorithm for linear mixed effect models for repeated -measures data. *Jasa* Dec. 1988, V83, N404, pp 1014-1022.
- [32] Littell R.C. ; Milliken G.A. ; Stroup W.W. & Wolfinger R.D. 1996. System for mixed models. SAS institute inc : Cary, N.C., 1996.
- [33] Littell R.C. ; Pendergast J. & Natarajan R. Modelling covariance Structure in the analysis of repeated measures data. *Statist. Med.* 2000 ; 19 : 1793-1819.
- [34] Lipsitz S.R. ; Laird N.M. & Harrington D.P. 1991. Generalized estimating equations for correlated binary data : using odds ratio as a measure of association. *Biometrika* 78, 153-160.
- [35] McCullagh P. & Nelder J.A. 1983. Generalized linear model. Chapman and Hall London.

- [36] Munoz A. ; Carey V. & al.1992. A parametric family of correlation structures for the analysis of longitudinal data. *Biometrics* 48, 733-742.
- [37] Nelder J.A. & Pregibon D.1987. An extended quasi-likelihood function. *Biometrika* , 74,2,pp221-232.
- [38] Park T.1993. A comparison of the GEE approach with the maximum likelihood approach for repeated measurements. *Stat.Med.* V12,1723-32.
- [39] Park T. & Shin D-Y.1999.On the use of working correlation matrices in the GEE approach for longitudinal data.*Commun.Statist.-Simula.*28(4),1011-1029.
- [40] Park T. Park J.K. & Davis C.S.2001. Effects of covariance model assumptions on hypothesis tests for repeated measurements : analysis of ovarian hormone data and pituitary-pterygomaxillary distance data. *Statist.Med.* ;20 :2441-2453.
- [41] Park,T.& Lee,Y.J. 2002. Covariance models for nested measures data : analysis of ovarian steroid secretion data. *Statist.Med.* ;21 :143-164.
- [42] Patterson,H.D. & Thompson,R.1971.Recovery of inter-block information when block sizes are unequal. *Biometrika* 58,545-54.
- [43] Potthof,R.F. & Roy,S.N.1964.A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika* 51,313-36.
- [44] Prentice R.L.1988.Correlated binary regression with covariates specific to each binary observation.*Biometrics* 44,1033-1048.
- [45] Prentice R.L. & Zhao L.P.1991. Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. *Biometrics* 47, 825-839.
- [46] Pregibon D.1984. Review of GLM by McCullagh and Nelder.*Ann.Stat.*12,1589-96.
- [47] Rao C.R.1965. The theory of least squares when the parameters are stochastics and its application to the analysis of growth curves. *Biometrika* 52, 447-458.
- [48] Robert-Granié C. & Foulley J.L.2002. Illustration des procédures de tests des effets fixes et des effets aléatoires en modèles mixtes. Journées didactiques INRIA Modèles Mixtes.
- [49] Rochon J. & Helms R.W.1989. Maximum likelihood estimation for incomplete repeated measures experiments under an ARMA covariance structure. *Biometrics* 45, 207-218, Mar.1989.
- [50] SAS Institute inc.SAS/STAT Software : Changes and Enhancements through Release 6.12. SAS Institute Inc. :Cary,NC,1997.
- [51] Searle,S.R.1971. Topics in variance component estimation. *Biometrics* 27, 1-76.
- [52] Searle,S.R. 1979. Notes on variances components estimation : a detailed account of maximum likelihood and kindred methodology. Paper BU 673-M, Cornell University Ithaca.

- [53] Stiratelli R., Laird N.M. & Ware J.H. 1984. Random-effects models for serial observations with binary responses. *Biometrics* 40, 961-971.
- [54] Verbeke G. & Molenberghs G. 2000. *Linear Mixed Models for longitudinal data*. New York : Springer.
- [55] Ware J.H. 1985. Linear models for the analysis of longitudinal studies. *American Statistician* 39, 95-101.
- [56] Wedderburn R.W.M. 1974. Quasi-likelihood function ; GLM and the Gauss-Newton method. *Biometrika*, 61,3,439-447.
- [57] Witkovsky V. 2002. Matlab algorithm mixed.m for solving Henderson's mixed model equations. Institute of Measurement Science Slovak Academy of Sciences.
- [58] Zhao L.P. & Prentice R.L. 1990. Correlated binary regression using a quadratic exponential model. *Biometrika* 77,642-648.
- [59] Zeger S.L & Liang K.Y. 1986. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* 42, 121-130.
- [60] Zeger S.L & Liang K.Y. 1992. An overview of methods for the analysis of longitudinal data. *Statistics in medicine* 11,1825-1839.
- [61] Zeger S.L & Liang K.Y. & Albert P.S. 1988. Models for longitudinal data : a generalized equation approach. *Biometrics* 44, 1049-1060.

Annexe A

Expressions de la moyenne et de la variance pour les GLM (Modèles Linéaires Généralisés).

ANNEXE A

Expressions de la moyenne et de la variance pour les GLM (Modèles Linéaires Généralisés).

Supposons que la densité marginale de y_i est donnée par :

$$f(y_i; \theta_i, \phi) = \exp \{ [y_i \theta_i - b(\theta_i)] / a(\phi) + c(y_i; \phi) \} \quad (\text{A1})$$

Pour une distribution normale, on a :

$$\begin{aligned} f(y_i; \theta_i, \phi) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp [y_i - \mu_i]^2 / \sigma^2 \\ &= \exp \{ [y_i \mu_i - \mu_i^2 / 2] / \sigma^2 - (1/2) [(y_i / \sigma^2) + \text{Log}(2\pi\sigma^2)] \} \end{aligned} \quad (\text{A2})$$

Donc, on peut poser :

$$\theta_i = \mu_i; b(\theta_i) = (\mu_i^2 / 2); \phi = \sigma^2; a(\phi) = \phi$$

et

$$c(y_i; \phi) = -(1/2) [(y_i / \sigma^2) + \text{Log}(2\pi\sigma^2)]$$

Le logarithme de la vraisemblance en utilisant la formule A1, noté par l_i est donné par :

$$l_i(\theta_i, \phi; y_i) = \text{Log}f(y_i; \theta_i, \phi) \quad (\text{A3})$$

Alors la moyenne et la variance de y_i , peuvent être déduites à partir des relations :

$$E\left(\frac{\partial l_i}{\partial \theta_i}\right) = 0 \quad (\text{A4})$$

et

$$E\left(\frac{\partial^2 l_i}{\partial \theta_i^2}\right) + E\left(\frac{\partial l_i}{\partial \theta_i}\right)^2 = 0 \quad (\text{A5})$$

Nous avons donc, à partir de (A3) :

$$l_i(\theta_i, \phi; y_i) = [[y_i \theta_i - b(\theta_i)] / a(\phi)] + c(y_i; \phi) \quad (\text{A6})$$

Donc :

$$\frac{\partial l_i}{\partial \theta_i} = [y_i - b'(\theta_i)] / a(\phi) \quad (\text{A7})$$

et

$$\frac{\partial^2 l_i}{\partial \theta_i^2} = -b''(\theta_i) / a(\phi) \quad (\text{A8})$$

D'où, à partir des équations (A4) et (A7) on a :

$$E\left(\frac{\partial l_i}{\partial \theta_i}\right) = [\mu_i - b'(\theta_i)] / a(\phi) = 0$$

et donc :

$$E(y_i) = \mu_i = b'(\theta_i)$$

De façon similaire à partir de (A5); (A7) et (A8) on a :

$$\frac{-b''(\theta_i)}{a(\phi)} + \frac{Var(y_i)}{a^2(\phi)} = 0$$

et donc :

$$Var(y_i) = b''(\theta_i).a(\phi)$$

Annexe B

Estimation des paramètres α et ϕ figurant dans la formule (4.5)

ANNEXE B

Estimation des paramètres α et ϕ figurant dans la formule (4.5)

Exemple 1 : ' Indépendance structure '

$R(\alpha) = R_0$, matrice de corrélation quelconque. Si $R(\alpha) = R_0 = I$, la matrice identité, on obtient les équations d'estimation dans le cas de l'indépendance (cf.équation (4.3)). Cependant pour n'importe quelle structure pour $R_0, \hat{\beta}_G$ et \hat{V}_G sont consistants.

On peut écrire :

$$\alpha_t = \text{corr}(y_{it}, y_{it'}) = \begin{cases} 1, & \text{si } t = t'' \\ 0, & \text{si } t \neq t'' \end{cases} \quad (\text{B1})$$

Exemple 2 :

$$\text{Soit : } \alpha = (\alpha_1, \dots, \alpha_{n-1})^t$$

$$\text{où } \alpha_t = \text{corr}(y_{it}, y_{i,t+1}), \quad \text{pour } t = 1, \dots, n-1$$

Un estimateur naturel de α_t , sachant β et ϕ est donné par :

$$\hat{\alpha}_t = \phi \sum_{i=1}^K \hat{r}_{it} \hat{r}_{i,t+1} / (K - p) \quad (\text{B2})$$

Dans cet exemple, $R(\alpha)$ est supposée être tridiagonale avec $R_{t,t+1} = \alpha_t$.

Un estimateur de ϕ n'est pas nécessaire pour estimer $\hat{\beta}_G$ et \hat{V}_G quand les α_t ci-dessus sont utilisés car les ϕ qui figurent dans la formule de $\hat{\alpha}_t$ sont supprimés dans le calcul de V_i . Comme cas particulier, prenons $s = 1$ et $\alpha_t = \alpha(t = 1, \dots, n-1)$. Alors, α peut être estimé par :

$$\hat{\alpha} = \frac{\sum_{t=1}^{n-1} \hat{\alpha}_t}{(n-1)} \quad (\text{B3})$$

Exemple 3 :

Soit $s = 1$ et supposons que $\text{corr}(y_{it}, y_{it'}) = \alpha$ pour tout $t \neq t'$. C'est une structure de corrélation échangeable obtenue à partir de modèles à effets aléatoires avec un niveau aléatoire pour chaque sujet (exemple dans Laird et Ware 1982). Pour ϕ donné, α peut être estimé par :

$$\hat{\alpha}_t = \frac{\phi \sum_{i=1}^K \sum_{t > t'} \hat{r}_{it} \hat{r}_{it'}}{\left\{ \sum_{i=1}^K \frac{1}{2} n_i (n_i - 1) - p \right\}} \quad (\text{B4})$$

Comme dans l'exemple 1 et 2, un estimateur de ϕ n'est pas nécessaire pour obtenir $\hat{\beta}_G$ et \hat{V}_G .

Exemple 4 :

Soit $\text{corr}(y_{it}, y_{it'}) = \alpha^{|t-t'|}$. Pour y_{it} gaussiennes, cette structure de corrélation est semblable à celle d'un processus autoregressif d'ordre un, AR(1) (Feller, 1971 p 89). Puisque sous ce modèle, $E(\hat{r}_{it}, \hat{r}_{it'}) \simeq \alpha^{|t-t'|}$, on peut estimer α par la pente de la regression de $\log(\hat{r}_{it} \hat{r}_{it'})$ sur $\log(|t-t'|)$. Dans cet exemple, un estimateur de ϕ est nécessaire pour déterminer $\hat{\beta}_G$ et \hat{V}_G .

Exemple 5 :

Considérons $R(\alpha)$ totalement non spécifiée, c'est à dire telle que $s = \frac{1}{2}n(n-1)$.

Dans cet exemple, R peut être estimée par :

$$\frac{\phi}{K} \sum_{i=1}^K A_i^{-1/2} S_i S_i^t A_i^{-1/2} \quad (\text{B5})$$

Notons que pour ce cas, les équations (4.6), (équations GEE données dans le chapitre 4) et (B5) ensemble donnent les équations de vraisemblance actuelles (ou courantes) si les Y_i suivent une distribution gaussienne multivariée. De plus, la covariance asymptotique, V_G , se réduit à :

$$\lim_{K \rightarrow \infty} \left\{ \frac{\sum_{i=1}^K D_i^T \text{cov}^{-1}(Y_i) D_i}{K} \right\}^{-1} \quad (\text{B6})$$

car R est la vraie matrice de corrélation. De même, aucun estimateur de ϕ n'est exigé pour obtenir $\hat{\beta}_G$.

Annexe C

Résolution des équations GEE

ANNEXE C

Résolution des équations GEE

1-Maximum de vraisemblance pour les GLM (Modèles Linéaires Généralisés).

Supposons que la densité marginale de y_i est donnée par :

$$f(y_i; \theta_i, \phi) = \exp \{ [[y_i \theta_i - b(\theta_i)] / a(\phi)] + c(y_i; \phi) \} \quad (\text{C1})$$

Le logarithme de la vraisemblance, l_i est donné par :

$$l_i(\theta_i, \phi; y_i) = \text{Log} f = [[y_i \theta_i - b(\theta_i)] / a(\phi)] + c(y_i; \phi) \quad (\text{C2})$$

Pour N observations, on a :

$$L(\beta) = \sum_{i=1}^N l_i$$

D'autre part,

$$\frac{\partial l_i}{\partial \beta_j} = \frac{\partial l_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j}, j = 1, \dots, t \quad (\text{C3})$$

puisque

$$\frac{\partial l_i}{\partial \theta_i} = [y_i - b'(\theta_i)] / a(\phi) \quad (\text{C4})$$

et puisque

$$\mu_i = b'(\theta_i) \text{ et } \text{Var}(y_i) = b''(\theta_i) \cdot a(\phi) \quad (\text{C5})$$

D'où

$$\frac{\partial \mu_i}{\partial \theta_i} = b''(\theta_i) = \text{Var}(y_i)/a(\phi) \quad (\text{C6})$$

Aussi, puisque

$$\eta_i = \sum_{j=1}^t \beta_j x_{ij}; \frac{\partial \eta_i}{\partial \beta_j} = x_{ij} \quad (\text{C7})$$

Finalement, puisque

$$\eta_i = g(\mu_i) \quad (\text{C8})$$

$\frac{\partial \mu_i}{\partial \eta_i}$ dépend de la fonction de lien du modèle.

En récapitulant , on a :

$$\begin{aligned} \frac{\partial l_i}{\partial \beta_j} &= \frac{(y_i - \mu_i)}{a(\phi)} \frac{a(\phi)}{\text{Var}(y_i)} \frac{\partial \mu_i}{\partial \eta_i} x_{ij} \\ &= \frac{(y_i - \mu_i)}{\text{Var}(y_i)} \frac{\partial \mu_i}{\partial \eta_i} x_{ij} \end{aligned} \quad (\text{C9})$$

Et les équations de vraisemblance sont :

$$\sum_{i=1}^N \frac{(y_i - \mu_i)}{\text{Var}(y_i)} \frac{\partial \mu_i}{\partial \eta_i} x_{ij}, j = 1, \dots, t \quad (\text{C10})$$

Ces équations de vraisemblance qui sont équivalentes aux équations GEE de Liang et Zeger 1986, sont non linéaires en fonction de β . Leur résolution pour trouver l'estimateur de β , noté par $\hat{\beta}$, nécessite des méthodes itératives, qui seront discutées ci-dessous.

L'algorithme que nous allons utiliser est celui des scores de Fisher et donc, la vitesse de convergence de $\hat{\beta}$ vers β , dépend de la matrice d'information. Soit donc :

$$\begin{aligned}
 E\left(\frac{\partial^2 l_i}{\partial \beta_h \partial \beta_j}\right) &= -E\left(\frac{\partial l_i}{\partial \beta_h}\right)\left(\frac{\partial l_i}{\partial \beta_j}\right) \\
 &= -E\left[\frac{(y_i - \mu_i)}{\text{Var}(y_i)} x_{ih} \frac{\partial \mu_i}{\partial \eta_i} \cdot \frac{(y_i - \mu_i)}{\text{Var}(y_i)} x_{ij} \frac{\partial \mu_i}{\partial \eta_i}\right] \\
 &= -\frac{x_{ih} x_{ij}}{\text{Var}(y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2
 \end{aligned} \tag{C11}$$

Donc

$$E\left(\frac{\partial^2 L(\beta)}{\partial \beta_h \partial \beta_j}\right) = -\sum_{i=1}^N \frac{x_{ih} x_{ij}}{\text{Var}(y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2, h, j = 1, \dots, t \tag{C12}$$

La matrice d'information, dont les éléments sont : $E\left(-\frac{\partial^2 L(\beta)}{\partial \beta_h \partial \beta_j}\right)$, est donnée par :

$$Inf = X'WX \tag{C13}$$

Où W est la matrice diagonale, dont les éléments suivant la diagonale principale, sont :

$$W_i = \frac{1}{\text{Var}(y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 \tag{C14}$$

et X' : transposé de X .

2-Algorithmme de résolution

On rappelle l'algorithme de Newton Raphson ;

$$\beta^{(k+1)} = \beta^{(k)} - (H^{(k)})^{-1} \cdot q^{(k)} \quad (\text{C15})$$

Les éléments de H sont : $(\frac{\partial^2 L(\beta)}{\partial \beta_h \partial \beta_j})$

Ceux de q , sont : $(\frac{\partial L(\beta)}{\partial \beta_j})$

L'algorithme des scores de Fisher est :

$$\beta^{(k+1)} = \beta^{(k)} + (Inf^{(k)})^{-1} \cdot q^{(k)} \quad (\text{C16})$$

D'où :

$$Inf^{(k)} \cdot \beta^{(k+1)} = Inf^{(k)} \cdot \beta^{(k)} + q^{(k)} \quad (\text{C17})$$

Où les éléments de $Inf^{(k)}$, sont donnés par : $E(-\frac{\partial^2 L(\beta)}{\partial \beta_h \partial \beta_j})$, évalués en $\beta^{(k)}$.

La partie droite de l'équation (C17), est le vecteur dont les éléments sont donnés par :

$$\sum_j \left[\sum_{i=1}^N \frac{x_{ih} x_{ij}}{Var(y_i)} (\frac{\partial \mu_i}{\partial \eta_i})^2 \beta_j^{(k)} \right] + \sum_{i=1}^N \frac{(y_i - \mu_i)}{Var(y_i)} x_{ih} \frac{\partial \mu_i}{\partial \eta_i} \quad (\text{C18})$$

Où μ_i et $\frac{\partial \mu_i}{\partial \eta_i}$ sont évalués en $\beta^{(k)}$.

Donc :

$$Inf^{(k)} \cdot \beta^{(k)} + q^{(k)} = X' W^{(k)} z^{(k)} \quad (\text{C19})$$

Où :

$W^{(k)}$ est la matrice W évaluée en $\beta^{(k)}$.

et

$z^{(k)}$ est le vecteur dont les éléments sont :

$$\begin{aligned} z_i^{(k)} &= \sum_{j=1}^t x_{ij} \beta_j^{(k)} + (y_i - \mu_i^{(k)}) \left(\frac{\partial \eta_i}{\partial \mu_i} \right)^{(k)} \\ &= \eta_i^{(k)} + (y_i - \mu_i^{(k)}) \left(\frac{\partial \eta_i}{\partial \mu_i} \right)^{(k)} \end{aligned} \quad (\text{C20})$$

D'où la forme de l'équation (C17) ;

$$(X'W^{(k)}X)\beta^{(k+1)} = X'W^{(k)}z^{(k)} \quad (\text{C21})$$

Ces équations, sont les équations normales dans la méthode d'ajustement des moindres carrés pondérés pour un modèles linéaire, avec $z^{(k)}$ comme variable dépendante ; X , la matrice du modèle et les matrices $W^{(k)}$, sont les poids.

Les solutions de ces équations sont données par :

$$\beta^{(k+1)} = (X'W^{(k)}X)^{-1}(X'W^{(k)}z^{(k)}) \quad (\text{C22})$$

Dans cette formulation, le vecteur z est une forme linéarisée de la fonction de lien au point μ , évaluée en y , c'est à dire :

$$g(y_i) \simeq g(\mu_i) + (y_i - \mu_i)g'(\mu_i) = \eta_i + (y_i - \mu_i) \frac{\partial \eta_i}{\partial \mu_i} = z_i \quad (\text{C23})$$

Cette variable dépendante ajustée z , ou variable dépendante de travail ('Working dependant variate'), a pour i ème élément, la valeur approximée par $z_i^{(k)}$ pour le ki ème

cycle du schéma itératif. Pour ce cycle, on fait une régression de $z^{(k)}$ en X avec les poids $W^{(k)}$ pour obtenir un nouveau estimateur $\beta^{(k+1)}$. Cet estimateur nous donne une nouvelle valeur du prédicteur linéaire et qui est donnée par : $\eta^{(k+1)} = X\beta^{(k+1)}$ et une nouvelle valeur ajustée pour la variable dépendante, $z^{(k+1)}$ pour le prochain cycle.

L'estimateur du maximum de vraisemblance est la limite de $\beta^{(k)}$ quand $k \rightarrow \infty$.

En conclusion, l'estimateur du maximum de vraisemblance résulte de la méthode des moindres carrés pondérés itérativement, dont les poids changent à chaque cycle. Ce processus est appelé méthode des moindres carrés répondérés itérativement ('IRLS pour iterative reweighted least squares').

Pour commencer le processus, on utilise les données y comme valeurs initiales de l'estimateur de μ . Ceci détermine l'estimateur initial de la matrice de pondération W et alors, l'estimateur initial de β .