JOURNAL DE LA SOCIÉTÉ FRANÇAISE DE STATISTIQUE

HÉLÈNE JACQMIN-GADDA
RODOLPHE THIÉBAUT

Modèles de sélection pour données longitudinales gaussiennes : application à l'étude du vieillissement cognitif

Journal de la société française de statistique, tome 145, n° 2 (2004), p. 19-32

<http://www.numdam.org/item?id=JSFS_2004__145_2_19_0>

MODÈLES DE SÉLECTION POUR DONNÉES LONGITUDINALES GAUSSIENNES : APPLICATION À L'ÉTUDE DU VIEILLISSEMENT COGNITIF

Hélène JACQMIN-GADDA * et Rodolphe THIÉBAUT

RÉSUMÉ

Cet article concerne les méthodes d'analyse de données longitudinales gaussiennes lorsque la variable réponse est observée de façon incomplète. Si la probabilité d'observation de la variable réponse ne dépend que des valeurs des réponses observées aux temps précédents et éventuellement de covariables, les données manquantes sont ignorables et la méthode du maximum de vraisemblance fournit des estimateurs asymptotiquement non biaisés. Nous montrons cependant sur un exemple que l'estimateur empirique de la moyenne est biaisé. Si la probabilité d'observation dépend des valeurs non observées de la variable réponse, une approche fréquente consiste à modéliser conjointement la réponse et la probabilité d'observation en utilisant un modèle de sélection. L'objectif de cet article est de présenter les deux catégories de modèle de sélection en insistant sur les hypothèses sous-jacentes. L'utilisation et les limites d'un modèle variable-réponse dépendant et d'un modèle effets-aléatoires dépendant sont illustrées sur une étude de la détérioration cognitive du sujet âgé. Le premier modèle apparaît beaucoup plus sensible au choix des variables d'ajustement que le second. Malgré leurs faiblesses, ces modèles sont utiles pour évaluer la sensibilité des estimations à différentes hypothèses concernant le processus d'observation.

Mots clés : Données longitudinales, Données manquantes, Modèles mixtes, Modèles conjoints, Modèles de sélection.

ABSTRACT

This paper focus on methods to analyse gaussian longitudinal data when the outcome is not completely observed. When the probability to observe the outcome depends only on the past observed values of the outcome and possibly on covariates, the missing data are ignorable and the maximum likelihood estimates are asymptotically unbiased. However, we show in an example that the empirical estimate of the mean is biased in this case. When the probability to be observed depends on the unobserved values of the outcome, a frequently used approach is to jointly model

* Institut National de la Santé et de la Recherche Médicale, Équipe de Biostatistique E0338, 146 rue Léo Saignat, 33076 Bordeaux cedex, France.
INSERM E0338, ISPED, case 11, 146 rue Léo Saignat, 33076 Bordeaux cedex, France.
Tel : (33) 5 57 57 45 18; Fax (33) 5 56 24 00 81;
e-mail : helene.jacqmin-gadda@bordeaux.inserm.fr

the outcome and the probability to be observed in a selection model. The aim of this article is to present the two types of selection models highlighting the underlying assumptions. The use and the limits of an outcome dependent model and a random-effect dependent model are illustrated on a study of cognitive ageing. The first model appears more sensitive to covariates adjustment than the second one. Despite some weaknesses, these models are useful to investigate the sensitivity of the estimates to various hypotheses regarding the observation process.

Keywords : Joint modelling, Longitudinal data, Missing data, Mixed models, Selection models.

1. Introduction

Les données manquantes sont fréquentes dans les enquêtes prospectives biomédicales et de nombreux travaux ont été consacrés à l'analyse de données longitudinales lorsque la variable réponse est observée de façon incomplète (Little, 1995). Il est maintenant démontré que les méthodes d'analyse doivent être adaptées aux hypothèses plausibles concernant le mécanisme d'observation de la variable réponse (que nous noterons Y) afin d'éviter des biais de sélection majeurs. Une typologie des données manquantes a été proposée par Little (1995) et est détaillée dans l'article de Minini et Chavance (2004) publié dans ce volume. Lorsque les données manquantes sont ignorables, les estimateurs des paramètres de l'évolution de Y obtenus par la méthode du maximum de vraisemblance sur l'ensemble des Y observés sont asymptotiquement sans biais. Par contre, lorsque la probabilité d'observation dépend des valeurs non observées de Y, les données manquantes sont dites non aléatoires ou informatives et il est nécessaire de modéliser conjointement la variable d'intérêt Y et le processus d'observation M. Deux approches ont été proposées : la modélisation de la probabilité d'observation en fonction de Y (modèle de sélection) et la modélisation de Y en fonction du schéma d'observation (modèle de mélange).

Dans cet article, nous présentons les deux types de modèles de sélection pour l'analyse de données longitudinales gaussiennes incomplètes désignés respectivement par variable-réponse dépendant et effets-aléatoires dépendant. Le modèle variable-réponse dépendant proposé par Diggle et Kenward (1994) et le modèle effets-aléatoires dépendant proposé par Wulfsohn et Tsiatis (1997) et Henderson *et al.* (2000) sont ensuite appliqués à une étude longitudinale du déclin cognitif du sujet âgé. L'objectif est de mettre en évidence, d'une part, les biais des analyses naïves et, d'autre part, les hypothèses sous-jacentes à ces différents modèles et leurs limites. Les différences entre ces 2 types de modèles sont ensuite discutées.

2. Modèles pour données longitudinales gaussiennes incomplètes

2.1. Modèle linéaire à effets mixtes

Soit $Y_i^{'} = (Y_{i1}, Y_{i2}, ..., Y_{in_i})$ le vecteur complet des réponses pour le sujet i, $i = 1, ..., N$, aux temps $t_i = (t_{i1}, ..., t_{in_i})^{'}$. La variable Y étant supposée quantitative gaussienne, le modèle classiquement utilisé pour l'analyse de données longitudinales est le modèle linéaire à effets mixtes. Si on note X_i la matrice $n_i \times p$ de variables explicatives pour le sujet i et Z_i une matrice $n_i \times q$ (souvent sous-matrice de X_i), le modèle linéaire à effets mixtes peut s'écrire :

$$Y_i = X_i\beta + Z_i\alpha_i + e_i \quad (1)$$

avec β un vecteur d'effets fixes et α_i le vecteur d'effets aléatoires spécifique à chaque individu. Le vecteur α_i de dimension q suit une distribution gaussienne de moyenne $\mathbf{0}$ et de matrice de covariance G. Le vecteur des erreurs e_i est supposé gaussien, $e_i \sim N(0, \sigma^2 I_{n_i})$, et indépendant de α_i.

Un exemple simple fréquemment utilisé est le modèle linéaire à pente et intercept aléatoires qui suppose une évolution linéaire pour tous les sujets avec une pente et un intercept spécifiques à chaque individu : $Y_{ij} = \beta_0 + \beta_1 t_{ij} + \alpha_{0i} + \alpha_{1i} t_{ij} + e_{ij}$ pour $j = 1, ..., n_i$.

2.2. Estimation sur données incomplètes

Lorsque les données sont incomplètes, le vecteur Y_i peut être scindé en deux vecteurs correspondant respectivement aux mesures effectivement réalisées et aux mesures manquantes : $Y_i^{'} = (Y_i^{obs'}, Y_i^{mis'})$. En reprenant les notations de Minini et Chavance (2004), nous noterons M_{ij} la variable indicatrice de non observation ($M_{ij} = 1$ si Y_{ij} est manquante et $M_{ij} = 0$ si Y_{ij} est observée) et $M_i^{'} = (M_{i1}, M_{i2}, ..., M_{in_i})$. Dans la suite, nous supposerons que les données manquantes sont monotones, c'est-à-dire qu'elles sont uniquement dues aux sorties d'étude : $P(M_{ij+1} = 1 | M_{ij} = 1) = 1$. L'objectif des méthodes présentées est donc d'obtenir des estimateurs asymptotiquement sans biais des paramètres de la distribution $f(Y|X)$ définie par le modèle mixte à partir des données observées : Y_i^{obs} et M_i pour $i = 1, ...N$.

Lorsque $P(M_{ij} = 1) = f(Y_{i1}, ..., Y_{ij-1}, X_i)$, les données manquantes sont aléatoires et la vraisemblance conjointe des données observées $L(Y^{obs}, M)$ se factorise en deux termes indépendants (données manquantes ignorables). Les paramètres du modèle mixte peuvent donc être estimés sans biais (asymptotiquement) en maximisant la vraisemblance sur les seuls Y observés : $L(Y^{obs})$.

Lorsque la probabilité d'observation peut dépendre des Y non observés Y_i^{mis}, soit directement à travers la réponse courante, qui est non observée si $M_{ij} = 1$, soit indirectement par l'intermédiaire des effets aléatoires, les données manquantes sont non-aléatoires. Il est alors nécessaire de maximiser la vraisemblance conjointe $L(Y^{obs}, M)$ sur l'ensemble des données observées sur

les deux processus Y et M. Plusieurs décompositions de cette vraisemblance ont été proposées.

Dans l'approche par les modèles de mélange (Little 1993, Michiels *et al.* 2002, Molenberghs *et al.*, 2004), la distribution de Y est spécifiée conditionnellement au schéma d'observation M_i. Cette méthode présente l'inconvénient de nécessiter des contraintes supplémentaires afin que l'ensemble des paramètres soit identifiable pour tous les schémas de réponse M_i. Par exemple la pente n'est pas identifiable pour les sujets n'ayant qu'une observation. Par ailleurs, la distribution d'intérêt $f(Y|X)$ n'est pas directement estimée. Ses paramètres sont estimés dans un second temps, par la moyenne des estimations des distributions conditionnelles $f(Y|X,M)$ pondérée par les proportions de chaque schéma d'observation.

La seconde approche proposée est celle des modèles de sélection. Elle consiste à spécifier la distribution de M_i conditionnellement aux données observées Y_i^{obs} et non observées Y_i^{mis} de la variable réponse Y (modèle de sélection variable-réponse dépendant) ou conditionnellement aux effets aléatoires α_i (modèle de sélection effets-aléatoires dépendant). La vraisemblance conjointe est alors calculée en utilisant l'une des deux décompositions suivantes :

$$f(Y_i^{obs}, M_i|X_i) = \int f(Y_i^{obs}, Y_i^{mis}|X_i) f(M_i|Y_i^{obs}, Y_i^{mis}, X_i) dY_i^{mis}$$

ou

$$f(Y_i^{obs}, M_i|X_i) = \int f(Y_i^{obs}|X_i, \alpha_i) f(M_i|X_i, \alpha_i) f(\alpha_i) d\alpha_i$$

L'intérêt majeur de cette approche est l'estimation directe des paramètres de $f(Y|X)$. De plus, la modélisation de la procédure d'observation en fonction de la variable Y peut sembler plus naturelle que la modélisation de Y sachant M qui n'est généralement pas intéressante en elle-même sauf lorsque la sortie d'étude est causée par un événement identifié.

3. Modèles de sélection

3.1. Modèles de sélection variable-réponse dépendant

Diggle et Kenward (1994) ont proposé un modèle de sélection variable-réponse dépendant en spécifiant la probabilité de sortie d'étude en fonction des réponses passées et présentes à l'aide d'un modèle logistique. Nous en donnons une formulation légèrement différente dans laquelle la probabilité de sortie d'étude peut dépendre d'un vecteur de covariables ζ_i mais la dépendance sur les réponses passées est limitée à la réponse précédente (hypothèse raisonnable retenue dans la majorité des travaux antérieurs). En effet, pour être inclus dans l'étude, un sujet a, au minimum, une réponse observée, donc $Y_{i,j-1}$ est toujours observée. L'estimation du modèle devient beaucoup plus complexe si la sortie d'étude peut dépendre de k réponses précédentes alors que tous les sujets n'ont pas un minimum de k observations. Le modèle logistique s'écrit donc :

$$logit\{P(M_{ij} = 1|y_{ij-1}, y_{ij})\} = \zeta_i'\gamma + y_{ij-1}\eta_1 + y_{ij}\eta_2 \qquad (2)$$

où γ est un vecteur de paramètres associé au vecteur de covariables ζ_i. Dans ce modèle, si $\eta_1 = \eta_2 = 0$, la sortie d'étude dépend seulement des covariables; si $\eta_1 \neq 0$ et $\eta_2 = 0$, les données manquantes sont aléatoires et, si $\eta_2 \neq 0$, les données manquantes sont non aléatoires (ou informatives suivant la terminologie de Diggle et Kenward). Ce modèle permet donc théoriquement un test de l'hypothèse de données manquantes aléatoires mais ce test est très peu robuste à l'hypothèse de normalité des résidus, de même que les estimations des paramètres du modèle mixte (Kenward, 1998, Jacqmin-Gadda *et al.* 1999).

Les paramètres du modèle mixte pour Y et du modèle logistique pour M sont estimés par maximisation de la vraisemblance conjointe $L(Y^{obs}, R)$. En notant n_{oi} le nombre de mesures réellement effectuées sur le sujet i et en posant $H_{ij} = (y_{i1}, ..., y_{ij-1})$, et $P_{ij}(y_{ij}) = P(M_{ij} = 1|y_{ij-1}, y_{ij})$, la vraisemblance conjointe peut s'écrire (Diggle and Kenward, 1994) :

$$L = \prod_{i=1}^{N} f_{Y_{i1}}(y_{i1}) \prod_{j=2}^{n_{oi}} (1 - P_{ij}(y_{ij})) f_{Y_{ij}|H_{ij}}(y_{ij}) \left(\int P_{in_{oi}+1}(y) f_{Y_{in_{oi}+1}|H_{in_{oi}+1}}(y) dy \right)^{I_{\{n_{oi}<n_i\}}}$$

Le premier terme représente la densité conjointe $f(Y_{i1}, ..., Y_{in_{oi}}, M_{i1}, ..., M_{in_{oi}})$ pour les n_{oi} réponses observées. Le dernier terme ne concerne que les sujets sortis d'études (c'est-à-dire tels que le nombre de mesures observées n_{oi} est inférieur au nombre de mesures prévues n_i) et représente la probabilité de sortie d'étude sachant le passé $P(M_{in_{oi}+1} = 1|H_{in_{oi}+1})$ calculée en intégrant sur la valeur manquante $Y_{in_{oi}+1}$. L'intégrale (de dimension 1) n'a pas de solution analytique et les auteurs proposent une approximation de l'intégrale par approximation probit de la transformation logit. Cette méthode est disponible dans le logiciel OSWALD sous Splus (http ://www.maths.lancs.ac.uk/Software/Oswald/.). L'application présentée dans la section 4 a été réalisée à l'aide d'un programme Fortran utilisant l'algorithme d'optimisation de Marquardt (1963).

Cette approche suppose que les temps de mesure soient définis *a priori* afin d'identifier le temps t où la première mesure est manquante. De plus, ce modèle est plus pertinent si le délai entre deux mesures est approximativement constant pour toutes les mesures et tous les sujets puisque le paramètre η_1 qui traduit la dépendance entre la sortie d'étude et la mesure précédente est constant. Cela implique également l'absence de données manquantes intermittentes. Cette méthode peut cependant être utilisée sur des échantillons comportant des données manquantes intermittentes en supposant, d'une part, que les données manquantes intermittentes sont aléatoires (ce qui est souvent réaliste) et, d'autre part, que l'association entre la probabilité d'observation en t_{ij} et la réponse en t_{ij-1} est indépendante du délai $t_{ij} - t_{ij-1}$. Troxel (1998) a proposé une extension du modèle de Diggle et Kenward pour les données manquantes intermittentes dans le cas où le nombre de mesures par sujet ne dépasse pas 3 ou 4 et en imposant une structure markovienne du premier ordre pour la covariance des scores.

3.2. Modèles de sélection effets-aléatoires dépendant

Certains auteurs ont proposé de modéliser la distribution de la sortie d'étude en fonction des effets aléatoires du modèle mixte plutôt que de la valeur courante de la variable réponse. Dans le modèle de Wu et Caroll (1988), la probabilité de sortie d'étude suit un modèle probit, tandis que DeGruttola et Tu (1994) et Schluchter (1992) supposent une distribution log-normale pour le temps de sortie d'étude. Plus récemment, des approches semi-paramétriques utilisant le modèle des risques proportionnels ont été développées (Wulfsohn et Tsiatis 1997, Henderson *et al.* 2000). Le risque instantané de sortie d'étude $\lambda(t)$ s'écrit alors sous sa forme générale :

$$\lambda(t) = \lambda_0(t)exp(\zeta_i^{'}\gamma + f(\alpha_i)\eta)$$

Les deux formulations les plus fréquemment utilisées sont :

$$\lambda(t) = \lambda_0(t)exp(\zeta_i^{'}\gamma + \alpha_i^{'}\eta) \quad (3)$$

ou, en supposant un modèle linéaire à intercept et pente aléatoires pour Y :

$$\lambda(t) = \lambda_0(t)exp(\zeta_i^{'}\gamma + (\alpha_{0i} + \alpha_{1i}t)\eta) \quad (4)$$

Dans ces modèles, η est un paramètre (ou un vecteur de paramètres) mesurant l'association entre la sortie d'étude et les effets aléatoires : si $\eta = 0$, la sortie d'étude dépend seulement des covariables. En définissant T_i le temps de sortie d'étude, C_i le temps de censure (fin d'étude programmée), $T_{oi} = min(T_i, C_i)$ et δ_i l'indicateur de l'événement « sortie d'étude », $\delta_i = I_{\{T_i<C_i\}}$, les paramètres du modèle mixte et du modèle de survie sont estimés par maximisation de la vraisemblance conjointe de (Y, T_o, δ) qui s'écrit :

$$L = \prod_{i=1}^{N} \int f_{Y_i|\alpha_i}(y_i|\alpha) S_{T_i|\alpha_i}(t_{oi}|\alpha) \left(\lambda_{T_i|\alpha_i}(t_{oi}|\alpha)\right)^{\delta_i} f_{\alpha_i}(\alpha) d\alpha$$

où $S_{T_i|\alpha_i}(t|\alpha)$ est la valeur en t de la fonction de survie pour la variable T_i, soit $P(T_i > t|\alpha)$. L'intégrale, dont la dimension est égale au nombre d'effets aléatoires q, n'a pas de solution analytique et doit être calculée numériquement. Les auteurs proposent un algorithme EM utilisant une intégration par quadrature gaussienne. Pour l'application, nous avons utilisé une macro SAS développée par D. Renard, et les variances des estimateurs ont été estimées par Bootstrap sur 100 rééchantillonnages. On note que dans l'algorithme EM, la vraisemblance observée L n'est pas calculée ; les modèles sont donc comparés d'après les estimations des paramètres et leurs variances.

Un aspect important commun à ces différents modèles effets-aléatoires dépendants est que le risque de sortie d'étude ne dépend que de la tendance à long terme de l'évolution du sujet représentée par les effets aléatoires (qui ne varient pas avec le temps). En particulier, dans le modèle (4), le risque de sortie d'étude ne dépend pas directement de la valeur courante de la variable étudiée comme dans le modèle de Diggle et Kenward, mais seulement de

l'écart courant entre la prédiction individuelle et la valeur moyenne prédite pour une population de mêmes caractéristiques. Henderson *et al.* (2000) ont proposé une extension dans laquelle le modèle mixte et le modèle de survie peuvent dépendre d'un processus gaussien stationnaire (généralement un processus autorégressif) qui décrit les écarts individuels au modèle d'évolution du marqueur à court terme. L'estimation de ce modèle est cependant difficile car elle nécessite l'approximation numérique d'une intégrale de grande dimension pour chaque sujet.

Les modèles effets-aléatoires dépendants ne nécessitent pas un calendrier des mesures communs à tous les sujets ni des mesures régulièrement espacées. Il est cependant nécessaire de disposer de critères de définition du temps de sortie d'étude (ce qui est parfois difficile lorsque les temps de visite ne sont pas déterminés *a priori*).

4. Application

4.1. Objectif

Nous allons illustrer sur un exemple les biais induits par les analyses naïves portant sur des données longitudinales incomplètes ainsi que l'intérêt et les limites de la prise en compte des données manquantes par des modèles de sélection. L'objectif de l'analyse était l'étude de la détérioration cognitive, mesurée par l'évolution du score au test des codes de Wechsler, chez le sujet âgé non dément. Le test des codes de Wechsler est un test de raisonnement logique simple qui nécessite une bonne capacité d'attention et est réalisé en temps limité (90 secondes). Ce dernier point le rend sensible au vieillissement en dehors de toute pathologie. Le score varie de 0 à 76 dans notre échantillon avec une distribution très proche d'une gaussienne, un score élevé traduisant un bon niveau cognitif.

4.2. Données

Les données sont issues de la cohorte Paquid mise en place en 1988 pour étudier le vieillissement cognitif normal et pathologique. Elle comporte 2792 sujets de 65 ans ou plus à l'inclusion et vivant à domicile au début de l'étude en Gironde. Les sujets ont été interviewés à domicile lors de la visite initiale (temps T0) puis 1, 3, 5, 8 et 10 ans plus tard (temps T1 à T10). Chaque visite comprenait la réalisation d'une série de tests psychométriques dont le test des codes de Wechsler (excepté à 3 ans) et un diagnostic de démence.

L'échantillon d'analyse comporte 2026 sujets non diagnostiqués déments entre T0 et T10 et ayant fait le test au moins 1 fois (à T0). Le temps de sortie d'étude (l'événement) est défini comme le suivi à partir duquel le sujet n'a plus fait le test des codes ; le temps de sortie d'étude est censuré lorsque le sujet a effectué le test à T10. L'échantillon comporte des données manquantes intermittentes qui sont supposées aléatoires. Pour le modèle de sélection variable-réponse dépendant, on suppose que l'association entre la probabilité de sortie d'étude et le score précédent ne dépend pas du temps écoulé entre les deux visites.

On peut noter qu'il s'agit d'une hypothèse forte car le délai entre deux visites devrait varier entre 1 et 4 ans d'après le protocole de l'étude Paquid et les données manquantes intermittentes induisent une variabilité supplémentaire. La répartition des sujets encore dans l'étude et des sujets ayant effectué le test à chaque visite est présentée dans le tableau 1.

TABLEAU 1. – Répartition des sujets selon le suivi (Cohorte Paquid, 1988-1998)

	T0	T1	T5	T8	T10
Dans l'étude	2 026	1 494	1 040	816	630
	(100 %)	(74 %)	(51 %)	(40 %)	(31 %)
Testés	2 026	1 315	919	720	630
	(100 %)	(65 %)	(45 %)	(35 %)	(31 %)

4.3 Modèle

L'évolution du score au test des codes de Wechsler est décrit par un modèle linéaire à pente et intercept aléatoires. Le temps de base est le temps depuis l'entrée dans la cohorte ($t_{ij} \in \{0, 1, 5, 8, 10\}$) et l'âge du sujet au début de l'étude est inclus en variable explicative en 4 classes : 65-69 ans (classe de référence), 70-74 ans, 75-79 ans, 80 ans et plus. Pour tenir compte d'un effet primo-passation précédemment mis en évidence (Jacqmin-Gadda *et al.*, 1997), le modèle inclut une variable indicatrice pour la visite initiale :

$$Y_{ij} = (\beta_0 + age_i'\gamma_0 + \alpha_{0i}) + (\beta_1 + age_i'\gamma_1 + \alpha_{1i}) \times t_{ij} + \beta_3 I_{\{t_{ij}=0\}} + e_{ij} \quad (5)$$

avec $\alpha_i = (\alpha_{0i}\ \alpha_{1i})' \sim N(0, G)$ et $e_{ij} \sim N\left(0, \sigma_e^2\right)$.

Le vecteur age_i est constitué des 3 variables indicatrices pour les classes d'âge et les vecteurs d'effets fixes γ_0 et γ_1 sont donc également de dimension 3 : $\gamma_0' = (\gamma_{01}\ \gamma_{02}\ \gamma_{03})$ et $\gamma_1' = (\gamma_{11}\ \gamma_{12}\ \gamma_{13})$.

4.4 Résultats

La figure 1 représente les moyennes observées à chaque suivi et par groupe d'âge de début d'étude, calculées d'une part, sur l'ensemble des scores disponibles et, d'autre part, sur les 432 sujets ayant effectué le test aux 5 visites. Il apparaît clairement que les sujets ayant des données complètes ne sont pas représentatifs de l'ensemble de l'échantillon car ils ont initialement un score moyen au test des codes nettement supérieur à celui de l'ensemble de l'échantillon. Par ailleurs, après une légère amélioration entre T0 et T1, le score moyen de ces sujets décline entre T1 et T10 dans toutes les classes d'âge. Si l'on considère l'évolution des scores moyens calculés sur l'ensemble des données disponibles, on constate que le déclin est nettement plus faible, voire inexistant, parmi les sujets de 80 ans et plus. Ces résultats traduisent la dépendance entre la sortie d'étude et le score au test des codes mais ne permettent pas de déterminer si les données manquantes sont aléatoires ou non.

Cette figure souligne également les biais potentiels d'une analyse réalisée sur les sujets ayant des données complètes puisqu'ils ne sont pas représentatifs

de l'ensemble de l'échantillon, ou d'une analyse naïve (n'utilisant pas les estimateurs du maximum de vraisemblance) sur les données disponibles : le déclin serait alors considérablement sous-évalué.

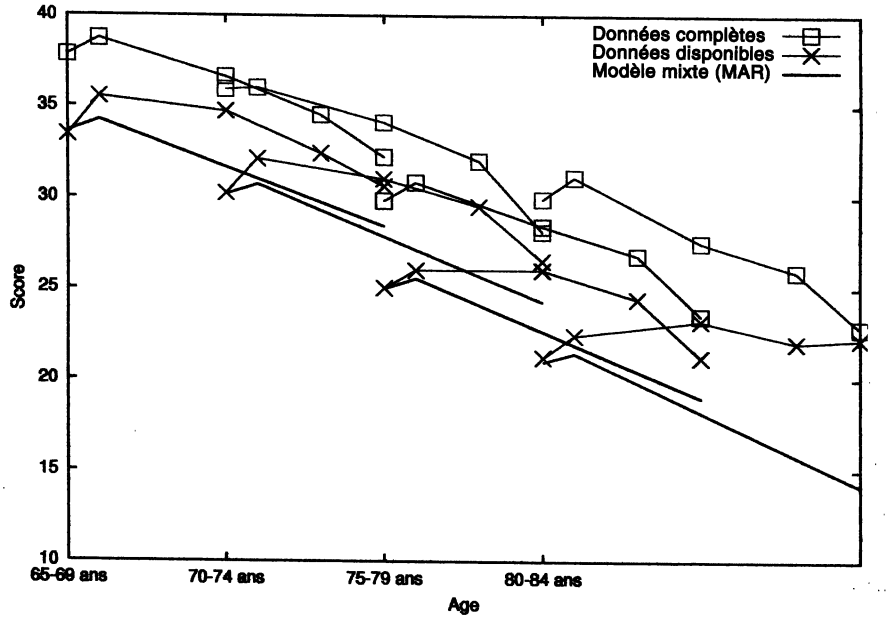

FIG 1. — Moyennes empiriques et moyennes estimées par le modèle mixte pour le score au test des codes de Wechsler.

Le modèle mixte (5) a été estimé par le maximum de vraisemblance, sous l'hypothèse de données manquantes ignorables et l'évolution moyenne estimée est représentée sur la figure 1. Ces courbes représentent l'évolution que l'on aurait dû observer si aucun sujet n'était sorti de l'étude avant T10. Le niveau initial correspond à la moyenne de l'ensemble de l'échantillon, mais la pente estimée est plus proche de la pente des sujets avec données complètes. L'écart entre les courbes estimées et les courbes observées n'est pas un indicateur d'une mauvaise spécification du modèle mais plutôt de l'impact des données manquantes. Pour évaluer l'ajustement du modèle, il serait préférable de calculer, pour chaque visite, la moyenne des valeurs prédites pour les sujets observés à cette visite, et de comparer la courbe obtenue à la courbe observée sur les données disponibles.

Le tableau 2 présente les estimations des pentes et différences de pente $(\beta_1, \gamma_{11}, \gamma_{12}, \gamma_{13})$ selon l'âge initial avec le modèle (5) et divers modèles de Diggle et Kenward. Dans le modèle MAR (pour Missing at Random) la probabilité de sortie d'étude ne dépend que du score précédent ; les données manquantes sont donc supposées ignorables et le modèle mixte et le modèle logistique sont estimés séparément. La somme des log-vraisemblances des deux sous-modèles est $-21\ 108.2$. La probabilité de sortie d'étude est significativement plus élevée lorsque le score précédent est faible. Dans le modèle M1, la probabilité de sortie d'étude dépend également du score courant, mais ce paramètre est non significatif, suggérant que les données

manquantes sont aléatoires. La log-vraisemblance est d'ailleurs identique à celle du modèle MAR et les paramètres du modèle mixte estimés sont très proches. Le modèle M2 comprend les mêmes variables plus un ajustement du risque de sortie d'étude sur le suivi et l'âge initial du sujet (en 4 classes) qui constituent les covariables ζ_i de la formulation générale (2). La log-vraisemblance globale est améliorée (−21 059.0) car le suivi et l'âge sont associés au risque de sortie d'étude : la probabilité de sortie d'étude augmente avec l'âge et est plus élevée au suivi à 5 ans. Le score précédent reste associé négativement au risque de sortie d'étude mais, contrairement au modèle non-ajusté M1, un score courant élevé augmente significativement le risque de sortie d'étude. Ajusté sur l'âge et le suivi, la sortie d'étude apparaît donc informative. Les associations négatives avec le score précédent et positives avec le score courant s'interprètent plus aisément en reparamétrant le modèle M2 :

$$logit(P_{ij}) = \zeta_i'\gamma - 0.047(\frac{Y_{ij-1} + Y_{ij}}{2}) + 0.041(Y_{ij} - Y_{ij-1})$$

Le risque de sortie d'étude est donc associé à un score moyen faible et à une amélioration du score entre t_{ij-1} et t_{ij}. En tenant compte de cette sortie d'étude informative, les détériorations estimées sont plus faibles, en particulier parmi les sujets les plus âgés ($\hat{\gamma}_{13} = -0.09$). Le modèle M3 est présenté pour insister sur la nécessité d'inclure le score précédent dans le modèle logistique. Sans cet ajustement, la probabilité de sortie d'étude apparaît significativement plus élevée lorsque le score courant est faible et cela induit une augmentation importante de la valeur absolue des pentes estimées. La valeur de la vraisemblance (-21079.8) montre cependant que le modèle M3 ajuste moins bien les données que le précédent.

TABLEAU 2. – Pente et différences de pentes estimées selon l'âge et le modèle avec un modèle de sélection variable-réponse dépendant (Modèle mixte + modèle logistique)

	Pente		Différence de pentes		Différence de pentes		Différence de pentes	
	65 − 69 ans		70 − 74 ans		75 − 79 ans		80 − 84 ans	
Modèle	$\hat{\beta}_1$	$\hat{\sigma}(\hat{\beta}_1)$	$\hat{\gamma}_{11}$	$\hat{\sigma}(\hat{\gamma}_{11})$	$\hat{\gamma}_{12}$	$\hat{\sigma}(\hat{\gamma}_{12})$	$\hat{\gamma}_{13}$	$\hat{\sigma}(\hat{\gamma}_{13})$
MAR	−0.65	(0.031)	−0.071	(0.047)	−0.082	(0.053)	−0.16	(0.081)
M1	−0.65	(0.031)	−0.071	(0.047)	−0.082	(0.054)	−0.17	(0.088)
M2	−0.63	(0.031)	−0.064	(0.047)	−0.053	(0.054)	−0.09	(0.086)
M3	−0.67	(0.031)	−0.080	(0.047)	−0.122	(0.053)	−0.27	(0.082)

MAR $logit(P_{ij}) = \gamma_0 - 0.061(0.0033)Y_{ij-1}$, L = −21108.2

M1 $logit(P_{ij}) = \gamma_0 - 0.061(0.011)Y_{ij-1} - 0.0003(0.013)Y_{ij}$, L = −21108.2

M2 $logit(P_{ij}) = \zeta_i'\gamma - 0.088(0.014)Y_{ij-1} + 0.041(0.016)Y_{ij}$, L = −21059.0

M3 $logit(P_{ij}) = \zeta_i'\gamma - 0.056(0.0042)Y_{ij}$, L = −21079.8

Le tableau 3 présente les résultats du modèle effets-aléatoires dépendant semi-paramétrique défini par (3) ou (4). Pour satisfaire aux contraintes du logiciel utilisé le modèle mixte inclut un effet aléatoire pour l'effet primo-passation mais les estimations des pentes sont peu modifiées (comme on peut le vérifier en comparant les modèles MAR des tableaux 2 et 3). Six formulations du modèle des risques proportionnels pour le risque instantané de sortie d'étude ont été comparées : les 3 premiers modèles sont non-ajustés tandis que les 3 derniers sont ajustés sur l'âge du sujet à la visite initiale (le temps de base étant toujours la durée depuis l'entrée dans la cohorte). Dans les modèles M4 et M7, le risque instantané de sortie d'étude est associé négativement à la déviation individuelle courante (calculée en incluant l'effet aléatoire pour l'effet primo-passation) : le risque de sortie d'étude est donc plus élevé pour les sujets ayant une prédiction individuelle au temps t inférieure à la moyenne pour la population. Dans les modèles M5 et M8, on constate que le risque de sortie d'étude est associé à un niveau initial inférieur à la moyenne et à un déclin plus marqué. Enfin, dans les modèles M6 et M9, le risque reste

TABLEAU 3. – Pente et différences de pentes estimées selon l'âge et le modèle avec un modèle de sélection effets-aléatoires dépendant (Modèle mixte + modèle de Cox)

	Pente		Différence de pentes		Différence de pentes		Différence de pentes	
	65 – 69 ans		70 – 74 ans		75 – 79 ans		80 – 84 ans	
Modèle	$\hat{\beta}_1$	$\hat{\sigma}(\hat{\beta}_1)$	$\hat{\gamma}_{11}$	$\hat{\sigma}(\hat{\gamma}_{11})$	$\hat{\gamma}_{12}$	$\hat{\sigma}(\hat{\gamma}_{12})$	$\hat{\gamma}_{13}$	$\hat{\sigma}(\hat{\gamma}_{13})$
MAR	−0.63	(0.030)	−0.071	(0.046)	−0.085	(0.053)	−0.16	(0.081)
M4	−0.69	(0.032)	−0.077	(0.047)	−0.085	(0.054)	−0.16	(0.088)
M5	−0.65	(0.030)	−0.078	(0.046)	−0.096	(0.052)	−0.16	(0.086)
M6	−0.64	(0.028)	−0.079	(0.048)	−0.096	(0.056)	−0.16	(0.090)
M7	−0.68	(0.029)	−0.081	(0.041)	−0.098	(0.050)	−0.19	(0.098)
M8	−0.65	(0.029)	−0.077	(0.049)	−0.095	(0.046)	−0.16	(0.082)
M9	−0.65	(0.028)	−0.077	(0.041)	−0.095	(0.046)	−0.16	(0.081)

$W_i(t) = \alpha_{0i} + \alpha_{1i}t + \alpha_{2i}I_{\{t=0\}}$

M4 $\lambda(t) = \lambda_0(t)exp(-0.032(0.005)W_i(t))$

M5 $\lambda(t) = \lambda_0(t)exp(-0.033(0.003)\alpha_{0i} - 0.12(0.052)\alpha_{1i})$

M6 $\lambda(t) = \lambda_0(t)exp(-0.0065(0.0026)W_i(t) - 0.031(0.0030)\alpha_{0i} + 0.095(0.095)\alpha_{1i})$

M7 $\lambda(t) = \lambda_0(t)exp(\zeta_i'\gamma - 0.031(0.007)W_i(t))$

M8 $\lambda(t) = \lambda_0(t)exp(\zeta_i'\gamma - 0.035(0.0028)\alpha_{0i} - 0.13(0.044)\alpha_{1i})$

M9 $\lambda(t) = \lambda_0(t)exp(\zeta_i'\gamma - 0.0060(0.0026)W_i(t) - 0.032(0.0034)\alpha_{0i} + 0.068(0.087)\alpha_{1i})$

fortement associé à un niveau initial bas et plus faiblement à une déviation individuelle courante négative mais il n'est plus associé à la pente individuelle. On notera que les estimations du modèle mixte dans cette approche sont beaucoup moins sensibles à la formulation du modèle de sortie d'étude que dans le modèle de Diggle et Kenward. Dans tous les modèles, la décroissance du score est d'autant plus marquée que le sujet est âgé et le test de la différence de pente par rapport à la classe de référence reste toujours proche du niveau de signification ($0.05 < p < 0.12$ pour les 70-74 ans, $0.04 < p < 0.11$ pour les 75-80 ans et $0.05 < p < 0.07$ pour les 80 ans et plus selon les modèles).

5. Discussion

L'analyse de l'évolution cognitive chez des sujets âgés non déments a permis d'illustrer l'impact des données manquantes sur des analyses naïves et de comparer deux types de modèles de sélection pour l'analyse de données longitudinales gaussiennes comportant des sorties d'étude non ignorables.

Les différences majeures entre ces deux approches doivent être gardées à l'esprit lors des analyses. Dans le modèle de Diggle et Kenward, le risque de sortie d'étude dépend de la valeur courante et des valeurs précédentes de la variable d'intérêt et donc plutôt de l'évolution récente. Il est très sensible aux variables d'ajustement incluses dans le modèle de sortie d'étude. En pratique, il est recommandé d'inclure dans le modèle logistique le score au temps précédent et les variables d'ajustement introduites dans le modèle mixte. En effet, si ces variables d'ajustement expliquent la dépendance entre la sortie d'étude et la réponse courante, les données manquantes sont ignorables. Dans notre exemple, on observe la situation inverse : la sortie d'étude n'est associée à la réponse courante qu'après ajustement sur l'âge et le suivi, elle devrait donc être traitée comme non ignorable. On doit cependant considérer ce résultat avec précaution car il a été montré précédemment que ce test était très sensible aux écarts à l'hypothèse de normalité des résidus qui est impossible à vérifier sur des données incomplètes. Les problèmes d'identifiabilité de ce modèle ont d'ailleurs été abondamment discutés (discussion de Diggle et Kenward, 1994). En effet, la probabilité de non réponse dépend de la valeur manquante et les données observées apportent peu d'information sur cette dernière : les estimations reposent donc essentiellement sur les hypothèses paramétriques du modèle. Etant donnée l'importance de l'hypothèse de normalité dans ce modèle, il pourrait être intéressant de l'étendre pour intégrer d'autres distributions de l'erreur résiduelle et ainsi permettre l'évaluation de la sensibilité des estimations à différentes hypothèses concernant le processus d'observation et la distribution de la variable d'intérêt.

Au contraire, les données observées apportent une information non négligeable pour l'estimation des effets aléatoires. Ceci peut expliquer que les modèles effets-aléatoires dépendants soient relativement robustes aux écarts à l'hypothèse de normalité des effets aléatoires (Song *et al.*, 2002). Par ailleurs, dans les modèles effets-aléatoires dépendants n'incluant pas de processus gaussien, le risque de sortie d'étude dépend de la tendance à long terme de l'évolution de la variable d'intérêt et uniquement des déviations individuelles par rapport aux prédictions moyennes. Les estimations issues de ces modèles sont donc moins sensibles aux variables d'ajustement incluses dans le modèle de sortie d'étude.

Un critère de choix entre les deux modèles présentés est également le schéma d'étude. Nous avons vu que le modèle de Diggle et Kenward nécessite que le calendrier des mesures soit déterminé *a priori* et que les mesures soient

régulièrement espacées tandis que les modèles effets-aléatoires dépendants sont plus souples sur ces aspects. Ce critère ne doit cependant pas être essentiel. Le modèle de Diggle et Kenward pourrait en effet être assoupli en introduisant par exemple une interaction entre la mesure précédente et le délai entre les mesures dans le modèle logistique. Il est en effet intéressant de laisser la possibilité de comparer les résultats de différents modèles.

Ces deux modèles de sélection ont en commun la limitation aux données manquantes monotones et, à des degrés différents, la nécessité d'hypothèses paramétriques sur le processus de non-réponse qui sont en pratique impossibles à vérifier. Ils sont cependant utiles, de même que les modèles de mélange, pour évaluer la sensibilité des estimations obtenues sous l'hypothèse de données manquantes ignorables à d'autres hypothèses concernant le processus d'observation.

Remerciements : Les auteurs remercient Didier Renard pour la mise à disposition de la macro SAS pour l'estimation des paramètres du modèle de sélection effets alétoires dépendant et l'équipe vieillissement de l'unité INSERM 593 pour l'utilisation des données de la cohorte Paquid.

Références

DE GRUTTOLA V. and TU X.M. (1994). Modelling progression of CD4-lymphocyte count and its relationship to survival time. *Biometrics*, **50**, 1003-14.

DIGGLE P.J. and KENWARD M.G. (1994). Informative dropout in longitudinal data analysis. *Applied statistics*, **43**, 49-93.

HENDERSON R., DIGGLE P. and DOBSON A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics*, **1**, 465-80.

JACQMIN-GADDA H., FABRIGOULE C., COMMENGES D. and DARTIGUES J.F. (1997). A five-year longitudinal study of Mini-Mental State Examination in normal aging. *American Journal of Epidemiology*, **145**, 498-506.

JACQMIN-GADDA H., COMMENGES D. and DARTIGUES J.F. (1999). Analyse de données longitudinales gaussiennes comportant des données manquantes sur la variable à expliquer. *Revue d'épidémiologie et Santé publique*, **47**, 525-34.

KENWARD M.G. (1998). Selection models for repeated measurements with non-random dropout : an illustration of sensitivity. *Statistics in Medicine*, **17**, 2723-32.

LITTLE R.J.A. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, **88**, 125-34.

LITTLE R.J.A. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association*, **90**, 112-21.

LITTLE R.J.A. and RUBIN D.B. (1987). *Statistical analysis with missing data.* Wiley : New-York.

MARQUARDT D.W. (1963). An algorithm for least squares estimation of nonlinear parameters. *SIAM Journal*, **11**, 431-41.

MICHIELS B., MOLENBERGHS G., BIJNENS L., VANGENEUGDEN T. and THIJS H. (2002). Selection models and pattern-mixture models to analyse longitudinal quality of life data subject to drop-out. *Statistics in Medicine*, **21**, 1023-41.

MININI P. et CHAVANCE M. (2004). Observations longitudinales incomplètes : de la modélisation des observations disponibles à l'analyse de sensibilité. *Journal de la Société Française de Statistique*, **145**, 2, 5-18.

MOLENBERGHS G., THIJS H., MICHIELS B., VERBEKE G. and KENWARD M.G. (2004). Pattern-Mixture Models (2004). *Journal de la Société Française de Statistique*, **145**, 2, 49-77.

SCHLUCHTER M.D. (1992). Methods for the analysis of informatively censored longitudinal data. *Statistics in Medicine*, **11**, 1861-70.

SONG X., DAVIDIAN M. and TSIATIS A.A. (2002). A semiparametric likelihood approach to joint modeling of longitudinal and time-to-event data. *Biometrics*, **58**, 742-53.

TROXEL A.B., HARRINGTON D.P. and LIPSITZ S.R. (1998). Analysis of longitudinal data with non-ignorable non-monotone missing values. *Applied statistics*, **57**, 425-38.

WU M.C. and CARROLL R.J. (1988). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*, **44**, 175-88.

WULFSOHN M.S. and TSIATIS A.A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics*, **53**, 330-9.