

PASCAL MININI

MICHEL CHAVANCE

**Observations longitudinales incomplètes : de
la modélisation des observations disponibles
à l'analyse de sensibilité**

Journal de la société française de statistique, tome 145, n° 2 (2004),
p. 5-18

http://www.numdam.org/item?id=JSFS_2004__145_2_5_0

© Société française de statistique, 2004, tous droits réservés.

L'accès aux archives de la revue « Journal de la société française de statistique » (<http://publications-sfds.math.cnrs.fr/index.php/J-SFdS>) implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques

<http://www.numdam.org/>

OBSERVATIONS LONGITUDINALES INCOMPLÈTES : DE LA MODÉLISATION DES OBSERVATIONS DISPONIBLES À L'ANALYSE DE SENSIBILITÉ

Pascal MININI et Michel CHAVANCE *

RÉSUMÉ

La modélisation d'observations longitudinales incomplètes pose des problèmes de biais et de précision. La typologie des mécanismes d'observation proposée par Little (1995) permet d'identifier les situations où il n'y a pas de problèmes de biais (données manquantes complètement au hasard), celles où le biais est évitable sans qu'il soit nécessaire de modéliser simultanément le processus d'observation et le vecteur des réponses (données manquantes ignorables) et celles où une telle modélisation est indispensable. Nous présentons les différentes stratégies d'analyse des observations longitudinales incomplètes en nous étendant sur les modèles de sélection et les modèles par mélange. Lorsque les données manquantes sont accidentelles, il est généralement impossible de s'assurer que les données manquantes sont ignorables et les approches usuelles sont potentiellement biaisées. Nous recommandons d'effectuer une analyse de sensibilité et nous présentons une méthode adaptée au cas d'observations longitudinales binaires. Nous montrons comment estimer les paramètres d'intérêt par imputation multiple et nous illustrons la méthode par un exemple.

Mots clés : Observations longitudinales, observations incomplètes, modèle linéaire généralisé, analyse de sensibilité.

ABSTRACT

The modelling of incomplete longitudinal observations raises problems of bias and precision. Little's typology of observation processes (1995) identifies situations where there is no risk of bias (data missing completely at random), those where the bias can be corrected without modelling simultaneously the observation process and the vector of outcomes (ignorable missing data) and those where such a modelling is required. We present the different modelling strategies for incomplete longitudinal data, in particular selection models and pattern-mixture models. When the missing data are fortuitous, it is generally impossible to assess whether they are ignorable. It is recommended to perform a sensitivity analysis and a method suited to longitudinal binary data is presented. The parameters of interest can be estimated using multiple imputation. An application is presented.

Keywords : Longitudinal observations, incomplete observations, generalized linear model, sensitivity analysis.

* INSERM U472 Paris XI 16 avenue P. Vaillant-Couturier F94807 Villejuif
chavance@vjf.inserm.fr

1. Introduction et notations

La modélisation d'observations incomplètes pose des problèmes de biais et de précision dont l'acuité dépend d'une part du mécanisme d'observation et d'autre part de la stratégie d'analyse utilisée. En particulier, la situation est fondamentalement différente selon que les données manquantes sont planifiées ou accidentelles. Nous nous intéresserons avant tout dans cet article au cas le plus complexe, celui des données manquantes accidentelles. Pour simplifier les notations, nous supposerons que le protocole d'étude prévoit un même nombre n d'observations pour chacun des N sujets de l'échantillon. Appelons Y_{ij} la jème valeur de la variable réponse, observée ou non ($j = 1, \dots, n$) du sujet i . Nous représenterons le processus d'observation à l'aide de l'indicatrice M_{ij} de l'absence de Y_{ij} ($M_{ij} = 1$ si Y_{ij} est manquant, $M_{ij} = 0$ si Y_{ij} est observé). Nous noterons Y_i^{obs} la composante observée et Y_i^{mis} la composante manquante du vecteur Y_i . Little (1995) a proposé une typologie des mécanismes d'observation qui élargit dans le cas longitudinal celle de Little et Rubin (1987) en ajoutant aux trois catégories des données manquantes complètement au hasard, au hasard et non au hasard un niveau intermédiaire entre les deux premiers : les données manquantes covariable-dépendantes. On peut décomposer la distribution conjointe des variables réponses et des indicatrices de données manquantes en produit de la probabilité conditionnelle de M_i sachant le vecteur complet Y_i et de la distribution marginale de Y_i

$$P_{\theta\beta} [M_i, Y_i | X_i, b_i] = P_{\theta} [M_i | Y_i, X_i, b_i] P_{\beta} [Y_i | X_i, b_i] \quad (1)$$

où les paramètres θ et β qui spécifient chacune des distributions sont entièrement distincts et où b_i représente le vecteur des effets aléatoires associé au sujet i . Nous supposerons que le vecteur X_{ij} des variables explicatives au temps j est entièrement observé. C'est le cas par exemple s'il correspond aux caractéristiques du sujet à l'entrée dans l'étude et au temps j de l'observation. Les observations successives sur le sujet i sont supposées indépendantes conditionnellement à b_i , ce qui correspond à un modèle linéaire généralisé mixte sur les vecteurs complets des réponses

$$g [E(Y_{ij} | X_{ij}, b_i)] = X_{ij}\beta + Z_{ij}b_i. \quad (2)$$

Les observations manquantes sont dites complètement au hasard si la probabilité qu'une observation soit manquante est constante

$$P_{\theta} [M_{ij} = 1 | Y_i, X_{ij}, b_i] = \pi. \quad (3)$$

Une supposition qui n'est tenable que si le protocole d'étude garantit une probabilité d'observation constante. Elles sont dites covariable-dépendantes si cette probabilité ne dépend que des covariables observées

$$P_{\theta} [M_{ij} | Y_i, X_{ij}, b_i] = P_{\theta} [M_{ij} | X_{ij}]. \quad (4)$$

Une supposition elle aussi irréaliste si le recueil incomplet n'est pas planifié dans ce but. Il est en général plus raisonnable de supposer que la probabilité

qu'une observation soit incomplète dépend aussi du vecteur des réponses à observer Y_i . Les données manquantes sont dites au hasard si cette probabilité ne dépend que de la composante observée

$$P_\theta [M_{ij} | Y_i, X_i, b_i] = P_\theta [M_{ij} | Y_i^{obs}, X_{ij}, b_i]. \quad (5)$$

C'est le cas par exemple dans les épreuves d'hypersensibilité bronchique où la fonction ventilatoire d'un sujet est mesurée après inhalation de doses croissantes de méthacholine, mais où l'administration de cette molécule est interrompue si la fonction ventilatoire du sujet tombe en dessous d'un certain seuil : les mesures correspondantes sont donc manquantes. C'est aussi le cas de schémas d'étude complexes, comme les enquêtes cas-témoins emboîtées dans une cohorte où la probabilité d'un recueil complet de l'information est fonction de l'état de santé du sujet. En revanche les données sont qualifiées de manquantes non au hasard, si la probabilité qu'une observation soit incomplète dépend aussi des réponses non observées Y_i^{mis} .

Remarquons que les quatre types de données manquantes sont emboîtés, chacun est un cas particulier des types suivants. L'intérêt de cette classification est de définir des situations où le risque de biais n'existe pas (données manquantes complètement au hasard) d'autres où il existe mais peut être évité sans qu'il soit besoin de modéliser le processus d'observation (données covariables dépendantes ou manquantes au hasard) et d'autres où la modélisation conjointe des réponses et du processus d'observation est nécessaire. Quand les données sont manquantes au hasard ou à plus forte raison covariable-dépendantes, et que les paramètres β et θ sont distincts, il est possible d'estimer sans biais le paramètre d'intérêt β sans modéliser le processus d'observation M_i car celui-ci n'apporte aucune information sur le paramètre β de la distribution marginale des réponses. La vraisemblance des paramètres de la distribution jointe des réponses et du processus d'observation s'écrit

$$\begin{aligned} L(\beta, \theta | Y_i^{obs}, M_i, X_i) &= \int \int P [M_i, Y_i | X_i, b_i] dY_i^{mis} db_i \\ &= \int \int P_\theta [M_i | Y_i^{obs}, X_i, b_i] P_\beta [Y_i | X_i, b_i] dY_i^{mis} db_i \\ &= \int P_\theta [M_i | Y_i^{obs}, X_i, b_i] \left\{ \int P_\beta [Y_i^{obs}, Y_i^{mis} | X_i, b_i] dY_i^{mis} \right\} db_i \\ &= L(\theta | Y_i^{obs}, M_i, X_i) L(\beta | Y_i^{obs}, X_i). \end{aligned} \quad (6)$$

C'est dans la mesure où l'estimateur du maximum de vraisemblance de β ne dépend que des réponses observées Y_i^{obs} que les données manquantes au hasard sont qualifiées d'ignorables. Notons qu'en revanche si les données manquantes ne sont pas au hasard ou si les paramètres β et θ ne sont pas distincts, il n'est pas possible de factoriser la vraisemblance comme ci-dessus ni, par conséquent d'effectuer des inférences sur β sans modéliser simultanément le processus d'observation et la distribution complète des réponses. Les données manquantes sont alors qualifiées d'informatives. Malgré

l'utilité conceptuelle indiscutable de cette typologie, son utilité pratique reste limitée par l'impossibilité de tester l'hypothèse de données manquantes au hasard contre celle de données manquantes non au hasard. Il faudrait pour cela connaître les valeurs des données manquantes, au moins pour un sous-échantillon représentatif. Il est donc généralement souhaitable d'évaluer les conséquences sur les conclusions de l'étude de données manquantes non au hasard.

Les observations longitudinales incomplètes peuvent également être caractérisées par la structure monotone ou intermittente des données manquantes. Quand les données manquantes résultent toutes d'une interruption du suivi, elles ont une structure monotone

$$M_{ij} = 1 \quad \Rightarrow \quad M_{ik} = 1 \quad \forall k > j. \quad (7)$$

Dans le cas contraire, elles sont dites intermittentes. Cette distinction n'a pas de conséquence en termes de biais. En revanche certaines procédures sont plus faciles à mettre en oeuvre sur des données manquantes monotones que sur des données manquantes intermittentes.

2. Stratégies d'analyse

Dans les études transversales, où une seule observation est prévue sur chaque sujet, la méthode d'analyse la plus courante est l'analyse des données complètes : que les données manquantes concernent la variable réponse modélisée Y ou les covariables X , les logiciels courants estiment les paramètres du modèle d'intérêt sur le sous-échantillon complètement observé. Cette approche expose au risque d'un biais de sélection. Dans le cas d'observations longitudinales avec réponses incomplètes, la stratégie la plus courante est l'analyse des données observées : dans la mesure où, comme nous le supposons ici, les observations manquantes sont ignorables et où les covariables sont entièrement observées, l'estimateur du maximum de vraisemblance appliqué aux données observées est asymptotiquement sans biais.

Une autre approche est utilisée fréquemment dans l'industrie pharmaceutique : le report de la dernière valeur observée. Cette méthode consiste à remplacer les valeurs manquantes, en particulier pour les sujets sortis d'étude ou perdus de vue par la dernière valeur observée sur le sujet. Elle est cependant difficile à justifier. Dans le meilleur des cas, elle présente l'avantage d'être conservatrice. Encore faut-il que la tendance observée dans les groupes de traitement comparés soit une amélioration monotone en fonction du temps. On ne peut cependant exclure qu'elle aggrave le biais qu'entraînerait l'analyse des données disponibles supposant les observations manquantes au hasard.

L'imputation de la valeur la plus favorable (respectivement la plus défavorable) fournit les limites que les estimateurs obtenus à partir des données complètes ne peuvent franchir, compte tenu des données observées. C'est une méthode facile à mettre en oeuvre mais rarement utilisée car peu d'études verraient leurs conclusions résister à un traitement aussi radical (Unnebrink et Windeler, 1999). Dans un essai thérapeutique longitudinal qui conclurait à

la supériorité du nouveau traitement en supposant les données manquantes au hasard, il faudrait à la fois un faible pourcentage de données manquantes et une grande différence entre les traitements, pour qu'en remplaçant les réponses manquantes dans le groupe de référence par la meilleure réponse possible et les données manquantes dans le groupe «nouveau traitement» par la moins bonne réponse possible, l'effet traitement reste significatif.

Lorsque les réponses Y_i suivent une distribution normale, il est possible d'utiliser le modèle proposé par Diggle et Kenward (1994) pour estimer simultanément les paramètres de la distribution marginale des réponses et ceux du processus d'observation supposé informatif. Ces auteurs se placent dans le cadre de données à structure monotone (sorties d'études). Ils utilisent la décomposition (1) en spécifiant pour la probabilité de sortie un modèle logistique. La date de sortie du sujet i est notée D_i

$$D_i = j \quad \Rightarrow \quad M_{ik} = 0, \forall k < j, \quad M_{ik} = 1, \forall k \geq j \quad (8)$$

et le modèle de sortie s'écrit

$$\text{logit}(P[D_i = j | Y_i]) = \Psi_0 + \Psi_1 Y_{i(j-1)} + \Psi_2 Y_{ij}. \quad (9)$$

Dans ce modèle, le cas particulier $\Psi_2 = 0$ correspond à des données manquantes au hasard puisque la probabilité de sortie ne dépend alors que de la dernière réponse observée et le cas particulier $\Psi_1 = \Psi_2 = 0$ correspond à des observations manquantes complètement au hasard. Tous les paramètres étant identifiables, il est formellement possible de tester l'hypothèse de données manquantes au hasard. Cependant, il est rapidement apparu que l'inférence sur l'ensemble des paramètres de ce modèle dépendait fortement de l'hypothèse de normalité des observations complètes, par définition invérifiable (Rubin, 1994, Little, 1994b, Kenward, 1998). Quand la distribution observée est asymétrique, par exemple en raison d'une queue de distribution plus lourde à droite qu'à gauche, le modèle implique une sélection des valeurs observées avec une probabilité de sortie élevée quand Y_{ij} est petit et une probabilité faible quand Y_{ij} est grand ($\Psi_2 < 0$). Cependant, l'asymétrie pourrait aussi bien correspondre à la réalité, par exemple en raison d'une distribution log-normale ou d'une hétérogénéité de la population (présence d'un sous groupe de sujets atypiques).

Les modèles qui utilisent la décomposition (1) sont appelés modèles par sélection puisqu'ils utilisent la probabilité qu'une observation soit sélectionnée conditionnellement à la valeur sous-jacente. La plupart des publications concernent des observations normales, mais les modèles par sélection peuvent également s'appliquer à des données catégorielles (Molenberghs *et coll.*, 1999). En se limitant à deux mesures, ils conduisent alors à utiliser un modèle log-linéaire pour représenter les réponses Y_{ij} et les indicatrices M_{ij} , $j=1,2$ sous la forme

$$P[M_{ij}, Y_{ij}] = P[M_{ij} | Y_{ij}] P[Y_{ij}]. \quad (10)$$

Il est également possible d'utiliser la décomposition de la distribution jointe de Y_i et M_i en produit de la distribution marginale de M_i et de la distribution

conditionnelle de Y_i sachant M_i . Les anglo-saxons parlent alors de « pattern-mixture model ». La terminologie française n'est pas fixée, mais nous utilisons l'expression « modèle par mélange » qui a l'avantage de la concision et qui exprime bien le fait que la distribution observée est un mélange de distributions. En outre les prépositions évitent le risque de confusion entre la modélisation par mélange et la modélisation des mélanges qui pourrait être à l'origine de la distinction entre « pattern-mixture models » et « mixture models ». Ces modèles sont surparamétrés, par exemple les sujets pour lesquels une seule observation est disponible ne permettent pas d'estimer l'effet du temps sur les réponses et les sujets avec seulement deux observations ne permettent d'estimer qu'un effet linéaire du temps. Il est cependant possible d'estimer tous les paramètres en choisissant des contraintes d'identification (Little 1993, 1994a, Molenberghs *et coll.*, 2004).

Bien qu'il s'agisse d'observations répétées et non longitudinales, un exemple éclairant concerne les données récoltées à l'occasion d'un sondage d'opinion réalisé en Slovénie à l'occasion du référendum sur l'indépendance (Rubin, Stern et Vehovar, 1995, Molenberghs Kenward et Goetghebeur, 2001). Son intérêt vient de ce que le résultat du référendum fournit une valeur de référence à laquelle comparer les résultats des différentes analyses, 88,5 % des inscrits ont voté et se sont prononcés pour l'indépendance. Trois questions concernaient le référendum, 1) Êtes-vous favorable à l'indépendance de la Slovénie ? 2) Êtes-vous favorable à la sécession de la Slovénie ? 3) Participerez vous au référendum ? Sur les 2074 électeurs interrogés 72,7 % ont répondu aux trois questions, mais individuellement le pourcentage de données manquantes variait de 15,8 % à 17 % selon la question. Outre les réponses complètes, les 7 combinaisons possibles de réponses incomplètes ont été observées. Pour estimer la participation et le pourcentage de votants en faveur de l'indépendance, Rubin, Stern et Vehovar (1995) ont comparé 5 méthodes. L'analyse des données complètes est limitée aux 1454 sujets qui ont répondu aux trois questions, et celle des données disponibles utilise les réponses des 1549 sujets qui ont répondu aux questions 1) et 3). L'imputation de la valeur la plus défavorable consiste à considérer que les non répondant à la question 1) sont des opposants à l'indépendance et les non répondant à la 3) des non votant. Pour effectuer une analyse sous l'hypothèse de données manquantes au hasard, Rubin, Stern et Vehovar (1995) ont utilisé un modèle log-linéaire saturé pour représenter les réponses (OUI ou NON) aux trois questions puis ont obtenu l'estimateur du maximum de vraisemblance appliqué aux données disponibles (OUI, NON, ou Non Réponse) par l'algorithme EM (Dempster, Laird et Rubin, 1977). Enfin, ils ont envisagé un modèle pour observations manquantes informatives qui suppose que la probabilité d'observer une réponse dépend de la valeur à observer mais non des autres réponses et qui ajuste un modèle à 17 paramètres aux 27 observations disponibles du tableau à 6 dimensions des réponses aux trois questions et des indicatrices de données manquantes. Par rapport au résultat du référendum, (88,5 % des inscrits en faveur de l'indépendance), l'analyse sous l'hypothèse d'observations manquantes au hasard fournissait une estimation satisfaisante (88,3 %) alors que l'analyse des données complètes (92,8 %) ou disponibles (92,9 %) le surestimait et que l'es-

timation obtenue en imputant la valeur la plus défavorable le sous estimait (69,4 %). Le modèle pour données manquantes informatives qu'ils ont utilisé sous-estimait également le pourcentage de oui (78,2 %). On notera en particulier qu'une analyse supposant que la probabilité d'observation à une question dépend des autres réponses mais non de la réponse manquante permet de prédire correctement le résultat du référendum. Cependant la mauvaise performance du modèle pour données manquantes informatives ne doit pas être interprétée dans le sens de l'inutilité des modèles pour données manquantes informatives, mais dans celui de la nécessité d'envisager plusieurs modèles et d'effectuer une analyse de sensibilité. Molenberghs Kenward et Goetghebeur (2001), analysant les mêmes données ont montré que plusieurs modèles pour données manquantes informatives, reposant sur d'autres hypothèses, conduisaient à des estimations compatibles avec le résultat final.

3. Analyse de sensibilité pour observations longitudinales incomplètes

Lorsqu'il n'est pas possible de tester de façon crédible l'hypothèse de données manquantes au hasard contre celle de données manquantes non au hasard, ce qui supposerait d'avoir pu récupérer les observations manquantes sur un sous-échantillon représentatif de l'ensemble des informations incomplètes, il est recommandé d'effectuer une analyse de sensibilité, c'est-à-dire d'analyser les données en envisageant différents scénarios pour le processus d'observation.

Une analyse de sensibilité peut être effectuée dans le cadre de modèles par mélange, L'analyse de sensibilité, telle qu'elle est définie par Little (1994a) consiste à effectuer des inférences sur les paramètres d'intérêt pour une série de valeurs plausibles d'un paramètre de sensibilité définissant la relation entre processus d'observation et valeurs manquantes. Il existe déjà une bibliographie relativement abondante sur le sujet (Rotnitzky, Robins et Scharfstein, 1998), Scharfstein, Rotnitzky, et Robins, 1999, Rotnitzky *et coll.*, 2001, Molenberghs Kenward et Goetghebeur (2001), Birmingham Rotnitzky et Fitzmaurice, 2003, Minini et Chavance, 2004a).

Reprenons les notations du premier paragraphe et utilisons un modèle log-linéaire (Bishop, Fienberg et Holland, 1975) pour représenter la distribution conjointe du vecteur complet des réponses binaires (observées ou non) $Y_i = (Y_{i1}, \dots, Y_{in})'$ et de l'indicatrice de données manquantes $M_i = (M_{i1}, \dots, M_{in})'$ qui peut être non structurée

$$\begin{aligned} \log P[Y_i, M_i] = & \mu + \sum_{j=1}^n \lambda_j Y_{ij} + \sum_{j < k} \sum_{k=2}^n \lambda_{jk} Y_{ij} Y_{ik} + \sum_{j=1}^n \theta_j M_{ij} \\ & + \sum_{j < k} \sum_{k=2}^n \theta_{jk} M_{ij} M_{ik} + \sum_{j=1}^n \psi y_{ij} M_{ij}. \end{aligned} \quad (11)$$

Un intérêt de ce modèle est qu'il s'applique aussi bien à des données manquantes non structurées que monotones. L'écriture (11) suppose que les termes

d'interaction d'ordre supérieur à deux sont nuls, une contrainte généralement raisonnable mais qu'il est évidemment possible de lever. Dans cette formule, μ est un terme de normalisation, le terme suivant modélise la distribution marginale des réponses, le troisième terme prend en compte l'association entre les réponses d'un même sujet, les deux termes suivants modélisent respectivement la fréquence marginale des données manquantes et les associations de données manquantes, enfin, le caractère informatif des données manquantes se reflète dans le terme d'interaction ψ entre Y et M que nous appellerons paramètre de sensibilité. En particulier les observations sont manquantes au hasard si $\psi = 0$. Un modèle plus général, incluant des termes d'interaction quelconques ψ_{jk} entre y_{ij} et m_{ik} a été envisagé par Baker, Rosenberger et Dersimonian (1992) et utilisé en analyse de sensibilité par Molenberghs, Kenward et Goetghebeur (2001).

Une spécificité de ce modèle est que le paramètre ψ s'interprète en termes de modèles par mélange comme en termes de modèles par sélection. Du point de vue des mélanges, on montre que (11) conduit à

$$\log P[Y_i|M_i] = \mu_0(M_i) + \sum_{j=1}^n \lambda_j Y_{ij} + \sum_{j < k} \sum_{k=2}^n \lambda_{jk} Y_{ij} Y_{ik} + \psi \sum_{j=1}^n Y_{ij} M_{ij} \quad (12)$$

où $\mu_0(M_i)$ est un terme de normalisation, spécifique au profil d'observation M_i . Un paramètre de sensibilité ψ positif (respectivement négatif) exprime ici que les réponses $Y_{ij} = 1$ sont d'autant plus probables (respectivement moins probables) que le sujet présente davantage de données manquantes.

Du point de vue des modèles par sélection, (11) conduit également à

$$\log P[M_i|Y_i] = \nu_0(M_i) + \sum_{j=1}^n \theta_j M_{ij} + \sum_{j < k} \theta_{jk} M_{ij} M_{ik} + \psi \sum_{j=1}^n Y_{ij} M_{ij} \quad (13)$$

d'où la probabilité que la jème observation soit manquante, sachant le vecteur des réponses et la valeur des $n - 1$ autres indicatrices

$$\log P[M_{ij} = 1 | Y_i, M_{ik} k \neq j] = \theta_j + \sum_{l=1}^n \theta_{il} M_{il} + \psi Y_{ij} \quad (14)$$

qui montre que le paramètre de sensibilité est également le log odds ratio de la probabilité qu'une observation soit manquante, entre sujets ayant répondu positivement et sujets ayant répondu négativement.

Il est évidemment possible d'introduire dans le modèle l'effet de covariables X sur la distribution des réponses Y et des indicatrices de données manquantes (Minini et Chavance, 2004b), voire d'introduire une interaction d'ordre trois entre une variable explicative, par exemple une indicatrice de traitement, les réponses Y et le processus d'observation M afin d'exprimer que la relation réponse-observation dépend du traitement et vaut ψ_g dans le groupe de traitement g.

4. Estimation

Soit (ϕ, ψ) le vecteur des paramètres du modèle où ϕ représente l'ensemble des paramètres à l'exception du paramètre de sensibilité. Ces deux composantes sont identifiables, et peuvent être estimées, par exemple en utilisant l'algorithme EM (Dempster, Laird et Rubin, 1977).

4.1. Estimation du paramètre de sensibilité

Bien que ψ soit identifiable, son estimation dépend fortement d'hypothèses non vérifiables et les observations contenant peu d'information sur sa valeur, la log-vraisemblance est généralement trop plate pour en obtenir une estimation utilisable, même quand le modèle est correctement spécifié, comme l'illustrent les simulations que nous avons effectuées. Nous avons généré des échantillons de $n=50$, et $n=500$ paires d'observations telles que $P[Y_1] = P[Y_2] = 0,5$, et reliées par un odds ratio valant $\exp(1)$. Les données manquantes étaient ignorables ($\psi = 0$) et leur proportion pouvait être de 10 %, 20 %, 30 %, 40 % ou 50 %. Pour chaque configuration, 10 000 échantillons ont été générés. Dans de nombreux cas, la vraisemblance ne présentait pas de maximum et l'estimateur du maximum de vraisemblance n'était pas défini (estimation infinie). De plus, même lorsque la vraisemblance présentait un maximum, l'intervalle de confiance à 95 % construit sur le test du rapport de vraisemblance était souvent de longueur infinie. Nous le qualifions de semi-infini quand une seule borne était finie et d'infini quand aucune des bornes ne l'était. Le tableau 1 montre que même pour un échantillon de 500 sujets il est impossible d'obtenir une estimation précise du paramètre de sensibilité. Il est donc préférable de le fixer et d'examiner comment l'estimation de ϕ varie en fonction des valeurs choisies pour ψ , c'est-à-dire d'effectuer une analyse de sensibilité.

TABLEAU 1. – Proportion de simulations avec estimation infinie de ψ , Intervalle de Confiance (IC) à 95 % infini ou semi-infini

| Nombre de sujets | % données manquantes | Estimation infinie | IC à 95 % infini | IC à 95 % semi-infini |
|------------------|----------------------|--------------------|------------------|-----------------------|
| 50 | 10 % | 53 % | 84 % | 16 % |
| | 20 % | 43 % | 77 % | 22 % |
| | 30 % | 40 % | 77 % | 21 % |
| | 40 % | 39 % | 78 % | 20 % |
| | 50 % | 40 % | 79 % | 19 % |
| 500 | 10 % | 4 % | 10 % | 72 % |
| | 20 % | 2 % | 6 % | 55 % |
| | 30 % | 2 % | 8 % | 49 % |
| | 40 % | 2 % | 14 % | 49 % |
| | 50 % | 5 % | 24 % | 49 % |

4.2. Estimation de ϕ à ψ fixé

L'estimation du maximum de vraisemblance des paramètres du modèle 11, à ψ fixé, peut être obtenue par l'algorithme EM. Les statistiques exhaustives sont les effectifs n_{MY} correspondant aux 2^{2n} modalités du vecteur (Y, M) . On n'observe que 3^n statistiques $n_{MY^{obs}}$, puisque chaque Y_{ij} peut prendre les modalités 0, 1 ou manquant. À l'étape E, on calcule les espérances des statistiques exhaustives conditionnellement aux observations, à la valeur courante ϕ^t du paramètre à estimer et à la valeur choisie pour ψ . À l'étape M, on cherche la valeur ϕ^{t+1} qui maximise l'espérance de la log vraisemblance (Minini et Chavance, 2004b). Cette valeur peut être obtenue à l'aide d'un logiciel pour modèles linéaires généralisés, comme SAS PROC GENMOD, en stipulant une distribution de Poisson et un terme de compensation (offset) $\log(N) + \psi \sum y_{ij} m_{ij}$, où N représente le nombre de sujets de l'échantillon. La variance de l'estimateur peut être obtenue par différenciation numérique (Jamshidian et Jenrich, 2000) et les probabilités marginales $P_{\hat{\phi}(\psi)}[Y]$ s'obtiennent par sommation des probabilités jointes de Y et M prédites par le modèle.

Une démarche analogue peut être mise en œuvre quand le modèle d'intérêt fait intervenir des variables explicatives X liées aux réponses et au processus d'observation. Cependant les paramètres d'intérêt que l'on aurait estimés à partir des observations complètes ne coïncident pas nécessairement avec les paramètres du modèle log linéaire spécifiant la distribution jointe de Y et M sachant X . En ce cas une approche utile est d'utiliser les probabilités déduites du modèle log linéaire pour réaliser des imputations multiples (Rubin 1987, Chavance et Manfredi, 2000).

5. Imputation multiple

L'imputation multiple consiste à fabriquer plusieurs jeux de données complétées en tenant compte des différents niveaux d'incertitude sur les valeurs manquantes. Un modèle de prédiction est estimé à partir des observations disponibles. Pour tenir compte du caractère stochastique du modèle, on ne remplace pas les données manquantes par la valeur prédite, mais par une valeur tirée au hasard dans la distribution prédite. Pour tenir compte de l'incertitude sur les paramètres du modèles d'imputation, on tire au hasard K ($K \geq 2$) vecteurs de paramètres dans la loi asymptotique de l'estimateur et on fabrique un jeu de données complétées avec chacun. Chacun de ces K jeux de données complétées est analysé et fournit un estimateur des paramètres d'intérêt de l'étude. L'estimateur de l'imputation multiple permet d'en effectuer la synthèse : son espérance est la moyenne arithmétique des K espérances, sa variance est la somme d'une composante inter-imputation et d'une composante intra-imputation.

Pratiquement la procédure permettant d'effectuer une analyse de sensibilité est la suivante

- 1) construire un modèle $P[Y, M|\phi, \psi]$;
- 2) choisir une valeur de ψ ;
- 3) obtenir l'estimation du maximum de vraisemblance $\hat{\phi}(\psi)$ et sa variance asymptotique $Var[\hat{\phi}(\psi)]$;

- 4) tirer une valeur ϕ^* de ϕ dans la loi asymptotique de son estimateur $N(\hat{\phi}(\psi), \widehat{Var}[\hat{\phi}(\psi)])$;
- 5) tirer les observations manquantes Y^{mis} dans la loi conditionnelle $P[Y^{mis}|Y^{obs}, M, \phi^*, \psi]$;
- 6) répéter $K \geq 2$ fois les étapes 4) et 5) pour obtenir K jeux de données complets;
- 7) analyser les K jeux de données complets;
- 8) synthétiser les K ensembles de résultats en utilisant l'estimateur de l'imputation multiple;
- 9) retourner en 2) et choisir une nouvelle valeur de ψ ou interrompre la procédure. Une macro SAS et un exemple sont disponibles sur le site

<http://ifr69.vjf.inserm.fr/~u472/EQUIPES/BIOSTATISTIQUE/savoir.html>

6. Illustration

Cette méthode a été appliquée aux données d'un essai thérapeutique comparant deux traitements de l'asthme. Pendant les 12 semaines de l'essai, le contrôle de l'asthme était évalué sur la base d'un questionnaire hebdomadaire concernant la présence de symptômes et la consommation médicale. L'asthme était déclaré contrôlé pendant un mois donné si le contrôle avait été effectif pendant au moins trois semaines de ce mois. Au total, 119 patients ont été randomisés dans le groupe « nouveau traitement » et 127 dans le groupe de « référence ». Les observations étaient complètes chez 69 % des sujets, mais on peut penser que les données manquantes n'étaient pas complètement au hasard car les patients complètement observés étaient plus nombreux dans le groupe nouveau traitement que dans le groupe de référence.

Le modèle utilisé pour l'imputation incluait des covariables (groupe de traitement, sexe, statut tabagique) ainsi qu'une interaction traitement \times contrôle \times observation, afin de pouvoir stipuler une valeur différente du paramètre de sensibilité dans chacun des groupes de traitement. Il semble naturel de supposer que les sujets du groupe nouveau traitement ont mieux participé à l'étude parce qu'ils recevaient un traitement plus efficace que ceux du groupe de référence. On ne peut cependant exclure que les patients les mieux contrôlés oublient de remplir leur questionnaire hebdomadaire. Il est donc souhaitable d'envisager des valeurs positives et négatives pour le paramètre de sensibilité. Des valeurs dans l'intervalle $[-1, 1]$ correspondant à des odds ratios compris entre 0,37 et 2,7 semblent raisonnables, mais des valeurs plus extrêmes sont envisageables.

Les données complètes ont été analysées à l'aide d'un modèle logistique marginal où l'effet du traitement, X_1 était ajusté sur le sexe, X_2 , le tabac, X_3 , et le mois ($j = 1, 3$, considéré comme une variable qualitative et d'indicatrices X_4 et X_5) :

$$\text{logit}P[Y_j = 1|X] = \mu_j + \sum_{k=1}^n \beta_k X_k. \quad (15)$$

Étant donné la faible dimension de la matrice des variances-covariances, une matrice non structurée a été utilisée pour modéliser les corrélations.

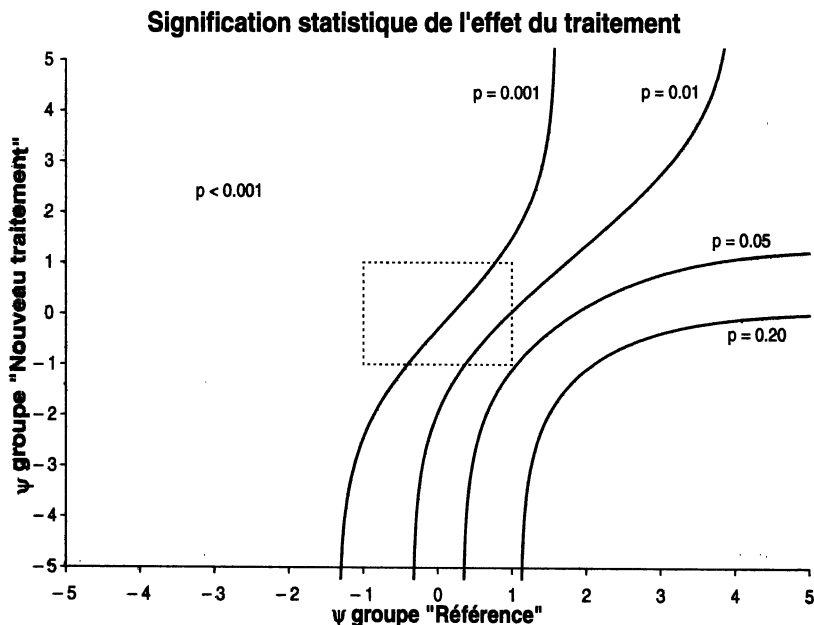


FIG 1. — Résultat de l'analyse de sensibilité : contour de la signification statistique de l'effet du traitement pour différentes valeurs de ψ dans chaque groupe.

Les résultats de l'analyse de sensibilité sont donnés sur la Figure 1. Si l'on suppose que les données sont manquantes au hasard, on peut conclure à la supériorité du nouveau traitement, avec un odds ratio en sa faveur de 2,15 ($p < 0,001$). Si l'on suppose que les données sont manquantes non au hasard, et que les paramètres de sensibilité dans chacun des groupes sont compris dans l'intervalle $[-1, 1]$, même dans l'hypothèse où le processus de sélection aurait favorisé le nouveau traitement en augmentant le pourcentage apparent de patients contrôlés ($\psi = -1$) et défavorisé le traitement de référence ($\psi = 1$), les observations permettent de conclure à la supériorité du nouveau traitement. Seul un processus de sélection impliquant à la fois des paramètres de sélection de signe opposé et de valeur absolue élevée autoriserait une remise en cause de cette conclusion.

7. Discussion

Il est nécessaire de s'interroger sur l'effet que peuvent avoir les données manquantes sur les conclusions d'une étude. Cependant, le développement d'outils de modélisation pour observations incomplètes ne permet évidemment pas de «sauver» une étude mal conduite. Ces outils constituent à l'inverse une incitation supplémentaire à soigner le recueil de l'information, car seules les études où l'essentiel de l'information est recueilli peuvent conduire à des conclusions susceptibles de résister aux hypothèses envisagées sur le processus d'observation.

Le modèle log-linéaire (11) a parfois été critiqué, en particulier parce que l'interprétation des paramètres dépend du nombre d'observations sur chaque

unité statistique. Il n'est cependant utilisé ici que pour obtenir un modèle d'imputation, et ce sont d'autres paramètres qui sont estimés dans la phase d'analyse des jeux de données complétés. Dans cette phase d'analyse, on aurait pu également utiliser une autre approche. D'une part on aurait pu préférer un modèle mixte à un modèle marginal ; l'estimateur de l'imputation multiple a été proposé comme une approximation de l'estimateur du maximum de vraisemblance et reste valide dans ce cadre. D'autre part, on aurait pu utiliser une analyse pondérée pour adapter les équations d'estimation généralisées (Diggle *et coll.*, 2002) au cas des observations incomplètes (Robins, Rotnitzky et Zhao, 1995). L'inverse des probabilités des différents profils de réponses complets, tels qu'ils sont prédits par le modèle sont alors utilisés comme poids. Il faut cependant souligner que lorsque ces probabilités sont estimées, comme ici, une variance correcte des estimateurs des paramètres d'intérêt doit prendre en compte l'incertitude sur les poids, ce qui n'est pas le cas dans la procédure GENMOD de SAS lorsque l'option WEIGHT est utilisée.

Les conclusions de l'étude de sensibilité présentée dépendent des présupposés du modèle utilisé dans la phase d'imputation des données, en particulier de ceux concernant le processus d'observation qui ne peuvent être évalués. Plus précisément, nous avons supposé que le mécanisme de sélection ne dépendait que de la réponse éventuellement manquante et non des réponses antérieures ou postérieures et que la valeur du paramètre de sensibilité ne pouvait varier qu'en fonction du groupe de traitement. Ce sont des hypothèses relativement restrictives qu'il est possible de lever au prix d'une complexification du modèle, et d'une plus grande difficulté d'interprétation. Il faut par ailleurs noter que les bornes obtenues pour le paramètre d'intérêt en supposant, par exemple, $\psi = -1$ dans le groupe nouveau traitement et $\psi = 1$ dans le groupe de référence restent valables pour tout modèle où le paramètre de sensibilité pourrait varier dans chacun des groupes tout en restant supérieur à -1 dans le groupe traitement et inférieur à 1 dans le groupe de référence.

Références

- BAKER S.G., ROSENBERGER W.F. and DERSIMONIAN R. (1992). Closed-form estimates for missing counts in two-way contingency tables. *Statistics in medicine* **11**, 643–657.
- BIRMINGHAM J., ROTNITZKY A. and FITZMAURICE G.M. (2003). Pattern-mixture and selection models for analysing longitudinal data with monotone missing patterns. *Journal of the royal statistical society, Serie B* **65**, 275–297.
- BISHOP Y.M.M., FIENBERG S.E. and HOLLANDP.W. (1975). *Discrete multivariate analysis : theory and practice*, MIT Press : Cambridge, Massachusetts.
- CHAVANCE M. et MANFREDI R. (2000). Modélisation d'observations incomplètes. *Revue d'épidémiologie et de santé publique* **48**, 389–400.
- DEMPSTER A.P., LAIRD N.M. and RUBIN D.B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *Journal of the royal statistical society, Serie B* **39**, 1–38.
- DIGGLE P.J. and KENWARD M.G. (1994). Informative drop-out in longitudinal data analysis (with discussion). *Applied statistics* **43**, 49–93.
- DIGGLE P.J., HEAGERTY P., LIANG K.Y. and ZEGER S.L. (2002). *Analysis of longitudinal data, 2nd edition*, Oxford University Press : Oxford.
- JAMSHIDIAN M. and JENNRICH R.I. (2000). Standard errors for EM estimation. *Journal of the royal statistical society, Serie B* **62**, 257–270.

- KENWARD M.G. (1998). Selection and models for repeated measurements with nonrandom dropout : An illustration of sensitivity. *Statistics in medicine* **17**, 2723–2732.
- LITTLE R.J.A. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the american statistical association* **88**, 125–134.
- LITTLE R.J.A. (1994a). A class of pattern-mixture models for normal data. *Biometrika* **81**, 471–483.
- LITTLE R.J.A. (1994b). Discussion to Diggle and Kenward : Informative drop-out in longitudinal data analysis. *Applied statistics* **43**, 78.
- LITTLE R.J.A. (1995). Modelling the drop-out mechanism in repeated measures studies. *Journal of the american statistical association* **90**, 1112–1121.
- LITTLE R.J.A. and RUBIN D.B. (1987). *Statistical analysis with missing data* Wiley : New-York.
- MININI P. and CHAVANCE M. (2004a). Sensitivity analysis of longitudinal normal data with drop-outs. *Statistics in medicine* **23**, 1039–1054.
- MININI P. and CHAVANCE M. (2004b). Sensitivity analysis of longitudinal binary data with non-monotone missing values. *Biostatistics* (sous presse).
- MOLENBERGHS G., GOETGHEBEUR E., LIPSITZ S.R. and KENWARD M.G. (1999). Nonrandom missingness in categorical data : strengths and limitations. *The American Statistician* **53**, 110–118.
- MOLENBERGHS G., KENWARD M.G. and GOETGHEBEUR E. (2001). Sensitivity analysis for incomplete contingency tables : the Slovenian plebiscite case. *Applied statistics* **50**, 15–29.
- MOLENBERGHS G., THIJS H., MICHIELS B., VERBEKE G. and KENWARD M.G. (2004). Pattern-mixture models. *Journal de la Société Française de Statistique* **145**, 2, 49–77.
- ROBINS J.M., ROTNITZKY A. and ZHAO L.P. (1995). Analysis of semi-parametric regression models for repeated outcomes in the presence of missing data. *Journal of the american statistical association* **90**, 106–121.
- ROTNITZKY A., ROBINS J.M. and SCHARFSTEIN D. (1998). Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the american statistical association* **93**, 1321–1339.
- ROTNITZKY A., SCHARFSTEIN D., SU T.L. and ROBINS J.M. (2001). Methods for conducting sensitivity analysis of trials with potentially nonignorable competing causes of censoring. *Biometrics* **57**, 103–113.
- RUBIN D.B. (1987). *Multiple imputations for nonresponse in surveys*. Wiley : New-York.
- RUBIN D.B. (1994). Discussion to Diggle and Kenward : Informative drop-out in longitudinal data analysis. *Applied statistics* **43**, 80–82.
- RUBIN D.B., STERN H.S., and VEHOVAR V. (1995). Handling “don’t know” in survey responses : the case of the Slovenian plebiscite. *Journal of the american statistical association* **90**, 822–828.
- SCHARFSTEIN D., ROTNITZKY A. and ROBINS J.M. (1999). Adjusting for nonignorable dropout using semiparametric nonresponse models (with discussion). *Journal of the american statistical association* **94**, 1096–1146.
- UNNEBRINK K. and WINDELER J. (1999). Sensitivity analysis by worst and best case assessment : is it really sensitive?. *Drug information journal* **33**, 835–839.