



HAL
open science

Modèles à effets aléatoires pour l'analyse de la Qualité de Vie en Oncologie

Antoine Barbieri, Caroline Mollevi, Christian Lavergne

► **To cite this version:**

Antoine Barbieri, Caroline Mollevi, Christian Lavergne. Modèles à effets aléatoires pour l'analyse de la Qualité de Vie en Oncologie. 46ièmes journées de Statistique (Société Française de Statistiques), Jun 2014, Rennes, France. hal-01099667

HAL Id: hal-01099667

<https://hal.science/hal-01099667>

Submitted on 5 Jan 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

MODÈLES À EFFETS ALÉATOIRES POUR L'ANALYSE DE LA QUALITÉ DE VIE EN ONCOLOGIE

Antoine Barbieri¹⁻² & Caroline Mollevi¹ & Christian Lavergne²

¹ *Institut régional de Cancérologie de Montpellier - Unité de Biostatistiques - 208, Rue des Apothicaires - Parc Euromédecine - 34298 Montpellier Cedex 5*

² *Institut de Mathématiques et de Modélisation de Montpellier - Place Eugène Bataillon - 34095 Montpellier Cedex 5*

*Antoine.Barbieri@icm.unicancer.fr / Caroline.Mollevi@icm.unicancer.fr /
Christian.Lavergne@univ-montp2.fr*

Résumé. En cancérologie, la qualité de vie (QdV) constitue un critère de jugement secondaire pertinent pour s'assurer de l'intérêt du traitement pour le patient et le système de santé. Cependant, des freins conceptuels (définition, évaluation subjective par le patient) et méthodologiques (nature des données) limitent l'évaluation de la QdV. En effet, on observe à la fois des réponses multiples, répétées et ordinales. Deux modèles concurrents sont présentés : un modèle linéaire mixte utilisé classiquement, et un modèle linéaire mixte généralisé pour données catégorielles ordinales. Ce dernier est une extension d'un modèle de la famille de Rasch. On notera ce nouveau modèle LPCM pour "longitudinal partial credit model". Nous comparerons des résultats obtenus via la procédure *nlmixed* de SAS et via une approche alternative basée sur la linéarisation de Schall (1991).

Mots-clés. Qualité de vie, Modèles mixtes, Linéarisation de Schall, Données ordinales.

Abstract. Nowadays, when it comes to Clinical trials in Oncology, the quality of life has become an essential criterion. However, the longitudinal analysis of this criterion is difficult due to the nature of the data. Indeed, there are both multiple responses, repeated and ordinal ones. Two competing models are presented : a linear mixed model classically used, and a generalized linear mixed model for ordinal categorical data. The latter is an extension of the Rasch model family. We'll name this new model 'LPCM' for 'longitudinal partial credit model'. We will compare the results obtained via the *nlmixed* procedure of SAS and through an alternative approach of linearization (Schall, 1991).

Keywords. Quality of Life, Mixed Models, Schall linearization, Ordinal Categorical Data.

1 Introduction

La Qualité de Vie relative à la santé (QdV) est un concept dynamique, multi-dimensionnel et subjectif incorporant au moins trois domaines : les fonctionnements physique, psychologique et social. Dans les essais cliniques, la QdV constitue un critère de jugement secondaire pertinent pour s'assurer de l'intérêt du traitement pour le patient et le système de santé. Cependant, des freins conceptuels (définition, évaluation subjective par le patient) et méthodologiques (données longitudinales, nature multidimensionnelle, données manquantes, etc...) limitent non seulement l'évaluation de la QdV mais également l'utilisation des résultats pour l'aide à la décision thérapeutique. Dans les essais cliniques, la QdV des patients est évaluée par des auto-questionnaires tout au long de leur prise en charge. La référence en cancérologie est le questionnaire QLQ-C30 développé par l'EORTC¹. Ce questionnaire est composé de 30 items à plusieurs modalités de réponse $\{c\}_{c=1,\dots,m_j}$ ($j = 1, \dots, 30$), regroupés en 15 dimensions uni ou multi-items. Les dimensions sont analysées indépendamment les unes des autres. Les réponses y aux items (qualitatives ordinales) nous permettent ainsi d'évaluer indirectement la QdV. La mesure de la QdV est généralement réalisée à différentes visites au cours d'un essai clinique. Il est alors possible d'étudier la cinétique d'évolution et de la comparer entre différents groupes de patients, ayant par exemple reçus différents traitements.

A une époque où les traitements anticancéreux sont de plus en plus efficaces et permettent de repousser le décès de manière significative, il est indispensable d'optimiser les méthodes d'analyse longitudinale de la QdV. Pour cela, il est nécessaire de considérer la structure des données qui peut être vue selon trois niveaux :

- réponse multiple pour une dimension
- répétition des mesures au cours du temps
- répétition des mesures par un même patient

2 Méthodologie utilisée

D'un point de vue statistique, la QdV n'est pas directement mesurable et est considérée comme un trait latent accessible par l'intermédiaire des réponses aux items. Différentes théories existent pour la modélisation et l'analyse des dimensions des questionnaires de QdV. On différencie particulièrement la théorie classique (Classical Test Theory, CTT) de la théorie moderne de réponse à l'item (Item Response Theory, IRT) (Fayers and Machin, 2000). La CTT se fonde principalement sur les observations d'un score (S), qui correspond à la moyenne pondérée des réponses aux items. Le score S associé à une dimension est alors supposé être proche du « vrai » score de QdV. De ce fait, le score qui est considéré comme une variable continue se substitue à l'utilisation des observations brutes. Sous ces

1. European Organisation for Research and Treatment of Cancer

conditions, le modèle linéaire mixte (L2M) est actuellement utilisé. Il permet de prendre en compte l'aspect longitudinal, la variabilité induite par les mesures répétées, ainsi que des covariables pouvant influencer sur la variable réponse. Le modèle L2M est donnée par :

$$S_{it} = \alpha + \mathbf{x}'_{it}\boldsymbol{\beta}^L + \mathbf{u}'_t\boldsymbol{\xi}_i^L + \varepsilon_{it} \quad (1)$$

avec :

- S_{it} le score du patient i au temps t
- α intercept du modèle
- $\boldsymbol{\beta}^L$ vecteur des paramètres des effets fixes de taille p , et \mathbf{x}_{it} son vecteur associé
- $\boldsymbol{\xi}_i^L$ vecteur d'effets aléatoires du patient i de taille q , et \mathbf{u}_t son vecteur associé
- $\varepsilon_{it} \sim N(0, \sigma_\varepsilon^2)$ terme d'erreur

Les modèles issus de l'IRT utilisés d'abord dans la validation des questionnaires sont proposés comme une alternative au premier. Ces modèles peuvent être vus comme des modèles linéaires mixtes généralisés pour données ordinales (Agresti, 2010) et supposent l'existence d'un modèle probabiliste liant les réponses aux items au trait latent représentant la QdV. En 1982, Masters a proposé le modèle à crédit partiel (PCM) pour données polytomiques (De Boeck et al., 2004). Celui-ci appartient à la famille des modèles de Rasch et différencie les paramètres liés au patient i (paramètre de "capacité" θ_i) de ceux liés à l'item j (paramètre d' "indécision" δ_j). Pour l'analyse longitudinale, nous avons étendu ce modèle afin de prendre en compte des variables explicatives (telles que le traitement, le centre, l'âge, ...) et les caractéristiques des données. La réponse cachée (trait latent θ_{it}) continue inobservable suit alors un modèle de régression linéaire incorporant des effets fixes et aléatoires. Le modèle longitudinal à crédit partiel (LPCM) est donné par :

$$\begin{aligned} \Pr\left(Y_{it}^{(j)} = c | \theta_{it}, \boldsymbol{\delta}_j\right) &= \frac{\exp\{c\theta_{it} - \sum_{k=1}^c \delta_{jk}\}}{\sum_{h=1}^{m_j} \exp\{h\theta_{it} - \sum_{k=1}^h \delta_{jk}\}}, \quad c = 1, \dots, m_j \\ &= \frac{\exp\{c [\mathbf{x}'_{it}\boldsymbol{\beta}^G + \mathbf{u}'_t\boldsymbol{\xi}_i^G] - \sum_{k=2}^c \delta_{jk}\}}{1 + \sum_{h=2}^{m_j} \exp\{h [\mathbf{x}'_{it}\boldsymbol{\beta}^G + \mathbf{u}'_t\boldsymbol{\xi}_i^G] - \sum_{k=2}^h \delta_{jk}\}} \end{aligned} \quad (2)$$

$$= \frac{\exp\{\eta_{itc}^{(j)}\}}{1 + \sum_{h=2}^{m_j} \exp\{\eta_{ith}^{(j)}\}} \quad (3)$$

où

- $Y_{it}^{(j)}$ est la variable réponse associée au patient i au temps t pour l'item j
- $\boldsymbol{\delta}_j$ correspond au vecteur des paramètres d'incertitude lié à l'item j ($\delta_{j1} = 0$)
- $\boldsymbol{\beta}^G$ vecteur des paramètres individuels des effets fixes de taille p
- $\boldsymbol{\xi}_i^G$ vecteur d'effets aléatoires du patient i de taille q
- $\eta_{ith}^{(j)}$ prédicteur linéaire

Dans la famille des modèles linéaires mixtes généralisés pour données ordinales, le LPCM correspond à un modèle logistique adjacent (non proportionnel) (Agresti, 2001). On retrouve dans les modèles (1) et (2) les mêmes vecteurs \mathbf{x}_{it} et \mathbf{u}_t associés aux effets similaires pour modéliser le trait latent représentant la QdV. Pour les deux modèles, la valeur estimée de ces effets est différente, mais leur interprétation reste la même.

3 Simulations

Dans le but d'étudier et de comparer le comportement de ces modèles, des simulations ont été réalisées afin d'analyser leur sensibilité aux effets aléatoires via les critères AIC et BIC. Nous simulons des modèles LPCM pour une dimension à deux items $j = 1, 2$ et $m_j = 4$ modalités, avec $\boldsymbol{\delta}_j$ et $\boldsymbol{\beta}^G$ fixés. Les données sont transformées pour le modèle (1) : le score S_{it} correspond à la moyenne des réponses aux deux items au temps t pour le patient i . Nous nous sommes basés sur le modèle LPCM le plus simple pour simuler nos jeux de données de taille égale à 500 patients :

$$\mathcal{M}_1 \quad : \quad \begin{cases} Y_{it}^{(j)} | \theta_{it}, \boldsymbol{\delta}_j \sim LPCM(\theta_{it}) \\ \theta_{it} = \xi_{i0}^G + v_t (\tau + \xi_{i1}^G) \end{cases} \quad (4)$$

où

- τ correspond à l'effet fixe associé au temps
- v_t correspond au temps entre le début de l'essai et la visite $\{t\}_{t=1,\dots,8}$ à laquelle le patient a rempli le questionnaire ($v = (0.00, 0.05, 0.10, 0.20, 0.40, 0.60, 0.80, 1.00)$)
- $\xi_{ia} \sim \mathcal{N}(0, \sigma_a^2)$ et $a = 0, 1$

Dans un premier temps, le logiciel SAS (proc mixed et proc nlmixed) a mis en évidence des résultats préliminaires (voir Table 1) qui ont révélé une trop forte sensibilité du LPCM aux effets aléatoires. En effet, le modèle \mathcal{M}_1 associé au LPCM est toujours préféré quel que soit le critère ; le modèle détecte automatiquement un effet aléatoire qui n'existe pas dans les données simulées. On notera \mathcal{M}_0 le modèle sans ξ_{i1}^G ($\sigma_1^2 = 0$).

Dans le but de mieux cerner ce problème et palier à une éventuelle erreur de calcul de la vraisemblance des LPCM, nous proposons d'utiliser une approche alternative à la procédure *nlmixed* pour estimer le modèle. Dans un GL2M, il est complexe et couteux d'estimer les coefficients de régression et de prédire les effets aléatoires. Sous l'objectif d'utiliser une approche moins couteuse, Schall (1991) propose une méthode d'estimation itérative des paramètres dans un GL2M, sans spécification particulière de la loi ou de la modélisation des effets. Sa démarche consiste en une linéarisation du modèle conditionnellement aux effets aléatoires $\boldsymbol{\xi}$, puis en l'estimation des paramètres du L2M par utilisation des équations de Henderson. Pour linéariser, il introduit une pseudo-variable réponse \tilde{Y}_i représentée par le développement de Taylor au premier ordre de la fonction de lien g . Le

TABLE 1 – Sélection de l’effet aléatoire lié à la pente sur 500 jeux de données simulés via LPCM avec $\tau = 1$

Modèle d’analyse		LPCM*		L2M	
Critère de sélection		BIC	AIC	BIC	AIC
$\sigma_1^2 = 0$	\mathcal{M}_0^*	4	4	497	474
	\mathcal{M}_1	496	496	3	26
$\sigma_1^2 = 1$	\mathcal{M}_0	0	0	0	0
	\mathcal{M}_1^*	500	500	500	500

* Modèle simulé

modèle linéarisé suivant est considéré :

$$\mathcal{M}_\xi \quad : \quad \tilde{Y}_i = \eta_{\xi,i} + (Y_i - \mu_{\xi,i}) g'(\mu_{\xi,i}) \quad (5)$$

où :

- $E(\tilde{Y}_i | \xi_i) = g(\mu_{\xi,i}) = \eta_{\xi,i}$ le prédicteur linéaire
- $E(Y_i | \xi_i) = \mu_{\xi,i} = g^{-1}(\eta_{\xi,i})$

L’étude et la programmation de cette seconde approche est en cours. Dans un premier temps, les estimations obtenues avec le logiciel SAS seront comparées avec celles de la méthode de Schall afin de les valider. Ensuite, les simulations seront réitérées afin de vérifier à nouveau la sensibilité du LPCM aux effets aléatoires. Les critères d’information seront cette fois-ci calculés via l’estimation du modèle linéaire final \mathcal{M}_ξ^f (Lavergne, 2008). Ce travail permettra de conclure sur l’analyse de la sensibilité des effets aléatoires du modèle LPCM.

Bibliographie

- [1] Fayers, M. et Machin, D. (2000) Quality of life : Assessment, Analysis and Interpretation. *Wiley*.
- [2] Agresti, A. (2010) Analysis of Ordinal Categorical Data. *John Wiley & Sons*.
- [3] De Boeck, P. et Mark Wilson, M. (2004) Explanatory item response models : a generalized linear and nonlinear approach. *Springer*.
- [4] Hartzel, J. Agresti, A. et Caffo, B. (2001) Multinomial logit random effects models. *Statistical Modelling*, 1, 81–102.
- [5] Schall, R. (1991), Estimation in generalized linear models with random effects. *Biometrika*, 78, 719-727.
- [6] Lavergne, C., Martinez, M.J. et Trottier, C. (2008), Empirical model selection in generalized linear mixed effects models. *Computational Statistics*, 23, 99-109.