REVUE DE STATISTIQUE APPLIQUÉE

A. GUÉGUEN
M. ZINS
J. P. NAKACHE

Utilisation des modèles marginaux et des modèles mixtes dans l'analyse de données longitudinales (1992-1996) concernant mariage et consommation d'alcool des femmes de la cohorte GAZEL

Revue de statistique appliquée, tome 48, n° 3 (2000), p. 57-73
<http://www.numdam.org/item?id=RSA_2000__48_3_57_0>

Rev. Statistique Appliquée, 2000, **XLVIII** (3), 57-73

UTILISATION DES MODÈLES MARGINAUX ET DES MODÈLES MIXTES DANS L'ANALYSE DE DONNÉES LONGITUDINALES (1992-1996) CONCERNANT MARIAGE ET CONSOMMATION D'ALCOOL DES FEMMES DE LA COHORTE GAZEL[1]

A. Guéguen, M. Zins, JP. Nakache

INSERM Unité 88, Hôpital National de Saint-Maurice
14, rue du Val d'Osne, 94410 Saint-Maurice.

RÉSUMÉ

Les données longitudinales sont fréquentes aussi bien en épidémiologie qu'en recherche clinique; les deux principales méthodes pour analyser ce type de données sont les modèles marginaux et les modèles mixtes. Ces deux méthodes sont présentées ici en s'appuyant sur un exemple épidémiologique : l'étude des consommations de boissons alcoolisées des femmes de la cohorte GAZEL au moment du mariage.

Mots-clés : *Données longitudinales, modèle marginal, modèle linéaire mixte.*

ABSTRACT

Longitudinal data are frequently encountered in epidemiology and clinical research as well; the main methods used to analyze such data are marginal models and mixed models which are presented here and illustrated using epidemiological data extracted from the cohort GAZEL and concerning the relationship between weeding and women alcohol consumption.

Keywords : *Longitudinal data, marginal model, linear mixed model.*

1. Introduction et hypothèses préliminaires

En épidémiologie et en recherche clinique, la plupart des problèmes peuvent se poser sous la forme de l'étude de la liaison entre une variable «réponse» ou phénomène d'intérêt, et un ensemble de variables explicatives (covariables, variables exogènes, prédicteurs). La construction de modèles mathématiques permet d'explorer la forme et l'intensité de cette liaison. Cependant, il est fréquent de recueillir un

[1] Ce travail a été réalisé en partie avec des crédits de l'ONIVINS, attribués par le Comité National Vin et Santé.

ensemble de données présentant une configuration telle qu'elles sont difficilement analysables au moyen des modèles statistiques «classiques». C'est le cas des données dites «longitudinales» qui constituent des données (variable à expliquer et variables explicatives) recueillies pour chaque sujet à différents temps. Il en résulte que les observations réalisées sur un même sujet sont, en général, fortement corrélées. On ne peut donc utiliser les modèles classiques, qui reposent sur l'indépendance des observations.

Il existe principalement deux méthodes d'analyse des données longitudinales permettant de prendre en compte la liaison entre les observations réalisées sur un même sujet : 1) les modèles marginaux qui sont utilisés quand l'objectif scientifique est de caractériser la population; on modélise alors la moyenne du phénomène à étudier, et 2) les modèles mixtes où l'inférence est individuelle.

L'objectif de cet article est de présenter et de confronter les modèles marginaux et les modèles mixtes dans une étude dont le but est d'étudier l'effet du mariage sur la consommation d'alcool des femmes de la cohorte GAZEL[4].

2. Matériel et méthodes

2.1. Les données analysées

En 1989, environ 20 000 volontaires, âgés de 35 à 50 ans, ont accepté de participer à une cohorte à visée épidémiologique, la cohorte GAZEL; chaque année les volontaires remplissent un auto-questionnaire sur leurs problèmes de santé et leurs habitudes de consommation. Les observations dont on dispose proviennent des auto-questionnaires de 1992, 1993, 1994, 1995 et 1996 et concernent la consommation d'alcool, exprimée en nombre de verres par semaine. Parmi les 5614 femmes de la cohorte GAZEL, 155 femmes se sont mariées entre 1992 et 1996, et on dispose de données sur leur consommation d'alcool au moment de leur mariage pour 148 d'entre elles.

2.2. Les modèles utilisés

2.2.1. Les modèles Marginaux

Les modèles marginaux ont été introduits par Liang et Zeger [6] pour analyser les données longitudinales normales, non normales ou discrètes. Ils sont l'analogue des modèles de quasi-vraisemblance pour observations non indépendantes. De même que pour les modèles de quasi-vraisemblance, il est suffisant de donner les deux premiers moments de la variable réponse, par opposition aux modèles linéaires généralisés dans lesquels toute la distribution de la variable réponse doit être définie. Par ailleurs, comme pour les modèles de quasi-vraisemblance, il est possible d'obtenir une estimation consistante du vecteur des paramètres, mais avec une perte d'efficacité, même si la variance de la variable réponse n'est pas correctement spécifiée.

L'utilisation des modèles marginaux nécessite que le nombre d'observations par sujet soit faible et que le nombre de sujets soit important, ce qui est souvent le cas

dans les études épidémiologiques. Par ailleurs, s'il existe des données manquantes, les estimations obtenues par le modèle marginal sont sans biais à condition que les données soient manquantes complètement au hasard selon la terminologie de Little et Rubin [7]. L'intérêt principal des modèles marginaux repose sur la robustesse des estimations obtenues. Ceci est du au fait qu'il n'est pas nécessaire que la matrice des variances-covariances soit correctement définie; il suffit de définir une matrice de corrélation de travail.

Les modèles marginaux sont décrits dans l'annexe A.1 en s'inspirant des présentations de Diggle et al [2] et de Fahrmeir et Tutz [3].

2.2.2. Les modèles mixtes

Les modèles mixtes linéaires [5] sont des extensions des modèles linéaires. Les modèles mixtes font intervenir des effets aléatoires spécifiques à chaque sujet et c'est l'espérance de la variable réponse conditionnellement à ces effets aléatoires qui est modélisée sous la forme d'une combinaison linéaire des variables explicatives, incluant à la fois des facteurs fixes et des facteurs aléatoires. Dans les modèles mixtes, contrairement aux modèles marginaux, la moyenne et la variance de la variable réponse sont modélisées en même temps. Les coefficients du modèle mesurent l'influence directe des variables explicatives sur la variable réponse pour des groupes de sujets hétérogènes et non pour la population.

L'utilisation des modèles mixtes est conseillée si le nombre d'observations par sujet est élevé, si on cherche à établir des prédictions individuelles ou si on souhaite appréhender la structure de la matrice des variances-covariances. Par ailleurs, s'il existe des données manquantes, les estimations seront sans biais à condition que les données soient manquantes au hasard selon la terminologie de Little et Rubin [7], ce qui constitue une condition moins contraignante que celle requise pour les modèles marginaux.

Les modèles mixtes sont présentés dans l'annexe A.2, en s'inspirant, comme pour les modèles marginaux, des ouvrages de Diggle et al [2] et de Fahrmeir et Tutz [3].

3. Application

Les données sont constituées de 5 cohortes de femmes caractérisées par l'année de leur mariage; 46, 41, 17, 24 et 20 femmes se sont respectivement mariées en 1992, 1993, 1994, 1995 et 1996. Les moyennes de consommation d'alcool de ces 5 cohortes sont représentées en figure 1.

On note y_{ic} la mesure de la consommation d'alcool du sujet i au temps c, où c représente le temps calendaire et prend les 5 valeurs allant de 1992 à 1996. Etant donné qu'on s'intéresse à l'effet du mariage sur la consommation d'alcool, il est nécessaire de prendre un autre repère pour le temps : l'origine du temps est l'année du mariage. Pour chaque sujet i et chaque année c, on crée une variable t_{ic} représentant le nombre d'années depuis le mariage. Cette variable prend ses valeurs dans l'intervalle variant entre – 4 et 4. Par exemple, les valeurs de t_{ic} pour une femme qui s'est mariée en 1994 sont : – 2, – 1, 0, 1, 2.

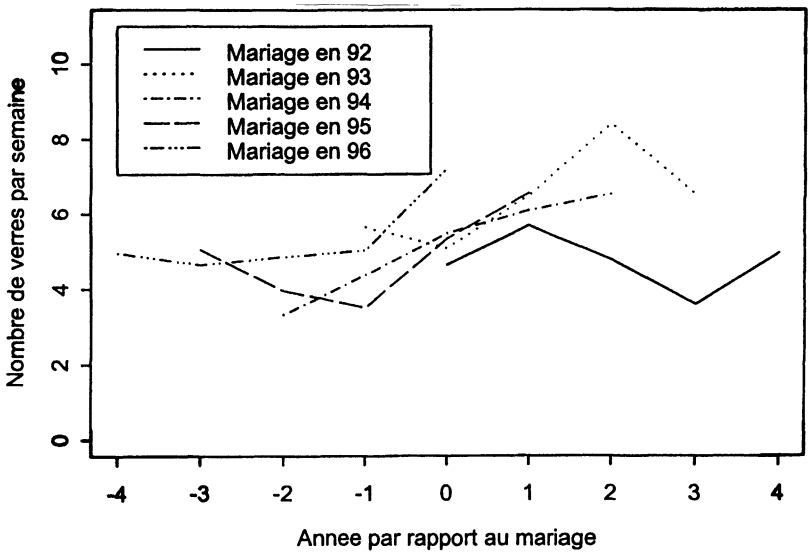

FIGURE 1
Consommation moyenne de boissons alcoolisées

3.1. Modèle marginal

3.1.1 Modélisation de la moyenne

L'examen de la figure 1 suggère de modéliser la moyenne marginale des consommations d'alcool au moyen de 3 portions de droite, les changements de pente ayant lieu un an avant et un an après le mariage.

On note $\mu_{ic} = E(y_{ic})$ l'espérance marginale de la consommation d'alcool. Par la suite l'indice i est omis de manière à simplifier l'écriture. Le modèle proposé est le suivant :

$$\begin{aligned} \mu_c &= \beta_0 + \beta_1(t_c + 4) && -4 \leqslant t_c \leqslant -1 \\ \mu_c &= \beta_0 + 3\beta_1 + \beta_2(t_c + 1) && -1 \leqslant t_c \leqslant 1 \\ \mu_c &= \beta_0 + 3\beta_1 + 2\beta_2 + \beta_3(t_c - 1) && 1 \leqslant t_c \leqslant 4 \end{aligned}$$

où t_c représente le nombre d'années depuis le mariage.

Ce modèle s'écrit, en fonction de 3 variables t_1, t_2 et t_3 créées à partir de t_c sous la forme équivalente suivante :

$$\mu_c = \beta_0 + \beta_1 t_1 + \beta_2 t_2 + \beta_3 t_3,$$

où les valeurs de t_1, t_2 et t_3 en fonction de t_c sont données dans le tableau 1.

β_0 représente la consommation d'alcool moyenne 4 ans avant le mariage, β_1, β_2 et β_3 représentent respectivement les augmentations moyennes (par année) des

TABLEAU 1
Valeurs de t_1, t_2 et t_3 en fonction de t_c

t_c	−4	−3	−2	−1	0	1	2	3	4
t_1	0	1	2	3	3	3	3	3	3
t_2	0	0	0	0	1	2	2	2	2
t_3	0	0	0	0	0	0	1	2	3

consommations d'alcool entre 4 ans et 1 an avant le mariage, entre 1 an avant et 1 an après le mariage et entre 1 an et 4 ans après le mariage.

3.1.2. Modélisation de la variance

Dans les modèles marginaux, il n'est pas nécessaire de définir la matrice des variances-covariances : la définition d'une matrice de corrélation de travail la plus proche possible de la réalité est suffisante. La matrice de corrélation de travail choisie dépend de 4 paramètres ρ_1, ρ_2, ρ_3 et ρ_4. Ces paramètres représentent respectivement les corrélations entre deux observations réalisées sur un même sujet et distantes respectivement d'un an, de deux ans, de trois ans et de quatre ans.

$$\mathrm{Var}(y_{ic}) = \sigma^2 \begin{pmatrix} 1 & \rho_1 & \rho_2 & \rho_3 & \rho_4 \\ \rho_1 & 1 & \rho_1 & \rho_2 & \rho_3 \\ \rho_2 & \rho_1 & 1 & \rho_1 & \rho_2 \\ \rho_3 & \rho_2 & \rho_1 & 1 & \rho_1 \\ \rho_4 & \rho_3 & \rho_2 & \rho_1 & 1 \end{pmatrix}.$$

3.1.3. Résultats

Les résultats sont donnés dans les tableaux 2 et 3 et ont été obtenus à l'aide de la procédure GENMOD de SAS [8]. Le fichier de commande utilisé est le suivant :

```
proc genmod data=donnees ;
class obs tnum ;
model verres=t1 t2 t3 ;
repeated subject=obs / within=tnum type=mdep(4) ;
run ;
```

où `donnees` représente le fichier des données dans lequel chaque ligne contient 6 variables : `obs, tnum, verres, t1, t2` et `t3`,

– `obs` représente le numéro d'identification de chaque femme,

– `tnum` représente le numéro de l'année calendaire (`tnum` varie de 1 à 5),

– `verres` représente le nombre de verres d'alcool consommés par semaine,

– `t1, t2` et `t3` sont les trois variables créées selon la correspondance donnée dans le tableau 1.

TABLEAU 2
Paramètres du modèle marginal modélisant la moyenne

Paramètre	Estimation	IC 95 %	
β_0	4.20	2.36	6.04
β_1	0.08	− 0.56	0.72
β_2	0.96	0.38	1.54
β_3	− 0.07	− 0.52	0.39

La consommation moyenne quatre années avant le mariage est de 4.20 verres par semaine. Entre quatre ans et un an avant le mariage, la consommation moyenne reste stable : $\widehat{\beta}_1$ n'est pas significativement différent de 0 ($p = 0.74$). Entre un an avant et un an après le mariage, la consommation d'alcool augmente significativement ($p < 0.001$), de 1.92 verres par semaine ($2\widehat{\beta}_2$). Entre un et quatre ans après le mariage, la consommation d'alcool reste stable : $\widehat{\beta}_3$ n'est pas significativement différent de 0 ($p = 0.58$).

TABLEAU 3
Paramètres du modèle marginal modélisant la variance

Paramètre	Estimation
ρ_1	0.719
ρ_2	0.473
ρ_3	0.524
ρ_4	0.470

Dans les modèles marginaux, les paramètres permettant de modéliser la variance sont des paramètres de nuisance. D'autres types de matrices des variances-covariances ont été utilisées et les ajustements obtenus donnent des résultats semblables en ce qui concerne l'estimation des paramètres β.

3.2. Modèle mixte

3.2.1. Modèle mixte à deux effets aléatoires

Etant donné qu'il y a au plus 5 observations par femme, deux effets aléatoires seulement ont été introduits dans le modèle : le premier est associé à la constante β_0 et traduit le fait que la consommation d'alcool quatre ans avant le mariage est variable d'une femme à l'autre; le second effet aléatoire est associé au paramètre β_2 et indique que l'évolution de la consommation d'alcool entre un an avant et un an après le mariage peut être différente selon les femmes. Par contre, le modèle ne

comprend pas d'effets aléatoires associés aux paramètres β_1 et β_3 : on considère que l'évolution de la consommation d'alcool entre quatre ans avant et un an avant est identique pour toutes les femmes, de même que l'évolution entre un an et quatre ans après le mariage. On suppose, d'autre part, qu'il peut exister une liaison entre la consommation initiale et l'augmentation de la consommation d'alcool entre un an avant et un an après le mariage.

Le modèle ajusté à la consommation d'alcool du sujet i est le suivant :

$$y_{ic} = (\beta_0 + b_{0i}) + \beta_1 t_{i1} + (\beta_2 + b_{2i}) t_{i2} + \beta_3 t_{i3} + e_{ic},$$

où $\begin{pmatrix} b_{0i} \\ b_{2i} \end{pmatrix} \propto N(0, Q)$, avec $Q = \begin{pmatrix} \sigma_0^2 & \sigma_{02} \\ \sigma_{02} & \sigma_2^2 \end{pmatrix}$, et $e_{ic} \propto N(0, \sigma_e^2 I)$.

β_0 représente la consommation d'alcool d'une femme «standard» 4 ans avant le mariage; β_0 représente également la moyenne des consommations d'alcool 4 ans avant le mariage,

β_2 représente l'augmentation de la consommation d'alcool entre 1 an avant et 1 an après le mariage pour une femme «standard»; β_2 représente également la moyenne de l'augmentation de la consommation d'alcool entre 1 an avant et 1 an après le mariage,

β_1 et β_3 représentent respectivement les moyennes des augmentations des consommations d'alcool entre 4 ans et 1 an avant le mariage et entre 1 an et 4 ans après le mariage.

σ_0^2 représente la variance des consommations d'alcool entre sujets, quatre ans avant le mariage,

σ_2^2 représente la variance des modifications de consommation d'alcool entre un an avant et un an après le mariage,

σ_{02} représente la covariance entre la consommation initiale et la modification de consommation d'alcool entre un an avant et un an après le mariage; une valeur élevée de ce paramètre indiquerait que les femmes qui sont initialement les plus grosses consommatrices sont également celles qui augmentent le plus leur consommation au moment du mariage,

σ_e^2 représente la variance de la consommation d'alcool d'un sujet autour de sa consommation théorique.

3.2.2. Résultats

Le modèle mixte à deux effets aléatoires (modèle M_1) a été ajusté aux données en utilisant la méthode REML. [1] (Le même modèle a été ajusté aux données en utilisant la méthode MLE [1] et donne sensiblement les mêmes résultats). Les résultats sont donnés dans les tableaux 4 et 5 et ont été obtenus à l'aide de la procédure MIXED de SAS [8]. Le fichier de commande utilisé est le suivant :

```
proc mixed data=donnees ;
class obs ;
```

[1] Cf. Annexe A.2.2.

```
model verres=t1 t2 t3 / solution ;
random intercept t2 /subject=obs type=un ;
run ;
```

où `donnees` représente le fichier des données dans lequel chaque ligne contient 5 variables : `obs`, `verres`, `t1`, `t2` et `t3`.

TABLEAU 4
Estimation des paramètres du modèle mixte modélisant la moyenne

Paramètre	Estimation	IC 95 %	
β_0	4.06	2.62	5.50
β_1	0.11	− 0.35	0.57
β_2	1.12	0.59	1.65
β_3	− 0.15	− 0.48	0.19

TABLEAU 5
Estimation des paramètres du modèle mixte modélisant la variance

Paramètre	Estimation
σ_0^2	19.53
σ_{02}	2.09
σ_2^2	4.65
σ_e^2	9.54

La consommation moyenne quatre années avant le mariage est de 4.06 verres par semaine. Entre quatre ans et un an avant le mariage, la consommation moyenne reste stable : $\widehat{\beta}_1$ n'est pas significativement différent de 0 ($p = 0.63$). Entre un an avant et un an après le mariage, la consommation d'alcool augmente significativement ($p < 0.001$), de 2.24 verres par semaine ($2\widehat{\beta}_2$). Entre un et quatre ans après le mariage, la consommation d'alcool reste stable : $\widehat{\beta}_3$ n'est pas significativement différent de 0 ($p = 0.41$). Ces résultats sont très proches de ceux obtenus par le modèle marginal. Par ailleurs, les paramètres $\widehat{\beta}_0$ et $\widehat{\beta}_2$ ont une deuxième interprétation; ils correspondent aux valeurs prises par une femme «standard».

D'autres modèles ont également été ajustés aux données; ces modèles comprennent les mêmes paramètres β que ceux du modèle M_1 mais des types d'effets aléatoires différents. Le modèle M_2 comprend 2 effets aléatoires b_{0i} et b_{2i} mais suppose que leur covariance σ_{02} est nulle. Le modèle M_3 comprend un seul effet aléatoire

associé à β_0 et le modèle M_4 comprend 4 effets aléatoires avec une matrice Q associée qui comprend 10 paramètres. Le tableau 6 donne les valeurs de – 2 log de la vraisemblance , du critère d'Akaike [1] et du critère de Scharwz [9] pour les modèles M_1 à M_4.

TABLEAU 6

Valeurs de plusieurs critères : – 2 log de la vraisemblance($-2\,log\,L$), critères d'Akaike (AIC) et de Scharwz (BIC) pour différents modèles. $AIC = -2\log L + 2(p+q)$ et $BIC = -2\log L + (p+q)\log(N-p)$, où p et q représentent le nombre de paramètres permettant de modéliser respectivement la moyenne et la variance et N le nombre d'observations

	– 2 log L	AIC	BIC
M_1	3791.85	3807.85	3843.65
M_2	3793.80	3807.80	3839.13
M_3	3847.16	3859.16	3886.01
M_4	3784.43	3814.43	3881.56

Les deux modèles présentant les meilleurs critères sont les modèles M_1 et M_2. Le modèle M_3 ne peut être retenu ; il est donc nécessaire d'introduire un effet aléatoire associé à β_2 et les modifications de consommation d'alcool au moment du mariage ne sont donc pas les mêmes pour toutes les femmes. Le modèle M_4 ne peut lui non plus être retenu. En conclusion, le modèle M_2 est celui qui présente le meilleur ajustement. (Les données ont été ajustées à ce modèle et donnent des résultats très proches de ceux du modèle M_1) : il n'existe pas de lien significatif entre la consommation d'alcool d'une femme 4 ans avant le mariage et la modification de sa consommation d'alcool au moment du mariage.

Effets aléatoires

Les valeurs des estimations des facteurs aléatoires $\{b_i\}$ ne présentent pas d'intérêt particulier en épidémiologie, mais peuvent en avoir en recherche clinique si on cherche à réaliser des prédictions individuelles. Les histogrammes des estimations des deux facteurs aléatoires sont représentés en figures 2 et 3 et montrent clairement des distributions non normales. La non normalité a une importance moindre si on s'intéresse essentiellement aux paramètres β; elle a, par contre, une influence plus grande si on s'intéresse aux effets aléatoires eux mêmes ou à l'estimation des paramètres qui permettent de modéliser la variance. Ceci suggère d'utiliser un autre modèle, le modèle mixte de Poisson [2] et [3]. Ce modèle, utilisé en général pour des données de dénombrement, suppose d'une part, que la consommation d'alcool y_{ic} conditionnellement à b_i suit une distribution de Poisson de paramètre $\mu_{ic} = E(y_{ic}/b_i)$, telle que $\log(\mu_{ic})$ est une combinaison linéaire des variables explicatives et d'autre part, que les b_i suivent des lois multinormales de moyenne nulle et sont indépendants entre eux. Appliqué aux données, ce modèle conduit à des distributions observées des effets aléatoires dont l'ajustement par des lois normales est bien meilleur que celui obtenu par le modèle mixte linéaire.

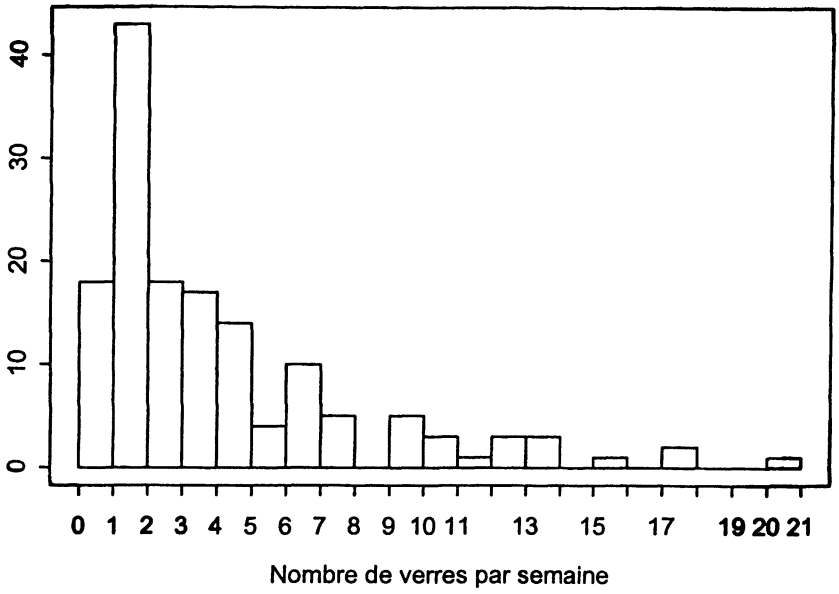

FIGURE 2
Histogramme des estimations des consommations d'alcool initiales

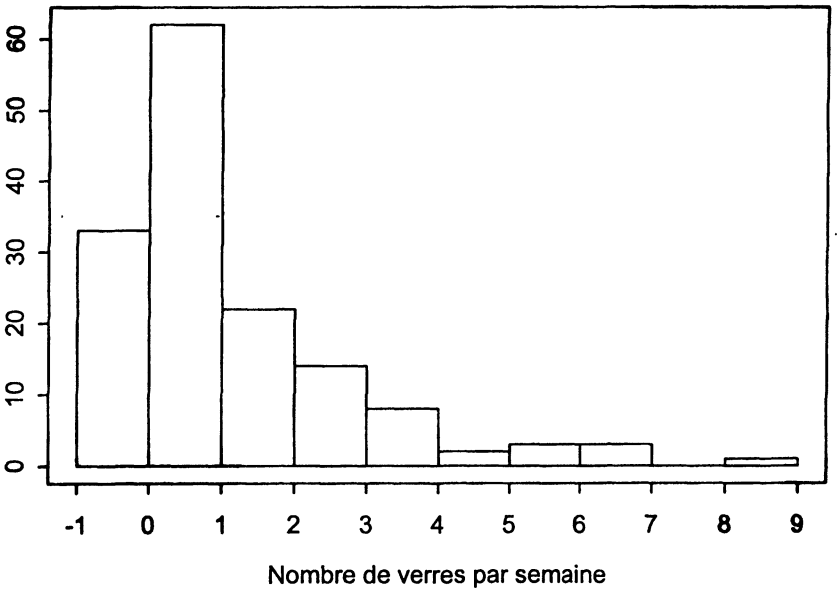

FIGURE 3
Histogramme des estimations des modifications de consommation d'alcool au moment du mariage

4. Discussion

En général, les modèles marginaux sont recommandés quand l'objectif est de réaliser des inférences sur la moyenne de la variable réponse, les corrélations entre observations n'étant que des paramètres de nuisance, alors que les modèles mixtes permettent d'étudier plus finement le phénomène étudié en fournissant une connaissance de la matrice des variances-covariances et de réaliser des prédictions individuelles, mais ceci au prix d'hypothèses plus contraignantes : il faut que le modèle, aussi bien dans sa partie fixe (qui modélise la moyenne) que dans sa partie aléatoire (qui modélise la variance), soit correctement spécifié. Par ailleurs, si le nombre d'observations réalisées sur un même sujet est important, les modèles marginaux sont particulièrement déconseillés : il est nécessaire d'utiliser les modèles mixtes qui permettent de paramétriser la matrice des variances-covariances et de diminuer ainsi le nombre de paramètres à estimer. Enfin, s'il existe des données manquantes, les modèles mixtes sont moins contraignants puisqu'ils reposent sur une hypothèse plus faible, à savoir que les données sont manquantes au hasard, alors qu'elles doivent être manquantes complètement au hasard si on veut utiliser un modèle marginal.

Dans l'application sur les consommations d'alcool, l'utilisation d'un modèle mixte est délicate car l'hypothèse de normalité des facteurs aléatoires n'est pas vérifiée. Cependant, il est intéressant d'examiner et de confronter les résultats apportés par les deux approches : en ce qui concerne les paramètres relatifs à la moyenne, le modèle mixte et le modèle marginal donnent des résultats très proches. De plus, le modèle mixte met en évidence le fait qu'il existe une grande variabilité entre sujets : les consommations initiales des femmes (quatre ans avant le mariage) sont très variables, et d'autre part, l'augmentation moyenne de la consommation d'alcool observée au moment du mariage (entre un an avant et un an après) correspond à des évolutions individuelles différentes selon les femmes. La mise en évidence de cette diversité conduit à de nouvelles questions : rechercher les facteurs déterminants d'une augmentation importante de la consommation d'alcool au moment du mariage.

Bibliographie

[1] AKAIKE H. (1974), A New Look at the Statistical Model Identification, IEEE Transaction on Automatic control, AC-19, pp. 716-723.

[2] DIGGLE P.J., LIANG K., ZEGER S.L. (1994), Analysis of Longitudinal Data, Oxford University Press, Oxford.

[3] FAHRMEIR L., TUTZ G. (1994), Multivariate statistical modelling based on generalised linear models, Springer, New york.

[4] GOLDBERG M., LECLERC A. (1994), Cohorte GAZEL, 20 000 volontaires d'EDF-GDF pour la recherche médicale, Bilan 1989-1993, Les Editions INSERM, Paris.

[5] LAIRD N.M., WARE J.H. (1982), Random-effects models for longitudinal data. Biometrics, 38, pp. 963-974.

[6] LIANG K.-Y., ZEGER S.L. (1986), Longitudinal data analysis using generalized linear models. Biometrika, 73, pp. 13-22.

[7] LITTLE R.J.A., RUBIN D.B. (1987), Statistical analysis with missing data, John Wiley, New York.

[8] SAS Institute, Inc. SAS/STAT software : changes and enhancements trough release 6.12, Cary, NC :SAS Institute Inc, 1997.

[9] SCHWARZ G. (1978), Estimating the dimension of a model, Annals of Statistics, 6, pp. 461-464.

[10] STRAM D.O., LEE J.W. (1994), Variance compononents testing in the longitudinal mixed effects model, Biometrics, 50, pp. 1171-1177.

[11] VERBEKE G., MOLENBERGHS G. (1997), Linear mixed models in practice, A SAS-oriented approach. Springer, New york.

Annexe

A.1. Modèle marginal

A.1.1. Le modèle

Soit T_i le nombre d'observations réalisées sur le sujet i; on note $y_i = (y_{i1}, ..., y_{iT_i})$ le vecteur des réponses et $x_i = (x_{i1}, ... x_{iT_i})$ le vecteur des covariables pour l'individu i, $i = 1, ..., n$, pour lequel chacune des T_i composantes est elle-même un vecteur de covariables. On note $N = \Sigma T_i$ le nombre total d'observations. Un modèle marginal repose sur les hypothèses suivantes :

1) la moyenne marginale μ_{it} de la variable réponse est correctement spécifiée à travers la relation :

$$\mu_{it} = h(z'_{it}\beta), \qquad (A.1.1)$$

où $\mu_{it} = E(y_{it}/x_{it})$ est la moyenne marginale de la variable réponse,

h est une fonction connue, dont l'inverse est appelée fonction de lien,

et $z_{it} = z_{it}(x_{it})$, est un vecteur de dimension p fonction appropriée du vecteur des covariables x_{it}. Dans la plupart des cas, 1 est la première composante de z_{it}.

2) la variance marginale de la variable réponse est fonction de la moyenne marginale :

$$\text{var}(y_{it}) = \phi\nu(\mu_{it}), \qquad (A.1.2)$$

où ν est une fonction connue, appelée fonction de variance, et ϕ est un paramètre d'échelle à estimer.

3) la covariance entre y_{it} et y_{is} est fonction des moyennes marginales μ_{it} et μ_{is} et de α, vecteur de paramètres inconnus à estimer :

$$\text{cov}(y_{it}, y_{is}) = c(\mu_{it}, \mu_{is}, \alpha), \qquad (A.1.3.)$$

où c est une fonction connue.

A.1.2. Matrice de corrélation de travail

Il n'est pas nécessaire que la matrice des variances-covariances de la variable réponse soit correctement spécifiée. Aussi, on suppose que la moyenne marginale est correctement spécifiée et à partir des relations données en 2) et 3) on définit une matrice des variances- covariances de travail :

$$\mathrm{cov}_{TR}(y_i) = \Sigma_i(\beta, \alpha, \phi),$$

qui dépend des paramètres β, α et ϕ. Cette matrice est en général différente de la vraie matrice des variances-covariances $\mathrm{cov}(y_i) = S_i$.

On note A_i la matrice diagonale dont les éléments diagonaux sont $\mathrm{var}_{TR}(y_{it})$; les éléments de A_i dépendent des paramètres β et ϕ. On définit une matrice de corrélation de travail $R_i(\alpha)$ de dimension $T_i \times T_i$. La matrice des variances-covariances de travail s'écrit :

$$\Sigma_i(\beta, \alpha, \phi) = A_i^{1/2} R_i(\alpha) A_i^{1/2}. \qquad (A.1.4)$$

Plusieurs choix sont possibles pour la matrice de corrélation de travail $R_i(\alpha)$:

- la matrice identité correspond à une hypothèse d'indépendance pour la matrice de travail : les observations réalisées sur un même sujet sont non corrélées;
- la matrice de corrélation de travail est complètement non spécifiée : $R_i(\alpha) = R(\alpha)$. (Ce choix n'est possible qu'à la condition que tous les T_i soient égaux);
- la matrice de corrélation de travail correspond à une structure échangeable : $(R_i)_{st} = \alpha$, pour $s \neq t$.
- La matrice de corrélation de travail est stationnaire : $(R_i)_{s_i t_i} = \alpha(|t_i - s_i|)$.

A.1.3. Estimation des paramètres β

β est solution de l'équation d'estimation généralisée :

$$S_\beta(\beta, \alpha, \phi) = \sum_{i=1}^{n} Z_i' D_i(\beta) \Sigma_i^{-1}(\beta, \alpha, \phi)(y_i - \mu_i) = 0, \qquad (A.1.5)$$

où la matrice Z_i' est définie par $Z_i' = (z_{i1}, ...z_{iT})$ et la matrice diagonale $D_i(\beta)$ est égale à $D_i(\beta) = \mathrm{diag}(D_{it}(\beta))$, où $D_{it}(\beta) = \dfrac{\partial h}{\partial \eta}$ évalué à $\eta = z_{it}'\beta$.

Quand la matrice de corrélation de travail est la matrice identité, on retrouve la fonction de score obtenue dans le cas d'indépendance des observations. L'estimation de β se fait à l'aide d'un processus itératif : algorithme de Fischer modifié. Etant données les estimations en cours des paramètres α et ϕ, l'estimation à l'itération $k+1$ est égale à :

$$\widehat{\beta}^{(k+1)} = \widehat{\beta}^{(k)} + \widehat{F}^{(k)^{-1}} \widehat{s}^{(k)}, \qquad (A.1.6)$$

où la matrice de travail de Fisher $\widehat{F}^{(k)}$ est égale à

$$\widehat{F}^{(k)} = \sum_{i=1}^{n} Z_i \, D_i(\widehat{\beta}^{(k)}) \Sigma_i^{-1}(\widehat{\beta}^{(k)}, \widehat{\alpha}, \widehat{\phi}) D_i(\widehat{\beta}^{(k)}) Z_i',$$

et où la fonction de quasi-score $\widehat{s}^{(k)}$ est égale à :

$$\widehat{s}^{(k)} = \sum_{i=1}^{n} Z_i' D_i(\widehat{\beta}^{(k)}) \Sigma_i^{-1}(\widehat{\beta}^{(k)}, \widehat{\alpha}, \widehat{\phi})(y_i - \widehat{\mu}_i),$$

où « ˆ » signifie évalué à $\beta = \widehat{\beta}^{(k)}$.

Etant donné l'estimation en cours de β, on estime les paramètres α et ϕ à l'aide des résidus de Pearson en cours :

$$\hat{r}_{it} = \frac{y_{it} - \widehat{\mu}_{it}}{\nu(\widehat{\mu}_{it})^{1/2}}. \qquad (A.1.7)$$

Le paramètre d'échelle ϕ est estimé par :

$$\widehat{\phi} = \frac{1}{N-p} \sum_{i=1}^{n} \sum_{t=1}^{T_i} \hat{r}_{it}^2. \qquad (A.1.8)$$

Les paramètres α sont estimés différemment selon la définition de la matrice de corrélation de travail $R_i(\alpha)$:

• dans le cas où la matrice de corrélation de travail est complètement non spécifiée, on estime $R(\alpha)$ par :

$$R(\widehat{\alpha}) = \frac{1}{n\widehat{\phi}} \sum_{i=1}^{n} A_i^{-1/2}(y_i - \widehat{\mu}_i)(y_i - \widehat{\mu}_i)' A_i^{-1/2}.$$

• dans le cas où la structure de la matrice de corrélation est la corrélation échangeable, on estime α par :

$$\widehat{\alpha} = \frac{1}{\widehat{\phi}\left\{\sum \frac{1}{2} T_i(T_i - 1) - p\right\}} \sum_{i=1}^{n} \sum_{t>s} \widehat{r}_{it}\widehat{r}_{is};$$

A.1.4. Distribution asymptotique de $\widehat{\beta}$

Sous des conditions de régularité «faibles», $\widehat{\beta}$ est asymptotiquement distribué selon une loi multinormale :

$$\widehat{\beta} \xrightarrow{a} N(\beta, \widehat{F}^{-1}\widehat{V}\widehat{F}^{-1}), \qquad (A.1.9)$$

où $F = \sum_{i=1}^{n} Z_i' D_i \Sigma_i^{-1} D_i Z_i$ et $V = \sum_{i=1}^{n} Z_i' D_i \Sigma_i^{-1} S_i \Sigma_i^{-1} D_i Z_i$, et où « $\hat{}$ » signifie évalué à $\beta = \widehat{\beta}$, $\alpha = \widehat{\alpha}$ et $\phi = \widehat{\phi}$.

Si la matrice des variances-covariances est correctement spécifiée (*i.e.* $\Sigma_i = S_i$), alors $V = F$ et on retrouve le résultat classique :

$$\widehat{\beta} \overset{a}{\longrightarrow} N(\beta, \widehat{F}^{-1});$$

et l'estimation de β est asymptotiquement efficace. Par contre, s'il s'agit juste d'une matrice de travail, alors il y aura perte d'efficacité.

A.2 Modèle mixte linéaire pour données normales

A.2.1. Le modèle

Soit T_i le nombre d'observations réalisées sur le sujet i; on note $y_i = (y_{i1}, ..., y_{iT_i})$ le vecteur des réponses et $x_i = (x_{i1}, ...x_{iT_i})$ le vecteur des covariables pour l'individu i, $i = 1, ..., n$, pour lequel chacune des T_i composantes est elle-même un vecteur de covariables. Le modèle mixte linéaire pour données normales est une extension du modèle linéaire pour données normales :

$$y_{it} = z_{it}'\beta = w_{it}'b_i + e_{it} \qquad (A.2.1)$$

où $z_{it} = z_{it}(x_{it})$ est un vecteur de dimension p, fonction appropriée du vecteur des covariables x_{it}. Dans la plupart des cas, 1 est la première composante de z_{it};

$w_{it} = w_{it}(x_{it})$ est un vecteur de dimension q, également fonction du vecteur des covariables x_{it}. Dans la plupart des cas, w_{it} est un sous-vecteur de z_{it};

β est un vecteur de dimension p représentant les effets dans la population,

b_i est un vecteur de dimension q représentant les effets pour le sujet i, $b_i \propto N(0, Q)$,

$e_i = (e_{i1}, ..., e_{iT_i})$ est un vecteur de dimension T_i, $e_i \propto N(0, R_i)$, les $\{e_i\}$ et les $\{b_i\}$ sont indépendants.

Le modèle précédent peut s'écrire sous forme matricielle :

$$y_i = Z_i\beta + W_i b_i + e_i, \qquad (A.2.2)$$

où Z_i' est la matrice $(z_{i1}, ..., z_{iT_i})$ et W_i' est la matrice $(w_{i1}, ...w_{iT_i})$.

En regroupant les deux derniers termes en un seul, le modèle s'écrit :

$$y_i = z_i\beta + e_i^*, \qquad (A.2.3)$$

où $e_i^* \propto N(0, \sigma^2 V_i(\alpha))$ avec $\sigma^2 V_i(\alpha) = W_i' Q W_i + R_i$,

α est un vecteur de dimension r comprenant les éléments de Q et les éléments permettant de paramétriser les matrices R_i,

les $\{e_i^*\}$ sont indépendants.

A.2.2. Estimation des paramètres β et α

Les estimations des paramètres peuvent être obtenues soit par la méthode du maximum de vraisemblance (MLE) soit par celle du maximum de vraisemblance restreint (REML). Il est recommandé d'utiliser la méthode REML car les estimations obtenues par la méthode MLE peuvent être biaisées, le biais étant d'autant plus important que p, la dimension du vecteur β, est importante.

Pour une valeur de α donnée, l'estimateur du maximum de vraisemblance de β est :

$$\widehat{\beta}(\alpha) = \left(\sum_{i=1}^{n} Z_i' V_i(\alpha)^{-1} Z_i \right)^{-1} \sum_{i=1}^{n} Z_i' V_i(\alpha) y_i. \qquad (A.2.4)$$

L'estimateur du maximum de vraisemblance de σ^2 est :

$$\widehat{\sigma}^2(\alpha) = SCR(\alpha)/N, \qquad (A.2.5)$$

avec $SCR(\alpha) = \sum_{i=1}^{n} (y_i - Z_i \widehat{\beta}(\alpha))' V_i(\alpha)^{-1} (y_i - Z_i \widehat{\beta}(\alpha))$.

L'estimateur du maximum de vraisemblance de α maximise :

$$L(\alpha) = -\frac{1}{2} \left[N \log SCR(\alpha) + \sum_{i=1}^{n} \log |V_i(\alpha)| \right]. \qquad (A.2.6)$$

Pour une valeur de α donnée, l'estimateur RMLE de β est la même que celle donnée en (A.2.4). L'estimateur du RMLE de σ^2 est :

$$\widehat{\sigma}^2(\alpha) = SCR(\alpha)/N - p, \qquad (A.2.7)$$

avec $SCR(\alpha) = \sum_{i=1}^{n} (y_i - Z_i \widehat{\beta}(\alpha))' V_i(\alpha)^{-1} (y_i - Z_i \widehat{\beta}(\alpha))$.

L'estimateur RMLE de α maximise :

$$L^*(\alpha) = -\frac{1}{2} \left[N \log SCR(\alpha) + \sum_{i=1}^{n} \log |V_i(\alpha)| + \log \left(\sum_{i=1}^{n} |Z_i' V_i(\alpha) Z_i| \right) \right]. \qquad (A.2.8)$$

L'estimation de α se fait à l'aide de procédures itératives : algorithme de Newton-Raphson ou algorithme d'Estimation-Maximisation. Une fois l'estimation α' obtenue en maximisant $L(\alpha)$ ou $L^*(\alpha)$, on obtient les estimations de β et de σ^2 en remplaçant α par α' dans les équations (A.2.4) et (A.2.5) ou (A.2.7) selon la méthode utilisée MLE ou REML.

A.2.3. Estimation des effets aléatoires individuels b_i

L'estimation des effets aléatoires $\{\widehat{b_i}\}$ est basée sur la fonction de densité des $\{b_i\}$ conditionnellement aux données $\{y_i\}$. La distribution *a posteriori* de b_i ne dépend que de y_i car les $\{b_i\}$ et les $\{e_i\}$ sont indépendants.

$$\widehat{b_i} = E(b_i/y_i) = \widehat{Q}W_i'\widehat{V}_i^{-1}(y_i - Z_i\widehat{\beta}). \qquad (A.2.9)$$

A.2.4. Inférences sur les paramètres du modèle

En utilisant (A.2.4), et en remplaçant α par $\widehat{\alpha}$ et σ^2 par $\widehat{\sigma}^2$, on obtient :

$$\widehat{\beta} \propto N(\beta, \widehat{V}) \qquad (A.2.10)$$

avec $\widehat{V} = \widehat{\sigma}^2 \left(\sum_{i=1}^{n} Z_i' V_i(\widehat{\alpha}) Z_i \right)^{-1}$.

Soit H_0 l'hypothèse nulle à tester :

$$\psi = L\beta = \psi_0, \qquad (A.2.11)$$

où L est une matrice de dimensions $l \times p$ et de rang l.

Sous l'hypothèse nulle,

$$(\widehat{\psi} - \psi_0)'(L\widehat{V}L')^{-1}(\widehat{\psi} - \psi_0) \qquad (A.2.12)$$

suit approximativement une distribution du χ^2 à l degrés de liberté.

En ce qui concerne les paramètres α définissant la structure de la matrice des variances- covariances, il est préférable d'utiliser les tests du rapport de vraisemblance, la vraisemblance du modèle étant soit celle du MLE soit celle du REML. Stram et Lee [10] ont montré que le test du rapport de vraisemblance est un test conservateur; en effet, la distribution asymptotique du rapport de vraisemblance sous l'hypothèse nulle est en fait un mélange de distributions du χ^2 plutôt qu'une distribution du χ^2.

D'autre part, le choix entre deux modèles partageant la même partie fixe, mais présentant des matrices des variances-covariances de structures différentes peut se faire en utilisant le critère d'information d'Akaike [8] ou le critère d'information de Schwarz [9]. Verbeke et Molenberghs [11] proposent une adaptation de ces deux critères prenant en compte le résultat de Stram et Lee.