Test du score pour l'exploration de modèles complexes: Application à la modélisation dynamique du VIH.

Julia Drylewicz, Daniel Commenges, Rodolphe Thiébaut

Test du score pour l'exploration de modèles complexes : Application à la modélisation dynamique du VIH.

Julia Drylewicz, Daniel Commenges & Rodolphe Thiébaut

INSERM, U897 Centre de Recherche Epidémiologie et Biostatistiques
Université Victor Segalen Bordeaux 2, 146 rue Léo Saignat, 33076 Bordeaux cedex, France

Résumé : Dans la modélisation des systèmes biologiques, des modèles de plus en plus complexes sont développés. L'estimation des paramètres de tel modèle implique des temps de calculs parfois longs. Un grand nombre de modèles est à explorer incluant des effets aléatoires ainsi que des variables explicatives sur les différents paramètres. Le test du score peut, dans ce contexte, être utile pour inclure de nouvelles variables dans le modèle. Cependant, l'hypothèse nulle du test peut être complexe, du à la présence d'effets aléatoires dans le modèle et il se peut également que la matrice d'information ne puisse pas être calculée et qu'une approximation basée sur les scores soit nécessaire. Nous proposons l'utilisation de ces tests du score dans le contexte de la modélisation dynamique du VIH à partir de systèmes d'équations différentielles. Nous nous intéressons aux statistiques du test du score pour tester l'effet de variables explicatives et la variance d'effets aléatoires. Pour chacun de ces tests nous étudions leur erreur de type 1 ainsi que leur puissance. Nous appliquons ces tests aux paramètres d'un modèle de l'interaction entre le VIH et le système immunitaire estimés sur un jeu de données de patients infectés par le VIH issu de la Collaboration CASCADE.

Mots clés : Test du score, homogénéité, effets aléatoires, distribution asymptotique, modèle dynamique

Abstract : To model biologic systems, more and more complex models are developed. Parameters estimation of such models can be very time-consuming. A lot of models including random effects and explanatory variables may be explored. In this context, the score tests can be useful in order to include new variables in the model. However, the null hypothesis can be complex due to random effects and it may happen that the information matrix cannot be computed and an approximation based on the scores is necessary. We propose to use score tests in the context of HIV dynamics models. We examine the score test statistics for testing the effect of explanatory variables and the variance of random effects and we study the type I error and the statistical power of this score test statistics. We apply the score tests to the estimates obtained from a real dataset of HIV infected patients from the CASCADE Collaboration.

Key words : Score test, homogeneity, random effects, asymptotic distribution, dynamical model

1 Test du score pour modèles complexes

1.1 Le modèle non-linéaire mixte

Pour un sujet i $(i = 1, ..., n)$, on considère un modèle qui donne la distribution du vecteur d'observations $Y_i = (Y_{ij}, j = 1, \ldots, n_i)$ en terme de paramètres individuels $\boldsymbol{\xi}^i = (\xi_l^i, l = 1, \ldots, p)$, eux-mêmes modélisés par $\tilde{\xi}_l^i = \psi_l(\xi_l^i)$ sous la forme linéaire : $\tilde{\xi}_l^i = \phi_l + \omega_l u_l^i + z_l^i \beta_l$, où ϕ_l est l'intercept, ω_l est l'écart-type de l'effet aléatoire, z_l^i est un vecteur de variables explicatives pour le l-ième paramètre. Les β_l sont les effets fixes et on suppose $\mathbf{u^i} \sim \mathcal{N}(\mathbf{0}, \mathbf{I_p})$, où $\mathbf{u^i} = (u_l^i, l = 1, \ldots, p)$ est le vecteur individuel des effets aléatoires. On note $\boldsymbol{\theta} = (\phi_l, \omega_l, \boldsymbol{\beta_l}, l = 1, \ldots, p)$ les paramètres du modèle, lui-même noté $(P^{\boldsymbol{\theta}})_{\boldsymbol{\theta} \in \Theta}$.

Pour un sujet i, la vraisemblance des observations conditionnellement aux effets aléatoires est notée $\mathcal{L}^{\boldsymbol{\theta}}_{\mathcal{Y}_i|\mathbf{u^i}}(Y_i|\mathbf{u})$, quand $\mathbf{u^i}$ prend la valeur $\mathbf{u}$. La logvraisemblance pour le sujet i est alors :

$$L_i^{\boldsymbol{\theta}} = \log \int_{\mathbb{R}^p} \mathcal{L}^{\boldsymbol{\theta}}_{\mathcal{Y}_i|\mathbf{u^i}}(Y_i|\mathbf{u}) \phi(\mathbf{u}) d\mathbf{u}$$

où ϕ est la densité d'une loi normale multivariée $N(0, I_p)$. Nous notons $L = \sum_{i=1}^{n} L_i^{\boldsymbol{\theta}}$.

1.2 Test du score pour variable explicative

Nous voulons tester l'effet possible d'une variable explicative z_l^i sur le paramètre ξ_l^i $(l = 1, ..., p)$. L'hypothèse nulle H_0 est "$\boldsymbol{\beta_l} = 0$" qui définit un sous-modèle $(P^{\boldsymbol{\theta}})_{\boldsymbol{\theta} \in \Theta_0}$, où Θ_0 est le sous-ensemble de Θ tel que $\boldsymbol{\beta_l} = 0$. Le score de $\boldsymbol{\beta_l}$, $U_{\beta_l}^{\cdot}(\theta)$, est $\frac{\partial L}{\partial \beta_l}|_\theta$. Comme $U_{\beta_l}^{\cdot} = \sum_{i=1}^{n} U_{\beta_l}^i$, en appliquant le théorème limite central, on obtient que $\frac{U_{\beta_l}(\theta_*)}{\sqrt{\text{var } U_{\beta_l}^{\cdot}(\theta_*)}}$ suit asymptotiquement une loi normale. On note $\hat{\boldsymbol{\theta}}_0$ l'estimateur du maximum de vraisemblance de $\boldsymbol{\theta}_*$ pour $(P^{\boldsymbol{\theta}})_{\boldsymbol{\theta} \in \Theta_0}$. Si H_0 est vraie, cet estimateur est consistant pour $\boldsymbol{\theta}_*$ et en appliquant le "continuous mapping theorem", on a la même distribution asymptotique pour la statistique du test du score :

$$S = \frac{U_{\beta_l}^{\cdot}(\hat{\boldsymbol{\theta}}_0)}{\sqrt{\widehat{\text{var}}\, U_{\beta_l}^{\cdot}(\hat{\boldsymbol{\theta}}_0)}} \tag{1.1}$$

où $\widehat{\text{var}}\, U_{\beta_l}^{\cdot}(\hat{\boldsymbol{\theta}}_0)$ est un estimateur consistant de var $U_{\beta_l}^{\cdot}(\hat{\boldsymbol{\theta}}_0)$.

Les scores $U^i_{\beta_l}(\hat{\boldsymbol{\theta}}_0)$ sont calculés facilement à partir des scores obtenus durant la procédure d'optimisation. On a $U^i_{\beta_l}(\hat{\boldsymbol{\theta}}_0) = z^i_l U^i_{\phi_l}(\hat{\boldsymbol{\theta}}_0)$; puisque l'algorithme a calculé le gradient de la logvraisemblance à chaque itération, les $U^i_{\phi_l}(\hat{\boldsymbol{\theta}}_0)(i = 1, \ldots, n)$ sont les valeurs obtenues à la dernière itération.

La variance du score $U_{\beta_l}(\hat{\boldsymbol{\theta}}_0)$ sous H_0 peut être asymptotiquement déterminée à partir de la matrice d'information (Cox et Hinkley, 1974). Dans les modèles complexes, la Hessienne ne peut pas toujours être calculée et un estimateur de la variance du score peut être utilisé (Guedj et al., 2007) :

$$\widehat{\text{var}}\, U_{\beta_l}(\hat{\boldsymbol{\theta}}_0) = \sum_{i=1}^{n} {U^i_{\beta_l}}^2(\hat{\boldsymbol{\theta}}_0)$$

1.3 Test du score d'homogénéité

Nous voulons tester la présence d'un effet aléatoire sur le paramètre l. L'hypothèse nulle H_0 est "$\omega_l = 0$". En nous basant sur la forme générale du test du score d'homogénéité développé par Commenges et Jacqmin-Gadda (1997), la statistique du test du score sous H_0 est $T_{WPC} = \mathbf{U}^T W^* \mathbf{U} - tr(\hat{I}_\epsilon W^*)$, où $\mathbf{U}$ est le vecteur des $U^{ij}_l = \frac{\partial \mathcal{L}(\varepsilon_{ij}, j=1,\ldots,n_i; i=1,\ldots,n)}{\partial \varepsilon_{ij}}|_{(\hat{\theta}_0)}$ avec $\varepsilon_{ij} = \omega_l u^i_l$, $\hat{I}_\varepsilon$ est la matrice d'éléments $-\frac{\partial^2 \log \mathcal{L}}{\partial \varepsilon_{ij} \partial \varepsilon_{i'j'}}(\hat{\theta}_0)$ et W^* est la matrice de corrélation de $\mathbf{u^i_l}$ moins la matrice identité.

Les scores U^{ij}_l sont calculés avec : $\frac{\partial \mathcal{L}(\varepsilon_{ij}, j=1,\ldots,n_i; i=1,\ldots,n)}{\partial \varepsilon_{ij}}|_{(\hat{\theta}_0)} = \frac{\partial L}{\partial \phi_l}|_{(\hat{\theta}_0)}$. Dans le cas où aucun effet aléatoire n'est présent sous H_0, la matrice $\hat{I}_\varepsilon$ est diagonale et $\text{Tr}(\hat{I}_\varepsilon W^*) = 0$. Dans les cas complexes (présence d'effets aléatoires sous H_0), nous devons calculer la dérivée seconde de la logvraisemblance; calcul qui peut s'avérer compliqué et qu'on souhaiterait éviter. Cependant, le terme informatif dans l'expression de la statistique de test est $\mathbf{U}^T W^* \mathbf{U}$ mais son espérance n'est pas nulle dans le cas complexe. Nous proposons de garder ce terme et nous calculerons cette espérance par simulations:

$$T = \mathbf{U^T} W^* \mathbf{U} = \sum_{i=1}^{n} T_i \quad \text{où} \quad T_i = \sum_{\substack{j,j' \leq n_i \\ j' \neq j}} U^{ij}_l U^{ij'}_l$$

La statistique de test est alors :

$$S_H = \frac{T - \widehat{\mathrm{E}}_{H_0}[T]}{\sqrt{\widehat{\text{var}}_{H_0}\, T}}$$

où $\widehat{\mathrm{E}}_{H_0}[T]$ et $\widehat{\text{var}}_{H_0}\, T$ sont des estimateurs de l'espérance et de la variance de T sous H_0. Cette statistique suit asymptotiquement une loi normale centrée réduite. Pour obtenir ces estimateurs, nous réalisons un bootstrap paramétrique : nous simulons K jeux de données de n individus sous $P^{\hat{\boldsymbol{\theta}}_0}$; pour chaque replique nous calculons les statistiques $T_{(k)}$ et nos estimateurs sont la moyenne et la variance empirique de ces valeurs.

2 Simulations et Application

2.1 Modèle dynamique

Nous considérons un modèle classique d'équations differentielles ordinaires (Perelson et al. 1996) pour la dynamique des concentrations des virions et des lymphocytes T CD4+ pour un sujet i dans le contexte de l'infection par le VIH. Il inclut trois compartiments : T^i (les cellules cibles : lymphocytes T CD4+), I^i (les cellules infectées) et V^i (les virions circulants). Le modèle peut s'écrire sous la forme suivante :

$$\begin{cases} \frac{dT^i}{dt} = \lambda^i - \gamma V^i T^i - \mu_T^i T^i \\ \frac{dI^i}{dt} = \gamma V^i T^i - \mu_I^i I^i \\ \frac{dV^i}{dt} = \pi^i I^i - \mu_V V^i \end{cases}$$

Les cellules cibles entrent dans le sang circulant à un taux λ^i, elles peuvent être infectées par un virion à un taux γ et meurent au taux μ_T^i. Les CD4 infectés meurent à un taux μ_I^i et produisent de nouveaux virions au taux π^i. Les virions meurent au taux μ_V et peuvent alors infecter d'autres cellules cibles. Nous supposons que γ et μ_V sont constant et commun à tous les sujets. Les paramètres individuels sont $\boldsymbol{\xi}^i = (\lambda^i, \mu_I^i, \pi^i, \mu_T^i)$; γ et μ_V sont supposés connus. Les Y_i sont les observations bruitées de $T^i + I^i$ et V^i. Nous prenons $\tilde{\xi}_l^i = \log(\xi_l^i)$ pour tout l pour assurer la positivité des paramètres.

2.2 Simulations

Nous avons calculé les erreurs de type 1 ainsi que les puissances des deux tests en fonction de l'effet de la variable explicative (binaire ou continue) et de la variance des effets aléatoires. Les erreurs de type 1 étaient de l'ordre de la taille nominale des tests (0.05). Leurs puissances augmentaient avec l'effet de la variable explicative et la variance de l'effet aléatoire et tendaient vers 1.

2.3 Application

Nous avons appliqué ces tests aux paramètres estimés du modèle présenté en section 2.1 sur un jeu de données de 761 patients infectés par le VIH issu de la Collaboration CASCADE. Le modèle incluait des effets aléatoires sur les paramètres λ^i et μ_I^i. Nous avons testé l'effet des variables *sexe* et *âge* sur les différents paramètres. La variable *âge* a été prise en variable continue et en variable binaire. Aucun effet de la variable *sexe* n'a été mis en évidence, de même pour la variable *âge* prise en continue. Nous avons trouvé un effet de l'*âge* pris en variable binaire codée 1 pour les patients âgés ($\geq$ 50 ans) et 0 pour les plus jeunes (50 ans est le 95-ième percentile) sur le paramètre μ_I ($S = -1.99$, p=0.04).

Cela signifie que les patients âgés ont une taux de décès des cellules infectées plus faible, ce qui pourrait refléter une moins bonne réponse immunitaire chez ces patients.

Nous avons également appliqué le test d'homogénéité à ces estimations. Nous n'avons mis en évidence aucun nouvel effet aléatoire : $S_H = 1.42$ (p=0.16) et $S_H = 1.55$ (p=0.12), respectivement pour π et μ_T.

3 Conclusion

Nous avons développé des tests du score dans un contexte non-standard. Les simulations montrent que ces tests fonctionnent bien. Ces tests peuvent être utilisés pour réaliser une exploration rapide d'une famille de modèles complexes en particulier pour les modèles dynamiques.

Bibliographie

[1] Cox, A. et Hinkley, D. (1974) *Asymptotic theory*, Theoretical statistics, Chapman & Hall, London, pages 279-363.

[2] Guedj, J., Thiébaut, R. et Commenges, D. (2007) *Maximum Likelihood Estimation in Dynamical Models of HIV*, Biometrics, 63, 1198-1206.

[3] Commenges, D. et Jacqmin-Gadda, H. (1997) *Generalized Score Test of Homogeneity Based on Correlated Random Effects Models*, Journal of the Royal Statistical Society: Serie B (Methodological), 59, 157-171.

[4] Perelson, A. S., Neumann, A. U., Markowitz, M., Leonard, J. M. et Ho, D. (1996) *Viral dynamics in human immunodeficiency virus type 1 infection*, Science, 271, 1582-1586.