Comparaison de modèles basés sur les données d'origine et des données transformées

JL Foulley

Pour simplifier, nous nous placerons dans le cadre d'une transformation simple de type « puissance » telle que celle de Box-Cox, soit

$$z_i = \left(y_i^{\alpha} - 1 \right) / \alpha$$

où $\alpha > 0$, la valeur limite $\alpha = 0$ conduisant à la transformation logarithmique

On peut par exemple se placer dans le cadre d'un modèle hiérarchique tel que :

$$i)\ \mathbf{z} \mid \boldsymbol{\theta}, \lambda_1 \sim \mathrm{N}\left(\mathbf{T}\boldsymbol{\theta}, \lambda_1 \mathbf{I}_N \right)$$

$$ii)\ \boldsymbol{\theta} \mid \boldsymbol{\beta}, \lambda_2 \sim \mathrm{N}\left[\mathbf{A}\boldsymbol{\beta}, \mathbf{G}\left(\lambda_2 \right) \right]$$

où $\mathbf{T}$ et $\mathbf{A}$ sont des matrices d'incidence connues, $\mathbf{G}\left(\lambda_2 \right)$ est une matrice de variance-covariance dépendant de λ_2 et $\boldsymbol{\beta}$, λ_1 et λ_2 sont des paramètres.

La comparaison des modèles est basée sur la déviance $L(\boldsymbol{\beta},\lambda_1,\lambda_2;\mathbf{y}) = \ln p(\mathbf{y}\mid\boldsymbol{\beta},\lambda_1,\lambda_2)$

$$p(\mathbf{y}\mid\boldsymbol{\beta},\lambda_1,\lambda_2) = p(\mathbf{z}\mid\boldsymbol{\beta},\lambda_1,\lambda)\left|\mathbf{J}\right|$$

où $\left|\mathbf{J}\right|$ est la valeur absolue du jacobien de la transformation $\mathbf{J} = \det\{\partial z_i / \partial y_i\}$ soit ici :

$$\mathbf{J} = \prod_{i=1}^{N} y_i^{\alpha-1}$$

Donc, si le paramètre de la transformation α est connu, l'expression de la déviance en fonction des données d'origine ne pose pas de difficulté particulière.

Il resterait à calculer $p(\mathbf{z}\mid\boldsymbol{\beta},\lambda_1,\lambda)$ qui nécessite une intégration par rapport à $\boldsymbol{\theta}$.

En définitive, $$\boxed{-2L(\mathbf{y}) = -2L(\mathbf{z}) \underbrace{-2(\alpha-1)\sum_{i=1}^{N}\ln(y_i)}_{\text{facteur correctif}}}$$

Exemple/modèles

Modèles d'analyse des scores de cellules somatiques

(Thèse M Duval)

$$1)\ y_{ij} = \phi_{1,i} \exp(\phi_{2,i} t_{ij} + \phi_{3,i} t_{ij}^2 + \phi_{4,i} / t_{ij}) + e_{ij}$$

$$\left(\phi_{1,i}, \phi_{2,i}, \phi_{3,i}, \phi_{4,i}\right)'_{iid}\ N\left(\left(\mu_1, \mu_2, \mu_3, \mu_4\right)', \Gamma\right)$$

$$e_{ij} \sim_{iid}\ N\left(0, \sigma^2\right)$$ (Morand & Gnanasakthy,1989)

Ne passe pas en Proc Mixed

$$2) z_{ij} = \ln\left(y_{ij}\right) = \phi_{1,i}^* + \phi_{2,i} t_{ij} + \phi_{3,i} t_{ij}^2 + \phi_{4,i} / t_{ij} + \varepsilon_{ij}$$

$$\phi_{1,i}^* = \ln\left(\phi_{1,i}\right)$$ Modèle **linéaire** mixte

$$\left(\phi_{1,i}^*, \phi_{2,i}, \phi_{3,i}, \phi_{4,i}\right)'_{iid}\ N\left(\left(\mu_1^*, \mu_2, \mu_3, \mu_4\right)', \Gamma^*\right)$$

$$\varepsilon_{ij} \sim_{iid}\ N\left(0, \sigma^2\right)$$

Exemple/-2L

	-2L (a)	-2L (b)	AIC (a)	AIC (b)
Modèle 1 (y)	17121	17121	17151	17151
Modèle 2 (logy)	3992	4555		
$2\sum_{i=1}^{N}\ln(y_i)$	2x6775	2x6742		
Modèle 2 (retour à y)	17542	18039	17572	18069

a) N=6405 du fait d'un seuil à -1.5 sur log(y)>-1.5
b) N=6424 du fait d'un seuil à -2 sur log(y)>-2.0