La loi négative binomiale pour le traitement des données des microbiotes

SAHBANE Abdesstar

Montpellier University

November 8, 2020

Table of contents

Presentation des données

Les données typiques sur le microbiome générées par le séquençage du gène, comprennent les composants suivants :

- C_{ij} : Counts.
- X_i : Host factors.
- T_i : Total sequence read.
- Z_i : Sample variables.

Loi négative binomiale

L'objectif est de détecter les associations entre les caractéristiques du microbiome C_{ij} et la variable X_i.
$\forall j$ on note $y_i = C_{ij}$ et on suppose que y_i suit la loi négative binomial :

$$y_i \sim NB\left(y_i \mid \mu_i, \theta\right) = \frac{\Gamma\left(y_i + \theta\right)}{\Gamma(\theta) y_i!} \cdot \left(\frac{\theta}{\mu_i + \theta}\right)^{\theta} \cdot \left(\frac{\mu_i}{\mu_i + \theta}\right)^{y_i} \quad (1)$$

où μ_i et θ sont les paramêtres de forme, et Γ et la fonction gamma.

Fonction de lien

Le modèle négative binomiale mixte relie la moyenne μ_i aux variables T_i, X_i et Z_i par le lien :

$$\log(\mu_i) = \log(T_i) + X_i\beta + Z_i b \tag{2}$$

avec $b \sim N_K\left(0, \tau^2 I\right)$

Famille exponentielle

Pour tout paramêtre θ fixe, la loi négative binomiale appartient à la famille exponentielle, et s'écrit sous la forme :

$$NB\left(y_i \mid \mu_i, \theta\right) = \exp\left\{\frac{y_i \vartheta_i - b\left(\vartheta_i\right)}{\phi} + c\left(y_i, \phi\right)\right\} \tag{3}$$

où $\vartheta_i = \log \frac{\mu_i}{\mu_i + \theta}$,

$b\left(\vartheta_i\right) = -\theta \log\left(1 - e^{\log \frac{\mu_i}{\mu_i+\theta}}\right) = -\theta \log\left(1 - e^{\vartheta_i}\right)$, et

$c\left(y_i, \phi\right) = \log\left(\frac{\Gamma(y_i+\theta)\theta^\theta}{\Gamma(\theta) y_i!}\right)$

IWLS

• Le modèle négative binomiale est donc un cas spéciale des modèles linéaires généralisés pout tout θ fixe.
• lorsque θ est inconnu, le modèle négative binomiale n'est pas un GLM. Or, les NBMMs peuvent être ajustés en mettant à jour de manière itérative les paramètres $\left(\beta, b, \tau^2\right)$ et θ.
• Conditionnellement à θ , le NBMM est un GLMM spécial et donc les paramètres (β, b, τ^2) peuvent être mis à jour en utilisant la procédure GLMMs.
• Conditionnellement à (β , b), le paramètre de forme θ peut être mis à jour en maximisant la vraisemblance NB en utilisant l'algorithme standard de Newton – Raphson.

IWLS

• Conditionnellement à θ , nous mettons à jour les paramètres (β, b, τ^2) en étendant l'algorithme IWLS.
• L'algorithme IWLS procède à une approximation de la vraisemblance du modèle linéaire généralisé par une vraisemblance normale pondérée, puis met à jour les paramètres à partir du modèle normal pondéré.
• Conditionnellement au paramètres θ, β, et b, la vraisemblance binomiale négative $NB(y_i \mid \mu_i, \theta)$ peut être approximée par la vraisemblance normale pondérée :

$$NB(y_i \mid \mu_i, \theta) \approx N\left(t_i \mid \eta_i, w_i^{-1}\right)$$

où $\eta_i = \log(T_i) + X_i\beta + Z_i b$, les «données de réponse normales» t_i et les «poids» w_i sont appelés respectivement pseudo-réponse et pseudo-poids.

IWLS

La pseudo-réponse t_i et les pseudo-poids w_i sont calculés par:

$$t_i = \hat{\eta}_i - \frac{L'\left(y_i \mid \hat{\eta}_i, \hat{\theta}\right)}{L''\left(y_i \mid \hat{\eta}_i, \hat{\theta}\right)}, \text{ et } w_i = -L''\left(y_i \mid \hat{\eta}_i, \hat{\theta}\right)$$

Où $\hat{\eta}_i = \log(T_i) + X_i\widehat{\beta} + Z_i\hat{b}$, $L\left(y_i \mid \hat{\eta}_i, \hat{\theta}\right) = \log NB\left(y_i \mid \widehat{\mu}_i, \hat{\theta}\right)$,
$L'(y_i \mid \eta_i, \theta) = dL(y_i \mid \eta_i, \theta)/d\eta_i$, $L''(y_i \mid \eta_i, \theta) =$
$d^2L(y_i \mid \eta_i, \theta)/d\eta_i^2$,
et $(\widehat{\beta}, \hat{b})$ et $\hat{\theta}$ sont les estimations actuelles de (β, b) et θ, respectivement.

IWLS

Par conséquent, les NBMM peuvent être approximés par le modèle mixte linéaire avec w_i comme poids:

$$t_i = \log(T_i) + X_i\beta + Z_i b + w_i^{-1/2} e_i, b \sim N_K\left(0, \tau^2\right), e \sim N_n\left(0, \sigma^2 I\right)$$

Les paramètres $\left(\beta, b, \tau^2, \sigma^2\right)$ sont ensuite mis à jour à partir de ce modèle mixte linéaire en utilisant l'algorithme standard d'ajustement des LMM.

Résumé IWLS

En résumé, l'IWLS pour l'ajustement des NBMM est un algorithme itératif et se déroule comme suit:

1. Initialiser β , b et θ quelques valeurs plausibles;
2. $j = 1, 2, \cdots$
   1. Sur la base des valeurs actuelles $\left(\beta^{(j-1)}, b^{(j-1)}, \theta^{(j-1)}\right)$, calculer la pseudo-réponse $t_i^{(j)}$ et les pseudo-poids $w^{(j)}$;
   2. Mettre à jour $\left(\beta, b, \tau^2, \sigma^2\right)$ en ajustant le LMM;
   3. Mettre à jour θ par l'algorithme standard de Newton – Raphson.
3. Répétez l'étape 2) jusqu'à convergence.

Nous utilisons le critère $\left(\eta^{(j)} - \eta^{(j-1)}\right)^2 < \varepsilon \left(\eta^{(j)}\right)^2$, pour évaluer la convergence, où $\eta^{(j)} = \sum_{i=1}^{n} \left(\log\left(T_i\right) + X_i\beta^{(j)} + Z_i b^{(j)}\right)$, et ε est une petite valeur (disant 10^{-5}).

Frame Title

Merci