Estimation bootstrap de l'efficacité relative de deux sérums

Projet MAP 433

Notre objectif est de comparer les niveaux d'anticorps anticoronavirus dans deux échantillons de sérum prélevés en mai et en juin sur une même vache. Les données sont disponibles à l'adresse suivante:
`http://www.cmap.polytechnique.fr/~giraud/MAP433/data.Rdata`
Ce sont des mesures de densité optique Y pour différents niveaux de dilution d. Pour chaque date et chaque dilution on dispose de deux mesures de densité optique du sérum.

Modèle statistique: on modélise la densité $Y_{i,j}^{(\text{mois})}$ pour le mois "mois", la log-dilution $x_i = \log_{10}(1/d_i)$ (où d_i est la dilution du sérum), et la mesure j par

$$Y_{i,j}^{(\text{mois})} = f(x_i, \theta^{(\text{mois})}) + \varepsilon_{i,j}^{(\text{mois})}, \quad i = 1, \dots, k, \quad j = 1, \dots, r, \text{ et mois} = \text{mai, juin}$$

avec

$$f(x, \theta) = \theta_1 + \frac{\theta_2 - \theta_1}{1 + \exp(\theta_3(x - \theta_4))}$$

et les $\varepsilon_{i,j}^{(\text{mois})}$ i.i.d. centrées et de variance σ^2 inconnue.
Le but de l'analyse est d'estimer l'efficacité relative ρ du sérum de juin par rapport au sérum de mai. Le sérum de juin a une efficacité relative ρ par rapport au sérum de mai s'il se comporte comme une dilution ρ du sérum de mai. Autrement dit, ρ est tel que

$$f(x, \theta^{(\text{mai})}) = f(x + \log_{10}\rho, \theta^{(\text{juin})}), \quad \text{pour tout } x \in [0, 1]. \tag{1}$$

On estime $\theta^{(\text{mois})}$ par $\hat{\theta}^{(\text{mois})}$ minimisant

$$M^{(\text{mois})}(\theta) = \frac{1}{kr} \sum_{i=1}^{k} \sum_{j=1}^{r} (Y_{i,j}^{(\text{mois})} - f(x_i, \theta))^2,$$

et on note $n = kr$, $\theta = (\theta^{(\text{mai})}, \theta^{(\text{juin})})^T$ et $\hat{\theta}_n = (\hat{\theta}^{(\text{mai})}, \hat{\theta}^{(\text{juin})})^T$.

1 Etude théorique préliminaire

Le nombre r de répétitions étant supposé fixe, on va étudier le comportement de $\hat{\theta}_n$ lorsque $n = kr \to \infty$. On note $\nabla_\theta f$ le gradient de f par rapport à θ et on suppose que

$$\frac{1}{k} \sum_{i=1}^{k} (\nabla_\theta f)(\nabla_\theta f)^T(x_i, \theta) = \underbrace{\int_0^1 (\nabla_\theta f)(\nabla_\theta f)^T(x, \theta)\, dx}_{=H(\theta)} + O(1/k)$$

1. Montrez que $\sqrt{n}(\hat{\theta}_n - \theta)$ converge en loi lorsque $n \to \infty$ et identifiez la loi limite en fonction de σ^2 et des $H(\theta^{(\text{mois})})$.

   Indications: comme $\hat{\theta}^{(\text{mai})}$ et $\hat{\theta}^{(\text{juin})}$ sont indépendants, étudiez les séparément. Et toute ressemblance avec le paragraphe 5.4.2 du poly n'est pas fortuite....

2. Pour estimer ρ, il faut commencer par vérifier que la relation (1) est en accord avec les données. Nous allons donc tester

$$\mathbf{H0}: \ \theta_i^{(\text{mai})} = \theta_i^{(\text{juin})} \ i = 1,2,3 \quad \text{contre } \mathbf{H1}: \exists i \in \{1,2,3\} \text{ tel que } \theta_i^{(\text{mai})} \neq \theta_i^{(\text{juin})}.$$

Remarquez que **H0** s'écrit $A\theta = 0$ pour une certaine matrice A de taille 3×8 et de rang 3. Si $A\theta = 0$, montrez que $\sqrt{n}A\hat{\theta}_n$ converge en loi vers une variable gaussienne $\mathcal{N}(0, \sigma^2 V(\theta))$.

3. Proposez un estimateur consistant $\hat{\sigma}_n^2$ de σ^2. On pose

$$T_n = \frac{n}{\hat{\sigma}_n^2}(A\hat{\theta}_n)^T \widehat{V}(\hat{\theta}_n)^{-1} A\hat{\theta}_n,$$

avec $\widehat{V}(\theta)$ la version empirique de $V(\theta)$. Quel est le comportement asymptotique de T_n lorsque $A\theta \neq 0$? Montrez que si $A\theta = 0$, la variable T_n converge en loi vers un $\chi^2(3)$.

4. Proposez un test de niveau asymptotique 5% de **H0** contre **H1**.

2 Estimation des paramètres et analyse du parallélisme

1. Calculez $\hat{\theta}_n$ et T_n. L'hypothèse **H0** est-elle rejetée?
2. On note $\tilde{\theta}_n$ le minimiseur de $M^{(\text{mai})}(\theta^{(\text{mai})}) + M^{(\text{juin})}(\theta^{(\text{juin})})$ sous la contrainte $A\theta = 0$. On estime ρ par $\hat{\rho} = 10^{(\tilde{\theta}_4^{(\text{juin})} - \tilde{\theta}_4^{(\text{mai})})}$. calculez $\hat{\rho}$.

3 Intervalle de confiance par ré-échantillonage

En adaptant les calculs de la première partie, on peut déduire la loi limite de $\sqrt{n}(\hat{\rho} - \rho)$ et batir des intervalles de confiance asymptotique pour ρ. Lorsque la taille n de l'échantillon est faible, cette approximation n'est cependant pas valide. Dans ce cas, une alternative intéressante est de construire un intervalle de confiance par bootstrap. Le principe du bootstrap est le suivant. Supposons qu'on puisse répéter B fois l'expérience biologique de façon indépendantes. Pour l'expérience b on obtiendrait un estimateur $\hat{\rho}_b$ et on aurait donc un échantillon de B estimateurs i.i.d. distribués comme $\hat{\rho}$. Pour B grand on pourrait donc estimer la distribution de $\hat{\rho}$ (et ainsi contruire un intervalle de confiance pour ρ).
Les méthodes de ré-échantillonage sont une manière de mimer la répétition d'une expérience. Les estimateurs $\hat{\rho}_b$ sont calculés à partir d'un échantillon bootstrap

$$Y_{i,j}^{(\text{mois}),b} = f(x_i, \tilde{\theta}_n^{(\text{mois})}) + \varepsilon_{i,j}^{(\text{mois}),b}$$

où les ε^b sont obtenus de la façon suivante. On note $\hat{\varepsilon}_{i,j}^{(\text{mois})} = Y_{i,j}^{(\text{mois})} - f(x_i, \tilde{\theta}_n^{(\text{mois})})$ et $\tilde{\varepsilon}_{i,j}$ les $\hat{\varepsilon}_{i,j}$ recentrés (en retranchant la moyenne empirique). Pour chaque b, chaque variable $\varepsilon_{i,j}^{(\text{mois}),b}$ est obtenue par un tirage aléatoire (avec remise) parmi les $\{\tilde{\varepsilon}_{i,j}^{(\text{mois})}, \ i = 1, \ldots, k, j = 1, \ldots, r, \text{mois} = \text{mai, juin}\}$.
Enfin, l'intervalle de confiance bootstrap de ρ de niveau α est donné par

$$I_B(\alpha) = [F_B^*(\alpha/2), F_B^*(1-\alpha/2)] \quad \text{où} \quad F_B^*(\alpha) = \inf\left\{ x \in \mathbf{R} : \ \frac{1}{B}\sum_{b=1}^{B} \mathbf{1}_{\{\hat{\rho}_b \leq x\}} \geq \alpha \right\}.$$

1. Calculer l'intervalle $I_B(\alpha)$ pour $B = 1000$ et $\alpha = 5\%$. (Il est possible de montrer que $I_B(\alpha)$ est un intervalle de confiance de niveau asymptotique α lorsque $B, n \to \infty$. La preuve de ce résultat est difficile.)

2. L'approche précédente est le bootstrap naif. On peut montrer qu'il vaut mieux chercher à estimer la loi de $\hat{Z} = \sqrt{\frac{n}{\hat{s}^2}}(\hat{\rho} - \rho)$ par bootstrap pour obtenir un intervalle de confiance plus précis.
a. Quelle est la loi asymptotique de $\sqrt{n}(\hat{\rho} - \rho)$? en déduire un estimateur $\hat{s}^2$ de sa variance asymptotique s^2.
b. Pour chaque échantillon bootstrap on peut calculer la variance $\hat{s}_b^2$ et $\hat{Z}_b = \sqrt{\frac{n}{\hat{s}_b^2}}(\hat{\rho}_b - \hat{\rho})$.
Les $\hat{Z}_b$ fournissent un échantillon bootstrap de $\hat{Z}$. On peut donc estimer pour $x \in \mathbf{R}$ la probabilité $\mathbf{P}(\hat{Z} \leq x)$ par $\frac{1}{B}\sum_{b=1}^{B} \mathbf{1}_{\{\hat{Z}_b \leq x\}}$. En déduire un intervalle de confiance $I'_B(\alpha)$ de ρ en fonction de $G^*(\alpha/2)$ et $G^*(1-\alpha/2)$ où

$$G_B^*(\alpha) = \inf\left\{ x \in \mathbf{R} : \ \frac{1}{B}\sum_{b=1}^{B} \mathbf{1}_{\{\hat{Z}_b \leq x\}} \geq \alpha \right\}.$$

Il est possible de montrer que cet intervalle est de meilleure qualité que $I_B(\alpha)$.