MAP433 Statistique

PC7 - Tests asymptotiques

1 Rapport de vraisemblance généralisé

On considère le modèle suivant. Soit $\beta \in \mathbb{R}$ et $t \in (0,1)$. On a n observations $X_1, \ldots, X_n$ modélisées par des variables indépendantes et telles que $X_i \sim \mathcal{N}(0,1)$ si $i/n \leq t$ et $X_i \sim \mathcal{N}(\beta,1)$ si $i/n > t$, où $\mathcal{N}(\mu,1)$ est la loi gaussienne scalaire de moyenne μ et de variance 1. Dans un premier temps, on suppose les paramètres β et t inconnus.

1. Montrer que la log-vraisemblance s'écrit comme une fonction croissante de $L_n(\beta,t)$, où

$$L_n(\beta,t) = -\left(\sum_{i=1}^{[tn]} X_i^2 + \sum_{[tn]+1}^{n} (X_i-\beta)^2\right),$$

   avec $[u]$ la partie entière de u.

2. On note β_0 et t_0 les vrais paramètres du modèle. Montrer que, pour tout $\beta \in \mathbb{R}$,

$$-\frac{1}{n} L_n(\beta,t) \overset{\mathbb{P}_{\beta_0,t_0}}{\longrightarrow} \begin{cases} 1+\beta_0^2\,|t-t_0| + (1-t)\,(\beta_0-\beta)^2 & \text{si } t_0 \leq t \\ 1+\beta^2\,|t-t_0| + (1-t_0)\,(\beta_0-\beta)^2 & \text{si } t_0 \geq t \end{cases}$$

3. En quelles valeurs de $t \in (0,1)$ et $\beta \in \mathbb{R}$ la limite ci-dessus est-elle minimale ?

On suppose désormais le paramètre t connu et fixé dans $(0,1)$, $t = t_0$.

4. Calculer l'estimateur $\hat{\beta}_n$ de β_0 par maximum de vraisemblance.
5. Montrer que le test du rapport de vraisemblance généralisé de

$$H_0 = \{\text{la moyenne des } X_i \text{ est nulle pour i=1,...,n}\}$$

   contre

$$H_1 = \{\text{la moyenne des } X_i \text{ devient non-nulle à partir de } i > [t_0 n]\}.$$

   s'exprime en fonction de $L_n(\hat{\beta}_n, t_0) - L_n(0, t_0)$. Donner le test de niveau asymptotique α et exprimer la p-valeur à l'aide de la fonction de répartition d'une loi classique.

2 Test asymptotique pour la régression logistique

Soit $\theta \in \mathbb{R}^p$ un paramètre vectoriel à estimer. On considère une suite $Y_1, \ldots, Y_n$ de variables de Bernoulli indépendantes de paramètres respectifs $p_1(\theta), \ldots, p_n(\theta)$ donnés par

$$p_i(\theta) = \varphi(\theta^T x_i)$$

où $\varphi(t) = (1 + e^{-t})^{-1}$ et où $x_1, \ldots, x_n$ est une suite de vecteurs déterministes de taille p. On suppose que la matrice $[x_1, \ldots, x_n]$ est de rang p. On note la fonction de log-vraisemblance $L_n(\theta)$ des observations $(Y_1, \cdots, Y_n)$. On supposera toujours que le supremum de L_n est atteint. On note respectivement $\nabla L_n(\theta)$ et $\nabla\nabla^T L_n(\theta)$ le gradient et la matrice hessienne de L_n au point θ.

1. Donner l'allure de la fonction φ.
2. Écrire $L_n(\theta)$ en fonction de θ et des observations $(x_i, Y_i)_{i=1\ldots n}$.
3. Montrer que

$$\nabla L_n(\theta) = \sum_{i=1}^{n} (Y_i - p_i(\theta)) x_i$$

$$-\nabla\nabla^T L_n(\theta) = \sum_{i=1}^{n} h(\theta^T x_i) x_i x_i^T \qquad \text{où} \;\; h = \varphi(1-\varphi)$$

   En déduire que $-\nabla\nabla^T L_n(\theta)$ est définie positive.
4. Montrer que l'estimateur du maximum de vraisemblance $\hat{\theta}_n$ est défini comme la solution d'un système d'équations que l'on précisera.

On souhaite établir la normalité asymptotique de $\sqrt{n}(\hat{\theta}_n - \theta)$ sous $\mathbb{P}_\theta$. On fait les hypothèses suivantes :

- $\hat{\theta}_n$ converge probabilité vers θ, ce qu'on écrit $\hat{\theta}_n = \theta + o_P(1)$.
- Pour tout θ, $Q_\theta := \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} h(\theta^T x_i) x_i x_i^T$ existe et est inversible.
- $\sup_n \frac{1}{n} \sum_{i=1}^{n} \|x_i\|^3 < \infty$.

5. En utilisant le développement de Taylor

$$\nabla L_n(\hat{\theta}_n) = \nabla L_n(\theta) + \nabla\nabla^T L_n(\tilde{\theta}_n)(\hat{\theta}_n - \theta)\,,$$

   pour $\|\tilde{\theta}_n - \theta\| \le \|\hat{\theta}_n - \theta\|$, montrer que

$$\left(\frac{\nabla\nabla^T L_n(\theta)}{n} + o_P(1)\right) \sqrt{n}(\hat{\theta}_n - \theta) = -\frac{\nabla L_n(\theta)}{\sqrt{n}}$$

   Indication : On pourra utiliser que h est 1-lipschitzienne.
6. Montrer que $-\frac{\nabla L_n(\theta)}{\sqrt{n}}$ est asymptotiquement normal.
7. Conclure que $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, Q_\theta^{-1})$.
8. On note $\beta_{n,k}$ le $k$ème coefficient de la diagonale de $(-\frac{\nabla\nabla^T L_n(\hat{\theta}_n)}{n})^{-1}$. Montrer que

$$\sqrt{\frac{n}{\beta_{n,k}}}\, (\hat{\theta}_{n,k} - \theta_k) \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)\,.$$

9. En déduire un test de niveau asymptotique α pour $H_0 \;:\; \theta_k = 0$ contre $H_0 \;:\; \theta_k \neq 0$.

3 Test d'adéquation du χ^2

Une étude sur l'utilisation d'une station Velib a établi que le nombre de clients ramenant un vélo entre 8h et 8h10 un jour de semaine suit une loi de Poisson de paramètre 4.
L'analyse des données de la même station récoltées sur $n = 200$ jours à la même tranche horaire indique les nombres de retours de vélo suivant (on a gardé uniquement des jours de semaine pendant lesquels des emplacements libres étaient disponibles pour le retour de vélos durant les 10 minutes considérées).

Nombre de retours	0	1	2	3	4	5	6	7	8	9	10	11
Effectifs observés	6	15	40	42	37	30	10	9	5	3	2	1

On note N_i le nombre de retours observés au jour i et on suppose les N_i indépendants. Enfin, on note $N'_i = \min(N_i, 8)$ et Q la loi de $\min(N, 8)$ lorsque $N \sim \mathcal{P}(4)$.

1. Calculer la p-valeur du test du χ^2 d'adéquation des N'_i à la loi Q.
2. On suppose que N_i sont indépendants et suivent une loi de Poisson de paramètre λ. Proposer un test de niveau asymptotique $\alpha = 5\%$ de $\lambda = 4$ contre $\lambda \neq 4$. Calculer la p-valeur associée.

Table de la loi de Poisson :

k	0	1	2	3	4	5	6	7	8	9	10	11	12	13
$\mathbb{P}_{3.7}[X = k]$	.025	.091	.169	.209	.193	.143	.088	.047	.021	.009	.003	.001	3.10^{-4}	10^{-4}
$\mathbb{P}_4[X = k]$	.018	.073	.146	.195	.195	.156	.104	.059	.03	.013	.005	.002	6.10^{-4}	2.10^{-4}

4 BONUS 1 Test de Wilcoxon

Une firme pharmaceutique a mis au point une nouvelle molécule pour faire chuter le taux de sucre dans le sang. Pour tester l'efficacité de cette molécule, elle le compare à un placebo. Elle réunit $n + m$ patients. A un premier groupe de m individus, elle administre un placebo (sans leur dire !). Au second groupe elle donne sa nouvelle molécule. Après un délai approprié, on mesure les taux de glycémie $\{X_i : i = 1, \ldots, n\}$ et $\{Y_i : i = 1, \ldots, m\}$ chez les deux groupes.

placebo : X	médicament : Y
1.0	1.4
1.41	0.94
0.61	3.1
0.22	0.54
5.9	1.2
0.84	0.043
0.49	3.0
	0.40
	0.075
	1.1

On suppose que les $X_1, \ldots, X_n$ (resp. $Y_1, \ldots, Y_m$) sont i.i.d. de loi de fonction de répartition F_X (resp. F_Y). On supposera que les lois sont diffuses, en particulier F_X et F_Y sont continues. On veut tester si les lois des X et des Y sont les mêmes ou si les Y_i son t stochastiquement plus petits que les X_i, c'est à dire $F_X < F_Y$. On va donc tester $H_0 : F_X = F_Y$ contre $H_1 : F_X < F_Y$.
On pose $Z_i = X_i$ pour $i = 1, \ldots, n$ et $Z_{n+i} = Y_i$ pour $i = 1, \ldots, m$. On note $R(i)$ le rang de Z_i dans la suite $(Z_1, \ldots, Z_{n+m})$, à savoir, $R(i) = 1$ si Z_i est la plus petite valeur, $R(i) = 2$ si

Z_i est la seconde plus petite valeur, etc. La statistique de Wilcoxon est définie pa r

$$W_{n,m} = \sum_{i=1}^{n} R(i).$$

L'idée est que sous H_1 la statistique $W_{n,m}$ sera plus grande que sous H_0.

1. Montrez que la permutation $R : \{1, \dots, n+m\} \to \{1, \dots, n+m\}$ suit sous H_0 une loi uniforme sur l'ensemble $\mathcal{S}_{n+m}$ des permutations de $\{1, \dots, n+m\}$. En déduire que sous H_0 la loi de $W_{n,m}$ ne dépend pas de la loi F_X. Quelle est la loi de $R(i)$ sous H_0 ?
2. Montrez que sous H_0 on a $\mathbb{E}(W_{n,m}) = n(n+m+1)/2$.
3. Montrez que sous H_0 on a $\text{Var}(W_{n,m}) = n\text{Var}(R(1)) + n(n-1)\text{Cov}(R(1), R(2))$ et
$$0 = \text{Var}\left(\sum_{i=1}^{n+m} R(i)\right) = (n+m)\text{Var}(R(1)) + (n+m)(n+m-1)\text{Cov}(R(1), R(2)).$$
En déduire que $\text{Var}(W_{n,m}) = nm(n+m+1)/12$ sous H_0.
4. En admettant que $T_{n,m} = (W_{n,m} - \mathbb{E}(W_{n,m}))/\sqrt{\text{Var}(W_{n,m})}$ converge en loi sous H_0 vers une loi normale $\mathcal{N}(0,1)$ lorsque $n \to \infty$, testez au niveau asymptotique 5% si la nouvelle molécule est plus efficace que le placebo.

5 BONUS 2 Test de naissance : l'approche bayésienne

Soient $X_1, \dots, X_n$ n v.a. i.i.d de Bernoulli : $\mathbb{P}_\theta(X_i) = \theta^{X_i}(1-\theta)^{1-X_i}$, $\theta \in \Theta = [0,1]$.

1. Définir un test UPP de risque de première espèce α pour $H_0 = \{\theta = \theta_0\}$ contre $H_1 = \{\theta = \theta_1\}$.
2. Proposer un test asymptotique pour l'hypothèse $H_0 = \{\theta \geq \theta_0\}$ contre $H_1 = \{\theta < \theta_0\}$ basé sur la statistique $S_n = \sum_{i=1}^{n} X_i$.
3. Notons π la densité de la loi a priori pour θ par rapport à la mesure de Lebesgue sur $[0,1]$. Ecrire la densité a posteriori $\pi(\theta | S_n = y)$.
4. Pour loi a priori pour θ, nous avons besoin d'une loi dont le support soit inclus dans l'intervalle $[0,1]$. Parmi les choix possibles de telles lois, il est intéressant de considérer la famille des lois Beta. Rappelons que les lois *Beta* dépendent de deux paramètres $\alpha, \beta > 0$ et que la densité de la loi $\text{Beta}(\alpha, \beta)$ est donnée par :
$$b_{\alpha,\beta}(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}, \quad 0 < x < 1,$$
où $B(\alpha, \beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)$ est la fonction Beta. Quel est la loi posteriori correspondante ?
5. En déduire le test Bayesien de l'hypothèse $H_0 = \{\theta \geq \theta_0\}$ contre $H_1 = \{\theta < \theta_0\}$ pour cette loi a priori.
6. Montrer que la moyenne de la loi a posteriori (i.e. l'espérance conditionnelle de θ sachant $S_n = y$) est donnée par
$$\mathbb{E}\left[\theta | S_n = y\right] = \frac{\alpha + y}{\alpha + \beta + n}.$$
On utilisera que $\Gamma(x+1) = x\Gamma(x)$ pour tout $x > 0$.

7. Montrer que la variance a posteriori vaut

$$\mathrm{Var}[\theta|S_n = y] = \frac{\mathbb{E}[\theta|S_n = y](1 - \mathbb{E}[\theta|S_n = y])}{\alpha + \beta + n + 1}.$$

8. Que remarque-t-on quand n et y sont grands ?
9. La loi $\beta(1, 1)$ est la loi uniforme sur $[0, 1]$. C'est la loi a priori utilisée par Bayes (1763) et redécouverte indépendamment par Laplace (1800), fondateurs de l'estimation bayésienne, pour l'analyse bayésienne du modèle de Bernoulli. La motivation première de Laplace était de déterminer si le nombre de garçons et de filles à la naissance suivait une loi de Bernoulli de paramètre 0.5. Un total de 241945 filles et de 251527 garçons sont nés à Paris de 1745 à 1770. En appelant "succès" la naissance d'un enfant de sexe féminin, calculer une borne de

$$\pi(\theta \geq 0.5|S_n = 241945) \text{pour } n = 241945 + 251527,$$

obtenue avec l'inégalité de Tchebyschev-Cantelli ($\mathbb{P}(Z - \mathbb{E}[Z] \geq \delta) \leq \mathrm{Var}(Z)/(\delta^2 + \mathrm{Var}(Z))$).

Laplace a en fait montré que

$$\text{pour } n = 241945 + 251527,\ \pi(\theta \geq 0.5|S_n = 241945) = 1.15 \times 10^{-42},$$

montrant qu'avec une probabilité très proche de 1, $\theta < 0.5$. Comme nous l'avons noté ci-dessus, pour des valeurs aussi grandes, l'influence de la loi a priori est tout à fait négligeable, la loi a posteriori étant extrêmement concentrée autour de la valeur $\hat{\theta}_n = 241945/(241945 + 251527) = 0.49$.