CB2 Maths II - Vendredi 8 mars 2024

La régression logistique permet de modéliser l'influence qu'exercent des facteurs exogènes sur une variable binaire, c'est-à-dire une variable ne pouvant prendre que deux valeurs.

Outre son domaine d'application privilégié qui est l'apprentissage automatique (machine learning), la régression logistique est couramment utilisée aussi bien en médecine qu'en actuariat et en économétrie.

Dans tous les programmes Python, on suppose que l'on a déjà importé `import numpy as np` et `import numpy.random as rd`.

Partie I. Fonction logistique et lois logistiques

On appelle fonction logistique la fonction Λ définie sur $\mathbb{R}$ par :

$$\forall x \in \mathbb{R}, \quad \Lambda(x) = \frac{1}{1+\mathrm{e}^{-x}}$$

1. (a) Montrer que Λ est une bijection de $\mathbb{R}$ sur $]0,1[$, dont la bijection réciproque est la fonction L définie par :

$$\forall x \in]0,1[, \quad L(x) = \ln\left(\frac{x}{1-x}\right)$$

   (b) Calculer la dérivée de la fonction Λ.

   (c) Justifier l'existence d'un unique réel x_0 tel que : $\Lambda(x_0) = x_0$.

   (d) Établir pour tout $x \in \mathbb{R}$, l'inégalité : $|\Lambda(x) - x| \leqslant |x - x_0|$.

2. Le script Python suivant, dont les lignes 1. et 2. définissent la fonction Λ, permet de calculer une valeur approchée de x_0 par la méthode de dichotomie.

```
1.def Lambda(x):
2.      return 1/(1+np.exp(-x))
3. a=0
4. b=1
5. eps=......
6. while b-a>eps :
7.     c=(a+b)/2
8.     if Lambda(c)>c:
9.          .........
10.   else :
11.        b=.........
12. x0=(a+b)/2
```

   Vous justifierez avec précision les réponses aux questions ci-dessous.

   (a) Compléter les lignes 9. et 11. et justifier le choix des valeurs affectées en lignes 3. et 4. aux variables a et b.

   (b) Quelle valeur <u>maximale</u> peut-on affecter en ligne 5. à la variable `eps` pour être assuré que l'erreur d'approximation commise ne dépasse pas 10^{-4} ?

   (c) Que peut-on dire de la valeur numérique affichée par l'instruction `13.` suivante ?
   `13. print(Lambda (x0)-x0)`

3. On note λ la dérivée de la fonction Λ.

   (a) Vérifier que λ est une densité de probabilité.

   (b) Préciser la parité de la fonction λ ; donner l'allure de sa courbe représentative dans le plan rapporté à un repère orthogonal et en déterminer les points d'inflexion.

   On dit qu'une variable aléatoire Z suit la **loi logistique standard** si elle admet la fonction λ pour densité. Pour tout couple $(r,s) \in \mathbb{R}\times\mathbb{R}_+^*$, on dit qu'une variable aléatoire Y suit **la loi logistique** $\mathcal{L}(r,s)$ si la variable aléatoire Z définie par $Z = \dfrac{Y-r}{s}$ suit la loi logistique standard.

4. (a) Justifier qu'une variable aléatoire Z qui suit la loi logistique standard admet des moments de n'importe quel ordre et vérifier que $E(Z) = 0$.

   (b) Soit Y une variable qui suit une loi logistique $\mathcal{L}(r,s)$. Calculer $E(Y)$.

   (c) En utilisant la méthode d'inversion, écrire le script d'une fonction Python intitulée `def randlogis(n,p,r,s):` fournissant pour tout couple (n,p) d'entiers strictement positifs, une matrice S à n lignes et p colonnes dont les coefficients sont des simulations de variables aléatoires indépendantes suivant la loi logistique $\mathcal{L}(r,s)$.

   (d) Décrire un procédé permettant de calculer une valeur approchée de la variance de la loi logistique standard à l'aide de la fonction `randlogis`.

5. Soit U_1 et U_2 deux variables aléatoires indépendantes suivant chacune la loi exponentielle de paramètre 1 .

   (a) Montrer que la variable aléatoire $Z = \ln\left(\dfrac{U_1}{U_2}\right)$ suit la loi logistique standard (on pourra utiliser un changement de variable exponentiel, c'est-à-dire de la forme $t = \mathrm{e}^x$).

   (b) En déduire un nouveau script Python permettant de simuler une variable aléatoire suivant la loi logistique standard à l'aide de la fonction `rd.exponential`.

Partie II. Variance de la loi logistique standard

On admet que $\displaystyle\sum_{k=1}^{+\infty} \frac{1}{k^2} = \frac{\pi^2}{6}$ $\quad(*)$.

6. Soit Z une variable aléatoire suivant la loi logistique standard.

   (a) À l'aide d'une intégration par parties, justifier que la variance de Z, notée $V(Z)$, vérifie l'égalité :

$$V(Z) = 4\int_0^{+\infty} \frac{x\mathrm{e}^{-x}}{1+\mathrm{e}^{-x}}\,\mathrm{d}x$$

   (b) Établir pour tout $n \in \mathbb{N}$, l'égalité :

$$\int_0^{+\infty} \frac{x\mathrm{e}^{-x}}{1+\mathrm{e}^{-x}}\,\mathrm{d}x = \sum_{k=0}^{n}(-1)^k \int_0^{+\infty} x\mathrm{e}^{-(k+1)x}\,\mathrm{d}x + I_n, \quad \text{où} \quad I_n = (-1)^{n+1}\int_0^{+\infty} \frac{x\mathrm{e}^{-(n+2)x}}{1+\mathrm{e}^{-x}}\,\mathrm{d}x$$

   (c) Montrer que l'intégrale I_n tend vers 0 lorsque n tend vers $+\infty$ et en déduire l'égalité :

$$\int_0^{+\infty} \frac{x\mathrm{e}^{-x}}{1+\mathrm{e}^{-x}}\,\mathrm{d}x = \sum_{k=0}^{+\infty} \frac{(-1)^k}{(k+1)^2}$$

(d) En utilisant la formule (*), déduire de l'égalité précédente que $V(Z) = \dfrac{\pi^2}{3}$.

7. Établir la convergence des deux intégrales $\displaystyle\int_0^{+\infty} \ln(x)\mathrm{e}^{-x}\,\mathrm{d}x$ et $\displaystyle\int_0^{+\infty} (\ln(x))^2\mathrm{e}^{-x}\,\mathrm{d}x$.

8. On pose $I = \displaystyle\int_0^{+\infty} \ln(x)\mathrm{e}^{-x}\,\mathrm{d}x$ et $J = \displaystyle\int_0^{+\infty} (\ln(x))^2\mathrm{e}^{-x}\,\mathrm{d}x$.
En utilisant le résultat de la question 5.(a), calculer $J - I^2$.

Partie III. Estimation à partir de données binaires

Dans cette partie, θ est un paramètre réel inconnu et F désigne la fonction de répartition d'une variable aléatoire à densité dont une densité f est continue et strictement positive sur $\mathbb{R}$.

Soit $(Y_n)_{n\in\mathbb{N}^*}$ une suite de variables aléatoires indépendantes définies sur un espace probabilisé $(\Omega, \mathcal{A}, \mathbf{P}_\theta)$ suivant chacune la loi de Bernoulli de paramètre $F(\theta)$.

9. Justifier que F est une bijection de $\mathbb{R}$ sur $]0,1[$. On note F^{-1} sa bijection réciproque.

10. Pour tout $n \in \mathbb{N}^*$, on pose : $\bar{Y}_n = \dfrac{1}{n}\displaystyle\sum_{j=1}^{n} Y_j$.

Montrer que la suite $\left(\sqrt{n}\left(\bar{Y}_n - F(\theta)\right)\right)_{n\in\mathbb{N}^*}$ converge en loi vers une variable aléatoire suivant une loi normale centrée dont on précisera la variance.

11. Pour tout $n \in \mathbb{N}^*$ et tout $\omega \in \Omega$, on pose : $T_n(\omega) = \begin{cases} F^{-1}\left(\bar{Y}_n(\omega)\right) & \text{si } 0 < \bar{Y}_n(\omega) < 1 \\ 0 & \text{sinon} \end{cases}$.

De plus, pour tout $n \in \mathbb{N}^*$, on note E_n l'événement $\left[0 < \bar{Y}_n < 1\right]$.

(a) Calculer $\mathbf{P}_\theta(E_n)$ et trouver la limite de cette probabilité lorsque n tend vers $+\infty$.

(b) Soit $x \in \mathbb{R}$ et $n \in \mathbb{N}^*$.

i. Établir l'égalité ensembliste $\{\omega \in E_n \ / T_n(\omega) \leqslant x\} = \left[\bar{Y}_n \leqslant F(x)\right] \cap E_n$.

ii. Justifier l'encadrement :

$$\mathbf{P}_\theta\left(\left[\bar{Y}_n \leqslant F(x)\right] \cap E_n\right) \leqslant \mathbf{P}_\theta\left([T_n \leqslant x]\right) \leqslant \mathbf{P}_\theta\left(\left[\bar{Y}_n \leqslant F(x)\right] \cap E_n\right) + 1 - \mathbf{P}_\theta(E_n)$$

(c) Montrer que pour tout $x \neq \theta$, on a : $\lim_{n\to+\infty} \mathbf{P}_\theta\left([T_n \leqslant x]\right) = \begin{cases} 0 & \text{si } x < \theta \\ 1 & \text{si } x > \theta \end{cases}$.

(d) En déduire que $(T_n)_{n\in\mathbb{N}^*}$ est une suite d'estimateurs du paramètre θ qui converge en loi vers θ.

(e) Montrer que $(T_n)_{n\in\mathbb{N}^*}$ est une suite convergente d'estimateurs du paramètre θ.

12. Pour tout $n \in \mathbb{N}^*$ et tout $\omega \in \Omega$, on pose :

$$U_n(\omega) = \begin{cases} \frac{T_n(\omega)-\theta}{\bar{Y}_n(\omega)-F(\theta)} & \text{si } \bar{Y}_n(\omega) \neq F(\theta) \\ \frac{1}{f(\theta)} & \text{si } \bar{Y}_n(\omega) = F(\theta) \end{cases}$$

On admet sans démonstration que pour tout $n \in \mathbb{N}^*$, U_n est une variable aléatoire sur $(\Omega, \mathcal{A}, \mathbf{P}_\theta)$.

(a) Soit $\varepsilon > 0$.

Pour tout $n \in \mathbb{N}^*$, on note $B_n(\varepsilon)$ l'événement $\left[\left|U_n - \frac{1}{f(\theta)}\right| \leqslant \varepsilon\right]$.

i. Établir l'existence d'un réel $\alpha > 0$ tel que : $\forall x \in [\theta-\alpha, \theta+\alpha], \left|\dfrac{1}{f(x)} - \dfrac{1}{f(\theta)}\right| \leqslant \varepsilon$.

ii. Pour un tel α, justifier l'inclusion : $[|T_n - \theta| \leqslant \alpha] \cap E_n \subset B_n(\varepsilon)$, où E_n a été défini dans la question 11.

iii. Montrer que la suite $(U_n)_{n\in\mathbb{N}^*}$ converge en probabilité vers $\dfrac{1}{f(\theta)}$.

(b) On admet le lemme de Slutsky : si (X_n) et (Y_n) sont deux suites de variables aléatoires telles que $(X_n)_{n\in\mathbb{N}}$ converge **en loi** vers X, et $(Y_n)_{n\in\mathbb{N}}$ converge **en probabilité** vers une variable aléatoire certaine égale à c, alors la suite $(X_nY_n)_{n\in\mathbb{N}}$ converge en loi vers la variable aléatoire cX.

Montrer alors que la suite $\left(\sqrt{n}\,(T_n-\theta)\right)_{n\in\mathbb{N}^*}$ converge en loi vers une variable aléatoire suivant une loi normale centrée dont on précisera la variance.

Partie IV. Régression logistique

— Dans toute cette partie, p désigne un entier supérieur ou égal à 2 .
— Pour tout couple $(n,m)\in(\mathbb{N}^*)^2$, on note $\mathcal{M}_{n,m}(\mathbb{R})$ l'ensemble des matrices à n lignes et m colonnes à coefficients réels et ${}^tM$ la transposée de toute matrice $M\in\mathcal{M}_{n,m}(\mathbb{R})$.
— Pour tout $m\in\mathbb{N}^*$, le produit scalaire usuel de deux vecteurs u et v de $\mathbb{R}^m$ est noté $\langle u,v\rangle$. Si U et V sont les matrices colonnes représentant u et v dans la base canonique, le produit scalaire $\langle u,v\rangle$ est donc l'unique coefficient de la matrice ${}^tUV$.
— On rappelle que les fonctions Λ et L ont été définies dans la partie I.

Dans cette partie, on note Y une variable aléatoire de Bernoulli, dite variable endogène, dont la loi dépend du niveau de p facteurs exogènes.

L'influence de ces facteurs sur la loi de Y est résumée par la fonction b qui associe à un vecteur $x\in\mathbb{R}^p$, la probabilité $b(x)$ que Y soit égale à 1 lorsque les niveaux des facteurs sont donnés par les composantes du vecteur x.

Dans le modèle de régression logistique envisagé dans cette partie, la fonction b est supposée de la forme :

$$b : x \longmapsto \Lambda(\langle\alpha, x\rangle)$$

où $\alpha=(\alpha_1,\alpha_2,\ldots,\alpha_p)$ est un vecteur de $\mathbb{R}^p$ dont les composantes $\alpha_1,\alpha_2,\ldots,\alpha_p$ sont des paramètres inconnus qui représentent les degrés d'influence des divers facteurs exogènes sur la variable endogène Y.

Pour estimer les paramètres du modèle, on dispose de k vecteurs $x^{(1)},x^{(2)},\ldots,x^{(k)}$ de $\mathbb{R}^p$ $(k\in\mathbb{N}^*)$ et pour tout $i\in[\![1,k]\!]$, d'une suite $(Y_{i,n})_{n\in\mathbb{N}^*}$ de variables aléatoires indépendantes suivant chacune la loi de Bernoulli de paramètre $b(x^{(i)})=\Lambda\left(\left\langle\alpha,x^{(i)}\right\rangle\right)$.

Pour chaque indice fixé i et pour tout $n\in\mathbb{N}^*$, les variables aléatoires $Y_{i,1},Y_{i,2},\ldots,Y_{i,n}$ définissent donc un n-échantillon associé à la loi de la variable endogène lorsque les niveaux des facteurs exogènes sont les composantes $x_1^{(i)},x_2^{(i)},\ldots,x_p^{(i)}$ du vecteur $x^{(i)}$ dans la base canonique de $\mathbb{R}^p$.

13. On note respectivement A et M la matrice du vecteur α et la matrice de la famille $\left(x^{(1)},x^{(2)},\ldots,x^{(k)}\right)$ dans la base canonique de $\mathbb{R}^p$:

$$A=\begin{pmatrix}\alpha_1\\ \alpha_2\\ \vdots\\ \alpha_p\end{pmatrix}\in\mathcal{M}_{p,1}(\mathbb{R})\quad\text{et}\quad M=\begin{pmatrix}x_1^{(1)} & \ldots & x_1^{(k)}\\ \vdots & & \vdots\\ x_p^{(1)} & \ldots & x_p^{(k)}\end{pmatrix}\in\mathcal{M}_{p,k}(\mathbb{R})$$

On suppose que le rang de la matrice M est égal à p.

(a) Montrer que la matrice M^tM est inversible.

(b) Soit H une matrice, $H \in \mathcal{M}_{k,1}(\mathbb{R})$. Montrer qu'il existe une unique matrice $U \in \mathcal{M}_{p,1}(\mathbb{R})$ pour laquelle l'unique coefficient de la matrice ${}^t\left({}^tMU - H\right)\left({}^tMU - H\right)$ est le plus petit possible. Quelle propriété vérifie alors cette matrice ?

On admet que $U = \left(M^tM\right)^{-1} MH$.

(c) Expliquer pourquoi les lois des variables aléatoires $Y_{i,n}$ ne suffiraient pas à définir le vecteur α si le rang de M n'était pas égal à p.

14. Pour tout $n \in \mathbb{N}^*$ et tout $i \in [\![1, k]\!]$, on pose $\bar{Y}_{i,n} = \frac{1}{n}\sum_{j=1}^{n} Y_{i,j}$ et pour tout $\omega \in \Omega$:

$$T_{i,n} = \begin{cases} L\left(\bar{Y}_{i,n}(\omega)\right) & \text{si } 0 < \bar{Y}_{i,n}(\omega) < 1 \\ 0 & \text{sinon} \end{cases}$$

(a) Soit $(c_1, c_2, \ldots, c_k) \in \mathbb{R}^k$. En utilisant les résultats de la partie III, montrer que $\left(\sum_{i=1}^{k} c_i T_{i,n}\right)_{n\in\mathbb{N}^*}$ est une suite convergente d'estimateurs du paramètre $\sum_{i=1}^{k} c_i \left\langle \alpha, x^{(i)} \right\rangle$.

(b) Pour tout $n \in \mathbb{N}^*$ et tout $\omega \in \Omega$, on pose

$$H_n(\omega) = \begin{pmatrix} T_{1,n}(\omega) \\ T_{2,n}(\omega) \\ \vdots \\ T_{k,n}(\omega) \end{pmatrix} \quad \text{et} \quad \begin{pmatrix} A_{1,n}(\omega) \\ A_{2,n}(\omega) \\ \vdots \\ A_{p,n}(\omega) \end{pmatrix} = \left(M^tM\right)^{-1} MH_n(\omega)$$

Montrer que pour tout $j \in [\![1, p]\!]$, la suite $(A_{j,n})_{n\in\mathbb{N}^*}$ est une suite convergente d'estimateurs de α_j.