



PRÉPASIMPLE

prepasimple.fr

HEC — 2019

Sujet de mathématiques approfondies

Mathématiques approfondies — Prépa ECG

Annales de concours · Sujet

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.

Partie I : Fonction logistique et lois logistiques

On appelle fonction logistique la fonction Λ définie sur $\mathbb{R}$ par : $\forall x \in \mathbb{R}, \Lambda(x) = \frac{1}{1 + e^{-x}}$.

- 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 : $|\Lambda(x) - x_0| \leq |x - x_0|$.

- 2) Le code Python suivant, dont les lignes (1) à (3) définissent la fonction Λ , permet de calculer une valeur approchée de x_0 par la méthode de dichotomie. Les numéros de lignes sont indiqués en commentaire.

Python

```
import numpy as np                # (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 # (13)

```

- a) Compléter les lignes (10) et (12) et justifier le choix des valeurs affectées en lignes (4) et (5) aux variables a et b.
- b) Quelle valeur maximale peut-on affecter en ligne (6) 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 obtenue par l'instruction (14) suivante?

Python

```
print(Lambda(x0) - x0) # (14)
```

- 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 $Z = \frac{Y - r}{s}$ suit la loi logistique standard.

- 4) a) Justifier qu'une variable aléatoire qui suit une loi logistique $\mathcal{L}(r, s)$ admet des moments de n'importe quel ordre et en indiquer l'espérance.
- b) On rappelle qu'après les instructions `import numpy as np` et `import numpy.random as rd`, l'appel `rd.random((n, p))` renvoie un tableau à n lignes et p colonnes dont les coefficients sont des simulations de variables aléatoires indépendantes de loi uniforme sur $[0, 1[$, et que les opérations $+$, $-$, $*$, $/$ ainsi que `np.log` appliquées à un tableau agissent coefficient par coefficient. En utilisant la méthode d'inversion, écrire le code d'une fonction Python `grandlogis(n, p, r, s)`, fournissant pour tout couple (n, p) d'entiers strictement positifs, un tableau S à n lignes et p colonnes dont les coefficients sont des simulations de variables aléatoires indépendantes suivant la loi logistique $\mathcal{L}(r, s)$.
- c) Décrire un procédé permettant de calculer une valeur approchée de la variance de la loi logistique standard à l'aide de la fonction `grandlogis`.
- 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 $Z = \ln\left(\frac{U_1}{U_2}\right)$ suit la loi logistique standard. On pourra déterminer une densité de chacune des variables aléatoires $\ln(U_1)$ et $-\ln(U_2)$, puis utiliser un produit de convolution et le changement de variable exponentiel $u = e^t$.
- b) On rappelle qu'après l'instruction `import numpy.random as rd`, l'appel `rd.exponential(1)` renvoie une simulation de la loi exponentielle de paramètre 1 (l'argument de `rd.exponential` est l'inverse du paramètre de la loi). En déduire un nouveau code 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

Pour tout polynôme

$$P = \sum_{k=0}^d a_k X^k \in \mathbb{R}[X]$$

de degré $d \in \mathbb{N}$, les termes non nuls $a_k X^k$ sont appelés les monômes de P et les a_k leurs coefficients.

Pour tout $n \in \mathbb{N}$, on pose :

$$P_n = \sum_{k=0}^n (-1)^k \binom{2n+1}{2k+1} (X-1)^{n-k}$$

- 6) a) Expliciter les polynômes P_0 et P_1 .
- b) Pour tout $n \in \mathbb{N}^*$, préciser le degré du polynôme P_n et donner les coefficients de ses deux monômes de plus hauts degrés.
- c) Soit $n \in \mathbb{N}^*$. On suppose que P_n admet n racines réelles deux à deux distinctes $t_1, t_2, \dots, t_n$. Justifier que

$$P_n = a \prod_{k=1}^n (X - t_k),$$

où a désigne le coefficient dominant de P_n , puis utiliser le résultat précédent pour montrer que

$$\sum_{k=1}^n t_k = \frac{2n(n+1)}{3}.$$

7) Soit $x \in \mathbb{R}$ et $n \in \mathbb{N}$.

a) Montrer par récurrence que pour tout $m \in \mathbb{N}$, on a :

$$\begin{aligned}\cos(mx) &= \sum_{j=0}^{\lfloor m/2 \rfloor} (-1)^j \binom{m}{2j} \cos^{m-2j}(x) \sin^{2j}(x) \quad \text{et} \quad \sin(mx) \\ &= \sum_{j=0}^{\lfloor (m-1)/2 \rfloor} (-1)^j \binom{m}{2j+1} \cos^{m-2j-1}(x) \sin^{2j+1}(x)\end{aligned}$$

avec les conventions $\cos^0(x) = 1$ et qu'une somme indexée par l'ensemble vide est nulle.
En déduire l'égalité :

$$\sin((2n+1)x) = \sum_{k=0}^n (-1)^k \binom{2n+1}{2k+1} \cos^{2(n-k)}(x) \sin^{2k+1}(x)$$

b) En déduire, pour tout $x \in]0, \pi[$, la relation : $\frac{\sin((2n+1)x)}{\sin^{2n+1}(x)} = P_n\left(\frac{1}{\sin^2(x)}\right)$.

c) À l'aide des résultats des questions 6) c) et 7) b), montrer que pour tout $n \in \mathbb{N}^*$, on a :

$$\sum_{k=1}^n \frac{1}{\sin^2\left(\frac{k\pi}{2n+1}\right)} = \frac{2n(n+1)}{3}$$

8) Soit $x \in \left]0, \frac{\pi}{2}\right[$.

a) Justifier les inégalités suivantes : $\sin(x) \leq x \leq \tan(x)$ et $\frac{1}{\sin^2(x)} - 1 \leq \frac{1}{x^2} \leq \frac{1}{\sin^2(x)}$.

b) En utilisant le résultat de la question 7) c), en déduire, pour tout $n \in \mathbb{N}^*$, l'encadrement :

$$\frac{n(2n-1)}{3} \leq \frac{(2n+1)^2}{\pi^2} \sum_{k=1}^n \frac{1}{k^2} \leq \frac{2n(n+1)}{3}$$

c) Établir l'égalité :

$$\sum_{k=1}^{+\infty} \frac{1}{k^2} = \frac{\pi^2}{6}.$$

9) 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 :

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

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

$$\begin{aligned} \int_0^{+\infty} \frac{x e^{-x}}{1 + e^{-x}} dx &= \sum_{k=0}^n (-1)^k \int_0^{+\infty} x e^{-(k+1)x} dx + I_n, \quad \text{où} \quad I_n \\ &= (-1)^{n+1} \int_0^{+\infty} \frac{x e^{-(n+2)x}}{1 + e^{-x}} dx \end{aligned}$$

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

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

d) En utilisant la formule établie en 8) c), déduire de l'égalité précédente que la variance de Z est égale à $\frac{\pi^2}{3}$.

10) a) Établir la convergence des deux intégrales $\int_0^{+\infty} \ln(x) e^{-x} dx$ et $\int_0^{+\infty} (\ln(x))^2 e^{-x} dx$.

b) On pose $I = \int_0^{+\infty} \ln(x) e^{-x} dx$ et $J = \int_0^{+\infty} (\ln(x))^2 e^{-x} dx$. 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 $(\Omega, \mathcal{A}, P_\theta)$ suivant chacune la loi de Bernoulli de paramètre $F(\theta)$.

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

12) Pour tout $n \in \mathbb{N}^*$, on pose :

$$\bar{Y}_n = \frac{1}{n} \sum_{j=1}^n Y_j.$$

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

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

$$T_n(\omega) = \begin{cases} F^{-1}(\bar{Y}_n(\omega)) & \text{si } 0 < \bar{Y}_n(\omega) < 1 \\ 0 & \text{sinon} \end{cases}$$

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

- a) Calculer $P_\theta(E_n)$ 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 $\{\omega \in E_n ; T_n(\omega) \leq x\} = [\bar{Y}_n \leq F(x)] \cap E_n$. On admet que T_n est une variable aléatoire sur $(\Omega, \mathcal{A}, P_\theta)$.
- (ii) Justifier l'encadrement :

$$P_\theta([\bar{Y}_n \leq F(x)] \cap E_n) \leq P_\theta([T_n \leq x]) \leq P_\theta([\bar{Y}_n \leq F(x)] \cap E_n) + 1 - P_\theta(E_n)$$

- c) Montrer que pour tout $x \neq \theta$, on a : $\lim_{n \rightarrow +\infty} P_\theta([T_n \leq x]) = \begin{cases} 0 & \text{si } x < \theta \\ 1 & \text{si } x > \theta \end{cases}$
- d) En déduire que $(T_n)_{n \in \mathbb{N}^*}$ est une suite convergente d'estimateurs du paramètre θ .

14) 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 $(\Omega, \mathcal{A}, P_\theta)$.

- a) Soit $\varepsilon > 0$. Pour tout $n \in \mathbb{N}^*$, on note $B_n(\varepsilon)$ l'événement $\left[\left| U_n - \frac{1}{f(\theta)} \right| \leq \varepsilon \right]$.

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

(ii) Pour un tel α , justifier l'inclusion : $[|T_n - \theta| \leq \alpha] \cap E_n \subset B_n(\varepsilon)$, où E_n a été défini dans la question 13).

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

c) On note $c = \frac{1}{f(\theta)}, \sigma^2$ la variance obtenue à la question 12) et, pour tout $n \in \mathbb{N}^*$,
 $W_n = \sqrt{n} (\bar{Y}_n - F(\theta))$.

(i) Vérifier que pour tout $n \in \mathbb{N}^*$, on a : $\sqrt{n} (T_n - \theta) = U_n W_n$.

(ii) Soit $x \in \mathbb{R}$ et $\varepsilon \in]0, c[$. On pose $m_\varepsilon = \min \left\{ \frac{x}{c - \varepsilon}, \frac{x}{c + \varepsilon} \right\}$ et
 $M_\varepsilon = \max \left\{ \frac{x}{c - \varepsilon}, \frac{x}{c + \varepsilon} \right\}$. Établir pour tout $n \in \mathbb{N}^*$ l'encadrement :

$$P_\theta([W_n \leq m_\varepsilon]) - P_\theta(\overline{B_n(\varepsilon)}) \leq P_\theta([U_n W_n \leq x]) \leq P_\theta([W_n \leq M_\varepsilon]) + P_\theta(\overline{B_n(\varepsilon)})$$

(iii) En déduire que la suite $(\sqrt{n} (T_n - \theta))_{n \in \mathbb{N}^*}$ 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 ${}^t M$ 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 ${}^t U V$.

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 :

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

où $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_p)$ est un vecteur de $\mathbb{R}^p$ dont les composantes $\alpha_1, \alpha_2, \dots, \alpha_p$ 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 $x^{(1)}, x^{(2)}, \dots, x^{(k)}$ de $\mathbb{R}^p$ ($k \in \mathbb{N}^*$) et pour tout $i \in \llbracket 1, k \rrbracket$, d'une suite $(Y_{i,n})_{n \in \mathbb{N}^*}$ de variables aléatoires indépendantes suivant chacune la loi de Bernoulli de paramètre $b(x^{(i)}) = \Lambda(\langle \alpha, x^{(i)} \rangle)$.

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

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

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

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

- a) Montrer que la matrice $M^t M$ est inversible.
- b) Montrer que pour toute matrice $H \in \mathcal{M}_{k,1}(\mathbb{R})$, la matrice $U \in \mathcal{M}_{p,1}(\mathbb{R})$ pour laquelle l'unique coefficient de la matrice ${}^t(MU - H)$ (${}^tMU - H$) est le plus petit possible, est la matrice $(M^t M)^{-1} M H$.
- 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 .

16) Pour tout $n \in \mathbb{N}^*$ et tout $i \in \llbracket 1, k \rrbracket$, on pose

$$\bar{Y}_{i,n} = \frac{1}{n} \sum_{j=1}^n Y_{i,j}$$

et pour tout $\omega \in \Omega$:

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

a) Soit $(c_1, c_2, \dots, c_k) \in \mathbb{R}^k$. 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

$$\sum_{i=1}^k c_i \langle \alpha, x^{(i)} \rangle.$$

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

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

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