TP : Mathématiques statistiques et Apprentissage

Régression linéaire

26 mars 2021

Partie Théorique

$$\hat{\beta}_n((y_i, x_i^{(1)}, \cdots, x_i^{(1)})_{i\in\{1,\cdots,n\}}) = \text{argmin}_{\beta\in\mathbb{R}^d}\|y - x\beta\|^2.$$

1. Donner le modèle statistique associé à (1) dans le cas où x est supposé déterministe et aléatoire.

   Réponse :

   - Si x est déterministe : on peut prendre $\mathcal{M} = (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \{Q_\epsilon\epsilon \in \mathbb{R}^n\})$ avec Q_ϵ est la densité de la variable aléatoire ϵ.

     Exemple : Le modèle linéaire gaussien $Y = X\beta + \epsilon$ tel que :

     - $\epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ est un vecteur de n réalisations indépendantes d'une v.a.r normale de moyenne 0 et de variance σ^2 inconnue.
     - X est une matrice (n, d) de rang d.
     - β est inconnu de $\mathbb{R}^d$.

     Donc on peut prendre $\mathcal{M}_1 = (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \{\mathcal{N}(0, \sigma^2 I_n), \sigma > 0\})$ ou $\mathcal{M}_2 = (\mathbb{R}, \mathcal{B}(\mathbb{R}), \{\mathcal{N}(0, \sigma^2), \sigma > 0\})^{\otimes n}$ le modèle n-échantillons.
   - Si x est aléatoire : $\mathcal{M} = (\mathbb{R}^n \times \mathbb{R}^n, \mathcal{B}(\mathbb{R}^n) \otimes \mathcal{B}(\mathbb{R}^n)), \{Q_\epsilon\epsilon \in \mathbb{R}^n\} \otimes \{P_x, x \in \mathbb{R}^n\})$ avec Q_ϵ et P_x les densités respectives des variables ϵ et de x.

2. Montrer que si $\epsilon \sim \mathcal{N}(0, \sigma^2 I_N)$ alors $\hat{\beta}_n$ est l'estimateur du maximum de vraisemblance du modèle où l'on suppose que x est déterministe.

Dans ce cas, on a $Y \sim \mathcal{N}(x\beta, \sigma^2 I_n)$. Soit donc $L(\beta, \sigma) = \frac{1}{\sqrt{2\pi}^n} \times \frac{1}{\sigma^n} e^{-\frac{\|y-x\beta\|^2}{2\sigma^2}}$ la fonction de vraisemblance correspondant à $Y = x\beta + \epsilon$.
Donc la fonction log-vraisemblance est donnée par :

$$l(\beta, \sigma) = -\frac{n}{2}\log(2\pi) - n\log(\sigma) - \frac{1}{2\sigma^2}\|y - x\beta\|^2.$$

On cherche

$$(\hat{\beta}_{MV}, \hat{\sigma}_{MV}) = \text{argmax}_{\beta\in\mathbb{R}^d,\sigma>0} l(y,\sigma) = \text{argmin}_{\beta\in\mathbb{R}^d,\sigma>0} -l(y,\sigma) = \text{argmin}_{\beta\in\mathbb{R}^d,\sigma>0} n\log(\sigma) + \frac{}{2}$$

Donc si σ ne fait pas partie des paramètres à estimer, alors on voit que l'on retombe sur l'estimateur $\hat{\beta}_n$ des moindres carrés, donc ils sont équivalents.

- On estime $\hat{\beta}_{MV}$:

l est lisse donc $\hat{\beta}_{MV}$ est extrémum de $\beta \to l(\beta, \sigma)$ ssi il est extrémum de $\beta \to \|y - x\beta\|$ une fonction convexe non bornée, donc ssi est un minimum GLOBAL de $\beta \to l(\beta, \sigma)$.
Puisque ce minimum existe $(\hat{\beta}_n)$, $\hat{\beta}_{MV}$ existe donc et il coïncide donc avec $\hat{\beta}_n$ l'estimateur de moindre carrée.

- On estime $\hat{\sigma}_{MV}$:

Comme l est assez lisse et donc admet des dérivées partielles on a :

$$\frac{\partial l(\hat{\beta}_{MV}, \sigma)}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\|y - x\hat{\beta}_{MV}\|^2$$

s'annule au point $\hat{\sigma}_n$ tel que :

$$\hat{\sigma}_n^2 = \frac{1}{n}\|y - x\hat{\beta}_{MV}\|^2.$$

Une étude de variation de la fonction $\sigma \to l(y, \sigma)$ montre qu'elle est croissante sur $]0, \hat{\sigma}_n]$ et décroissante sur $[\hat{\sigma}_n, +\infty[$ donc on déduit que : $\hat{\sigma}_n = \hat{\sigma}_{MV}$ avec

$$\hat{\sigma}_n^2 = \hat{\sigma}_{MV}^2 = \frac{1}{n}\|y - x\hat{\beta}_{MV}\|^2.$$

On a trouvé donc $(\hat{\beta}_{MV}, \hat{\sigma}_{MV}) = \text{argmax}_{\beta\in\mathbb{R}^d,\sigma>0} l(y,\sigma) = (\hat{\beta}_n, \frac{1}{n}\|y - x\hat{\beta}_n\|^2)$.

Par suite l'estimateur $\hat{\beta}_n$ est l'estimateur du maximum de vraisemblance.

3. Montrer que l'estimateur est toujours bien défini.

   $F = x(\mathbb{R}^d)$ est un sous-espace vectoriel d'espace euclidien $\mathbb{R}^n$ en tant qu'image d'un espace vectoriel par une application linéaire. Puisque tout espace vectoriel de dimension finie est fermé (et convexe aussi), la projection de tout point y de $\mathbb{R}^n$ sur F existe et est unique, et elle est définie par l'unique point $p_F(y) \in F$ tel que $\|y - p_F(y)\| = \inf_{z\in F} \|y - z\|$.

   L'application x est par définition surjective de $\mathbb{R}^d$ à F, donc il existe $\hat{\beta}_n((y_i, x_i^{(1)}, \cdots, x_i^{(1)})_{i\in\{1,\cdots,n\}})$ unique tel que $x\hat{\beta}_n((y_i, x_i^{(1)}, \cdots, x_i^{(1)})_{i\in\{1,\cdots,n\}}) = p_F(y)$.

   Et qui vérifie donc,

   $$\hat{\beta}_n((y_i, x_i^{(1)}, \cdots, x_i^{(1)})_{i\in\{1,\cdots,n\}}) = \text{argmin}_{\beta\in\mathbb{R}^d}\|y - x\beta\|^2.$$

4. Montrer que si $d \leq n$, alors x est injective si et seulement si x^Tx est inversible.

   Supposons $d \leq n$:

   - Si x est injective : la matrice x^Tx est positive car $y^Tx^Txy = \|xy\|^2 \geq 0$. Et puisque x est injective alors $\text{Ker}(x) = \{0\}$. Par suite, x^Tx est définie positive. Et puisqu'elle est symétrique, donc par le théorème spectral, x^Tx est diagonalisable ayant ses valeurs propres toutes strictement positives.

     En particulier $\det(x^Tx) \neq 0$, donc x^Tx est inversible.
   - Si x^Tx est inversible : supposons qu'il existe y non nul tel que $xy = 0$, alors $x^Txy = 0$, ce qui est impossible car x^Tx est inversible (en particulier injective). Donc, x est injective.

5. Pour $z \in \mathbb{R}^d$, on pose $f(z) = \|y - xz\|^2$. On a f est différentiable et

   $$f(z+h) = f(\beta) - 2\langle y - xz, xh\rangle + \|xh\|^2 = f(\beta) - 2\langle x^T(y - xz), h\rangle + o(h).$$

   Donc, $\nabla f(z) = 2x^T(xz - y)$. f est fonction convexe comme composition des fonctions convexes, donc atteint son minimum global en $\hat{\beta}_n$, donc $\nabla f(\hat{\beta}_n) = 2x^T(x\hat{\beta}_n - y) = 0$.

Par suite, sous l'hypothèse que x est injective, on a $x^T x$ est inversible et par suite :

$$\hat{\beta}_n = (x^T x)^{-1} x^T y.$$

6. On est dans le cas où x est déterministe.

- On a $\epsilon \sim \mathcal{N}(0, \sigma^2)$, donc comme $Y = x\beta + \epsilon$, $Y \sim \mathcal{N}(x\beta, \sigma^2)$ par suite :

$$\hat{\beta}_n(Y) = (x^T x)^{-1} x^T Y \sim \mathcal{N}((x^T x)^{-1} x^T x\beta, \sigma^2((x^T x)^{-1} x^T)((x^T x)^{-1} x^T)^T).$$

Donc,

$$\hat{\beta}_n(Y) \sim \mathcal{N}(\beta, \sigma^2 (x^T x)^{-1}).$$

- $\hat{\epsilon}_n(Y) = Y - x\hat{\beta}_n = [I_n - x(x^T x)^{-1} x^T]Y$, donc comme $Y \sim \mathcal{N}(x\beta, \sigma^2)$, on déduit que

$$\hat{\epsilon}_n(Y) \sim \mathcal{N}([I_n - x(x^T x)^{-1} x^T]x\beta, \sigma^2 [I_n - x(x^T x)^{-1} x^T][I_n - x(x^T x)^{-1} x^T]^T).$$

D'où :

$$\hat{\epsilon}_n(Y) \sim \mathcal{N}(0, \sigma^2 [I_n - x(x^T x)^{-1} x^T]).$$

7. En déduire un estimateur non biaisé de β et de σ^2.

- $\hat{\beta}_n(Y) \sim \mathcal{N}(\beta, \sigma^2 (x^T x)^{-1})$ est un estimateur linéaire non biaisé de β.
- $\hat{\gamma}_n = (x^T x)(\hat{\beta}_n - \beta)(\hat{\beta}_n - \beta)^T$ est un estimateur non biaisé de $\sigma^2 I_n$.

En effet,

$$E[\hat{\gamma}_n] = E[(x^T x)(\hat{\beta}_n - \beta)(\hat{\beta}_n - \beta)^T] = (x^T x)\mathrm{cov}(\hat{\beta}_n) = (x^T x)\sigma^2 (x^T x)^{-1} = \sigma^2 I_n.$$

Soit $(E_{i,j})_{1 \leq i,j \leq n}$ la base canonique de $M_{n,n}(\mathbb{R})$. On a $\forall i, 1 \leq i \leq n$,

$$\hat{\gamma}_n . E_{i,i} = \hat{\alpha}_{i,n} E_{i,i} \text{ (on a de plus, } \hat{\alpha}_{i,n} = \mathrm{Tr}(\hat{\gamma}_n E_{i,i})).$$

$$E[\hat{\gamma}_n . E_{i,i}] = E[\hat{\alpha}_{i,n}] E_{i,i} = \sigma^2 I_n . E_{i,i} = \sigma^2 E_{i,i}.$$

Par suite $E[\hat{\alpha}_{i,n}] = \sigma^2$, donc $\hat{\alpha}_{i,n} = \mathrm{Tr}(\hat{\gamma}_n E_{i,i})$ est un estimateur non biaisé de σ^2.

Et ceci $\forall i, 1 \leq i \leq n$, on obtient donc n estimateurs non biaisés de σ^2.

8. On a $\text{cov}(\tilde{\beta}_A) = \sigma^2 AA^T$ est symétrique comme matrice de covariance, et $\text{cov}(\hat{\beta}_n) = \sigma^2(x^T x)^{-1}$ est symétrique car c'est une matrice de covariance (ou comme inverse d'une matrice symétrique réelle).

   Donc $R = \text{cov}(\tilde{\beta}_A) - \text{cov}(\hat{\beta}_n) = \sigma^2[AA^T - (x^T x)^{-1}]$, une matrice symétrique.

   Montrons que R est positive :

   On pose $K = A - (x^T x)^{-1} x^T$.

   $KK^T = [A-(x^T x)^{-1}x^T][A-(x^T x)^{-1}x^T]^T = AA^T - Ax(x^T x)^{-1} - (x^T x)^{-1}x^T A^T + (x^T x)^{-1}x^T x($

   Comme l'estimateur $\tilde{\beta}_A$ est non biaisé, donc on a $Ax = I$, la matrice identité. Par suite :

$$\sigma^2 KK^T = \sigma^2[AA^T - (x^T x)^{-1} - (x^T x)^{-1} + (x^T x)^{-1}] = R.$$

   R est donc positive, car pour tout vecteur colonne y on a $\langle y, Ry\rangle = y^T Ry = \sigma^2 y^T KK^T y = \sigma^2 \|K^T y\|^2 \geq 0$.

   Remarque : R est symétrique positive, donc en tant que formes quadratiques on a $\text{cov}(\tilde{\beta}_A) \geq \text{cov}(\hat{\beta}_n)$.

   Donc sous condition que x est injective, l'estimateur de moindre carrés est le meilleur estimateur linéaire non biaisé, et il présente une (co)variance minimale.