PC 8 : Estimation statistique

On corrigera les exercices (1), (4) et (6) en PC.

1 Estimation de paramètres

Exercice 1 (Loi de Poisson). On observe $\mathbf{x} = (x_1, \ldots, x_n)$ que l'on considère comme la réalisation du vecteur aléatoire $\mathbf{X} = (X_1, \ldots, X_n)$, où les X_i sont des variables aléatoires i.i.d. de la loi de Poisson $\mathcal{P}(\theta)$ de paramètre inconnu $\theta > 0$.

1. Calculer l'estimateur par la méthode des moments $\hat{\theta}_n^{\mathrm{MM}}$ de θ (s'il existe). On utilisera le moment d'ordre 1.
2. Calculer l'estimateur du maximum de vraisemblance $\hat{\theta}_n^{\mathrm{MV}}$ de θ (s'il existe).
3. Vérifier si l'estimateur $\hat{\theta}_n^{\mathrm{MV}}$ est convergent, et déterminer son risque quadratique moyen $\mathrm{RQM}_\theta(\hat{\theta}_n^{\mathrm{MV}})$.
4. Montrer que $\hat{\theta}_n^{\mathrm{MV}}$ est asymptotiquement normal et préciser sa vitesse de convergence.

Solution.

1. Soit $X \sim \mathcal{P}(\theta)$ avec $\theta > 0$. On a $\mathbb{E}_\theta[X] = \theta$. Selon la méthode des moments, on approche $\mathbb{E}_\theta[X]$ par la moyenne empirique $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. On obtient alors que l'estimateur par la méthode des moments (EMM) est donné par $\hat{\theta}^{\mathrm{MM}} = \bar{X}_n$. Mais attention, l'espace de paramètre Θ est $\mathbb{R}_+$ et $\bar{X}_n = 0$ se produit avec probabilité nonnulle. En effet,

$$\mathbb{P}_\theta(\bar{X}_n = 0) = \mathbb{P}_\theta(X_i = 0, i = 1, \ldots, n) = (\mathbb{P}_\theta(X_1 = 0))^n = \mathrm{e}^{-n\theta} > 0.$$

Donc, lorsque $\bar{x}_n = 0$, l'EMM n'est pas défini. Néanmoins, $\mathbb{P}_\theta(\bar{X}_n = 0) \to 0$ lorsque $n \to \infty$ pour tout $\theta > 0$. Donc, si n est suffisamment grand, le problème ne se pose pas.

2. La fonction de vraisemblance associée à $\mathbf{x} = (x_1, \ldots, x_n)$ est

$$\mathcal{L}(\mathbf{x}; \theta) = \prod_{i=1}^n \mathbb{P}_\theta(X_i = x_i) = \prod_{i=1}^n \frac{\theta^{x_i}}{x_i!}\mathrm{e}^{-\theta} = \mathrm{e}^{-n\theta}\theta^{\sum_{i=1}^n x_i}\frac{1}{\prod_{i=1}^n (x_i!)}.$$

On peut passer à la fonction de log-vraisemblance

$$\ell(\theta) = \log \mathcal{L}(\mathbf{x}; \theta) = -n\theta + \log\theta \sum_{i=1}^n x_i - \sum_{i=1}^n \log(x_i!).$$

Cette fonction est deux fois dérivable avec

$$\ell'(\theta) = -n + \frac{1}{\theta}\sum_{i=1}^n x_i, \qquad \ell''(\theta) = -\frac{1}{\theta^2}\sum_{i=1}^n x_i.$$

Comme $\ell''(\theta) \le 0$ pour tout $\theta > 0$, la fonction $\ell(\theta)$ est concave. Pour la maximiser il suffit alors de trouver un point critique :

$$\ell'(\theta) = 0 \iff -n + \frac{1}{\theta}\sum_{i=1}^n x_i = 0 \iff \theta = \bar{x}_n.$$

L'EMV est alors $\hat{\theta}_n^{\mathrm{MV}} = \hat{\theta}_n^{\mathrm{MM}} = \bar{X}_n$.
La même remarque que pour l'EMM s'applique : si $\bar{x}_n = 0$, alors l'EMV n'existe pas.

3. La LFGN implique la convergence de $\hat{\theta}_n^{\mathrm{MV}}$. Le risque quadratique moyen est donné par

$$\mathrm{RQM}_\theta(\hat{\theta}^{\mathrm{MV}}) = (\mathbb{E}_\theta[\hat{\theta}^{\mathrm{MV}}] - \theta)^2 + \mathrm{Var}(\hat{\theta}^{\mathrm{MV}}) = 0 + \frac{1}{n}\mathrm{Var}(X_1) = \frac{1}{n}.$$

4. Par le TCL, on a $\sqrt{n}(\hat{\theta}_n^{\mathrm{MV}} - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \theta)$ $(n \to \infty)$. Donc, l'EMV est bien asymptotiquement normal et sa vitesse de convergence est de $n^{-1/2}$.

Exercice 2 (Loi géométrique)**.** On observe $\mathbf{x} = (x_1, \ldots, x_n)$ que l'on considère comme la réalisation du vecteur aléatoire $\mathbf{X} = (X_1, \ldots, X_n)$, où les X_i sont des variables aléatoires i.i.d. de la loi géométrique $\mathrm{Geo}(\theta)$ de paramètre inconnu $0 < \theta < 1$.

1. Calculer l'estimateur par la méthode des moments $\hat{\theta}_n^{\mathrm{MM}}$ de θ (s'il existe). On utilisera le moment d'ordre 1.
2. Calculer l'estimateur du maximum de vraisemblance $\hat{\theta}_n^{\mathrm{MV}}$ de θ (s'il existe).
3. Vérifier si l'estimateur $\hat{\theta}_n^{\mathrm{MV}}$ est convergent.
4. Montrer que $\hat{\theta}_n^{\mathrm{MV}}$ est asymptotiquement normal et préciser sa vitesse de convergence.

Exercice 3 (Loi de Cauchy)**.** On observe $\mathbf{x} = (x_1, \ldots, x_n)$ que l'on considère comme la réalisation du vecteur aléatoire $\mathbf{X} = (X_1, \ldots, X_n)$, où les X_i sont des variables aléatoires i.i.d. de la loi de Cauchy avec un paramètre d'échelle dont la densité est donnée par

$$f_\theta(x) = \frac{\theta}{\pi(\theta^2 + x^2)},$$

pour $\theta > 0$ inconnu.

1. Calculer l'estimateur par la méthode des moments $\hat{\theta}_n^{\mathrm{MM}}$ de θ (s'il existe). On utilisera le moment d'ordre 1. Vérifier si l'estimateur est convergent et asymptotiquement normal, et calculer son risque quadratique moyen.
2. Calculer l'estimateur du maximum de vraisemblance $\hat{\theta}_n^{\mathrm{MV}}$ de θ (s'il existe). Que dire des propriétés de cet estimateur ?

2 Modèle statistique

Exercice 4 (Méthode de capture-recapture)**.** On souhaite estimer le nombre N de poissons vivant dans un bassin. Pour cela, on en pêche k que l'on marque. Ensuite, on pêche et on relâche successivement n poissons et l'on compte le nombre X de poissons marqués parmi les n. Donner l'estimateur du maximum de vraisemblance $\hat{N}$ de N en fonction de X, n et k. Montrer que $\hat{N}$ est convergent et donner une approximation de sa loi quand n est grand. Comment la variance de $\hat{N}$ dépend du choix de k ?

Solution. Notons Y_i la v.a. à valeurs dans $\{0, 1\}$ telle que $Y_i = 1$ si le i-me poisson pêché est marqué. Comme les poissons sont toute de suite relâchés, on peut supposer que les $Y_i, i = 1, \ldots, n$ sont des v.a. i.i.d. de loi Bernoulli de paramètre k/N avec k connu et N inconnu. Le nombre X de poissons marqués parmi les n est $X = \sum_{i=1}^n Y_i$ de loi binomiale $\mathrm{Bin}(n, k/N)$.
La fonction de vraisemblance est donnée par

$$\mathcal{L}(x; N) = \mathbb{P}(X = x) = \binom{n}{x} \left(\frac{k}{N}\right)^x \left(1 - \frac{k}{N}\right)^{n-x} = \binom{n}{x} k^x (N-k)^{n-x} \left(\frac{1}{N}\right)^n,$$

et sa fonction de log-vraisemblance par

$$\ell(N) = \log(\mathcal{L}(x; N)) = (n - x)\log(N - k) - n \log N + \text{constante}.$$

Ainsi

$$\ell'(N) \leq 0 \quad \Leftrightarrow \quad \frac{n-x}{N-k} - \frac{n}{N} \leq 0 \quad \Leftrightarrow \quad N \geq \frac{kn}{x}.$$

Donc l'EMV de N est donné par $\hat{N} = kn/X$.
Remarquons que $\hat{N} = k/\bar{Y}_n$. Par la LFGN et $h(y) = k/y$ fonction continue, on a $\hat{N} = h(\bar{Y}_n) \longrightarrow h(\mathbb{E}[Y_1]) = N$ *p.s.*. Donc, l'estimateur est convergent.
Par le TCL et la méthode delta (h est C^1), on a

$$\sqrt{n}(\hat{N} - N) = \sqrt{n}\left(h(\bar{Y}_n) - h\left(\frac{k}{N}\right)\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \mathrm{Var}(Y_1)\left[h'\left(\frac{k}{N}\right)\right]^2\right) = \mathcal{N}\left(0, N^2\left(\frac{N}{k} - 1\right)\right).$$

On en déduit que, pour n grand, $\hat{N} \approx \mathcal{N}\left(N, \frac{N^2}{n}\left(\frac{N}{k} - 1\right)\right)$. Donc, plus k est grand, plus la variance de $\hat{N}$ diminue.

Exercice 5 (Ampoules défaillantes)**.** Un statisticien observe pendant n jours le nombre d'ampoules défaillantes par jour à la sortie d'une chaîne de fabrication, noté $x_1, \dots, x_n$. Il considère que $x_1, \dots, x_n$ sont des réalisations i.i.d. d'une loi $\mathbb{P}$, et il souhaite estimer la probabilité p de n'avoir aucune ampoule défaillante par jour ($p = \mathbb{P}(X = 0)$).

1. Dans un premier temps, il compte le nombre de jours où aucune ampoule n'est défaillante, noté N_n, qui est alors le nombre de X_i, $i = 1, \dots, n$, égaux à 0. Il propose d'estimer la probabilité p par
$$\hat{p}_1 = \frac{1}{n} N_n.$$
Montrer que l'estimateur est convergent et sans biais. Calculer son risque quadratique, et donner sa loi limite.
2. Désormais le statisticien suppose que X_i suivent une loi de Poisson $\mathrm{Poi}(\lambda)$ de paramètre $\lambda > 0$ inconnu. Il propose comme estimateur de p
$$\hat{p}_2 = \mathrm{e}^{-\bar{X}_n},$$
où $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$.
   (i) Expliquer sa démarche. Montrer que $\hat{p}_2$ est convergent. Calculer sa variance et son biais. Déterminer des équivalents asymptotiques des quantités précédentes.
   (ii) Montrer que l'on peut choisir t_n tel que $\hat{p}_3 = \mathrm{e}^{-t_n \bar{X}_n}$ soit sans biais.
   (iii) Lequel de $\hat{p}_1$, $\hat{p}_2$ et $\hat{p}_3$ choisiriez-vous pour estimer $\mathrm{e}^{-\lambda}$?

3 Estimation statistique dans le modèle linéaire

Exercice 6 (Régression linéaire)**.** Soit $X \in M_{n,p}(\mathbb{R})$, $\beta \in \mathbb{R}^p$ avec $p < n$. On définit
$$Y = X\beta + \varepsilon,$$
où $\varepsilon \sim \mathcal{N}_n(0, \sigma^2 I_n)$. La matrice X est connue et on observe Y. L'objectif est d'estimer β. On supposera que X est de rang maximal, ce qui implique en particulier que $X'X$ est inversible.

1. Quelle est la loi du vecteur $Y = (Y_1, \dots, Y_n)'$? Les cordonnées Y_i sont-elles indépendantes et identiquement distribuées ?
2. La définition de l'estimateur du maximum de vraisemblance se généralise à des observations non i.i.d.. La fonction de vraisemblance est alors définie comme la densité du vecteur d'observations. Ainsi, l'estimateur du maximum de vraisemblance $\hat{\beta}$ de β est défini comme le point où la densité de Y, notée f_β, est maximale (on maximise $\beta \mapsto f_\beta(Y)$).
Montrer que $\hat{\beta}$ minimise $\beta \mapsto \|Y - X\beta\|_2^2$. En déduire que
$$\hat{\beta} = (X'X)^{-1}X'Y.$$
Donner la loi de $\hat{\beta}$.
3. On définit un estimateur de σ^2 par
$$\hat{\sigma}^2 = \frac{\|Y - X\hat{\beta}\|_2^2}{n-p}.$$
Montrer que $\hat{\beta}$ et $\hat{\sigma}^2$ sont indépendants.
4. Donner la loi de $\hat{\sigma}^2$ et en déduire qu'il s'agit d'un estimateur sans biais qui converge en probabilité vers σ^2.
Rappel du Théorème de Cochran : Soient $(V_i)_{1\le i\le n}$ une décomposition orthogonale en somme directe de $\mathbb{R}^n$ et $(\Pi_{V_i})_{1\le i\le n}$ les matrices de projection orthogonale associées. Pour un vecteur gaussien $Z \sim \mathcal{N}_n(0, \sigma^2 I_n)$, on a
— $\Pi_{V_1}Z, \cdots, \Pi_{V_k}Z$ *sont des vecteurs aléatoires indépendants de loi* $\mathcal{N}_n(0, \sigma^2\Pi_{V_i})$.
— *Pour tout* $1 \le i \le n$, $\|\Pi_{V_j}Z\|_2^2/\sigma^2$ *suit la loi* $\chi^2_{p_i}$ *où* p_i *est la dimension de* V_i.

Solution. 1. Comme ε est un vecteur gaussien de loi $\mathcal{N}_n(0, \sigma^2 I_n)$,
$$Y = X\beta + \varepsilon \sim \mathcal{N}_n(X\beta + \mathbb{E}[\varepsilon], \mathrm{Var}(\varepsilon)) = \mathcal{N}_n(X\beta, \sigma^2 I_n).$$
La matrice de covariance de Y étant diagonale, les Y_i sont bien indépendants. En revanche, les entrées de $X\beta$ ne sont pas identiques, donc les Y_i ne suivent pas la même loi.

2. La log-vraisemblance associée à $Y \sim \mathcal{N}_n(X\beta, \sigma^2 I_n)$ s'écrit

$$\ell(\beta) = \log(f_{\mathcal{N}_n(X\beta,\sigma^2 I_n)}(Y)) = -n\log(\sigma) - \frac{p}{2}\log(2\pi) - \frac{\|Y - X\beta\|_2^2}{2\sigma^2}.$$

Ainsi, maximiser la log-vraisemblance en β revient à minimiser $\|Y - X\beta\|_2^2$. On remarque que minimiser $\|Y - X\beta\|_2^2$ revient à projeter linéairement Y sur le sous-espace vectoriel engendré par les colonnes de X. Ainsi $\hat{\beta}$ vérifie

$$\langle Y - X\hat{\beta}, X\beta\rangle_{\mathbb{R}^n} = 0, \quad \forall \beta \in \mathbb{R}^p \iff \langle X'(Y - X\hat{\beta}), \beta\rangle_{\mathbb{R}^p} = 0, \quad \forall \beta \in \mathbb{R}^p.$$

On en déduit que $\hat{\beta}$ vérifie :

$$X'Y = X'X\hat{\beta},$$

et comme $X'X$ est supposé inversible,

$$\hat{\beta} = (X'X)^{-1}X'Y.$$

Le vecteur Y est gaussien, alors $\hat{\beta}$ l'est aussi, de moyenne

$$\mathbb{E}[\hat{\beta}] = (X'X)^{-1}X'\mathbb{E}[Y] = (X'X)^{-1}X'X\beta = \beta,$$

et variance

$$\mathrm{Var}(\hat{\beta}) = (X'X)^{-1}X'\mathrm{Var}(Y)((X'X)^{-1}X')' = \sigma^2(X'X)^{-1}X'((X'X)^{-1}X')' = \sigma^2(X'X)^{-1}.$$

Donc, $\hat{\beta} \sim \mathcal{N}_p\left(\beta, \sigma^2(X'X)^{-1}\right)$.

3. L'estimateur $\hat{\sigma}^2$ est une fonction du vecteur $Y - X\hat{\beta}$. Pour montrer que $\hat{\sigma}^2$ et $\hat{\beta}$ sont indépendants, il suffit de montrer que $Y - X\hat{\beta}$ et $\hat{\beta}$ sont indépendants. Or, $Y - X\hat{\beta}$ est un vecteur gaussien, et même le vecteur $(\hat{\beta}, Y - X\hat{\beta})'$ est vecteur gaussien comme transformation linéaire de ε. Plus précisément,

$$\begin{aligned} \hat{\beta} &= (X'X)^{-1}X'(X\beta + \varepsilon) = \beta + (X'X)^{-1}X'\varepsilon \\ Y - X\hat{\beta} &= X\beta + \varepsilon - X\beta - X(X'X)^{-1}X'\varepsilon = (I - X(X'X)^{-1}X')\varepsilon. \end{aligned}$$

Pour montrer l'indépendance, il suffit de montrer que la matrice de covariance de $\hat{\beta}$ et $Y - X\hat{\beta}$ est nulle :

$$\begin{aligned} \mathrm{Cov}(\hat{\beta}, Y - X\hat{\beta}) &= \mathbb{E}\left[(\hat{\beta} - \mathbb{E}[\hat{\beta}])(Y - X\hat{\beta} - \mathbb{E}[Y - X\hat{\beta}])'\right] \\ &= \mathbb{E}\left[((X'X)^{-1}X'\varepsilon)((I - X(X'X)^{-1}X')\varepsilon)'\right] \\ &= (X'X)^{-1}X'\mathbb{E}\left[\varepsilon\varepsilon'\right](I - X(X'X)^{-1}X')' \\ &= \sigma^2\left((X'X)^{-1}X' - (X'X)^{-1}X'X(X'X)^{-1}X'\right) \\ &= 0. \end{aligned}$$

4. Par construction, $X\hat{\beta}$ est la projection orthogonale de Y sur l'espace vectoriel engendré par les colonnes de X. On a la décomposition orthogonale :

$$\varepsilon = Y - X\beta = (X\hat{\beta} - X\beta) + (Y - X\hat{\beta}),$$

où $X\hat{\beta} - X\beta$ est la projection orthogonale de $Y - X\beta$ sur l'espace vectoriel engendré par X et $Y - X\hat{\beta}$ sur son supplémentaire orthogonal de dimension $n - p$. On en déduit, par le théorème de Cochran, que

$$\frac{\|Y - X\hat{\beta}\|_2^2}{\sigma^2} \sim \mathcal{X}_{n-p}^2.$$

Puisque $\mathbb{E}\left(\mathcal{X}_{n-p}^2\right) = n - p$ et $\mathrm{Var}(\mathcal{X}_{n-p}^2) = 2(n-p)$, on en déduit que $\mathbb{E}\left(\hat{\sigma}^2\right) = \sigma^2$ et $\mathrm{Var}(\hat{\sigma}^2) = \frac{2\sigma^4}{n-p}$. L'erreur quadratique moyenne tend vers 0 ce qui implique la convergence L^2 de l'estimateur et donc en probabilité.