Second Examen de statistiques (STF8)
Simon Coste

Jeudi 20 juin 2024 — de 9h30 à 12h30.

Exercice 1. $\sim$ Questions de cours.

1. Énoncer et démontrer la décomposition biais-variance d'un estimateur.
2. Dans un modèle exponentiel en dimension 1 dont la densité par rapport à la mesure de Lebesgue est $p_\theta(x) = e^{\theta T(x)}/Z(\theta)$, calculer la dérivée de $\ln Z(\theta)$.

Exercice 2. $\sim$ Soit (E_i) une suite de variables aléatoires iid de loi exponentielle de paramètre $c > 0$, et soit (X_i) une suite de variables aléatoires indépendantes de loi de Poisson, avec $X_i \sim \text{Poisson}(E_i)$. Autrement dit, X_i suit une loi de Poisson de paramètre E_i qui est lui-même aléatoire et dépend de c. L'objectif est d'estimer c.

1. Calculer $\mathbb{P}(X_i = n)$ pour tout n. Quelle est la loi de X_i ?
2. Proposer un estimateur convergent de c.
3. Proposer un intervalle de confiance asymptotique de c de niveau de risque α.

Exercice 3. $\sim$ On dispose de n variables $x_1, \dots, x_n \in \mathbb{R}$, et n variables à expliquer $y_1, \dots, y_n \in \mathbb{R}$. L'objectif est de trouver le meilleur polynôme possible de degré d qui relie les x_i aux y_i.

1. Soit $\mathscr{F}_d$ l'ensemble des polynômes à coefficients réels, de degré inférieur ou égal à d. L'objectif est de trouver $f \in \mathscr{F}_d$ tel que

$$f \in \arg\min_{g \in \mathscr{F}_d} \sum_{i=1}^{n} (y_i - g(x_i))^2.$$

   Exprimer ce problème de minimisation comme un problème de moindres carrés et exprimer la solution f en fonction des x_i et des y_i (on pourra faire des hypothèses raisonnables sur les x_i).

2. On suppose qu'il existe un polynôme $f_\star$ de degré au plus d tel que $y_i = f_\star(x_i) + \varepsilon_i$ où les ε_i sont des variables aléatoires indépendantes de loi $N(0, \sigma^2)$, avec σ^2 une variance inconnue. On note $\hat{f}$ l'estimateur trouvé à la question précédente. Quelle est la loi des coefficients de $\hat{f}$? Donner aussi l'estimateur de σ^2 et sa loi.
3. Tester l'hypothèse selon laquelle $f_\star$ est en réalité de degré strictement inférieur à d.
4. Maintenant, on suppose que les x_i sont de dimension 2. On souhaite faire une régression polynomiale de degré d comme ci-dessus des x_i vers les y_i.
   (a) Quelle est la dimension de cette nouvelle régression ?
   (b) ($\star$) Tester l'hypothèse selon laquelle les y_i dépendent seulement des $|x_i|^2$.

EXERCICE 4. $\sim$ Calculer l'information de Fisher d'une loi $\mathscr{N}(\mu, \sigma^2)$.

EXERCICE 5. $\sim$ Kylian Mbappé, recruté en août 2017 au Paris-Saint-Germain, a joué son dernier match pour ce club en mai 2024. Sur cette période, le PSG a joué 237 matches, mais Mbappé n'était pas toujours présent :

	Mbappé présent	Mbappé absent
Victoires	144	29
Défaites ou nuls	51	13

Mbappé a-t-il été statistiquement utile au PSG ?

Solution de l'exercice 2

En conditionnant, on voit que $\mathbb{P}(X=n) = \int_0^\infty \mathbb{P}(X=n|E=x)dx = \int_0^\infty \frac{x^n e^{-x}}{n!} \times ce^{-cx}dx$. On reconnaît la fonction Gamma, et on trouve

$$\mathbb{P}(X=n) = \frac{c}{(1+c)^{n+1}} = q^n p$$

avec $p = c/(1+c)$. On reconnaît une loi géométrique de paramètre p, sur $\mathbb{N}$. Par conséquent, on a $\mathbb{E}[X] = 1/p = (1+c)/c$. Par la loi forte des grands nombres, $\bar{X}_n$ converge donc vers $1/p$ presque sûrement, et donc $\hat{c} = 1/(\bar{X}_n - 1)$ converge presque sûrement vers c. Par le théorème central limite, $\sqrt{n}(\bar{X}_n - 1/p)$ converge en loi vers $\mathcal{N}(0, \sigma^2)$ où $\sigma^2 = \mathrm{Var}(X) = q/p^2 = (1/(1+c))((1+c)/c)^2 = (1+c)/c^2$. On peut estimer asymptotiquement σ^2 par $(1+\hat{c})/\hat{c}$, et le lemme de Slutsky nous dit que

$$\frac{\sqrt{n}(\hat{c}-c)}{\sqrt{\frac{1+\hat{c}}{\hat{c}}}} \to \mathcal{N}(0,1).$$

Si q_α est le quantile symétrique d'ordre $1-\alpha$ de la loi normale centrée réduite, alors un intervalle de confiance asymptotique de niveau de confiance $1-\alpha$ pour c est donné par

$$\left[\hat{c} \pm q_\alpha \sqrt{\frac{\hat{c}}{n(1+\hat{c})}}\right].$$

Solution de l'exercice 3

L'exercice entier est trivial une fois que l'on a compris que la régression polynomiale est linéaire. En effet, si l'on note $X = (1, x, x^2, \ldots, x^d)$, alors le problème revient à résoudre $\hat{\beta} = \arg\min_\beta \|Y - X\beta\|^2$. Le vecteur colonne $\hat{\beta}$ contient les $d+1$ coefficients de $\hat{f}$, puisque si l'on note $X_i = (1, x_i, x_i^2, \ldots, x_i^d)$, alors

$$X_i\hat{\beta} = \sum_{k=0}^{d} \beta_k x_i^k = \hat{f}(x_i).$$

Par conséquent, la formule des MCO nous donne

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

L'entrée k,ℓ de la matrice $X^\top X$ est $\sum_{i=1}^n x_i^{k+\ell}$, et l'entrée k du vecteur $X^\top Y$ est $\sum_{i=1}^n x_i^k y_i$. Bien sûr, la formule n'est valable que si la matrice $X^\top X$ est effectivement inversible.

Lorsqu'on suppose que le modèle sous-jacent est gaussien (à savoir $Y_i \sim \mathcal{N}(f_\star(x_i), \sigma^2)$), alors $\hat{\beta}$ est gaussien, et sa loi est donnée par

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

Tester si $f_\star$ est de degré strictement inférieur à d, c'est tester si β_d est nul. Il ne s'agit donc de rien d'autre que d'un test de significativité de Student, qui se formule sous la forme

$$\left\{\frac{|\hat{\beta}_d|}{\hat{\sigma}_d} > t\right\}$$

où t est le quantile symétrique d'ordre $1-\alpha$ d'une loi de Student de paramètre $n-d-1$ (il y a bien $d+1$ paramètres à étudier).

En dimension supérieure (ici, 2), l'espace des polynômes de degré d est

$$\binom{2+d}{d}.$$

Tout polynôme peut s'écrire sous la forme

$$\sum_{k+\ell\leqslant d} c_{k,\ell} x(i)^k x(i)^\ell$$

Si le polynôme $f(a,b)$ ne dépend que de a^2+b^2, alors cela veut dire qu'il y a des coefficients h_k tels que

$$f(a,b) = \sum_{k\leqslant d/2} h_k (a^2+b^2)^k = \sum_{k\leqslant d/2} h_k \sum_{\ell=0}^{k} \binom{k}{\ell} a^{2k-2\ell} b^{2\ell}.$$

Autrement dit les seuls monômes qui contribuent à la somme sont ceux qui sont de la forme

$$a^{2k-2\ell} b^{2\ell}$$

pour $0 \leqslant \ell \leqslant k \leqslant \lfloor d/2 \rfloor$. Ces monômes sont au nombre de $(D+1)(D+2)/2$ avec $D = \lfloor d/2 \rfloor$. On teste donc $\binom{d+2}{d} - (D+1)(D+2)/2$ contraintes linéaires (on teste si tous les autres coefficients sont nuls). Noter que lorsque d est grand cela revient à tester environ $7d^2/8$ contraintes.

Solution de l'exercice 4

La densité de la loi $\mathscr{N}(\mu,\sigma^2)$ est donnée par

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

L'information de Fisher est donnée par la matrice

$$\mathbb{E}\left[\nabla_{\mu,\sigma^2} \ln p(x) \nabla_{\mu,\sigma^2} \ln p(x)^\top\right].$$

On a

$$\nabla_{\mu,\sigma^2} p(x) = \begin{pmatrix} \frac{x-\mu}{\sigma^2} p(x) \\ \frac{(x-\mu)^2}{2\sigma^4} p(x) - \frac{1}{2\sigma^2} p(x) \end{pmatrix} = \begin{pmatrix} \frac{x-\mu}{\sigma^2} \\ \frac{(x-\mu)^2}{2\sigma^4} - \frac{1}{2\sigma^2} \end{pmatrix} p(x).$$

Par conséquent les quatre moments dont il faut calculer l'espérance sont

$$\left(\frac{x-\mu}{\sigma^2}\right)^2,$$

$$\left(\frac{x-\mu}{\sigma^2}\right)\left(\frac{(x-\mu)^2}{2\sigma^4} - \frac{1}{2\sigma^2}\right),$$

$$\left(\frac{(x-\mu)^2}{2\sigma^4} - \frac{1}{2\sigma^2}\right)^2$$

Or, lorsque $X \sim \mathscr{N}(\mu, \sigma^2)$, la variable aléatoire $Y = (X - \mu)/\sigma$ n'est autre qu'une $\mathscr{N}(0, 1)$. L'espérance des trois termes ci-dessus est donc

$$\mathbb{E}[Y^2/\sigma^2] = \frac{1}{\sigma^2}$$

$$\mathbb{E}[Y/\sigma \times (Y^2/\sigma^2 - 1/2\sigma^2)] = 0$$

$$\mathbb{E}[(Y^2/\sigma^2 - 1/2\sigma^2)^2] = \frac{1}{2\sigma^4}.$$

Par conséquent l'information de Fisher de ce modèle est

$$\begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 1/2\sigma^4 \end{pmatrix}. \tag{1}$$

Solution de l'exercice 5

C'est un test du chi-deux d'indépendance. Mes calculs donnent une statistique de Pearson proche de 0.2, ce qui est très faible ; la loi limite donnée par le théorème du cours est une loi du chi-deux à 1 degré de liberté. Il n'y a pas de raison forte pour rejeter l'hypothèse nulle d'indépendance (pour rappel le quantile à 90% est proche de 2.71).