TD 6 : Modèle linéaire gaussien

Exercice 1 (Modèle gaussien ordinaire) On considère le modèle

$$Y_i = m + \sigma\varepsilon_i, \quad 1 \leq i \leq n$$

où les v.a. ε_i sont i.i.d. de loi commune $\mathcal{N}(0,1)$, pour des paramètres $m \in \mathbb{R}$ et $\sigma > 0$. On note $\overline{Y}_n = n^{-1}\sum_{i=1}^n Y_i$.

1. On suppose que σ est connu.
   (a) Déterminer un intervalle de confiance symétrique pour m de niveau $1-\alpha$ ($\alpha \in]0,1[$).
   (b) Pour $\sigma = 3$, combien d'observations doit-on avoir pour que la longueur de l'intervalle de confiance de niveau 95% soit inférieure à 2 ? Donner la forme de cet intervalle au niveau 95% pour $\sigma = 3$, $n = 25$ et $\overline{y}_{25} = \overline{Y}_{25}(\omega) = 20$. Indication : $\Phi^{-1}(0.975) \approx 2$.
   (c) Proposer un test de niveau α pour l'hypothèse $H_0 : m = m_0$ contre $H_1 : m \neq m_0$. Pour $\sigma = 3$, $n = 25$, $\overline{y}_{25} = \overline{Y}_{25}(\omega) = 20$ et $m_0 = 18.9$, quelle est la p-valeur de ce test ? Peut-on accepter l'hypothèse H_0 aux niveaux 1%, 5% et 10% ? Indication : $\Phi\left(\frac{5.5}{3}\right) \simeq \Phi(1.83) \simeq 0.97$.
2. On ne suppose plus que σ est connu. On pose $\widehat{\sigma}_n = \sqrt{\frac{1}{n-1}\sum_{i=1}^n (Y_i - \overline{Y}_n)^2}$.
   (a) Écrire le modèle de régression associé aux hypothèses énoncées plus haut et donner l'estimateur des moindres carrés.
   (b) Énoncer le théorème de Cochran dans ce cas.
   (c) Montrer que $\widehat{\sigma}_n^2$ est un estimateur sans biais et consistant de σ^2.
   (d) Tester l'hypothèse $H_0 : \sigma^2 = 3$ contre $H_1 : \sigma^2 \neq 3$ au niveau α.
   (e) Donner la loi exacte de $\sqrt{n}\frac{\overline{Y}_n - m}{\widehat{\sigma}_n}$.
   (f) Déterminer un intervalle de confiance de niveau $1-\alpha$ pour m. En déduire un test pour l'hypothèse $H_0 : m = m_0$ contre $H_1 : m \neq m_0$ au niveau α.
   (g) Tester maintenant $H_0 : m \geq m_0$ contre $H_1 : m < m_0$ au niveau α. Calculer la p-valeur lorsque $m_0 = 12.5$, $n = 25$, $\overline{y}_{25} = \overline{Y}_{25}(\omega) = 12$ et $\widehat{\sigma}_n^2(\omega) = 1.69$? Peut-on accepter l'hypothèse H_0 au niveau 5% ? Indication : $F_{\mathcal{T}(24)}(-1.92) \simeq 0.03$.

Exercice 2 (Régression linéaire simple) On considère le modèle suivant

$$Y_i = a + bt_i + \varepsilon_i, \quad 1 \leq i \leq n,$$

où les variables aléatoires ε_i sont i.i.d. $\mathcal{N}(0,\sigma^2)$, les réels $(t_i)_{1\leq i\leq n}$ sont connus et a, b, σ^2 sont trois paramètres réels inconnus. On suppose que $\sum_{i=1}^n t_i = 0$ et on note

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^n Y_i, \quad v_t = \frac{1}{n}\sum_{i=1}^n t_i^2 > 0, \quad v_Y = \frac{1}{n}\sum_{i=1}^n Y_i^2 - \overline{Y}^2 \quad \text{et} \quad \rho = \frac{1}{n}\sum_{i=1}^n Y_i t_i.$$

1. Préciser les conditions d'identifiabilité du modèle.
2. Calculer les estimateurs des moindres carrés $\widehat{a}$, $\widehat{b}$ et $\widehat{\sigma}^2$ de a, b et σ^2 en fonction de $\overline{Y}$, v_t, v_Y et ρ. Quelle est leur loi jointe ?
3. Soit $\alpha \in]0,1[$. Donner un intervalle de confiance de niveau $1-\alpha$ pour chacun des paramètres a et b. En déduire un rectangle de confiance de niveau 95% pour le paramètre (a,b).
4. Construire une ellipse de confiance $\mathcal{E}$ de niveau 95% pour le paramètre (a,b).
5. Donner un intervalle de confiance pour $5a-8b$, de niveau 95%, lorsque $n=18$.
6. Tester l'hypothèse H_0 : "$a=b$" contre H_1 : "$a \neq b$" au niveau 1% lorsque $n=22$.
7. On appelle $\Theta_0 = \big\{(a,b) \in \mathbb{R}^2 : a=b\big\}$. Montrer que le test consistant à rejeter H_0 si $\Theta_0 \cap \mathcal{E} = \emptyset$ est de niveau $\alpha \in]0,1[$ lorsque que $\mathcal{E}$ est une ellipse de confiance de niveau $1-\alpha$. Comparer ce test à celui de la question précédente.
8. On considère une nouvelle donnée $t'=2$. Quelle est la valeur prédite $\widehat{Y}^{(p)}$ pour la réponse associée Y ? Donner un intervalle de prédiction de niveau 98% pour Y.

Exercice 3 (Régression linéaire multiple) On considère le modèle linéaire

$$Y_i = a + bW_i + cZ_i + \varepsilon_i, \quad i=1,\dots,n$$

où les W_i et les Z_i sont des réels fixés, les ε_i sont des variables aléatoires réelles indépendantes de même loi $\mathcal{N}(0,\sigma^2)$, et a, b, c, σ^2 sont des paramètres réels inconnus. On note W, Y, Z les vecteurs colonnes de $\mathbb{R}^n$ de coordonnées respectives $(W_i), (Y_i), (Z_i)$, et $\overline{W} = \frac{1}{n}\sum_{i=1}^n W_i$, $\langle W, Z\rangle = \sum_{i=1}^n W_i Z_i$, $\|W\|^2 = \langle W, W\rangle$ et ainsi de suite pour les autres variables. On suppose dans tout ce qui suit que $\overline{W} = \overline{Z} = 0$,

$$\|W\| = \|Z\| = r > 0, \quad \langle W, Z\rangle = r^2 \sin\theta, \quad \theta \in]-\pi/2, \pi/2[.$$

1. Calculer les estimateurs des moindres carrés $\widehat{a}, \widehat{b}, \widehat{c}$, de a, b, c en fonction de r, θ et de produits scalaires du type ci-dessus. Déterminer la loi jointe de $(\widehat{a}, \widehat{b}, \widehat{c}, \widehat{\sigma}^2)$, où $\widehat{\sigma}^2$ est l'estimateur sans biais usuel de σ^2.
2. Donner un intervalle de confiance pour c de niveau $1-\alpha$ ($\alpha \in]0,1[$). Quelle est l'espérance du carré de sa longueur ? Comment varie-t-elle en fonction de θ ?
3. Donner un parallélépipède rectangle de confiance pour (a,b,c) de niveau 97% lorsque $n=27$.
4. Construire un ellipsoïde de confiance pour (a,b,c) de niveau 97% pour $n=27$.

Exercice 4 ($\star$) (Régression linéaire multiple avec orthogonalité) On considère $G = (g_i^j)_{1\leq i\leq n,\, 1\leq j\leq p}$, matrice de taille $n \times p$ dont les p colonnes sont des vecteurs de $\mathbb{R}^n$ non nuls et deux à deux orthogonaux. On note $\delta_j = \left[\sum_{i=1}^n (g_i^j)^2\right]^{1/2}$, $1 \leq j \leq p$, les normes euclidiennes des p colonnes de G et $\delta^{-2} = \sum_{j=1}^p \delta_j^{-2}$. On suppose $n > p \geq 2$. On observe un vecteur aléatoire Z dans $\mathbb{R}^n$ donné par

$$Z = G\gamma + \varepsilon, \quad \gamma \in \mathbb{R}^p, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I_n),$$

les paramètres γ et σ^2 étant inconnus.

1. Expliciter les estimateurs des moindres carrés $\widehat{\gamma}$ et $\widehat{\sigma}^2$ de γ et σ^2. Déterminer leur loi jointe.
2. Si $n=25$ et $p=5$, donner un intervalle de confiance pour σ^2 de niveau 95%.
3. Que peut-on dire des composantes $\widehat{\gamma}_1, \dots, \widehat{\gamma}_p$ de $\widehat{\gamma}$? Quelle est la loi de $\sum_{j=1}^p \widehat{\gamma}_j$?

4. On suppose que $n = 32$ et $p = 7$. Déduire de la question précédente un intervalle de confiance pour $\sum_{j=1}^{p} \gamma_j$ au niveau 99% en fonction de δ, puis un test de H_0 : "$\sum_{j=1}^{p} \gamma_j = 0$" au niveau 1%.

Exercice 5 ($\star$) (Regression non-paramétrique) Pour $1 \leq i \leq n$, on observe les variables aléatoires $Y_i = f(i/n) + \varepsilon_i$, où les ε_i sont i.i.d. $\mathcal{N}(0, \sigma^2)$ (σ connu) et $f : [0,1] \to \mathbb{R}$ est une fonction inconnue qui est le paramètre d'intérêt.

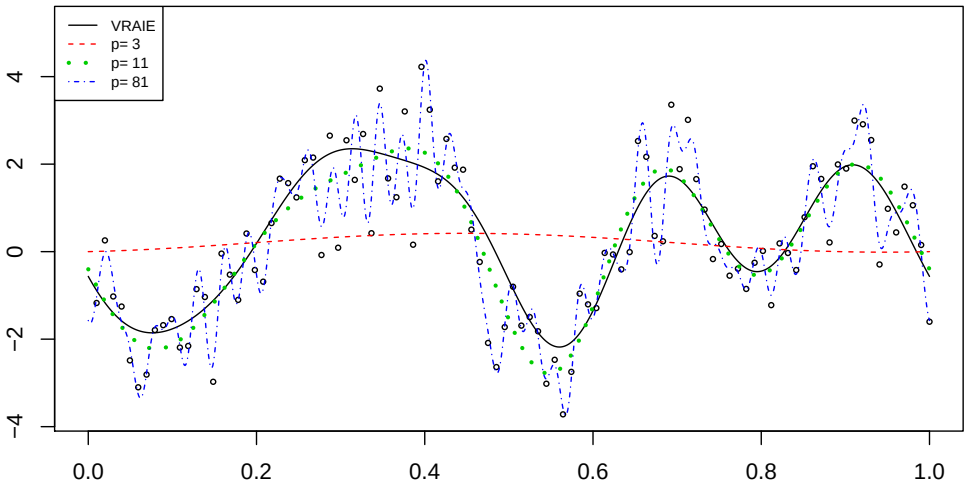
1. Réécriture sous forme de modèle linéaire.
   (a) Quel est la difficulté particulière de ce modèle statistique ?
   Pour simplifier le problème, on suppose que f s'écrit comme le début d'un développement en série de Fourier : pour tout $x \in [0,1]$,

$$f(x) = a_0 + \sum_{k=1}^{K} \big(a_k \cos(2\pi k x) + b_k \sin(2\pi k x)\big),$$

   pour a_0, a_k, b_k, $1 \leq k \leq K$, des réels inconnus.
   (b) Sous cette hypothèse, écrire le modèle comme un modèle linéaire gaussien $Y = X\beta + \varepsilon$ et préciser X, β et p.
   (c) On suppose à présent $2K + 1 \leq n$. Vérifier que le modèle est identifiable et donner l'estimateur des moindres carrés $\widehat{\beta}$ de β. En déduire un estimateur $\widehat{\mu}$ de $\mu = (f(i/n))_{1 \leq i \leq n}$. Proposer finalement un estimateur $\widehat{f}$ de la *fonction* f. On rappelle les formules suivantes :

$$\begin{cases} \cos(A)\sin(B) = & \Im\left(e^{\mathbf{i}(A+B)} - e^{\mathbf{i}(A-B)}\right)/2 \\ \cos(A)\cos(B) = & \Re\left(e^{\mathbf{i}(A-B)} + e^{\mathbf{i}(A+B)}\right)/2 \\ \sin(A)\sin(B) = & \Re\left(e^{\mathbf{i}(A-B)} - e^{\mathbf{i}(A+B)}\right)/2 \end{cases}$$

2. Overfitting et choix de modèle.
   (a) Calculer la somme des carrés résiduelle normalisée $r_n = n^{-1}\mathbb{E}(\|Y - X\widehat{\beta}\|^2)$. Que se passe-t-il lorsque p est fixe et n tend vers l'infini ?
   (b) On suppose maintenant $p = n$. Donner la valeur de r_n. Que dire de la fonction $\widehat{f}$ aux points i/n, $1 \leq i \leq n$?
   (c) Commenter la figure ci-dessous, pour laquelle $\beta_j^\star \neq 0$ pour $j \leq 11$, $\beta_j^\star = 0$ pour $j > 11$ et $n = 101$. Quels phénomènes observe-t-on ? Comment choisir un modèle bien ajusté ?



Exercice 6 (⋆) ("Réciproque" du Théorème de Cochran)

1. Soit $X = (X_1, \ldots, X_n)^t \sim \mathcal{N}(0, I)$. On pose $\overline{X}_n = n^{-1}\sum_{i=1}^n X_i$ et $s_n^2 = n^{-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$. Montrer que les variables aléatoires $\overline{X}_n$ et s_n^2 sont indépendantes et déterminer leurs lois.
2. Soit $X_1, \ldots, X_n$ des v.a. i.i.d. dont la loi commune est supposée inconnue. On suppose que $\overline{X}_n = n^{-1}\sum_{i=1}^n X_i$ et $s_n^2 = n^{-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$ sont indépendantes et que $\mathbb{E}(X_1^2)$ est finie. On note $m = \mathbb{E}(X_1)$ et $\sigma^2 = \mathrm{Var}(X_1)$. On cherche à montrer que X_i suit la loi $\mathcal{N}(m, \sigma^2)$, pour tout $i = 1, \ldots, n$, ce qui peut être vu comme une réciproque du Théorème de Cochran. Notons ϕ la fonction caractéristique de X_1 que l'on suppose non nulle.
   (a) Calculer $\mathbb{E}(ns_n^2)$ en fonction de σ^2 et montrer que, pour tout réel t,
   $$\mathbb{E}(s_n^2 e^{itn\overline{X}_n}) = \phi^n(t)\mathbb{E}(s_n^2).$$
   (b) En développant s_n^2, trouver une autre expression de $\mathbb{E}(s_n^2 e^{itn\overline{X}_n})$ en fonction de $\phi'(t)$ et $\phi''(t)$. En déduire que ϕ est solution de
   $$\frac{\phi''}{\phi} - \left(\frac{\phi'}{\phi}\right)^2 = -\sigma^2, \quad \phi(0) = 1, \phi'(0) = im.$$
   (c) En déduire que $X_1 \sim \mathcal{N}(m, \sigma^2)$. Indication : $(\ln \phi)'' = \frac{\phi''}{\phi} - \left(\frac{\phi'}{\phi}\right)^2$.