Variables aléatoires - Lois de probabilité

Exercice 1 (Modèle Statistique)

1) On note d'abord que

$$\begin{aligned}\hat{\beta} &= (X^\top X)^{-1}X^\top Y = (X^\top X)^{-1}X^\top(X\beta+\epsilon) = (X^\top X)^{-1}X^\top X\beta + (X^\top X)^{-1}X^\top\epsilon \\ &= \beta + (X^\top X)^{-1}X^\top\epsilon.\end{aligned} \tag{1}$$

Par définition de l'OLS et d'après l'équation ci-dessus, et puisque ϵ est de moyenne nulle,

$$\mathbb{E}[\hat{\beta}] = \beta + (X^\top X)^{-1}X^\top\,\mathbb{E}[\epsilon] = \beta.$$

De plus, puisque $\mathrm{Var}[\epsilon] = \Sigma$,

$$\mathrm{Var}[\hat{\beta}] = (X^\top X)^{-1}X^\top\,\mathrm{Var}[\epsilon]\left((X^\top X)^{-1}X^\top\right)^\top = (X^\top X)^{-1}X^\top\Sigma X(X^\top X)^{-1}.$$

2) Puisque Σ est définie positive, elle est diagonalisable dans une base orthonormale. Soit $P^\top DP$ la décomposition en valeurs propres de Σ, avec P orthonormale ($P^{-1} = P^\top$) et $D = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ la matrice diagonale contenant les valeurs propres de Σ :

$$\Sigma = P^\top DP.$$

Soit $\tilde{D} = \mathrm{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n})$. On a $\Sigma = P^\top\tilde{D}\tilde{D}P$. Soit $\Omega = P^\top\tilde{D}$, alors $\Omega^\top = (P^\top\tilde{D})^\top = \tilde{D}P$ et

$$\Sigma = \Omega\Omega^\top.$$

De plus, $\Omega = P^\top\tilde{D}$ est inversible et

$$\Omega^{-1} = \tilde{D}^{-1}(P^\top)^{-1} = \tilde{D}^{-1}P = \mathrm{diag}(1/\sqrt{\lambda_1}, \ldots, 1/\sqrt{\lambda_n})P.$$

(Vérifier par calcul direct)

3) Cela revient à montrer que l'endomorphisme associé à $\Omega^{-1}X$ est injectif.Soit $u \in \mathbb{R}^p$ tel que $\Omega^{-1}Xu = 0$. Alors en multipliant par Ω on a $Xu = 0$. Or, comme X est de rang complet, on obtient $u = 0$.

4) En multipliant tous les termes du modèle de régression initial par Ω^{-1} on obtient

$$\Omega^{-1}Y = \Omega^{-1}X\beta + \Omega^{-1}\epsilon,$$

ce qui équivaut à

$$Y^*X^*\beta + \epsilon^*.$$

D'après la question 2,

$$\begin{aligned}\mathrm{Var}[\epsilon^*] &= \Omega^{-1}\,\mathrm{Var}[\epsilon]\Omega^{-1} = \Omega^{-1}\Sigma(\Omega^{-1})^\top \\ &= \Omega^{-1}\Omega\Omega^\top(\Omega^{-1})^\top = \Omega^\top(\Omega^\top)^{-1} = I_n.\end{aligned} \tag{2}$$

De plus, $\mathbb{E}[\epsilon^*] = \Omega^{-1}\,\mathbb{E}[\epsilon] = 0$. On trouve donc le modèle de régression linéaire dans lequel les erreurs sont centrées et homoscédastiques.

5) Dans ce nouveau modèle, on peut calculer l'estimateur des moindres carrés ordinaires :

$$\begin{aligned}\hat{\beta^*} &= \left((X^*)^\top X^*\right)^{-1}(X^*)^\top Y^* \\ &= \left((\Omega^{-1}X)^\top\Omega^{-1}X\right)^{-1}(\Omega{-}1X)^\top\Omega^{-1}Y \\ &= \left(X^\top(\Omega^{-1})^\top\Omega^{-1}X\right)^{-1}X^\top(\Omega^{-1})^\top\Omega^{-1}Y \\ &= \left(X^\top(\Omega\Omega^\top)^{-1}X\right)^{-1}X^\top(\Omega\Omega^\top)^{-1}Y \\ &= (X^\top\Sigma^{-1}X)^{-1}X^\top\Sigma^{-1}Y \\ &:= \hat{\beta}_G.\end{aligned} \tag{3}$$

6) $\hat{\beta}_G$ est l'estimateur des moindres carrés ordinaires pour le nouveau modèle. Il satisfait donc $\mathbb{E}[\hat{\beta}_G] = \beta$. Par ailleurs, on a

$$\mathrm{Var}[\hat{\beta}_G] = \hat{\beta^*} = \left((X^*)^\top X^*\right)^{-1} = \left(X^\top\Sigma^{-1}X\right)^{-1}.$$

7) Notons d'abord que tout estimateur linéaire T dans les observations Y est linéaire dans les observations Y^* :

$$T = AY = A\Omega Y^* = A^*Y^* = T^*.$$

L'estimateur $\hat{\beta}_G$ est linéaire (dans les observations Y et Y^*) et sans biais. Par Gauss-Markov, il vient que $\hat{\beta}_G$ est optimal parmi tous les estimateurs linéaires sans biais T^* dans le nouveau modèle car les hypothèses du théorème sont satisfaites. Pour tout $u \in \mathbb{R}^p$:

$$u^\top\,\mathrm{Var}[T^*]u \geq u^\top\,\mathrm{Var}[\hat{\beta}_G]u.$$

8) En particulier pour l'estimateur $\hat{\beta}$, on obtient que

$$u^\top\,\mathrm{Var}[\hat{\beta}]u \geq u^\top\,\mathrm{Var}[\hat{\beta}_G]u.$$

Exercice 2 (Régression ridge)

1) Notons ma fonction $\Phi(\beta) = \|Y - X\beta\|^2 + \lambda\|\beta\|^2$. Cette fonction objectif est différentiable et convexe, on peut donc écrire les conditions d'optimalité du premier ordre :

$$\nabla\Phi(\beta) = -2X^\top(Y - X\beta) + 2\lambda\beta.$$

L'estimateur ridge satisfait $\nabla\Phi(\hat{\beta}_\lambda^R) = 0$. On en déduit la relation voulue :

$$(X^\top X + \lambda I_p)\hat{\beta}_\lambda^R = X^\top Y.$$

2) Une matrice symétrique A est inversible si et seulement si $\forall v \in \mathbb{R}^p \setminus \{0\}$, $v^\top A v \neq 0$. Ici, pour $v \in \mathbb{R}^p$ non nul,

$$v^\top(X^\top X + \lambda I_p)v = \|Xv\|^2 + \lambda\|v\|^2.$$

Puisque $\lambda > 0$, $\|Xv\|^2 + \lambda\|v\|^2 > 0$ ce qui prouve que $X^\top X + \lambda I_p$ est définie positive. En conséquence, $(X^\top X + \lambda I_p)^{-1}$ est bien définie et on obtient

$$\hat{\beta}_\lambda^R = (X^\top X + \lambda I_p)^{-1} X^\top Y.$$

3) $X^\top X$ admet une décomposition spectrale $X^\top X \sum_{j=1}^p \mu_j \theta_j \theta_j^\top$, où les μ_j sont les valeurs propres et θ_j sont les vecteurs propres correspondants. Notons que $(\theta_1, \ldots, \theta_p)$ est une base orthonormale. On obtient donc

$$X^\top X + \lambda I_p = \sum_{j=1}^p (\mu_j + \lambda)\theta_j \theta_j^\top, \text{ et}$$

$$(X^\top X + \lambda I_p)^{-1} = \sum_{j=1}^p \frac{1}{\mu_j + \lambda}\theta_j \theta_j^\top.$$

Par construction de la décomposition en valeurs singulières $X = \sum_{j=1}^r \sqrt{\mu_j}\tilde{\theta}_j \theta_j^\top$, où r est le rang de la matrice X et $(\tilde{\theta}_j)_{j=1}^r$ est une famille orthonormale de $\mathbb{R}^n$. Donc

$$\begin{aligned}(X^\top X + \lambda I_p)^{-1} X^\top &= \sum_{j=1}^r \frac{\mu_j^{1/2}}{\mu_j + \lambda}\theta_j \tilde{\theta}_j^\top \\ &= \sum_{j=1}^r \frac{1}{\mu_j + \lambda}\theta_j \theta_j^\top X^\top.\end{aligned} \tag{4}$$

Donc $(X^\top X + \lambda I_p)^{-1} \to \sum_{j=1}^r \frac{1}{\mu_j}\theta_j\theta_j^\top$ lorsque $\lambda \to 0$; on définit $(X^\top X)^* = \sum_{j=1}^r \frac{1}{\mu_j}\theta_j\theta_j^\top$ la pseudo-inverse de $X^\top X$. Finalement, $\hat{\beta}_\lambda^R \to (X^\top X)^* X^\top Y$ lorsque $\lambda \to 0$ et $\hat{\beta}_\lambda^R \to 0$ lorsque $\lambda \to +\infty$.

4)

$$\begin{aligned}\text{Bias}(\hat{\beta}_\lambda^R) &= \mathbb{E}[\hat{\beta}_\lambda^R] - \beta \\ &= \left((X^\top X + \lambda I_p)^{-1} X^\top X - I_p\right)\beta.\end{aligned} \tag{5}$$

5)
$$
\begin{aligned}
\operatorname{Cov}(\hat{\beta}^R_\lambda) &= A \operatorname{Cov}(Y) A^\top \\
&= \sigma^2 (X^\top X + \lambda I_p)^{-1} X^\top X (X^\top X + \lambda I_p)^{-1}.
\end{aligned} \tag{6}
$$

6) Supposons que X est de rang plein (i.e. $X^\top X$ est de rang p).

(a) L'estimateur des moindres carrés classique est donné par :

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

avec $\operatorname{Cov}(\hat{\beta}) = \sigma^2 (X^\top X)^{-1}$.

(b) La formule découle d'un calcul simple à l'aide des expressions $\operatorname{Cov}(\hat{\beta}) = \sigma^2 (X^\top X)^{-1}$ et $\operatorname{Cov}(\hat{\beta}^R_\lambda) = \sigma^2 (X^\top X + \lambda I_p)^{-1} X^\top X (X^\top X + \lambda I_p)^{-1}$.

(c) On va prouver que pour tout $x \in \mathbb{R}^p$,

$$x^\top \operatorname{Cov}(\hat{\beta}) x \geq x^\top \operatorname{Cov}(\hat{\beta}^R_\lambda) x.$$

Puisque $(\theta_j)_{1 \leq j \leq p}$ est une base orthonormale de $\mathbb{R}^p$, il suffit de montrer que l'inégalité ci-dessus est vraie pour tout θ_j. Soit $\lambda_j > 0$ une valeur propre de P et θ_j un vecteur propre associé ; on a $P\theta_j = \lambda \theta_j$.

$$(P + I_p)^{-1} \theta_j = \frac{1}{\lambda_j + 1} \theta_j,\ P^{-1} \theta_j = \frac{1}{\lambda_j} \theta_j$$

On obtient donc

$$\theta_j^\top \left(\operatorname{Cov}(\hat{\beta}) - \operatorname{Cov}(\hat{\beta}^R_\lambda) \right) \theta_j = \frac{\sigma^2}{\lambda_j(\lambda_j + 1)} \left[1 + \frac{1}{\lambda_j} - \frac{\lambda_j}{\lambda_j + 1} \right],$$

et

$$\theta_j^\top \left(\operatorname{Cov}(\hat{\beta}) - \operatorname{Cov}(\hat{\beta}^R_\lambda) \right) \theta_j = \frac{\sigma^2}{\lambda_j(\lambda_j + 1)} \frac{(\lambda_j + 1)^2 - \lambda_j^2}{\lambda_j(1 + \lambda_j)} \geq 0.$$

(d) L'avantage principal de l'estimateur ridge est d'avoir une variance plus petite que celle de l'estimateur des moindres carrés classique.