EXAMEN 2019 - 2020

(3 HEURES – DOCUMENTS NON AUTORISÉS)

Exercice 1 (Régression avec features orthogonales, 8pts)

On considère une matrice (déterministe) $\mathbf{X} = [X_i^j]_{1\le i\le n,\, 1\le j\le d}$ de taille $n\times d$ dont les d colonnes $X^j \in \mathbb{R}^n$ sont des vecteurs non nuls et deux à deux orthogonaux. On note $\delta_j = \|X^j\|$ les normes euclidiennes des d colonnes de $\mathbf{X}$ et $\delta^{-2} = \sum_{j=1}^d \delta_j^{-2}$. On suppose $n > d \ge 2$. On observe un vecteur aléatoire $Y \in \mathbb{R}^n$ donné par $Y = \mathbf{X}\beta + \varepsilon$ où $\beta \in \mathbb{R}^d$ est un vecteur de paramètres inconnus et $\varepsilon \in \mathbb{R}^n \sim \mathcal{N}(0, \sigma^2\mathbf{I})$ où $\mathbf{I}$ est la matrice identité dans $\mathbb{R}^n$ avec $\sigma^2 > 0$ inconnu également.

1. (1pt) Explicitez les estimateurs des moindres carrés $\widehat{\beta}$ et $\widehat{\sigma^2}$ de β et σ^2 respectivement, rappelez leurs lois et rappelez rapidement pourquoi ils sont indépendants.
2. (1pt) Donnez un intervalle de confiance pour σ^2 de niveau 95%.
3. (1,5pt) Que peut-on dire des coordonnées $\widehat{\beta}_1, \dots, \widehat{\beta}_d$ de $\widehat{\beta}$? Quelle est la loi de $\sum_{j=1}^d \widehat{\beta}_j$? Construisez un intervalle de confiance pour $\sum_{j=1}^d \beta_j$ au niveau 99% puis un test de $H_0 : \sum_{j=1}^d \beta_j = 0$ contre $H_1 : \sum_{j=1}^d \beta_j \neq 0$ au niveau 1%.
4. (1,5pt) Soit q un entier fixé, $1 \le q \le d-1$. Proposez un test de niveau 5% de l'hypothèse $H_0 : \beta_j = 0$ pour tout $j = q+1, \dots, d$.
5. (1pt) On considère maintenant l'estimateur des moindres carrés pénalisé par la norme ℓ_1 (LASSO) donné pour $\lambda > 0$ par
$$\widehat{\beta}_\lambda \in \operatorname{argmin}_{\beta\in\mathbb{R}^d} \left\{ \frac{1}{2}\|Y - \mathbf{X}\beta\|^2 + \lambda\|\beta\|_1 \right\}$$
où $\|\beta\|_1 = \sum_{j=1}^d |\beta_j|$. Expliquez pourquoi $\widehat{\beta}_\lambda$ est unique dans le cadre considéré ici.
6. (2pt) Montrez que
$$\widehat{\beta}_\lambda \in \operatorname{argmin}_{\beta\in\mathbb{R}^d} \left\{ \sum_{j=1}^d \left(\frac{\delta_j^2}{2}\left(Y^\top X^j/\delta_j^2 - \beta_j\right)^2 + \lambda|\beta_j| \right) \right\}$$
puis montrez que $\widehat{\beta}_j = \delta_j^{-2}\operatorname{sign}(Y^\top X^j)(|Y^\top X^j| - \lambda)_+$ où $x_+ = \max(x, 0)$ pour $x \in \mathbb{R}$ et en déduire une condition suffisante sur λ pour que $\widehat{\beta}_\lambda = 0_d$.

Exercice 2 (Loi binomiale négative, 5pt)

Si $T \sim \text{Gamma}(\alpha, \beta)$ alors T a pour densité $f_T(t) = \frac{\beta^\alpha}{\Gamma(\alpha)} t^{\alpha-1} e^{-\beta t}\mathbf{1}_{t\ge 0}$ où $\alpha, \beta > 0$ et Γ désigne la fonction gamma. On rappelle que l'espérance et la variance de $\text{Gamma}(\alpha, \beta)$ valent α/β et α/β^2 respectivement. Si $Z \sim \text{NegBin}(r, \gamma)$ (binomiale négative) alors Z a pour densité $f_Z(z) = \frac{\Gamma(z+r)}{z!\,\Gamma(r)}\gamma^r(1-\gamma)^z\mathbf{1}_{z\in\mathbb{N}}$ avec $r > 0$ et $\gamma \in (0, 1)$.

1. (1pt) Vérifiez que pour deux variables aléatoires réelles Y et Θ avec Y dans L^2, on a
$$\mathbb{V}(Y) = \mathbb{V}(\mathbb{E}(Y|\Theta)) + \mathbb{E}(\mathbb{V}(Y|\Theta)).$$
2. (1pt) On suppose que $\Theta \sim \text{Gamma}(\alpha, \alpha)$ avec $\alpha > 0$ et que la loi conditionnelle de Y sachant Θ est une loi de Poisson $\mathcal{P}(\lambda\Theta)$ avec $\lambda > 0$. Calculez $\mathbb{V}(Y)$ et $\mathbb{E}(Y)$.
3. (1,5pt) Montrez que $Y \sim \text{NegBin}(\alpha, \alpha/(\alpha+\lambda))$.
4. (1,5pt) On suppose α connu. Montrez que la loi $\text{NegBin}(\alpha, \alpha/(\alpha+\lambda))$ est dans la famille exponentielle naturelle et utilisez cela pour retrouver le résultat de la Question 2.

Exercice 3 (Résidus "studentisés" et données aberrantes en régression, 12pts)

Continuons avec le modèle de régression linéaire, mais cette fois-ci on sait juste que la matrice $\mathbf{X} = [X_i^j]_{1\le i\le n,\,1\le j\le d}$ est de rang plein. On considère à nouveau le modèle linéaire $Y = \mathbf{X}\beta + \varepsilon$ avec $\beta \in \mathbb{R}^d$ inconnu et $\varepsilon \in \mathbb{R}^n \sim \mathcal{N}(0, \sigma^2\mathbf{I})$ où $\sigma^2 > 0$ est inconnu. On considère $\widehat{\beta}$ l'estimateur des moindres carrés et le vecteur des prédictions $\widehat{Y} = \mathbf{X}\widehat{\beta}$. On considère le vecteur des résidus de régression $E = Y - \widehat{Y}$ de coordonnées $E_i = Y_i - \widehat{Y}_i = Y_i - X_i^\top\widehat{\beta}$.

1. (1pt) Montrez que $E \sim \mathcal{N}(0, \sigma^2(\mathbf{I} - \mathbf{H}))$ où $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ est la matrice "chapeau".

2. (1pt) On introduit $\widehat{\sigma}^2 = \frac{1}{n-d}\sum_{i=1}^n (Y_i - \widehat{Y}_i)^2$. Expliquez rapidement pourquoi on ne peut pas dire que les résidus "standardisés"
$$R_i = E_i/\sqrt{\widehat{\sigma}^2(1 - h_{i,i})}$$
sont de loi de student où $h_{i,i}$ est le "levier" du i-ème échantillon, c'est-à-dire le i-ème élément diagonal de $\mathbf{H}$.

3. (1pt) On fait alors la chose suivante : on considère $\mathbf{X}^{(i)} \in \mathbb{R}^{(n-1)\times d}$ la matrice $\mathbf{X}$ sans sa i-ème ligne X_i et $Y^{(i)} \in \mathbb{R}^{n-1}$ le vecteur Y sans sa i-ème coordonnée. On calcule $\widehat{\beta}^{(i)}$ l'estimateur des moindres carrés sans le i-ème échantillon et le vecteur des prédictions $\widehat{Y}^{(i)} = \mathbf{X}\widehat{\beta}^{(i)}$ associé, de coordonnées $\widehat{Y}_{i'}^{(i)} = X_{i'}^\top\widehat{\beta}^{(i)}$ pour $i' = 1, \ldots, n$. Rappelons la formule de Shermann et Morrison, qui dit que pour toute matrice $\mathbf{A} \in \mathbb{R}^{k\times k}$ inversible et tous vecteurs $u, v \in \mathbb{R}^k$, on a
$$(\mathbf{A} + uv^\top)^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}uv^\top\mathbf{A}^{-1}}{1 + v^\top\mathbf{A}^{-1}u}$$
si $1 + v^\top\mathbf{A}^{-1}u \neq 0$. Utilisez cette formule pour montrer que
$$((\mathbf{X}^{(i)})^\top\mathbf{X}^{(i)})^{-1} = (\mathbf{X}^\top\mathbf{X})^{-1} + \frac{(\mathbf{X}^\top\mathbf{X})^{-1}X_iX_i^\top(\mathbf{X}^\top\mathbf{X})^{-1}}{1 - h_{i,i}}.$$

4. (1,5pt) En déduire que l'on peut exprimer les résidus en fonction des résidus dits "externes" via l'équation
$$Y_i - \widehat{Y}_i = (1 - h_{i,i})(Y_i - \widehat{Y}_i^{(i)}).$$

5. (2pt) Montrez que l'on a
$$Y_i - \widehat{Y}_i^{(i)} \sim \mathcal{N}\Big(0, \sigma^2(1 + X_i^\top((\mathbf{X}^{(i)})^\top\mathbf{X}^{(i)})^{-1})X_i\Big)$$
et que
$$1 + X_i^\top((\mathbf{X}^{(i)})^\top\mathbf{X}^{(i)})^{-1})X_i = \frac{1}{1 - h_{i,i}}.$$

6. (1pt) On considère alors les résidus dits "studentisés"
$$T_i = E_i/\sqrt{\widehat{\sigma}^2_{(i)}(1 - h_{i,i})}$$
où $\widehat{\sigma}^2_{(i)}$ est l'estimateur de σ^2 dans le modèle linéaire privé de l'observation i donné par
$$\widehat{\sigma}^2_{(i)} = \frac{1}{n-d-1}\sum_{i'\neq i}(Y_{i'} - \widehat{Y}_{i'}^{(i)})^2.$$
Montrez que $(n-d-1)\widehat{\sigma}^2_{(i)}/\sigma^2$ suit une loi du $\chi^2(n-d-1)$ (Chi-deux à $n-d-1$ degrés de liberté) et que $Y_i - \widehat{Y}_i^{(i)}$ et $\widehat{\sigma}^2_{(i)}$ sont indépendants et en déduire que T_i suit une loi $t(n-d-1)$ (loi de student à $n-d-1$ degrés de liberté).

7. (1pt) Considérons maintenant que $Y^{(i)} = (\mathbf{X}^{(i)})^\top\beta + \varepsilon^{(i)}$ où $\varepsilon^{(i)}$ est le vecteur ε privé de sa i-ème coordonnée et que $Y_i = X_i^\top\beta + \mu_i + \varepsilon_i$ où $\mu_i \in \mathbb{R}$. On dit que l'échantillon i est un outlier si $\mu_i \neq 0$. Proposez un test pour $H_0 : \mu_i = 0$ contre $H_1 : \mu_i \neq 0$ de niveau $\alpha \in (0,1)$ à partir de la Question 6 et calculez sa p-valeur.

8. (3,5pt) Montrez que l'équation suivante, permettant de calculer les résidus studentisés à partir des résidus normalisés, est vérifiée:
$$T_i = R_i\sqrt{\frac{n-d-1}{n-d-R_i^2}}.$$