Exercices relatifs au cours MAT7381

J.-F. Coeurjolly & A. Charpentier

3 Janvier 2020

Contents

1 Algèbre linéaire

Exercice 1 (Un peu d'algèbre en utilisant R)**.**
Définir la matrice $\mathbf{X}$ *de taille* $(8, 3)$ *définie par*

$$\mathbf{X} = \begin{pmatrix} 39 & 49 & 60 \\ 60 & 14 & 30 \\ 51 & 66 & 42 \\ 59 & 51 & 50 \\ 67 & 58 & 52 \\ 51 & 26 & 55 \\ 55 & 41 & 43 \end{pmatrix}$$

1. *Calculer* $\mathbf{A} = \mathbf{X}^\top \mathbf{X}$*.*
2. *Calculer* $\mathbf{B} = \mathbf{A}^{-1}$*.*
3. *Déterminer les valeurs propres et les vecteurs propres de* $\mathbf{A}$*, et vérifier que le déterminant de* $\mathbf{A}$ *est égal au produit des valeurs propres.*
4. *Vérifier numériquement que les valeurs propres de* $\mathbf{A}^{-1}$ *sont les réciproques de celles de* $\mathbf{A}$*.*
5. *Déterminer* $\mathbf{P}$ *et* $\mathbf{D}$ *dans la décomposition* $\mathbf{X}^\top \mathbf{X} = \mathbf{P}\mathbf{D}\mathbf{P}^\top$ *où* $\mathbf{P}$ *est une matrice de colonnes orthonormales et* $\mathbf{D}$ *est diagonale et vérifier que* $\mathbf{X}^\top \mathbf{X}$ *est bien égale à* $\mathbf{P}\mathbf{D}\mathbf{P}^\top$*.*
6. *Calculer la matrice* $\mathbf{P} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}$*. Vérifier que* $\mathbf{P}$ *est idempotente.*
7. *Déterminer les valeurs propres et les vecteurs propres de* $\mathbf{P}$*. (On connaît la réponse : 3 des valeurs propres sont égales à 1, et les autres sont nulles).*
8. *Vérifier également que la matrice* $\mathbf{I}_n - \mathbf{P}$ *est idempotente.*

Exercice 2 (Matrice de projection)**.**
Soient $\mathbf{x}_1, \ldots, \mathbf{x}_p$ p, vecteurs de $\mathbb{R}^n$ et soit $Q \subset \{1, \ldots, p\}$ un sous-ensemble d'indices. On note $\hat{y}_Q$ le projeté orthogonal de y sur l'espace $\mathcal{V}_Q$ engendré par les vecteurs $\mathbf{x}_j$ pour $j \in Q$ et on note $\mathbf{X}_Q = (\mathbf{x}_j, j \in Q)$ la matrice de taille $(n, \#Q)$. Démontrez les propriétés suivantes:

1. $\hat{y}_Q = \mathbf{X}_Q(\mathbf{X}_Q^\top \mathbf{X}_Q)^{-1}\mathbf{X}_Q^\top \mathbf{y}$.
2. *La matrice de projection est égale à* $\mathcal{P}_Q = \mathbf{X}_Q(\mathbf{X}_Q^\top \mathbf{X}_Q)^{-1}\mathbf{X}_Q^\top$.
3. *La matrice $\mathcal{P}_Q$ est idempotente et symétrique. En particulier* $\mathcal{P}_Q\mathbf{X}_Q = \mathbf{X}_Q$.
4. *Si $Q \subset Q' \subset \{1, \ldots, p\}$, alors* $\mathcal{P}_Q\mathcal{P}_{Q'} = \mathcal{P}_{Q'}\mathcal{P}_Q = \mathcal{P}_Q$.
5. *La matrice $\mathcal{P}_Q^\perp = \mathbf{I}_n - \mathcal{P}_Q$ est aussi une matrice de projection. Il s'agit du projecteur orthogonal sur $\mathcal{V}_Q^\perp$, l'orthogonal de $\mathcal{V}_Q$.*
6. $\mathcal{P}_Q\mathcal{P}_Q^\perp = \mathcal{P}_Q^\perp\mathcal{P}_Q = 0$ *et en particulier,* $\mathcal{P}_Q^\perp\mathbf{X}_Q = 0$.

Exercice 3 (Autour des valeurs propres pour $\mathbf{J}_n = \mathbf{e}\mathbf{e}^\top$)**.**
Soit $\mathbf{I}_n$ l'a matrice identité de taille (n, n) et $\mathbf{e} = (1, \ldots, 1)^\top$ un vecteur de taille n et $\mathbf{J}_n = \mathbf{e}\mathbf{e}^\top$.

1. *Montrer que les valeurs propres de $\mathbf{J}_n$ sont n (de multiplicité 1) et 0 (de multiplicité $n-1$). Identifiez les vecteurs propres correspondants.*
2. *Montrer que $\mathbf{I}_n - k\mathbf{J}_n$ est inversible à moins que $k = 1/n$. Déterminer l'inverse $\mathbf{I}_n - k\mathbf{J}_n$ pour $k \neq 1/n$. [Essayez une matrice de la forme $\mathbf{I}_n + m\mathbf{e}\mathbf{e}^\top$].*

Exercice 4 (Valeurs propres de $(\mathbf{I}_n + \mathbf{C})^{-1}C$)**.**
Soit $\mathbf{C}$ une matrice semi-définie positive de taille (n, n). Montrer que si λ est une valeur propre de $\mathbf{C}$, alors $\lambda/(1+\lambda)$ est une valeur propre de $(\mathbf{I}_n+\mathbf{C})^{-1}C$. Montrer ensuite que si $\mathbf{A}$ est définie positive, $\mathbf{B}$ semi-définie positive, et λ est une valeur propre de $\mathbf{A}^{-1}\mathbf{B}$, alors $\lambda/(1+\lambda)$ est une valeur propre de $(\mathbf{A}+\mathbf{B})^{-1}\mathbf{B}$.

Exercice 5 (Encadrement valeurs propres)**.**
Soit $\mathbf{A}$ une matrice réelle symétrique de taille (p, p) et soient $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_p$ ses valeurs propres ordonnées. Montrez alors que pour tout $\mathbf{x} \in \mathbb{R}^p$

$$\lambda_1 \leq \frac{\mathbf{x}^\top \mathbf{A}\mathbf{x}}{\mathbf{x}^\top \mathbf{x}} \leq \lambda_p.$$

Par ailleurs, montrez qu'une matrice $\mathbf{A}$ *réelle symétrique est définie positive (resp. semi-définie positive) si et seulement si toutes ses valeurs propres sont toutes positives (resp. positives ou nulles).*

Exercice 6 (Théorème de Cochran version matricielle)**.** *Si* $\mathbf{A}$ *est une matrice idempotente alors*

- $\mathrm{rg}(\mathbf{A}) = \mathrm{tr}(\mathbf{A})$.
- *Les valeurs propres de* $\mathbf{A}$ *valent soit 0 soit 1.*

Exercice 7 (Matrice de Helmert)**.**
On définit la matrice de Helmert de taille (n, n) *par la matrice*

$$\mathbf{H} = \begin{pmatrix} 1/\sqrt{n} & 1/\sqrt{n} & \dots & \dots & 1/\sqrt{n} \\ 1/\sqrt{2} & -1/\sqrt{2} & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & 0 \\ \dots & \dots & \dots & 0 & 0 \\ 1/\sqrt{n(n-1)} & 1/\sqrt{n(n-1)} & \dots & 1/\sqrt{n(n-1)} & -(n-1)/\sqrt{n(n-1)} \end{pmatrix}$$

Plus simplement, on réécrit $\mathbf{H} = \begin{pmatrix} \mathbf{e}^\top/\sqrt{n} \\ \mathbf{P}_2 \end{pmatrix}$ *où* $\mathbf{e} = (1, \dots, 1)^\top$ *est un vecteur de taille* $(n, 1)$ *et* $\mathbf{P}_2$ *une matrice de taille* $(n-1, n)$.

1. *Montrez que la matrice de Helmert* $\mathbf{H}$ *est une matrice orthogonale. Par conséquent*

$$\mathbf{I}_n = \mathbf{H}\mathbf{H}^\top = \frac{1}{n}\mathbf{e}\mathbf{e}^\top + \mathbf{P}_2^\top\mathbf{P}_2 \Leftrightarrow \mathbf{I}_n - \frac{1}{n}\mathbf{e}\mathbf{e}^\top = \mathbf{P}_2^\top\mathbf{P}_2$$

2. *Partons de* p *covariables observées pour* n *individus, i.e.* $\mathbf{x}_1, \dots, \mathbf{x}_p$ *où chaque* $\mathbf{x}_i$ *est un vecteur de taille* $(n, 1)$. *Construisons* $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_p)$ *la matrice de covariables (appelée matrice de design) et soit* $\check{\mathbf{X}}$ *la version centrée (chaque covariable est recentrée). Montrez que*

$$\check{\mathbf{X}} = \left(\mathbf{I}_n - \frac{1}{n}\mathbf{e}\mathbf{e}^\top\right)\mathbf{X} = \mathbf{Q}\mathbf{X}, \text{ avec } \mathbf{Q} = \mathbf{P}_2^\top\mathbf{P}_2 \qquad (\text{💡})$$

3. *Montrez que les matrices* $\frac{1}{n}\mathbf{e}\mathbf{e}^\top$ *et* $\mathbf{Q}$ *sont idempotentes et* $\mathrm{rg}(\frac{1}{n}\mathbf{e}\mathbf{e}^\top) = 1$ *et* $\mathrm{rg}(\mathbf{Q}) = n - 1$.

4. *Puisque* $\mathbf{Q} = \mathbf{P}_2^\top(\mathbf{P}_2\mathbf{P}_2^\top)\mathbf{P}_2 = \mathbf{P}_2^\top(\mathbf{P}_2\mathbf{P}_2^\top)^{-1}\mathbf{P}_2$, *déduire* $\mathbf{Q}$ *est une matrice de projection. Et donc que* $\mathbf{QX}$ *est le projeté ortogonal de* $\mathbf{X}$ *sur les colonnes de* $\mathbf{P}_2$.

Exercice 8 (Application des matrices de Helmert à la loi de la norme d'un vecteur)**.**
Soit $\mathbf{Y} = (Y_1, \ldots, Y_n)$ *un vecteur* n *v.a.i.i.d. de moyenne* μ *et de variance* σ^2. *Et définissons,* $\mathbf{X} = \mathbf{HY}$ *où* $\mathbf{H}$ *est la matrice de Helmert. Montrez que*

1. $\mathrm{E}(\mathbf{X}) = (\sqrt{n}\mu, 0, \ldots, 0)^\top$ *et* $\mathrm{Var}(\mathbf{X}) = \sigma^2\mathbf{I}_n$ *et donc que les variables* X_i *sont non corrélées,*

2. $X_1 = \sqrt{n}\overline{Y}$, $\sum_{i=1}^n (Y_i - \overline{Y})^2 = \sum_{i=2}^n X_i^2$. *En effet:*

$$\sum_{i=1}^n (Y_i - \overline{Y})^2 = \|(\mathbf{I}_n - \mathbf{e}\mathbf{e}^\top/n)\mathbf{Y}\|^2 = \ldots\ldots$$

3. *En déduire que si les* Y_i *sont de loi normale, alors*

   (a) $S^2 = \frac{1}{n-1}\sum_{i=1}^n (Y_i - \overline{Y})^2$ *et* $\overline{Y}$ *sont indépendantes.*

   (b) $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$.

2 Statistique mathématique, statistique inférentielle

Exercice 9 (Convergence de T_n et $F_{n,p}$)**.**
Les deux résultats ci-dessous utilisent le théorème de Slustsky.

1. *Soient $X_n \sim T_n$ et $Z \sim \mathcal{N}(0,1)$, montrez que lorsque $n \to \infty$, $X_n \to Z$ en distribution (rappel: la convergence en distribution correspond à la convergence des fonctions de répartition);*
2. *Soient $X_n \sim F_{p,n}$ et $U_p \sim \chi^2_p$, montrez que lorsque $n \to \infty$, $X_n \to U_p/p$.*
3. *Vérifiez ces deux propriétés par simulation en `R`.*

Exercice 10 (Application de la δ-méthode)**.**
Soient $X_1, \ldots, X_n$ n v.a. i.i.d. de loi $\mathcal{B}(p)$ (ou $\mathcal{P}(\lambda)$) et soit $\overline{X}_n = n^{-1}\sum_i X_i$.

1. *Quelle est la loi de $\sqrt{n}(\overline{X}_n - \mu)$ pour $\mu = p$ ou λ selon que $X_i \sim \mathcal{B}(p)$ ou $X_i \sim \mathcal{P}(\lambda)$.*
2. *dans le cas $\mathcal{B}(p)$, appliquez la delta méthode à la fonction $g(t) = \log(t)$.*
3. *dans le cas $\mathcal{P}(\lambda)$, appliquez la delta méthode à une fonction g, telle que $\sqrt{n}(g(\overline{X}_n) - g(\mu)) \xrightarrow{d} \mathcal{N}(0,1)$? À quoi peut servir ce genre de résultat?*

Exercice 11 (Loi marginale et loi conditionnelle)**.**
Soient X et Y deux variables aléatoires continues dont la densité jointe est donnée par $f_{(X,Y)}(x,y) = 2y\mathrm{e}^{-(2+x)y}$ si $y \geq 0$ et 0 sinon.

1. *Vérifiez qu'il s'agit bien d'une distribution de probabilité.*
2. *Calculez la densité marginale de X.*
3. *Calculez la densité marginale de Y.*
4. *Pour $x > 0$, calculez la densité conditionnelle de Y sachant $X = x$.*
5. *Pour $y > 0$, calculez la densité conditionnelle de Y sachant $Y = y$.*

Exercice 12 (Estimation pour une loi $\mathcal{N}(\theta, \theta)$)**.**
Pour cette question, on travaille avec n variables aléatoires Y_i de loi normale de moyenne et variance égales à θ, où θ est un paramètre réel strictement positif ($\theta > 0$) qu'on se propose d'estimer à partir d'un échantillon $\mathbf{Y} = (Y_1, \ldots, Y_n)$ de n variables aléatoires indépendantes. On définit $S^2 = (n-1)^{-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2$ et on rappelle que $(n-1)S^2/\theta \sim \chi^2_{n-1}$.

1. *Trouver une statistique exhaustive pour cette famille ? Cette statistique est-elle minimale?*

2. *Montrer que $\bar{Y}$ et S^2 sont deux estimateurs non biaisés de θ. En déduire que pour tout $\alpha \in (0,1)$ l'estimateur défini par*

$$\hat{\theta}(\mathbf{Y}) = \alpha \overline{Y} + (1-\alpha) S^2$$

est sans biais.

3. *Calculez la variance de $\bar{Y}$, de S^2 puis (en se basant sur un résultat connu) celle de $\hat{\theta}(\mathbf{Y})$. En déduire que $\hat{\theta}(\mathbf{Y})$ converge en moyenne quadratique vers θ lorsque $n \to \infty$.*

4. *Quelle est la valeur de α minimisant la variance de $\hat{\theta}(\mathbf{Y})$? Parmi les trois estimateurs quel est le plus efficace ?*

Exercice 13 (Statistique exhaustive)**.**

1. *Soient $X_1, \ldots, X_n$ n v.a.i.i.d. de loi $\mathcal{U}([\theta - 1/2, \theta + 1/2])$. Déterminez une statistique exhaustive pour θ.*

2. *Soient $X_1, \ldots, X_n$ n v.a.i.i.d. de loi $\mathcal{N}(\mu, \sigma^2)$. Construisez à partir de cet échantillon une statistique exhaustive*

   (a) *pour μ, σ^2 étant connu.*

   (b) *pour σ^2, μ étant connu.*

   (c) *pour $\theta = (\mu, \sigma^2)^\top$.*

3. *Soient $X_1, \ldots, X_n$ un échantillon de n v.a.i.i.d. de densité $f(x;\theta) = \theta e^{-x+\theta}$ si $x \geq 0$ et 0 sinon, où $\theta > 0$. Déterminez une statistique exhaustive pour le paramètre θ.*

Exercice 14 (Autour d'une loi de Poisson)**.**
On considère le modèle statistique $\{\mathcal{P}(\lambda)\}$ *où* $\mathcal{P}(\lambda)$ *désigne la loi de Poisson de paramètre* $\lambda > 0$*. L'objet de l'exercice est d'étudier des estimateurs du paramètre* $e^{-\lambda}$*. Pour un échantillon* $(X_1, \ldots, X_n)$ *de* n *v.a.i.i.d. de loi* $\mathcal{P}(\lambda)$*, on introduit les estimateurs de* $e^{-\lambda}$ *suivants:*

$$\hat{\theta}_1 = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(X_i = 0) \qquad et \qquad \hat{\theta}_2 = \left(\frac{n-1}{n}\right)^{S_n}$$

avec $S_n = \sum X_i$*.*

1. *Montrer que* $\hat{\theta}_1$ *et* $\hat{\theta}_2$ *sont sans biais et consistants.*
2. *Prouver que* S_n *est une statistique exhaustive complète.*
3. *Etablir l'égalité* $\mathrm{E}(\hat{\theta}_1 \mid S_n) = \theta^2$*. Conclusion ?*
4. *Etablir la normalité asymptotique des estimateurs* $\hat{\theta}_1$ *et* $\hat{\theta}_2$*.*

3 Vecteurs et vecteurs gaussiens

Exercice 15 (Propriétés espérance et matrice de covariance)**.**
Soit $\mathbf{Y}$ *un vecteur aléatoire de dimension* d*, de moyenne* $\boldsymbol{\mu}$ *et de matrice de covariance* $\boldsymbol{\Sigma}$*. Soient* $\mathbf{A}$ *et* $\mathbf{B}$ *deux matrices réeeles de taille* (d, p) *et* (d, q) *et enfin soit* $\mathbf{a} \in \mathbb{R}^p$*, montrez que*

1. $\mathrm{Var}(\mathbf{Y}) = \mathrm{E}(\mathbf{Y}\mathbf{Y}^\top) - \boldsymbol{\mu}\boldsymbol{\mu}^\top$.
2. $\mathrm{E}\left(\mathbf{A}^\top\mathbf{Y} + \mathbf{a}\right) = \mathbf{A}^\top\boldsymbol{\mu} + \mathbf{a}$.
3. $\mathrm{Var}\left(\mathbf{A}^\top\mathbf{Y} + \mathbf{a}\right) = \mathbf{A}^\top\boldsymbol{\Sigma}\mathbf{A}$.
4. $\mathrm{Cov}\left(\mathbf{A}^\top\mathbf{Y}, \mathbf{B}^\top\mathbf{Y}\right) = \mathbf{A}^\top\boldsymbol{\Sigma}\mathbf{B}$.

Exercice 16.
Soit $\mathbf{A}$ *une matrice réelle symétrique de taille* (d, d) *et* $\mathbf{Y}$ *un vecteur aléatoire de moyenne* $\boldsymbol{\mu}$ *et de matrice de covariance* $\boldsymbol{\Sigma}$*, montrez que*

$$\mathrm{E}\left(\mathbf{Y}^\top\mathbf{A}\mathbf{Y}\right) = \boldsymbol{\mu}^\top\mathbf{A}\boldsymbol{\mu} + \mathrm{tr}(\mathbf{A}\boldsymbol{\Sigma}).$$

Exercice 17 (Théorème de Cochran (version simplifiée))**.**
Soit $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ *et* F *un sous-espace vectoriel de* $\mathbb{R}^d$ *de dimension* $p < d$*. Soient* $F^\perp$ *son orthogonal,* $\mathbf{P}_F$ *et* $\mathbf{P}_{F^\perp}$ *les matrices de projection sur* F *et* $F^\perp$*, alors*

- $\mathbf{P}_F\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{P}_F)$ *et* $\mathbf{P}_{F^\perp}\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{P}_{F^\perp})$ *et* $(\mathbf{P}_F\mathbf{X}) \perp (\mathbf{P}_{F^\perp}\mathbf{X})$.
- $\|\mathbf{P}_F\mathbf{X}\|^2 \sim \chi^2_p$*,* $\|\mathbf{P}_{F^\perp}\mathbf{X}\|^2 \sim \chi^2_{d-p}$ *et ces deux variables sont indépendantes.*

Exercice 18 (Application simple du théorème de Cochran)**.**
Soient $Y_1, \ldots, Y_n$ *n v.a. i.i.d. de loi* $\mathcal{N}(0, 1)$ *et soit* $\overline{Y} = n^{-1}\sum Y_i$ *et* $S^2 = (n-1)^{-1}\sum_i (Y_i - \overline{Y})$*. Montrez que* $\overline{Y}$ *et* S^2 *sont indépendantes et que* $(n-1)S^2\chi^2_{n-1}$ *en utilisant le théorème de Cochran avec* $F = \mathrm{Vec}(\mathbf{1}_n)$*. Pour cela, vous détaillerez* $\mathbf{P}_F, \mathbf{P}_{F^\perp}$ *et leurs propriétés.*

Exercice 19 (Transformation linéaire d'un vecteur gaussien)**.**
Soient Y_1, Y_2 et Y_3 des variables aléatoires indépendantes, de moyenne 0 et de variance σ^2. Soit $Z_1 = Y_1$, $Z_2 = Y_1 + Y_2$ et $Z_3 = Y_1 + Y_2 + Y_3$. Déterminez la matrice de covariance de $\mathbf{Z} = (Z_1, Z_2, Z_3)^\top$.

Exercice 20 (Normalisation d'une gaussienne multivariée)**.**
Supposons que $\mathbf{X} = (X_1, X_2)^\top \sim \mathcal{N}(0, \mathbf{\Sigma})$, où $\mathbf{\Sigma} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$. Déterminez une matrice $\mathbf{A}$ telle que $\mathbf{Y} = \mathbf{AX} \sim \mathcal{N}(0, \mathbf{I}_2)$.

Exercice 21 (À faire en R!)**.**
Soit $\mathbf{X}$ un vecteur gaussien de taille 4 de moyenne $\boldsymbol{\mu} = (2, 4, 5, 6)^\top$ et de matrice de covariance

$$\mathbf{\Sigma} = \begin{pmatrix} 11.69 & 6.45 & 8.32 & -2.32 \\ 6.45 & 26.00 & 13.06 & 8.45 \\ 8.32 & 13.06 & 29.36 & 13.43 \\ -2.32 & 8.45 & 13.43 & 17.61 \end{pmatrix}$$

1. *Déterminer la distribution conditionnelle de X_1 sachant $(X_2, X_3, X_4)^\top = (1, 4, 3)^\top$*
2. *Déterminer la distribution conditionnelle de $\mathbf{Y} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ sachant $\begin{pmatrix} X_3 \\ X_4 \end{pmatrix} = \begin{pmatrix} 5 \\ 1 \end{pmatrix}$.*
3. *Déterminer la distribution de $\mathbf{Z} = \mathbf{AX}$ avec $\mathbf{A} = \begin{pmatrix} 1 & 2 & 1 & 6 \\ 3 & 4 & 2 & 4 \end{pmatrix}$.*
4. *Définir une transformation $\mathbf{Y} = \mathbf{B}^\top \mathbf{X}$ telles que les composantes de $\mathbf{Y}$ soient indépendantes, où $\mathbf{B}$ est une matrice de taille $(4, 4)$.*
5. *Déterminer la loi de $\overline{X}$ et calculer en R, $\mathrm{P}(\overline{X} \leq 14)$.*

Exercice 22 (Loi d'une forme quadratique d'un vecteur gaussien)**.**
Soit $\mathbf{Y} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$ et soit $\mathbf{X}$ une matrice de taille (n, q) de rang plein.

1. *Montrez que $\mathbf{Y}^\top (\mathbf{I}_n - \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top) \mathbf{Y} \sim \chi^2_{n-q}$.*
2. *Montrez que $\mathbf{Y}^\top (\mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top) \mathbf{Y} \sim \chi^2_q$.*

Exercice 23 (Forme quadratique particulière)**.**
Soit $\mathbf{Y} = (Y_1, \ldots, Y_q)^\top$ *un vecteur gaussien centré de matrice de covariance* $\mathbf{I}_q$*. Déterminer l'espérance et la variance de* $(Y_1 - Y_2)^2 + (Y_2 - Y_3)^2 + \cdots + (Y_{q-1} - Y_q)^2$*. [Déterminer la matrice* $\mathbf{A}$ *telle que* $\mathbf{Y}^\top \mathbf{A} = (Y_1 - Y_2, \ldots, Y_{q-1} - Y_q)^\top$*.]*

Exercice 24 (Vérification de quelques propriétés en `R`)**.**

1. *Soit* $\mathbf{Y}$ *un vecteur aléatoire de dimension* d *(non nécessairement gaussien) de moyenne* $\boldsymbol{\mu}$ *et de matrice de covariance* $\boldsymbol{\Sigma}$*. Enfin soit* $\mathbf{A}$ *une matrice réelle de taille* (d, d) *, montrez que*

$$\mathrm{E}(\mathbf{Y}^\top \mathbf{A} \mathbf{Y}) = \boldsymbol{\mu}^\top \mathbf{A} \boldsymbol{\mu} + \mathrm{tr}(\mathbf{A}\boldsymbol{\Sigma}).$$

   [indication: on remarquera que $\mathrm{E}(\mathbf{Y}^\top \mathbf{A} \mathbf{Y}) = \mathrm{E}(\mathrm{tr}(\mathbf{Y}^\top \mathbf{A} \mathbf{Y}))$*]*

2. *Soit*

$$\boldsymbol{\mu} = \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.4 & 0.4 \\ 0.4 & 1 & 0.4 \\ 0.4 & 0.4 & 1 \end{pmatrix} \quad \textit{et} \quad \mathbf{A} = \begin{pmatrix} 1 & 4 & 7 \\ 2 & 5 & 8 \\ 3 & 6 & 9 \end{pmatrix}.$$

   En `R`*, vérifiez que* $\boldsymbol{\Sigma}$ *est inversible, et calculez* $\mathrm{tr}(\mathbf{A}\boldsymbol{\Sigma})$*,* $\boldsymbol{\mu}^\top \mathbf{A} \boldsymbol{\mu}$ *puis* $\mathrm{E}(\mathbf{Y}^\top \mathbf{A} \mathbf{Y})$*.*

3. *Simulation en* `R`*:*

   - *Créez une fonction permettant de simuler* $\mathbf{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *avec* $\boldsymbol{\mu}$ *et* $\boldsymbol{\Sigma}$ *fixés aux valeurs précédentes.*
   - *Simulez* $m = 1000$ *réalisations de* $\mathbf{Y}$*, puis calculez pour chacune d'entre elles la réalisation de* $\mathbf{Y}^\top \mathbf{A} \mathbf{Y}$*.*
   - *Calculez la moyenne empirique des* m *réalisations de* $\mathbf{Y}^\top \mathbf{A} \mathbf{Y}$*. Cette quantité est sensée approcher le résultat de la question 2.*

Exercice 25 (Forme quadratique d'un vecteur gaussien, toujours)**.**
Soit $\mathbf{Y} = (Y_1, Y_2, Y_3)^\top$ *un vecteur gaussien centré de matrice de covariance* $\mathbf{I}_3$*.*

1. *Déterminer l'espérance de* $Z = (Y_1 - Y_2)^2 + (Y_2 - Y_3)^2$ *(en utilisant la formule de l'exercice précédent) [indication: déterminer la matrice* $\mathbf{A}$ *de taille* $(q-1, q)$ *telle que* $A\mathbf{Y} = (Y_1 - Y_2, \ldots, Y_{q-1} - Y_q)^\top$*.] Déterminer ensuite* $\mathrm{Var}(Z)$ *en utilisant une formule du cours (dans le cas gaussien, on rappelle que* $\mu_4 - 3\sigma^4 = 0$*).*

2. *Fixons* $q = 3$*. Déterminer la loi de* $(Y_1 + Y_2 - Y_3)^2/9 + (Y_1 + Y_2 - Y_3)^2/9 + (-Y_1 - Y_2 + Y_3)^2/9$*. [indication: déterminer la matrice* $\mathbf{A}$ *de taille* $(3, 3)$ *telle que* $A\mathbf{Y} = (Y_1 + Y_2 - Y_3, Y_1 + Y_2 - Y_3, -Y_1 - Y_2 + Y_3)^\top/3$*. Quelles sont les propriétés de* $\mathbf{A}$*?]*

Exercice 26. 1. *Soit* S *la v.a. définie par*

$$\begin{aligned} 16S =& (3X_1 + X_2 - X_3 + X_4)^2 + (X_1 + 3X_2 + X_3 - X_4)^2 \\ &+ (-X_1 + X_2 + 3X_3 + X_4)^2 + (X_1 - X_2 + X_3 + 3X_4)^2, \end{aligned}$$

où $X_1, \ldots, X_4$ *sont des v.a. indépendantes de loi* $\mathcal{N}(0, 1)$*. Déterminer la loi de* S*. Indication: trouvez la matrice* $\mathbf{P}$ *telle que* $S = \|\mathbf{PX}\|^2$*. Pour cette matrice* $\mathbf{P}$*, on a*

```
> sum((P%*%P-P)^2)

[1] 0
```

2. *Quelle est la loi de* $\|(\mathbf{I}_4 - \mathbf{P})\mathbf{X}\|^2$*? Calculez également* $\mathrm{E}((\|(\mathbf{I}_4 - \mathbf{P})\mathbf{X}\|^2 - 1)\exp(\|\mathbf{PX}\|^2))$*.*

Exercice 27 (Construction d'intervalles et région de confiance)**.**
Les deux questions ci-dessous peuvent ensuite être traitées par simulation en `R`*.*

1. *Soit* $\mathbf{Y}$ *un échantillon de* $\mathcal{P}(\lambda)$*, construire un IC approximatif de* λ *en utilisant le fait que* $\sqrt{n}(\overline{Y} - \lambda) \xrightarrow{d} \mathcal{N}(0, \lambda)$ *ou que* $\sqrt{n}(g(\overline{Y}_n - g(\lambda)) \xrightarrow{d} \mathcal{N}(0, 1)$ *lorsque* $n \to \infty$ *avec* $g(t) = 2\sqrt{t}$*.*

2. *Soient* $\mathbf{Y}_1$ *et* $\mathbf{Y}_2$ *deux échantillons gaussiens indépendants de moyenne* μ_1 *et* μ_2 *et de variance 1 et 2 respectivement. Construire une région de confiance pour* (μ_1, μ_2)*.*

   (a) *soit* $\mathbf{A}_n = \sqrt{n}(\overline{Y}_1 - \mu_1, \overline{Y}_2 - \mu_2)$*. Quelle est la loi de* $\mathbf{A}_n$ *puis de* $\mathbf{A}_n\mathbf{A}_n^\top$*.*

   (b) *Montrez alors que* $\mathrm{P}(\mathbf{A}_n\mathbf{\Sigma}^{-1}\mathbf{A}_n^\top \leq \chi^2_{1-\alpha}) = 1 - \alpha$ *où* $\mathbf{\Sigma} = \mathrm{diag}(1, 2)$*. Et en déduire une région de confiance pour* (μ_1, μ_2)*.*

4 Estimation dans les modèles linéaires homoscédastiques

Exercice 28 (Régresseurs orthogonaux)**.**
Dans le contexte (et les notations du chapitre du cours associé), si les variables $\mathbf{x}_j$*,* $j = 1, \dots, p$ *sont orthogonales,* $(\mathbf{X}^\top \mathbf{X}) = \mathrm{diag}(\|\mathbf{x}_j\|^2, j = 1, \dots, p)$*, montrez que*

$$\hat{\beta}_j = \frac{\mathbf{x}_j^\top \mathbf{Y}}{\|\mathbf{x}_j\|^2}, \quad j = 1, \dots, p.$$

Exercice 29 (Estimateurs MCO $p = 2$)**.**
Dans le contexte (et les notations du chapitre du cours associé), montrez que l'estimateur des MCO du modèle de régression linéaire simple $Y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$ *(sous l'hypothèse que* $\widehat{\mathrm{Var}}(\mathbf{x}) \neq 0$*) est donné par*

$$\hat{\beta}_2 = \frac{\widehat{\mathrm{Cov}}(\mathbf{x}, \mathbf{Y})}{\widehat{\mathrm{Var}}(\mathbf{x})} \quad \text{et} \quad \hat{\beta}_1 = \overline{Y} - \hat{\beta}_2 \overline{x}. \qquad [\widehat{\mathrm{Cov}}(\mathbf{a}, \mathbf{b}) = (n-1)^{-1} \sum (a_i - \overline{a})(b_i - \overline{b})]$$

Vérifiez que ces estimateurs sont sans biais et

$$\mathrm{Var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum (x_i - \overline{x})^2}, \ \mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2 \sum x_i^2}{n \sum (x_i - \overline{x})^2}, \ \mathrm{Cov}(\hat{\beta}_1, \hat{\beta}_2) = -\frac{\sigma^2 \overline{x}}{\sum (x_i - \overline{x})^2}$$

Exercice 30 (Théorème de Gauss-Markov)**.**
Sous les hypothèses $\mathcal{H}_1 - \mathcal{H}_2$*, montrez que l'estimateur des MCO est optimal (sans biais et de variance minimale) dans la classe des estimateurs linéaires sans biais de* $\boldsymbol{\beta}$*.*

Exercice 31 (Convergence en moyenne quadratique)**.**
Sous les hypothèses, $\mathcal{H}_1 - \mathcal{H}_3$ *montrez que l'estimateur des MCO* $\hat{\boldsymbol{\beta}}$ *est convergent en moyenne quadratique*

Exercice 32 (Propriété de $\hat{\mathbf{Y}}$ et $\hat{\boldsymbol{\varepsilon}}$)**.**
Sous les hypothèses $\mathcal{H}_1 - \mathcal{H}_2$, montrez que

$$\begin{aligned} \mathrm{E}(\hat{\mathbf{Y}}) = \mathbf{X}\boldsymbol{\beta}, \quad \mathrm{Var}(\hat{\mathbf{Y}}) = \sigma^2 \mathcal{P}_{\mathbf{X}} = \sigma^2 \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \\ \mathrm{E}(\hat{\boldsymbol{\varepsilon}}) = 0, \mathrm{Var}(\hat{\boldsymbol{\varepsilon}}) = \sigma^2 \mathcal{P}_{\mathbf{X}^\perp} = \sigma^2(\mathbf{I}_n - \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top) \\ \mathrm{Cov}(\hat{\mathbf{Y}}, \hat{\boldsymbol{\varepsilon}}) = 0. \end{aligned}$$

Exercice 33 (Estimation de σ^2 et de $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}}$)**.**
Montrez que

- *la statistique $\hat{\sigma}^2 = \frac{1}{n-p}\sum \hat{\varepsilon}_i^2 = \frac{\|\hat{\boldsymbol{\varepsilon}}\|^2}{n-p}$ est un estimateur sans biais de σ^2.*
- *un estimateur sans biais de $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}} = \mathrm{Var}(\hat{\boldsymbol{\beta}})$ est donné par $\widehat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}} = \hat{\sigma}^2(\mathbf{X}^\top \mathbf{X})^{-1}$.*

Montrez également que sous les hypothèses $\mathcal{H}_1, \mathcal{H}_{2'}$ et $\mathcal{H}_3$, les estimateurs $\hat{\sigma}^2$ et $\widehat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}}$ sont respectivement des estimateurs convergents en moyenne quadratique de σ^2 et $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}}$.

Exercice 34 (Erreur de prédiction)**.**
Montrez que l'erreur de prédiction $e_{n+1} = Y_{n+1} - \hat{Y}_{n+1}$ vérifie

$$e_{n+1} = \mathbf{x}'_{n+1}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \boldsymbol{\varepsilon} + \varepsilon_{n+1}, \quad \mathrm{Var}(e_{n+1}) = \sigma^2 \left(1 + \mathbf{x}'_{n+1}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{x}'^{\top}_{n+1}\right).$$

Montrez également que dans le cas $p = 2$

$$\mathrm{Var}(e_{n+1}) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_{n+1} - \overline{x})^2}{\sum (x_i - \overline{x})^2}\right).$$

Exercice 35 (Coeffcient de corrélation $p = 2$)**.**
Vérifiez que lorsque $p = 2$ et que $\mathbf{1} \in \mathcal{V}(\mathbf{X})$,

$$R^2 = \frac{\widehat{\mathrm{Cov}}(\mathbf{x}, \mathbf{Y})^2}{\widehat{\mathrm{Var}}(\mathbf{x})\widehat{\mathrm{Var}}(\mathbf{Y})} = \widehat{\mathrm{Corr}}(\mathbf{x}, \mathbf{Y})^2.$$

Exercice 36 (Evolution de R^2 avec p)**.**
Lorsque l'on augmente le nombre de covariables, le coefficient R^2 augmente nécessairement.

Exercice 37 (Pratique en R sur les données ozone)**.**
Commencez par charger le jeu de données `ozone`*, du fichier* `ozone.RData`*. On se concentrera sur une sous-partie de ce jeu de données constituée des variables O3 variable à expliquer (concentration en ozone mesurée sur plusieurs jours à Rennes, France) et les régresseurs T12, Ne12 et Vx qui sont respectivement la température à 12h, la nébulosité à 12h et la vitesse du vent Ouest-Est.*

1. *Combien y a-t-il d'observations. Représentez ces données sous la forme d'une matrice de grahique et calculez la matrice de corrélation. Quelles vous semblent être les variables apportant de l'information pour expliquer O3?*

2. *Considérez le modèle $\mathcal{M}_1$ défini par la formule* `R`

   ```
   > O3~T12
   ```

   Comment s'écrit mathématiquement ce modèle? Calculez les estimations par MCO en utilisant les formules du cours puis en utilisant la fonction `lm`*.*

3. *Tracez O3 en fonction de T12, puis la droite de régression. Vérifiez qu'elle passe bien par le point moyen.*

4. *Quelle valeur de la concentration prévoit-on pour une journée dont la température à 12h serait de 23 degrés? (utilisez la fonction* `predict`*). Représentez cette valeur sur le graphique.*

5. *Considérez maintenant le modèle $\mathcal{M}_2$ défini par la formule* `R`

   ```
   > O3~T12+Ne12+Vx
   ```

   Comment s'écrit mathématiquement ce modèle? Soit $\mathbf{X}$ la matrice de design. Vérifiez que $\mathbf{X}^\top\mathbf{X}$ est inversible. Calculez les estimations par MCO en utilisant les formules du cours puis en utilisant la fonction `lm`*.*

6. *Pour le modèle $\mathcal{M}_2$, calculez en utilisant les formules du cours une estimation de σ^2 puis de $\mathbf{\Sigma}_{\hat{\beta}}$. Vérifiez ce calcul en utilisant la fonction* `vcov`*.*

7. *Toujours pour le modèle $\mathcal{M}_2$, calculez le coefficient de détermination multiple R^2 en utilisant les formules du cours. Puis vérifiez ce calcul en utilisant les fonctions internes de* `R` *(*`summary(lm(...))`*).*

8. *Parmi les deux modèles, lequel vous semble être le plus prédictif?*

Exercice 38 (Données centrées ou non)**.**
On considère un modèle de régression linéaire homoscédastique qui s'écrit

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \beta_1 + \beta_2\mathbf{x}_2 + \cdots + \beta_p\mathbf{x}_p + \boldsymbol{\varepsilon} \tag{1}$$

où $\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$ *et* $\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}_n$*. Nous supposerons que la matrice* $\mathbf{X}$ *est de rang plein (*$\mathcal{H}_1$*). Le coefficient* β_1 *s'interprète dans ce modèle comme la valeur de* $\mathrm{E}(Y_i)$ *lorsque tous les régresseurs sont mis à zéro. Parfois, la notion de covariable égale à zéro n'a pas de sens (imaginez que* $\mathbf{x}_2$ *soit la variable taille d'un individu). Dans ce genre de situations, on peut travailler avec le modèle centré*

$$\mathbf{Y} = \beta'_1 + \beta'_2\mathring{\mathbf{x}}_2 + \cdots + \beta'_p\mathring{\mathbf{x}}_p + \boldsymbol{\varepsilon} \tag{2}$$

où $\mathring{\mathbf{x}}_j = \mathbf{x}_j - \overline{x}_j$ *est la covariable* $\mathbf{x}_j$ *recentrée.*

1. *Comment s'interprète maintenant le paramètre* β'_1 *dans le modèle* (2)*?*

2. *Montrez qu'il existe une matrice* $\mathbf{A}$ *de taille* (p, p) *telle que* $\boldsymbol{\beta} = \mathbf{A}\boldsymbol{\beta}'$ *où* $\boldsymbol{\beta}' = (\beta'_1, \ldots, \beta'_p)^\top$*.*

3. *Ainsi si l'on estime le modèle* (1) *et* (2) *par l'estimateur MCO, nous devons avoir nécessairement que* $\hat{\boldsymbol{\beta}} = \mathbf{A}\hat{\boldsymbol{\beta}}'$*. En déduire en particulier que* $\hat{\beta}_j = \hat{\beta}'_j$ *pour* $j = 2, \ldots, p$*.*

4. *Quelle relation lie* $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}} = \mathrm{Var}(\hat{\boldsymbol{\beta}})$ *à* $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}'} = \mathrm{Var}(\hat{\boldsymbol{\beta}}')$*?*

5. *Montrez que* $\hat{\sigma}^2$ *et* $\hat{\sigma}'^2$ *les estimateurs de* σ^2 *des deux modèles sont les mêmes. En déduire une relation entre* $\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}}$ *et* $\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}'}$*.*

6. *Vérification en* `R`*: considérez le jeu de données* `COSanDiego` *du fichier* `COSanDiego.RData`*.On considèrera les modèles*

$$CO_i = \beta_1 + \beta_2\, Traffic_i + \beta_3\, Wind_i + \varepsilon_i \tag{3}$$
$$CO_i = \beta_1 + \beta_2\, TrafficC_i + \beta_3\, WindC_i + \varepsilon_i \tag{4}$$

où les variables TrafficC et WindC sont les versions centrées. En utilisant les fonctions internes de `R (lm,vcov)`

- *Estimez ces deux modèles. Vérifiez qu'en effet* $\hat{\boldsymbol{\beta}} = \mathbf{A}\hat{\boldsymbol{\beta}}'$.
- *Calculez les matrices de covariance estimées* $\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}}$ *et* $\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}'}$ *et vérifiez leur relation.*

5 Inférence pour les modèles linéaires homoscédastiques

Exercice 39 (Maxixmum de vraisemblance - Modèle linéaire gaussien)**.**
Soit un modèle linéaire homoscédastique vérifiant les hypothèses $\mathcal{H}_1$–$\mathcal{H}_2^{\text{Gauss}}$*, montrez que*

- *l'estimateur du MV de* $\boldsymbol{\beta}$ *vaut* $\hat{\boldsymbol{\beta}}^{\text{MV}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$*;*
- *l'estimateur du MV de* σ^2 *vaut* $\hat{\sigma}^2_{\text{MV}} = \frac{\|\hat{\boldsymbol{\epsilon}}\|^2}{n}$*.*

Sous les hypothèses $\mathcal{H}_1 - \mathcal{H}_2^{\text{Gauss}}$*, démontrez alors que*

1. $\hat{\boldsymbol{\beta}}$ *est un vecteur gaussien centré en* $\boldsymbol{\beta}$ *et de variance* $\sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}$*.*
2. $(n-p)\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-p}$*.*
3. $\hat{\boldsymbol{\beta}}$ *et* $\hat{\sigma}^2$ *sont indépendants.*

Exercice 40 (Loi de T_j et $\hat{\boldsymbol{\beta}}$ normalisé - Modèle linéaire gaussien)**.**
Sous les hypothèses $\mathcal{H}_1$ *et* $\mathcal{H}_2^{\text{Gauss}}$*, on a*

1. *Pour* $j = 1, \dots, p$

$$T_j = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}} \sim T_{n-p} \qquad \text{où } \hat{\sigma}_{\hat{\beta}_j} = \hat{\sigma}\sqrt{(\mathbf{X}^\top\mathbf{X})^{-1}_{jj}}.$$

2. *Soit* $\mathbf{R}$ *une matrice de taille* (q, p) *(avec* $q \leq p$*) alors*

$$\frac{1}{q\hat{\sigma}^2}(\mathbf{R}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}))^\top(\mathbf{R}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{R}^\top)^{-1}\mathbf{R}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim \mathcal{F}_{q,n-p}.$$

Exercice 41 (Intervalle de prévision - Modèle linéaire gaussien)**.**
*Montrez qu'un "*IC *pour* Y_{n+1}*" au niveau* $1-\alpha$ *est donné par*

$$\left[\mathbf{x}'_{n+1}\hat{\boldsymbol{\beta}} \pm t_{1-\alpha/2,n-p}\hat{\sigma}\sqrt{1 + \mathbf{x}'_{n+1}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}'^{\top}_{n+1}}\right]$$

Exercice 42 (Statistique de Fisher - Modèle linéaire gaussien)**.**
Sous les hypothèses $\mathcal{H}_1 - \mathcal{H}_2^{\text{Gauss}}$*, soit* $Q \subset \{1, \ldots, p\}$ *de dimension* $1 \leq q < p$*,* $\mathbf{X}_Q = (\mathbf{x}_j, j \in Q)$*,* $\boldsymbol{\beta}_Q = (\beta_j)_{j \in Q}$ *et soit* $p_0 = p - q$*. Pour tester les hypothèses*

$$H_0 : \boldsymbol{\beta}_Q = \mathbf{0} \Leftrightarrow \mathrm{E}(\mathbf{Y}) \in \mathcal{V}(\mathbf{X}_{Q^c}) \quad \textit{contre} \quad H_1 : \exists j \in Q, \beta_j \neq 0 \Leftrightarrow \mathrm{E}(\mathbf{Y}) \in \mathcal{V}(\mathbf{X})$$

montrez que la statistique F *sous* H_0 *satisfait*

$$F = \frac{\|\hat{\mathbf{Y}}_0 - \hat{\mathbf{Y}}\|^2/(p - p_0)}{\|\mathbf{Y} - \hat{\mathbf{Y}}\|^2/(n - p)} \sim F_{p-p_0, n-p}.$$

Exercice 43 (Loi asymptotique de $\hat{\boldsymbol{\beta}}$ et $\hat{\sigma}^2$)**.**
En s'appuyant sur le fait que $\hat{\boldsymbol{\beta}}$ *converge vers une loi normale, montrez que sous les hypothèses* $\mathcal{H}_1$*,* $\mathcal{H}_2$ *et* $\mathcal{H}_3^{\text{tcl}}$*, alors lorsque* $n \to \infty$

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{d} \mathcal{N}(0, \sigma^2 \mathbf{Q}^{-1})$$

De plus $\hat{\sigma}^2 \to \sigma^2$ *en probabilité, ce qui permet d'avoir*

$$\hat{\sigma}^{-1}\mathbf{Q}^{1/2}\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{d} \mathcal{N}(0, \mathbf{I}_p) \textit{ ou encore } \hat{\sigma}^{-1}(\mathbf{X}^\top\mathbf{X})^{1/2}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{d} \mathcal{N}(0, \mathbf{I}_p).$$

Exercice 44 (Procédure bootstrap)**.**
Proposez une procédure bootstrap pour

- *construire un intervalle de prédiction pour une future observation;*
- *construire une région de confiance pour* $\boldsymbol{\beta}$*;*
- *pour faire un test d'hypothèses de Fisher ou un test sur* $\lambda^\top\boldsymbol{\beta}$*,* $\lambda \in \mathbb{R}^p$*.*

Exercice 45 (Autour du jeu de données `ozone`)**.**
On considère le jeu de données ***`ozone`*** *(source, Cornillon et Matzner-Løber, 2009) du fichier* ***`ozone.RData`*** *(cf devoir pour une description des variables). Considérons le modèle*

```
O3 ~ T12
```

Les questions suivantes sont à traiter en **`R`***. Les niveaux de confiance et risque d'erreur de première expèce sont fixés à* $1 - \alpha$ *et* α *pour* $\alpha = 5\%$ *respectivement.*

1. *Ecrivez mathématiquement ce modèle et estimez le modèle par MCO. Calculez $\hat{\Sigma}_{\hat{\boldsymbol{\beta}}}$ la matrice estimée de la variance de $\hat{\boldsymbol{\beta}}$.*

2. *On supposera pour commencer l'hypothèse $\mathcal{H}_2^{\text{Gauss}}$. Quel test d'hypothèses permet de montrer que ce modèle est globalement significatif? Mettez en place ce test.*

3. *Montrez que la variable `T12` est significative? Pourrait-on montrer que $\beta_2 > 2$? (que cela signifierait-il?) Construire également un intervalle de confiance du paramètre β_2 (en utilisant la formule du cours, puis la fonction interne confint).*

4. *Reconsidérez la question précédente sans supposer l'hypothèse $\mathcal{H}_2^{\text{Gauss}}$ (en mettant en place une méthode de rééchantillonnage).*

5. *Considérez maintenant le modèle*

```
O3 ~ T12 + Vx + Ne12
```

   Ce modèle est-il globalement significatif? Vérifiez la relation liant la statistique du test de Fisher global au coefficient de détermination multiple au carré.

6. *Si l'hypothèse de normalité ne vous semble pas aberrante. Tracez les ellipses de confiance des paramètres (β_2, β_3), (β_2, β_4) et (β_3, β_4) (cf cours, il faut installer le paquet `ellipse`)*

7. *Quel niveau d'ozone un jour où `T12`=20, `Ne12`=5 et `Vx`=0 peut-on prévoir? Construire l'intervalle de prédiction associé.*

8. *Effectuez une analyse des résidus, des points leviers et points influents.*

Exercice 46 (Variance lorsque les régresseurs sont orthogonaux)**.**
On travaille ici avec le modèle linéaire homoscédastique $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ satisfaisant les hypothèses $\mathcal{H}_1 - \mathcal{H}_2$ du cours, et soit $\hat{\boldsymbol{\beta}}$ l'estimateur des MCO:

1. *Si les covariables sont orthogonales, rappelez ce que vaut $\text{Var}(\hat{\beta}_j)$ pour $j = 1, \ldots, p$.*

2. *Supposons $p = 2$, montrer que $\text{Var}(\hat{\beta}_1) \geq \frac{\sigma^2}{\mathbf{x}_1^\top \mathbf{x}_1}$ (indication: calculer le 1er terme diagonal de $(\mathbf{X}^\top \mathbf{X})^{-1}$).*

Exercice 47 (Estimation sous contraintes)**.**
Considérons le modèle homoscédastique $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ *satisfaisant les hypothèses* $\mathcal{H}_1 - \mathcal{H}_2$*. L'objectif est d'estimer* $\boldsymbol{\beta}$ *sous la contrainte que* $\mathbf{R}\boldsymbol{\beta} = \mathbf{b}$ *où* $\mathbf{R}$ *est une matrice de rang* $r \leq p$ *et* $\mathbf{b} \in \mathbb{R}^r$*.*

1. *Considérons un exemple où* $p = 3$ *et où la contrainte est* $\beta_2 + \beta_3 = 1$*. Montrer qu'il existe* $\gamma \in \mathbb{R}^2$ *tel que le modèle se réécrive* $\mathbf{Y}' = \mathbf{X}'\gamma + \boldsymbol{\varepsilon}$*, où* $\mathbf{X}'$ *est maintenant une matrice de taille* $(n, 2)$*. Estimer alors* γ *par MCO puis en déduire* $\hat{\boldsymbol{\beta}}$*.*

2. *Nous disposons pour* n *entreprises de leur valeur du capital* K_i *de l'emploi* L_i *et de la valeur ajoutée* V_i*. Pour ces entreprises, nous supposerons qu'elle sont régies par un modèle de Cobb-Douglas à rendement d'échelle constant.*

$$V_i = \lambda K_i^{\alpha_1} L_i^{\alpha_2} \eta_i$$

où $\lambda > 0$*,* $\alpha_1, \alpha_2 \in \mathbb{R}$ *tels que* $\alpha_1 + \alpha_2 = 1$ *et où* η_i *est une variable aléatoire de moyenne 1. Linéarisez ce modèle pour le ramener au précédent. Quelle hypothèse doit-on faire sur les variables* η_i*? Estimez* λ, α_1 *et* α_2 *à partir du jeu de données du fichier* `cobb.RData` *regroupant les valeurs de capitaux, d'emploi et de valeur ajoutée de 10 entreprises.*

3. *D'un point de vue général, nous pouvons évidemment résoudre le problème d'optimisation sous contraintes, et on peut montrer (la preuve est laissée en exercice si vous êtes curieux) que l'estimateur sous contraintes noté* $\hat{\boldsymbol{\beta}}^C$ *est une transformation linéaire de* $\hat{\boldsymbol{\beta}}$ *l'estimateur par MCO obtenu sans contrainte. En effet, on a*

$$\hat{\boldsymbol{\beta}}^C = \hat{\boldsymbol{\beta}} + (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{R}^\top \left(\mathbf{R}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{R}^\top\right)^{-1} (r - \mathbf{R}\hat{\boldsymbol{\beta}})$$

Vérifiez sur l'exemple précédent.

4. *Qu'aurait donné l'estimation de* α_1 *et* α_2 *si on avait en plus ajouté la contrainte* $\lambda = 1$*?*

Exercice 48 (Lois de vecteurs projetés)**.**
Soit $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ *un modèle linéaire homoscédastique tel que* $\mathbf{X}$ *est une matrice de taille* (n, p) $(n > p)$ *de rang plein et tel que* $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2\mathbf{I}_n)$ *avec* $\sigma = 1$*. Pour* $I \subset \{1, \ldots, p\}$*, on note* $\mathcal{P}_I$ *et* $\mathcal{P}_{I^\perp}$ *les matrices de projection sur* $\mathcal{V}(\mathbf{x}_j, j \in I)$ *et* $\mathcal{V}(\mathbf{x}_j, j \in I)^\perp$ *respectivement. Enfin on notera* $I^c = \{1, \ldots, p\} \setminus I$*. Les deux questions sont indépendantes.*

1. *Montrez que si toutes les covariables sont orthonormées alors pour tout* $j = 1, \dots, p$

$$\|\mathcal{P}_{\{j\}}(\mathbf{X}\boldsymbol{\beta})\|^2 = \beta_j^2.$$

Déterminez alors la loi de $\|\mathcal{P}_{\{j\}}\mathbf{Y} - \beta_j \mathbf{x}_j\|^2$.

2. *On fixe* $p = 10$. *Soient* $\mathbf{U}$ *et* $\mathbf{V}$ *les deux vecteurs définis par* $\mathbf{U} = \mathcal{P}_{\{1,2\}}\mathcal{P}_{\{3,4\}}\boldsymbol{\varepsilon}$ *et* $\mathbf{V} = \mathcal{P}_{\{3,4\}^\perp}\boldsymbol{\varepsilon}$. *Déterminez la loi de* $\|\mathbf{U}\|^2$ *et* $\|\mathbf{V}\|^2$. *Les vecteurs* $\mathbf{U}$ *et* $\mathbf{V}$ *sont-ils indépendants?*

Exercice 49. 1. *Considérons les deux modèles linéaires suivants:*

$$\begin{aligned} M_1 : &\quad Y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i \\ M_2 : &\quad Y_i = \mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\gamma} + \eta_i \end{aligned}$$

où pour $i = 1, \dots, n$ $\mathbf{x}_i', \mathbf{z}_i'$ *sont des vecteurs lignes de de dimension* p *et* q *respectivement, où* $\boldsymbol{\beta} \in \mathbb{R}^p$ *et* $\boldsymbol{\gamma} \in \mathbb{R}^q$ $(p, q \geq 1)$ *et où les erreurs aléatoires* ε_i, η_i *sont non corrélées et de variance* σ^2. *Soient* $\hat{\boldsymbol{\beta}}_1$ *et* $\hat{\boldsymbol{\beta}}_2$ *les estimateurs de* $\boldsymbol{\beta}$ *par la méthode des moindres carrés issus des modèles 1 et 2 respectivement. Enfin, soit* $\mathbf{c} \in \mathbb{R}^p$. *Même si aucun des deux modèles n'est vrai, montrez que*

$$\mathrm{Var}(\mathbf{c}^\top \hat{\boldsymbol{\beta}}_1) \leq \mathrm{Var}(\mathbf{c}^\top \hat{\boldsymbol{\beta}}_2)$$

2. *Soient* $\hat{Y}_i^1$ *et* $\hat{Y}_i^2$ *les valeurs prédites par les modèles 1 et 2 respectivement. Montrez que* $\mathrm{Var}(\hat{Y}_i^1) \leq \mathrm{Var}(\hat{Y}_i^2)$.

Indication*: on pourra éventuellement mais non nécessairement utiliser le fait que pour une matrice par blocs sous la forme* $\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}$ *bien définie, et inversible alors le bloc supérieur gauche de l'inverse est donné par* $\mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{B}\left(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}\right)^{-1}\mathbf{C}\mathbf{A}^{-1}$.

6 Analyse des résidus et validation du modèle linéaire

Exercice 50 (Résidus studentisés par validation croisée)**.**
Sous $\mathcal{H}_1$ et si la suppression de la $i$ème ligne ne modifie pas le rang de $\mathbf{X}$, on définit

$$t_{(i)} = \frac{\hat{\varepsilon}_{(i)}}{\hat{\sigma}_{(i)}\sqrt{1 + \mathbf{x}_i'(\mathbf{X}_{(i)}^\top \mathbf{X}_{(i)})^{-1}\mathbf{x}_i}} = \frac{Y_i - \hat{Y}_{(i)}}{\hat{\sigma}_{(i)}\sqrt{1 + \mathbf{x}_i'(\mathbf{X}_{(i)}^\top \mathbf{X}_{(i)})^{-1}\mathbf{x}_i'^\top}}$$

où $\hat{Y}_{(i)} = \mathbf{x}_{(i)}'\hat{\boldsymbol{\beta}}_{(i)}$. Sous $\mathcal{H}_2$[Gauss], montrez que $t_{(i)} \sim T_{n-1-p}$ (💡).

Exercice 51 (Propriétés de $\mathbf{H}$)**.**
Soit $\mathbf{H}$ la matrice de projection d'un modèle linéaire. Montrez que

1. *$h_{ii} \in [0, 1]$, $h_{ij} \in [-1/2, 1/2]$ pour $j \neq i$;*
2. *si $h_{ii} \in \{0, 1\}$ alors $h_{ij} = 0$ pour $j \neq i$;*
3. *$\mathrm{tr}(\mathbf{H}) = \sum h_{ii} = p$.*

Exercice 52 (Formulation alternative - $\hat{Y}_{(i)}$ et C_i)**.**
Soit $\mathbf{X}$ la matrice de taille (n, p) de design d'un modèle linéaire homoscédastique. Rappelons les notations: $\mathbf{x}_i'$ désigne le vecteur ligne de taille $(1, p)$ correspondant à la $i$ème ligne de $\mathbf{X}$ et soit $\mathbf{X}_{(i)}$ la matrice $\mathbf{X}$ privée de sa $i$ème ligne.

1. *Montrez que $\mathbf{X}^\top\mathbf{X} = \mathbf{X}_{(i)}^\top\mathbf{X}_{(i)} + \mathbf{x}_i'^\top\mathbf{x}_i'$, puis que $\mathbf{X}_{(i)}^\top\mathbf{Y}_{(i)} = \mathbf{X}^\top\mathbf{Y} - \mathbf{x}_i'Y_i$, où $\mathbf{Y}_{(i)}$ est le vecteur $\mathbf{Y}$ privé de sa $i$ème ligne. Et enfin en vous servant de l'indication en fin de question que*

$$(\mathbf{X}_{(i)}^\top\mathbf{X}_{(i)})^{-1} = (\mathbf{X}^\top\mathbf{X})^{-1} + \frac{1}{1 - h_{ii}}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_i'^\top\mathbf{x}_i'(\mathbf{X}^\top\mathbf{X})^{-1}$$

où on rappelle que h_{ii} est le $i$ème terme diagonal de la matrice de projection sur $\mathcal{V}(\mathbf{X})$.

2. *Soit $\hat{Y}_{(i)}$ la prévision de la variable réponse pour le $i$ème individu basé sur le modèle estimé sans le $i$ème individu. En utilisant le résultat de la question précédente, montrez que*

$$\hat{Y}_{(i)} = \frac{1}{1-h_{ii}}\hat{Y}_i - \frac{h_{ii}}{1-h_{ii}}Y_i. \tag{5}$$

3. *Comment interprèteriez-vous le résultat* (5) *en fonction de la valeur de h_{ii}.*

4. *Montrez également que*

$$\hat{\boldsymbol{\beta}}_{(i)} = \hat{\boldsymbol{\beta}} - \frac{1}{1-h_{ii}}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_i'^\top(Y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}).$$

5. *En déduire que la distance de Cook, définie par*

$$C_i = \frac{1}{p\hat{\sigma}^2}(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})^\top(\mathbf{X}^\top\mathbf{X})(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})$$

sécrit sous la forme

$$C_i = \frac{1}{p}\,\frac{h_{ii}}{1-h_{ii}}\,\frac{\hat{\varepsilon}_i^2}{\hat{\sigma}^2(1-h_{ii})}.$$

Indication (lemme d'inversion matricielle)*: soient* $\mathbf{M}$ *une matrice symétrique inversible de taille* (p,p) *et* $\mathbf{u},\mathbf{v}$ *deux vecteurs de taille* p*. Alors*

$$(\mathbf{M}+\mathbf{u}\mathbf{v}^\top)^{-1} = \mathbf{M}^{-1} - \frac{\mathbf{M}^{-1}\mathbf{u}\mathbf{v}^\top\mathbf{M}^{-1}}{1+\mathbf{u}^\top\mathbf{M}^{-1}\mathbf{v}}.$$

Exercice 53 (Impact d'une variable)**.**
Sans rappeler les notations, considérons les deux modèles suivants (modlèe linéaire standard et projeté sur l'orthogonal de $\mathcal{V}(\mathbf{X}^{(j)})$):

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \mathbf{X}^{(j)}\boldsymbol{\beta}^{(j)} + \mathbf{x}_j\beta_j + \boldsymbol{\varepsilon} \tag{6}$$

$$\mathcal{P}_{\mathbf{X}^{(j)\perp}}\mathbf{Y} = \mathcal{P}_{\mathbf{X}^{(j)\perp}}\mathbf{X}^{(j)}\beta^{(j)} + \beta_j\mathcal{P}_{\mathbf{X}^{(j)\perp}}\mathbf{x}_j + \mathcal{P}_{\mathbf{X}^{(j)\perp}}\boldsymbol{\varepsilon}$$

$$\mathcal{P}_{\mathbf{X}^{(j)\perp}}\mathbf{Y} = \beta_j\mathcal{P}_{\mathbf{X}^{(j)\perp}}\mathbf{x}_j + \boldsymbol{\eta} \tag{7}$$

Montrez que le coefficient estimé $(\hat{\boldsymbol{\beta}})_j$ provenant de (8) *est égal au coefficient estimé $\hat{\beta}_j$ provenant de* (9)*.*

Exercice 54 (Exercice de synthèse - Jeu de données `internet`)**.** *Cette question s'appuie sur le jeu de données* ***`internet`****. L'objectif est d'étudier les facteurs influençant l'utilisation mondiale d'internet (via le nombre d'internautes). On dispose de l'information pour 37 pays et des variables:*

- ***`int`****: nombre d'internautes.*
- ***`ord`****: nombre d'ordinateurs individuels.*
- ***`pop`****: taille de la population.*
- ***`pib`****: produit intérieur brut.*

```
> load('../R/internet.RData');internet[1:3,]
        pays      int       pop      ord  pib
1 etats-unis 123326.0 278058.88 49896.20 7746
2      japon  63955.2 126771.66  7485.78 4202
3  allemagne  28876.9  83029.54  3914.06 2100
```

Pour information, l'ensemble du jeu de données se trouve en fin de question. Les variables seront toutes transformées logarithmiquement par la suite pour gommer des effets d'échelle dûs à l'hétérogénéité des différents pays étudiés.

1. *Dans un premier temps, on regarde l'influence de la variable log(pib) sur la variable log(int). Écrivez ce modèle mathématiquement ainsi que l'hypothèse permettant au moins de pouvoir calculer l'estimateur par moindres carrés ordinaires.*

2. *Calculez l'estimation par MCO $\hat{\boldsymbol{\beta}}$ de $\boldsymbol{\beta}$ en vous servant des indications suivantes:*

```
> with(internet,cov(log(int),log(pib)))

[1] 1.458265

> with(internet,c(mean(log(int)),mean(log(pib))))

[1] 8.603284 5.689294

> with(internet,c(var(log(int)),var(log(pib))))

[1] 1.784450 1.405787

> n=nrow(internet);n

[1] 37
```

3. *Calculez le coefficient de détermination multiple du modèle précédent.*

4. *Etant donnée la taille d'échantillon et le fait que* ***qnorm****(0.975)*$\simeq 1.96$ *et que* $\hat{\sigma}_{\hat{\beta}_2} \simeq 0.07$*, la variable log(pib) vous semble-t-elle être significative au seuil de 5%?*

5. *On décide de rajouter au modèle la variable log(ord) (vous tenterez d'expliquer plus tard pourquoi log(pop) n'a pas été incluse). Voilà la sortie résumée:*

```
> out.lm= lm(log(int)~log(ord)+log(pib),data=internet)
> summary(out.lm)


Call:
lm(formula = log(int) ~ log(ord) + log(pib), data = internet)

Residuals:
    Min      1Q  Median      3Q     Max
-0.9242 -0.2619 -0.1007  0.2438  1.0186

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.58132    0.39136   6.596 1.47e-07 ***
log(ord)     0.23515    0.07827   3.005  0.00497 **
log(pib)     0.79690    0.10438   7.634 7.15e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4768 on 34 degrees of freedom
Multiple R-squared:  0.8797,Adjusted R-squared:  0.8726
F-statistic: 124.3 on 2 and 34 DF,  p-value: 2.327e-16
```

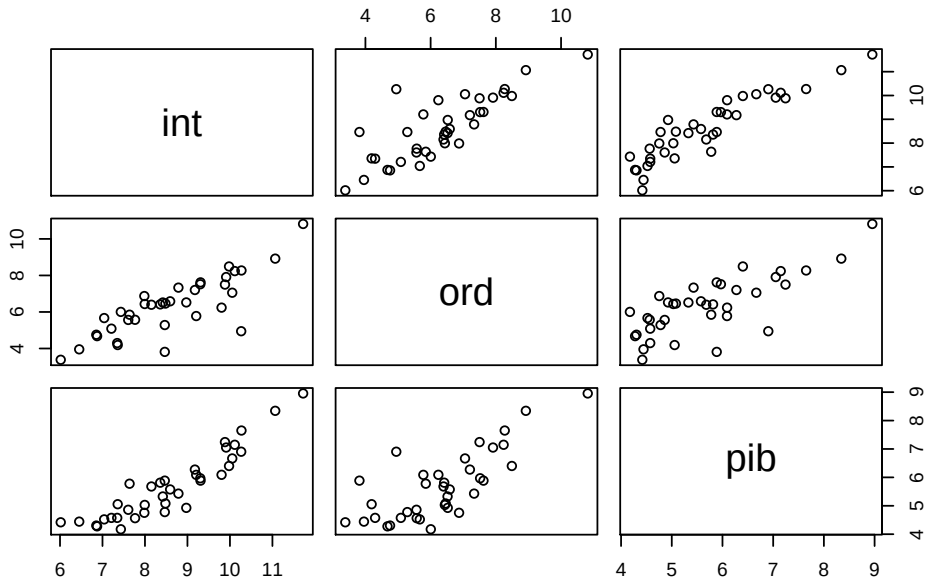
Écrivez mathématiquement ce modèle. Testez séparément et au seuil de 5% les hypothèses suivantes: $H_1 : \beta_2 > 0$, $H_1 : \beta_3 > 0.4$. *A l'aide des informations ci-dessous testez également* $H_1 : \beta_3 > \beta_2$ *(détaillez les étapes).*

```
> round(vcov(out.lm),3)

            (Intercept) log(ord) log(pib)
(Intercept)       0.153   -0.003   -0.022
log(ord)         -0.003    0.006   -0.006
log(pib)         -0.022   -0.006    0.011

> qt(0.95,n-3)

[1] 1.690924
```

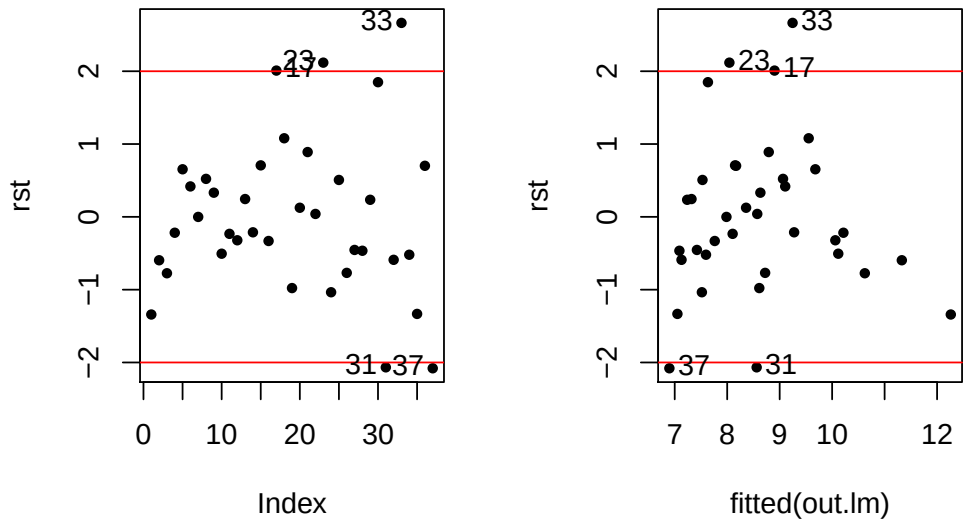
6. *Faites une synthèse de* **15-20 lignes maximum** *de la sortie résumée précédente et des sorties présentées en annexe. Éventuellement, vous pouvez suggérer d'autres analyses que vous seriez tentés de faire.*

Annexes (sorties R) - Jeu de données `internet`

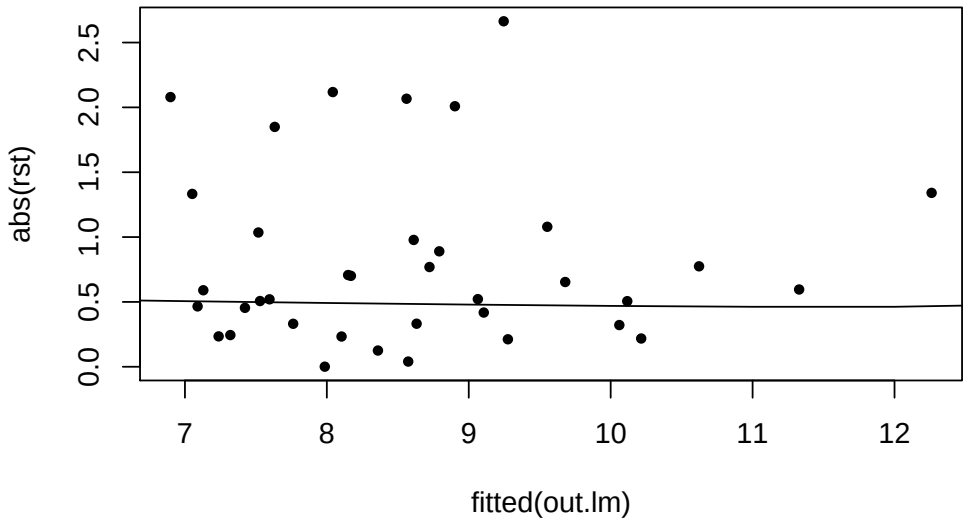
```
> plot(log(internet[-c(1,3)]))
```



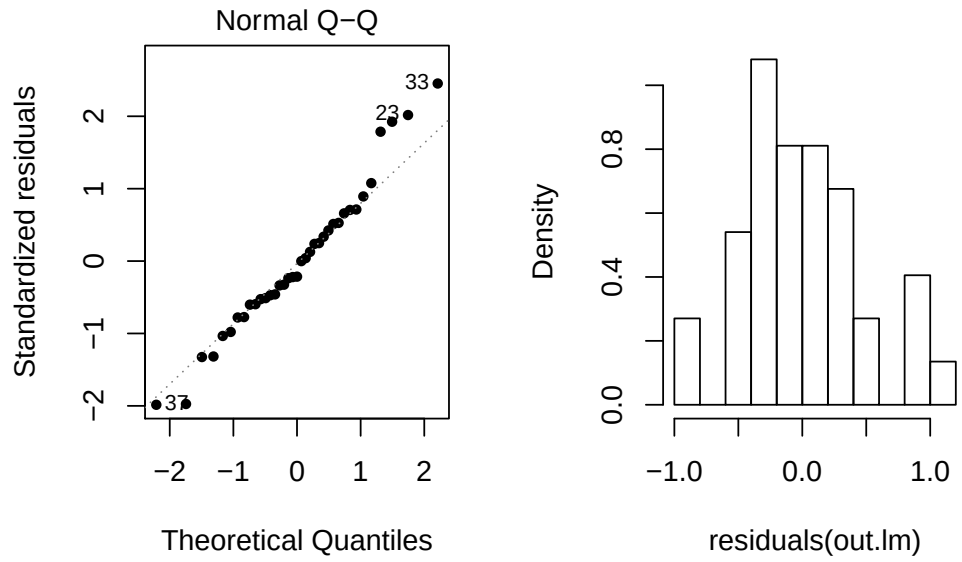

```
> cor(log(internet[-1]))
          int        pop        ord       pib
int 1.0000000 0.41353001 0.82059928 0.9207138
pop 0.4135300 1.00000000 0.08709955 0.5730704
ord 0.8205993 0.08709955 1.00000000 0.7666041
pib 0.9207138 0.57307038 0.76660408 1.0000000
> require(ppcor);pcor(log(internet[-1]))$estimate
            int         pop        ord       pib
int  1.00000000 -0.07602491  0.3158625 0.6532911
pop -0.07602491  1.00000000 -0.6092173 0.6471739
ord  0.31586246 -0.60921733  1.0000000 0.4242685
pib  0.65329112  0.64717386  0.4242685 1.0000000
> require(car);par(mfrow=c(1,2))
> rst=rstudent(out.lm)
> plot(rst,pch=20);abline(h=c(-2,2),col="red")
> tmp=showLabels(1:n,rst,method=which(abs(rst)>2))
> plot(fitted(out.lm),rst,pch=20);abline(h=c(-2,2),col="red")
> tmp=showLabels(fitted(out.lm),rst,method=which(abs(rst)>2))
```



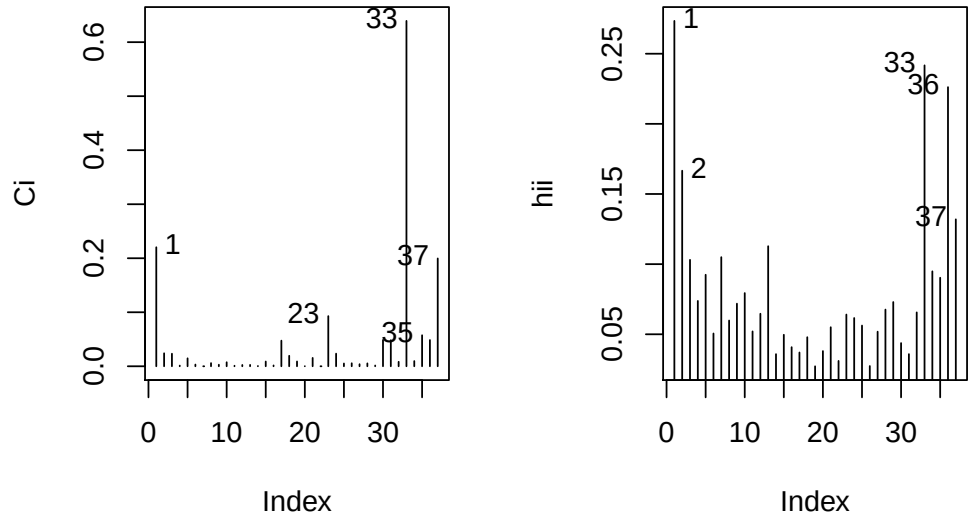

```
> par(mfrow=c(1,1))
> plot(fitted(out.lm),abs(rst),pch=20);lines(lowess(abs(rst)))
```



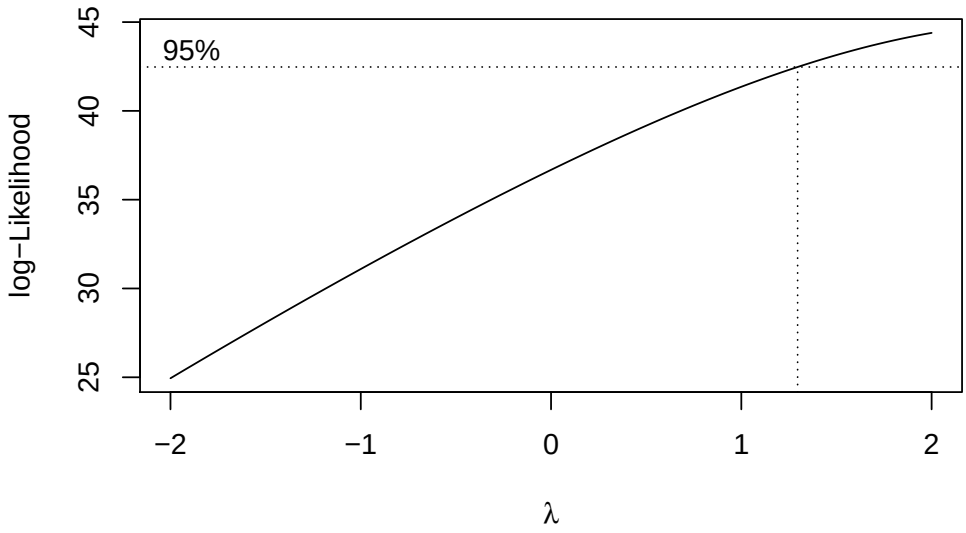
```
> par(mfrow=c(1,2))
> plot(out.lm,which=2,pch=20)
> hist(residuals(out.lm),prob=TRUE,main="")
```




```
> par(mfrow=c(1,2))
> Ci=cooks.distance(out.lm);plot(Ci,type="h")
> tmp=showLabels(1:n,Ci,n=5,method="r")
> hii=hatvalues(out.lm);plot(hii,type="h")
> tmp=showLabels(1:n,hii,n=5,method="r")
```



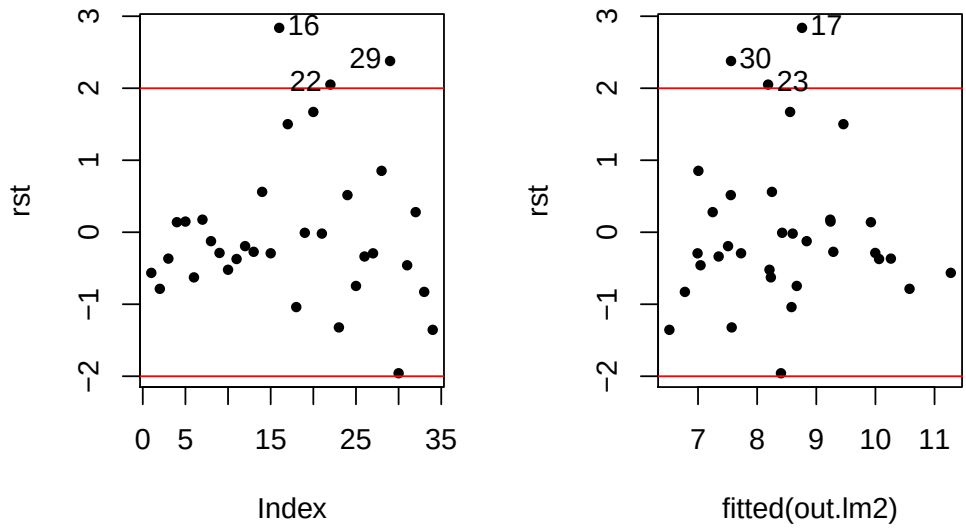

```
> p=3
> c(2*p/n,3*p/n,.5,qf(c(.1,.5),p,n-p))
[1] 0.1621622 0.2432432 0.5000000 0.1936416 0.8047134
> require(MASS)
> par(mfrow=c(1,1));boxcox(out.lm);par(mfrow=c(1,2))
```



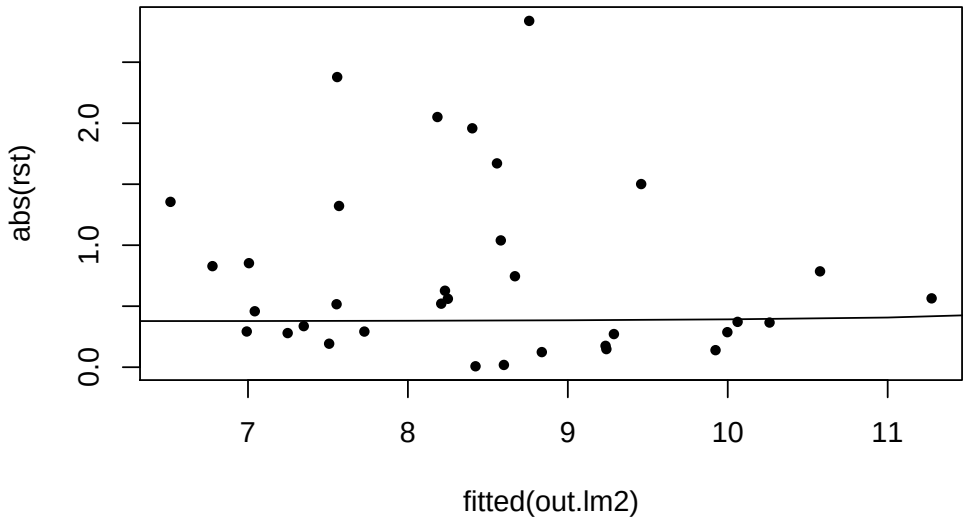

```
> internet[c(1,33,36),]
         pays       int       pop       ord  pib
1  etats-unis 123326.00  278058.9 49896.200 7746
33      chine  28697.20 1280775.5   140.599  996
36       inde   4748.76 1029991.2    45.420  360
```

```
> internet2=internet[-c(1,33,36),]
> out.lm2=lm(log(int)~log(ord)+log(pib),data=internet2)
> summary(out.lm2)

Call:
lm(formula = log(int) ~ log(ord) + log(pib), data = internet2)

Residuals:
     Min       1Q   Median       3Q      Max
-0.76522 -0.21321 -0.11398  0.09989  1.04450

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.39539    0.38912   6.156 7.88e-07 ***
log(ord)     0.43436    0.09153   4.745 4.45e-05 ***
log(pib)     0.59983    0.11690   5.131 1.48e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4197 on 31 degrees of freedom
Multiple R-squared:  0.8932,Adjusted R-squared:  0.8863
F-statistic: 129.6 on 2 and 31 DF,  p-value: 8.826e-16
> n2=nrow(internet2);par(mfrow=c(1,2))
```

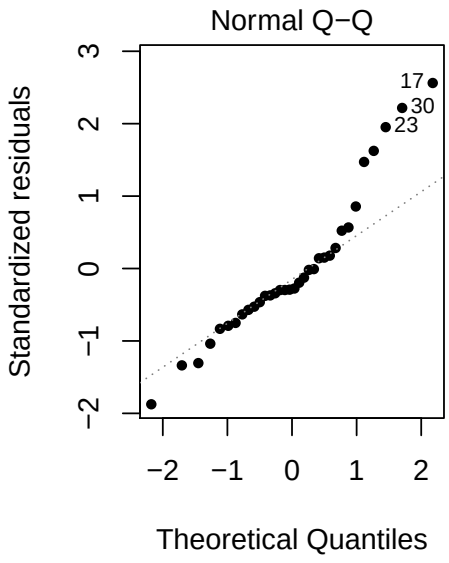
```
> rst=rstudent(out.lm2)
> plot(rst,pch=20);abline(h=c(-2,2),col="red")
> tmp=showLabels(1:n2,rst,method=which(abs(rst)>2))
> plot(fitted(out.lm2),rst,pch=20);abline(h=c(-2,2),col="red")
> tmp=showLabels(fitted(out.lm2),rst,method=which(abs(rst)>2))
```

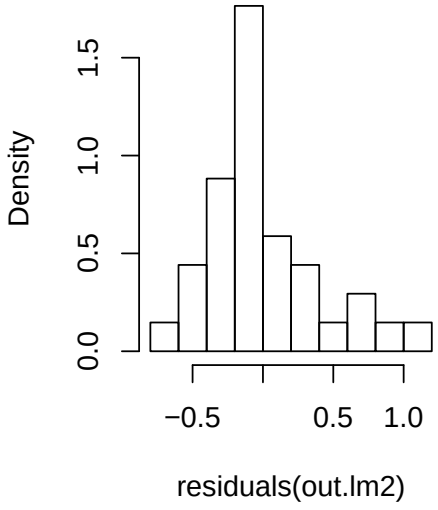



```
> par(mfrow=c(1,1))
> plot(fitted(out.lm2),abs(rst),pch=20);lines(lowess(abs(rst)))
```

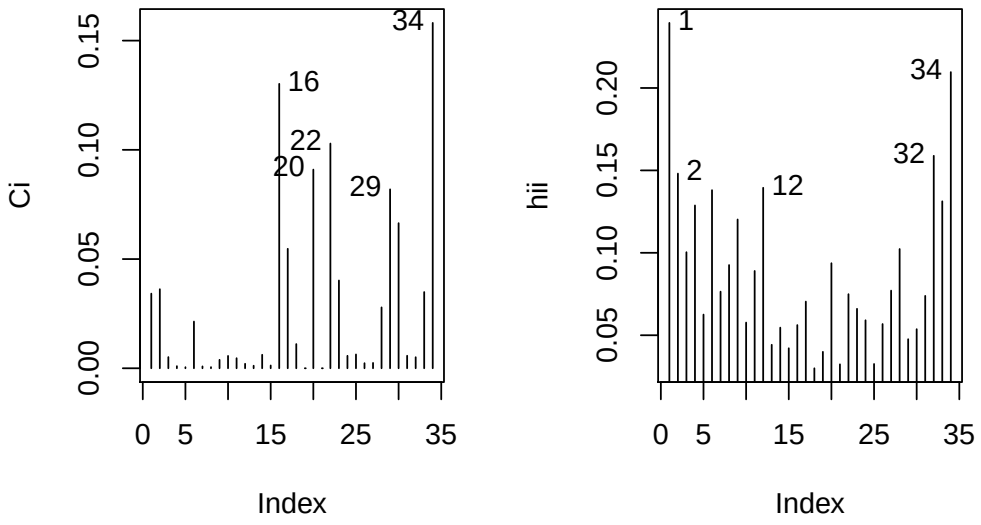



```
> par(mfrow=c(1,2))
> plot(out.lm2,which=2,pch=20)
> hist(residuals(out.lm2),prob=TRUE,main="")
```

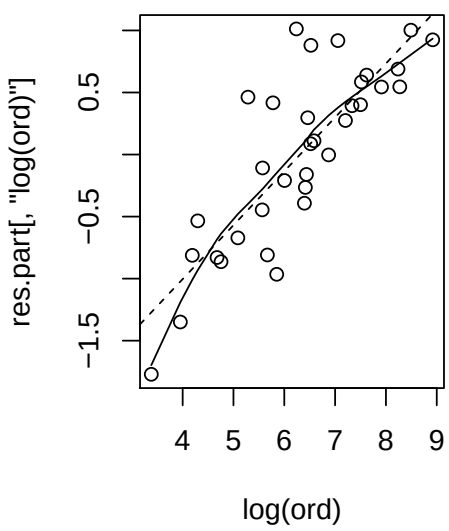


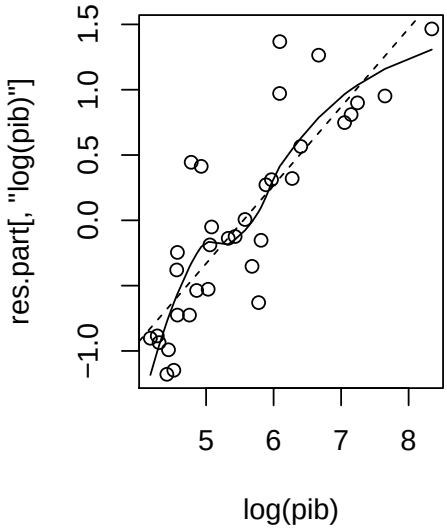



```
> par(mfrow=c(1,2))
> Ci=cooks.distance(out.lm2);plot(Ci,type="h")
> tmp=showLabels(1:n2,Ci,n=5,method="r")
> hii=hatvalues(out.lm2);plot(hii,type="h")
> tmp=showLabels(1:n2,hii,n=5,method="r")
```




```
> #
> res.part=resid(out.lm2,type="partial")
> lisseur= with(internet2,loess(res.part[,"log(ord)"]~log(ord)))
> ordre=with(internet2,order(log(ord)))
> with(internet2,plot(log(ord),res.part[,"log(ord)"]))
> with(internet2,matlines(log(ord)[ordre],predict(lisseur)[ordre]))
> abline(lm(res.part[,"log(ord)"]~log(ord),data=internet2),lty=2)
> #
> lisseur= with(internet2,loess(res.part[,"log(pib)"]~log(pib)))
> ordre=with(internet2,order(log(pib)))
> with(internet2,plot(log(pib),res.part[,"log(pib)"]))
> with(internet2,matlines(log(pib)[ordre],predict(lisseur)[ordre]))
> abline(lm(res.part[,"log(pib)"]~log(pib),data=internet2),lty=2)
```


res.part[, "log(ord)"]
−1.5
−0.5
0.5
4
5
6
7
8
9
log(ord)


res.part[, "log(pib)"]
−1.0
0.0
0.5
1.0
1.5
5
6
7
8
log(pib)

7 Introduction de variables qualitatives dans un modèle linéaire

Exercice 55 (Régression linéaire simple et variable qualitative)**.**
Soit un modèle de régression linéaire cherchant à expliquer une variable continue Y, par une variable quantitative et une variable qualitative possédant I modalités. Les références aux modèles ci-dessous sont celles associées au cours sur l'introduction de variables qualitatives dans un modèle de régression linéaire.

1. *Comment s'écrit la matrice de design $\mathbf{X}$ pour l'équivalent des modèles (2), (3), (4) du cours.*
2. *Calculer l'estimateur des MCO obtenu pour le modèle (2).*
3. *Montrez que cet estimateur peut être obtenu en faisant I régression linéaires simples et ce quelle que soit l'hypothèse faite sur la variance des fluctuations aléatoires de chacune des variables.*

Exercice 56 (Variable qualitative - Jeu de données `mtcars`)**.**
On considère ici le jeu de données interne `mtcars`*. Les références aux modèles ci-dessous sont celles associées au cours sur l'introduction de variables qualitatives dans un modèle de régression linéaire.*

1. *Cherchons à expliquer la variable* `mpg` *en fonction des variables* `wt` *et* `cyl`*. La variable* `cyl` *décrit le nombre de cylindrés 4,6 ou 8 de la voiture. Il est donc naturel de la considérer comme une variable indicatrice. Transformez la classe de* `cyl` *afin qu'elle soit reconnue comme un facteur. Tracez le jeu de données en identifiant les observations par leur cylindrée. Qu'observe-t-on?*
2. *Considérez l'équivalent des trois modèles principaux (2), (3) et (4) vus en cours, estimez ceux-ci et tracez les droites estimées sur le même graphique. Quels sont vos commentaires?*

3. *Mettez en place un test de Fisher pour comparer les modèles (3) à (2), (4) à (2).*

4. *Introduisons maintenant la variable* **`hp`**. *Tracez le jeu de données des trois variables* **`mpg,wt,hp`** *en fonction de leur cylindrée. Comment s'écrit mathématiquement un modèle linéaire avec un effet moyen différent par cylindrée, une interaction entre* **`cyl`** *et* **`wt`** *et interaction entre* **`cyl`** *et* **`hp`**. *Comparez ce modèle au précédent (en l'absence de* **`hp`** *donc).*

5. *Considérer le test d'hypothèses précédent en utilisant une technique de rééchantillonnage.*

Exercice 57 (Variable qualitative - Jeu de données `iris`)**.**
On considère le jeu de données **`iris`** *interne à* **`R`**. *Nous nous focalisons sur les variables* **`width`**, **`length`** *et* **`species`** *qui donnent pour trois variétés de plantes (la variable* **`species`** *a trois modalités: versicolor, virginica et setosa), les longueurs et largeurs de pétales de 150 iris. Les instructions relatives à cette question se trouvent juste après celle-ci.*

1. *Les instructions* **`lm1,lm2`** *et* **`lm3`** *envisagent trois modèles différents. Écrivez précisément et mathématiquement ces trois modèles.*

2. *Comment faut-il interpréter graphiquement ces trois modèles?*

3. *En analysant brièvement les sorties* **`R`**, *quel modèle choisiriez-vous? Justifiez votre réponse.*

```
> Iris=iris[c(3,4,5)]
> head(Iris)
  Petal.Length Petal.Width Species
1          1.4         0.2  setosa
2          1.4         0.2  setosa
3          1.3         0.2  setosa
4          1.5         0.2  setosa
5          1.4         0.2  setosa
6          1.7         0.4  setosa
> length=Iris$Petal.Length;width=Iris$Petal.Width; species=Iris$Species
> unique(species)
[1] setosa     versicolor virginica
Levels: setosa versicolor virginica
> lm1=lm(width~length)
> lm2=lm(width~length+species)
> lm3=lm(width~length*species)
> ## summary outputs
> summary(lm1)
```

```
Call:
lm(formula = width ~ length)

Residuals:
     Min       1Q   Median       3Q      Max
-0.56515 -0.12358 -0.01898  0.13288  0.64272

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.363076   0.039762  -9.131  4.7e-16 ***
length       0.415755   0.009582  43.387  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2065 on 148 degrees of freedom
Multiple R-squared:  0.9271,Adjusted R-squared:  0.9266
F-statistic:  1882 on 1 and 148 DF,  p-value: < 2.2e-16
> summary(lm2)

Call:
lm(formula = width ~ length + species)

Residuals:
     Min       1Q   Median       3Q      Max
-0.63706 -0.07779 -0.01218  0.09829  0.47814

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       -0.09083    0.05639  -1.611    0.109
length             0.23039    0.03443   6.691 4.41e-10 ***
speciesversicolor  0.43537    0.10282   4.234 4.04e-05 ***
speciesvirginica   0.83771    0.14533   5.764 4.71e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1796 on 146 degrees of freedom
Multiple R-squared:  0.9456,Adjusted R-squared:  0.9445
F-statistic: 845.5 on 3 and 146 DF,  p-value: < 2.2e-16
> summary(lm3)

Call:
lm(formula = width ~ length * species)

Residuals:
    Min      1Q  Median      3Q     Max
-0.6337 -0.0744 -0.0134  0.0866  0.4503

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)             -0.04822    0.21472  -0.225 0.822627
length                   0.20125    0.14586   1.380 0.169813
speciesversicolor       -0.03607    0.31538  -0.114 0.909109
speciesvirginica         1.18425    0.33417   3.544 0.000532 ***
length:speciesversicolor  0.12981    0.15550   0.835 0.405230
length:speciesvirginica  -0.04095    0.15291  -0.268 0.789244
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1773 on 144 degrees of freedom
Multiple R-squared:  0.9477,Adjusted R-squared:  0.9459
F-statistic: 521.9 on 5 and 144 DF,  p-value: < 2.2e-16
> ## BIC et anova
> k=log(nrow(iris))
> bic1=extractAIC(lm1,k=k)[2]
> bic2=extractAIC(lm2,k=k)[2]
> bic3=extractAIC(lm3,k=k)[2]
> c(bic1,bic2,bic3)
[1] -465.2514 -499.0472 -495.0086
> anova(lm1,lm2)
Analysis of Variance Table

Model 1: width ~ length
Model 2: width ~ length + species
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    148 6.3101
2    146 4.7116  2    1.5984 24.766 5.482e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> anova(lm2,lm3)
Analysis of Variance Table

Model 1: width ~ length + species
Model 2: width ~ length * species
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    146 4.7116
2    144 4.5274  2   0.18422 2.9297 0.0566 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

8 Phénomène de colinéarité : appréhension, détection et traitement

Exercice 58 (Coefficient de corrélation partielle)**.**
Soit $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$*, un modèle linéaire homoscédastique satisfaisant au moins l'hypothèse* $\mathcal{H}_1$*. Supposons que* $p > 2$ *et définissons* $\mathbf{X}^{(j)} = (\mathbf{x}_1, \ldots, \mathbf{x}_{j-1}, \mathbf{x}_{j+1}, \ldots, \mathbf{x}_p)$*, c'est-à-dire la matrice de design* $\mathbf{X}$ *privée de la* j*ème covariable. De la même façon, on définit* $\beta^{(j)}$ *le vecteur de* $\mathbb{R}^{p-1}$*. Pour essayer de comprendre l'impact de l'ajout d'une covariable, nous allons travailler dans cet exercice sur le coefficient de corrélation partielle entre* $\mathbf{Y}$ *et* $\mathbf{x}_p$ *(par exemple), qui se trouve être le coefficient de corrélation entre* Y *et* $\mathbf{x}_p$ *pour lequel l'effet linéaire des autres variables aurait été retiré. Pour cela, on définit*

$$\hat{\boldsymbol{\varepsilon}}_{\mathbf{Y}|\mathbf{X}^{(p)}} = \mathbf{Y} - \mathbf{X}^{(p)}\hat{\boldsymbol{\beta}}^{(p)} \quad \text{et} \quad \hat{\boldsymbol{\varepsilon}}_{\mathbf{x}_p|\mathbf{X}^{(p)}} = \mathbf{x}_p - \hat{\mathbf{X}}^{(p)}\hat{\boldsymbol{\beta}'}^{(p)}$$

où $\hat{\boldsymbol{\beta}}^{(p)}$ *et* $\hat{\boldsymbol{\beta}'}^{(p)}$ *sont les estimateurs par moindres carrés des modèles*

$$\mathbf{Y} = \mathbf{X}^{(p)}\boldsymbol{\beta}^{(p)} + \boldsymbol{\varepsilon}_1 \text{ et } \mathbf{x}_p = \mathbf{X}^{(p)}\boldsymbol{\beta'}^{(p)} + \boldsymbol{\varepsilon}_2$$

Ensuite, on définit le coefficient de corrélation partielle entre $\mathbf{Y}$ *et* $\mathbf{x}_p$ *sachant* $\mathbf{X}^{(p)}$*, noté* $r_{\mathbf{Y},\mathbf{x}_p|\mathbf{X}^{(p)}}$ *par le coefficient de corrélation linéaire entre* $\hat{\boldsymbol{\varepsilon}}_{\mathbf{Y}|\mathbf{X}^{(p)}}$ *et* $\hat{\boldsymbol{\varepsilon}}_{\mathbf{x}_p|\mathbf{X}^{(p)}}$*, i.e.*

$$r_{\mathbf{Y},\mathbf{x}_p|\mathbf{X}^{(p)}} = \frac{\hat{\boldsymbol{\varepsilon}}_{\mathbf{Y}|\mathbf{X}^{(p)}}^\top \hat{\boldsymbol{\varepsilon}}_{\mathbf{x}_p|\mathbf{X}^{(p)}}}{\|\hat{\boldsymbol{\varepsilon}}_{\mathbf{Y}|\mathbf{X}^{(p)}}\| \, \|\hat{\boldsymbol{\varepsilon}}_{\mathbf{x}_p|\mathbf{X}^{(p)}}\|}$$

1. *Montrez que*

$$r^2_{\mathbf{Y},\mathbf{x}_p|\mathbf{X}^{(p)}} = \frac{\mathbf{Y}^\top(\mathcal{P}_{\mathbf{X}} - \mathcal{P}_{\mathbf{X}^{(p)}})\mathbf{Y}}{\mathbf{Y}^\top(\mathbf{I}_n - \mathcal{P}_{\mathbf{X}^{(p)}})\mathbf{Y}}.$$

   Ce résultat montre que le coefficient de corrélation partielle représente la réduction relative de la sommme de carrés résiduelle due à l'introduction de la variable $\mathbf{x}_p$ *dans un modèle qui comprend déjà les variables* $\mathbf{X}^{(p)}$*.*

2. *Montrez que*

$$r^2_{\mathbf{Y},\mathbf{x}_p|\mathbf{X}^{(p)}} = \frac{R^2_p - R^2_{p-1}}{1 - R^2_{p-1}}$$

où R^2_p est le coefficient de détermination multiple de la régresssion avec p covariables. Ceci montre que le coefficient de corrélation partielle donne l'augmentation de R^2 relative à la portion de la variation de $\mathbf{Y}$ inexpliquée par les variables déjà dans le modèle.

3. *En* **R**, *à partir du jeu de données* **mtcars**, *calculez (en utilisant la définition initiale ou l'une des trois formules précédentes) $r_{mpg,wt}$, $r_{mpg,hp}$, $r_{mpg,wt|hp}$, $r_{mpg,hp|wt}$. Vous pourrez vérifier les sorties ci-dessous:*

```
> cor(mtcars[,c("mpg","wt","hp")])

           mpg         wt         hp
mpg  1.0000000 -0.8676594 -0.7761684
wt  -0.8676594  1.0000000  0.6587479
hp  -0.7761684  0.6587479  1.0000000

> require(ppcor)
> pcor(mtcars[,c("mpg","wt","hp")])$estimate

           mpg          wt          hp
mpg  1.0000000 -0.75120490 -0.54699262
wt  -0.7512049  1.00000000 -0.04690019
hp  -0.5469926 -0.04690019  1.00000000
```

4. *Analysez encore la sortie ci-dessous pour le jeu de données voiture (fichier voiture.RData) donnant pour 10 voiture différentes leur âge, prix et nombre de kilomètres au compteur. Tracez ce jeu de données pour mettre en parallèle les calculs de matrice de corrélation et corrélation partielle. On cherche dans ce jeu de données à expliquer la variable* **prix**.

```
> load('../R/voiture.RData')
> cor(voiture)

            age         km       prix
age   1.0000000  0.9932458 -0.9631406
km    0.9932458  1.0000000 -0.9492412
prix -0.9631406 -0.9492412  1.0000000

> pcor(voiture)$estimate

            age        km       prix
age   1.0000000 0.9335777 -0.5565049
km    0.9335777 1.0000000  0.2369003
prix -0.5565049 0.2369003  1.0000000
```

5. *La notion de corrélation partielle est en réalité bien plus riche et complexe que celle qui a été définie plus haut. Ainsi par exemple lorsque $p \geq 3$, on peut très bien définir $r_{\mathbf{Y},\mathbf{x}_1|\mathbf{x}_2}$, corrélation entre $\mathbf{Y}$ et $\mathbf{x}_1$ dans laquelle l'effet de $\mathbf{x}_2$ uniquement a été enlévé. Appuyez-vous sur la définition initiale pour définir ce nouveau coefficient en utilisant les résidus puis les matrices de projection.*

Exercice 59 (VIF et R^2)**.**
Dans le contexte du cours sur le phénomène de colinéarité (voir le cours pour les différentes notations). On suppose évidemment disposer d'un modèle linéaire homoscédastique. Montrez que

$$\mathrm{VIF}_j = \frac{1}{1-R_j^2} = S_j^2 \times \left(\mathbf{X}^\top \mathbf{X}\right)^{-1}_{jj}.$$

Exercice 60. *Soit $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ un modèle de régression linéaire homoscédastique gaussien. On suppose que la matrice de design $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2)$ est de taille $(n, 2)$. Le vecteur des paramètres $\boldsymbol{\beta}$ est donc de dimension 2. On notera σ^2 le paramètre de variance du bruit. Enfin, on supposera en outre que $\mathbf{x}_1$, $\mathbf{x}_2$ sont centrés et que*

$$\mathbf{X}^\top \mathbf{X} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

où ρ est un paramètre réel.

1. *Quelle est la condition sur n et ρ pour que la matrice de design $\mathbf{X}$ soit de rang plein et que $\mathbf{X}^\top\mathbf{X}$ soit définie positive? Cette condtion sera supposée par la suite.*

2. *Soit $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \hat{\beta}_2)$ l'estimateur obtenu par moindres carrés ordinaires de ce modèle. Montrez que pour $j = 1, 2$*

$$\hat{\beta}_j = \frac{1}{1-\rho^2}\mathbf{z}_j^\top \mathbf{Y}, \qquad \text{avec } \mathbf{z}_1 = \mathbf{x}_1 - \rho\mathbf{x}_2 \text{ et } \mathbf{z}_2 = \mathbf{x}_2 - \rho\mathbf{x}_1.$$

3. *Calculez $\mathrm{Var}(\hat{\beta}_j)$ en fonction de σ^2 et ρ, pour $j = 1, 2$.*

4. *On rappelle que le critère du facteur d'augmentation de la variance du $j$ème régresseur VIF_j vaut $\left((\mathbf{X}^\top\mathbf{X})^{-1}\right)_{jj}$. Pour quelle condition sur ρ, VIF_j est-il supérieur à 4? supérieur à 10?*

5. *Soit $\hat{\sigma}^2$ l'estimateur de la variance du bruit ($= RSS/(n-2)$). Définir les statistiques des tests de significativité des paramètres β_1 et β_2 en fonction de ρ. Que se passe-t-il lorsque $|\rho| \to 1$? Commentez.*

9 Sélection de variables

Exercice 61 (Biais dans un sous-modèle)**.**
Dans le contexte du chapitre du cours et de ses notations, montrez que

1. $\hat{\boldsymbol{\beta}}_\xi$ *et* $\hat{Y}_\xi$ *sont en général biaisés.*
2. $\hat{\sigma}^2_\xi$ *est en général positivement biaisé.*

Exercice 62 (Augmentation de la variance d'un sous-modèle)**.**
Dans le contexte du chapitre du cours et de ses notations, montrez que $\mathrm{Var}([\hat{\boldsymbol{\beta}}]_\xi) - \mathrm{Var}(\hat{\boldsymbol{\beta}}_\xi)$ *est une matrice semi-définie positive (la notation implique ici que* ξ *est nécessairement un modèle plus petit que le vrai modèle).*

Exercice 63 (EQM d'un sous-modèle)**.**
Montrez que

1. $$\mathrm{tr}(\mathrm{EQM}(\hat{\mathbf{Y}}_\xi)) = |\xi|\sigma^2 + \|\mathcal{P}_{\mathbf{X}_\xi^\perp}\mathbf{X}\boldsymbol{\beta}\|^2.$$
2. *(si nouvelles données sont indépendantes des observations)*
$$\mathrm{tr}(\mathrm{EQMP}(\hat{\mathbf{Y}}^*_\xi)) = n^*\sigma^2 + \mathrm{tr}(\mathrm{EQM}(\mathbf{X}^*_\xi\hat{\boldsymbol{\beta}}_\xi))$$

Exercice 64 (log-vraisemblance au point maximum)**.**
Sous l'hypothèse de normalité du bruit additif, montrez que la log-vraisemblance du modèle à ξ *variables explicatives évaluée à l'EMV vaut*

$$\log \mathcal{L}(\xi) = -\frac{n}{2}\log\frac{\mathrm{SCR}(\xi)}{n} - \frac{n}{2}\left(1 + \log 2\pi\right).$$

Exercice 65 (Exercice pratique - Jeu de données `swiss`)**.**
Le jeu de données `swiss` *disponible sous* `R` *est décrit de la façon suivante sous* `R`*:*

"A data frame with 47 observations on 6 variables, each of which is in percent, i.e., in [0, 100].

- *Fertility Ig, 'common standardized fertility measure'*
- *Agriculture % of males involved in agriculture as occupation*
- *Examination % draftees receiving highest mark on army examination*
- *Education % education beyond primary school for draftees.*
- *Catholic % 'catholic' (as opposed to 'protestant').*
- *Infant.Mortality live births who live less than 1 year.*

"

L'idée est d'expliquer la variable `Fertility` *en fonction des autres. Analysez ce jeu de données en utilisant les outils du cours. Cette analyse doit rentrer dans le format suivant:*

1. *Une synthèse d'une page maximum (essayez d'être synthétique tout en étant précis!)*
2. *Annexes: vous présentez des instructions* `R`*, leurs évaluations, graphiques, qui justifient votre synthèse.*

10 Moindres carrés généralisés

Exercice 66 (Introduction aux MCG)**.**
Avec les notations du cours, soit $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$*, un modèle linéaire tel que la matrice* $\mathbf{X}$ *de taille* (n, p) *est de rang plein. En revanche, concernant le vecteur aléatoire* $\boldsymbol{\varepsilon}$*, on suppose que*

$$\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0} \quad \textit{et} \quad \mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2\boldsymbol{\Omega}, \tag{8}$$

où $\boldsymbol{\Omega}$ *est une matrice de taille* (n, n) *symétrique, définie positive.*

1. *Soit* $\widehat{\boldsymbol{\beta}}$ *l'estimateur par moindres carrés ordinaires de ce modèle. Après avoir rappelé la forme de cet estimateur, montrez que bien que ce modèle ne soit plus homoscédastique,* $\widehat{\boldsymbol{\beta}}$ *reste un estimateur sans biais.*

2. *Calculez la variance de* $\widehat{\boldsymbol{\beta}}$*. En supposant que* $\boldsymbol{\varepsilon}$ *est un vecteur gaussien, déterminez la loi de* $\widehat{\boldsymbol{\beta}}$*?*

3. *On rappelle que l'estimateur de* σ^2 *est classiquement défini par*

$$\hat{\sigma}^2 = \frac{\|\widehat{\boldsymbol{\varepsilon}}\|^2}{n-p} = \frac{\boldsymbol{\varepsilon}^\top \mathcal{P}_{\mathbf{X}^\perp}\boldsymbol{\varepsilon}}{n-p}.$$

   où $\mathcal{P}_{\mathbf{X}}^{\perp}$ *est la matrice de projection sur* $\mathcal{V}(\mathbf{X})^\perp$*. Calculez* $\mathrm{E}(\hat{\sigma}^2)$*.*

4. *Sur le plan pratique, quelles sont les conséquences des trois questions précédentes?*

5. *Soit* $\boldsymbol{\Omega}^{1/2}$ *la racine carrée de la matrice* $\boldsymbol{\Omega}$ *(rappel: ceci a un sens car* $\boldsymbol{\Omega}$ *est symétrique définie positive). On transforme le modèle* (8) *de la façon suivante:*

$$\underbrace{\boldsymbol{\Omega}^{-1/2}\mathbf{Y}}_{:=\mathbf{Y}'} = \underbrace{\boldsymbol{\Omega}^{-1/2}\mathbf{X}}_{:=\mathbf{X}'}\boldsymbol{\beta} + \underbrace{\boldsymbol{\Omega}^{-1/2}\boldsymbol{\varepsilon}}_{:=\boldsymbol{\varepsilon}'}. \tag{9}$$

   Montrez que le modèle (9)*,* $\mathbf{Y}' = \mathbf{X}'\boldsymbol{\beta} + \boldsymbol{\varepsilon}'$*, est un modèle linéaire homoscédastique standard.*

6. *Soit $\widetilde{\boldsymbol{\beta}}$ l'estimateur par moindres carrés ordinaires du modèle* (9)*. Exprimez $\widetilde{\boldsymbol{\beta}}$ en fonction de $\mathbf{X}$, $\mathbf{Y}$ et $\boldsymbol{\Omega}$.*

7. *Montrez que $\widetilde{\boldsymbol{\beta}}$ est un estimateur sans biais. Calculez sa variance. Déterminez sa loi (en supposant en plus que $\boldsymbol{\varepsilon}$ est un vecteur gaussien).*

8. *Quel résultat justifie que $\mathrm{Var}(\widetilde{\boldsymbol{\beta}}) \leq \mathrm{Var}(\widehat{\boldsymbol{\beta}})$?*

9. *Soit $\tilde{\sigma}^2 = \|\widetilde{\boldsymbol{\varepsilon}}'\|^2/(n-p)$ où $\widetilde{\boldsymbol{\varepsilon}}'$ est le vecteur des résidus du modèle* (9)*. Exprimez $\tilde{\sigma}^2$ en fonction de $\boldsymbol{\varepsilon}$ et d'une certaine matrice de projection que vous définirez. Montrez alors que $\tilde{\sigma}^2$ est un estimateur sans biais de σ^2.*

10. *En supposant que $\boldsymbol{\varepsilon}$ est un vecteur gaussien, en déduire la loi de $(n-p)\tilde{\sigma}^2$ et que $\widetilde{\boldsymbol{\beta}}$ et $\tilde{\sigma}^2$ sont indépendants.*

11. *Sur le plan pratique, y a-t-il des limites à mettre en place l'estimateur $\widetilde{\boldsymbol{\beta}}$?*

12. *Supposons dans cette question uniquement que $\boldsymbol{\Omega}$ soit une matrice diagonale et que $\Omega_{ii} = 1$ pour $i = 1, \ldots, n/2$ (supposons n pair) et δ pour $i = n/2+1, \ldots, n$ avec $\delta > 0$ un paramètre réel inconnu. Comment feriez-vous pour mettre en place une procédure approchant le modèle* (9)*?*

Exercice 67 (Simulation et estimation dans un modèle hétéroscédastique)**.**

Considérons un modèle linéaire $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ de taille $n = 200$, avec $p = 2$ et tel que $\mathbf{1} \in \mathcal{V}(\mathbf{X})$. Supposons que $\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma_1^2$ pour $i = 1, \ldots, 100$ et σ_2^2 pour $i = 101, \ldots, 200$. Par une étude en simulation en `R` (basée sur un grand nombre de réalisations):

- *Fixez-vous un régresseur*
- *Construire un modèle linéaire suivant le modèle précédent. Et simulez ce modèle m fois.*
- *Calculez pour chaque réalisation du modèle $\hat{\boldsymbol{\beta}}^{\mathrm{MCO}}$ et $\hat{\boldsymbol{\beta}}^{\mathrm{MCG}}$ (et stockez ces valeurs); pour la méthode MCG on utilisera la variante MCQG présentée en cours.*

Le but de l'exercice est de vérifier le théorème de Gauss-Markov. Aussi

1. *Vérifiez que les estimations obtenues par MCO et MCQG sont sans biais.*

2. *Montrez que* $\text{Var}(\hat{\boldsymbol{\beta}}^{\text{MCG}}) \leq \text{Var}(\hat{\boldsymbol{\beta}}^{\text{MCO}})$ *en montrant que la différence* $\text{Var}(\hat{\boldsymbol{\beta}}^{\text{MCG}}) - \text{Var}(\hat{\boldsymbol{\beta}}^{\text{MCO}})$ *estimée empiriquement entre les deux matrices de covariance empiriques est semi-définie négative.*

Exercice 68. *Considérez la fonction* `R` *ci-dessous. Cette simulation est éxécutée pour* $n = 40$ *et* $n = 200$ *dans la page suivante. En 10-15 lignes, décrivez précisément ce qui est réalisé dans cette simulation, son objectif, les résultats de cette simulation, vos réflexions, etc.*

```
> sim=function(n,m=1000){
+   x1=runif(n,0,1);x2=runif(n,5,6)
+   v1=.05;v2=20;v=c(rep(v1,n/2),rep(v2,n/2))
+   est0=est1=est2=est.th=matrix(0,nrow=m,ncol=3)
+   se0=se1=se2=se.th=matrix(0,nrow=m,ncol=3)
+   for (i in 1:m){
+     eps=rnorm(n,0,sqrt(v))
+     y=1+.5*x1+.5*x2+eps
+     lm0=lm(y~x1+x2);res0<-residuals(lm0)

+ #### stratégie 1
+     v1est=var(res0[1:(n/4)])
+     v2est=var(res0[(3*n/4+1):n])
+     weights1=1/c(rep(v1est,n/4),rep(v2est,n/4))
+     ind.est=(n/4+1):(3*n/4)
+     lm1=lm(I(y[ind.est])~I(x1[ind.est])+I(x2[ind.est]),weights=weights1)
+ #### stratégie 2
+     v1est=var(res0[1:(n/2)])
+     v2est=var(res0[(n/2+1):n])
+     weights2=1/c(rep(v1est,n/2),rep(v2est,n/2))
+     lm2=lm(y~x1+x2,weights=weights2)
+ #### Théoriquement
+         lm.th=lm(y~x1+x2,weights=1/v)

+     est0[i,]=lm0$coeff;est1[i,]=lm1$coeff
+     est2[i,]=lm2$coeff;est.th[i,]=lm.th$coeff
+     se0[i,]=summary(lm0)$coeff[,2]
+     se1[i,]=summary(lm1)$coeff[,2]
+     se2[i,]=summary(lm2)$coeff[,2]
+     se.th[i,]=summary(lm.th)$coeff[,2]
+   }
+   f=function(vec,fun) apply(vec,2,fun)
+   moy=rbind(f(est0,mean),f(est1,mean),f(est2,mean),f(est.th,mean))
+   et=rbind(f(est0,sd),f(est1,sd),f(est2,sd),f(est.th,sd))
+   se=rbind(f(se0,mean),f(se1,mean),f(se2,mean),f(se.th,mean))
+   colnames(moy)=colnames(et)=colnames(se)=c('Intercept','b1','b2')
+   rownames(moy)=rownames(et)=rownames(se)=c('est0','est1','est2','est.th')
+   return(list(moy=moy,et=et,se=se))
+ }
```

Cette simulation est maintenant exécutée:

```
> lapply(sim(n=40,m=500),round,3)
$moy
       Intercept    b1    b2
est0       1.112 0.396 0.487
est1       1.036 0.483 0.495
est2       1.061 0.489 0.490
est.th     1.026 0.495 0.496

$et
       Intercept    b1    b2
est0      11.445 1.759 2.043
est1       4.737 0.478 0.857
est2       2.353 0.359 0.413
est.th     1.233 0.175 0.221

$se
       Intercept    b1    b2
est0      10.594 1.606 1.914
est1       5.928 0.647 1.077
est2       2.828 0.437 0.503
est.th     1.248 0.193 0.222
```

```
> lapply(sim(200),round,3)
$moy
       Intercept    b1    b2
est0       0.870 0.495 0.523
est1       0.988 0.496 0.503
est2       1.004 0.501 0.499
est.th     1.009 0.501 0.498

$et
       Intercept    b1    b2
est0       3.903 0.793 0.715
est1       0.544 0.111 0.101
est2       0.412 0.083 0.076
est.th     0.408 0.081 0.075

$se
       Intercept    b1    b2
est0       3.992 0.775 0.720
est1       0.733 0.151 0.135
est2       0.550 0.110 0.100
est.th     0.397 0.080 0.072
```

11 Régressions ridge et lasso

Exercice 69 (Valeurs et vecteurs propres de $\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_p$)**.**
Soit $\lambda \geq 0$, montrez que $\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_p$ a les mêmes vecteurs propres que $\mathbf{X}^\top\mathbf{X}$ et pour valeurs propres $\mu_j + \lambda$ pour $j = 1, \ldots, p$.

Exercice 70 (Estimateur ridge et propriétés)**.**
Montrez que l'estimateur ridge satisfait:

1. $\hat{\boldsymbol{\beta}}_{\text{ridge}} = \left(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_p\right)^{-1}\mathbf{X}^\top\mathbf{X}\hat{\boldsymbol{\beta}}$.
2. $\mathrm{E}\left(\hat{\boldsymbol{\beta}}_{\text{ridge}}\right) = \boldsymbol{\beta} - \lambda\left(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_p\right)^{-1}\boldsymbol{\beta}$.
3. $\mathrm{Var}(\hat{\boldsymbol{\beta}}_{\text{ridge}}) = \sigma^2\left(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_p\right)^{-1}\mathbf{X}^\top\mathbf{X}\left(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_p\right)^{-1}$.

Montrez également qu'en termes d'erreur quadratique moyenne

- $\mathrm{EQM}(\hat{\boldsymbol{\beta}}) = \mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}$.
- $\mathrm{EQM}(\hat{\boldsymbol{\beta}}_{\text{ridge}})\left(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_p\right)^{-1}\left(\sigma^2(\mathbf{X}^\top\mathbf{X}) + \lambda^2\boldsymbol{\beta}\boldsymbol{\beta}^\top\right)\left(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_p\right)^{-1}$

Exercice 71 (Matrice de projection pour l'estimateur ridge)**.**
Soit $\mathbf{H}(\lambda) = \tilde{\mathbf{X}}\left(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} + \lambda\mathbf{I}_p\right)^{-1}\tilde{\mathbf{X}}^\top$ la matrice de projection faisant passer $\mathbf{Y}$ à $\hat{\mathbf{Y}}_{\text{ridge}}$. Montrez que

$$\mathrm{tr}(\mathbf{H}(\lambda)) = \sum_{j=1}^{p}\frac{\tilde{\mu}_j}{\tilde{\mu}_j + \lambda}$$

où les $\tilde{\mu}_j$ sont les v.p. de $\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}}$, version centrée-réduite de $\mathbf{X}$.

Exercice 72 (Estimateur ridge adaptatif)**.**
Dans cet exercice, on considère un modèle linéaire de la forme $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ avec $\boldsymbol{\varepsilon}$ un vecteur centré de variance $\sigma^2\mathbf{I}_n$ et où $\mathbf{X}$ est une matrice de design de taille (n, p) (avec $n > p$) de rang plein. Nous allons généraliser l'estimateur

ridge à une version adaptative: soient $\lambda_1, \dots, \lambda_p$ des réels positifs et $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \dots, \lambda_p)$. Soit $\hat{\boldsymbol{\beta}}_{\mathrm{ridge}}$ l'estimateur minimisant le problème

$$\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \|\mathbf{\Lambda}^{1/2}\boldsymbol{\beta}\|^2 = \sum_{i=1}^{n}(Y_i - \mathbf{x}_i'\boldsymbol{\beta})^2 + \sum_{j=1}^{p}\lambda_j\beta_j^2.$$

La solution de ce problème est appelé estimateur ridge adaptatif (il correspond évidemment à l'estimateur ridge standard dans le cas où $\lambda_1 = \cdots = \lambda_p$).

1. *On rappelle que $\frac{\partial\|\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}\|^2}{\partial\boldsymbol{\beta}} = -2\mathbf{X}^\top(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$. Montrez que*

$$\hat{\boldsymbol{\beta}}_{\mathrm{ridge}} = \left(\mathbf{X}^\top\mathbf{X} + \mathbf{\Lambda}\right)^{-1}\mathbf{X}^\top\mathbf{Y}.$$

2. *Notons $\hat{\boldsymbol{\beta}}$ l'estimateur des MCO. Rappelez sans démonstration ce que valent $\mathrm{E}(\hat{\boldsymbol{\beta}})$, $\mathrm{Var}(\hat{\boldsymbol{\beta}})$ et $\mathrm{EQM}(\hat{\boldsymbol{\beta}})$ (l'erreur quadratique moyenne de $\hat{\boldsymbol{\beta}}$).*

3. *Calculez $\mathrm{E}(\hat{\boldsymbol{\beta}}_{\mathrm{ridge}})$ et $\mathrm{Var}(\hat{\boldsymbol{\beta}}_{\mathrm{ridge}})$.*

4. *En déduire le biais de l'estimateur ridge adaptatif puis montrez que l'erreur quadratique moyenne de l'estimateur ridge adaptatif s'écrit sous la forme*

$$\mathrm{EQM}(\hat{\boldsymbol{\beta}}_{\mathrm{ridge}}) = \left(\mathbf{X}^\top\mathbf{X} + \mathbf{\Lambda}\right)^{-1}\left(\sigma^2(\mathbf{X}^\top\mathbf{X}) + (\mathbf{\Lambda}\boldsymbol{\beta})(\mathbf{\Lambda}\boldsymbol{\beta})^\top\right)\left(\mathbf{X}^\top\mathbf{X} + \mathbf{\Lambda}\right)^{-1}$$

5. *Dans cette question uniquement, nous ferons les simplifications suivantes: supposons que $\mathbf{1} \notin \mathcal{V}(\mathbf{X})$, que les covariables $\mathbf{x}_1, \dots, \mathbf{x}_p$ sont orthogonales. On notera $v_j = \|\mathbf{x}_j\|^2 = \sum_{i=1}^{n}(\mathbf{x}_j)_i^2$. Et on supposera en outre que $\lambda_1 = \cdots = \lambda_{p-1} = 0$. Calculez $\mathbf{X}^\top\mathbf{X}$ en fonction des v_j puis en déduire une expression simple pour $\mathrm{EQM}(\hat{\boldsymbol{\beta}})$ et $\mathrm{EQM}(\hat{\boldsymbol{\beta}}_{\mathrm{ridge}})$. En supposant maintenant que $\lambda_p = \sigma\sqrt{v_p}/\beta_p$, en déduire une condition pour que $\mathrm{tr}(\mathrm{EQM}(\hat{\boldsymbol{\beta}}_{\mathrm{ridge}})) \leq \mathrm{tr}(\mathrm{EQM}(\hat{\boldsymbol{\beta}}))$.*

Exercice 73. *Soit $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ un modèle linéaire de n observations. La matrice de design $\mathbf{X}$ de taille (n, p) (avec $p < n$) de plein rang est supposée constituée de régresseurs orthonormés, i.e. $\mathbf{X}^\top\mathbf{X} = \mathbf{I}_p$. On cherche à estimer $\boldsymbol{\beta}$ par minimisation du critère:*

$$f(\boldsymbol{\beta}) = \frac{1}{2}\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\left(\alpha\frac{1}{2}\|\boldsymbol{\beta}\|_2^2 + (1-\alpha)\|\boldsymbol{\beta}\|_1\right)$$

où on rappelle que $\|\boldsymbol{\beta}\|_p^p = \sum_j |\beta_j|^p$ pour $p > 0$, et où $\lambda \geq 0$ est un paramètre de pénalisation. Dans la suite $j \in \{1, \dots, p\}$.

1. *Soit $\hat{\boldsymbol{\beta}}$ l'estimateur par moindres carrés ordinaires, qui correspond à $\lambda = 0$. Que vaut $\hat{\beta}_j$ (en fonction de $\mathbf{X}$ et $\mathbf{Y}$)?*

2. *Soit $\alpha = 1$, on note $\hat{\boldsymbol{\beta}}^R$ l'estimateur ridge associé. Montrez que*

$$\hat{\beta}_j^R = \frac{\hat{\beta}_j}{1+\lambda}.$$

3. *Soit $\alpha = 0$, on note $\hat{\boldsymbol{\beta}}^L$ l'estimateur lasso associé. Montrez que*

$$\hat{\beta}_j^L = \text{sign}(\hat{\beta}_j)\max(|\hat{\beta}_j| - \lambda, 0).$$

   Pour cette question, commencez par montrer que minimiser f est équivalent à minimiser pour tout j

$$-\hat{\beta}_j\beta_j + \frac{1}{2}\beta_j^2 + \lambda|\beta_j|$$

   et remarquez également que $\text{sign}(\hat{\beta}_j) = \text{sign}(\beta_j)$.

4. *Pour $0 < \alpha < 1$, soit $\hat{\boldsymbol{\beta}}^E$ l'estimateur elastic net. Suivez la même stratégie pour établir une relation entre $\hat{\beta}_j^E$ et $\hat{\beta}_j$.*

12 Modèles linéaires généralisés: introduction et théorie

Exercice 74 (Propriétés des lois de la famille exponentielle)**.**
Pour rappel, la densité d'une loi de la famille exponentielle s'écrit

$$dF(x;\theta,\phi) = \exp\left(\frac{x\theta - b(\theta)}{\phi} + c(x,\phi)\right)$$

On appelle transformée d'Esscher de paramètre h d'une fonction de répartition F est F_h définie par

$$dF_h(x) = \frac{e^{hx}dF(x)}{\int e^{hy}dF(y)}$$

Soit Y une variable dont la distribution est dans la famille exponentielle.

1. *Montrez que la fonction génératrice des moments de Y est*

$$m_Y(t) = \exp\left(\frac{b(\theta + t\phi) - b(\theta)}{\phi}\right)$$

2. *En déduire que si on dispose de n variables $Y_1, \cdots, Y_n$ suivant cette loi, alors $\overline{Y} = \dfrac{Y_1 + \cdots + Y_n}{n}$ admet une distribution qui est également dans la famille exponentielle (dont on donnera les paramètres).*

3. *Montrer que la transformée d'Esscher de Y est aussi une loi de la famille exponentielle (dont on donnera les paramètres).*

4. *Donner la transformée d'Esscher de paramètre h*

   - *pour une variable $Y \sim \mathcal{N}(0,1)$*
   - *our une variable $Y \sim \mathcal{P}(1)$*
   - *pour une variable binomiale $Y \sim \mathcal{B}(n, 1/2)$*

- *pour une variable Gamma* $Y \sim \mathcal{G}(1,1)$

Exercice 75 (Propriétés de $\hat{\boldsymbol{\eta}}$ et $\hat{\boldsymbol{\mu}}$)**.**
En admettant la convergence en loi de $\hat{\boldsymbol{\beta}}$*, montrez que lorsque* $n \to \infty$*, on a les deux convergences en distribution suivantes:*

- $\sigma^{-1}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}) \to \mathcal{N}(0, \mathbf{X}\boldsymbol{\Sigma}\mathbf{X}^\top)$*;*
- $\sigma^{-1}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}) \to \mathcal{N}(0, \boldsymbol{\Delta}^{-2}\mathbf{X}\boldsymbol{\Sigma}\mathbf{X}^\top)$.

Exercice 76 (Fonction de lien très particulière)**.**
On se donne deux vecteurs de taille n*,* $\mathbf{y}$ *et* $\mathbf{x}_1$*. Ci-dessous* $n = 10$ *pour fixer les idées mais peu importe.*

1. *Ici on suppose que* $\mathbf{x}_1$ *et* $\mathbf{y}$ *sont à valeurs positives. Expliquez en détail pourquoi mathématiquement, les deux sorties ci-dessous donnent le même résultat.*

```
> y=runif(10,5,10);x1=runif(10,8,12)
> glm(y~x1-1,family=gaussian(link="inverse"))coef

       x1
0.01310987

> 1/(lm(y~I(1/x1)-1)$coef)

   I(1/x1)
0.01310987
```

2. *On se donne un second vecteur* $\mathbf{x}_2$ *et ici, on prendra* $\mathbf{y}$ *à valeurs entières. Expliquez mathématiquement ce que vous observez sur la colonne* ***Std. Error****:*

```
> y=rpois(10,1);x1=runif(10,8,12);x2=runif(10,4,9)
> summary(glm(y~x1+x2,family=poisson(link=sqrt)))$coefficients

              Estimate Std. Error    z value  Pr(>|z|)
(Intercept)  1.16338255  1.7394362  0.6688274 0.5036056
x1          -0.04208880  0.1534676 -0.2742521 0.7838909
x2           0.02809503  0.1085161  0.2589020 0.7957109

> X=cbind(1,x1,x2); M = solve(t(X)%*%X)/4
> sqrt(diag(M))

                x1        x2
1.7394362 0.1534676 0.1085161
```

Exercice 77. *Les mêmes notations qu'en cours sont utilisées. On rappelle qu'une variable aléatoire* Y *appartient à la famille de dispersion si sa densité (ou fonction de masse) s'écrit:*

$$f_Y(y;\theta) = \exp\left(\frac{y\theta - v(\theta)}{\sigma^2} + c(y,\sigma^2)\right).$$

où θ est le paramètre canonique, v une fonction deux fois continûment dérivable, c une fonction ne dépendant que de y et σ^2. Soit Y une variable aléatoire de loi inverse gaussienne de moyenne $\mu > 0$ et de paramètre de forme $\lambda > 0$. Sa densité est donnée par:

$$f_Y(y) = \left(\frac{\lambda}{2\pi y^3}\right)^{1/2} \exp\left(-\frac{\lambda(y-\mu)^2}{2\mu^2 y}\right), \qquad y > 0.$$

1. *Montrez que la loi de Y appartient à la famille exponentielle de dispersion. Que valent θ, $v(\theta)$, σ^2?*

2. *Vérifiez que $E(Y) = \mu$ et $\mathrm{Var}(Y) = \mu^3/\lambda$ (et donc que $V(\mu) = \mu^3$).*

3. *Soient $Y_1, \ldots, Y_n$ n v.a.i.i.d. de loi inverse gaussienne de moyenne μ_i et de paramètre de forme $\lambda = 1$ fixé. Soit $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$ le prédicteur linéaire associé à p covariables $\mathbf{x}_1, \ldots, \mathbf{x}_p$. On envisage un modèle GLM liant en particulier $\eta_i = g(\mu_i)$. Quelle est la fonction de lien canonique g de ce modèle?*

4. *Sous les hypothèses énoncées en cours, on rappelle que $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \approx \mathcal{N}(0, (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1}$. Que vaut cette matrice dans le cas où g correspond au lien canonique?*

5. *Quelle fonction de lien devrait-on choisir pour que $\mathbf{W} = \mathbf{I}_n$?*

6. *Dans cette question, on suppose disposer de deux covariables `x1` et `x2` (`x2` à valeurs positives) en `R` (vecteurs de taille n) et d'une variable réponse `y` (à valeurs positives) de n observations issues de $Y_1, \ldots, Y_n$ modélisées par des variables de loi inverse gaussienne (`family=inverse.gaussian` en `R`). On souhaite mettre en place un modèle GLM pour lequel*

$$\mathrm{E}(Y_i) = \sqrt{x_{i2}} \exp\left(\beta_1 x_{i1}\right), \qquad i = 1, \ldots, n$$

où β_1 serait à estimer. Quelle est l'instruction `R` complète qui permet d'estimer ce modèle? Justifiez précisément.

13 Modèles linéaires généralisés: pratiques sur quelques modèles particuliers

Exercice 78 (Pratique d'un modèle GLM - Jeu de données `titanic`).
Cette question porte sur le jeu de données ***`titanic`*** *et en particulier sur les variables:*

- ***`survived`****: 0 ou 1 (a survécu au naufrage du titanic);*
- ***`sex`****: male ou female;*
- ***`pclass`****: classe du billet de l'individu (1,2 ou 3).*
- ***`age`****: âge de l'individu.*

Voilà les premières observations:

```
> nrow(titanic)
[1] 1046
> head(titanic)
  survived    sex pclass   age
1        1 female      1 29.00
2        1   male      1  0.92
3        0 female      1  2.00
4        0   male      1 30.00
5        0 female      1 25.00
6        1   male      1 48.00
```

Sans précision les tests se feront au seuil de 5%.

1. *On envisage le modèle ci-dessous:*

```
> glm.titanic=glm(survived~age+pclass+sex,data=titanic,
+         family=binomial)
```

Écrire mathématiquement et précisément ce modèle. En particulier, précisez la loi de la variable réponse, la fonction de lien utilisée.

2. *L'estimation de ce modèle a donné les résultats suivants*

```
> glm.titanic$coefficients

(Intercept)         age      pclass     sexmale
 4.58926850 -0.03388495 -1.13324397 -2.49737670
```

Interprétez la valeur du coefficient associé à **`pclass`** *et* **`sexmale`**.

3. *Voilà maintenant le résumé de cette sortie ainsi que deux quantiles utiles pour la suite:*

```
> summary(glm.titanic)


Call:
glm(formula = survived ~ age + pclass + sex, family = binomial,
    data = titanic)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6159  -0.7162  -0.4321   0.6572   2.4041

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.58927    0.40572  11.311  < 2e-16 ***
age         -0.03388    0.00628  -5.395 6.84e-08 ***
pclass      -1.13324    0.11173 -10.143  < 2e-16 ***
sexmale     -2.49738    0.16612 -15.034  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1414.62  on 1045  degrees of freedom
Residual deviance:  983.02  on 1042  degrees of freedom
AIC: 991.02

Number of Fisher Scoring iterations: 4

> n=nrow(titanic);n

[1] 1046

> c(qchisq(.95,n-4),qchisq(0.95,3))

[1] 1118.208685    7.814728
```

Peut-on rejeter le modèle GLM basé sur le test de déviance? Peut-on affirmer que que les variables apportent de l'information?

4. *Calculez l'équivalent du R^2 pour ce modèle. Ce modèle vous semble-t-il prédictif?*

5. *On construit la table de prédiction ci-dessous ("observés" en lignes et "prédits" en colonne) en utilisant un seuil d'affectation à la modalité 1 à 0.5.*

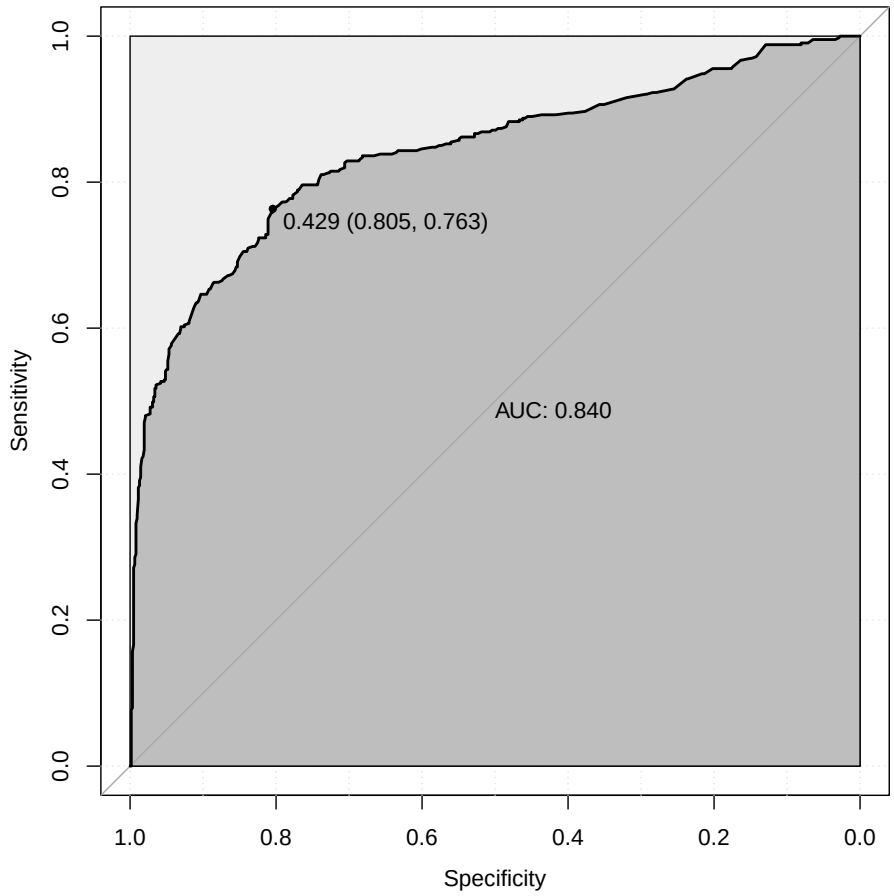
```
> yobs=as.numeric(titanic$survived)
> ypred=as.numeric(fitted(glm.titanic)>0.5)
> table(yobs,ypred)[2:1,2:1]

    ypred
```

```
yobs   1   0
   1 301 126
   0  96 523
```

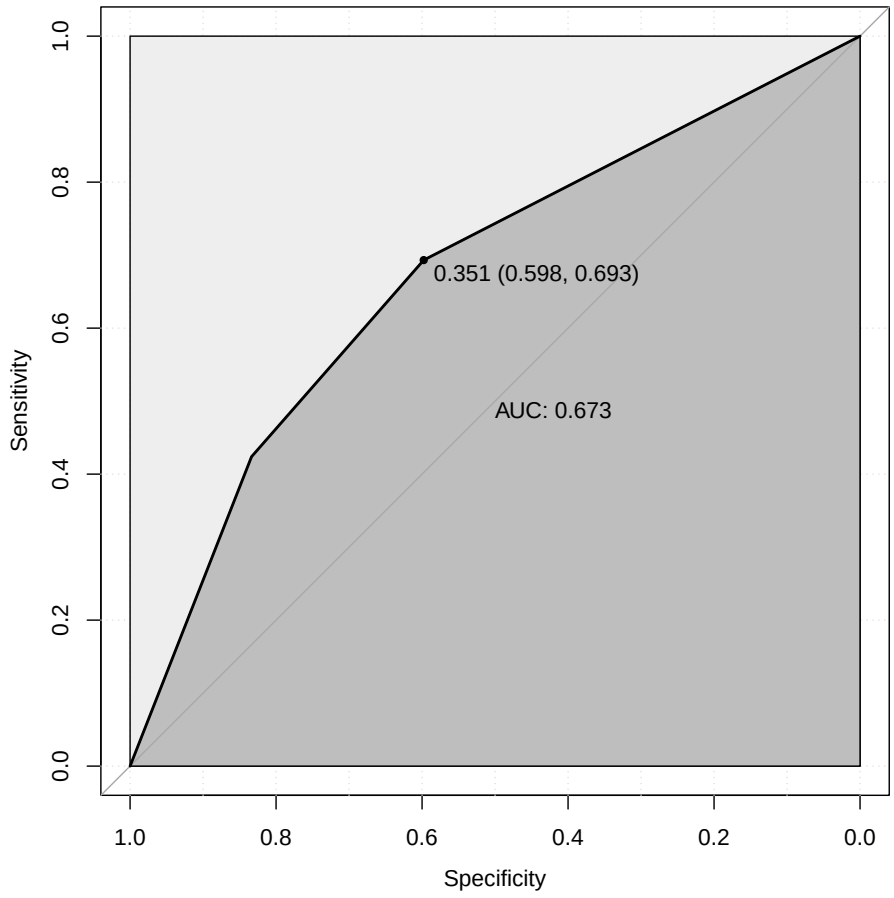
Calculez pour cette valeur du seuil les quantités TPR et FPR (True positive Rate et False Positive Rate).

6. *On rappelle que les indicateurs sensibilité et spécificité sont définis par: sensibilité=TPR et spécificité=1-FPR. Interprétez le graphique ci-dessous (courbe grise et surface grise). La valeur du seuil* $s = 0.5$ *vous semble-t-elle optimale? Que diriez-vous du dernier graphique où seule la variable* **`pclass`** *a été intégrée au modèle?*

```
> require(pROC)
> roc.titanic=roc(yobs,fitted(glm.titanic))
> plot(roc.titanic, print.auc=TRUE, auc.polygon=TRUE,
+         grid=c(0.1, 0.2), max.auc.polygon=TRUE,
+         auc.polygon.col="gray", print.thres=TRUE)
```




```
> glm.titanic2=glm(survived~pclass,data=titanic,
+ family=binomial)
> roc.titanic2=roc(yobs,fitted(glm.titanic2))
> plot(roc.titanic2, print.auc=TRUE, auc.polygon=TRUE,
+         grid=c(0.1, 0.2), max.auc.polygon=TRUE,
+         auc.polygon.col="gray", print.thres=TRUE)
```



Exercice 79. *Dans cet exercice, nous nous intéressons à un jeu de données recueilli par la ville de Montréal en 2015 du 1er janvier au 15 novembre 2015, recensant quotidiennement le nombre de cyclistes. Ces nombres sont observés sur différentes pistes cyclables de Montréal: Pont Jacques Cartier, Berri, Boyer, Brébeuf,. . . . Le jeu de données brut contient 22 localisations. Comme vous le savez sans doute, de plus en plus de cyclistes circulent toute l'année sur les pistes cyclables, hormis celles qui sont fermées, comme par exemple le pont Jacques Cartier. Du jeu de données complet, on s'intéresse au jeu de données, noté* `velo2` *ci-dessous contenant la date et 4 pistes cyclables: Maisonneuve, Berri, Rachel-Papineau et Pont Jacques Cartier (*`PJC`*). Voici les premières observations.*

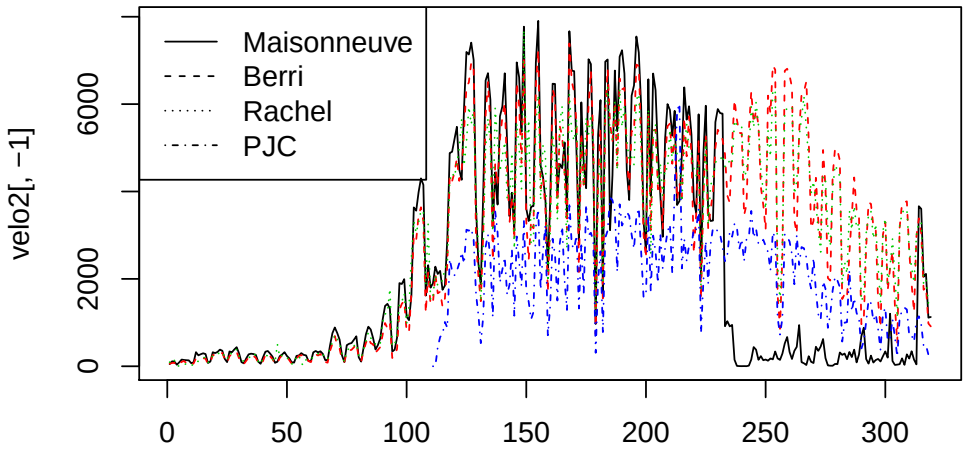
```
> load('velo.RData')
> velo2[c(1:3,(nrow(velo2)-2):nrow(velo2)),]
          Date Maisonneuve Berri Rachel PJC
1   01/01/2015          49    58     91  NA
2   02/01/2015         113    75    177  NA
3   03/01/2015         107    79    131  NA
317 13/11/2015        2115  1818   1980 482
318 14/11/2015        1112   979   1448 266
319 15/11/2015        1128   913   1491 380
```

Comme on peut le voir, `PJC` *est en effet fermé l'hiver. Les premières observations ont été faites mi-avril. La problématique de l'exercice est la suivante: nous voulons donner une estimation du nombre de cyclistes moyens qui passerait par le* `PJC` *s'il était ouvert toute l'année. Pour cela, on divise le jeu de donées en deux (certaines observations entre le 1er et 22 avril ne seront pas utilisées, manque d'information) :*

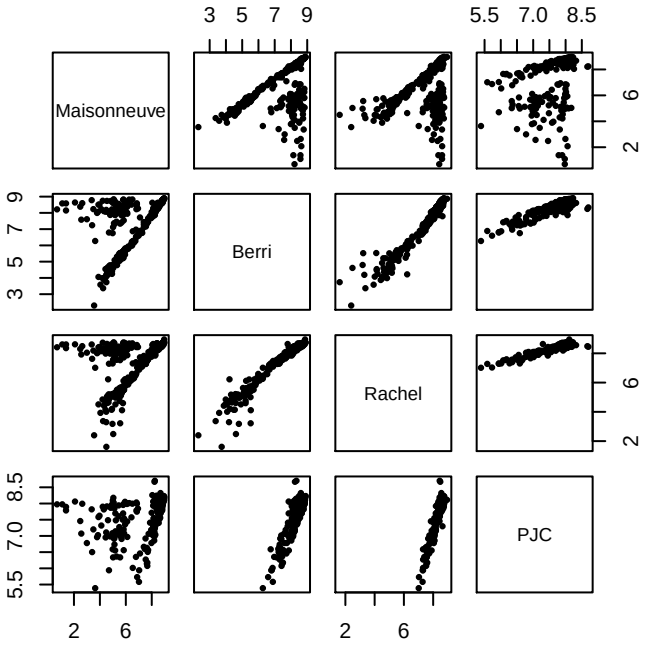
```
> i.test=1:90;i.app=113:nrow(velo2)
> velo2$Date[i.test[c(1,length(i.test))]]
[1] 01/01/2015 31/03/2015
319 Levels: 01/01/2015 01/02/2015 01/03/2015 01/04/2015 ... 31/10/2015
> velo2$Date[i.app[c(1,length(i.app))]]
[1] 23/04/2015 15/11/2015
319 Levels: 01/01/2015 01/02/2015 01/03/2015 01/04/2015 ... 31/10/2015
```

Le jeu de données `velo2[i.app,]` *sera utilisé pour estimer un modèle liant* `PJC` *aux autres variables et le jeu de données* `velo2[i.test,]` *sera utilisé pour prévoir la fréquentation du* `PJC`*. Voyons une première visualisation montrant qu'une telle modèlisation ne semble pas insensée:*

```
> matplot(velo2[,-1],type="l",lty=1:4)
> legend('topleft',names(velo2[-1]),lty=1:4)
```




```
> plot(log(velo2[,-1]),pch=20,cex=.7)
```



1. *Intéressons-nous aux données d'apprentissage, et modélisons* **PJC** *par des réalisations d'un modèle de Poisson. Commençons par modéliser* **PJC** *en fonction de* **Berri**. *En regardant le graphique précédent, on propose un premier modèle:*

```
> glm1=glm(PJC~log(Berri),data=velo2[i.app,],family=poisson)
```

   Justifiez ce modèle. En particulier comment sont modélisées $\mathrm{E}(\mathrm{PJC}_i)$ *et* $\mathrm{Var}(\mathrm{PJC}_i)$*?*

2. *Interprétez les coefficients estimés.*

```
> summary(glm1)


Call:
glm(formula = PJC ~ log(Berri), family = poisson, data = velo2[i.app,
    ])

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-20.538  -10.013   -3.140    4.735   69.644

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
```

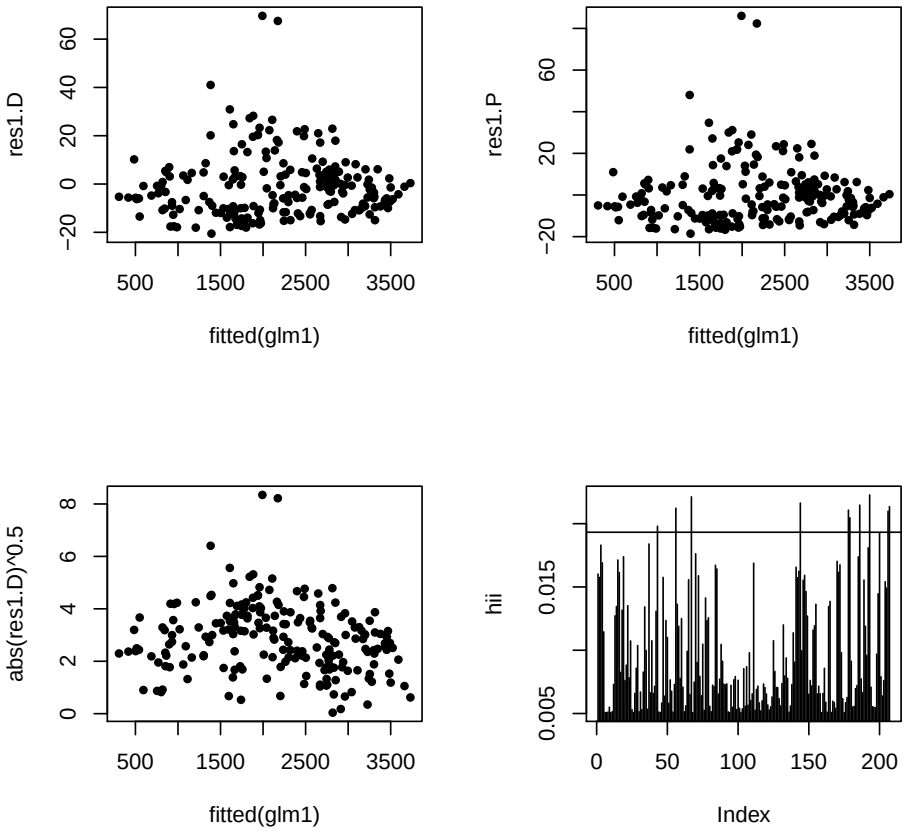
```
(Intercept) -0.21766    0.03282  -6.632 3.31e-11 ***
log(Berri)   0.94777    0.00389 243.645  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 109434  on 206  degrees of freedom
Residual deviance:  36436  on 205  degrees of freedom
AIC: 38380

Number of Fisher Scoring iterations: 4
```

3. *A l'aide de deux tests basés sur la déviance: peut-on montrer que la covariable* ***log(Berri)*** *apporte de l'information? peut-on rejeter le modèle GLM?*

```
> n=nrow(velo2[i.app,]);p=2
> 1-pchisq(glm1$deviance,n-p)

[1] 0

> qchisq(.95,df=c(1,n-p))

[1]   3.841459 239.403439
```

4. *Cela ne semble pas bien parti. Au vu des graphiques ci-dessous, pourriez-vous en donner une possible cause?*

```
> hii=hatvalues(glm1);res1.D=residuals(glm1,type="deviance");
> res1.P=residuals(glm1,type="pearson")
> par(mfrow=c(2,2))
> plot(fitted(glm1),res1.D,cex=.7,pch=19)
> plot(fitted(glm1),res1.P,cex=.7,pch=19)
> plot(fitted(glm1),abs(res1.D)^.5,cex=.7,pch=19);plot(hii,type='h')
> abline(h=c(2,3)*p/n)
```



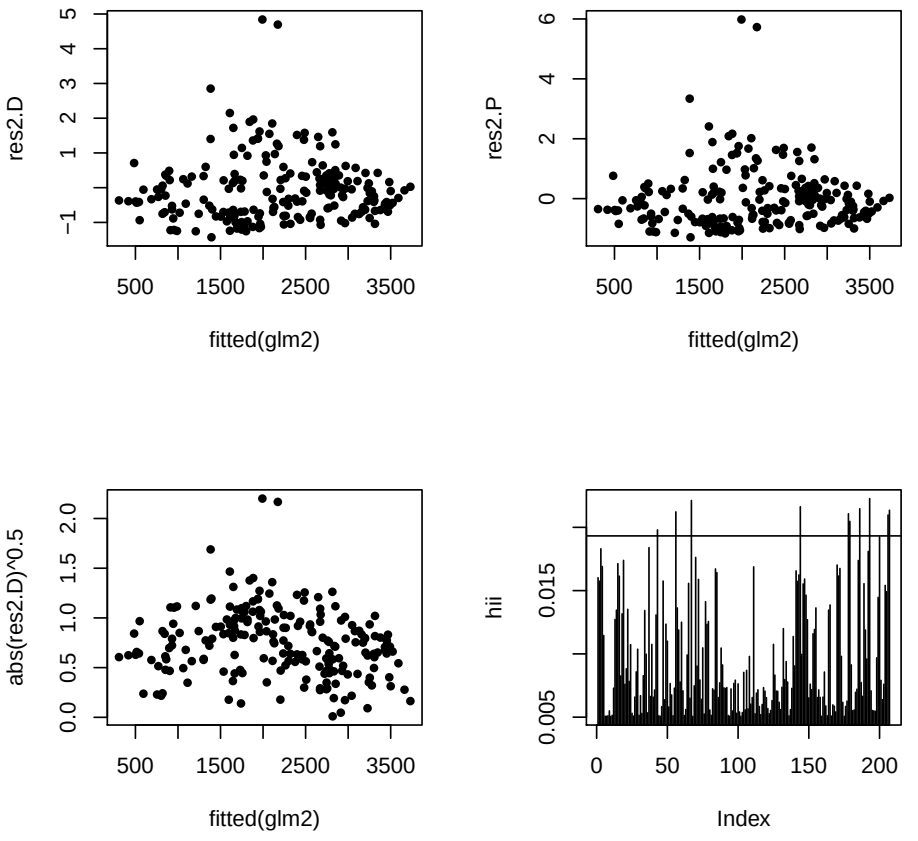

```
> par(mfrow=c(1,1))
```

5. *Pour contrer l'un des problèmes précédents, on propose la procédure ci-dessous. Détaillez cette procédure, et son effet (5 lignes maximum)*

```
> s2=sum(res1.P^2)/(n-p)
> s2

[1] 206.9092

> wt=rep(1/s2,n)
> glm2=glm(PJC~log(Berri),data=velo2[i.app,],family=poisson,weights=wt)
```

```
> hii=hatvalues(glm2)
> res2.D=residuals(glm2,type="deviance")
> res2.P=residuals(glm2,type="pearson")
> par(mfrow=c(2,2))
> plot(fitted(glm2),res2.D,cex=.7,pch=19)
```

```
> plot(fitted(glm2),res2.P,cex=.7,pch=19)
> plot(fitted(glm2),abs(res2.D)^.5,cex=.7,pch=19);plot(hii,type='h')
> abline(h=c(2,3)*p/n)
```




```
> par(mfrow=c(1,1))
```

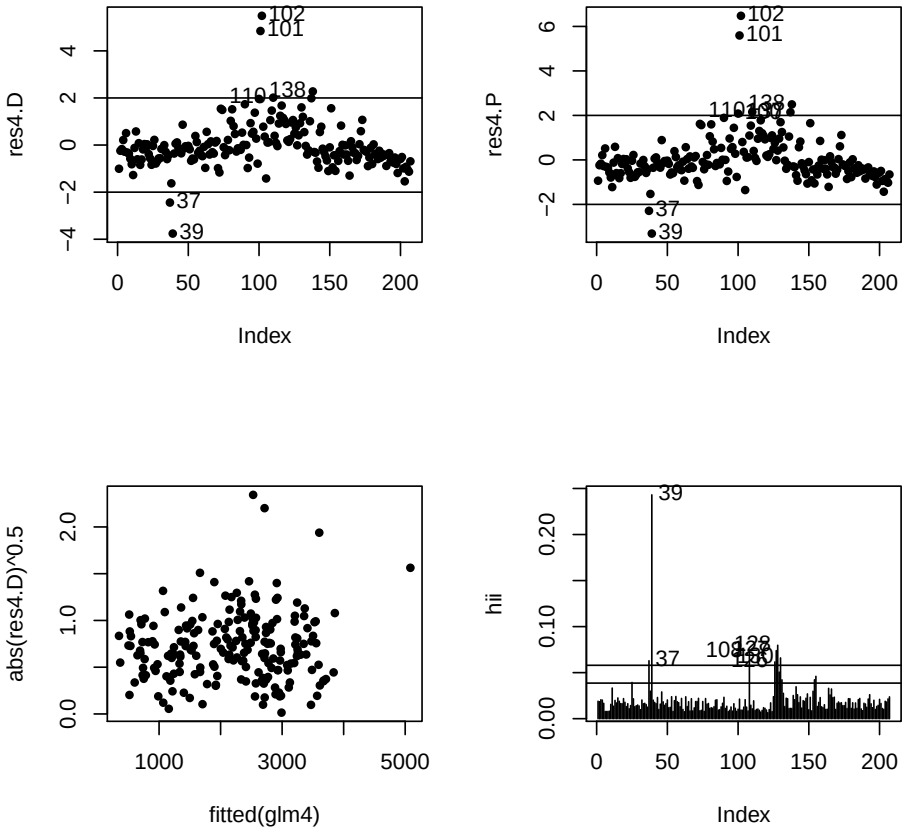
6. *On décide d'incorporer maintenant les deux autres variables en adoptant la même stratégie. Au vu de l'instruction* **R** *ci-dessous, comment est modélisée* $\mathrm{E}(\mathrm{PJC}_i)$ *et* $\mathrm{Var}(\mathrm{PJC}_i)$*?*

```
> glm3=glm(PJC~log(Berri)+log(Maisonneuve)+log(Rachel),
+          data=velo2[i.app,],family=poisson)
> p=4
> s2=sum(residuals(glm3,type="pearson")^2)/(n-p);wt=rep(1/s2,n)
> s2

[1] 114.4341

> glm4=glm(PJC~log(Berri)+log(Maisonneuve)+log(Rachel),
```

```
+         data=velo2[i.app,],family=poisson,weights=wt)
```

7. *Analysez les graphiques ci-dessous (5 lignes maximum)*

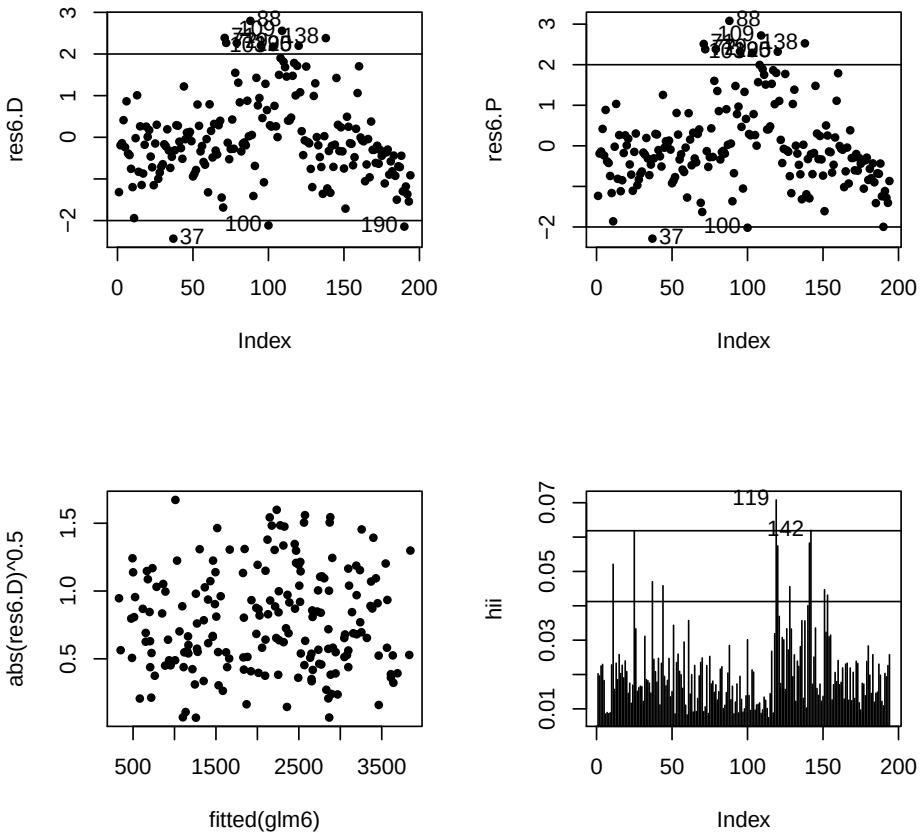
```
> require(car)
> hii=hatvalues(glm4)
> res4.D=residuals(glm4,type="deviance")
> res4.P=residuals(glm4,type="pearson")
> par(mfrow=c(2,2))
> plot(res4.D,cex=.7,pch=19)
> abline(h=c(-2,2))
> id.D=which(abs(res4.D)>2); tmp=showLabels(1:n,res4.D,method=id.D)
> plot(res4.P,cex=.7,pch=19)
> abline(h=c(-2,2))
> id.P=which(abs(res4.P)>2); tmp=showLabels(1:n,res4.P,method=id.P)
> plot(fitted(glm4),abs(res4.D)^.5,cex=.7,pch=19);plot(hii,type='h')
> abline(h=c(2,3)*p/n)
> id.H=which(hii>3*p/n); tmp=showLabels(1:n,hii,method=id.H)
```



```
> par(mfrow=c(1,1))
```

```
> id.del=unique(c(id.D,id.P,id.H))
> glm5=glm(PJC~log(Berri)+log(Maisonneuve)+log(Rachel),
+         data=velo2[i.app,][-id.del,],family=poisson)
> p=4;n=nrow(velo2[i.app,][-id.del,])
> s2=sum(residuals(glm5,type="pearson")^2)/(n-p);wt=rep(1/s2,n)
> s2

[1] 53.0044

> glm6=glm(PJC~log(Berri)+log(Maisonneuve)+log(Rachel),
+         data=velo2[i.app,][-id.del,],family=poisson,weights=wt)
```

```
> hii=hatvalues(glm6)
> res6.D=residuals(glm6,type="deviance")
> res6.P=residuals(glm6,type="pearson")
> par(mfrow=c(2,2))
> plot(res6.D,cex=.7,pch=19)
> abline(h=c(-2,2))
> id.D=which(abs(res6.D)>2); tmp=showLabels(1:n,res6.D,method=id.D)
> plot(res6.P,cex=.7,pch=19)
> abline(h=c(-2,2))
> id.P=which(abs(res6.P)>2); tmp=showLabels(1:n,res6.P,method=id.P)
> plot(fitted(glm6),abs(res6.D)^.5,cex=.7,pch=19);plot(hii,type='h')
> abline(h=c(2,3)*p/n)
> id.H=which(hii>3*p/n); tmp=showLabels(1:n,hii,method=id.H)
```

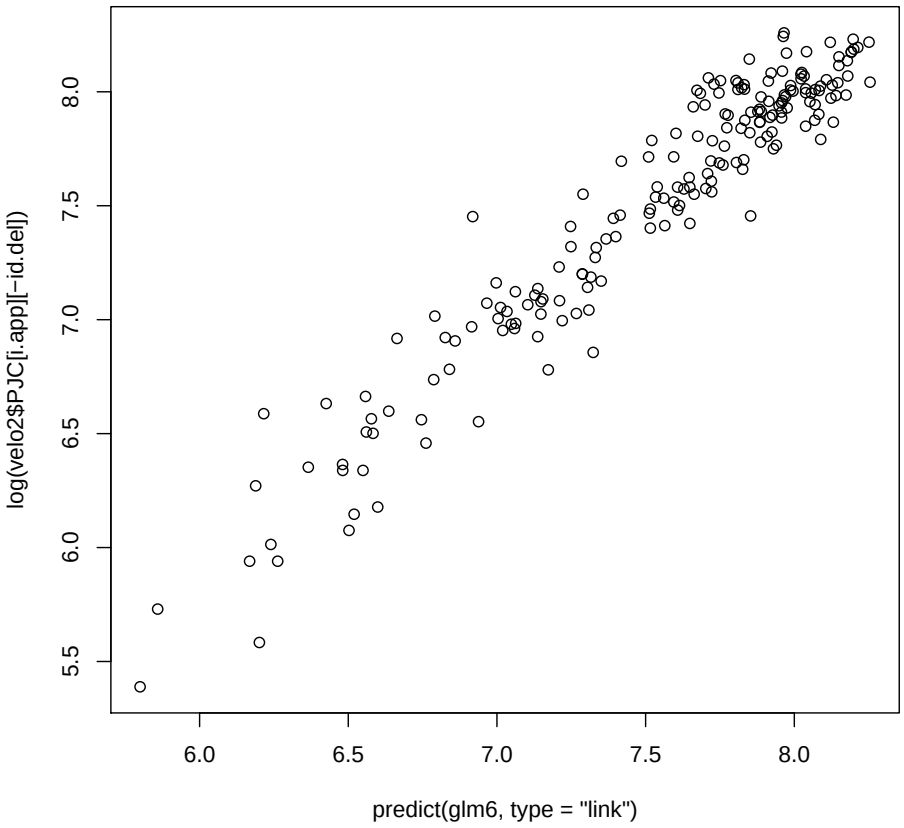



```
> par(mfrow=c(1,1))
```

8. *Considérant la dernière étape, comme une étape intéressante (même si elle peut être améliorée), au vu de la sortie, quelle fréquentation approximative du* **PJC** *prévoierait-on si, sur chacune des trois autres pistes on observait 2500 cyclistes. Votre résultat est-il cohérent avec le graphique ci-dessous?*

```
> summary(glm6)


Call:
glm(formula = PJC ~ log(Berri) + log(Maisonneuve) + log(Rachel),
    family = poisson, data = velo2[i.app, ][-id.del, ], weights = wt)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4332  -0.6684  -0.1717   0.3643   2.7961
```

```
Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      -4.335864   0.370247 -11.711  < 2e-16 ***
log(Berri)       -0.201080   0.083069  -2.421   0.0155 *
log(Maisonneuve)  0.029161   0.006959   4.191 2.78e-05 ***
log(Rachel)       1.609953   0.109245  14.737  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1802.68  on 193  degrees of freedom
Residual deviance:  185.41  on 190  degrees of freedom
AIC: 227.59

Number of Fisher Scoring iterations: 4

> plot(predict(glm6,type='link'),log(velo2$PJC[i.app][-id.del]))
```



9. *Il est maintenant temps de proposer une réponse à la problématique. Donnez l'instruction qui a permis de calculer le vecteur* **`p6`** *correspondant aux prédictions de la fréquentation du* **`PJC`** *pour les trois premiers mois de 2015.*

```
> ## p6= instruction R à produire
> summary(p6)

     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
  0.09394  15.24881  32.34132  47.43020  56.21228 296.51178

> round(mean(p6),1)

[1] 47.4
```

Ainsi, on aurait pu prévoir environ 47.4 cyclistes en moyenne par jour entre le 1er et le 31 mars 2015.