TD DE MODÈLES LINÉAIRES - FEUILLE 5
Projection

Exercice 1 Soit X de loi $\mathcal{N}(0, I_n)$ et δ un vecteur fixé de $\mathbb{R}^n$. On définit la variable aléatoire réelle Y par

$$Y = \| X + \delta \|^2 = \sum_{i=1}^{n} (X_i + \delta_i)^2.$$

On dit que Y suit un $\chi^2(n, \| \delta \|^2)$. Supposons $\delta \neq 0$. Calculer l'espérance et la variance de Y.

Exercice 2 Soit X de loi $\mathcal{N}(0, I_n)$. Soit P une projection orthogonale de rang r. Montrer que $X'PX$ suit un χ^2_r.

Exercice 3 Soit Z une matrice de taille $n \times q$ de rang q et soit X une matrice $n \times p$ de rang p composée des q vecteurs colonne de Z et de $p - q$ autres vecteurs linéairement indépendants. Nous considérons les deux modèles suivants

$$\begin{aligned} Y &= Z\beta + \varepsilon \\ Y &= X\beta^* + \eta \end{aligned}$$

Supposons pour simplifier que la constante ne fait partie d'aucun modèle. Notons respectivement P_X et P_Z les projections orthogonales sur les sous-espaces $\mathcal{M}(X)$ et $\mathcal{M}(Z)$ engendrés par les p colonnes de X et les q colonnes de Z. Notons enfin $P_{X \cap Z^\perp}$ la projection orthogonale sur $\mathcal{M}(X) \cap \mathcal{M}(Z)^\perp$, orthogonal de $\mathcal{M}(Z)$ dans $\mathcal{M}(X)$, soit

$$\mathbb{R}^n = \mathcal{M}(X) \overset{\perp}{\oplus} \mathcal{M}(X)^\perp = \left(\mathcal{M}(Z) \overset{\perp}{\oplus} (\mathcal{M}(X) \cap \mathcal{M}(Z)^\perp)\right) \overset{\perp}{\oplus} \mathcal{M}(X)^\perp.$$

1. Exprimer $\|P_X Y\|^2$ en fonction de $\|P_Z Y\|^2$ et $\|P_{X \cap Z^\perp} Y\|^2$.
2. Comparer alors les coefficients de détermination des deux modèles, soit R^2_Z et R^2_X.
3. De facon générale, qu'en déduire quant à l'utilisation du R^2 pour le choix de variables ?

Exercice 4 (Régression sur variables centrées) Nous considérons le modèle de régression linéaire

$$Y = X\beta + \varepsilon, \tag{1}$$

où $Y \in \mathbb{N}^n$, X est une matrice de taille $n \times p$ de rang p, $\beta \in \mathbb{N}^p$ et $\varepsilon \in \mathbb{N}^n$. La première colonne de X est le vecteur constant $\mathbb{1}$. X peut donc s'écrire $X = [\mathbb{1}, Z]$ où $Z = [X_2, \ldots, X_p]$ est la matrice $n \times (p-1)$ des $(p-1)$ derniers vecteurs colonnes de X. Le modèle peut donc s'écrire sous la forme :

$$Y = \beta_1 \mathbb{1} + Z\beta_{(1)} + \varepsilon,$$

où β_1 est la première coordonnée du vecteur β et $\beta_{(1)}$ représente β privé de sa première coordonnée.

1. Donner $P_{\mathbb{1}}$, matrice de projection orthogonale sur le sous-espace engendré par le vecteur $\mathbb{1}$.
2. En déduire la matrice de projection orthogonale $P_{\mathbb{1}^\perp}$ sur le sous-espace $\mathbb{1}^\perp$ orthogonal au vecteur $\mathbb{1}$.
3. Calculer $P_{\mathbb{1}^\perp} Z$.
4. En déduire que l'estimateur de β des Moindres Carrés Ordinaires du modèle (1) peut être obtenu en minimisant par les MCO le modèle suivant :

$$\tilde{Y} = \tilde{Z}\beta_{(1)} + \eta, \tag{2}$$

   où $\tilde{Y} = P_{\mathbb{1}^\perp} Y$ et $\tilde{Z} = P_{\mathbb{1}^\perp} Z$.

5. Ecrire la SCR estimée dans le modèle (2) en fonction des variables du modèle (2). Vérifier que la SCR du modèle (2) est identique à celle qui serait obtenue par l'estimation du modèle (1).

Exercice 5 On considère le modèle de régression $y_i = \beta_1 + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \varepsilon_i, \quad 1 \leq i \leq n$, que l'on écrit sous la forme $Y = X + \varepsilon$. Les $x_{i,j}$ sont des variables exogénes du modèle, les ε_i sont des variables aléatoires indépendantes, de loi normale centrée admettant la même variance σ^2. On a observé :

$$X^T X = \begin{bmatrix} 30 & 20 & 0 \\ 20 & 20 & 0 \\ 0 & 0 & 10 \end{bmatrix}, \quad X^T Y = \begin{bmatrix} 15 \\ 20 \\ 10 \end{bmatrix}, \quad Y^T Y = 59.5.$$

1. Déterminer n, la moyenne des $x_{i,3}$, le coefficient de corrélation des $x_{i,2}$ et des $x_{i,3}$.
2. Estimer $\beta_1, \beta_2, \beta_3, \sigma^2$ par la méthode des moindres carrés ordinaires.
3. Calculer pour β_2 un intervalle de confiance à 95% et tester l'hypothése $\beta_3 = 0.8$ au niveau 10%.
4. Tester $\beta_2 + \beta_3 = 3$ contre $\beta_2 + \beta_3 \neq 3$, au niveau 5%.
5. Calculer y et déduire le coefficient de détermination ajusté R_a^2.
6. Construire un intervalle de prévision à 95% de y_{n+1} si $x_{n+1,2} = 3$ et $x_{n+1,3} = 0.5$.