TD DE MODÈLES LINÉAIRES - FEUILLE 4
Régression linéaire multiple

Exercice 1 (Rôle de la constante) Soit X une matrice de dimensions $n \times p$. Soit $\hat{Y}$ la projection orthogonale d'un vecteur Y de $\mathbb{N}^n$ sur l'espace engendré par les colonnes de X. On note $\mathbb{1}$ le vecteur de $\mathbb{N}^n$ uniquement composé de la valeur 1.

1. Exprimer le produit scalaire $\langle Y, \mathbb{1} \rangle$ en fonction des y_i.
2. Soit $\hat{\varepsilon} = Y - \hat{Y}$ et supposons que la constante fait partie du modéle, c'est-à-dire que la première colonne de X est $\mathbb{1}$. Que vaut $\langle \hat{\varepsilon}, \mathbb{1} \rangle$?
3. En déduire que lorsque la constante fait partie du modèle, $\sum_{i=1}^n y_i = \sum_{i=1}^n \hat{y}_i$.

Exercice 2 On considère les trois modèles de régression suivants

(a) $Y_k = b_1 + b_2 \ln X_k + \varepsilon_k$, $k = 1, ..., n$,

(b) $Y_k = b_1 + b_2 X_k + \varepsilon_k$, $k = 1, ..., n$,

(c) $Y_k = b_1 + b_2 X_k + b_3 X_k^2 + \varepsilon_k$, $k = 1, ..., n$,
où les ε_k sont des $\mathcal{N}(0, \sigma^2)$ indépendantes, $k = 1, ..., n$.

1. Calculer les estimateurs linéaires sans biais de variance minimale pour chacun des trois modèles.
2. Donner des estimateurs de σ^2 et leurs propriétés (ne pas expliciter les calculs).

Exercice 3 Soient m et λ deux paramètres réels. Soit le modèle linéaire gaussien défini par les équations

$\mathbb{E}(X_1) = m + \lambda$ et $\forall i \in \{2, ..., n\}, \mathbb{E}(X_i) = m$.

Calculer l'ESBVM par deux méthodes différentes.

Exercice 4 On considère un vecteur aléatoire normal $X = (X_1, ..., X_n)'$ de loi $\mathcal{N}(M, I_n)$, où $M = (\mu_1, ..., \mu_n)', n \geq 2$.

On suppose que M vérifie les équations

$$\sum_{i=1}^n \mu_i = 0 \qquad \text{et} \qquad \sum_{i=1}^n (-1)^i \mu_i = 0.$$

1. Déterminer l'ESBMV de M.
2. Tester l'hypothèse M = 0 au seuil α.
3. A.N. $n = 5, X_1 = 2, X_2 = 0, X_3 = 4, X_4 = -1, X_5 = -2, \alpha = 0.05$.

Exercice 5 (Minimisation de l'erreur de prévision) Soit un échantillon de n couples de réels $(x_i, y_i)_{1 \leq i \leq n}$ pour le modèle de régression linéaire simple $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, où les erreurs ε_i sont supposées centrées décorrélées et de même variance σ^2. On estime $\beta = (\beta_0, \beta_1)$ par la méthode des moindres carrés ordinaires, ce qui donne $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)$.

Soit x_{n+1} une nouvelle valeur de la variable explicative pour laquelle on veut prédire la variable réponse y_{n+1}. L'erreur de prévision est par définition

$$\hat{\varepsilon}_{n+1} = y_{n+1} - \hat{y}_{n+1} = y_{n+1} - (\hat{\beta}_0 + \hat{\beta}_1 x_{n+1}).$$

On montre que sa variance vaut

$$\mathrm{Var}(\hat{\varepsilon}_{n+1}) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right).$$

En notant X la matrice $n \times 2$ définie par

$$X = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

nous avons de facon générale

$$\mathrm{Var}(\hat{\varepsilon}_{n+1}) = \sigma^2 \left(1 + [1, x_{n+1}](X'X)^{-1}[1, x_{n+1}]'\right).$$

A partir de cette formule, il est clair que l'erreur de prévision est minimale (en moyenne) lorsque $x_{n+1} = \bar{x}$, la variance de l'erreur valant alors $\sigma^2(1 + 1/n)$.

Le but de cet exercice est de généraliser le résultat précédent. Nous considérons désormais un échantillon $(x_i', y_i)_{1 \leq i \leq n}$, où $x_i' = [1, z_i']$ avec $z_i' = [x_{i1}, \ldots, x_{ip}]$. En notant $\mathbb{1}$ le vecteur de taille n uniquement composé de 1, nous adoptons l'écriture matricielle :

$$X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix} = \left[\begin{array}{cc} 1 & z_1' \\ \hline \vdots & \vdots \\ \hline 1 & z_n' \end{array}\right] = \left[\begin{array}{c|c|c|c} \mathbb{1} & Z_1 & \cdots & Z_p \end{array}\right] = \left[\begin{array}{c|c} \mathbb{1} & Z \end{array}\right],$$

où Z est donc une matrice de taille $n \times p$. Les moyennes de ses colonnes $Z_1, \ldots, Z_p$ sont regroupées dans le vecteur ligne $\bar{x}' = [\bar{x}_1, \ldots, \bar{x}_p]$. Enfin, on considère comme précédemment le modèle de régression linéaire

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = x_i'\beta + \varepsilon_i,$$

où les erreurs ε_i sont supposées centrées indépendantes et de même variance σ^2. Matriciellement, ceci s'écrit donc $Y = X\beta + \varepsilon$, avec X donnée ci-dessus et supposée telle que $X'X$ est inversible.

1. Ecrire la matrice $X'X$ sous forme de 4 blocs faisant intervenir Z, $\bar{x}$ et la taille n de l'échantillon.
2. On rappelle la formule d'inversion matricielle par blocs : Soit M une matrice inversible telle que

   $$M = \left[\begin{array}{c|c} T & U \\ \hline V & W \end{array}\right]$$

   avec T inversible, alors $Q = W - VT^{-1}U$ est inversible et l'inverse de M est :

   $$M^{-1} = \left[\begin{array}{c|c} T^{-1} + T^{-1}UQ^{-1}VT^{-1} & -T^{-1}UQ^{-1} \\ \hline -Q^{-1}VT^{-1} & Q^{-1} \end{array}\right].$$

   Ecrire la matrice $(X'X)^{-1}$ sous forme de 4 blocs dépendant de n, $\bar{x}$ et Γ^{-1}, où $\Gamma = \frac{1}{n}Z'Z - \bar{x}\bar{x}'$.
3. Soit $x_{n+1}' = [1, z_{n+1}']$ une nouvelle donnée. Montrer que la variance de l'erreur de prévision est égale à

   $$\mathrm{Var}(\hat{\varepsilon}_{n+1}) = \sigma^2 \left(1 + \frac{1}{n} + \frac{1}{n}(z_{n+1} - \bar{x})'\Gamma^{-1}(z_{n+1} - \bar{x})\right).$$

4. On admet pour l'instant que $\Gamma = \frac{1}{n}Z'Z - \bar{x}\bar{x}'$ est symétrique définie positive (on rappelle que S est symétrique définie positive si $S' = S$ et si pour tout vecteur x non nul, $x'Sx > 0$). Pour quelle nouvelle donnée x_{n+1}' la variance de l'erreur de prévision est-elle minimale ? Que vaut alors cette variance ?
5. Justifier le fait que si $X'X$ est inversible, alors Γ est bien symétrique définie positive.