Université de Lyon 2 — Modèle linéaire
Licence 3 MIASHS — Année 2022-23

Contrôle 1

Exercice 1 On considère le

$$Y_i = \beta_0 + \beta_1 d_i + \beta_2 h_i + \beta_3 a_i + \varepsilon_i \text{ pour } i = 1 \dots 20,$$

où Y_i est une variable aléatoire représentant le nombre de cas d'une certaine pathologie dans la $i^{\text{ème}}$ région française, d_i, h_i et a_i désignent respectivement la densité urbaine moyenne, l'humidité moyenne et l'altitude moyenne de cette $i^{\text{ème}}$ région, les ε_i sont des termes d'erreur aléatoires supposés indépendants et Gaussiens de variance $\text{var}(\varepsilon_i) = \sigma^2$ pour tout $i = 1, \dots, n = 20$.

Ce modèle s'écrit sous la forme matricielle classique :

$$Y = X\beta + \varepsilon$$

avec $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)^\top$ et $\varepsilon = (\varepsilon_1, \dots, \varepsilon_{20})^\top$. On a calculé

$$X^\top y = (40.7378, 527.2645, 111.4137, 22.89)^\top. \tag{1}$$

$$X^\top X = \begin{bmatrix} 20 & 226.8 & 54.1 & 12.7 \\ 226.8 & 3135.8 & 757.06 & 133.98 \\ 54.1 & 757.06 & 190.97 & 32.339 \\ 12.7 & 133.98 & 32.339 & 9.3258 \end{bmatrix}, \tag{2}$$

$$(X^\top X)^{-1} = \begin{bmatrix} 0.9711 & -0.0541 & 0.0769 & -0.8115 \\ -0.0541 & 0.011 & -0.0338 & 0.0333 \\ 0.0769 & -0.0338 & 0.1281 & -0.064 \\ -0.8115 & 0.0333 & -0.064 & 0.9557 \end{bmatrix} \tag{3}$$

et

$$\varepsilon = \begin{bmatrix} \varepsilon_1, \varepsilon_2 \end{bmatrix}^\top \tag{4}$$

où

$$\varepsilon_1 = \begin{bmatrix} 1.4 & 1.2 & 1.2 & -0.8 & 2.2 & -0.2 & -3.1 & 6.2 & 0.1 & 0 \end{bmatrix}$$

et

$$\varepsilon_2 = \begin{bmatrix} 1 & -2 & 0.5 & 0.7 & -0.9 & 1.5 & -3 & -3 & 1.2 & 0.4 \end{bmatrix}$$

1. Après avoir rappelé la définition de l'estimateur des moindres carrés $\hat{\beta}$ de β, calculer sa valeur.
2. Quelles sont les propriétés statistiques connues de β (biais, matrice de variance-covariance $\mathrm{Var}(\hat{\beta})$,...) ?
3. Déterminer un estimateur $\widehat{\sigma^2}$ sans biais de la variance σ^2, donner sa valeur et en déduire une estimation sans biais de la matrice de variance-covariance $\mathrm{Var}(\beta)$.
4. Faire le test de nullité pour chaque coefficient β_0, β_1, β_2 et β_3. On fera un test au risque d'erreur $\alpha = 5\%$
5. Calculez le BIC pour le modèle où toutes les variables sont considérées et décrire quels calculs vous pourriez faire pour comparer tous les modèles possibles obtenus en incluant ou pas chacune des variables.

Exercice 2 On s'intéresse dans cet exercice à la construction d'une fonction d'influence pour la détection d'outliers.

On s'intéresse plus particulièrement au modèle linéaire :

$$y_i = x_i^\top \beta + \varepsilon_i \tag{5}$$

$i = 1, \ldots, n$, où x_i est un vecteur colonne de $\mathbb{R}^p$. On voudrait pouvoir détecter la donnée en rouge dans la figure ci-dessous, qui est une donnée abérrante. Notons

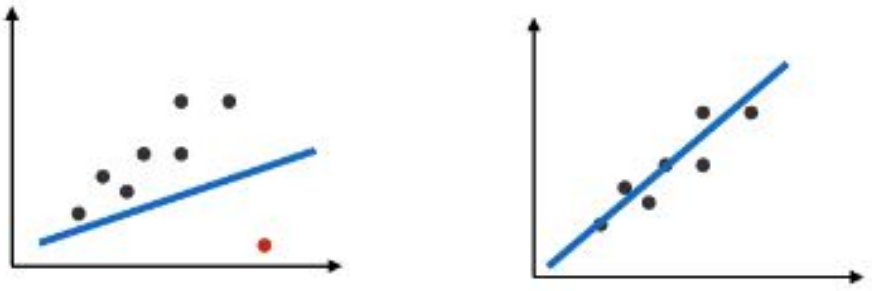

FIGURE 1 – Indice de Cook

$$h_{i,i} = x_i^\top \left(X^\top X\right)^{-1} x_i.$$

Cette quantité est appelée "leverage score" et nous permet d'introduire la "distance de Cook" comme le nombre

$$D_i = \frac{(Y_i - x_i^\top \hat{\beta})^2}{(p+1)s^2} \; \frac{h_{ii}}{(1-h_{ii})^2}$$

pour $i = 1, \ldots, n$, où

$$s^2 = \frac{1}{n-p-1} \sum_{i=1}^{n} (Y_i - x_i^\top \hat{\beta})^2 \tag{6}$$

est l'estimation non-biasée de la variance σ^2 du bruit. Cette distance nous indique la sensibilité de la régression vis à vis de la $i^{\text{ième}}$ donnée (x_i, y_i) et focalise habituellement sur la détection de la présence d'outliers (voir la Figure 1).

Passons maintenant à un exemple. Un père a deux garçons, et s'inquiète de la croissance de son cadet qu'il trouve petit. Il décide de faire un modèle familial à partir des mesures de taille en fonction de l'âge de l'aîné :

age	3	4	5	7	8	9	10	11	12
taille	96	104.8	110.3	121.9	127.4	130.8	136	139.7	144.5

1. Calculer les paramètre du modèle de régression pour ces données.

2. Calculer l'estimation de la variance à partir de l'estimateur non-biaisé.

3. Calculer l'indice de Cook pour la donnée numéro 3 et la donnée numéro 9. Attention, ici il faut comprendre les données des individus dont l'âge est égal à 3 et 9 respectivement.

4. Recalculer le modèle de régression sans ces données.

5. Comparer l'erreur de prédiction pour une nouvelle donnée $(6, 115.3)$ pour
   — le modèle avec les données 3 et 9
   — le modèle sans les données 3 et 9
   .