TD n° 3 : Régression linéaire multiple 1

Exercice 1. Soient Y_1, Y_2 et Y_3 3 *var* indépendantes telles que

$$\begin{cases} Y_1 = 2a - b + \epsilon_1, \\ Y_2 = a + b + \epsilon_2, \\ Y_3 = -a + 2b + \epsilon_3, \end{cases}$$

où ϵ_1, ϵ_2 et ϵ_3 sont 3 *var iid* suivant chacune la loi normale $\mathcal{N}(0, \sigma^2)$. Les paramètres a, b et σ sont des réels inconnus.

1. Écrire le modèle linéaire associé sous la forme matricielle usuelle : $Y = X\beta + \epsilon$, en indiquant ce que sont ici Y, X, β et ϵ.
2. Calculer l'*emco* $\widehat{a}$ de a et l'*emco* $\widehat{b}$ de b.
3. Calculer $\mathbb{C}(\widehat{a}, \widehat{b})$. Est-ce que $\widehat{a}$ et $\widehat{b}$ sont indépendantes ?

Exercice 2. On souhaite expliquer le nombre d'années d'études d'un enfant (unique) (variable Y) à partir du nombre d'années d'études de la mère (variable X_m) et du nombre d'années d'études du père (variable X_p). On dispose du nombre d'années d'études de $n = 20$ enfants ainsi que des données concernant leurs parents. En notant $(y_1, x_{m,1}, x_{p,1}), \ldots, (y_n, x_{m,n}, x_{p,n})$ ces n observations de (Y, X_m, X_p), on a les résultats numériques suivants :

$\sum_{i=1}^{n} x_{m,i}^2$	$\sum_{i=1}^{n} x_{p,i}^2$	$\sum_{i=1}^{n} y_i x_{m,i}$	$\sum_{i=1}^{n} y_i x_{p,i}$	$\sum_{i=1}^{n} x_{m,i} x_{p,i}$
228	202	184	158	144

On adopte le modèle de *rlm* (sans terme constant pour simplifier) : pour tout $i \in \{1, \ldots, n\}$, y_i est une réalisation de

$$Y_i = a x_{m,i} + b x_{p,i} + \epsilon_i,$$

où $\epsilon_1, \ldots, \epsilon_n$ sont n *var iid* suivant chacune la loi normale $\mathcal{N}(0, \sigma^2)$. Les paramètres a, b et σ sont des réels inconnus.

1. Écrire le modèle linéaire associé sous la forme matricielle usuelle : $Y = X\beta + \epsilon$, en indiquant ce que sont ici Y, X, β et ϵ.
2. Calculer l'*emco* $\widehat{\beta}$ de β. Donner l'*emco* ponctuel b de β.
3. Donner un estimateur $\widehat{\sigma}^2$ sans biais de σ^2. En posant $y = (y_1, \ldots, y_n)^t$ et en adoptant les notations de la question 1, on donne $\|y - X\left(X^tX\right)^{-1}X^ty\|^2 = 16.776$, où $\|.\|$ désigne la norme euclidienne dans $\mathbb{R}^n$. Donner l'estimation ponctuelle de σ.
4. Quelles sont les hypothèses adaptées à la situation suivante : "on souhaite affirmer, avec un faible risque de se tromper, qu'en moyenne, le nombre d'années d'études de la mère et celui du père n'ont pas le même effet sur le nombre d'années d'études de leur enfant" ?

Exercice 3. On souhaite expliquer une variable quantitative Y à partir de 3 variables quantitatives X_1, X_2 et X_3. On observe $n = 18$ valeurs de (Y, X_1, X_2, X_3) notées $(y_1, x_{1,1}, x_{2,1}, x_{3,1}), \ldots,$ $(y_n, x_{1,n}, x_{2,n}, x_{3,n})$. On adopte le modèle de *rlm* : pour tout $i \in \{1, \ldots, n\}$, y_i est une réalisation de

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \epsilon_i,$$

où $\epsilon_1, \ldots, \epsilon_n$ sont n *var iid* suivant chacune la loi normale $\mathcal{N}(0, \sigma^2)$. Les paramètres β_0, β_1, β_2, β_3 et σ sont des réels inconnus.

1. Écrire le modèle linéaire associé sous la forme matricielle usuelle : $Y = X\beta + \epsilon$, en indiquant ce que sont ici Y, X, β et ϵ.

En posant $y = (y_1, \ldots, y_n)^t$ et en adoptant les notations de la question 1, les données recueillies donnent :

$$(X^tX)^{-1} = 10^{-4} \begin{pmatrix} 555.5 & 0 & 0 & 0 \\ 0 & 3.6 & -14.7 & 9.2 \\ 0 & -14.7 & 1219 & 3.8 \\ 0 & 9.2 & 3.8 & 1000 \end{pmatrix}, \quad b = (X^tX)^{-1} X^t y = \begin{pmatrix} -5.92 \\ 0.133 \\ 0.55 \\ 2.1 \end{pmatrix}$$

et $\|y - Xb\|^2 = 1.4$, où $\|.\|$ désigne la norme euclidienne dans $\mathbb{R}^n$. Soit $\widehat{\beta} = (\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2, \widehat{\beta}_3)^t$, l'*emco* de $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)^t$.

2. Exprimer $\widehat{\beta}$ en fonction de X et Y. Donner l'*emco* ponctuel de β.
3. Quelle est la loi de $\widehat{\beta}$, ainsi que celle de chacune de ses composantes : $\widehat{\beta}_0$, $\widehat{\beta}_1$, $\widehat{\beta}_2$ et $\widehat{\beta}_3$?
4. Est-ce que $\widehat{\beta}_2$ et $\widehat{\beta}_3$ sont indépendantes ? Calculer $\mathbb{V}(\widehat{\beta}_2 - 2\widehat{\beta}_3)$.
5. Donner un estimateur sans biais de σ^2 en fonction de X et Y. En déduire une estimation ponctuelle de σ^2.
6. Donner l'*emco* ponctuel de β_2 et une estimation ponctuelle de la variance de $\widehat{\beta}_2$.
7. On considère les hypothèses :

$$H_0 : \beta_2 = 0 \qquad \text{contre} \qquad H_1 : \beta_2 \neq 0.$$

   Peut-on rejeter H_0 au risque 5% ?

Exercice 4. On considère le modèle : pour tout $(u, i) \in \{1, 2\} \times \{1, \ldots, 2m\}$, avec $m \in \mathbb{N}^*$,

$$Y_{u,i} = a_u + b_u x_{u,i} + \epsilon_{u,i},$$

où $x_{u,i} = (-1)^{u+i}$ et $\epsilon_{1,1}, \ldots, \epsilon_{1,2m}, \epsilon_{2,1}, \ldots, \epsilon_{2,2m}$ sont $4m$ *var iid* suivant chacune la loi normale $\mathcal{N}(0, \sigma^2)$. Les paramètres a_1, a_2, b_1, b_2 et σ sont inconnus.

1. Déterminer l'*emco* $\widehat{\beta}$ de $\beta = (a_1, a_2, b_1, b_2)^t$.
2. Donner la loi de $K = \dfrac{2m}{\sigma^2}\|\widehat{\beta} - \beta\|^2$, et la loi de $K^* = K + \dfrac{1}{\sigma^2}\|Y - X\widehat{\beta}\|^2$, où $\|.\|$ désigne la norme euclidienne dans $\mathbb{R}^{4m}$.