Université Mohamed Khider – Biskra
Département de Mathématiques
Module: Analyse de Données-Master 1 -2021/2022
TD N°2 : Régression Linéaire Multiple

Exercice 1. Dans une étude de régression multiple comportant quatre variables explicatives, nous avons obtenu le tableau d'AV pour 19 observations, comme suit :

Variation	*ddl*	*SC*	*MC*	*F*
Expliquée				
Résiduelle		2152.7		
Total		7654		

1) Complétez le tableau d'AV et testez la signification globale du modèle au seuil 5%.

2) Calculez le coefficient de détermination du modèle.

Exercice 2. On s'intéresse à l'effet du nombre d'années d'études des parents (M : *mère* et P : *père*) sur le nombre d'années d'études de leurs enfants noté Y. On dispose du nombre d'années d'études de 26 familles (enfant, mère et père). On décide d'ajuster ces données par le modèle linéaire suivant

$$y_i = aM_i + bP_i + \varepsilon_i, \quad \varepsilon_i \sim N\left(0, \sigma_\varepsilon^2\right), \qquad i = 1, 2, ..., 26.$$

On vous donne les sommes suivantes :

$$\sum M_i^2 = 288, \qquad \sum P_i^2 = 202, \qquad \sum M_i P_i = 144, \qquad \sum y_i M_i = 184, \qquad \sum y_i P_i = 158.$$

1) Ecrire ce modèle sous forme matricielle : $Y = XB + \varepsilon$, en précisant Y, X et B.

2) Donnez les matrices $(X^t X)$ et $X^t Y$ à l'aide des valeurs numériques des sommes ci-dessus.

3) Calculez l'estimateur $\widehat{B}$ de B.

4) Calculez $\hat{\sigma}_\varepsilon^2 = S^2$, sachant que $\sum \hat{\varepsilon}_i^2 = 36$.

5) Calculez les variances estimées de chacun des coefficients : $\hat{\sigma}_a^2$ et $\hat{\sigma}_b^2$.

6) Tester la signification des paramètres a et b à 90%.

7) Donner un intervalle de confiance à 90% des paramètres a et b.

8) Le nombre d'années d'études des mères à un effet sur celui de leur enfants ?

9) Le nombre d'années d'études des pères à un effet sur celui de leur enfants ?

10) Quel est le nombre d'années d'études prévu pour un enfant, sachant que le nombre d'années d'études de ces parents a été 17 ans pour le père et 14 ans pour la mère ?

Exercice 3. Considérons la régression linéaire multiple avec constante et deux variables explicatives U et V :

$$Y_j = a + bU_j + cV_j + \varepsilon_j, \quad j = 1, ..., n \qquad \varepsilon \sim N\left(0, \sigma_\varepsilon^2\right)$$

et soient les résultats empiriques :

$$X^t Y = \begin{pmatrix} 11,5 \\ 24,8 \\ 17.3 \end{pmatrix}, \quad X^t X = \begin{pmatrix} 30 & 0 & 0 \\ 0 & 20 & 5 \\ 0 & 5 & 15 \end{pmatrix} \quad et \quad \sum_{i=1}^{n} \widehat{\varepsilon}_i^2 = 15,7.$$

1) Calculer la matrice inverse $\left(X^tX\right)^{-1}$. Puis estimer les paramètres du modèle : $B = (a, b, c)^t$.
2) Calculer les moyenne empiriques : $\overline{U}$, $\overline{V}$.
3) Calculer $\hat{\sigma}_{\varepsilon}^2 = S^2$.

4) Calculer les variances estimées de chacun des coefficients : $\hat{\sigma}_a^2, \hat{\sigma}_b^2$ et $\hat{\sigma}_c^2$

5) Donner les valeurs des covariances :

$$cov\left(\widehat{b}, \widehat{a}\right), \quad cov\left(\widehat{a}, \widehat{c}\right) \quad et \quad cov\left(\widehat{b}, \widehat{c}\right).$$

6) La variable explicative U est-elle à 90% significativement contributive dans l'explication de Y ?
7) Le coefficient c est-il à 90% significativement différent de 1 ?
8) Sachant que : $U_h = 13.3$, $V_h = 16.6$. Calculer la prévision $\widehat{Y}_h$ et son intervalle de confiance à 90%.