QCM et exercices : Modèle de régression linéaire simple et multiple

Questions 1

Soient deux variables X et Y de moyennes respectives 20 et 70 et d'écart-types respectifs 10 et 20.
On veut établir l'équation de la droite de régression linéaire de Y en fonction de X.

1. Quelle est éventuellement l'équation de cette droite ?
   (a) Y=1.29 X +43.5
   (b) Y=-1.29 X + 95.8
   (c) Y= -0.64 X +82
   (d) Aucune des 3 équations.
2. Donner le coefficient de détermination R^2 dans ce cas.

Devinette

Des étudiants voient leurs moyennes générale s'afficher. Ils disposent des notes de toutes les matières (contrôle continue et examens) mais n'ont aucune idée des coefficients des matières ni des pourcentages appliqués entre CC et Examens. Comment peuvent-t-ils procéder ?

Questions 2

On effectue une régression de y sur deux variables explicatives x et z à partir d'un échantillon de n individus, c.a.d que $X = \begin{pmatrix} 1 & x_1 & z_1 \\ \vdots & \vdots & \vdots \\ 1 & x_n & z_n \end{pmatrix}$. On a obtenu le résultat suivant : $X'X = \begin{pmatrix} 5 & 3 & 0 \\ 3 & 3 & 1 \\ 0 & 1 & 1 \end{pmatrix}$.

1. Que vaut n ?
2. Donner les moyennes de x et z.
3. Donner les variances de x et z.
4. Donner le coefficient de corrélation linéaire entre x et z.
5. La régression selon la MCO a donné le résultat suivant $y_i = -1 + 3x_i + 4z_i + e_i$ et la $SCR = 3$.
   (a) Calculer $X'Y$ et en déduire la moyenne de Y.
   (b) Calculer $\sum_{i=1}^{n}(y_i - \overline{y})^2$.
   (c) Calculer le coefficient de détermination R^2 et le coefficient de détermination ajusté.

Question 3

En regression multiple, après une estimation des paramètres, il s'avère qu'une variable explicative a un coefficient très très faible pourtant la variable est corrélée linéairement avec la variable réponse. Expliquez ?

• Contact : imen.kammoun@gmail.com

Corrections

Questions 1

$$\beta_0 = \overline{y} - \beta_1 \overline{x}$$

$$70 + 1.29 \cdot 20 = 95.8$$

$R^2 = \rho^2(x,z)$ avec $\rho(x,z) = \dfrac{cov(x,z)}{\sigma_x \sigma_z}$

or $\beta_1 = \dfrac{cov(x,z)}{\sigma_x^2}$
$R^2 = (-0.64)^2$

Devinette

Comme la moyenne est une combinaison linéaire des notes (CC et examens), on peut réaliser une régression linéaire multiple avec toutes les notes comme variables explicatives de la moyenne (variable réponse). A partir de l'estimation $\widehat{\beta}$ des paramètres, on peut déduire les coefficients des matières.

Remarque : La constante dans ce cas doit être égale à 0 et le coefficient de détermination R^2 égal à 1 vu que la moyenne est expliquée à 100% par toutes les notes.

Questions 2

$$X'X = \begin{pmatrix} n & \sum_{i=1}^n x_i & \sum_{i=1}^n z_i \\ - & \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i z_i \\ - & - & \sum_{i=1}^n z_i^2 \end{pmatrix}$$

$n = 5$
$\overline{x} = \frac{3}{5}$
$\overline{z} = 0$

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \overline{x})^2$$

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \overline{x})^2 = \frac{1}{n}(\sum_{i=1}^n x_i^2 - n\overline{x}^2) = 0.24$$

$$\sigma_z^2 = \frac{1}{n}\sum_{i=1}^n (z_i - \overline{z})^2 = \frac{1}{n}\sum_{i=1}^n z_i^2 = 0.2$$

$$\rho(x,z) = \frac{cov(x,z)}{\sigma_x \sigma_z}$$

$$cov(x,z) = \frac{1}{n}\sum_{i=1}^n (x_i - \overline{x})(z_i - \overline{z})$$

$$cov(x,z) = \frac{1}{n}\sum_{i=1}^n (x_i - \overline{x})(z_i - \overline{z}) = \frac{1}{n}(\sum_{i=1}^n x_i\ z_i - n\overline{x}\overline{z}) = 0.2\ \rho(x,z) = \frac{0.2}{0.489 \cdot 0.447} = 0.9.$$

$$\beta = \left({}^tX \cdot X\right)^{-1} \cdot\ {}^tX \cdot Y \text{ donc } {}^tX \cdot Y = \left({}^tX \cdot X\right) \cdot \beta = \begin{pmatrix} 5 & 3 & 0 \\ 3 & 3 & 1 \\ 0 & 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} -1 \\ 3 \\ 4 \end{pmatrix} = \begin{pmatrix} 4 \\ 10 \\ 7 \end{pmatrix}$$

• Contact : imen.kammoun@gmail.com

$$^tX \cdot Y = \begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i\ y_i \\ \sum_{i=1}^{n} x_i\ z_i \end{pmatrix}$$

donc $\overline{y} = \frac{4}{5} \sum_{i=1}^{n} (y_i - \overline{y})^2 = SCT = SCE + SCR.$

Or $SCE = \sum_{i=1}^{n} (\widehat{y_i} - \overline{y})^2$

avec $\widehat{y_i} = -1 + 3x_i + 4z_i$
A.N $SCE = 34.4$ donc $SCT = 37.4$
$R^2 = 1 - \frac{SCR}{SCT} = 0.91$

$\overline{R}^2 = 1 - \frac{\mathrm{SCR}/(n-(p+1))}{\mathrm{SCT}/(n-1)} = 0.84.$

Questions 3

En régression multiple, on peut avoir une variable explicative fortement corrélée avec la variable réponse quand on la prend toute seule. Si on rajoute à cette variable une autre variable explicative, en plus s'il y' a une forte corrélation linéaire entre ces variables explicatives et aussi entre la variable rajoutée et la variable réponse, le modèle peut garder seulement une de ces variables explicatives pour expliquer la variable réponse (ce qui s'explique par un coefficient très très faible d'une des variables) ! C'est normal il ne doit pas y avoir redondance.

• Contact : imen.kammoun@gmail.com