Régression multiple

Exemple et exercice

1

Exemple:

- Supposons que les services de police souhaitent établir un modèle de régression linéaire reliant la variable endogène « taux de criminalité juvénile » mesuré par un indicateur Y, à la densité de la population urbaine mesurée par un indicateur X_1 et au taux de scolarité X_2. On a relevé 5 observations:

2

Y	X_1	X_2
1	2	4
1	3	2
2	5	2
3	7	1
3	8	1

3

Question:

- La spécification de l'équation de régression retenue est $Y=a_1X_1+a_2X_2+b+\varepsilon$; donnez une estimation des paramètres de cette équation.

4

Réponse:

- Nous avons donc à déterminer les paramètres de l'équation estimée $\hat{Y} = \hat{a}_1 X_1 + \hat{a}_2 X_2 + \hat{b}$. Pour simplifier les calculs matriciels, nous opérons un changement de variables y=Y-2, $x_1=X_1-5$; $x_2=X_2-2$.
- On obtient

5

y	x_1	x_2
-1	-3	2
-1	-2	0
0	0	0
1	2	-1
1	3	-1

On sait que $$\hat{a} = (x'x)^{-1} x'y$$

6

$$x'x = \begin{pmatrix} -3 & -2 & 0 & 2 & 3 \\ 2 & 0 & 0 & -1 & -1 \end{pmatrix} \cdot \begin{pmatrix} -3 & 2 \\ -2 & 0 \\ 0 & 0 \\ 2 & -1 \\ 3 & -1 \end{pmatrix} = \begin{pmatrix} 26 & -11 \\ -11 & 6 \end{pmatrix}$$

$$(x'x)^{-1} = \frac{1}{35}\begin{pmatrix} 6 & 11 \\ 11 & 26 \end{pmatrix} = \begin{pmatrix} \frac{6}{35} & \frac{11}{35} \\ \frac{11}{35} & \frac{26}{35} \end{pmatrix} = \begin{pmatrix} 0{,}17 & 0{,}31 \\ 0{,}31 & 0{,}74 \end{pmatrix}$$

$$x'y = \begin{pmatrix} -3 & -2 & 0 & 2 & 3 \\ 2 & 0 & 0 & -1 & -1 \end{pmatrix} \begin{pmatrix} -1 \\ -1 \\ 0 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 10 \\ -4 \end{pmatrix}$$

7

$$\hat{a} = \begin{pmatrix} 0{,}17 & 0{,}31 \\ 0{,}31 & 0{,}74 \end{pmatrix} \begin{pmatrix} 10 \\ -4 \end{pmatrix} = \begin{pmatrix} 0{,}46 \\ 0{,}14 \end{pmatrix}$$

$$\hat{b} = \overline{Y} - \hat{a}_1 \overline{X}_1 - \hat{a}_2 \overline{X}_2 = 2 - (0{,}46 \cdot 5) - (0{,}14 \cdot 2) = -0{,}58$$

D'où l'expression finale de l'équation de régression multiple estimée:

$$\hat{Y} = 0{,}46\ X_1 + 0{,}14\ X_2 - 0{,}58$$

8

- Ceci signifie qu'il existe une relation positive assez forte entre le taux de criminalité juvénile et la densité urbaine, l'augmentation de l'indicateur de la densité urbaine d'une unité entraine l'augmentation de la criminalité juvénile de 46%, (alors que l'augmentation du taux de scolarisation d'une unité de mesure entraine la baisse de la criminalité juvénile de …!)

9

Exercice:

- On veut exprimer l'évolution de l'indice du revenu nominale moyen (Y) d'un ménage de salariés en fonction de l'indice général des prix (X_1) et de l'indice du produit intérieur brut réel (X_2). On se limite à 9 observations par simplifications:

10

i	y_i	X_{1i}	X_{2i}
1	100	100	100
2	106	104	99
3	107	106	110
4	120	111	126
5	111	111	113
6	116	115	103
7	123	120	102
8	133	124	103
9	137	126	98

11

Questions:

1. Après avoir calculé la matrice $(X'X)^{-1}$, calculez X'Y et en déduire le vecteur â des estimations des paramètres du modèle:

$$Y=a_1X_1+a_2X_2+a_3+\varepsilon$$

2. Calculez $\hat{Y} = X\hat{a}$ et le résidu $\hat{\varepsilon} = Y - \hat{Y}$
3. Calculez une estimation de la variance des résidus $\left(\hat{\sigma}_\varepsilon^2\right)$

12

Questions: (suite)

4. Calculez l'estimation $\hat{\sigma}_{\hat{a}}$, effectuez le test de Student pour a_1. construire un intervalle de confiance à 0,95% pour a_1.
5. Même question que précédemment pour a_2.

13

Réponses:

1. Pour simplifier les calculs on peut centrer les données. Le modèle s'écrira alors:

$$y-\overline{Y}=a_1\left(x_1-\overline{X}_1\right)+a_2\left(x_2-\overline{X}_2\right)+\varepsilon-\overline{\varepsilon}$$

on pose: $\left(x_1-\overline{X}_1\right)=X_1;$

$\left(x_2-\overline{X}_2\right)=X_2;$

$y-\overline{Y}=Y$

Le tableau des valeurs centrées est le suivant:

14

i	y_i	x_{1i}	x_{2i}
1	-17	-13	-6
2	-11	-9	-7
3	-10	-7	4
4	3	-2	20
5	-6	-2	7
6	-1	2	-3
7	6	7	-4
8	16	11	-3
9	20	13	-8

15

- L'équation estimée du modèle est

$$\hat{Y} = \hat{a}_1 X_1 + \hat{a}_2 X_2 + \hat{a}_3 + \hat{\varepsilon}$$

$$\hat{a} = \begin{pmatrix} \hat{a}_1 \\ \hat{a}_2 \end{pmatrix} = (X'X)^{-1}(X'Y)$$

$$= \begin{pmatrix} \sum X_{1t}^2 & \sum X_{1t}X_{2t} \\ \sum X_{1t}X_{2t} & \sum X_{2t}^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum X_{1t}Y_t \\ \sum X_{2t}Y_t \end{pmatrix}$$

$$= \begin{pmatrix} 650 & -112 \\ -112 & 648 \end{pmatrix}^{-1} \begin{pmatrix} 872 \\ -72 \end{pmatrix}$$

$$= \frac{1}{408656} \begin{pmatrix} 648 & 112 \\ 112 & 650 \end{pmatrix} \begin{pmatrix} 872 \\ -72 \end{pmatrix} = \begin{pmatrix} 1{,}3629 \\ 0{,}1245 \end{pmatrix}$$

16

- La détermination du coefficient $\hat{a}_3$ fait appel à la relation:

$$\overline{Y} = \hat{a}_1 \overline{X}_1 + \hat{a}_2 \overline{X}_2 + \hat{a}_3$$

- D'où

$$\hat{a}_3 = \overline{Y} - \hat{a}_1 \overline{X}_1 - \hat{a}_2 \overline{X}_2$$

2. $\hat{Y} = X\hat{a} = 1{,}3629 X_1 + 0{,}1245 X_2 - 50{,}206$

$$\hat{\varepsilon} = Y - \hat{Y} = Y - X\hat{a} = Y - 1{,}3629 X_1 - 0{,}1245 X_2 + 50{,}206$$

3.

$$\hat{\sigma}^2 = \frac{1}{n-k}\sum(\hat{\varepsilon}_i^2) = \frac{\left[(Y - X\hat{a})'(Y - X\hat{a})\right]}{n-k} = \frac{(Y'Y - \hat{a}'X'Y)}{9-3}$$

17

$$\hat{a}'X'Y = \begin{pmatrix} 1{,}3629 & 0{,}1245 \end{pmatrix}\begin{pmatrix} 872 \\ -72 \end{pmatrix} = 1179{,}5$$

$$Y'Y = \sum y_i^2 = 1248$$

D'où l'on obtient:

$$\hat{\sigma}_\varepsilon^2 = \frac{68{,}5}{6} = 11{,}4$$

4. Les calculs permettent d'obtenir $\hat{\sigma}_{\hat{a}_1} = 0{,}1344$, on sait que $\frac{(\hat{a}_1 - a_1)}{\hat{\sigma}_{\hat{a}_1}}$ Suit une loi de Student à 6 degrés de liberté. Pour vérifier est ce que a_1=0, l'indicateur est alors égale à 1,3629/0,1344=10,14. L'intervalle d'acceptation est fourni par la table de Student au seuil de signification de 0,05; ***I*=[-2,447; +2,447].**

18

Donc $a_1 \neq 0$, c'est-à-dire la variable X_1 intervient dans la détermination de Y.

5. Les calculs permettent d'obtenir $\hat{\sigma}_{\hat{a}_2} = 0{,}1347$, on sait que $\frac{(\hat{a}_2 - a_2)}{\hat{\sigma}_{\hat{a}_2}}$ Suit une loi de Student à

6 degrés de liberté. Pour vérifier est ce que a_2=0, l'indicateur est alors égale à 0,1245/0,1347=0,924. L'intervalle d'acceptation est fourni par la table de Student au seuil de signification de 0,05; ***I*=[-*2,447; +2,447*].**

19

- Donc a_2=0 au seuil de signification de 0,05. d'où la variable X_2 (l'indice de PIB réel) n'intervient pas dans la détermination de Y (le salaire nominal des ménages).
- Il faut donc éliminer la variable X_2 du modèle, et procéder à une nouvelle estimation de a_1, à moins que les données soient erronées, ou qu'il y ait d'autres variables explicatives omises.

20