Fiche sur les régressions
Correction de l'exercice 6

9 janvier 2007

1. La variable expliquée est la valeur ajoutée de l'entreprise Y. Les variables explicatives sont la valeur du capital utilisé X et le nombre de salariés employés Z. En prenant de log du modèle M, nous obtenons :

$$\log(Y) = a\log(X) + (1-a)\log(Z) + \log(k). \tag{1}$$

a et $1-a$ s'interprètent comme les élasticités relatives de la valeur ajoutée de Y par rapport à X et Z :

$$a = \frac{d\log(Y)}{d\log(X)} = \frac{dY/Y}{dX/X}, \quad (1-a) = \frac{d\log(Y)}{d\log(Z)} = \frac{dY/Y}{dZ/Z}.$$

Si on multiplie X et Z par la même constante $c > 0$, alors la valeur ajoutée de l'entreprise correspondante, prévue par le modèle M est :

$$k(cX)^a(cY)^{1-a} = kc^aX^ac^{1-a}Y^{1-a} = c\left(kX^aY^{1-a}\right),$$

et on constate qu'elle est elle également multipliée par c : f telle que $Y = f(X,Z)$ est homogène de degré 1.

2. Dans le modèle M, Y n'est pas linéaire en X ni en Z. En prenant le log comme en (1), et en utilisant les notations de l'énoncé :

$$y = ax + (1-a)z + c, \quad \Leftrightarrow \quad (y-z) = a(x-z) + c \quad [M'].$$

Le modèle $[M']$ est un modèle de régression linéaire où la variable expliquée est $y-z$ et la variable explicative $x-z$, et on peut estimer les coefficients a et c par la méthode des moindres carrés.

3. Nous pouvons obtenir en prenant le log de nos données (X_j, Y_j, Z_j), pour $j \in [\![1,n]\!]$, des données $(x_j = \log(X_j), y_j = \log(Y_j), z_j = \log(Z_j))$, pour $j \in [\![1,n]\!]$. Nous avons :

$$\begin{aligned}
\mathrm{cov}(x-z, y-z) =& \frac{1}{n}\sum_{j=1}^{n}(x_j - z_j - \bar{x} + \bar{z})(y_j - z_j - \bar{y} + \bar{z}) = \frac{1}{n}\sum_{j=1}^{n}\left[(x_j-\bar{x}) - (z_j - \bar{z})\right]\left[(y_j - \bar{y}) - (z_j - \bar{z})\right] \\
=& \frac{1}{n}\sum_{j=1}^{n}\left[(x_j-\bar{x})(y_j-\bar{y}) - (x_j-\bar{x})(z_j-\bar{z}) - (z_j-\bar{z})(y_j-\bar{y}) + (z_j-\bar{z})^2\right] \\
=& \mathrm{cov}(x,y) - \mathrm{cov}(x,z) - \mathrm{cov}(y,z) + \mathrm{Var}(z).
\end{aligned} \tag{2}$$

Comme :

$$\mathrm{Var}(x-z) = \mathrm{cov}(x-z, x-z),$$

en utilisant le calcul fait en (2) et en y remplaçant y par x, on obtient :

$$\mathrm{Var}(x-z) = \mathrm{Var}(x) - 2\mathrm{cov}(x,z) + \mathrm{Var}(z). \tag{3}$$

De même, en remplaçant dans (3) x par y, $\mathrm{Var}(y-z) = \mathrm{Var}(y) - 2\mathrm{cov}(y,z) + \mathrm{Var}(z)$.

4. En utilisant les formules du cours et les résultats des questions 2 et 3 :

$$\begin{aligned}
\widehat{a} =& \frac{\mathrm{cov}(x-z, y-z)}{\mathrm{Var}(x-z)} = \frac{\mathrm{cov}(x,y) - \mathrm{cov}(x,z) - \mathrm{cov}(y,z) + \mathrm{Var}(z)}{\mathrm{Var}(x) - 2\mathrm{cov}(x,z) + \mathrm{Var}(z)} \\
=& \frac{0,1836 - 0,1791 - 0,1760 + 0,1720}{0,1887 - 2\times 0,1791 + 0,1720} = 0,20 \\
\widehat{c} =& \overline{y-z} - \widehat{a}\overline{x-z} = \bar{y} - \bar{z} - \widehat{a}\,(\bar{x} - \bar{z}) \\
=& 9,97 - 4,71 - 0,20 \times (4,76 - 4,71) = 5,25.
\end{aligned}$$

Pour une entreprise de valeur ajoutée Y, de capital X et ayant Z salarié, la nouvelle valeur ajoutée Y' obtenue avec le même capital X et un nombre de salarié Z' sera $Y' = kX^a Z'^{1-a} = Y(Z'/Z)^{1-a}$. On en déduit $Y'/Y = (Z'/Z)^{1-a}$. Si $Z'/Z = 0,01$, l'augmentation prévue en remplaçant a par $\widehat{a}$ est : $Y'/Y = 2,5\%$.

On définit :

$$\widehat{y} = z + \widehat{a}(x-z) + \widehat{c} = \bar{y} + (z - \bar{z}) + \widehat{a}((x-z) - \overline{x-z}) \tag{4}$$

en remplaçant $\widehat{c}$ par $\overline{y-z} - \widehat{a}\overline{x-z}$. On pose $\widehat{\varepsilon} = (y-z) - (\widehat{y} - z) = y - \widehat{y}$. En utilisant la définition de la variance et (4), on a :

$$\begin{aligned} \mathrm{Var}\,(\widehat{y} - z) &= \frac{1}{n}\sum_{j=1}^{n} (\widehat{y}_j - z_j - \bar{y} + \bar{z})^2 = \frac{\widehat{a}^2}{n}\sum_{j=1}^{n} ((x_j - z_j) - \overline{x_j - z_j})^2 \\ &= \widehat{a}^2 \mathrm{Var}(x-z) = \widehat{a}^2\left(\mathrm{Var}(x) - 2\mathrm{cov}(x,z) + \mathrm{Var}(z)\right), \end{aligned}$$

par (3). L'équation d'analyse de la variance s'écrit alors :

$$\begin{aligned} \mathrm{Var}(\widehat{\varepsilon}) &= \mathrm{Var}(y-z) - \mathrm{Var}(\widehat{y} - z) \\ &= \left(\mathrm{Var}(y) - 2\mathrm{cov}(y,z) + \mathrm{Var}(z)\right) - \widehat{a}^2\left(\mathrm{Var}(x) - 2\mathrm{cov}(x,z) + \mathrm{Var}(z)\right) \\ &= (0,1804 - 2 \times 0,1836 + 0,1720) - 0,20^2 \times (0,1887 - 2 \times 0,1791 + 0,1720) = 3,6\,10^{-4}. \end{aligned} \tag{5}$$

On en déduit :

$$\begin{aligned} R^2 &= \frac{\mathrm{Var}(\widehat{y} - z)}{\mathrm{Var}(y-z)} = 1 - \frac{\mathrm{Var}(\widehat{\varepsilon})}{\mathrm{Var}(y-z)} \quad \textbf{Ceci est une conséquence de (5) à retenir !} \\ &= 1 - \frac{3,6\,10^{-4}}{0,1804 - 2 \times 0,1760 + 0,1720} = 0,10. \end{aligned}$$

La part de variance expliquée, selon le modèle M' est très mauvaise.

$$\sqrt{\frac{1}{n}\sum_{j=1}^{n} \widehat{\varepsilon}_j^2} = \sqrt{\mathrm{Var}(\widehat{\varepsilon})} = 1,90\%,$$

ce qui est très mauvais.