Fiche sur les régressions
Correction de l'exercice 2 (Première partie)

7 janvier 2007

1. Sur le graphique a, les points semblent légèrement moins alignés que sur le graphique b. On choisit donc dans un premier temps d'étudier le modèle M1 qui correspond au graphique b.

Comme :

$$\frac{\Delta_t \log q}{\Delta_t \log p} = \frac{\log q_t - \log q_0}{\log p_t - \log p_0} = a,$$

le coefficient a dans $M1$ est l'élasticité absolue de la demande q de micro-ordinateurs par rapport à l'indice de prix p. Comme les points du graphique b semblent être alignés suivant une droite de pente négative, on s'attend à ce que $a < 0$.

Le paramètre b du modèle $M1$ a bien un sens : il correspond au log de la demande q lorsque $\log p = 0$, c'est à dire $p = 1$. C'est le log de la demande lorsque l'indice de prix p vaut 1, c'est à dire lorsque l'on a le même prix qu'à la date de référence : le second trimestre de 1992.

En prenant l'exponentielle dans $M1$:

$$q_t = p_t^a \times e^b,$$

qui correspond à la forme recherchée avec $k = e^b$.

2. En utilisant les formules vues en cours :

$$\widehat{a} = \frac{\text{cov}(\log p_t, \log q_t)}{\text{Var}(\log p_t)}, \quad \widehat{b} = \overline{\log q} - \widehat{a}\overline{\log p}.$$

Comme on a 20 observations :

$$\begin{aligned}
\overline{\log p} =& \frac{1}{20}\sum_{t=1}^{20} \log p_t = \frac{-5,149}{20} = -0,257 \\
\overline{\log q} =& \frac{1}{20}\sum_{t=1}^{20} \log q_t = \frac{37,888}{20} = 1,894 \\
\text{cov}(\log p_t, \log q_t) =& \frac{1}{20}\sum_{t=1}^{20} \left(\log p_t - \overline{\log p}\right)\left(\log q_t - \overline{\log q}\right) = \frac{1}{20}\sum_{t=1}^{20} \left(\log p_t \times \log q_t\right) - \overline{\log p} \times \overline{\log q} \\
=& \frac{-10,413}{20} - (-0,257) \times 1,894 = -0,033 \\
\text{Var}(\log p_t) =& \frac{1}{20}\sum_{t=1}^{20} \left(\log p_t\right)^2 - \overline{\log p}^2 = \frac{1,912}{20} - (-0,257)^2 = 0,029 \\
\text{Var}(\log q_t) =& \frac{1}{20}\sum_{t=1}^{20} \left(\log q_t\right)^2 - \overline{\log q}^2 = \frac{72,532}{20} - (1,894)^2 = 0,038.
\end{aligned}$$

On en déduit :

$$\widehat{a} = -1,124, \quad \widehat{b} = 1,605.$$

Le modèle $M1$ dont on a estimé les coefficients est donc :

$$[\log \widehat{q}_t = -1,124 \times \log p_t + 1,605] \Leftrightarrow \left[q_t = 4,978 \times p_t^{-1,124}\right], \text{ où } \widehat{k} = e^{\widehat{b}} = 4,978. \tag{1}$$

3. Nous définissons classiquement :

$$\widehat{\varepsilon}_t = \log(q_t) - \log \widehat{q}_t = \log(q_t) - \left(\widehat{a}\log(p_t) + \widehat{b}\right) = \left(\log(q_t) - \overline{\log q}\right) - \widehat{a}\left(\log(p_t) - \overline{\log p}\right),$$

la dernière égalité étant obtenue en remplaçant $\widehat{b}$ par $\overline{\log q} - \widehat{a}\overline{\log p}$. Remarquons (c'est classique) que $\overline{\widehat{\varepsilon}} = 0$. Nous avons donc :

$$\begin{aligned}
\mathrm{Var}(\widehat{\varepsilon}) &= \frac{1}{20}\sum_{t=1}^{20}\widehat{\varepsilon}_t^2 = \frac{1}{20}\sum_{t=1}^{20}\left[\left(\log(q_t) - \overline{\log q}\right) - \widehat{a}\left(\log(p_t) - \overline{\log p}\right)\right]^2 \\
&\text{en utilisant l'identité remarquable } (x-y)^2 = x^2 - 2xy + y^2 \text{ :} \\
&= \frac{1}{20}\sum_{t=1}^{20}\left(\log(q_t) - \overline{\log q}\right)^2 - \frac{1}{20}\times 2\widehat{a}\sum_{t=1}^{20}\left(\log(p_t) - \overline{\log p}\right)\left(\log(q_t) - \overline{\log q}\right) + \widehat{a}^2\frac{1}{20}\sum_{t=1}^{20}\left(\log(p_t) - \overline{\log p}\right)^2 \\
&= \mathrm{Var}(\log q) - 2\widehat{a}\,\mathrm{cov}(\log p, \log q) + \widehat{a}^2\mathrm{Var}(\log p) \\
&= \mathrm{Var}(\log q) - 2\frac{\left(\mathrm{cov}(\log p, \log q)\right)^2}{\mathrm{Var}(\log p)} + \frac{\left(\mathrm{cov}(\log p, \log q)\right)^2}{\mathrm{Var}(\log p)^2}\mathrm{Var}(\log p) \text{ en remplaçant } \widehat{a} \text{ par } \frac{\mathrm{cov}(\log p, \log q)}{\mathrm{Var}(\log p)} \\
&= \mathrm{Var}(\log q) - \frac{\left(\mathrm{cov}(\log p, \log q)\right)^2}{\mathrm{Var}(\log p)} \\
&= \mathrm{Var}(\log q) - \widehat{a}^2\mathrm{Var}(\log p) \text{ en utilisant toujour la même relation} \\
&= \mathrm{Var}(\log q) - \mathrm{Var}(\log \widehat{q}) \text{ car } \mathrm{Var}(\log \widehat{q}) = \widehat{a}^2\mathrm{Var}(\log p). \qquad (2)
\end{aligned}$$

En effet, par la definition de $\widehat{a}$ et $\widehat{b}$, on a $\overline{\log \widehat{q}} = \overline{\log q}$ et :

$$\begin{aligned}
\mathrm{Var}(\log \widehat{q}) &= \frac{1}{20}\sum_{t=1}^{20}\left(\log \widehat{q}_t - \overline{\log q}\right)^2 = \frac{1}{20}\sum_{t=1}^{20}\left(\widehat{a}\log p_t + \widehat{b} - \overline{\log q}\right)^2 \\
&= \frac{1}{20}\sum_{t=1}^{20}\left(\widehat{a}\log p_t + \overline{\log q} - \widehat{a}\overline{\log p} - \overline{\log q}\right)^2 = \frac{1}{20}\sum_{t=1}^{20}\widehat{a}^2\left(\log p_t - \overline{\log q}\right)^2 = \widehat{a}^2\mathrm{Var}(\log p). \qquad (3)
\end{aligned}$$

(L'expression $\mathrm{Var}(\log \widehat{q}) = \widehat{a}^2\mathrm{Var}(\log p)$ démontrée en (3) est **à retenir** !) Remarquons qu'en (2) nous avons re-démontré la décomposition de la variance :

$$\mathrm{Var}(\log q) = \mathrm{Var}(\widehat{q}) + \mathrm{Var}(\widehat{\varepsilon}) \qquad (4)$$

qui peut s'utiliser directement en devoir. Cette formule exprime le fait que la variance de la variable expliquée (ici $\log q$) se décompose en une variance expliquée par la variable explicative $\mathrm{Var}(\widehat{q})$ et une variance résiduelle $\mathrm{Var}(\widehat{\varepsilon})$.

Calculons l'erreur type de la régression en partant de la formule du cours :

$$\begin{aligned}
s(\widehat{\varepsilon}) &= \sqrt{\frac{1}{20-2}\sum_{t=1}^{20}\widehat{\varepsilon}_t^2} = \sqrt{\frac{20}{18}\mathrm{Var}(\widehat{\varepsilon})} = \sqrt{\frac{20}{18}\left(\mathrm{Var}(\log q) - \mathrm{Var}(\widehat{q})\right)} \text{ par (2)} \\
&= \sqrt{\frac{20}{18}\left(\mathrm{Var}(\log q) - \widehat{a}\mathrm{Var}(p)\right)} \text{ par (3)} \\
&= \sqrt{\frac{20}{18}\left(0,038 - (-1,124)^2 \times 0,029\right)} = 0,023.
\end{aligned}$$

Comme $s(\widehat{\varepsilon})$ est l'écart-type corrigé de l'erreur fait en remplaçant $\log q$ par $\log \widehat{q}$, et comme $s(\widehat{\varepsilon})$ est petit par rapport à $\sqrt{\mathrm{Var}(\log q)} = 0,195$, alors on en déduit que $M1$ est un bon modèle.

A $t = 20$, on a : $\widehat{\varepsilon}_{20} = 0,037$ et $\widehat{q}_{20} = 9,362$ correspondant à une erreur relative de $0,037$ de l'ordre de $s(\widehat{\varepsilon})$. Ceci peut s'expliquer. Lorsque $q_t/\widehat{q}_t$ est proche de 1, et lorsqu'on peut approcher $\log(q_t/\widehat{q}_t)$ par son développement limité à l'ordre 1 en 1 $q_t/\widehat{q}_t - 1$, alors :

$$\widehat{\varepsilon}_t = \log(q_t) - \log(\widehat{q}_t) = \log(q_t/\widehat{q}_t) \sim q_t/\widehat{q}_t - 1 = \frac{q_t - \widehat{q}_t}{q_t} \times \frac{q_t}{\widehat{q}_t} \sim \frac{q_t - \widehat{q}_t}{q_t},$$

puisque $q_t/\widehat{q}_t$ est proche de 1. Ainsi, $s(\widehat{\varepsilon})$ est-elle proche de la moyenne des erreurs relatives.

4. D'après le cours, le coefficient de détermination est :

$$R^2 = \frac{\mathrm{Var}(\log \widehat{q})}{\mathrm{Var}(\log q)} = \frac{\widehat{a}^2\mathrm{Var}(\log p)}{\mathrm{Var}(\log q)} = 0,99, \qquad (5)$$

en utilisant (3). Comme R^2 est le pourcentage de la variance de la variable $\log q$ expliquée par l'explicative $\log p$, on en déduit que l'ajustement (1) fourni par $M1$ est très bon.

5. Entre le second trimestre de 1992 et le second trimestre de 1997, nous avons 20 trimestres. Le taux trimestriel moyen de variation du prix des micro-ordinateurs, qui vaut 1 en $t = 0$ et 0,57 en $t = 20$ est donc :

$$r_m = \left(\frac{0,57}{1}\right)^{1/20} - 1 - 2,77\%.$$

Si l'on suppose que la baisse des prix se poursuit à ce même rythme r_m jusqu'à la fin de 1997 ($t = 21$ et $t = 22$), alors les ventes prévues par $M1$, déterminées à l'aide de (1) en remplaçant p_{21} par $(1 + r_m)p_{20}$ et p_{22} par $(1 + r_m)^2 p_{20}$ sont :

$$\widehat{q}_{21} = 9,664, \quad \widehat{q}_{22} = 9,974 \text{ (en milliers)}.$$

6. Nous avons :

$$\log V = \log(pq) = \log p + \log q = (1 + a)\log p + b,$$

qui correspond à la forme recherchée avec $c = 1 + a$ et $d = b$.

L'estimateur des moindres carrés de c est :

$$\widehat{c} = \frac{\text{cov}(\log p, V)}{\text{Var}(\log p)} = \frac{\text{cov}(\log p, \log p + \log q)}{\text{Var}(\log p)} = \frac{\text{Var}(\log p)}{\text{Var}(\log p)} + \frac{\text{cov}(\log p, \log q)}{\text{Var}(\log p)} = 1 + \widehat{a} = -0,124.$$

On en déduit que :

$$\widehat{d} = \overline{\log V} - \widehat{c}\overline{\log p} = \overline{\log p} + \overline{\log q} - \overline{\log p} - \widehat{a}\overline{\log p} = \overline{\log p} - \widehat{a}\overline{\log p} = \widehat{b}.$$

L'équation de l'ajustement de $M3$ est donc :

$$\widehat{V} = 4,978 \times p_t^{-0,124}.$$

Les résidus de cet ajustement sont :

$$\widehat{\varepsilon'}_t = \log(V_t) - \log(\widehat{V}_t) = \log p_t + \log q_t - \left((1 + \widehat{a})\log p_t + \widehat{b}\right) = \log q_t - \widehat{a}\log p_t - \widehat{b} = \widehat{\varepsilon}_t.$$

7. En utilisant la formule $(a + b)^2 = a^2 + b^2 + 2ab$, nous avons :

$$\begin{aligned}
\text{Var}(x + y) =& \frac{1}{20}\sum_{t=1}^{20}(x + y - \bar{x} - \bar{y})^2 = \frac{1}{20}\sum_{t=1}^{20}(x - \bar{x})^2 + \frac{1}{20}\sum_{t=1}^{20}(y - \bar{y})^2 + \frac{2}{20}\sum_{t=1}^{20}(x - \bar{x})(y - \bar{y}) \\
=& \text{Var}(x) + \text{Var}(y) + 2\text{cov}(x, y) = (1 + 2\widehat{a})\text{Var}(x) + \text{Var}(y), \qquad (6)
\end{aligned}$$

en utilisant le fait que $\widehat{a} = \text{cov}(x, y)/\text{Var}(x)$ implique que $\text{cov}(x, y) = \widehat{a}\text{Var}(x)$.

On en déduit le coefficient de détermination associé à $M3$:

$$\begin{aligned}
R'^2 =& \frac{\text{Var}(\log(\widehat{V}))}{\text{Var}(\log V)} = \frac{(1 + \widehat{a})^2\text{Var}(\log p)}{\text{Var}(\log p + \log q)} \text{ en utilisant (3)} \\
=& \frac{(1 + \widehat{a})^2\text{Var}(\log p)}{(1 + 2\widehat{a})\text{Var}(\log p) + \text{Var}(\log q)} \text{ en utilisant (6)} \\
=& \frac{(1 + 2\widehat{a} + \widehat{a}^2)\text{Var}(\log p)}{(1 + 2\widehat{a})\text{Var}(\log p) + \text{Var}(\log q)}, \qquad (7) \\
=& \frac{(1 - 2 \times 1,124 + (-1,124)^2) \times 0,029}{(1 - 2 \times 1,124) \times 0,029 + 0,038} = 0,35.
\end{aligned}$$

Cette valeur est bien inférieure à celle trouvée en (5).

On déduit de (5) et (7) que $R^2 \geq R'^2$ si et seulement si :

$$\begin{aligned}
&\frac{\widehat{a}^2\text{Var}(\log p)}{\text{Var}(\log q)} \geq \frac{(1 + \widehat{a})^2\text{Var}(\log p)}{(1 + 2\widehat{a})\text{Var}(\log p) + \text{Var}(\log q)} \\
\Leftrightarrow& \widehat{a}^2\text{Var}(\log p)\left((1 + 2\widehat{a})\text{Var}(\log p) + \text{Var}(\log q)\right) \geq \text{Var}(\log q)(1 + \widehat{a})^2\text{Var}(\log p) \\
\Leftrightarrow& \widehat{a}^2(1 + 2\widehat{a})(\text{Var}(\log p))^2 \geq \text{Var}(\log p)\text{Var}(\log q) \\
\Leftrightarrow& \widehat{a}^2(1 + 2\widehat{a})\frac{\text{Var}(\log p)}{\text{Var}(\log q)} \geq 0 \\
\Leftrightarrow& \widehat{a} \geq -0,5, \text{ car tous les autres termes sont positifs.}
\end{aligned}$$

Comme $\widehat{a} < -0,5$, ceci explique pourquoi $R^2 = 0,99 > R'^2 = 0,35$.

Conclusion : les modèles $M1$ et $M3$ qui sont équivalents ne donnent pas lieu à la même précision d'ajustement suivant la valeur de $\widehat{a}$. On ne devra donc pas apprécier la pertinence d'une modélisation à partir de la seule précision de l'ajustement aux données.