

L'exemple du prix des ordinateurs

Antoine MOREAU
Insee

Sept constructeurs participent à l'indice au début de 1991. Un modèle reliant les caractéristiques des produits au prix de vente a été construit séparément pour chacun des constructeurs. Ce document s'articule en trois parties :

- un bref rappel des méthodes utilisées pour construire l'indice
- deux exemples, sur deux constructeurs différents, de modélisation.

I - Deux méthodes de calcul

On présente ici les résultats de deux méthodes de calcul d'un indice hédonique pour les micro-ordinateurs.

Première méthode : la méthode dite "des caractéristiques". Supposons qu'on s'intéresse à un micro-ordinateur apparaissant au 2ème trimestre 1990 (90T2). On estime une équation de prix au 1er trimestre 1990 (90T1) sous la forme :

$$\text{Log}(P_i) = \sum_{k=1}^K b_k x_{ki} + u_i \quad (1)$$

où P_i est le prix de la machine i

x_{ki} la k ème caractéristique de la machine i

On estime alors directement le prix qu'aurait eu en 90T1 le micro apparaissant en 90T2 par

$$\text{Log}(p) = \sum_{k=1}^K \hat{b}_k x_k$$

où les \hat{b}_k sont les estimations des b_k obtenues par les moindres carrés ordinaires sur (1)

les x_k sont les caractéristiques du nouveau micro

A noter que plusieurs spécifications de l'équation (1) sont possibles. On pourrait utiliser p et non son logarithme ; de même, toutes les caractéristiques de type continu (horloge interne, capacité sur le disque dur - par opposition aux caractéristiques de type indicatrice - processeur, type de lecteur de disquettes) peuvent être prises en logarithme.

On a retenu ici trois spécifications possibles :

- log-log : toutes variables continues en logarithme
- log-linéaire : prix en logarithme, toutes les caractéristiques continues en linéaire
- log-semilog : prix en logarithme, toutes les variables continues, en linéaire, sauf la capacité sur le disque dur.

Réserver un traitement spécial à cette caractéristique peut se justifier par le fait qu'elle a une variabilité beaucoup plus grande que les autres. Par ailleurs, en consultant les catalogues d'option on constate que, si une extension nécessaire de 2 Mo vaut à peu près deux fois moins cher qu'une extension mémoire de 4 Mo, ce n'est pas le cas pour les disques durs : un disque de 300 Mo vaut nettement moins cher que 3 disques de 100 Mo.

Il est clair que d'autres spécifications sont aussi envisageables mais on ne les étudiera pas ici.

Deuxième méthode : la méthode dite "des indicatrices". C'est cette méthode là que les statisticiens d'IBM ont proposé initialement en 85 à l'OCDE. Elle consiste à utiliser les prix de produits sur deux périodes consécutives t et $t-1$ pour estimer l'équation reliant le prix aux caractéristiques. Ainsi on estime :

$$\text{Log}(p_i) = \sum_{k=1}^K b_k x_{ki} + C_{1i} + C_{2i} + u_i \quad (2)$$

C_{1i} et C_{2i} sont des indicatrices, valant 0 ou 1 :

- C_{2i} vaut 0 pour le prix en $t-1$ du modèle i
- 0 pour le prix en t du modèle i si celui-ci apparaît en t
- 1 pour le prix en t du modèle i si celui-ci est présent en t et $t-1$

- . C_{1i} vaut 0 pour le prix en t-1 du modèle i
- 0 pour le prix en t du modèle i si celui-ci est présent en t et t-1
- 1 pour le prix en t du modèle i si celui-ci apparaît en t

Pour un modèle présent en t et en t-1, on a

$$P_{it} / P_{it-1} = \exp (C_{2i})$$

Donc, $\exp (C_2)$ est une estimation de la variation de prix moyenne entre t et t-1 des modèles présents dans l'échantillon en t et t-1.

De la même façon, $\exp (C_1)$ est une estimation de la variation de prix moyenne entre t et t-1 des modèles présents seulement en t. Comme pour la méthode précédente, plusieurs spécifications sont possibles pour l'équation (2). On a retenu les trois déjà décrites.

II - Constructeur n° 1

On a appliqué les deux méthodes pour estimer les prix, au troisième trimestre 1990 (90T3) de deux micro-ordinateurs apparaissant au catalogue d'un grand constructeur au quatrième trimestre (90T4).

a) Résultats de la méthode des indicatrices

Les trois spécifications donnent les résultats suivants

68 observations	Log-Log	Log-Lin	Log-Semi Log
Constante	8.896 (53.534)	9.459 (350.972)	9.445 (288.855)
Processeur 386SX	- 1.183 (- 2.214)	0	0
Processeurs 386DX	- 1.266 (- 2.359)	0	0
Processeurs 486	0.402 (1.965)	0	0
Capacité disque dur 286	0.199 (5.732)	0	0.208 (6.560)
Capacité disque dur 386	0.184 (7.150)	0.001 (7.075)	0.186 (7.374)
Capacité disque dur 486	0.146 (6.889)	0.001 (5.532)	0.147 (6.925)
Présence d'un disque dur	- 0.456 (- 4.005)	0.212 (8.963)	- 0.467 (- 4.225)
Horloge 386	0.495 (2.576)	0.015 (13.257)	0.015 (10.037)
RAM minimale	0	0	0
RAM maximale	0.350 (4.528)	0.019 (26.661)	0.020 (22.664)
Modèle tour	0.416 (9.395)	0.469 (18.213)	.0468 (17.327)
C1	- 0.147 (- 2.422)	- 0.128 (- 2.245)	- 0.156 (- 2.576)
C2	- 0.075 (- 4.136)	- 0.074 (- 4.383)	- 0.072 (- 4.040)
R2	0.9889	0.9900	0.9889

L'estimation de C2 donne une variation de - 7.1 % pour les modèles présents en 90T3 et 90T4, à comparer avec une variation observée de - 6.4 %.

b) Résultats de la méthode des caractéristiques

Les résultats des trois spécifications sont les suivants :

35 observations	Log-Log	Log-Lin	Log-Semi Log
Constante	8.808 (111.320)	9.476 (279.060)	9.448 (227.438)
Processeur 386SX	0	0	0
Processeurs 386DX	0	0	0
Processeurs 486	0	0	0
Capacité disque dur 286	0.136 (3.086)	0	0.167 (3.789)
Capacité disque dur 386	0.150 (4.296)	0.002 (6.265)	0.162 (4.625)
Capacité disque dur 486	0.136 (4.718)	0.001 (4.464)	0.134 (4.693)
Présence d'un disque dur	- 0.327 (- 2.149)	0.183 (5.870)	- 0.382 (- 2.511)
Horloge 386	0.044 (3.052)	0.015 (10.396)	0.015 (7.815)
Modèle tour	0.427 (13.190)	0.425 (13.601)	0.421 (13.281)
RAM minimale	0.101 (2.861)	0	0.041 (2.461)
RAM maximale	0.423 (14.789)	0.019 (20.658)	0.018 (15.827)
R2	0.9920	0.9928	0.9920

c) Comparaison des prix estimés

Les deux modèles apparaissant en 90T4 ont le même prix : 47 940 F.

Les diverses estimations pour le prix en 90T3 obtenues sont les suivantes :

<u>Caractéristiques</u>	<u>Prix</u>	<u>Variation</u>
Log-Log	55 605	- 13.8 %
Log-Lin	56 453	- 15.1 %
Log-semilog	56 052	- 14.5 %
<u>Indicatrices</u>		
Log-Log	55 531	- 13.6 %
Log-Lin	54 486	- 12 %
Log-semilog	56 033	- 14.4 %

III - Constructeur n° 2

Des régressions sur deux périodes, qui ont posé des problèmes différents, sont présentées dans la suite.

1) Deuxième trimestre 90a) Méthode des indicatrices

On a tenté d'inclure dans l'équation le nombre de connecteurs et la RAM maximale sur carte mère. Le nombre de connecteurs ne peut pas être pris en compte du fait de colinéarités entre les variables. La RAM maximale ne ressort pas significativement. Les résultats des trois spécifications sont très proches. On trouvera ci-dessous ceux de la spécification log-linéaire.

63 observations

Constante	9.962 (143.79)
Processeur 386SX	- 0.277 (9.49)
Capacité disque dur 286	0.213 (6.42)
Capacité disque dur 386	0.194 (10.33)
Présence d'un disque dur (1 si disque dur)	0.194 (3.89)
Horloge interne 386	0.021 (7.04)
RAM minimale	0.045 (3.82)
Modèle Tour	0.119 (3.82)
286 apparaissant en 90T2	- 0.395 (10.12)
286 présents en 90T1	- 0.109 (2.844)
386 présents en 90T1	- 0.057 (3.63)
R^2	0.9923

Les variations estimées des produits présents en 90T1 et 90T2 sont - 10,3 pour les 286 et - 5,5 pour les 386, à comparer avec des variations observées de - 9,9 et - 5,8. La variation estimée des produits apparaissant en 90T2 (- 32,6 %) est très forte. Elle correspond effectivement à ce que l'on observe dans l'échantillon : les 286 qui apparaissent sont nettement moins chers que les anciens. Ainsi, un 296 à 12 Mhz et 20 Mégaoctets sur le disque dur et 1 Mégaoctet de RAM vaut 17 502 en 90T2, son équivalent plus ancien valant 25 002 en 90T1 (et 22 802 en 90T2).

b) Méthode des caractéristiques

L'application de la méthode des caractéristiques pour cette période ne semble pas adéquate. En effet, tous les 286 en 90T1 ont la même capacité mémoire (20 Mégaoctets), alors que certains de ceux qui apparaissent en 90T2 ont 40 Mégaoctets sur le disque dur. On aurait pu :

- estimer sur 89T4, où d'anciens 286 ont des disques durs de 40 Mégaoctets

- estimer sur 90T1 en contraignant le coefficient des capacités mémoire à être le même pour les 286 et 386.

Dans le premier cas, il est vraisemblable que l'on surestimerait la baisse des prix (même phénomène qu'en 89T2) ; on a préféré s'en tenir à la méthode des indicatrices.

2) Troisième trimestre 90

Dix références nouvelles au catalogue en 90T3 : identifiants 63 et 68, 69 et 71, 76 et 77, 82 à 84.

Cette période a posé de très gros problèmes d'estimation. Seule la méthode des indicatrices a été mise en oeuvre.

Une première estimation des trois formulations utilisées habituellement a donné le résultat suivant (voir page 199, après élimination des variables non significatives).

On peut d'emblée éliminer la spécification log-log : le coefficient de la vitesse d'horloge pour les 286 est négatif, les coefficients des RAM minimale et maximale sont du même ordre de grandeur, alors que le second devrait être plus faible que le premier.

Par contre, rien ne permet de départager les deux autres spécifications, dont les implications sur l'indice sont pourtant fort différentes : - 20 % environ pour l'une, 0 pour l'autre.

On a donc tenté d'estimer l'équation par le maximum de vraisemblance, en appliquant la transformation de BOX-COX aux trois variables de capacité sur le disque dur : $x \rightarrow (x^d - 1)/d$.

Ceci a été réalisé un balayage sur les d : une fois les d fixés, l'estimation du modèle peut se faire par les MCO. Entre deux triplets de d , le meilleur modèle est celui qui a la valeur de la vraisemblance la plus forte. Celle ci se calcule facilement, à partir de l'estimation s^2 de la variance des résidus par les MCO :

Log-vraisemblance :

$$- \frac{N}{2} \left(\ln \left(s^{12} \frac{N-k}{N} \right) + 1 \right)$$

88 observations	Log-log	Log-linéaire	Log-semi log
Constante	9.441 (121.02)	8.990 (230.23)	9.167 (443.91)
Processeur 386DX	- 1.405 (9.35)	0.631 (28.18)	0
Processeur 386SX	- 1.660 (12.31)	0	- 0.396 (27.66)
Processeur 486	0	1.765 (36.45)	0.697 (9.99)
Capacité disque dur 286	0.132 (9.07)	0.004 (10.05)	0.154 (18.71)
Capacité disque dur 386	0.196 (22.67)	0.001 (27.61)	0.198 (32.01)
Capacité disque dur 486	0.237 (18.18)	0.001 (12.45)	0.220 (17.80)
Présence d'un disque dur (1 si disque dur)	- 0.199 (3.91)	0.062 (2.65)	- 0.292 (9.63)
Horloge 286	- 0.167 (5.89)	- 0.010 (6.01)	0
Horloge 386	0.316 (7.78)	0.013 (11.82)	0.011 (10.03)
RAM minimale	0.103 (8.82)	0.056 (10.54)	0.046 (8.49)
RAM maximale	0.148 (11.95)	0.006 (6.72)	0.006 (15.912)
Nombre de connecteur 8 bits	0.213 (21.59)	0.219 (25.37)	0.195 (31.09)
Nombre de connecteur 16 bits	0.130 (15.44)	0.146 (18.58)	0.125 (20.93)
Nombre de connecteur 32 bits	0.071 (11.44)	0	0.058 (9.81)
286 apparaissant en 90T3	0	- 0.068 (2.75)	0
386 apparaissant en 90T3	0	- 0.213 (10.44)	0
486 apparaissant en 90T3	0	- 0.231 (5.74)	0
R ²	0.9988	0.9990	0.9991

où N est le nombre d'observations et k le nombre de variables exogènes.

Cette procédure est coûteuse en temps : 1 600 modèles ont été estimés. On a essayé aussi bien des valeurs de d comprises entre 0 et 1 qu'à l'extérieur de cet intervalle. Le point le plus élevé de la vraisemblance a été obtenu pour

- Capacité des 286 à la puissance $\frac{1}{10}$
- Capacité des 386 à la puissance $\frac{2}{5}$
- Capacité des 486 à la puissance $\frac{19}{20}$

Cela revient à un modèle où la capacité des 286 serait en logarithme, celle des 386 en racine carré, celle des 486 en niveau.

Après élimination des coefficients non significatifs, on obtient

88 observations

Constante	8.128 (140.74)
Processeur 386DX	0.480 (33.43)
Processeur 486	1.763 (50.49)
Capacité disque dur 286	1.012 (25.15)
Capacité disque dur 386	0.073 (42.44)
Capacité disque dur 486	0.001 (18.44)
Présence d'un disque dur (1 si disque dur)	- 0.180 (9.47)
Horloge interne 386	0.012 (16.46)
RAM minimale	0.054 (12.93)
RAM maximale	0.004 (7.09)
Modèle Tour	0.043 (2.37)
Nombre de connecteurs 8 bits	0.182 (24.96)
Nombre de connecteurs 16 bits	0.110 (13.45)
Nombre de connecteurs 32 bits	0.017 (2.01)
386 apparaissant en 90T3	- 0.073 (5.34)
486 apparaissant en 90T3	- 0.168 (6.98)
R^2	0.9996

La nullité des indicatrices pour les modèles présents en 90T2 et 90T3 correspond bien au fait qu'aucune variation de prix n'est enregistrée. C'est cette régression qui a servi au calcul de l'indice.

Summary

Antoine MOREAU
Insee

The basic principle of a price index is to price at regular intervals a fixed basket of goods. In practice, the basket of goods can never be fixed over a long period of time : the statistician needs to derive methods to take account of new products. When technological changes happen continuously and fastly, as it is the case in the computer industry, an index in which new products are not properly integrated can be seriously biased. Some classical methods to compute a price index are surveyed, particularly the so called matched-models method, which doesn't take any account of quality changes.

The main features of the methods used in the United States to compute a price index for computers are then presented. The basic principle of these so-called hedonic methods is to desagregate a product into stable characteristics and then to price these characteristics. Several ways of implementing hedonic methods are reviewed. The american experience in the field is described.

Finally, some examples of hedonic equations used by INSEE for the price index of microcomputers are given.