Compléments sur les modèles linéaires

Compléments sur les modèles linéaires

Ce chapitre aborde plusieurs points relatifs à l'inférence statistique basée sur un modèle linéaire. Il traite des tests d'hypothèses sur les coefficients du modèle, de l'estimation du modèle linéaire par maximum de vraisemblance, de la multicolinéarité, de l'utilisation de variables indicatrices, des estimateurs de variance robustes, de l'estimateur des moindres carrés généralisés, des modèles linéaires sur des données en coupe instantanée et de la méthode des variables instrumentales.

1 Tests d'hypothèses linéaires sur plusieurs coefficients

1.1 Test général de restrictions linéaires sur les coefficients

On suppose que Y et X sont liés par un modèle linéaire $Y = X\beta + u$, où u est indépendant de X et est un bruit blanc normal : $u \sim N(0, \Sigma_u)$ où $\Sigma_u = \sigma_u^2 I_n$. On souhaite tester un ensemble de restrictions linéaires portant chacune sur un ou plusieurs coefficients du vecteur β.

Exemple

On veut tester la véracité simultanée des deux restrictions suivantes : $\beta_2 = 1$ et $\beta_3 = \beta_4$ dans le modèle :

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \beta_4 X_{4t} + u_t$$

Pour cela, trois étapes :

1. On estime le modèle par *MCO* sans aucune restriction, et l'on garde la somme des carrés des résidus ainsi obtenue, notée SE.
2. On estime le modèle par *MCO en imposant les restrictions*, et l'on garde la somme des carrés des résidus ainsi obtenue, notée $SE2$.
3. On montre que, *si les restrictions sont vraies*, alors :

$$\frac{(n-k)\,(SE2 - SE)}{mSE} \sim F(m, n-k) \tag{3.1}$$

et il suffit de calculer la valeur de ce test et de la comparer aux valeurs critiques d'une table de la distribution de Ficher à m degrés de liberté au numérateur et $n-k$ degrés de liberté au dénominateur. Il s'agit d'un test de petit échantillon, donc d'un test exact quelle que soit la taille de l'échantillon, si le terme d'erreur est normal. C'est en fait une transformation monotone du test du rapport de vraisemblance des mêmes restrictions, mais ici la distribution est exacte puisqu'on a fait l'hypothèse que le terme d'erreur est normal.

Exemple (suite)

Pour tester les deux restrictions de l'exemple précédent, on estime donc le modèle sans restrictions :

$$Y_t = \hat{\beta}_1^{MCO} + \hat{\beta}_2^{MCO} X_{2t} + \cdots + \hat{\beta}_4^{MCO} X_{4t} + e_t$$

et l'on garde la valeur $SE = e'e = \sum_{t=1}^{n} e_t^2$. Puis on estime le modèle avec restrictions :

$$Y_t - X_{2t} = \hat{\beta}_1^{MCO} + \hat{\beta}_3^{MCO}(X_{3t} + X_{4t}) + \hat{e}_t$$

et l'on garde la valeur $SE2 = \hat{e}'\hat{e} = \sum_{t=1}^{n} \hat{e}_t^2$. Puis on calcule :

$$\frac{(n-k)\,(SE2 - SE)}{mSE}$$

et l'on compare la valeur obtenue aux valeurs critiques d'une table $F(2, n-4)$, ou bien on utilise un logiciel pour obtenir la probabilité critique.

1.2 Test d'une ou de plusieurs restrictions linéaires sur les coefficients

On suppose que Y et X sont liés par un modèle linéaire $Y = X\beta + u$, où u est indépendant de X, est un bruit blanc : $\Sigma_u = \sigma_u^2 I_n$ et a une distribution normale. Un ensemble de m restrictions linéaires sur les coefficients β peut être représenté de manière générale par :

$$R\beta = r \tag{3.2}$$

où R, β et r sont des matrices de type $m \times k$ pour R, $k \times 1$ pour β, et $m \times 1$ pour r.

Exemple (suite)

Les restrictions $\beta_2 = 1$ et $\beta_3 = \beta_4$ peuvent être représentées par :

$$\begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

Sous l'hypothèse H_0 que ces restrictions sont vraies, c'est-à-dire que $R\beta = r$, on montre aisément que :

$$\frac{W}{m} = (R\beta - r)' \left(R \left(X'X\right)^{-1} R'\right)^{-1} (R\beta - r) \frac{1}{m\left(\dfrac{e'e}{n-k}\right)} \sim F(m, n-k) \tag{3.3a}$$

Il s'agit d'un test de Wald *modifié* par une division par m, qui nécessite uniquement une estimation du modèle sans restrictions. On montre que la valeur obtenue est identique à celle du test calculé selon la formule de la section précédente pour les mêmes restrictions, et qui nécessite les estimations du modèle sans restrictions et avec restrictions.

Dès lors que le terme d'erreur est un bruit blanc et normal, il s'agit d'un test de petit échantillon, donc d'un test exact quelle que soit la taille de l'échantillon.

Si u n'a pas une distribution normale, le test est quand même asymptotique. Puisque $\sqrt{n}\left(\hat{\beta} - \beta\right) \xrightarrow{L} N(0, \sigma_u^2 Q^{-1})$ et $p\lim \dfrac{X'X}{n} = Q$, on montre aisément que :

$$W = (R\beta - r)' \left(R \left(X'X\right)^{-1} R'\right)^{-1} (R\beta - r) \frac{1}{\left(\dfrac{e'e}{n-k}\right)} \xrightarrow{L} \chi^2_m \tag{3.3b}$$

Il s'agit d'un vrai test de Wald. Bien que la distribution asymptotique de W soit χ^2, la distribution F appliquée à $\dfrac{W}{m}$ fournit en pratique une meilleure approximation pour n petit. Pour tester l'hypothèse $R\beta = r$, il suffit donc de calculer le membre de gauche de l'expression précédente, de diviser le résultat par m, et de comparer le tout aux valeurs critiques d'une table de la distribution Fischer à m degrés de liberté au numérateur et $n-k$ au dénominateur. Pour réaliser ce test, on n'a besoin uniquement d'estimer le modèle non contraint par MCO, et ensuite de calculer la valeur du test par la formule précédente.

1.3 Test d'une seule restriction linéaire sur les coefficients

On souhaite tester une seule hypothèse H_0 de restriction linéaire sur les coefficients, qui peut se formaliser par :

$$w'\beta = x \tag{3.4}$$

où w est un vecteur $k \times 1$ et x un scalaire 1×1.

Exemple

La restriction $\beta_3 = \beta_4$ peut être représentée par $(0\ 0\ -1\ 1)\begin{pmatrix}\beta_1\\ \beta_2\\ \beta_3\\ \beta_4\end{pmatrix} = (0)$. Dans ce cas, $r = 0$ et $w = \begin{pmatrix}0\\ 0\\ -1\\ 1\end{pmatrix}$.

On montre que sous H_0, donc si $w'\beta = x$ est vraie, alors :

$$\frac{w'\hat{\beta}^{MCO} - x}{\sqrt{\dfrac{e'e}{n-k}}\sqrt{w'\,(X'X)^{-1}\,w}} \sim t_{n-k} \tag{3.5}$$

Pour tester l'hypothèse $w'\beta = x$, il suffit d'estimer le modèle sans restrictions, $Y = X\beta + u$, par MCO. On obtient le modèle estimé $Y = X\beta^{MCO} + e$. On calcule alors la valeur de la formule du test et on la compare aux valeurs critiques d'une distribution t_{n-k}. Il s'agit là d'un cas particulier du test précédent : celui où $m = 1$. On sait en effet qu'une variable aléatoire ayant une distribution $F(1, n-k)$ a une racine carrée qui a une distribution t_{n-k}.

1.4 Tests de valeurs particulières pour tous les coefficients

On souhaite tester l'hypothèse que le vecteur β est égal à un vecteur de valeurs particulières $\bar{\beta}$. On sait que, pour les valeurs « vraies inconnues » β de la réalité, $\dfrac{\left(\hat{\beta}^{MCO} - \beta\right)' X'X \left(\hat{\beta}^{MCO} - \beta\right)(n-k)}{(e'e)\,k} \sim F(k, n-k)$ (voir chapitre 2). Cela implique que si $\beta = \bar{\beta}$, alors $\dfrac{\left(\hat{\beta}^{MCO} - \bar{\beta}\right)' X'X \left(\hat{\beta}^{MCO} - \bar{\beta}\right)(n-k)}{(e'e)\,k} \sim F(k, n-k)$. Pour vérifier l'hypothèse que $\bar{\beta} = \beta$, il suffit donc d'estimer le modèle linéaire $Y = X\beta + u$ sans restrictions. On obtient un modèle estimé $Y = X\beta^{MCO} + e$, on calcule la formule du test et on la compare aux valeurs critiques d'une distribution $F(k, n-k)$. Dès lors que le terme d'erreur est un bruit blanc et normal, il s'agit d'un test de petit échantillon, donc d'un test exact quelle que soit la taille de l'échantillon, c'est-à-dire même avec peu d'observations. C'est un cas particulier du test de Wald modifié, ou test F, de plusieurs restrictions linéaires sur les coefficients lorsque le terme d'erreur est un bruit blanc.

2 Estimation du modèle linéaire par maximum de vraisemblance

L'objet de cette section est d'expliquer comment on peut estimer un modèle linéaire du type $Y = X\beta + u$ par la technique du maximum de vraisemblance [1], sous l'hypothèse que u est indépendant de X et que $u \sim N(0, \Sigma_u)$, où $\Sigma_u = \sigma_u^2 I_n$ (u bruit blanc). Les hypothèses faites sur le terme d'erreur impliquent les égalités suivantes :

$$f_u(u) = f_{u_1 u_2 \ldots u_n}(u_1, u_2, \ldots u_n) = f_{u_1}(u_1) f_{u_2}(u_2) \ldots f_{u_n}(u_n) = \prod_{t=1}^{n} f_{u_t}(u_t) \tag{3.6}$$

$$f_{u_t}(u_t) = \frac{1}{\sigma_u \sqrt{2\pi}} e^{-\frac{u_t^2}{2\sigma_u^2}} \quad \text{pour tout } t \tag{3.7}$$

On observe la réalisation du vecteur Y et celle de la matrice X. La fonction de vraisemblance de l'échantillon est donc $f_{Y, X_2, \ldots X_n}(Y, X_2, \ldots X_n)$. Si l'on raisonne conditionnellement à X, la fonction de vraisemblance de l'échantillon est $f_{Y|X_2, \ldots X_n}(Y|X_2, \ldots X_n)$, notée simplement $f_{Y|X}(Y|X)$. Puisque Y est une fonction de u, la distribution de Y conditionnellement à X doit être calculée à partir de la distribution de u conditionnellement à X, qui est égale à la distribution marginale de u, puisque u et X sont indépendants : $f_u(u|X) = f_u(u)$. Il suffit donc d'appliquer la formule habituelle permettant d'obtenir la fonction de densité d'une variable aléatoire, qui est fonction d'une autre variable aléatoire dont on connaît la densité. Puisque le Jacobien vaut 1, on obtient :

$$f_{Y|X_2, \ldots X_n}(Y|X_2, \ldots X_n) = \prod_{t=1}^{n} \frac{1}{\sigma_u \sqrt{2\pi}} e^{-\frac{\left(Y_t - \beta_1 - \sum_{i=2}^{k} \beta_i X_{it}\right)^2}{2\sigma_u^2}} \underset{\text{déf}}{=} L\left(\beta, \sigma_u^2\right) \tag{3.8}$$

Par conséquent, le logarithme de la fonction de vraisemblance est fourni par :

$$\begin{aligned} \ln L\left(\beta, \sigma_u^2\right) &= -\frac{n}{2} \ln 2\pi - \frac{n}{2} \ln \sigma_u^2 - \sum_{t=1}^{n} \frac{\left(Y_t - \beta_1 - \sum_{i=2}^{k} \beta_i X_{it}\right)^2}{2\sigma_u^2} \\ &= -\frac{n}{2} \ln 2\pi - \frac{n}{2} \ln - \frac{(Y - X\beta)'(Y - X\beta)}{2\sigma_u^2} \end{aligned} \tag{3.9}$$

Les estimateurs de maximum de vraisemblance de β et σ_u^2 sont les solutions du système d'équations formé par les deux conditions de premier ordre :

$$\begin{aligned} 0 &= \frac{\partial \ln L\left(\beta, \sigma_u^2\right)}{\partial \beta} \\ 0 &= \frac{\partial \ln L\left(\beta, \sigma_u^2\right)}{\partial \sigma_u^2} \end{aligned} \tag{3.10}$$

Ces solutions sont :

$$\begin{aligned} \hat{\beta}^{MV} &= \left(X'X\right)^{-1} X'Y \\ \hat{\sigma}_u^{2^{MV}} &= \frac{e'e}{n} \end{aligned} \tag{3.11}$$

1. Pour quelques rappels utiles sur l'estimation par la méthode du maximum de vraisemblance, le lecteur peut se référer par exemple au chapitre 5 du livre de Patrick Roger, Probabilités, statistique et processus stochastiques, publié chez Pearson Education France dans la même collection.

Les conditions de deuxième ordre pour un maximum sont respectées. L'estimateur de maximum de vraisemblance de β, noté $\hat{\beta}^{MV}$, est ainsi égal à l'estimateur des moindres carrés ordinaires de β, noté $\hat{\beta}^{MCO}$. Par contre, l'estimateur de maximum de vraisemblance de σ_u^2, noté $\hat{\sigma}_u^{2^{MV}}$, n'est pas égal à l'estimateur des moindres carrés ordinaires de σ_u^2, noté $\hat{\sigma}_u^{2^{MCO}}$, car $\hat{\sigma}_u^{2^{MV}} = \frac{n-k}{n}\sigma_u^{2^{MCO}}$. Bien sûr, cette différence est d'autant plus petite que n est grand, et elle devient insignifiante quand n est très grand.

La matrice d'information est :

$$R = -E\left(\frac{\partial \ln L\left(\beta, \sigma_u^2\right)}{\partial\begin{pmatrix}\beta\\ \sigma_u^2\end{pmatrix}\partial\begin{pmatrix}\beta\\ \sigma_u^2\end{pmatrix}}\right) = \begin{pmatrix}\sigma_u^{-2}\left(X'X\right) & 0\\ 0 & \frac{n}{2\sigma_u^4}\end{pmatrix} \tag{3.12}$$

Par conséquent, la borne de Rao Cramer est :

$$R^{-1}(\beta, \sigma_u^2) = \begin{pmatrix}\sigma_u^2\left(X'X\right)^{-1} & 0\\ 0 & \frac{2\sigma_u^4}{n}\end{pmatrix} \tag{3.13}$$

L'estimateur de *MCO* de β, égal à l'estimateur de maximum de vraisemblance de β, a une matrice de variance et de covariance égale à la borne de Rao Cramer. Il est donc forcément le plus précis de tous les estimateurs sans biais. Il est en outre efficient. On remarque que $\hat{\sigma}_u^{2^{MV}}$ est biaisé : $E\left(\hat{\sigma}_u^{2^{MV}}\right) \neq \sigma_u^2$ alors que $\hat{\sigma}_u^{2^{MCO}}\sigma_u^2$ est sans biais : $E\left(\hat{\sigma}_u^{2^{MCO}}\right) = \sigma_u^2$. Toutefois $\hat{\sigma}_u^{2^{MCO}}$ a une variance supérieure à la borne de Rao Cramer : $V\left(\hat{\sigma}_u^{2^{MCO}}\right) = \frac{2\sigma_u^4}{n-4} > \frac{2\sigma_u^4}{n}$. Cependant, pour un grand échantillon, on se rapproche de la borne : $\lim_{n\to\infty} V\left(\hat{\sigma}_u^{2^{MCO}}\right) = \frac{2\sigma_u^4}{n}$. Par conséquent, $\hat{\sigma}_u^{2^{MCO}}$ est asymptotiquement efficient.

Test du rapport de vraisemblance. La fonction de vraisemblance fournit un test asymptotique pratique et général pour tester simultanément plusieurs contraintes portant chacune sur un ou plusieurs coefficients du modèle linéaire. Le principe est d'estimer d'abord le modèle linéaire sans restrictions sur ses coefficients. La valeur de la fonction de vraisemblance maximisée de ce modèle est notée L_1. Ensuite, on estime le modèle linéaire en imposant à ses coefficients toutes les restrictions que l'on souhaite tester. La valeur de la fonction de vraisemblance maximisée de ce modèle contraint est notée L_0. On montre que, si les restrictions sont vraies, la valeur absolue de $2\big(\ln(L_1) - \ln(L_0)\big)$ suit asymptotiquement une loi Chi-2 à p degrés de liberté, où p est le nombre de restrictions testées sur les coefficients. On rejette les restrictions quand la valeur de ce test est supérieure aux valeurs critiques à un seuil de 5 % ou de 1 % d'une loi de Chi-2 à p degrés de liberté. En d'autres termes, c'est lorsque les restrictions provoquent une forte chute de la fonction de vraisemblance : L_1 est très supérieure à L_0, ce qui rend la différence de leurs logarithmes trop grande. Le test du rapport de vraisemblance est asymptotique, c'est-à-dire qu'il suit une distribution de Chi-2 quand le nombre d'observations tend vers l'infini. Pour des échantillons réduits, sa distribution n'est qu'approximative ; des tests de petit échantillon peuvent être plus fiables.

3 Multicolinéarité

On parle de *multicolinéarité* lorsqu'il y a des corrélations assez élevées entre certaines variables explicatives, ce qui affecte la matrice $(X'X)^{-1}$. On peut calculer l'estimateur des moindres carrés ordinaires, mais il est d'autant plus imprécis que la multicolinéarité est forte. Plus celle-ci augmente, plus les variances des coefficients estimés croissent aussi. Si des variables explicatives sont très liées linéairement, il est difficile de mesurer l'impact respectif de chacune d'elles sur la variable à expliquer. Souvent, en cas de corrélation forte entre deux variables explicatives, on ne peut rejeter l'hypothèse de nullité de leurs coefficients pris individuellement, alors que l'on rejette l'hypothèse de nullité conjointe. L'observation de ce phénomène indique la présence de multicolinéarité.

La multicolinéarité est dite *parfaite* quand une colonne de la matrice X est une combinaison linéaire d'une ou de plusieurs autres colonnes de X, en d'autres termes quand une variable explicative est une combinaison linéaire exacte d'autres variables explicatives. Dans ce cas, le rang de X est inférieur à k, ce qui implique que le rang de $X'X$ est également inférieur à k. La matrice $X'X$ n'est donc pas inversible ; il est impossible de calculer l'estimateur des moindres carrés ordinaires.

4 Variables indicatrices

Une variable est dite **indicatrice** (ou **dummy**) quand elle est artificielle ; on lui attribue des valeurs particulières aux différentes observations, pour faire varier la valeur de la constante de manière déterminée, en fonction des observations. En économétrie des séries temporelles, on emploie le plus souvent des variables indicatrices d'impulsion, de saut et de saison.

4.1 Variable indicatrice d'impulsion

Une variable indicatrice d'impulsion est telle que toutes ses observations valent 0, sauf une observation, qui vaut 1. Si t_i est la date de l'impulsion, alors :

$$\begin{aligned} DU_t &= 0 \quad \text{si} \quad t \neq t_i \\ DU_t &= 1 \quad \text{si} \quad t = t_i \end{aligned} \tag{3.14}$$

Lorsqu'on introduit une telle variable parmi les k variables explicatives d'un modèle linéaire, celui-ci devient :

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \beta_4 X_{4t} + \cdots + \beta_{k-1} X_{kt-1} + \beta_k DU_t + u_t \quad \text{pour} \quad t = 1 \ldots n \tag{3.15}$$

et :

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \beta_4 X_{4t} + \cdots + \beta_{k-1} X_{kt-1} + u_t \quad \text{pour} \quad t < t_i \ \text{ et } \ t > t_i \tag{3.16}$$

et

$$Y_t = (\beta_1 + \beta_k) + \beta_2 X_{2t} + \beta_3 X_{3t} + \beta_4 X_{4t} + \cdots + \beta_{k-1} X_{kt-1} + u_t \quad \text{pour} \quad t = t_i \tag{3.17}$$

La constante du modèle vaut β_1 en toute période, sauf à la période t_i où elle vaut $\beta_1 + \beta_k$. On utilise une variable indicatrice d'impulsion lorsque la variable dépendante présente un saut à partir d'une date bien particulière, qui ne peut pas être expliqué par un saut comparable, à partir de cette date, d'une variable explicative.

4.2 VARIABLE INDICATRICE DE SAUT

Une variable indicatrice de saut est telle que toutes ses observations valent 0 avant une certaine date, et toutes ses observations valent 1 à partir de cette date :

$$\begin{aligned} DU_t &= 0 \quad \text{si} \quad t < t_i \\ DU_t &= 1 \quad \text{si} \quad t \geqslant t_i \end{aligned} \tag{3.18}$$

Lorsqu'on introduit une telle variable parmi les k variables explicatives d'un modèle linéaire, celui-ci devient :

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \beta_4 X_{4t} + \cdots + \beta_{k-1} X_{kt-1} + \beta_k DU_t + u_t \quad \text{pour} \quad t = 1 \ldots n \tag{3.19}$$

ou encore :

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \beta_4 X_{4t} + \cdots + \beta_{k-1} X_{(k-1)t} + u_t \quad \text{pour} \quad t < t_i \tag{3.20}$$

et

$$Y_t = (\beta_1 + \beta_k) + \beta_2 X_{2t} + \beta_3 X_{3t} + \beta_4 X_{4t} + \cdots + \beta_{k-1} X_{(k-1)t} + u_t \quad \text{pour} \quad t \geqslant t_i \tag{3.21}$$

La constante du modèle vaut β_1 en toute période avant la date t_i, et $\beta_1 + \beta_k$ à la date t_i et ultérieurement. On utilise une variable indicatrice d'impulsion lorsque la variable dépendante présente un saut à partir d'une date bien particulière, qui ne peut pas être expliqué par un saut comparable, à partir de cette date, d'une variable explicative.

4.3 VARIABLE INDICATRICE SAISONNIÈRE

En données mensuelles, on introduit des variables indicatrices saisonnières pour obtenir une constante ayant les propriétés suivantes :

- Elle est différente pour chacun des douze mois d'une même année.
- Elle est identique pour un même mois, quelle que soit l'année.

On définit des variables indicatrices $DU1, DU2 \ldots DU11$ de la manière suivante, quelle que soit l'année :

$DU1_t = 1$ si t est un mois de janvier.

$DU1_t = 0$ si t n'est pas un mois de janvier ou de décembre.

$DU1_t = -1$ si t est un mois de décembre.

$DU2_t = 1$ si t est un mois de février.

$DU2_t = 0$ si t n'est pas un mois de février ou de décembre.

$DU2_t = -1$ si t est un mois de décembre.

...

$DU11_t = 1$ si t est un mois de novembre.

$DU11_t = 0$ si t n'est pas un mois de novembre ou de décembre.

$DU11_t = -1$ si t est un mois de décembre.

Lorsqu'on introduit de telles variables parmi les k variables explicatives d'un modèle linéaire, celui-ci devient :

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \beta_4 X_{4t} + \cdots + \beta_{k-11} X_{(k-11)t} + \beta_{k-10} DU1_t + \beta_{k-9} DU1_t + \cdots + \beta_k DU11_t + u_t \quad \text{pour} \quad t = 1 \ldots n \tag{3.22}$$

et :

$$Y_t = (\beta_1 + \beta_{k-10}) + \beta_2 X_{2t} + \beta_3 X_{3t} + \beta_4 X_{4t} + \cdots + \beta_{k-1} X_{kt-1} + u_t$$
si t est un mois de janvier,
$$Y_t = (\beta_1 + \beta_{k-9}) + \beta_2 X_{2t} + \beta_3 X_{3t} + \beta_4 X_{4t} + \cdots + \beta_{k-11} X_{kt-11} + u_t$$
si t est un mois février,

$\ldots$

$$Y_t = \left(\beta_1 - (\beta_{k-10} + \beta_{k-9} + \cdots + \beta_k)\right) + \beta_2 X_{2t} + \beta_3 X_{3t} + \beta_4 X_{4t} + \cdots + \beta_{k-11} X_{(k-11)t} + u_t \quad \text{si } t \text{ est un mois de décembre.}$$

On utilise des variables indicatrices saisonnières lorsque la variable dépendante présente un cycle saisonnier qui ne peut être expliqué par l'évolution des variables explicatives, ces dernières n'ayant pas de cycle comparable. On évite ainsi que cette composante cyclique inexpliquée se retrouve dans les résidus calculés du modèle estimé, qui sera autocorrélé. La variation de la valeur de la constante, selon les mois de l'année, prend en compte le cycle saisonnier de la variable dépendante, palliant le fait que l'évolution des variables explicatives ne peut l'expliquer.

Remarque

Pour les données trimestrielles, le principe est le même, mais on définit seulement trois variables indicatrices.

5 Estimateurs de variance robustes

5.1 Estimateurs de variance robustes à l'hétéroscédasticité

Si le terme d'erreur u est hétéroscédastique, la vraie matrice $\Sigma_{\hat{\beta}^{MCO}|X}$ de variance et de covariance des estimateurs de *MCO* n'est pas égale à $\sigma_u^2 \left(X'X\right)^{-1}$ et ne peut donc être estimée par $\hat{\Sigma}_{\hat{\beta}^{MCO}|X} = \dfrac{e'e}{n-k}(X'X)^{-1}$, qui est la formule utilisée par défaut par les logiciels d'économétrie. Le plus fréquemment, la forme de l'hétéroscédasticité est inconnue, tout comme la formule précise de $\Sigma_{\hat{\beta}^{MCO}|X}$. Il est donc impossible d'établir un estimateur sur mesure de cette matrice. White [WHI 1980] a toutefois défini un estimateur convergent de $\Sigma_{\hat{\beta}^{MCO}|X}$ quelle que soit la forme de l'hétéroscédasticité. MacKinnon et White [MAC 1985] en ont proposé une version corrigée pour les degrés de liberté, mieux adaptée pour les petits échantillons [1]

$$\hat{\Sigma}^{White^{adj}}_{\hat{\beta}^{MCO}|X} = \frac{n}{n-k}\left(X'X\right)^{-1}\left(\sum_{t=1}^{n} e_t^2 x_t x_t'\right)\left(X'X\right)^{-1}$$

1. La formule originale de White [WHI 1980] est $\hat{\Sigma}^{White}_{\hat{\beta}^{MCO}|X} = (X'X)^{-1}\left(\sum_{t=1}^{n} e_t^2 x_t x_t'\right)(X'X)^{-1}$.

où x_t est un vecteur dont les éléments sont les observations de la ligne t de la matrice X. Les racines carrées des éléments de la diagonale principale de X sont donc des estimateurs convergents des écarts types (standard errors) des estimateurs de *MCO* en cas d'hétéroscédasticité des erreurs, contrairement aux écarts types estimés calculés par défaut par les logiciels d'économétrie. Bien entendu, la convergence est une propriété de grands échantillons : quand ils sont petits, le résultat peut être médiocre.

5.2 Estimateurs de variance robustes à l'hétéroscédasticité et à l'autocorrélation

Si le terme d'erreur u est autocorrélé, la vraie matrice $\Sigma_{\hat{\beta}^{MCO}|X}$ de variance et de covariance des estimateurs de MCO n'est pas égale à $\sigma_u^2 \left(X'X\right)^{-1}$ et ne peut donc être estimée par $\hat{\Sigma}_{\hat{\beta}^{MCO}|X} = \dfrac{e'e}{n-k}\left(X'X\right)^{-1}$, qui est la formule utilisée par défaut par les logiciels d'économétrie.

Le plus fréquemment, la forme de l'autocorrélation est inconnue, tout comme la formule précise de $\Sigma_{\hat{\beta}^{MCO}|X}$. Il est donc impossible d'établir un estimateur sur mesure de cette matrice, d'autant plus que l'autocorrélation s'accompagne souvent d'hétéroscédasticité, dont la forme est également inconnue. Newey et West [NEW 1987] ont toutefois proposé un estimateur convergent de $\Sigma_{\hat{\beta}^{MCO}|X}$ quelles que soient les formes de l'autocorrélation et de l'hétéroscédasticité. On peut lui appliquer la correction de petit échantillon de MacKinnon et White [MAC 1985]. On obtient la formule [1]

$$\hat{\Sigma}^{NW^{adj}}_{\hat{\beta}^{MCO}|X} = \frac{n}{n-k}\left(X'X\right)^{-1}\left(\hat{\Omega}_0 + \sum_{j=1}^{m} w(j,m)\left(\hat{\Omega}_j + \hat{\Omega}'_j\right)\right)\left(X'X\right)^{-1} \tag{3.23}$$

où $\hat{\Omega}_j = \sum_{t=j+1}^{n} e_t e_{t-j} x_t x'_{t-j}$ pour tout $j = 0 \ldots m$ et où les coefficients $w(j,m)$ déterminent une fenêtre de retards (tronquée à m retards). On distingue plusieurs formes de fenêtre :

- la fenêtre uniforme :

$$w(j,m) = 1 \quad \text{pour tout } \; j = 1 \ldots m$$

- la fenêtre de Bartlett :

$$w(j,m) = 1 - \frac{j}{m+1} \quad \text{pour tout } \; j = 1 \ldots m$$

- la fenêtre de Parzen :

$$w(j,m) = 1 - 6\left(\frac{j}{m+1}\right)^2 + 6\left(\frac{j}{m+1}\right)^3 \quad \text{si } \; 1 \leqslant j \leqslant \frac{m+1}{2}$$

et

$$w(j,m) = 2\left(1 - \frac{j}{m+1}\right)^2 \quad \text{si } \; m \geqslant j > \frac{m+1}{2}$$

1. La formule initiale de Newey et West est $\hat{\Sigma}^{NW}_{\hat{\beta}^{MCO}|X} = \left(X'X\right)^{-1}\left(\hat{\Omega}_0 + \sum_{j=1}^{m} w(j,m)\left(\hat{\Omega}_j + \hat{\Omega}'_j\right)\right)\left(X'X\right)^{-1}$.

La valeur de m doit être choisie arbitrairement. Un terme d'erreur autocorrélé en moyenne mobile d'un ordre connu impose l'usage d'une fenêtre uniforme. C'est le cas en finance, lorsqu'on veut tester l'hypothèse d'efficience d'un marché et que l'horizon des anticipations dépasse la longueur de l'intervalle temporel entre les données successives. Dans les autres cas, la fenêtre de Parzen est généralement préférable aux autres formes. Les fenêtres de Bartlett et de Parzen garantissent l'obtention d'une matrice de variance et de covariance estimée qui soit semi-définie positive, même si m est grand par rapport à n.

Les racines carrées des éléments de la diagonale principale de la matrice de Newey et West sont donc des estimateurs convergents des écarts types des estimateurs de *MCO* en cas d'autocorrélation ou d'hétéroscédasticité des erreurs, contrairement aux écarts types estimés calculés par défaut par les logiciels d'économétrie. Bien entendu, la convergence est une propriété de grands échantillons : quand ils sont petits, le résultat peut être médiocre.

La matrice de White décrite précédemment correspond en fait à celle de Newey et West quand $m = 0$.

6 Estimateur des moindres carrés généralisés

En cas d'autocorrélation ou d'hétéroscédasticité du terme d'erreur u d'un modèle linéaire $Y = X\beta + u$, u n'est pas un bruit blanc : $\Sigma_u \neq \sigma_u^2 I_n$ mais $\Sigma_u = \sigma^2\Omega$ où Ω est définie positive. Il en résulte que, même si le terme d'erreur est indépendant des variables explicatives, la matrice de variance et de covariance de l'estimateur des moindres carrés ordinaires n'est plus égale à $\sigma_u^2 \left(X'X\right)^{-1}$ et ne peut donc être estimée par $\frac{e'e}{n-k}\left(X'X\right)^{-1}$. Les écarts types des coefficients de *MCO* publiés par défaut par les logiciels sont donc faux, puisqu'ils sont calculés d'après la formule $\frac{e'e}{n-k}\left(X'X\right)^{-1}$, valable uniquement si u est un bruit blanc.

Des estimateurs linéaires sans biais, plus précis que l'estimateur de *MCO* sont disponibles, et celui-ci n'est plus un estimateur linéaire sans biais de variance minimale.

Cela dit, si u est indépendant des variables explicatives, les estimateurs de *MCO* restent sans biais et convergents. Leur précision est en fait fortement altérée par l'autocorrélation et l'hétéroscédasticité. Ces dernières augmentent la probabilité que, sur un échantillon particulier, une valeur estimée soit éloignée de la valeur vraie du coefficient.

Si $Y = X\beta + u$, que $\Sigma_u = \sigma^2\Omega$, que $E(u) = 0$, et si u est indépendant des X_i et Ω connu (σ^2 étant un paramètre inconnu, tout comme le vecteur β), la méthode des moindres carrés généralisés (MCG) permet de mieux estimer β que celle des *MCO*.

L'estimateur de MCG est défini par :

$$\hat{\beta}^{MCG} = \left(X'\Omega^{-1}X\right)^{-1} X'\Omega^{-1}Y \tag{3.24}$$

Si l'on pose $e = Y - X\beta^{MCG}$, l'expression $\frac{e'e}{n-k}$ est un estimateur sans biais de σ^2. On montre aussi que sous les hypothèses décrites précédemment :

$$\Sigma_{\hat{\beta}^{MCG}} = \sigma^2 \left(X'\Omega^{-1}X\right)^{-1} \tag{3.25}$$

L'application pratique directe des estimateurs des moindres carrés généralisés est assez limitée car elle exige la connaissance préalable de la matrice de variance et de covariance Σ_u du vecteur u des termes d'erreur successifs.

7 Prévision

Après avoir estimé un modèle linéaire, on peut l'utiliser pour produire des prévisions de la variable dépendante sur des périodes postérieures à celles de l'échantillon ayant servi à l'estimation. Ces prévisions sont évidemment conditionnelles à des valeurs supposées des variables explicatives durant les périodes concernées. De manière générale, si le modèle linéaire a été estimé, la prévision ponctuelle de la valeur de la variable dépendante à la période $n + l$ est donnée par :

$$\hat{Y}_{n+l} = \hat{\beta}_1^{MCO} + \hat{\beta}_2^{MCO} X_{2,n+l} + \cdots + \hat{\beta}_k^{MCO} X_{k,n+l} = \hat{\beta}^{MCO'} X_{n+l} \qquad (3.26)$$

où $X_{n+l} = (1 X_{2,n+l} \ldots X_{kn+l})'$. On a obtenu les coefficients estimés en utilisant les matrices Y et X qui rassemblent les données des périodes 1 à n. Sous l'hypothèse que le terme d'erreur u est un bruit blanc et normal, un intervalle de confiance de la valeur de Y_{n+l} est fourni par :

$$P\left(\hat{\beta}^{MCO'} X_{n+l} - t_{\frac{\alpha}{2}} \sqrt{\left(\frac{e'e}{n-k}\right)\left(1 + X'_{n+l}(X'X)^{-1} X_{n+l}\right)}\right.$$

$$\left. \leqslant Y_{n+l} \leqslant \hat{\beta}^{MCO'} X_{n+l} + t_{\frac{\alpha}{2}} \sqrt{\left(\frac{e'e}{n-k}\right)\left(1 + X'_{n+l}(X'X)^{-1} X_{n+l}\right)}\right) = 1 - \alpha$$

où $t_{\alpha/2}$ est la valeur critique d'une loi de Student à $n - k$ degrés de liberté au seuil de signification $\alpha/2$ et où e est le vecteur des résidus issus de l'estimation du modèle linéaire sur les périodes 1 à n.

8 Modèles linéaires en coupe instantanée

8.1 Généralités

Le concept de modèle linéaire peut s'appliquer à des variables dépendante et explicatives qui n'ont pas la dimension du temps, mais pour lesquelles les différentes observations correspondent à des agents distincts durant une même période de temps. Un modèle linéaire suppose alors l'existence de coefficients non aléatoires $\beta_1, \beta_2 \ldots \beta_k$ identiques quel que soit l'agent concerné, tels que :

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} + u_i \quad \text{pour tout} \quad i = 1 \ldots n \qquad (3.27)$$

où n est l'effectif des agents différents.

Dans une régression en coupe instantanée, le concept d'autocorrélation, désignant une dépendance du terme d'erreur vis-à-vis de son passé, n'a plus aucun sens tel quel. Les tests d'autocorrélation comme celui de Durbin et Watson ne sont donc pas appliqués en coupe instantanée. Mais cela n'écarte pas le risque que les erreurs de certains agents soient liées entre elles. Le concept d'autocorrélation est simplement remplacé par l'absence d'indépendance entre les erreurs d'agents différents. En série temporelle, l'oubli de variables explicatives importantes, comme une variable dépendante retardée, peut impliquer une dépendance entre les erreurs d'une période particulière t et des périodes précédentes, comme $t - 1$. En coupe instantanée, si l'étude porte sur des agents pouvant être répartis en différentes classes, l'erreur d'un agent particulier d'une classe quelconque risque d'être liée aux erreurs des agents de cette même classe.

8.2 Variables indicatrices et méthodes ANOVA

Comme en série temporelle, on peut utiliser dans des régressions en coupe instantanée des variables indicatrices qui valent, pour chaque agent, 1 ou 0 selon que l'agent concerné appartient ou pas à une certaine catégorie. Le coefficient d'une telle variable exprime alors dans quelle mesure l'appartenance de l'agent à cette catégorie influe sur la valeur de la variable dépendante. Cette analyse est très répandue en sciences de gestion, où elle est connue sous le nom d'ANOVA (ANalysis Of VAriance, en français « analyse de variance »). On distingue plusieurs types d'ANOVA : celle à un facteur (la plus simple) ou encore celle à deux facteurs avec répétition d'expérience. L'objectif général d'une ANOVA est de déterminer si la valeur d'une variable métrique prise (par une personne, une institution, pays, région, entreprise...) dépend des modalités prises par une ou plusieurs caractéristiques qualitatives de cet individu (plus précisément, par une caractéristique dans le cas de l'ANOVA à un facteur, et par deux dans le cas de l'ANOVA à deux facteurs avec répétition d'expérience). Une ANOVA à un facteur revient statistiquement à tester si l'espérance de la variable métrique est égale pour des individus ayant une même modalité de la caractéristique qualitative et pour d'autres individus partageant aussi une même modalité, mais différente de la précédente. Pour l'ANOVA à deux facteurs avec répétition d'expérience, on doit disposer, pour chaque modalité de la première caractéristique qualitative, de groupes d'observations différents de la variable pour chaque modalité différente de la deuxième caractéristique. Vérifier l'hypothèse nulle revient ici à tester si l'espérance de la variable pour un individu pris au hasard est la même quelles que soient :

- la modalité qu'il a pour la caractéristique 1 ;
- la modalité qu'il a pour la caractéristique 2 ;
- la combinaison de modalités qu'il a pour les caractéristiques 1 et 2 (interaction).

9 Méthode des variables instrumentales

9.1 Principe de la méthode

La méthode des variables instrumentales, ou auxiliaires, s'applique lorsque le terme d'erreur d'une équation linéaire n'est pas indépendant de certaine(s) variable(s) explicative(s). Dans ce cas en effet, l'estimateur des moindres carrés ordinaires est biaisé et n'est pas convergent (le biais ne tend donc pas à disparaître si la taille de l'échantillon est grande). Il produirait donc une erreur **systématique** des valeurs estimées. Le cas le plus fréquent où le terme d'erreur n'est pas indépendant de certaine(s) variable(s) explicative(s) est le cas où il y a simultanéité, et donc un impact de la variable dépendante sur une variable explicative.

Soit un modèle linéaire habituel :

$$Y_t = \beta_1 + \beta_2 X_{2t} + \ldots \beta_k X_{kt} + u_t \quad \text{ou} \quad Y = X\beta + u \tag{3.28}$$

où u_t est bruit blanc et où certaine(s) variable(s) explicative(s), parmi les X_i, ne sont pas indépendantes du terme d'erreur u_t. On ne peut donc pas estimer les β_i en calculant par la méthode des moindres carrés ordinaires une régression linéaire des données de Y sur les données des X_i.

L'idée est alors de « purger » les variables explicatives X_i de leur partie liée linéairement à u, c'est-à-dire de remplacer les X_i par des variables transformées $\hat{X}_i$ proches des X_i initiales, mais non liées à u. Bien entendu, on ne touche pas aux variables explicatives X_j, parmi les X_i, qui ne semblent pas liées à u. Pour ces quelques variables X_j, $\hat{X}_j = X_j$.

Comment fabriquer les transformations $\hat{X}_i$? Pour chaque variable explicative X_i, on prend simplement la partie expliquée d'une régression linéaire de X_i sur un ensemble de p variables instrumentales (ou auxiliaires) que l'on suppose de façon arbitraire non liées linéairement à u. Forcément, la partie expliquée de X_i par ces variables instrumentales n'est pas liée à u puisqu'il s'agit d'une fonction linéaire de variables non liées à u. On sélectionne donc *a priori* p variables instrumentales W_r, pour $r = 1 \ldots p$. W_1 est une simple colonne de 1 et donc égale à X_1. Les variables W_r peuvent contenir certaines variables X_i originales : celles supposées indépendantes de u. Il faut que p soit plus grand ou égal à k. Les p colonnes de la matrice W, qui contient n lignes, sont les différentes variables instrumentales W_r, pour $r = 1 \ldots p$.

Pour chaque variable X_i, on calcule par *MCO* le modèle estimé suivant :

$$X_{it} = \gamma_1^{MCO} + \gamma_2^{MCO} W_{2t} + \cdots \gamma_p^{MCO} W_{pt} + e_{it} \tag{3.29}$$

Cela permet de définir la partie expliquée de X_i par les variables instrumentales :

$$\hat{X}_{it} = \hat{\gamma}_{i1}^{MCO} + \hat{\gamma}_{i2}^{MCO} W_{2t} + \cdots \hat{\gamma}_{ip}^{MCO} W_{pt} \tag{3.30}$$

Bien entendu, dans les cas particuliers où X_i figure parmi les variables instrumentales, les coefficients estimés sont tous nuls, sauf celui de la variable elle-même, qui est égal à 1 ; on obtient par ailleurs un résidu toujours nul, ce qui implique que $\hat{X}_{it} = X_{it}$. On estime alors par *MCO* une régression de Y sur les $\hat{X}_i$:

$$Y_t = \hat{\varphi}_1^{MCO} + \hat{\varphi}_2^{MCO} \hat{X}_{2t} + \cdots \hat{\varphi}_k^{MCO} \hat{X}_{kt} + e_t \tag{3.31}$$

Les coefficients estimés obtenus sont les estimateurs de variables instrumentales des coefficients « vrais inconnus » β_i :

$$\hat{\beta}^{VI} = \begin{pmatrix} \hat{\beta}_1^{VI} \\ \hat{\beta}_2^{VI} \\ \vdots \\ \hat{\beta}_k^{VI} \end{pmatrix} = \begin{pmatrix} \hat{\varphi}_1^{MCO} \\ \hat{\varphi}_2^{MCO} \\ \vdots \\ \hat{\varphi}_k^{MCO} \end{pmatrix} \tag{3.32}$$

Si l'on regroupe dans une matrice X les k variables X_i, X a n lignes et k colonnes. Si l'on regroupe dans une matrice W les p variables W_i, W a n lignes et p colonnes. L'ensemble des régressions de chaque variable X_i sur les variables instrumentales peut donc être représenté de manière matricielle :

$$X = W\hat{\gamma}^{MCO} + E \tag{3.33}$$

où $\hat{\gamma}^{MCO}$ est une matrice à p lignes et k colonnes, chaque colonne étant l'un des vecteurs $\hat{\gamma}_i^{MCO}$ qui regroupe les p coefficients $\hat{\gamma}_{ir}^{MCO}$ de la régression de X_i sur les variables instrumentales W_r, pour $r = 1 \ldots p$. E est une matrice à n lignes et à k colonnes correspondant aux k vecteurs de résidus e_i. On utilise la définition des estimateurs de *MCO* :

$$\hat{\gamma}^{MCO} = (W'W)^{-1} W'X \tag{3.34}$$

On regroupe les k variables $\hat{X}_i$ dans une matrice $\hat{X}$ à n lignes et k colonnes :

$$\hat{X} = W\hat{\gamma}^{MCO} = W(W'W)^{-1}W'X \tag{3.35}$$

L'estimateur de variables instrumentales de β est alors l'estimateur de *MCO* de la régression de Y sur les variables $\hat{X}_i$ de $\hat{X}$. Par conséquent, $\hat{\beta}^{VI} = (\hat{X}'\hat{X})^{-1}\hat{X}'Y$, ce qui implique :

$$\hat{\beta}^{VI} = \left(X'W(W'W)^{-1}W'X\right)^{-1}X'W(W'W)^{-1}W'Y \tag{3.36}$$

Il s'agit de la formule générale de l'estimateur de variables instrumentales. Dans le cas particulier où $p = k$, $W'X$ est une matrice carrée inversible et l'estimateur s'écrit ainsi :

$$\hat{\beta}^{VI} = (W'X)^{-1}W'Y \tag{3.37}$$

Dans tous les cas, dès lors que les variables instrumentales sont indépendantes de u, on montre que ces estimateurs sont convergents :

$$p\lim\hat{\beta}^{VI} = \beta$$

Les résidus d'une estimation par variables instrumentales sont définis par $e^{VI} = Y - X\hat{\beta}^{VI}$ et un estimateur convergent de σ_u^2 est fourni par la formule $e^{VI'}e^{VI}/(n-k)$ si u est bruit blanc. Un estimateur convergent de la matrice de variance et de covariance de $\hat{\beta}^{VI}$ est alors donné par $\left((e^{VI'}e^{VI})/(n-k)\right)\left(X'W(W'W)^{-1}W'X\right)^{-1}$.

Remarque

Il ne faut pas confondre les résidus d'une estimation par variables instrumentales avec les résidus d'une estimation par moindres carrés ordinaires en deux étapes, ou doubles moindres carrés, définis par $e^{2MC} = Y - \hat{X}\hat{\beta}^{VI}$.

9.2 Qualité de l'ajustement et tests

Les statistiques R^2 et R^2 ajusté, quoique fournies par la plupart des logiciels dans les tableaux de résultats obtenus par variables instrumentales, ne sont en fait pas valables et peuvent conduire à des conclusions erronées [PES 1994]. Dans le contexte des variables instrumentales, on doit leur préférer les statistiques R^2 et R^2 ajustée généralisées. Celles-ci sont calculées d'après les mêmes formules que les statistiques R^2 et R^2 ajustée ordinaires, mais on remplace la série de résidus e par la série $e^{VI} + (X - \hat{X})\hat{\beta}^{VI}$.

Pour vérifier l'hypothèse d'absence d'autocorrélation des erreurs, on ne peut se servir du test de Durbin et Watson car il n'est pas valable sur des résidus estimés par variables instrumentales. On peut par contre utiliser le test LM d'autocorrélation des résidus estimés par variables instrumentales de Sargan, qui est distribué comme une χ^2 à p degrés de liberté sous l'hypothèse nulle d'absence d'autocorrélation du terme d'erreur contre l'hypothèse d'autocorrélation d'ordre p. Ce test est décrit par Breusch et Godfrey [BRE 1981]. Pesaran et Taylor [PES 1997] présentent en détail les tests applicables dans le contexte des variables instrumentales.

Résumé

Dès lors que le terme d'erreur est un bruit blanc et normal, on dispose de la distribution exacte de plusieurs tests qui permettent de vérifier des hypothèses portant simultanément sur plusieurs coefficients d'un modèle linéaire. Sous l'hypothèse de normalité, un modèle linéaire peut aussi être estimé par la méthode du maximum de vraisemblance, qui fournit les mêmes estimateurs des coefficients que ceux obtenus avec les moindres carrés ordinaires. La multicolinéarité réduit la précision des estimateurs de MCO. L'utilisation de variables indicatrices en séries temporelles permet de modéliser la saisonnalité et des événements particuliers non pris en compte par les autres variables explicatives. Alors que les écarts types obtenus de manière classique ne sont corrects que si les erreurs sont un bruit blanc, on dispose d'estimateurs des écarts types des coefficients estimés qui sont robustes même en cas d'hétéroscédasticité ou d'autocorrélation. C'est aussi dans ces situations que les estimateurs de MCG sont plus précis que les estimateurs de MCO. On peut utiliser un modèle estimé par *MCO* pour générer des prévisions ponctuelles ou des intervalles de confiance pour la variable dépendante. La méthode des *MCO* peut s'appliquer à des modèles linéaires reliant des données en coupe instantanée ; dans ce contexte, l'utilisation de certaines variables indicatrices produit des analyses ANOVA. Lorsque le terme d'erreur est lié à certaines variables explicatives, il faut recourir à la méthode des variables instrumentales pour obtenir des estimateurs convergents.

Problèmes et exercices

Exercice 1 Ventes d'une entreprise de grande distribution

Énoncé

Vous disposez de données historiques sur les ventes de l'une des principales chaînes de grande distribution des États-Unis et sur le revenu disponible des ménages américains, dans le fichier de type tableur DISTRIB.xls téléchargeable sur le site Internet www.pearsoneducation.fr. Les définitions des variables de ce fichier sont les suivantes :

VENTES : les ventes totales de tous les magasins de l'entreprise ;

REVDISMN : le revenu disponible des ménages aux États-Unis.

Sur la base de ces données, répondez aux questions suivantes :

- Les achats des ménages dans les magasins de cette entreprise augmentent-ils plus vite, aussi vite ou moins vite que le revenu disponible des ménages ?
- Comment chiffrer à l'avance la croissance des ventes de l'entreprise résultant d'une croissance donnée du revenu disponible des ménages ?

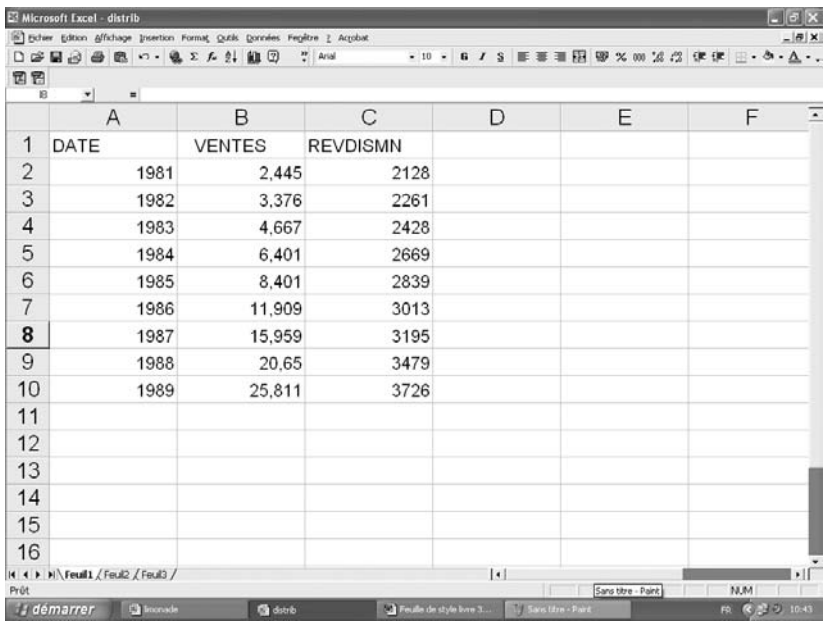
Résolvez ce problème avec un logiciel tableur et TSP.

Solution

Solution avec Excel

Les données se présentent de la manière suivante, dans le fichier DISTRIB.XLS (voir figure 3.1).

Figure 3.1

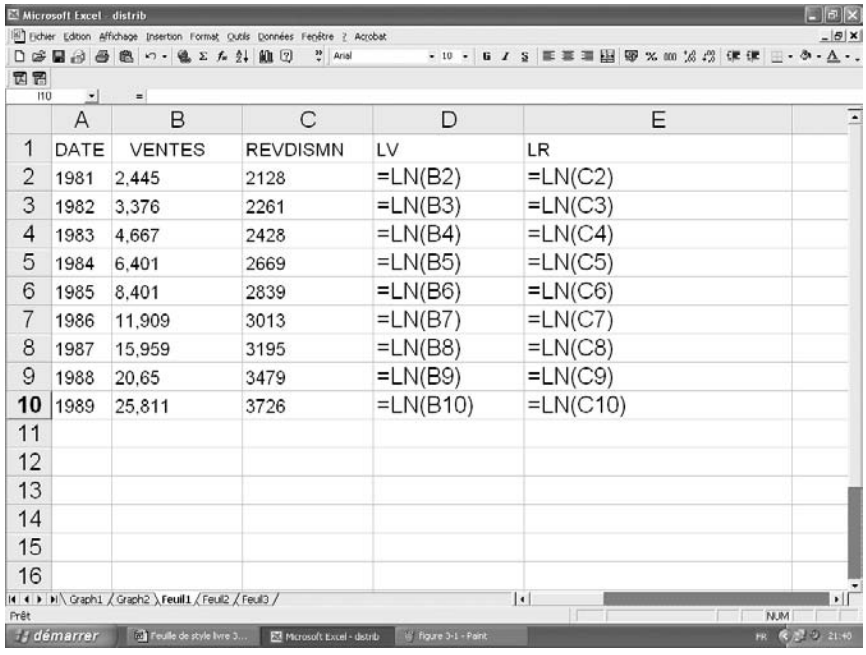


Les questions de l'exercice portent sur la relation entre les ventes de l'entreprise et le revenu disponible des ménages américains. Pour mesurer cette relation, il est utile de visualiser les données au moyen d'un graphique.

Une relation existe clairement entre les deux variables : les ventes augmentent quand le revenu disponible augmente. Elle ne semble pas tout à fait linéaire, mais plutôt légèrement exponentielle. Il est donc intéressant de visualiser aussi le graphique des logarithmes de ces variables. Pour cela, créez des variables LV et LR égales aux logarithmes népériens des variables initiales (voir figure 3.2).

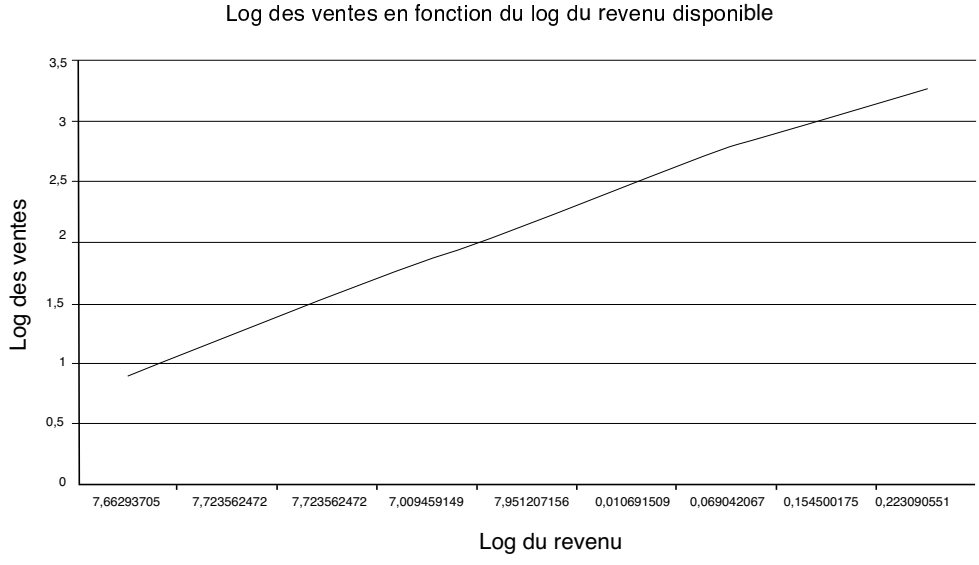
Vous obtenez le graphique de LV en fonction de LR (voir figure 3.3).

Figure 3.2



	A	B	C	D	E
1	DATE	VENTES	REVDISMN	LV	LR
2	1981	2,445	2128	=LN(B2)	=LN(C2)
3	1982	3,376	2261	=LN(B3)	=LN(C3)
4	1983	4,667	2428	=LN(B4)	=LN(C4)
5	1984	6,401	2669	=LN(B5)	=LN(C5)
6	1985	8,401	2839	=LN(B6)	=LN(C6)
7	1986	11,909	3013	=LN(B7)	=LN(C7)
8	1987	15,959	3195	=LN(B8)	=LN(C8)
9	1988	20,65	3479	=LN(B9)	=LN(C9)
10	1989	25,811	3726	=LN(B10)	=LN(C10)
11					
12					
13					
14					
15					
16					

Figure 3.3



La linéarité est plus nette dans la relation entre les logarithmes des séries que dans la relation entre les séries brutes. Par ailleurs, la problématique de départ porte sur une élasticité :

- Les ventes de l'entreprise augmentent plus (respectivement moins) vite que le revenu si l'élasticité des ventes au revenu est supérieure (respectivement inférieure) à 1.
- La réponse de la croissance des ventes à celle du revenu dépend de l'élasticité des ventes au revenu.

Or, l'élasticité des ventes au revenu est la dérivée partielle du logarithme des ventes au logarithme du revenu. Conclusion : c'est la relation entre les logarithmes des ventes et du revenu qui est intéressante. Il faut donc estimer les coefficients du modèle suivant :

$$\ln(VENTES_t) = \beta_1 + \beta_2 \ln(REVDISMN_t) + u_t$$

où u_t est le terme d'erreur, c'est-à-dire la partie de l'évolution du logarithme des ventes qui ne s'explique pas linéairement par l'évolution du logarithme du revenu.

Pour réaliser cette estimation par moindres carrés ordinaires avec Excel, procédez ainsi :

Cliquez sur Outils et vérifiez que l'option Utilitaire d'analyse est dans le menu déroulant. Si elle ne s'y trouve pas, cliquez sur l'option Macros complémentaires et, dans le menu déroulant, sélectionnez Utilitaire d'analyse puis cliquez sur OK. Cette fois, au prochain clic sur Outils, vous accéderez à l'Utilitaire d'analyse.

Cliquez donc sur Outils puis sur Utilitaire d'analyse. Dans le menu déroulant, sélectionnez Régression linéaire et cliquez sur OK. Dans Plage pour la variable *Y*, placez les cellules des données de la variable dépendante : $D$1:$D$10. Dans Plage pour la variable *X*, placez les cellules des données de la variable explicative : $E$1:$E$10. Cochez l'option Intitulé présent pour indiquer que ces plages contiennent les noms des variables (voir figure 3.4).

Figure 3.4

Cliquez sur OK. Vous obtenez les résultats (voir figure 3.5, page suivante).

La valeur estimée de β_2, calculée par la formule de l'estimateur $\hat{\beta}_2^{MCO}$, est donc 4,25711. Il s'agit d'une estimation de l'élasticité β_2 des ventes de l'entreprise au revenu disponible des

Figure 3.5

Microsoft Excel - distrib

Fichier Edition Affichage Insertion Format Outils Données Fenêtre ? Acrobat

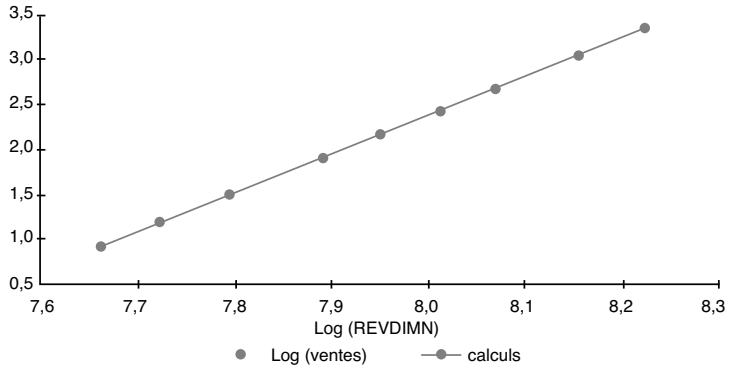
K26 =

	A	B	C	D	E	F	
1	RAPPORT DÉTAILLÉ						
2							
3	*Statistiques de la régression*						
4	Coefficient de	0,99753472					
5	Coefficient de	0,99507552					
6	Coefficient de	0,99437202					
7	Erreur-type	0,06150006					
8	Observations	9					
9							
10	ANALYSE DE VARIANCE						
11		*Degré de libert*	*mme des car*	*yenne des car*	*F*	*Valeur critique de F*	
12	Régression	1	5,34988437	5,34988437	1414,46845	2,44359E-09	
13	Résidus	7	0,0264758	0,00378226			
14	Total	8	5,37636017				
15							
16		*Coefficients*	*Erreur-type*	*Statistique t*	*Probabilité*	*Limite inférieure pour seuil de confiance = 95%*	*Limite supérieure pour*
17	Constante	-31,6815938	0,89922904	-35,2319514	3,8514E-09	-33,80793112	
18	LR	4,25710517	0,11319252	37,6094197	2,4436E-09	3,989447574	
19							
20							

ménages, sur la période 1981 à 1989. Par conséquent, de 1981 à 1989, une augmentation de 1 % du revenu disponible des ménages induit en moyenne une augmentation de 4,25711 % des ventes de l'entreprise (voir figure 3.1, page 91).

Le graphique de la figure 3.6, compare les vraies valeurs des ventes en logarithme et les valeurs calculées par la régression ; l'ajustement est excellent. Les erreurs sont très faibles.

Figure 3.6



Réponses aux questions de l'exercice

Durant la période étudiée, les achats des ménages auprès de cette entreprise de grande distribution ont augmenté plus vite que leur revenu disponible puisque l'élasticité estimée des ventes au revenu est supérieure à 1.

En supposant que cette élasticité reste stable dans le temps, vous pouvez l'utiliser pour prévoir la croissance des ventes de l'entreprise conditionnellement à une prévision donnée de croissance du revenu disponible des ménages. Par exemple, si vous prévoyez une croissance de 1,5 % du revenu disponible des ménages, vous pouvez en inférer une prévision de croissance de $4{,}25711 \times 1{,}5 = 6{,}385665$ % des ventes de l'entreprise.

Quelle est la fiabilité de ces conclusions?

D'après les résultats précédents, l'élasticité des ventes au revenu est estimée à 4,25711 durant la période 1980–1989. Cette valeur estimée est-elle fiable? Pour répondre à cette question, il faut tenir compte des points suivants :

- Les valeurs calculées de $\hat{\beta}_1^{MCO}$ et $\hat{\beta}_2^{MCO}$ ne sont pas les vraies valeurs de β_1 et β_2, qui restent inconnues. En effet, $\hat{\beta}_1^{MCO}$ et $\hat{\beta}_2^{MCO}$ sont des estimateurs.
- La précision de ces estimations est inversement proportionnelle aux écarts types des estimateurs. Il faut donc examiner les écarts types estimés des coefficients pour connaître leur précision.

Les logiciels donnent des estimations de ces écarts types, calculées selon une méthode qui n'est correcte que si le terme d'erreur est un bruit blanc.

Avant de se fier aux écarts types affichés, il faut donc tester deux hypothèses :

- L'absence d'autocorrélation : le terme d'erreur n'est pas lié à son passé.
- L'absence d'hétéroscédasticité : le terme d'erreur a une dispersion constante.

Pour apprécier la fiabilité des conclusions qui ont été tirées, il faut examiner les résultats de ces tests.

Solution avec TSP

Ouvrez le fichier DISTRIB.XLS initial, contenant les données de la variable VENTES et de la variable REVDISMN, tel qu'il était avant les manipulations réalisées avec le tableur. Il n'est pas nécessaire d'insérer une colonne avec les dates ; mais si vous le faites, nommez-la DATE. Pour que le fichier DISTRIB.XLS puisse être lu par TSP, il faut l'enregistrer comme un fichier de type feuille de calcul Excel (quelle que soit la version), et non comme un classeur. Enregistrez ce fichier dans le répertoire C:\ de l'ordinateur, par exemple. Fermez-le à présent pour pouvoir le lire avec TSP. Pour les raisons évoquées dans le chapitre 2, il faut estimer les coefficients du modèle suivant :

$$\ln(VENTES_t) = \beta_1 + \beta_2 \ln(REVDISMN_t) + u_t$$

Pour travailler avec TSP, suivez la procédure expliquée au chapitre 2. Ici le programme d'instructions est le suivant :

```
FREQ A;
SMPL 1981 1989;
READ(FILE='C:\DISTRIB.XLS');
LV=LOG(VENTES);
LR=LOG(REVDISMN);
REGOPT(PVPRINT,STARS) ALL;
NOPLOT;
OLSQ LV C LR;
```

L'instruction `FREQ A;` indique que la fréquence des données est annuelle. L'instruction `SMPL 1981 1989;` définit la période sur laquelle portent les données. L'instruction `READ(FILE='C:\DISTRIB.XLS');` lit les données dans le fichier DISTRIBT.XLS qui se trouve dans C:\. `LV=LOG(VENTES);` et `LR=LOG(REVDISMN);` fabriquent de nouvelles variables `LV` et `LR` égales au logarithme népérien des variables lues. `REGOPT(PVPRINT,STARS)ALL;` demande tous les tests disponibles (`ALL`), avec mention des probabilités critiques (`PVPRINT`) et présence d'étoiles (`STARS`) si l'hypothèse testée

est rejetée. `NOPLOT;` indique qu'aucun graphique sur les résultats de la régression n'est souhaité. `OLSQ LV C LR;` calcule la régression linéaire de `LV` sur une constante et sur `LR` par MCO.

Les résultats sont les suivants :

```
                                Equation 1
                               ============
                  Method of estimation = Ordinary Least Squares
Dependent variable : LV
Current sample : 1981 to 1989
Number of observations : 9
 Mean of dependent variable = 2.12910
Std. dev. of dependent var. = .819784
   Sum of squared residuals = .026475
      Variance of residuals = .378217E-02
   Std. error of regression = .061499
                  R-squared = .995076
         Adjusted R-squared = .994372
    Durbin-Watson statistic = 1.47226
  Wald nonlin. AR1 vs. lags = .122628 [.726]
                  ARCH test = .536009 [.464]
                 CuSum test = .361090 [1.00]
               CuSumSq test = .195113 [.733]
                  Chow test = .343988 [.724]
     LR het. test (w/ Chow) = -1.14689 [1.00]
           White het.  test = 1.67129 [.434]
 Jarque-Bera normality test = .619370 [.734]
  F-statistic (zero slopes) = 1414.50 ** [.000]
   Akaike Information Crit. = -2.54645
 Schwarz Bayes. Info. Crit. = -5.34050
 Log of likelihood function = 13.4590
            Estimated    Standard
Variable   Coefficient    Error       t-statistic     P-value
C          -31.6816       .899219     -35.2323     ** [.000]
LR         4.25711        .113191     37.6098      ** [.000]
```

L'interprétation des coefficients estimés est étudiée au chapitre 2. Des tests supplémentaires permettent d'apprécier la fiabilité des résultats. L'hypothèse d'absence d'autocorrélation du terme d'erreur [1] n'est pas rejetée puisque le test de Durbin et Watson vaut 1,47, pas plus que l'hypothèse d'absence d'hétéroscédasticité du terme d'erreur au seuil de 5 % car les probabilités critiques (entre [ ]) des tests White het, LR het, et ARCH test sont toutes supérieures à 0,05. La validité de ces tests asymptotiques est évidemment très incertaine sur un échantillon aussi petit, mais on admet ici que les erreurs sont homoscédastiques, essentiellement pour des raisons pédagogiques. Vous pouvez donc utiliser les écarts types estimés pour apprécier la précision de l'estimation de β_1 et β_2. L'écart type d'un estimateur est une mesure de l'imprécision de l'estimateur. La précision de l'estimation de $\beta 2$ est très

1. Pour $n = 15$ et $k = 2$, les valeurs critiques sont $dL = 1{,}08$ et $dU = 1{,}36$ au seuil de 5 %. La valeur du test est supérieure à dU et inférieure à $4 - dU$, ce qui permet de ne pas rejeter l'hypothèse d'absence d'autocorrélation.

grande puisque le coefficient estimé vaut 37 fois son écart type. Un intervalle de confiance pour β2 est donc très étroit autour de 4,25711. Les conclusions tirées sont fiables.

La précision d'un estimateur est d'autant plus faible que son écart type est élevé par rapport à la valeur du coefficient. Inversément, la précision d'un estimateur est d'autant plus élevée que son écart type est faible par rapport à la valeur du coefficient. La colonne *t*-statistic donne, pour le coefficient de chaque variable explicative, le rapport entre sa valeur estimée et l'écart type estimé de l'estimateur. La précision avec laquelle un coefficient est estimé est donc d'autant plus grande que sa *t*-stat est élevée. Ici les coefficients sont estimés avec une très grande précision : les valeurs estimées de β_1 et de β_2 valent, pour la première, 35 fois son écart type et, pour la seconde, 37 fois.

Exercice 2 Ventes de boisson au cola

Énoncé

Il s'agit d'établir, début 1981, une prévision des ventes d'une marque de boisson au cola pour les années 1981 à 1990. Les données de l'entreprise dont vous disposez sont les séries suivantes, de 1970 à 1980 :

SALES : les ventes de l'entreprise à prix courants (nominales) ;

PRICE : le prix de vente de boisson au cola de l'entreprise.

Sont également disponibles des données **externes** qui peuvent être utiles à l'analyse, pour expliquer les ventes de l'entreprise :

Y : le revenu disponible *réel* des ménages (c'est-à-dire le revenu disponible nominal divisé par l'indice des prix à la consommation) ;

CT : la consommation des ménages à prix constants ;

POP : la population ;

GNP : le produit national brut à prix courants ;

CPI : l'indice des prix à la consommation toutes catégories ;

CPI-FB : l'indice des prix à la consommation pour la catégorie « food and beverages » (aliments et boissons).

Ces données se trouvent dans la feuille 1 du fichier limonade.xls téléchargeable sur le site Internet www.pearsoneducation.fr.

Résolvez ce problème d'une part en utilisant un tableur, d'autre part en utilisant TSP.

Solution

Méthodologie

Procédez comme suit :

1. Spécifiez un modèle explicatif des ventes de la boisson au cola en fonction de certaines variables explicatives.
2. Estimez les coefficients « vrais inconnus » de ce modèle sur les données disponibles de 1970 à 1980.
3. Posez des hypothèses sur les valeurs des variables explicatives de 1981 à 1990. Conditionnellement à ces valeurs, utilisez le modèle pour calculer les valeurs des ventes de la boisson au cola pour les années 1981 à 1990.

Spécification du modèle

Supposez que les ventes de cette marque de boisson au cola se déterminent de la manière suivante : les quantités vendues par habitant sont fonction du revenu réel des ménages par habitant et du prix relatif de cette boisson par rapport au prix moyen de l'ensemble des biens et des services de consommation. Cette fonction est vraisemblablement linéaire d'un point de vue logarithmique :

$$\ln\left(\frac{SALES_t}{POP_t PRICE_t}\right) = \beta_1 + \beta_2 \ln\left(\frac{Y_t}{POP_t}\right) + \beta_3 \ln\left(\frac{PRICE_t}{CPI_t}\right) + u_t$$

Que se passe-t-il si vous n'utilisez pas le logarithme de ces variables ? Dans ce cas, vous supposez que :

$$\frac{SALES_t}{POP_t PRICE_t} = \beta_1 + \beta_2 \frac{Y_t}{POP_t} + \beta_3 \frac{PRICE_t}{CPI_t} + u_t$$

Le modèle implique alors qu'à prix relatif inchangé, une augmentation de 1 dollar du revenu entraîne une augmentation de $\beta_{2\,\text{dollars}}$ de la consommation de la boisson, quel que soit le montant de départ du revenu. En d'autres termes, si un riche et un pauvre reçoivent un surplus de revenu de 1 000 dollars, ils vont tous deux augmenter leur consommation de ladite boisson d'un même montant. Cela n'a évidemment aucun sens !

Le modèle en logarithmes peut encore être présenté de la manière suivante :

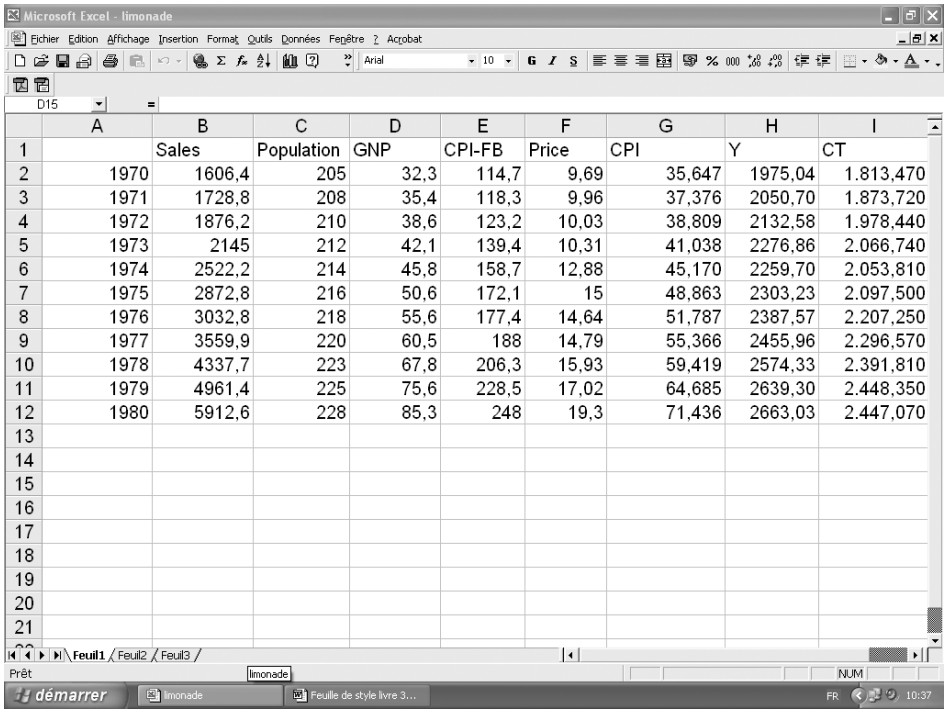
$$LRS_t = \beta_1 + \beta_2 LYP_t + \beta_3 LRPR_t + u_t$$

où $LRS_t = \ln\left(\frac{SALES_t}{POP_t PRICE_t}\right)$, $LY_t = \ln\left(\frac{Y_t}{POP_t}\right)$, $LRPR_t = \ln\left(\frac{PRICE_t}{CPI_t}\right)$.

Estimation du modèle avec un tableur

Le fichier de données se présente comme à la figure 3.7.

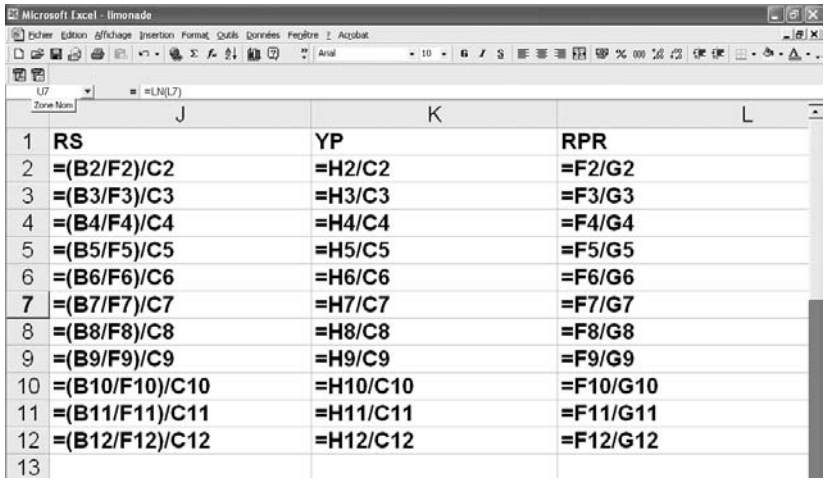
Figure 3.7



	A	B	C	D	E	F	G	H	I
1		Sales	Population	GNP	CPI-FB	Price	CPI	Y	CT
2	1970	1606,4	205	32,3	114,7	9,69	35,647	1975,04	1.813,470
3	1971	1728,8	208	35,4	118,3	9,96	37,376	2050,70	1.873,720
4	1972	1876,2	210	38,6	123,2	10,03	38,809	2132,58	1.978,440
5	1973	2145	212	42,1	139,4	10,31	41,038	2276,86	2.066,740
6	1974	2522,2	214	45,8	158,7	12,88	45,170	2259,70	2.053,810
7	1975	2872,8	216	50,6	172,1	15	48,863	2303,23	2.097,500
8	1976	3032,8	218	55,6	177,4	14,64	51,787	2387,57	2.207,250
9	1977	3559,9	220	60,5	188	14,79	55,366	2455,96	2.296,570
10	1978	4337,7	223	67,8	206,3	15,93	59,419	2574,33	2.391,810
11	1979	4961,4	225	75,6	228,5	17,02	64,685	2639,30	2.448,350
12	1980	5912,6	228	85,3	248	19,3	71,436	2663,03	2.447,070
13									
14									
15									
16									
17									
18									
19									
20									
21									

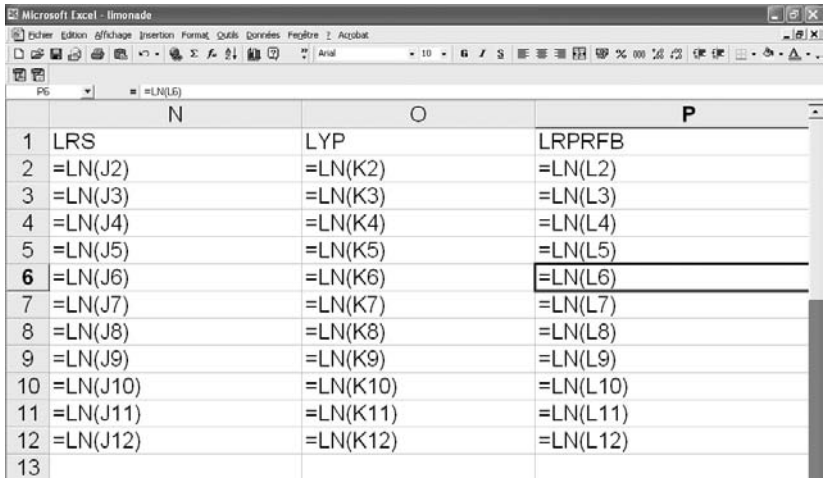
Commencez par créer dans les colonnes J, K et L les variables $RS_t = \frac{SALES_t}{POP_t PRICE_t}$, $Y_t = \frac{Y_t}{POP_t}$, $RPR_t = \left(\frac{PRICE_t}{CPI_t}\right)$. Les formules se présentent comme à la figure 3.8.

Figure 3.8



Puis calculez les valeurs de LRS, LYP et LRPR dans les colonnes M, N et O (voir figure 3.9).

Figure 3.9



Pour estimer les coefficients du modèle par MCO, sélectionnez successivement Outils/Utilitaire d'analyse/Régression linéaire et cliquez sur OK. Remplissez certaines cases de l'écran de régression à la figure 3.10, page suivante.

Cliquez sur OK. Les résultats sont identiques à ceux de la figure 3.11, page suivante.

Le logiciel a attribué à la feuille de résultats le nom arbitraire que vous avez choisi de lui donner : OLS. Vous obtenez des valeurs estimées pour les coefficients « vrais inconnus » β_1, β_2 et β_3 : $\hat{\beta}_1^{MCO} = -6{,}84817$, $\hat{\beta}_2^{MCO} = 2{,}492627$, $\hat{\beta}_3^{MCO} = -0{,}71303$.

L'élasticité des quantités vendues par habitant au revenu disponible réel par habitant est mesurée par β_2 et sa valeur estimée vaut 2,492627 : par conséquent, si le revenu augmente

Figure 3.10

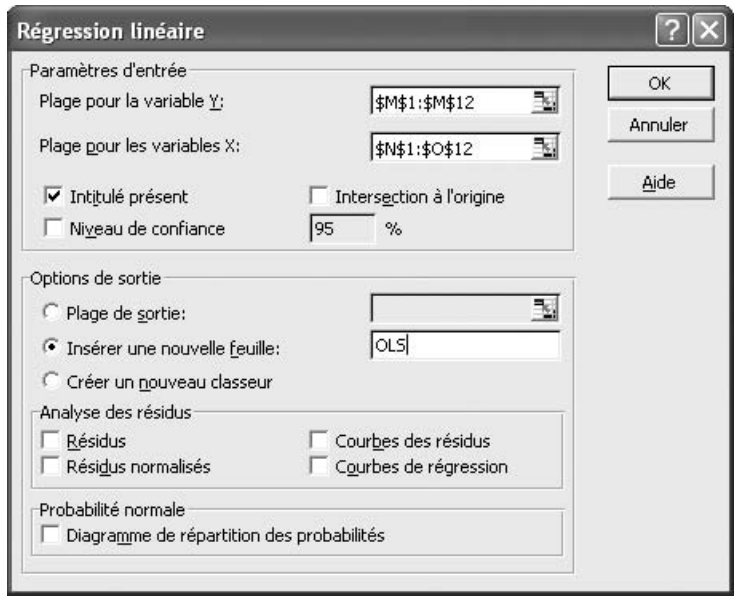


Figure 3.11

Microsoft Excel - limonade

Fichier Edition Affichage Insertion Format Outils Données Fenêtre ? Acrobat

N29

	A	B	C	D	E	F
1	RAPPORT DÉTAILLÉ					
2						
3	*Statistiques de la régression*					
4	Coefficient de	0,96411899				
5	Coefficient de	0,92952543				
6	Coefficient de	0,91190678				
7	Erreur-type	0,05297086				
8	Observations	11				
9						
10	ANALYSE DE VARIANCE					
11		*Degré de libert*	*mme des car*	*yenne des car*	*F*	*Valeur critique de F*
12	Régression	2	0,2960689	0,14803445	52,7580592	2,46678E-05
13	Résidus	8	0,02244729	0,00280591		
14	Total	10	0,31851619			
15						
16		*Coefficients*	*Erreur-type*	*Statistique t*	*Probabilité*	*Limite inférieure pour seuil de confiance = 95%*
17	Constante	-6,84817203	0,70792782	-9,67354555	1,0869E-05	-8,480657568
18	LYP	2,49262665	0,25147946	9,91184998	9,0679E-06	1,912713611
19	LRPR	-0,71303432	0,30943693	-2,30429616	0,05013349	-1,42659761
20						
21						

de 1 %, les quantités en question augmentent de 2,492627 %. L'élasticité des quantités vendues de boisson au cola par habitant à son prix relatif est mesurée par β_3 et sa valeur estimée vaut −0,71303 : par conséquent, si le prix augmente de 1 %, les quantités en question diminuent de 0,71303 %. R^2 est égal à 0,929525 : en d'autres termes, 92,9525 % des fluctuations des quantités vendues en logarithmes sont expliquées par les fluctuations du revenu disponible réel en logarithme et du prix relatif en logarithme, sur la période 1970–1980.

L'écart type estimé de l'estimateur $\hat{\beta}_1^{MCO}$ est égal à 0,707928, celui de $\hat{\beta}_2^{MCO}$ à 0,251479 et celui de $\hat{\beta}_3^{MCO}$ à 0,309437. Plus l'écart type de l'estimateur d'un coefficient est élevé, plus grande est l'imprécision avec laquelle le coefficient est estimé. Pour évaluer si l'écart type

estimé est grand ou petit, comparez-le à la valeur estimée du coefficient. La colonne *t* Stat permet cette comparaison puisque *t*-stat = coefficient estimé/écart type estimé. Donc, plus une *t*-stat est élevée, plus la précision de l'estimation du coefficient concerné est grande. De manière générale, à partir de 2, la précision est bonne et, en dessous de 1, elle est très insuffisante. Les écarts types estimés par les logiciels ne sont fiables que si le terme d'erreur n'est ni autocorrélé ni hétéroscédastique car ils sont calculés à l'aide d'une formule qui n'est correcte que si le terme d'erreur est un bruit blanc. Par ailleurs, en supposant que ce dernier est distribué normalement, et en l'absence d'autocorrélation et d'hétéroscédasticité, vous ne pouvez rejeter l'hypothèse qu'un coefficient « vrai inconnu » est nul si sa *t*-stat est inférieure à la valeur critique à 5 % d'une distribution Student à 8 = 11(observations) − 3(coefficients) degrés de liberté. Dans ce cas, la probabilité critique (*p*-value en anglais) fournit le résultat puisque vous ne rejetez pas l'hypothèse qu'un coefficient « vrai inconnu » est nul si la probabilité critique est supérieure à 0,05. Mais comment savoir si le terme d'erreur n'est ni autocorrélé, ni hétéroscédastique, et s'il est distribué normalement ? Utilisez les tests d'absence d'autocorrélation, d'absence d'hétéroscédasticité et de normalité du terme d'erreur (voir plus loin).

Extrapolation avec un tableur

Vous allez réaliser une extrapolation des valeurs des quantités vendues par l'entreprise de boisson au cola pour les années 1981 à 1990, en utilisant Excel.

Pour cela, il faut d'abord poser les hypothèses suivantes sur l'évolution des variables explicatives, de 1981 à 1990 :

- L'entreprise envisage d'augmenter son prix de vente (PRICE) de 10 % chaque année.
- L'indice des prix à la consommation (CPI) va augmenter de 8 % annuellement. Ces hypothèses supposent que le prix relatif (RPR) augmente de 2 % par an.
- Le revenu disponible réel par habitant (YP) va augmenter de 2 % chaque année.
- La population va augmenter de 1,5 % chaque année.

Dans le fichier limonade.xls, affichez la feuille initiale (appelée ici limonade). Prolongez les dates de la colonne A, inscrivez les valeurs supposées de YP, PRICE, CPI pour les années 1981 à 1990 en colonnes G, G et K, et calculez leurs implications sur RPR en colonne L, à partir de la ligne 15. Calculez les valeurs supposées de LYP et LRPR de 1981 à 1990, en colonnes N et O (voir figures 3.12, page suivante et 3.13, page suivante).

Maintenant que les hypothèses sur les variables explicatives, de 1981 à 1990, sont posées, il est possible d'utiliser le modèle estimé pour calculer les valeurs de LRS qui en résultent, pour 1981 à 1990. Les résultats de l'estimation du modèle sont dans la feuille OLS du fichier limonade.xls. En particulier, les coefficients estimés sont donnés par les cellules suivantes de cette feuille : $\hat{\beta}_1^{MCO}$ = OLS!$B$17, $\hat{\beta}_2^{MCO}$ = OLS!$B$18, $\hat{\beta}_3^{MCO}$ = OLS!$B$19. Il faut calculer les prévisions de LRS à l'aide des formules suivantes :

$$LRS_{1981}^{p} = \hat{\beta}_1^{MCO} + \hat{\beta}_2^{MCO} LYP_{1981}^{hyp} + \hat{\beta}_3^{MCO} LRPR_{1981}^{hyp}$$
$$LRS_{1982}^{p} = \hat{\beta}_1^{MCO} + \hat{\beta}_2^{MCO} LYP_{1982}^{hyp} + \hat{\beta}_3^{MCO} LRPR_{1982}^{hyp}$$
$$\cdots$$
$$LRS_{1990}^{p} = \hat{\beta}_1^{MCO} + \hat{\beta}_2^{MCO} LYP_{1990}^{hyp} + \hat{\beta}_3^{MCO} LRPR_{1990}^{hyp}$$

Faites les modifications qui s'imposent dans la feuille limonade du fichier limonade.xls (voir figure 3.14, page 103).

Figure 3.12

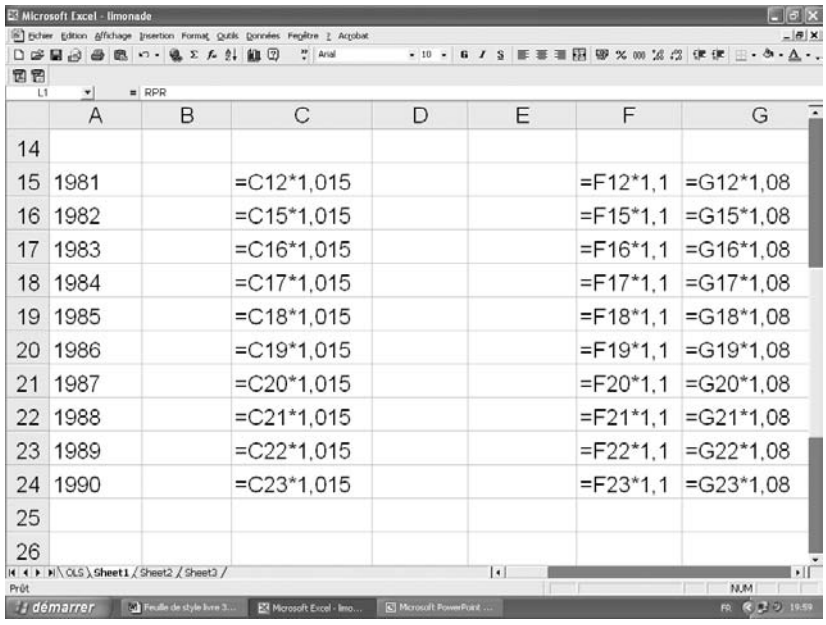
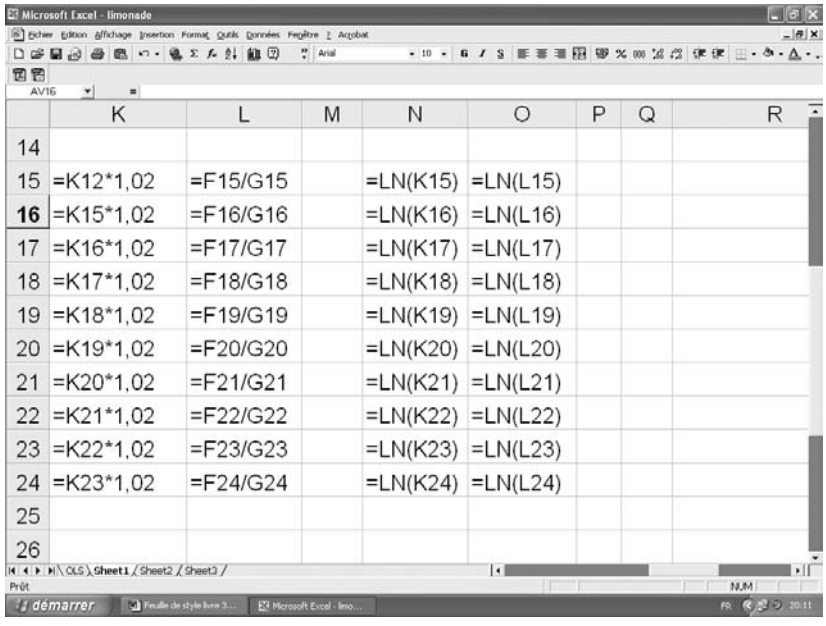


Figure 3.13



Il reste à transformer les prévisions de LRS (les cellules M15 à M24) en prévisions pour les ventes nominales totales (SALES) de la manière suivante :

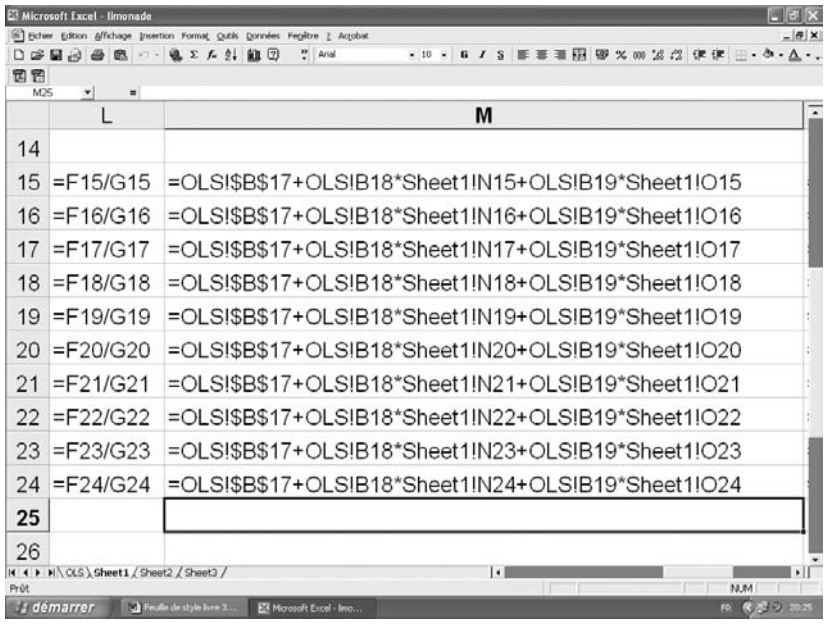
$$SALES^{p}_{1981} = POP_{1981} * PRICE_{1981} * \left(e^{LRS^{p}_{1981}}\right)$$

$$SALES^{p}_{1982} = POP_{1982} * PRICE_{1982} * \left(e^{LRS^{p}_{1982}}\right)$$

$$\ldots$$

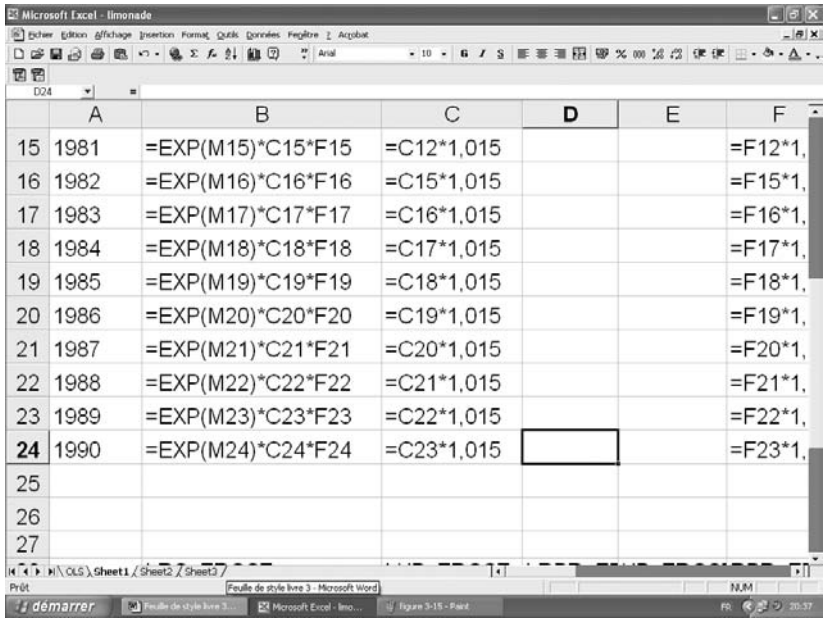
$$SALES^{p}_{1990} = POP_{1990} * PRICE_{1990} * \left(e^{LRS^{p}_{1990}}\right)$$

Figure 3.14



Pour cela, il faut introduire pour commencer les formules de calcul des valeurs supposées de la population de 1981 à 1990, puis insérer les formules de calcul des prévisions de SALES (voir figure 3.15).

Figure 3.15



Inscrivez dans une cellule surmontant les valeurs extrapolées de SALES la mention SALES-FRCST pour les désigner.

Vous obtenez les valeurs suivantes :

```
6294,69848
7287,66761
```

```
8437,27454
9768,22839
11309,1361
13093,1172
15158,5157
17549,7245
20318,139
23523,2623
```

Estimation et extrapolation avec TSP

Avec TSP, le principe et la procédure sont différents : créez un nouveau fichier Excel, dans lequel vous recopiez certaines données historiques du fichier initial limonade.XLS, en l'occurrence seulement celles dont vous avez besoin dans l'analyse : les dates, les données de SALES, Population, PRICE, CPI et Y. Sont donc exclues les données de GNP, CPI-FB et CT (vous pourriez insérer ces variables dans le fichier, même sans les utiliser par la suite, mais il faudrait alors changer le nom CPI-FB en CPIFB : TSP n'accepte pas les noms de variables contenant des tirets).

Ajoutez l'en-tête DATE au-dessus de la colonne de dates. Changez le nom Population en POP parce que certaines versions anciennes de TSP n'acceptent que des noms de huit caractères maximum. Enregistrez le document comme un fichier de type feuille de calcul Excel (quelle que soit la version), sous le nom LIM (l'ordinateur lui donne automatiquement l'extension .xls), dans le répertoire C:\ de l'ordinateur par exemple. Fermez-le pour pouvoir le lire avec TSP. Pour travailler sous TSP, suivez la procédure décrite au chapitre 2. Ici le programme d'instructions est le suivant :

```
FREQ A;
SMPL 1970 1980;
? LECTURE DES DONNEES
READ(FILE='C :\LIM.XLS');
? CREATION DE VARIABLES
RS=(SALES/POP)/PRICE;
YP=Y/POP;
RPR=PRICE/CPI;
LRS=LOG(RS);
LYP=LOG(YP);
LRPR=LOG(RPR);
? CALCUL DE REGRESSION
REGOPT(PVPRINT,STARS) ALL;
NOPLOT;
OLSQ LRS C LYP LRPR;
? HYPOTHESES EN PREVISION
SMPL 1981 1990;
YP=YP(-1)*1.02;
POP=POP(-1)*1.015;
CPI=CPI(-1)*1.08;
PRICE=PRICE(-1)*1.10;
RPR=PRICE/CPI;
```

```
LYP=LOG(YP);
LRPR=LOG(RPR);
? CALCUL DE PREVISION EN LOG SUR LES VENTES REELLES PAR HABITANT
FORCST(PRINT) LRSP;
? CALCUL DE LA PREVISION SUR LES VENTES NOMINALES TOTALES
SALESP=EXP(LRSP)*POP*PRICE;
PRINT SALESP;
```

Les lignes de code commençant par un point d'interrogation (?) sont des commentaires qui décrivent le programme (vous pouvez donc les ignorer lorsque vous réalisez cet exercice). Les résultats sont les suivants :

```
 Mean of dependent variable = .510351E-02
Std. dev. of dependent var. = .178470
   Sum of squared residuals = .022447
      Variance of residuals = .280592E-02
   Std. error of regression = .052971
                 R-squared = .929525
        Adjusted R-squared = .911907
    Durbin-Watson statistic = .961980
Breusch/Godfrey LM : AR/MA1 = 1.97215 [.160]
Ljung-Box Q-statistic1 = 1.12998 [.288]
  Wald nonlin. AR1 vs. lags = 5.16845 [.075]
                 ARCH test = 1.66231 [.197]
               CuSum test = .770092 [.166]
             CuSumSq test = .487032 * [.037]
                Chow test = 8.68049 * [.020]
LR het.  test (w/ Chow) = 26.9613 ** [.000]
White het. test = 1.11500 [.953]
Jarque-Bera normality test = .118607 [.942]
  F-statistic (zero slopes) = 52.7580 ** [.000]
  Akaike Information Crit.  = -2.81115
 Schwarz Bayes. Info. Crit. = -5.54051
 Log of likelihood function = 18.4613
          Estimated    Standard
Variable  Coefficient     Error      t-statistic     P-value
C         -6.84817       .707928     -9.67354     ** [.000]
LYP       2.49263        .251480     9.91184      ** [.000]
LRPR      -.713034       .309437     -2.30429        [.050]
              SALESP
1981       6294.69873
1982       7287.66650
1983       8437.27344
1984       9768.23047
1985      11309.13770
1986      13093.11816
1987      15158.51367
1988      17549.73047
1989      20318.14258
1990      23523.26367
```

Vous retrouvez certains résultats déjà obtenus précédemment, mais des tests permettent de mieux apprécier leur validité. Deux des trois tests d'hétéroscédasticité ont une probabilité critique supérieure à 0,05 et ne rejettent donc pas l'hypothèse que le terme d'erreur n'est pas hétéroscédastique. Il s'agit des tests ARCH (qui est peu pertinent ici sur des données annuelles) et du test White het. Les tests de Breusch et Godfrey, de LM : AR/MA1 et de Ljung et Box (Q-statistic1) ont une probabilité critique supérieure à 0,05 et ne rejettent donc pas l'hypothèse que le terme d'erreur n'est pas autocorrélé. Le terme d'erreur semble être un bruit blanc, ce qui permet d'affirmer que les écarts types estimés de la colonne standard errors sont fiables. Les *t*-stats élevées montrent que les coefficients sont estimés d'une manière précise.

Exercice 3 Le cas Banque Régionale Française

Énoncé

Une agence urbaine d'une grande banque régionale française [1] souhaite, en octobre 2001, mettre au point une méthode de prévision de ses volumes mensuels de production de crédits octroyés sous forme de prêts personnels. Cette agence est située dans le centre d'une grande métropole. Les données disponibles en provenance de cette banque sont :

- une série temporelle mensuelle avec le volume observé de production de crédits de l'agence Lille Métropole, de janvier 1999 à septembre 2001 (35 mois) ;
- des séries temporelles mensuelles de l'encours, des rachats et des échus de crédits de cette agence pour les mêmes mois ;
- des séries temporelles mensuelles de la marge sur les crédits produits et de la marge sur les crédits en cours de l'agence pour les mêmes mois.

Ces données sont sur la première feuille du fichier BA.XLS :

- en colonne A, la variable date : les dates successives de novembre 1998 à septembre 2001 ;
- en colonne B, la variable pp_pro : la production de prêts personnels ;
- en colonne C, la variable pp_enc : l'encours des prêts personnels ;
- en colonne D, la variable pp_ra : les rachats de prêts personnels ;
- en colonne E, la variable pp_ec : les échus de prêts personnels ;
- en colonne F, la variable pp_pro_m : la marge sur la production des prêts personnels ;
- en colonne G, la variable pp_enc_m : la marge sur l'encours des prêts personnels.

Pour chaque colonne, le nom de la variable est en ligne 1 ; les trente-cinq observations successives viennent sur les lignes suivantes.

Solution

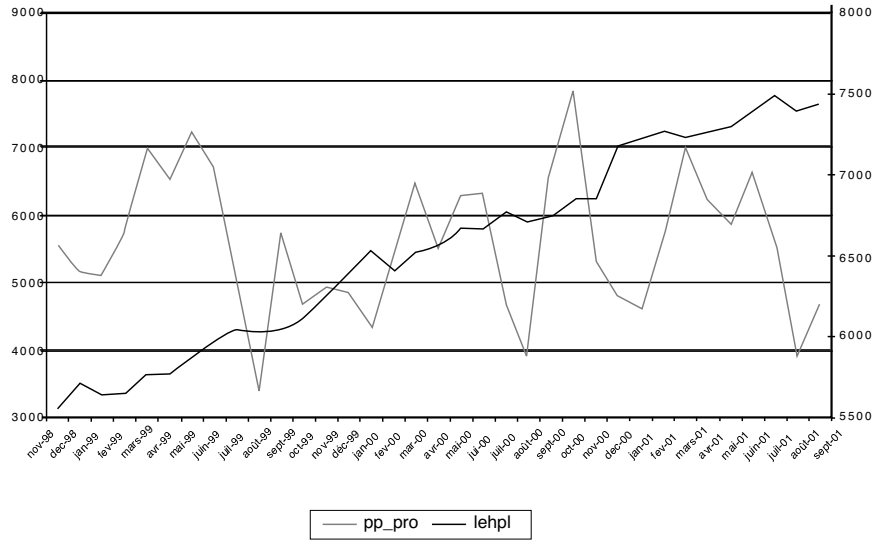
Exploration d'une première piste

Pour élaborer une méthode de prévision de la production de crédits prêts personnels pour l'agence bancaire, il faut d'abord comprendre les facteurs qui peuvent expliquer l'évolution de cette production.

1. Il s'agit d'une demande réelle traitée à l'IESEG pour le compte d'une agence d'une grande banque régionale française. Celle-ci ayant souhaité la confidentialité, son identité ne sera pas divulguée ici. La méthodologie de prévision devait être mise au point pour trois catégories différentes de crédits : les prêts acquéreurs, qui sont des prêts hypothécaires, les prêts personnels et les prêts équipements MLT (Moyen Long Terme), qui sont des prêts accordés à des entreprises et à des artisans pour l'achat de biens d'investissement.

Si les productions de crédits prêts personnels de l'agence et de l'ensemble des banques de France semblent suivre globalement la même évolution, vous pouvez expliquer la production de crédits de l'agence par les mêmes facteurs que ceux qui influencent la production totale de crédits en France. Il s'agit de facteurs nationaux globaux relativement bien connus, faisant l'objet de statistiques publiées par des organismes officiels comme l'INSEE, le FMI, l'OCDE. Pour vérifier que la production de crédits de l'agence concernée suit la tendance nationale, vous allez comparer graphiquement son évolution à celle de la production de crédits prêts personnels de l'ensemble des banques de France. Cette variable nationale figure dans les statistiques financières internationales du FMI : elle s'intitule FR BANK LENDING TO HOUSEHOLDS AS PERSONAL LOANS CUR. Ajoutez ces données à la colonne H du fichier CN.XLS, avec comme nom l'abréviation lehpl. Comparez alors la production de prêts personnels de l'agence avec la production nationale, représentées sur un même graphique (voir figure 3.16).

Figure 3.16



La production de crédits en prêts personnels de l'agence ne suit pas la tendance nationale.

Remarque

Les divergences entre évolution de l'agence et évolution nationale ne viennent pas de la différence entre le caractère lisse de la série nationale et le caractère erratique de la série de l'agence, qui s'explique par le fait que la série nationale est « corrigée des variations saisonnières », ce qui n'est pas le cas de la série de l'agence. Elles sont dues à la tendance croissante affichée par la série nationale (et non par la série de l'agence) tout au long de la période.

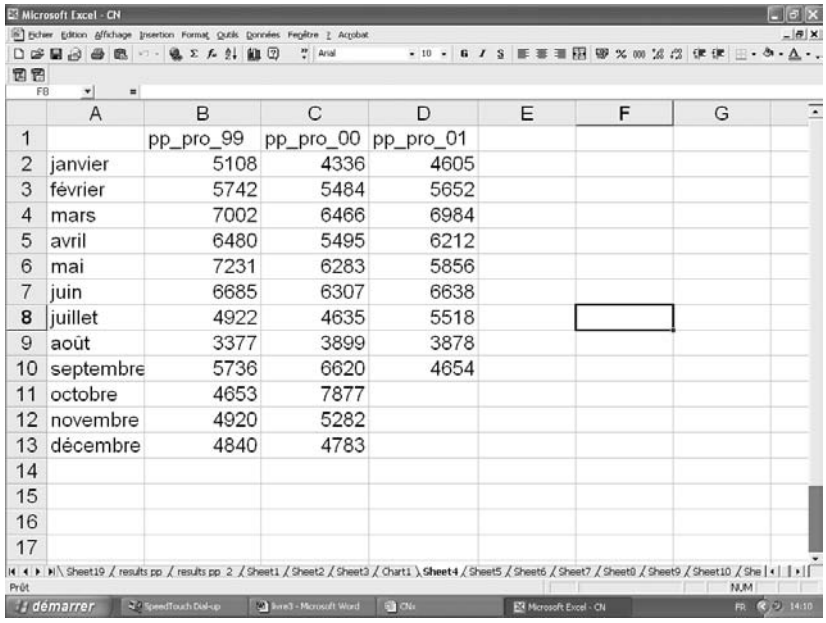
Les conclusions de l'examen graphique sont les suivantes :

1. L'évolution mensuelle des crédits de l'agence semble essentiellement due à des facteurs locaux plutôt que nationaux, ou à des facteurs spécifiques à l'entreprise et affectant sa part de marché dans la région.
2. L'évolution des crédits de l'agence semble présenter des fluctuations saisonnières.

Fluctuations saisonnières

Pour vérifier la présence de fluctuations saisonnières dans la série de production de crédits de l'agence bancaire, vous allez réaliser une nouvelle analyse graphique. Recopiez les données de la série pp_pro dans une nouvelle feuille du fichier BA, en la subdivisant en sous-variables correspondant à chaque année (voir figure 3.17).

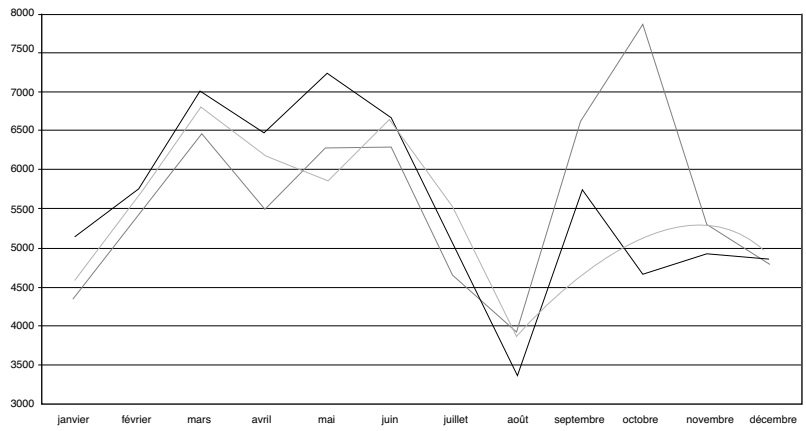
Figure 3.17



	A	B	C	D	E	F	G
1		pp_pro_99	pp_pro_00	pp_pro_01			
2	janvier	5108	4336	4605			
3	février	5742	5484	5652			
4	mars	7002	6466	6984			
5	avril	6480	5495	6212			
6	mai	7231	6283	5856			
7	juin	6685	6307	6638			
8	juillet	4922	4635	5518			
9	août	3377	3899	3878			
10	septembre	5736	6620	4654			
11	octobre	4653	7877				
12	novembre	4920	5282				
13	décembre	4840	4783				
14							
15							
16							
17							

Représentez ces sous-séries sous forme graphique en fonction de la première colonne, à savoir le mois (voir figure 3.18).

Figure 3.18



Un profil saisonnier systématique se répète clairement chaque année.

Modélisation et extrapolation

Sur la base des observations graphiques et d'un raisonnement économique élémentaire, spécifiez un modèle où l'évolution de pp_pro est expliquée par :

- des fluctuations saisonnières représentées par des variables dummy saisonnières ;
- le taux de marge, appliqué par la banque sur le coût de ses fonds et utilisé pour fixer le taux d'intérêt des crédits ;
- la valeur de pp_pro au mois précédent.

Le coefficient du taux de marge devrait être *a priori* négatif : si la banque réduit sa marge, et est donc prête à proposer des taux plus bas, elle peut attirer davantage d'emprunteurs. Vous devez d'abord préparer les données pour pouvoir estimer le modèle. À la suite des colonnes déjà remplies dans la feuille de données, recopiez en colonne K les dates à partir de décembre 1998, en commençant à la ligne 2. Puis, pour chaque date, complétez le tableau de la façon suivante :

- Recopiez la variable dépendante pp_pro en colonne L.
- Écrivez les onze variables dummy saisonnières, appelées $D1, D2 \ldots D11$, en colonnes M à W.
- Recopiez la variable pp_pro_m en colonne X.
- Recopiez la variable dépendante retardée d'une période – appelez-la pp_pro(−1) – en colonne Y.

Les variables dummy saisonnières sont définies de la manière suivante :

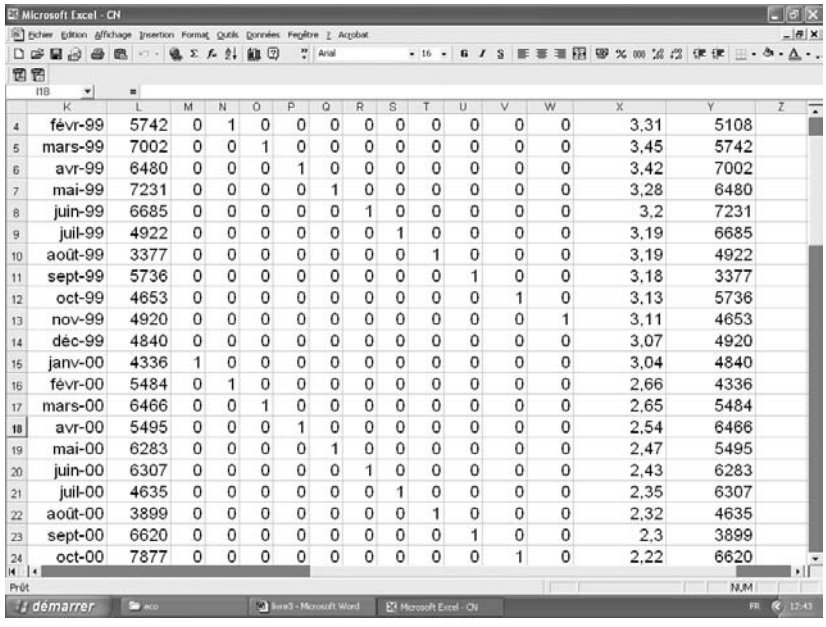
$D1$ vaut 1 pour janvier et 0 pour tous les autres mois.

$D2$ vaut 1 pour février et 0 pour tous les autres mois.

. . .

$D11$ vaut 1 pour novembre et 0 pour tous les autres mois.

Figure 3.19



	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y
4	févr-99	5742	0	1	0	0	0	0	0	0	0	0	0	3,31	5108
5	mars-99	7002	0	0	1	0	0	0	0	0	0	0	0	3,45	5742
6	avr-99	6480	0	0	0	1	0	0	0	0	0	0	0	3,42	7002
7	mai-99	7231	0	0	0	0	1	0	0	0	0	0	0	3,28	6480
8	juin-99	6685	0	0	0	0	0	1	0	0	0	0	0	3,2	7231
9	juil-99	4922	0	0	0	0	0	0	1	0	0	0	0	3,19	6685
10	août-99	3377	0	0	0	0	0	0	0	1	0	0	0	3,19	4922
11	sept-99	5736	0	0	0	0	0	0	0	0	1	0	0	3,18	3377
12	oct-99	4653	0	0	0	0	0	0	0	0	0	1	0	3,13	5736
13	nov-99	4920	0	0	0	0	0	0	0	0	0	0	1	3,11	4653
14	déc-99	4840	0	0	0	0	0	0	0	0	0	0	0	3,07	4920
15	janv-00	4336	1	0	0	0	0	0	0	0	0	0	0	3,04	4840
16	févr-00	5484	0	1	0	0	0	0	0	0	0	0	0	2,66	4336
17	mars-00	6466	0	0	1	0	0	0	0	0	0	0	0	2,65	5484
18	avr-00	5495	0	0	0	1	0	0	0	0	0	0	0	2,54	6466
19	mai-00	6283	0	0	0	0	1	0	0	0	0	0	0	2,47	5495
20	juin-00	6307	0	0	0	0	0	1	0	0	0	0	0	2,43	6283
21	juil-00	4635	0	0	0	0	0	0	1	0	0	0	0	2,35	6307
22	août-00	3899	0	0	0	0	0	0	0	1	0	0	0	2,32	4635
23	sept-00	6620	0	0	0	0	0	0	0	0	1	0	0	2,3	3899
24	oct-00	7877	0	0	0	0	0	0	0	0	0	1	0	2,22	6620

Commencer à la colonne K est un choix arbitraire. Vous pourriez commencer à la colonne I (voir figure 3.19).

Vous pouvez maintenant estimer le modèle. Cliquez successivement sur Outils (Tools)/Utilitaire d'analyse (Data analysis)/Régression linéaire (Regression) et remplissez

le tableau. Dans Plage pour la variable *Y*, écrivez $L$1:$L$35 pour définir les cellules des observations de la variable dépendante. Dans Plage pour les variables *X*, entrez $M$1:$Y$35 pour définir les cellules des observations des variables explicatives. Cochez Intitulé présent. Cochez Insérer une nouvelle feuille et écrivez results_pp. Cliquez sur OK. Les résultats s'affichent dans la feuille results_pp.

Puisque $R^2 = 0{,}765$, 76,5 % de la variabilité observée de pp_pro est expliquée par la variabilité des variables explicatives. Le coefficient estimé de la variable explicative vaut −137 ; il est donc négatif comme prévu : si la banque augmente ses marges, elle décourage des candidats emprunteurs. Toutefois, l'imprécision de l'estimation de ce coefficient, donnée par son écart type (standard errors) est énorme. Sa *t*-stat est donc petite et la probabilité critique (*p*-value) de sa *t*-stat est supérieure à 0,05 : vous ne pouvez rejeter l'hypothèse que le coefficient vrai de ll_pro_m soit nul. Pour autant, il n'est pas sûr qu'il soit nul : cela signifie simplement que 0 fait partie de l'ensemble de valeurs qui ne sont pas rejetées par le test (les valeurs d'un intervalle de confiance à 90 %). L'interprétation correcte du test est que le coefficient de ll_pro_m est compris dans un certain intervalle de confiance, avec une probabilité de 0,9. Toutefois, cet intervalle est trop large, si bien que vous ne pouvez montrer statistiquement si ce coefficient est négatif, nul ou positif.

Il peut alors être légitime de supprimer pp_pro_m des variables explicatives et d'estimer de nouveau le modèle. Supprimez donc la colonne X dans la feuille de données : la variable pp_pro(−1) vient occuper la colonne X. Cliquez sur Outils puis sur Régression linéaire. Dans Plage pour la variable *Y*, écrivez $L$1:$L$35 pour définir les cellules des observations de la variable dépendante. Dans Plage pour les variables *X*, écrivez $M$1:$X$35 pour définir les cellules des observations des variables explicatives. Cochez Intitulé présent puis Insérer une nouvelle feuille, et écrivez results_pp2. Cliquez sur OK. Les résultats s'affichent dans la feuille results_pp2.

Ensuite, utilisez ce modèle estimé pour prévoir la production de crédits prêts personnels d'octobre 2001 à décembre 2002. Dans la feuille de données, prolongez les dates à la colonne K et les variables dummy aux colonnes M à W, pour les lignes 36 à 50. Recopiez aussi les coefficients estimés de la feuille results_pp2 dans la feuille de données, aux cellules J36 à J48. Calculez alors la valeur de pp_pro en utilisant les coefficients estimés du modèle. En ce sens, dans la cellule L36, écrivez la formule pour pp_pro en octobre 2001 :

```
=$J$36+$J$37*M36+$J$38*N36+$J$39*O36+$J$40*P36+$J$41*Q36+$J$42*R36
+$J$43*S36+$J$44*T36+$J$45*U36+$J$46*V36+$J$47*W36+$J$48*L35
```

Dans la cellule L37, écrivez la formule pour pp_pro en novembre 2001 :

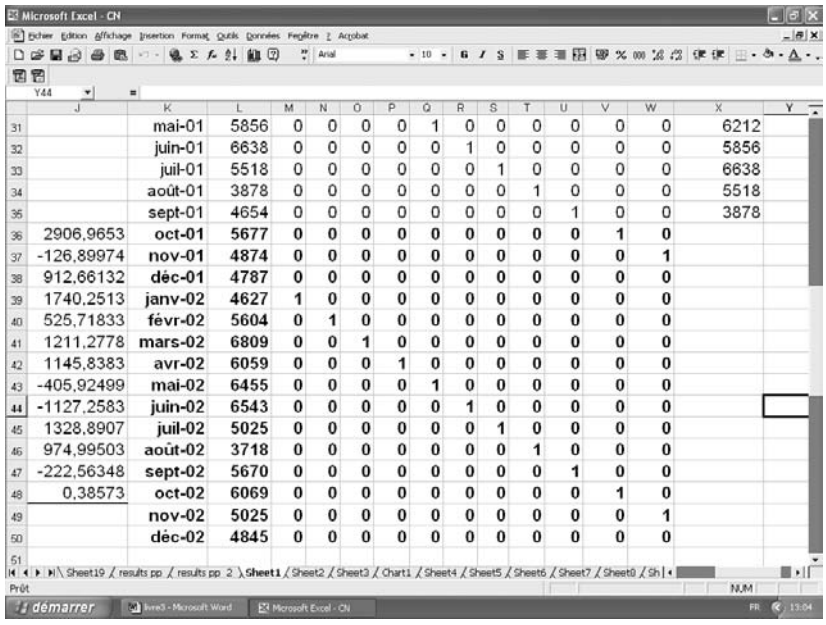
```
=$J$36+$J$37*M37+$J$38*N37+$J$39*O37+$J$40*P37+$J$41*Q37+$J$42*R37
+$J$43*S37+$J$44*T37+$J$45*U37+$J$46*V37+$J$47*W37+$J$48*L36
```

Réitérez jusqu'à la cellule L50, dans laquelle vous écrivez la formule pour pp_pro en décembre 2002 :

```
=$J$36+$J$37*M50+$J$38*N50+$J$39*O50+$J$40*P50+$J$41*Q50+$J$42*R50
 +$J$43*S50+$J$44*T50+$J$45*U50+$J$46*V50+$J$47*W50+$J$48*L49
```

Vous obtenez des valeurs extrapolées pour la production de prêts personnels (voir figure 3.20, page ci-contre).

Figure 3.20



	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
31		mai-01	5856	0	0	0	0	1	0	0	0	0	0	0	6212
32		juin-01	6638	0	0	0	0	0	1	0	0	0	0	0	5856
33		juil-01	5518	0	0	0	0	0	0	1	0	0	0	0	6638
34		août-01	3878	0	0	0	0	0	0	0	1	0	0	0	5518
35		sept-01	4654	0	0	0	0	0	0	0	0	1	0	0	3878
36	2906,9653	oct-01	5677	0	0	0	0	0	0	0	0	0	1	0	
37	-126,89974	nov-01	4874	0	0	0	0	0	0	0	0	0	0	1	
38	912,66132	déc-01	4787	0	0	0	0	0	0	0	0	0	0	0	
39	1740,2513	janv-02	4627	1	0	0	0	0	0	0	0	0	0	0	
40	525,71833	févr-02	5604	0	1	0	0	0	0	0	0	0	0	0	
41	1211,2778	mars-02	6809	0	0	1	0	0	0	0	0	0	0	0	
42	1145,8383	avr-02	6059	0	0	0	1	0	0	0	0	0	0	0	
43	-405,92499	mai-02	6455	0	0	0	0	1	0	0	0	0	0	0	
44	-1127,2583	juin-02	6543	0	0	0	0	0	1	0	0	0	0	0	
45	1328,8907	juil-02	5025	0	0	0	0	0	0	1	0	0	0	0	
46	974,99503	août-02	3718	0	0	0	0	0	0	0	1	0	0	0	
47	-222,56348	sept-02	5670	0	0	0	0	0	0	0	0	1	0	0	
48	0,38573	oct-02	6069	0	0	0	0	0	0	0	0	0	1	0	
49		nov-02	5025	0	0	0	0	0	0	0	0	0	0	1	
50		déc-02	4845	0	0	0	0	0	0	0	0	0	0	0	

EXERCICE 4 LE CAS PRODUCTEURS D'ÉLECTRICITÉ

Énoncé

Des investisseurs privés envisagent de financer une nouvelle usine de production d'électricité, d'une certaine capacité de production de kilowatts à l'heure. Avant de se décider et de fixer la taille de l'usine, ils souhaitent avoir la réponse à plusieurs questions, dont celles-ci : comment évoluent le coût total d'une centrale électrique en fonction de la quantité produite ? En d'autres termes, de quelle nature sont les rendements ? Quel est l'impact d'une augmentation de 1 % du prix du fuel sur le coût total, après une réorganisation optimale suite à cette augmentation ?

Comme informations, ils disposent en particulier de la production et du coût total de production des usines électriques déjà installées. Il s'agir des données de cent vingt-trois sociétés productrices d'électricité américaines rassemblées en 1970 et utilisées par L. Christensen et W. Greene [CHR 1984]. Elles se trouvent dans le fichier EL.xls téléchargeable sur le site Internet www.pearsoneducation.fr. Les variables sont les suivantes :

OUTPUT : la production de la centrale électrique ;

PRICELAB : le prix d'une unité de travail ;

PRICECAP : le prix d'une unité de capital physique ;

PFUEL : le prix d'une unité de fuel ;

COST : le coût total de production de la centrale.

Travaillez avec TSP.

Solution

Une production $Q = OUTPUT$ est fabriquée à partir de trois inputs : $x_1 = LABOR$, $x_2 = CAPITAL$, $x_3 = FUEL$, selon une technologie représentée par une certaine fonction de production : $Q = F(x_1, x_2, x_3)$. Supposez que la technologie de production d'électricité peut être représentée par une fonction de production Cobb-Douglas : $Q = Ax_1^{\gamma_1} x_2^{\gamma_2} x_3^{\gamma_3}$.

Les rendements d'échelle sont constants si $\gamma_1 + \gamma_2 + \gamma_3 = 1$. Ils sont croissants si $\gamma_1 + \gamma_2 + \gamma_3 > 1$ et décroissants si $\gamma_1 + \gamma_2 + \gamma_3 < 1$. En toute logique, vous pouvez supposer que les entreprises productrices d'électricité agissent rationnellement, et donc qu'elles minimisent leurs coûts de production. Cela veut dire qu'une entreprise, pour produire un montant Q d'électricité, choisit une quantité x_1 de travail, une quantité x_2 de capital et une quantité x_3 de fuel de manière à minimiser le coût de production. Chaque entreprise résout donc le problème : $\min_{x_1,x_2,x_3} p_1x_1 + p_2x_2 + p_3x_3$ sous la contrainte $Q = Ax_1^{\gamma_1}x_2^{\gamma_2}x_3^{\gamma_3}$, où p_1 est le prix d'une unité de travail, p_2 le prix d'une unité de capital et p_3 le prix d'une unité de fuel. Vous pouvez encore représenter ce problème avec un multiplicateur de Lagrange : $\min_{x_1,x_2,x_3,\lambda} p_1x_1 + p_2x_2 + p_3x_3 + \lambda\left(Q - Ax_1^{\gamma_1}x_2^{\gamma_2}x_3^{\gamma_3}\right)$. Les conditions de premier ordre de ce problème de choix optimal sont : $p_1 = \lambda\gamma_1 Ax_1^{\gamma_1-1}x_2^{\gamma_2}x_3^{\gamma_3}$, $p_2 = \lambda\beta_2 Ax_1^{\gamma_1}x_2^{\gamma_2-1}x_3^{\gamma_3}$, $p_3 = \lambda\beta_3 Ax_1^{\beta_1}x_2^{\gamma_2}x_3^{\gamma_3-1}$ et $Q = Ax_1^{\gamma_1}x_2^{\gamma_2}x_3^{\gamma_3}$. Ces conditions forment un système d'équations dont les solutions sont :

$$
\begin{aligned}
x_1 &= \left(QA^{-1}\left(\frac{p_1}{p_2}\frac{\gamma_2}{\gamma_1}\right)^{-\gamma_2}\left(\frac{p_1}{p_3}\frac{\gamma_3}{\gamma_1}\right)^{-\gamma_3}\right)^{\frac{1}{\gamma_1+\gamma_2+\gamma_3}}\\
x_2 &= \frac{p_1}{p_2}\frac{\gamma_2}{\gamma_1}\left(QA^{-1}\left(\frac{p_1}{p_2}\frac{\gamma_2}{\gamma_1}\right)^{-\gamma_2}\left(\frac{p_1}{p_3}\frac{\gamma_3}{\gamma_1}\right)^{-\gamma_3}\right)^{\frac{1}{\gamma_1+\gamma_2+\gamma_3}}\\
x_3 &= \frac{p_1}{p_3}\frac{\gamma_3}{\gamma_1}\left(QA^{-1}\left(\frac{p_1}{p_2}\frac{\gamma_2}{\gamma_1}\right)^{-\gamma_2}\left(\frac{p_1}{p_3}\frac{\gamma_3}{\gamma_1}\right)^{-\gamma_3}\right)^{\frac{1}{\gamma_1+\gamma_2+\gamma_3}}\\
\lambda &= \frac{1}{\gamma_1+\gamma_2+\gamma_3}\frac{Q}{A}\left(p_1^{\gamma_1}p_2^{\gamma_2}p_3^{\gamma_3}\right)^{\frac{1}{\gamma_1+\gamma_2+\gamma_3}}\left(\frac{p_1}{p_3}\frac{\gamma_3}{\gamma_1}\right)^{-\gamma_3}\\
&\quad\times\left(\left(\gamma_1^{\gamma_2+\gamma_3}\gamma_2^{-\gamma_2}\gamma_3^{-\gamma_3}\right)^{\frac{1}{\gamma_1+\gamma_2+\gamma_3}} + \left(\gamma_1^{-\gamma_1}\gamma_2^{\gamma_1+\gamma_3}\gamma_3^{-\gamma_3}\right)^{\frac{1}{\gamma_1+\gamma_2+\gamma_3}}\right.\\
&\quad\left.+\left(\gamma_1^{-\gamma_1}\gamma_2^{-\gamma_2}\gamma_3^{\gamma_1+\gamma_2}\right)^{\frac{1}{\gamma_1+\gamma_2+\gamma_3}}\right)
\end{aligned}
$$

Le coût total de production d'une quantité Q d'électricité, noté COST, est donné par :

$$COST \underset{\text{déf}}{=} p_1x_1 + p_2x_2 + p_3x_3$$

Cela implique que :

$$
\begin{aligned}
COST = \frac{Q}{A}\left(p_1^{\gamma_1}p_2^{\gamma_2}p_3^{\gamma_3}\right)^{\frac{1}{\gamma_1+\gamma_2+\gamma_3}}\left(\frac{p_1}{p_3}\frac{\gamma_3}{\gamma_1}\right)^{-\gamma_3}&\left(\left(\gamma_1^{\gamma_2+\gamma_3}\gamma_2^{-\gamma_2}\gamma_3^{-\gamma_3}\right)^{\frac{1}{\gamma_1+\gamma_2+\gamma_3}}\right.\\
&\left.+\left(\gamma_1^{-\gamma_1}\gamma_2^{\gamma_1+\gamma_3}\gamma_3^{-\gamma_3}\right)^{\frac{1}{\gamma_1+\gamma_2+\gamma_3}} + \left(\gamma_1^{-\gamma_1}\gamma_2^{-\gamma_2}\gamma_3^{\gamma_1+\gamma_2}\right)^{\frac{1}{\gamma_1+\gamma_2+\gamma_3}}\right)
\end{aligned}
$$

Vous pouvez en déduire la relation en logarithmes suivante :

$$\ln COST = \beta_1 + \beta_2 \ln Q + \beta_3 \ln p_1 + \beta_4 \ln p_2 + \beta_5 \ln p_3$$

où

$$
\begin{aligned}
\beta_1 = \ln&\left(\left(\gamma_1^{\gamma_2+\gamma_3}\gamma_2^{-\gamma_2}\gamma_3^{-\gamma_3}\right)^{\frac{1}{\gamma_1+\gamma_2+\gamma_3}} + \left(\gamma_1^{-\gamma_1}\gamma_2^{\gamma_1+\gamma_3}\gamma_3^{-\gamma_3}\right)^{\frac{1}{\gamma_1+\gamma_2+\gamma_3}}\right.\\
&\left.+\left(\gamma_1^{-\gamma_1}\gamma_2^{-\gamma_2}\gamma_3^{\gamma_1+\gamma_2}\right)^{\frac{1}{\gamma_1+\gamma_2+\gamma_3}}\right) - \frac{\ln A}{\gamma_1+\gamma_2+\gamma_3}
\end{aligned}
$$

$$\beta_2 = \frac{1}{\gamma_1 + \gamma_2 + \gamma_3}, \quad \beta_3 = \frac{\gamma_1}{\gamma_1 + \gamma_2 + \gamma_3},$$
$$\beta_4 = \frac{\gamma_2}{\gamma_1 + \gamma_2 + \gamma_3}, \quad \text{et} \quad \beta_5 = \frac{\gamma_3}{\gamma_1 + \gamma_2 + \gamma_3}$$

Si toutes les entreprises i ont la même technologie de production, la détermination de leurs coûts peut donc être représentée ainsi :

$$\ln COST_i = \beta_1 + \beta_2 \ln Q_i + \beta_3 \ln p_{1i} + \beta_4 \ln p_{2i} + \beta_5 \ln p_{3i} + u_i$$

pour tout $i = 1 \ldots 123$, où le terme d'erreur u_i représente la partie du logarithme du coût de l'entreprise i qui ne s'explique pas en fonction des logarithmes de la quantité produite et des prix des inputs, de la manière décrite par le raisonnement précédent. Il s'agit d'une régression linéaire sur des données en coupe instantanée. L'estimation des coefficients inconnus β_i, par la méthode des MCO, se fait au moyen des instructions TSP suivantes :

```
FREQ N;
SMPL 1 123;
READ(FILE='C:\EL.XLS');
LC=LOG(COST);
LQ=LOG(OUTPUT);
LL=LOG(PRICELAB);
LK=LOG(PRICECAP);
LP=LOG(PFUEL);
NOPLOT;
OLSQ LC C LQ LL LK LP;
```

Les résultats sont les suivants :

```
                          Equation 1
                        ============
          Method of estimation = Ordinary Least Squares

Dependent variable : LC
Current sample : 1 to 123
Number of observations : 123
       Mean of dep. var. = 2.96715      LM het. test = 1.93386 [.164]
  Std. dev. of dep. var. = 1.52498     Durbin-Watson = 1.33427 [<.000]
Sum of squared residuals = 5.35831  Jarque-Bera test = 63.3478 [.000]
   Variance of residuals = .045409   Ramsey's RESET2 = 150.297 [.000]
Std. error of regression = .213095   F (zero slopes) = 1532.51 [.000]
               R-squared = .981114    Schwarz B.I.C. = -6.15258
      Adjusted R-squared = .980474    Log likelihood = 18.1830
          Estimated     Standard
Variable  Coefficient     Error       t-statistic    P-value
C         -7.87891      1.29467       -6.08567       [.000]
LQ        .826268       .010949       75.4667        [.000]
LL        .099547       .130458       .763058        [.447]
LK        .192253       .130796       1.46987        [.144]
LP        .709348       .073380       9.66676        [.000]
```

Ignorez la valeur du test de Durbin et Watson puisque le concept d'autocorrélation des erreurs ne s'applique pas à des régressions utilisant des données en coupe instantanée. Les

99,11 % de la dispersion des logarithmes des coûts entre les différentes entreprises sont expliqués par la dispersion des logarithmes des variables explicatives (prix du travail, du capital et du fuel, production). Cette excellente performance du modèle le rend fiable pour la problématique considérée : vous pouvez l'utiliser pour répondre aux questions initiales. Par ailleurs, les rendements sont croissants dans l'industrie de production d'électricité puisque :

$$\hat{\gamma}_1 + \hat{\gamma}_2 + \hat{\gamma}_3 = \frac{1}{\hat{\beta}_2^{MCO}} = \frac{1}{0{,}826268} = 1{,}21026 > 1$$

Une augmentation de 1 % du prix du fuel entraîne une faible augmentation de 0,71 % du coût total de production, mais seulement après que la combinaison des quantités de facteurs ait été modifiée pour minimiser les coûts, en réponse à la nouvelle structure des prix.

Exercice 5 Le cas Prix des maisons

Énoncé

Vous disposez des données suivantes pour cinq cent quarante-six maisons mises en vente à Windsor, Canada :

PRIX : le prix de vente de la maison.
SUPERFICIE : la superficie de la parcelle sur laquelle se trouve la maison.
CHAMBRES : le nombre de chambres.
SDB : le nombre de salles de bain.
ETAGES : le nombre d'étages.
ALLEE : cette variable vaut 1 si la maison est dotée d'une allée privative, et 0 sinon.
SALLEJEU : cette variable vaut 1 si une salle de jeux est disponible, et 0 sinon.
GAZ : cette variable vaut 1 si un raccordement au gaz est disponible, et 0 sinon.
AIR : cette variable vaut 1 si un système d'air conditionné est installé, et 0 sinon.
GARAGES : le nombre de garages.
SITUATION : cette variable vaut 1 si la situation de la maison est particulièrement agréable, et 0 sinon.

Toutes ces variables se trouvent dans le fichier MAISON.xls téléchargeable à partir du site Internet www.pearsoneducation.fr. Ces données ont été utilisées par Paul Anglin et Ramazan Gencay [ANG 1996]. Mettez au point un outil permettant d'évaluer le prix de vente potentiel d'une maison, en fonction des valeurs de toutes les autres variables. Travaillez avec TSP.

Solution

Il faut d'abord spécifier un modèle linéaire reliant le prix de vente d'une maison à ses propriétés. Supposez que le prix de vente d'une maison est déterminé par le modèle suivant :

$$\begin{aligned} PRIX_i = {} & \beta_1 + \beta_2 SUPERFICIE_i + \beta_3 CHAMBRES_i + \beta_4 SDB_i + \beta_5 ETAGES_i + \beta_6 ALLEE_i \\ & + \beta_7 SALLEJEU_i + \beta_8 GAZ_i + \beta_9 AIR_i + \beta_{10} GARAGES_i + \beta_{11} SITUATION_i + u_i \end{aligned}$$

pour tout $i = 1 \ldots 546$. La variable u_i est un terme d'erreur inobservable.

Il s'agit d'un modèle linéaire reliant des variables en coupe instantanée. Il faut estimer les coefficients β_i par la méthode des MCO. Le programme d'instructions TSP est le suivant :

```
freq n;
smpl 1 546;
read(file='c :\maison.xls');
noplots;
olsq prix c superficie chambres sdb etages allee sallejeu gaz air garages situation;
```

Les résultats de l'estimation sont les suivants :

```
                                  Equation 1
                                  ==========
                Method of estimation = Ordinary Least Squares

Dependent variable : PRIX
Current sample : 1 to 546
Number of observations : 546
       Mean of dep. var. = 68121.6          LM het. test = 36.2854 [.000]
  Std. dev. of dep. var. = 26702.7         Durbin-Watson = 1.61923 [<.000]
Sum of squared residuals = .129829E+12  Jarque-Bera test = 263.929 [.000]
   Variance of residuals = .242672E+09   Ramsey's RESET2 = 20.5948 [.000]
Std. error of regression = 15577.9       F (zero slopes) = 106.635 [.000]
              R-squared = .665908         Schwarz B.I.C. = 6074.72
      Adjusted R-squared = .659663         Log likelihood = -6040.06
               Estimated     Standard
Variable    Coefficient      Error      t-statistic   P-value
C           -3127.96       3433.25      -.911079      [.363]
SUPERFICIE  3.45250         .352737      9.78776       [.000]
CHAMBRES    2341.89        1046.81      2.23716       [.026]
SDB         14819.3        1498.12      9.89194       [.000]
ETAGES      5674.82        897.823      6.32064       [.000]
ALLEE       6886.53        2064.94      3.33498       [.001]
SALLEJEU    6793.14        1797.79      3.77861       [.000]
GAZ         13016.0        3249.42      4.00564       [.000]
AIR         12855.3        1569.26      8.19195       [.000]
GARAGES     4287.96        848.882      5.05130       [.000]
SITUATION   10460.9        1654.98      6.32085       [.000]
```

Ce modèle peut être utilisé par des agents immobiliers lorsqu'ils doivent fixer le prix de vente de départ d'une habitation. Il faut d'abord rassembler les valeurs des différentes variables de droite (superficie...) pour la maison à vendre. Il suffit ensuite d'introduire les valeurs de ces variables dans la formule suivante :

$$\begin{aligned}\text{Prix de vente de départ} = {} & \beta_1 + \beta_2 SUPERFICIE_i + \beta_3 CHAMBRES_i + \beta_4 SDB_i \\ & + \beta_5 ETAGES_i + \beta_6 ALLEE_i + \beta_7 SALLEJEU_i \\ & + \beta_8 GAZ_i + \beta_9 AIR_i + \beta_{10} GARAGES_i \\ & + \beta_{11} SITUATION_i\end{aligned}$$

Remplacez les coefficients inconnus β_i par leurs valeurs estimées du tableau de résultats.

EXERCICE 6 LE CAS PRIX DES HÔTELS

Énoncé

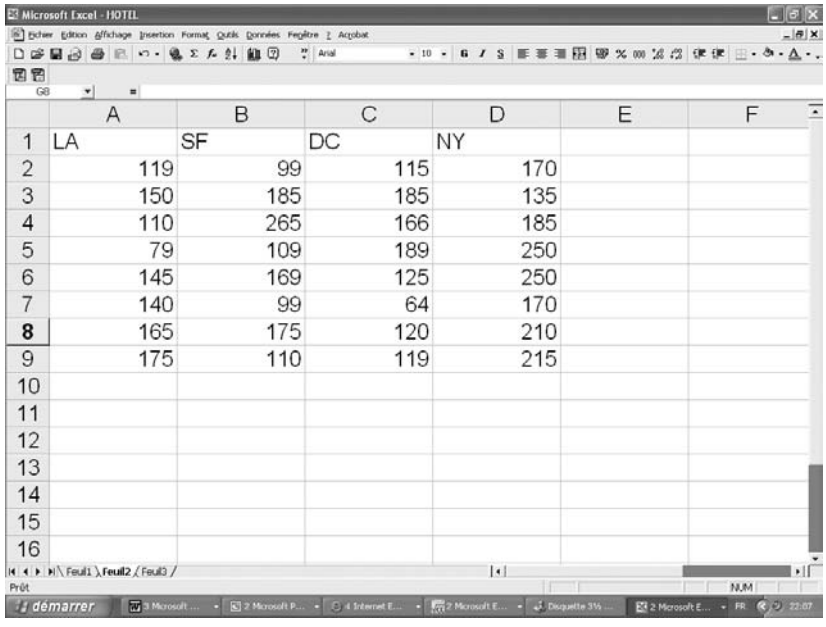
Vous souhaitez savoir si le prix d'une nuitée dans une chambre d'hôtel d'une grande ville américaine dépend de la ville dans laquelle l'hôtel est situé, à qualité égale. Pour répondre à cette question, vous allez analyser des données des villes Los Angeles, Sans Francisco, Washington DC et New York. Pour chacune d'elles, vous disposez du prix d'une nuitée et du nombre d'étoiles de huit hôtels choisis au hasard dans cette ville. Les données se trouvent initialement dans la feuille 1 du classeur Excel Hotel.xls. Ce fichier est téléchargeable sur le site www.pearsoneducation.fr. Les colonnes B, A et C contiennent respectivement les en-têtes Hotel, City, et Price en première ligne, puis les noms des différents hôtels, la ville où ils se situent et le prix d'une nuitée qu'ils tarifent sur les lignes suivantes.

Solution

Solution avec un tableur

Il s'agit d'une application de l'analyse de variance, ou ANOVA ; elle n'est qu'un cas particulier de la méthode des *MCO* avec variables indicatrices (ou dummy). Cette ANOVA est à un facteur. Pour pouvoir utiliser la fonction automatique Analyse de variance : un facteur d'Excel, recopiez d'abord les données nécessaires dans une autre feuille du classeur, chaque modalité de la caractéristique occupant une colonne (voir figure 3.21).

Figure 3.21



	A	B	C	D	E	F
1	LA	SF	DC	NY		
2	119	99	115	170		
3	150	185	185	135		
4	110	265	166	185		
5	79	109	189	250		
6	145	169	125	250		
7	140	99	64	170		
8	165	175	120	210		
9	175	110	119	215		
10						
11						
12						
13						
14						
15						
16						

Sur cette nouvelle feuille appelée Feuil2, cliquez sur Outils puis sur Utilitaire d'analyse. Dans le menu déroulant, sélectionnez Analyse de variance : 1 facteur et cliquez sur OK. Cochez Intitulé en première ligne parce que les modalités (intitulés) de la caractéristique qualitative sont bien en première ligne. Écrivez A1$:$D$9 dans Plage d'entrée puisque ce sont les cellules contenant les données. Saisissez 0,05 comme Seuil de signification. Cochez Colonnes puis Insérer une nouvelle feuille, et nommez-la Anova 1 facteur. Cliquez sur OK. Les résultats s'affichent dans cette nouvelle feuille. La statistique *F* vaut 3,60702982.

La probabilité critique est 0,02548955, inférieure à 0,05. Rejetez donc l'hypothèse que l'espérance du prix d'une nuitée est identique pour les quatre villes. Ce prix dépend de la ville d'accueil.

Solution avec TSP

L'ANOVA à un facteur est une simple application de la régression linéaire. Sachant que :

- La variable dépendante est le vecteur des prix observés pour les trente-deux hôtels.
- Il n'y a pas de constante.
- Les variables explicatives sont quatre variables dummy se rapportant chacune à une ville.
- La variable dummy d'une ville particulière a trente-deux observations qui valent 0 si l'hôtel n'est pas dans cette ville et 1 si l'hôtel est dans cette ville.

... estimez par *MCO* le modèle suivant :

$$Prix_i = \beta_1 DU1_i + \beta_2 DU2_i + \beta_3 DU3_i + \beta_4 DU4_i + u_i, \quad \text{pour} \quad i = 1 \ldots 32$$

... et testez l'hypothèse $\beta_1 = \beta_2 = \beta_3 = \beta_4$ par un test *F* habituel. Ce test n'est valable que si *u* est distribué normalement. Chaque coefficient β_j représente l'espérance de prix, $E(Prix_i)$, pour les hôtels de la ville *j*.

Pour estimer ce modèle avec TSP, suivez la procédure décrite au chapitre 2. Recopiez la colonne C du fichier Hotel.xls dans la colonne A d'un nouveau fichier que vous enregistrerez dans C:\ comme une feuille de calcul Excel, sous le nom de HO. Ensuite, utilisez les instructions suivantes sous TSP :

```
FREQ N;
SMPL 1 32;
READ(FILE='C :\HO.XLS');
LOAD DU1;
1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0;
LOAD DU2;
0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0;
LOAD DU3;
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0;
LOAD DU4;
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1;
NOPLOT;
OLSQ PRICE DU1 DU2 DU3 DU4;
```

Ce programme produit les résultats suivants :

```
Dependent variable: PRICE
Current sample: 1 to 32
Number of observations: 32
       Mean of dep. var. = 155.063      LM het. test = .390756E-03 [.984]
  Std. dev. of dep. var. = 49.4681     Durbin-Watson = 1.95603 [<.643]
Sum of squared residuals = 54714.5  Jarque-Bera test = .833548 [.659]
   Variance of residuals = 1954.09   Ramsey's RESET2 = .850706E+38 [.000]
Std. error of regression = 44.2051   F (zero slopes) = 3.60703 [.025]
```

```
                R-squared = .278743    Schwarz B.I.C. = 171.444
      Adjusted R-squared = .201465    Log likelihood = -164.512
            Estimated    Standard
Variable   Coefficient    Error     t-statistic    P-value
DU1        135.375       15.6289     8.66186        [.000]
DU2        151.375       15.6289     9.68561        [.000]
DU3        135.375       15.6289     8.66186        [.000]
DU4        198.125       15.6289     12.6769        [.000]
```

Vous retrouvez dans ces résultats la statistique F avec sa probabilité critique.

EXERCICE 7 CONSOMMATION ET SIMULTANÉITÉ

Énoncé

Comme au chapitre 1, il faut estimer un modèle linéaire expliquant le taux de croissance de la consommation réelle en fonction d'une constante, du taux de croissance du revenu disponible réel et du taux d'inflation, en utilisant les données du fichier USA.XLS. Tenez compte d'une probable simultanéité entre consommation et revenu, qui pourrait entraîner une dépendance entre le terme d'erreur du modèle et une variable explicative (le taux de croissance du revenu). Il faut donc estimer le modèle en utilisant des variables instrumentales, qui représentent la constante, les taux de croissance retardés de la consommation, du revenu et des prix, ainsi que le taux de chômage UR. Réalisez cette application avec TSP.

Solution

Le programme permettant de réaliser cette estimation avec TSP est le suivant :

```
FREQ A;
SMPL 1960 1994;
READ(FILE='C :\USA.XLS');
SMPL 1961 1994;
DLC=LOG(CT)-LOG(CT(-1));
DLY=LOG(Y)-LOG(Y(-1));
DLP=LOG(P)-LOG(P(-1));
REGOPT(PVPRINT,STARS,LMLAGS=2,QLAGS=2) ALL;
NOPLOT;
SMPL 1962 1994;
INST DLC C DLY DLP INVR C DLY(-1) DLP(-1) DLC(-1) UR;
OLSQ DLC C DLY DLP;
```

Certaines variables instrumentales choisies étant des variables en taux de croissance retardés d'une période, l'estimation par VI doit se faire à partir de 1962 plutôt que 1961. Pour pouvoir comparer valablement les résultats à ceux obtenus par variables instrumentales, vous devez donc réaliser l'estimation par *MCO* à partir de 1962.

Les résultats de l'estimation par variables instrumentales sont :

```
                          Equation 1
                          ==========
            Method of estimation = Instrumental Variable
Dependent variable: DLC
```

```
Endogenous variables: DLY DLP
Included exogenous variables: C
Excluded exogenous variables: DLY(-1) DLP(-1) DLC(-1) UR
Current sample: 1962 to 1994
Number of observations: 33
       Mean of dep. var. = .032166
  Std. dev. of dep. var. = .016776
Sum of squared residuals = .279486E-02
   Variance of residuals = .931620E-04
Std. error of regression = .965205E-02
              R-squared = .696993
      Adjusted R-squared = .676793
           Durbin-Watson = 1.77465 [.066,.565]
                  E'PZ*E = .256476E-03
           Estimated    Standard
Variable  Coefficient     Error      t-statistic     P-value
C          .023945     .011354      2.10896       * [.035]
DLY        .561161     .224709      2.49728       * [.013]
DLP        -.761937    .170785      -4.46138     ** [.000]
```

Comparez ces résultats avec ceux obtenus par la méthode des moindres carrés ordinaires :

```
                                Equation 2
                               ============
                Method of estimation = Ordinary Least Squares

Dependent variable: DLC
Current sample: 1962 to 1994
Number of observations: 33
          Mean of dep. var. = .032166
    Std.  dev. of dep. var. = .016776
   Sum of squared residuals = .269899E-02
      Variance of residuals = .899663E-04
   Std. error of regression = .948506E-02
                 R-squared = .700307
         Adjusted R-squared = .680327
               LM het. test = .401618 [.526]
              Durbin-Watson = 1.98694 [.334,.638]
Breusch/Godfrey LM : AR/MA1 = .135113E-02 [.971]
Breusch/Godfrey LM : AR/MA2 = .064560 [.968]
     Ljung-Box Q-statistic1 = .121645E-02 [.972]
     Ljung-Box Q-statistic2 = .089390 [.956]
  Wald nonlin. AR1 vs. lags = 1.36760 [.505]
                  ARCH test = .421448 [.516]
                 CuSum test = .423037 [.853]
               CuSumSq test = .086548 [1.00]
                  Chow test = .513721 [.676]
     LR het. test (w/ Chow) = -1.40845 [1.00]
            White het. test = 5.31376 [.379]
```

```
          Jarque-Bera test = .515258 [.773]
         Shapiro-Wilk test = .965642 [.370]
           Ramsey's RESET2 = 4.49155 * [.043]
           F (zero slopes) = 35.0512 ** [.000]
            Schwarz B.I.C. = -103.218
   Akaike Information Crit. = -105.463
            Log likelihood = 108.463
            Estimated    Standard
Variable   Coefficient    Error         t-statistic    P-value
C          .019425        .652456E-02    2.97724     ** [.006]
DLY        .671728        .110019        6.10557     ** [.000]
DLP        -.849897       .101655        -8.36062    ** [.000]
```

Les coefficients estimés par la méthode des variables instrumentales sont présentés dans la colonne estimated coefficient du tableau de résultats ayant comme titre equation 1. Les coefficients estimés par la méthode des *MCO* sont présentés dans la colonne estimated coefficient du tableau de résultats ayant comme titre equation 2. Les coefficients estimés par VI et par *MCO* sont différents (même si cette différence est d'une ampleur modérée), ce qui suggère un éventuel problème de simultanéité.

Exercice 8 Consommation par la méthode du maximum de vraisemblance

Énoncé

Utilisez de nouveau les données du fichier USA.xls, téléchargeable à partir du site Internet afférent à ce livre. Estimez le modèle linéaire expliquant la variation logarithmique de la consommation en fonction des variations logarithmiques des revenus et des prix, mais cette fois en utilisant la méthode du maximum de vraisemblance. Supposez que le terme d'erreur est indépendant des variables de droite, qu'il s'agit d'un bruit blanc et qu'il est normal.

Solution

Le modèle à estimer est $DLC_t = \beta_1 + \beta_2 DLY_t + \beta_3 DLP_t + u_t$, où $u_t N(0, \sigma_u^2)$. Les u_t sont supposés être des bruits blancs et indépendants des DLC_t et DLY_t. Le logarithme de la fonction de vraisemblance est donc égal à :

$$\ln L\left(\beta, \sigma_u^2\right) = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln \sigma_u^2 - \sum_{t=1}^{n} \frac{(Y_t - \beta_1 - \beta_2 DLY_t - \beta_3 DLP_t)^2}{2\sigma_u^2}$$

Il faut écrire un programme d'instructions TSP qui trouve les valeurs des coefficients qui maximisent cette fonction. Bien entendu, les valeurs qui maximisent $\ln\left(L(\beta, \sigma_u^2)\right)$ sont les mêmes que celles qui maximisent sa transformation monotone croissante suivante :

$$2\ln L\left(\beta, \sigma_u^2\right) + n\ln 2\pi = -n\ln \sigma_u^2 - \sum_{t=1}^{n} \frac{(Y_t - \beta_1 - \beta_2 DLY_t - \beta_3 DLP_t)^2}{\sigma_u^2}$$

Le programme d'instructions TSP est le suivant :

```
FREQ A;
SMPL 60 94;
READ(FILE='c :\USA.XLS');
LC=LOG(CT);
LY=LOG(Y);
LP=LOG(P);
SMPL 61 94;
DLC=LC-LC(-1);
DLY=LY-LY(-1);
DLP=LP-LP(-1);
FRML EQ1 LOGL=LOG(SIGINV)+LNORM((DLC--XB)*SIGINV);
FRML EQXB1 XB=B1+B2*DLY+B3*DLP;
EQSUB(NAME=OLS) EQ1 EQXB1;
PARAM B1 B2 B3 SIGINV;
SET SIGINV=1;
ML(HITER=N,HCOV=NBW) OLS;
```

Les coefficients à estimer sont `B1`, `B2` et `B3` qui représentent β_1, β_2 et β_3. Il faut également estimer σ_u. Cela revient à estimer σ_u^{-1}, représenté ici par `SIGINV`. `LOGL` représente la partie du logarithme de la fonction de vraisemblance qui correspond à une observation. La fonction `LNORM(x)` représente le logarithme de la fonction de densité d'une $N(0, 1)$ évaluée en x. Vous obtenez les résultats suivants :

```
                    MAXIMUM LIKELIHOOD ESTIMATION
                    =============================

EQUATION: OLS
Working space used: 645
                            STARTING VALUES
                     B1             B2            B3         SIGINV
VALUE           0.00000        0.00000       0.00000        1.00000
F=    31.266    FNEW=   7.9344    ISQZ=  0 STEP=  1.0000    CRIT=  34.236
F=    7.9344    FNEW=  -10.607    ISQZ=  0 STEP=  1.0000    CRIT=  38.141
F=  -10.607     FNEW=  -11.461    ISQZ=  3 STEP= 0.12500    CRIT=  378.75
F=  -11.461     FNEW=  -48.508    ISQZ=  1 STEP= 0.50000    CRIT=  117.74
F=  -48.508     FNEW=  -64.420    ISQZ=  1 STEP= 0.50000    CRIT=  52.250
F=  -64.420     FNEW=  -80.388    ISQZ=  2 STEP= 0.25000    CRIT=  87.521
F=  -80.388     FNEW=  -90.705    ISQZ=  0 STEP=  1.0000    CRIT=  34.624
F=  -90.705     FNEW=  -105.48    ISQZ=  2 STEP= 0.25000    CRIT=  126.22
F=  -105.48     FNEW=  -110.49    ISQZ=  1 STEP= 0.50000    CRIT=  19.698
F=  -110.49     FNEW=  -110.74    ISQZ=  0 STEP=  1.0000    CRIT= 0.50034
F=  -110.74     FNEW=  -110.75    ISQZ=  0 STEP=  1.0000    CRIT= 0.22757E-02
F=  -110.75     FNEW=  -110.75    ISQZ=  0 STEP=  1.0000    CRIT= 0.80716E-07
CONVERGENCE ACHIEVED AFTER 12 ITERATIONS

  34 FUNCTION EVALUATIONS.
Number of observations = 34        Log likelihood = 110.746
        Schwarz B.I.C. = -103.693
                         Standard
Parameter   Estimate      Error       t-statistic   P-value
B1          .016282     .613489E-02   2.65396       [.008]
```

```
B2          .696019     .107093       6.49918       [.000]
B3          -.832804    .099321      -8.38500       [.000]
SIGINV      107.359     13.0191      8.24621        [.000]
Standard Errors computed from analytic second derivatives
(Newton)
                        Standard
Parameter   Estimate      Error       t-statistic   P-value
B1          .016282     .620506E-02   2.62395       [.009]
B2          .696019     .110242       6.31354       [.000]
B3          -.832804    .123473       -6.74481      [.000]
SIGINV      107.359     16.6155       6.46136       [.000]
Standard Errors computed from covariance of analytic first derivatives
(BHHH)
                        Standard
Parameter   Estimate      Error       t-statistic   P-value
B1          .016282     .641869E-02   2.53662       [.011]
B2          .696019     .108838       6.39501       [.000]
B3          -.832804    .082664       -10.0746      [.000]
SIGINV      107.359     11.1076       9.66537       [.000]
Standard Errors computed from analytic first and second derivatives
(Eicker-White)
```

EXERCICE 9 MODÉLISATION DE LA POLITIQUE MONÉTAIRE

Énoncé

Estimez un modèle expliquant le niveau des taux d'intérêts américains en fonction du taux d'inflation et du taux de chômage. Pour que les écarts types estimés des coefficients estimés soient corrects, même en présence d'hétéroscédasticité, utilisez une estimation robuste de la matrice de variance et de covariance des coefficients estimés. Comparez avec les écarts types obtenus par l'estimation habituelle de cette matrice.

Le fichier USA.xls téléchargeable sur le site www.pearsoneducation.fr rassemble plusieurs variables, de 1960 à 1994, dont :

R : le taux d'intérêt ;

UR : le taux de chômage ;

P : l'indice des prix à la consommation.

Travaillez avec TSP.

Solution

Vous allez estimer un modèle linéaire dont la variable dépendante est le taux d'intérêt R_t, et les variables explicatives sont l'inflation INF_t, le taux de chômage UR_t et le taux d'intérêt retardé R_{t-1}. Le taux d'inflation INF_t doit être généré à partir de l'indice des prix P, à l'aide de la formule $INF_t = 100(P_t - P_{t-1})/P_{t-1}$. Le programme permettant d'estimer cette relation par moindres carrés ordinaires est présenté ci-après. Le modèle est estimé successivement sans et avec une matrice de variance et de covariance robuste.

```
freq a;
smpl 1960 1994;
read(file='c :\usa.xls');
smpl 1961 1994;
inf=100*(p-p(-1))/p(-1);
regopt(pvprint,stars) all;
```

```
supres vcov vcor csmax csqmax chow;
noplot;
olsq r c inf ur r(-1);
olsq(robustse, hctype=1) r c inf ur r(-1);
```

L'option `robustse` commande une estimation robuste des écarts types. L'option `hctype=` permet de choisir entre différentes formules. Ici, vous utilisez une formule dont les degrés de liberté sont ajustés pour un petit échantillon. L'instruction `supres` et ses arguments suppriment certains résultats, dont l'impression des tests de constance des coefficients, qui sont inutiles.

```
                            Equation 1
                            ==========
            Method of estimation = Ordinary Least Squares

Dependent variable: R
Current sample: 1961 to 1994
Number of observations: 34
        Mean of dep. var. = 7.64232
  Std. dev.  of dep. var. = 2.61989
 Sum of squared residuals = 23.2655
    Variance of residuals = .775517
 Std. error of regression = .880634
               R-squared = .897285
       Adjusted R-squared = .887014
            LM het. test = 4.81378 * [.028]
           Durbin-Watson = 1.97024 [.247,.694]
              Durbin's h = -.298590 [.765]
         Durbin's h alt. = -.291388 [.771]
               ARCH test = .644711E-02 [.936]
   LR het. test (w/ Chow) = 18.7015 ** [.000]
          White het. test = 12.1299 [.206]
        Jarque-Bera test = 4.35856 [.113]
       Shapiro-Wilk test = .936582 * [.049]
         Ramsey's RESET2 = 2.89648 [.099]
         F (zero slopes) = 87.3571 ** [.000]
         Schwarz B.I.C.  = 48.8470
Akaike Information Crit.  = 45.7943
           Log likelihood = -41.7943
           Estimated     Standard
Variable  Coefficient      Error       t-statistic     P-value
C         .916789        .655369       1.39889         [.172]
INF       .279208        .069364       4.02527      ** [.000]
UR        -.159064       .151533      -1.04969         [.302]
R(-1)     .846701        .091481       9.25547      ** [.000]
```

```
                          Equation 2
                         ============
           Method of estimation = Ordinary Least Squares
            Estimated     Standard
Variable   Coefficient      Error      t-statistic     P-value
C          .916789        .384744      2.38286       * [.024]
INF        .279208        .082467      3.38568      ** [.002]
UR         -.159064       .083728      -1.89976        [.067]
R(-1)      .846701        .098165      8.62526      ** [.000]
Standard Errors are heteroskedastic-consistent (HCTYPE=1).
```

Très logiquement, le coefficient de INF est positif tandis que celui de UR est négatif. Plusieurs tests d'hétéroscédasticité ont des probabilités critiques inférieures à 0,05, ce qui suggère une hétéroscédasticité du terme d'erreur. Cela implique que les écarts types estimés de manière conventionnelle, présentés sous le titre equation 1, ne sont pas fiables. Par contre, les écarts types estimés de manière robuste, présentés sous le titre equation 2, sont fiables. Par ailleurs, la précision de l'estimation du coefficient du taux de chômage semble bien meilleure quand les écarts types sont estimés de manière robuste.

Références bibliographiques

[ANG 1996] P. Anglin, R Gencay, Semiparametric Estimation of a Hedonic Price Function, *Journal of Applied Econometrics*, 11 (6), p. 633–648, 1996.

[BRE 1981] T.S. Breusch, L.G. Godfrey, A Review of Recent Work on Testing for Autocorrelation in Dynamic Simultaneous Models, dans *Macroeconomic Analysis : Essays in Macroeconomics and Econometrics*, eds D. Currie, R. Nobay and D. Peel, Croom Helm, London, 1981.

[BRE 1980] T.S. Breusch, A.R. Pagan, The Lagrange Multiplier Test and its Application to Model Specifications in Econometrics, *Review of Economic Studies*, 47, p. 239–253, 1980.

[CHR 1984] L. Christensen, W. Greene, Economies of Scale in US Electric Power Generation, *Journal of Political Economy*, 84 (4), p. 655–676, 1984.

[DAV 1993] R. Davidson, J.G. MacKinnon, *Estimation and Inference in Econometrics*, Oxford University Press, New York, 1993.

[GRE 2000] W.H. Greene, *Econometric Analysis*, Fourth Edition, Prentice Hall International, New Jersey, 2000.

[MAC 1985] J.G. MacKinnon, H. White, Some Heteroskedasticity-Consistent Matrix Estimators with Improved Finite Sample Properties, *Journal of Econometrics*, 29, p. 305–325, 1985.

[NEW 1987] W.K. Newey, K.D. West, A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix, *Econometrica*, 55, p. 703–708, 1987.

[PES 1999] M.H. Pesaran, L.W. Taylor, Diagnostics for IV Regressions, *Oxford Bulletin of Economics and Statistics*, 61, p. 255–281, 1999.

[PES 1994] M.H. Pesaran, R.J. Smith, A Generalized R Criterion for Regression Models Estimated by the Instrumental Variables Method, *Econometrica*, 62, p. 705–710, 1994.

[SPA 1986] A. Spanos, *Statistical Foundations of Econometric Modelling*, Cambridge University Press, Cambridge, 1986.

[WHI 1980] H. White, A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity, *Econometrica*, 48, p. 817–838, 1980.