Chapitre 3

Modèle de régression linéaire multiple

1. Le modèle linéaire général

1.1. Présentation

Le modèle linéaire général est une généralisation du modèle de régression simple dans lequel figurent plusieurs variables explicatives :

$$y_t = a_0 + a_1 x_{1t} + a_2 x_{2t} + \ldots + a_k x_{kt} + \varepsilon_t \text{ pour } t = 1, \ldots, n$$

avec :

y_t = variable à expliquer à la date t ;
x_{1t} = variable explicative 1 à la date t ;
x_{2t} = variable explicative 2 à la date t ;
...
x_{kt} = variable explicative k à la date t ;
$a_0, a_1, \ldots, a_k$ = paramètres du modèle ;
ε_t = erreur de spécification (différence entre le modèle vrai et le modèle spécifié), *cette erreur est inconnue et restera inconnue* ;
n = nombre d'observations.

1.2. Forme matricielle

En écrivant le modèle, observation par observation, nous obtenons :

$$
\begin{aligned}
y_1 &= a_0 + a_1 x_{11} + a_2 x_{21} + \ldots + a_k x_{k1} + \varepsilon_1 \\
y_2 &= a_0 + a_1 x_{12} + a_2 x_{22} + \ldots + a_k x_{k2} + \varepsilon_2 \\
&\ldots \\
y_t &= a_0 + a_1 x_{1t} + a_2 x_{2t} + \ldots + a_k x_{kt} + \varepsilon_t \\
&\ldots \\
y_n &= a_0 + a_1 x_{1n} + a_2 x_{2n} + \ldots + a_k x_{kn} + \varepsilon_n
\end{aligned}
$$

Soit, sous forme matricielle :

$$
\underset{(n,1)}{Y} = \underset{(n,k+1)}{X} \quad \underset{(k+1,1)}{a} + \underset{(n,1)}{\varepsilon}
$$

avec :

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_t \\ \vdots \\ y_n \end{pmatrix} ; \quad X = \begin{pmatrix} 1 & x_{11} & x_{21} & \dots & x_{k1} \\ 1 & x_{12} & x_{22} & \dots & x_{k2} \\ \vdots & \vdots & \vdots & \dots & \\ 1 & x_{1t} & x_{2t} & \dots & x_{kt} \\ \vdots & \vdots & \vdots & \dots & \\ 1 & x_{1n} & x_{2n} & \dots & x_{kn} \end{pmatrix} ; \quad a = \begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_k \end{pmatrix} ; \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_t \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

L'écriture sous forme matricielle rend plus aisée la manipulation du modèle linéaire général, c'est pourquoi nous l'adoptons par la suite.

2. Estimation et propriétés des estimateurs

2.1. Estimation des coefficients de régression

Soit le modèle sous forme matricielle à k variables explicatives et n observations :

$$Y = Xa + \varepsilon$$

Afin d'estimer le vecteur a composé des coefficients $a_0, a_1 \ldots a_k$, nous appliquons la méthode des Moindres Carrés Ordinaires (MCO) qui consiste à minimiser la somme des carrés des erreurs, soit :

$$\text{Min} \sum_{t=1}^{n} \varepsilon_t^2 = \text{Min}\, \varepsilon'\, \varepsilon = \text{Min}\, (Y - Xa)'(Y - Xa) = \text{Min}\, S$$

$$\frac{\partial S}{\partial a} = -2\,X'Y + 2\,X'X\,\widehat{a} = 0 \rightarrow \boxed{\widehat{a} = (X'\,X)^{-1}X'\,Y}$$

Remarques:

- En cas de colinéarité parfaite entre deux variables explicatives, la matrice $X'X$ est singulière et la méthode des MCO défaillante.
- On appelle équations normales les équations issues de la relation :

A- Forme matricielle:

$$(X'X)\hat{a} = X'Y$$

$$\begin{pmatrix} n & \sum x_{1t} & \sum x_{2t} & \dots & \sum x_{kt} \\ \sum x_{1t} & \sum x_{1t}^2 & \sum x_{1t}\,x_{2t} & \dots & \sum x_{1t}\,x_{kt} \\ \sum x_{2t} & \sum x_{2t}\,x_{1t} & \sum x_{2t}^2 & \dots & \sum x_{2t}\,x_{kt} \\ \vdots & \dots & \dots & \dots & \dots \\ \sum x_{kt} & \sum x_{kt}\,x_{1t} & \sum x_{kt}\,x_{2t} & \dots & \sum x_{kt}^2 \end{pmatrix} \begin{pmatrix} \hat{a}_0 \\ \hat{a}_1 \\ \hat{a}_2 \\ \dots \\ \hat{a}_k \end{pmatrix} = \begin{pmatrix} \sum y_t \\ \sum x_{1t}\,y_t \\ \sum x_{2t}\,y_t \\ \dots \\ \sum x_{kt}\,y_t \end{pmatrix}$$

Le modèle estimé s'écrit :

$$y_t = \hat{a}_0 + \hat{a}_1 x_{1t} + \hat{a}_2 x_{2t} + \ldots + \hat{a}_k\, x_{kt} + e_t$$

avec $e_t = y_t - \widehat{y}_t$ où e_t est le résidu, c'est-à-dire l'écart entre la valeur observée de la variable à expliquer et sa valeur estimée (ajustée).

B- Cas particulier : *Cas des données centrées sur la moyenne*

$$\begin{pmatrix} \widehat{a}_1 \\ \widehat{a}_2 \\ \widehat{a}_3 \\ \dots \\ \widehat{a}_k \end{pmatrix} = \begin{pmatrix} \text{Var}(x_1) & \text{Cov}(x_1, x_2) & \text{Cov}(x_1, x_3) & \dots & \text{Cov}(x_1\ x_k) \\ \text{Cov}(x_2, x_1) & \text{Var}(x_2) & \text{Cov}(x_2, x_3) & \dots & \text{Cov}(x_2, x_k) \\ \text{Cov}(x_3, x_1) & \text{Cov}(x_3, x_2) & \text{Var}(x_3) & \dots & \text{Cov}(x_3, x_k) \\ \dots & & & & \\ \text{Cov}(x_k, x_1) & \text{Cov}(x_k, x_2) & \text{Cov}(x_k, x_3) & \dots & \text{Var}(x_k) \end{pmatrix}^{-1} \times \begin{pmatrix} \text{Cov}\ (x_1, y) \\ \text{Cov}(x_2, y) \\ \text{Cov}(x_3, y) \\ \dots \\ \text{Cov}(x_k, y) \end{pmatrix}$$

avec:

$$\widehat{a}_0 = \overline{y} - \widehat{a}_1\, \overline{x}_1 - \widehat{a}_2\, \overline{x}_2 - \dots - \widehat{a}_k\, \overline{x}_k$$

C- Hypothèses et propriétés des estimateurs

Hypothèses Stochastiques

- H1 : les valeurs $x_{i,t}$ sont observées sans erreur.
- H2 : $E(\varepsilon_t) = 0$, l'espérance mathématique de l'erreur est nulle.
- H3 : $E(\varepsilon_t^2) = \sigma_\varepsilon^2$, la variance de l'erreur est constante $(\forall t)$ (homoscédasticité).
- H4 : $E(\varepsilon_t\ \varepsilon_{t'}) = 0$ si $t \neq t'$, les erreurs sont non corrélées (ou encore indépendantes).
- H5 : $\mathrm{Cov}(x_{it}, \varepsilon_t) = 0$, l'erreur est indépendante des variables explicatives.

Hypothèses Structurelles

- H6 : absence de colinéarité entre les variables explicatives, cela implique que la matrice $(X'\ X)$ est régulière et que la matrice inverse $(X'\ X)^{-1}$ existe.
- H7 : $(X'\ X)/n$ tend vers une matrice finie non singulière.
- H8 : $n > k + 1$, le nombre d'observations est supérieur au nombre des séries explicatives.

Propriétés des estimateurs

Rappel :

Estimateur sans biais ⇒ $E(\hat{a}) = a$

Estimateur convergent ⇒ $\lim_{T \to \infty} V(\hat{a}) = 0$

avec $\hat{a} = (X'X)^{-1} X'Y = (X'X)^{-1} X'(Xa + \varepsilon) = a + (X'X)^{-1} X'\varepsilon$

et $V(\hat{a}) = E\left(\hat{a} - E(\hat{a})\right)^2 = E\left[(\hat{a} - a)(\hat{a} - a)'\right] = \sigma_\varepsilon^2 (X'X)^{-1}$

$$\boxed{V(\hat{a}) = \sigma_\varepsilon^2 \frac{1}{T}\left(\frac{1}{T} X'X\right)^{-1}}$$

On calcule $V(\hat{a}) = \sigma_\varepsilon^2 (X'X)^{-1}$ à l'aide de l'estimateur de σ_ε^2 qui s'écrit comme suit:

$$\hat{\sigma}_\varepsilon^2 = \frac{\sum_{t=1}^{T} e_t^2}{T-k-1} = \frac{SCR}{T-k-1}$$

$$\Rightarrow \quad V(\hat{a}) = \Omega_{\hat{a}} = \sigma_\varepsilon^2 (X'X)^{-1} \qquad \text{et} \qquad \hat{\Omega}_{\hat{a}} = \hat{\sigma}_\varepsilon^2 (X'X)^{-1}$$

Théorème de Gauss-Markov : L'estimateur $\hat{a} = (X'X)^{-1} X'Y$ des moindres carrés est qualifié de **BLUE** (*Best Linear Unbiaised Estimator*), car il s'agit du meilleur estimateur linéaire sans biais (au sens qu'il fournit les variances les plus faibles pour les estimateurs).

2.2 Équation d'analyse de la variance et qualité d'un ajustement

Comme pour le modèle de régression simple, nous avons :

$$a) \quad \sum_t y_t = \sum_t \hat{y}_t \rightarrow \bar{y}_t = \bar{\hat{y}}_t$$

$$b) \quad \sum_t e_t = 0$$

De ces deux relations, nous en déduisons l'équation fondamentale d'analyse de la variance :

$$\sum_t \left(y_t - \bar{y}_t\right)^2 = \sum_t \left(\hat{y}_t - \bar{\hat{y}}_t\right)^2 + \sum_t e_t^2$$

$$SCT \quad = \quad SCE \quad + \; SCR$$

Qualité d'un ajustement

$$R^2 = \frac{SCE}{SCT} = \frac{\sum_t \left(\hat{y}_t - \bar{\hat{y}}_t\right)^2}{\sum_t \left(y_t - \bar{y}_t\right)^2} = 1 - \frac{SCR}{SCT}$$

R^2 : coefficient de détermination

R : coefficient de correlation multiple

Dans le cas de données centrées (moyenne nulle) et seulement dans ce cas, le coefficient de détermination est égal à :

$$R^2 = \frac{\hat{Y}'\hat{Y}}{Y'Y} = 1 - \frac{e'e}{Y'Y}$$

Cette qualité de l'ajustement et l'appréciation que l'on a du R^2 doivent être tempérées par le degré de liberté de l'estimation.

On calcule le R^2 ajusté noté $\bar{R}^2$:

$$\bar{R}^2 = 1 - \frac{T-1}{T-k-1}\left(1-R^2\right)$$

On a $\bar{R}^2 < R^2$ et si T et grand $\bar{R}^2 \simeq R^2$

Remarques

- ✓ Le R^2 ne permet de comparer que des modèles ayant le même nombre de variables explicatives, le même nombre d'observations et la même forme (on ne peut pas comparer un modèle simple avec un modèle en log).

- ✓ Lorsque l'on ajoute des variables explicatives supplémentaires dans un modèle, le R^2 a tendance à augmenter sans qu'il y ait forcément amélioration du modèle. C'est pourquoi, lorsque l'on veut comparer des modèles qui n'ont pas le même nombre de variables explicatives, on utilise le $\bar{R}^2$.

- ✓ En général, lorsque les modèles n'ont pas le même nombre de variables explicatives, on utilise pour comparer les modèles le critère du :

$$S = \hat{\sigma}_\varepsilon = \sqrt{\frac{e'e}{T-k-1}}$$

Le meilleur modèle est celui qui a le S le plus petit.

Exercice 1:

Un chercheur veut étudier la relation existant entre le chiffre d'affaire (considéré comme variable endogène) de la société et le niveau de production de 2 produits (considérés comme variables exogènes). 5 observations ont été relevées.

Y	X1	X2
1	2	4
1	3	2
2	5	2
3	7	1
3	8	1

1. Mettre le modèle sous forme matricielle en spécifiant bien les dimensions de chacune des matrices.
2. Estimer les paramètres du modèle.
3. Calculer l'estimation de la variance de l'erreur ainsi que les écarts types de chacun des coefficients.
4. Calculer R^2 le et le $\bar{R}^2$ corrigé.

3. Les tests statistiques

3.1. Rôle des hypothèses

$$\frac{\sum_{t=1}^{T} e_t^2}{\sigma_\varepsilon^2} = (T-k-1)\frac{\hat{\sigma}_\varepsilon^2}{\sigma_\varepsilon^2} = (T-k-1)\frac{\hat{\sigma}_{a_i}^2}{\sigma_{a_i}^2}$$ suit une loi du χ^2 (chi-deux) à $(T-k-1)$ degrés de liberté

n–k–1 degrés de liberté (somme au carré de n–k–1 variables aléatoires indépendantes normales centrées réduites).

⇨ $\frac{\hat{a}_i - a_i}{\hat{\sigma}_{\hat{a}_i}}$ *(l'écart type théorique est remplacé par l'écart type empirique) suit une loi de Student à n–k–1 degrés de liberté.*

Conclusions

⇨ $(\widehat{a}-a)'\Omega_{\hat{a}}^{-1}(\widehat{a}-a)$ suit une loi du χ^2 (chi-deux) à $k+1$ degrés de liberté (somme au carré de $k+1$ variables aléatoires normales centrées réduites, les $k+1$ coefficients).

⇨ Si on remplace la matrice des variances covariances théoriques des coefficients, par son estimateur $\widehat{\Omega}_{\hat{a}} = \widehat{\sigma}_\varepsilon^2 (X'X)^{-1}$, la loi de probabilité de $\frac{1}{k+1}(\widehat{a}-a)'\widehat{\Omega}_{\hat{a}}^{-1}(\widehat{a}-a)$ est alors un Fisher à $k+1$ et $n-k-1$ degrés de liberté.

En effet, $$F = \frac{\frac{1}{k+1}(\widehat{a}-a)'\left[\sigma_\varepsilon^2 (X'X)^{-1}\right]^{-1}(\widehat{a}-a)}{(n-k-1)\frac{\widehat{\sigma}_\varepsilon^2}{\sigma_\varepsilon^2}\frac{1}{(n-k-1)}}$$ est le rapport de deux chi-deux divisés par leurs degrés de liberté (caractéristique d'une loi de Fisher à $k+1$ et $n-k-1$ degrés de liberté).

3.2. Comparaison d'un paramètre a_i à une valeur fixée $\overline{a}$

Le test d'hypothèses est le suivant :

$$\text{H0} : a_i = \overline{a}$$

$$\text{H1} : a_i \neq \overline{a}$$

Nous savons que :

$\dfrac{\widehat{a_i} - a_i}{\widehat{\sigma}_{\hat{a}i}}$ suit une loi de Student à $n - k - 1$ degrés de liberté.

Sous l'hypothèse H0, cette relation devient :

$$\frac{|\widehat{a_i} - \overline{a}|}{\widehat{\sigma}_{\hat{a}i}} = t^*_{\hat{a}i} \rightarrow \text{ loi de Student } n - k - 1 \text{ degrés de liberté}$$

Si $t^*_{\hat{a}i} > t^{\alpha/2}_{n-k-1}$ alors nous rejetons l'hypothèse H0, a_i est significativement différent de $\overline{a}$ (au seuil de α).

Si $t^*_{\hat{a}i} \leqslant t^{\alpha/2}_{n-k-1}$ alors nous acceptons l'hypothèse H0, a_i n'est pas significativement différent de $\overline{a}$ (au seuil de α).

3.3. Comparaison d'un ensemble de paramètres à un ensemble de valeurs fixées

Nous cherchons à tester *simultanément* l'égalité d'un sous-ensemble de coefficients de régression à des valeurs fixées.

$$\text{H0} : a_q = \overline{a}_q$$

$$\text{H1} : a_q \neq \overline{a}_q$$

q étant le nombre de coefficients retenus, c'est-à-dire la dimension de chacun des vecteurs a_q.

Nous avons démontré que $\frac{1}{k+1}(\widehat{a}-a)'\widehat{\Omega}_{\hat{a}}^{-1}(\widehat{a}-a)$ suit une loi de Fisher à $k+1$ et $n-k-1$ degrés de liberté ; pour un sous-ensemble de paramètres q, l'expression $\frac{1}{q}(\widehat{a}_q-a_q)'\widehat{\Omega}_{\hat{a}q,q}^{-1}(\widehat{a}_q-a_q)$ suit alors une loi de Fisher à q et $n-k-1$ degrés de liberté.

Pour accepter H0, il suffit que :

$$\frac{1}{q}(\widehat{a}_q-\bar{a}_q)'\,\widehat{\Omega}_{\hat{a}q,q}^{-1}(\widehat{a}_q-\bar{a}_q) \leqslant F^{\alpha}(q, n-k-1)$$

$F^{\alpha}(q, n-k-1) =$ loi de Fisher au seuil α à q et $n-k-1$ degrés de liberté.

Remarque

Le test est très important ; en effet, si dans un modèle estimé, un des coefficients (hormis le terme constant) n'est pas significativement différent de 0, ***il convient d'éliminer cette variable 1 et de ré-estimer les coefficients du modèle***.

La cause de cette non-significativité, est due :

- soit à une absence de corrélation avec la variable à expliquer,
- soit à une colinéarité trop élevée avec une des variables explicatives.

3.3. Intervalle de confiance de la variance de l'erreur

L'intervalle de confiance de la variance de l'erreur permet de déterminer une fourchette de variation de l'amplitude de l'erreur. Pour un intervalle à $(1 - \alpha)$ %, il est donné par :

$$IC = \left[\frac{(n-k-1)\hat{\sigma}_{\varepsilon}^2}{\chi_1^2} ; \frac{(n-k-1)\hat{\sigma}_{\varepsilon}^2}{\chi_2^2} \right]$$

Avec χ_1^2 à n – k – 1 degrés de liberté et $\alpha/2$ de probabilité 1 d'être dépassée et χ_2^2 à n – k – 1 degrés de liberté et $(1 - \alpha/2)$ de probabilité d'être dépassée.

3.4. Tests sur les résidus : valeur anormale, effet de levier et point d'influence

a. Matrice HAT

La matrice « *HAT* », notée H, joue un rôle essentiel dans la détection de l'effet de levier.

Nous calculons la matrice « *HAT* » $H = X(X'X)^{-1}X'$.

Les éléments de la première diagonale de cette matrice H sont appelés les leviers, qui déterminent l'influence de l'observation i sur les estimations obtenues par la régression.

Le levier est situé sur la première diagonale de cette matrice soit $h_i = x_i(X'X)^{-1}x_i'$

⇨ Deux propriétés : $0 \leqslant h_i \leqslant 1$ et $\sum_{i=1}^{n} h_i = k + 1$, la somme des éléments de la première diagonale de la matrice H est égale au nombre de paramètres estimés du modèle.

Si chaque observation pèse le même poids, alors les valeurs des h_i doivent être proches de $\frac{k+1}{n}$.

Le levier d'une observation i est donc anormalement élevé si : $h_i > 2\frac{k+1}{n}$, l'observation est alors considérée comme un point de levier (*leverage point*) ou point d'influence.

b. Point de levier et valeur anormale

- Une observation exerce un effet de levier si elle est éloignée des autres en termes de combinaison des variables explicatives ;

  Exemple

  un pays dont la population est faible mais le PIB élevé, chaque facteur explicatif pris individuellement n'est pas surprenant, mais la survenance de deux valeurs à la fois pour un pays est insolite.

- Le point d'influence est une observation qui contribue très fortement au pouvoir explicatif du modèle (sans cette valeur la régression peut être non significative !)

- La valeur prise par la variable explicative est anormale si le résidu de cette observation est beaucoup plus élevé que les autres résidus, pour identifier une valeur anormale nous pouvons calculer le résidu standardisé (ou encore appelé le résidu studentisé).

c. Résidu standardisé (ou studentisé)

Les résidus standardisés notés e_i^S permettent de détecter des valeurs anormales. Le résidu e_i est divisé par son écart type estimé pondéré par le levier : $e_i^S = \frac{e_i}{\hat{\sigma}_e \sqrt{1 - h_i}}$ suit une loi de Student à $n - k - 1$ degrés de liberté, avec $\hat{\sigma}_e = \sqrt{\frac{\sum_t e_i^2}{n - k - 1}}$. Si, par exemple, les résidus standardisés e_i^S sont compris dans l'intervalle $\pm t_{n-k-1}^{0,025}$, on ne suspecte pas de valeurs anormales pour un seuil de confiance 95 %.

Exercice 2:

Soit le modèle à trois variables explicatives :

$$y_t = a_0 + a_1 x_{1t} + a_2 x_{2t} + a_3 x_{3t} + \varepsilon_t$$

Nous disposons des données du tableau 1.

1) Mettre le modèle sous forme matricielle en spécifiant bien les dimensions de chacune des matrices.

2) Estimer les paramètres du modèle.

3) Calculer l'estimation de la variance de l'erreur ainsi que les écarts types de chacun des coefficients.

4) Calculer le R^2 et le $\overline{R}^2$ corrigé.

t	y	x_1	x_2	x_3
1	12	2	45	121
2	14	1	43	132
3	10	3	43	154
4	16	6	47	145
5	14	7	42	129
6	19	8	41	156
7	21	8	32	132
8	19	5	33	147
9	21	5	41	128
10	16	8	38	163
11	19	4	32	161
12	21	9	31	172
13	25	12	35	174
14	21	7	29	180

Solution

1) Forme matricielle

Nous disposons de 14 observations et trois variables explicatives, le modèle peut donc s'écrire :

$$Y = Xa + \varepsilon \quad \text{avec :}$$

$$Y = \begin{pmatrix} 12 \\ 14 \\ 10 \\ \vdots \\ 21 \end{pmatrix} ; \; X = \begin{pmatrix} 1 & 2 & 45 & 121 \\ 1 & 1 & 43 & 132 \\ 1 & 3 & 43 & 154 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 7 & 29 & 180 \end{pmatrix} ; \; a = \begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{pmatrix} ; \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_t \\ \vdots \\ \varepsilon_{14} \end{pmatrix}$$

Dimensions :

$$(14,1) \qquad (14,4) \qquad (4,1) \qquad (14,1)$$

2) Estimation des paramètres

$$\widehat{a} = (X'\ X)^{-1}\ X'\ Y.$$

Calcul de $X'\ X$ et de $(X'\ X)^{-1}$

$$\underset{X'}{\begin{pmatrix} 1 & 1 & 1 & \dots & 1 \\ 2 & 1 & 3 & \dots & 7 \\ 45 & 43 & 43 & \dots & 29 \\ 121 & 132 & 154 & \dots & 180 \end{pmatrix}} \underset{X}{\begin{pmatrix} 1 & 2 & 45 & 121 \\ 1 & 1 & 43 & 132 \\ 1 & 3 & 43 & 154 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 7 & 29 & 180 \end{pmatrix}} =$$

$$= \begin{pmatrix} 14 & 85 & 532 & 2\,094 \\ 85 & 631 & 3\,126 & 13\,132 \\ 532 & 3\,126 & 20\,666 & 78\,683 \\ 2\,094 & 13\,132 & 78\,683 & 317\,950 \end{pmatrix}$$

$$(X'\ X)^{-1} = \begin{pmatrix} 20,16864 & 0,015065 & -0,23145 & -0,07617 \\ 0,015065 & 0,013204 & 0,001194 & -0,00094 \\ -0,23145 & 0,001194 & 0,003635 & 0,000575 \\ -0,07617 & -0,00094 & 0,000575 & 0,000401 \end{pmatrix}$$

Calcul de $X'\ Y$

$$\begin{matrix} & X' & & & \\ \begin{pmatrix} 1 & 1 & 1 & \dots & 1 \\ 2 & 1 & 3 & \dots & 7 \\ 45 & 43 & 43 & \dots & 29 \\ 121 & 132 & 154 & \dots & 180 \end{pmatrix} & \begin{pmatrix} 12 \\ 14 \\ 10 \\ \vdots \\ 21 \end{pmatrix} & = & \begin{pmatrix} 248 \\ 1\,622 \\ 9\,202 \\ 37\,592 \end{pmatrix} \\ & Y & & & \end{matrix}$$

Calcul de $\widehat{a}$

$$(X'\ X)^{-1} \qquad\qquad X'\ Y$$

$$\begin{pmatrix} 20,16864 & 0,015065 & -0,23145 & -0,07617 \\ 0,015065 & 0,013204 & 0,001194 & -0,00094 \\ -0,23145 & 0,001194 & 0,003635 & 0,000575 \\ -0,07617 & -0,00094 & 0,000575 & 0,000401 \end{pmatrix} \begin{pmatrix} 248 \\ 1\,622 \\ 9\,202 \\ 37\,592 \end{pmatrix} = \widehat{a}$$

$$= \begin{pmatrix} 32,89132 \\ 0,801900 \\ -\ 0,38136 \\ -\ 0,03713 \end{pmatrix}$$

Soit $\widehat{a}_0 = 32,89$; $\widehat{a}_1 = 0,80$; $\widehat{a}_2 = -0,38$; $\widehat{a}_3 = -0,03$

3) Calcul de $\widehat{\sigma}_{\varepsilon}^2$ et de $\widehat{\sigma}_{\hat{a}}^2$

$$\widehat{\sigma}_{\varepsilon}^2 = \frac{e'\,e}{n-k-1}$$, nous devons donc calculer le résidu e.

$$e = Y - \widehat{Y} = Y - X\widehat{a}$$

Soit $$e_t = y_t - (\widehat{a}_0 + \widehat{a}_1 x_{1t} + \widehat{a}_2 x_{2t} + \widehat{a}_3 x_{3t})$$

$$e_t = y_t - 32{,}89 - 0{,}80\, x_{1t} + 0{,}38\, x_{2t} + 0{,}03\, x_{3t}$$

Par exemple pour e_1 :

$$e_1 = y_1 - 32{,}89 - 0{,}80\, x_{11} + 0{,}38\, x_{21} + 0{,}03\, x_{31}$$

$$e_1 = 12 - 32{,}89 - 0{,}80 \times 2 + 0{,}38 \times 45 + 0{,}03 \times 121 = -0{,}84$$

Le tableau 2 présente l'ensemble des résultats.

Par construction, la somme des résidus est bien nulle.

$$\widehat{\sigma}_{\varepsilon}^{2} = \frac{e'\ e}{n-k-1} = \frac{\sum_{t=1}^{t=14} e_t^2}{14-3-1} = \frac{67,45}{10} = 6,745$$

Tableau 2 – Calcul du résidu

t	y_t	$\widehat{y}_t$	e_t	e_t^2
1	12	12,84	– 0,84	0,71
2	14	12,39	1,61	2,58
3	10	13,18	– 3,18	10,11
4	16	13,39	1,61	2,58
5	14	17,70	– 3,70	13,67
6	19	17,88	1,12	1,26
7	21	22,20	– 1,20	1,44
8	19	18,86	0,14	0,02
9	21	16,51	4,49	20,14
10	16	18,76	– 2,76	7,63
11	19	17,92	1,08	1,17
12	21	21,90	– 0,90	0,81
13	25	22,71	2,29	5,27
14	21	20,76	0,24	0,06
		Somme	0	67,45

La matrice des variances et covariances estimées des coefficients nous est donnée par

$$\widehat{\Omega}_{\hat{a}} = \widehat{\sigma}_{\varepsilon}^{2}(X'\ X)^{-1}$$

$$\widehat{\Omega}_{\hat{a}} = 6{,}745 \times \begin{pmatrix} 20{,}16864 & 0{,}015065 & -0{,}23145 & -0{,}07617 \\ 0{,}015065 & 0{,}013204 & 0{,}001194 & -0{,}00094 \\ -0{,}23145 & 0{,}001194 & 0{,}003635 & 0{,}000575 \\ -0{,}07617 & -0{,}00094 & -0{,}000575 & 0{,}000401 \end{pmatrix}$$

Les variances des coefficients de régression se trouvent sur la première diagonale :

$$\begin{aligned} \widehat{\sigma}_{\hat{a}_0}^{2} &= 6{,}745 \times 20{,}17 = 136{,}04 \rightarrow \widehat{\sigma}_{\hat{a}_0} = 11{,}66 \\ \widehat{\sigma}_{\hat{a}_1}^{2} &= 6{,}745 \times 0{,}013 = 0{,}087 \rightarrow \widehat{\sigma}_{\hat{a}1} = 0{,}29 \\ \widehat{\sigma}_{\hat{a}2}^{2} &= 6{,}745 \times 0{,}0036 = 0{,}024 \rightarrow \widehat{\sigma}_{\hat{a}2} = 0{,}15 \\ \widehat{\sigma}_{\hat{a}3}^{2} &= 6{,}745 \times 0{,}0004 = 0{,}0026 \rightarrow \widehat{\sigma}_{\hat{a}3} = 0{,}05 \end{aligned}$$

4) Le calcul du R^2 est effectué à partir de la formule

Nous connaissons $e'\ e = 67,45$, il convient de calculer $\sum_t (y_t - \overline{y})^2 = 226,86$.

$$R^2 = 1 - \frac{\sum_t e_t^2}{\sum_t (y_t - \overline{y})^2} = 1 - \frac{67,45}{226,86} = 0,702$$

Le $\overline{R}^2$ corrigé est donné par

$$\overline{R}^2 = 1 - \frac{n-1}{n-k-1}(1 - R^2) = 1 - \frac{14-1}{14-4}(1 - 0,702) = 0,613$$

Exercice 3:

Tests statistiques sur les coefficients et la variance de l'erreur : calcul des leviers et des résidus studentisés

En reprenant les données du tableau 1 et les résultats de l'exercice 1, on demande de répondre aux questions suivantes :

1. **Les variables explicatives sont-elles significativement contributives pour expliquer la variable endogène ?**
2. **Le coefficient a1 est-il significativement inférieur à 1 ?**
3. **Les coefficients a1 et a2 sont-ils simultanément et significativement différents de 1 et – 0,5 ?**
4. **Quel est l'intervalle de confiance pour la variance de l'erreur ?**
5. **Calculer les leviers et les résidus standardisés, existe-t-il des valeurs aberrantes ?**

   (Les seuils choisis seront de 5 %.)

Solution

1) Il convient de calculer les trois ratios de Student et de les comparer à la valeur lue dans la table pour un seuil de 5 %

$$\frac{\widehat{a}_1}{\widehat{\sigma}_{\hat{a}1}} = \frac{0,80}{0,29} = t^*_{\hat{a}1} = 2,75 > t^{0,05}_{10} = 2,228 \rightarrow a_1 \neq 0$$, la variable explicative x_1 est

contributive à l'explication de y ; de même :

$$\frac{\widehat{a}_2}{\widehat{\sigma}_{\hat{a}2}} = \left|\frac{-0,38}{0,15}\right| = t^*_{\hat{a}2} = 2,53 > t^{0,05}_{10} = 2,228 \rightarrow a_2 \neq 0$$

$$\frac{\widehat{a}_3}{\widehat{\sigma}_{\hat{a}3}} = \left|\frac{-0,03}{0,05}\right| = t^*_{\hat{a}3} = 0,60 < t^{0,05}_{10} = 2,228 \rightarrow a_3 = 0$$

$$IC_{a1} = \widehat{a}_1 \pm t_{n-k-1}^{0,05} \cdot \widehat{\sigma}_{\hat{a}1} = 0,80 \pm 2,228 \times 0,29 = [0,14\,;\ 1,45]$$

De même nous obtenons :

$$IC_{a2} = [-0,71\,;\ -0,04] \quad \text{et} \quad IC_{a3} = [-0,14\,;\ 0,08]$$

La valeur 0 n'appartient pas à l'intervalle de confiance à 95 % de a_1 et a_2, donc ces deux coefficients sont significativement différents de 0 ; en revanche, 0 appartient à l'intervalle de confiance de a_3, ce coefficient n'est pas significativement différent de 0.

2) Nous posons le test d'hypothèses suivant :

$$\text{H0} : a_1 = 1$$
$$\text{H1} : a_1 < 1$$

Sous H0, la relation s'écrit :

$$\frac{\widehat{a}_1 - a_1}{\widehat{\sigma}_{\hat{a}1}} = \frac{0{,}80 - 1}{0{,}29} = -0{,}68 > -t_{10}^{0{,}05} = -1{,}81^{1} \Rightarrow \text{acceptation de H0}$$

Nous sommes bien dans la zone d'acceptation de H0

Par souci de simplification, nous pouvons procéder au test de Student en profitant de la symétrie de cette loi, soit à calculer :

$$\frac{|\widehat{a}_1 - a_1|}{\widehat{\sigma}_{\hat{a}1}} = \frac{|0{,}80 - 1|}{0{,}29} = 0{,}68 < t_{10}^{0{,}05} = 1{,}81 \Rightarrow \text{acceptation de H0}$$

Le fait de raisonner sur la valeur absolue du numérateur entraîne une lecture directe de la table et ainsi une construction et interprétation immédiate du test.

3) Le test d'hypothèses est le suivant :

$$\text{H0} : \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} 1 \\ -0,5 \end{pmatrix}$$

$$\text{H1} : \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \neq \begin{pmatrix} 1 \\ -0,5 \end{pmatrix}$$

Examinons les différents éléments de la relation sous H0 :

$$\frac{1}{q}(\widehat{a}_q - a_q)' \, \widehat{\Omega}_{\hat{a}q}^{-1} (\widehat{a}_q - a_q)$$

Nous avons : $q = 2$, $\widehat{a}_q = \begin{pmatrix} 0,80 \\ -0,38 \end{pmatrix}$ et $a_q = \begin{pmatrix} 1 \\ -0,5 \end{pmatrix}$. La matrice des variances et covariances des coefficients a été calculée lors de l'exercice 1, nous ne retenons que la sous-matrice de dimension 2×2 correspondant aux deux coefficients de régression faisant l'objet du test.

$$\widehat{\Omega}_{\hat{a}q} = 6,745 \cdot \begin{pmatrix} 0,013204 & 0,001194 \\ 0,001194 & 0,003635 \end{pmatrix} \rightarrow \widehat{\Omega}_{\hat{a}q}^{-1} = \begin{pmatrix} 11,57140 & -3,80213 \\ -3,80213 & 42,03506 \end{pmatrix}$$

$$F^* = \frac{1}{2}\,(0{,}80-1;\ -0{,}38+0{,}5)\begin{pmatrix} 11{,}57140 & -3{,}80213 \\ -3{,}80213 & 42{,}03506 \end{pmatrix} \times \begin{pmatrix} 0{,}80-1 \\ -0{,}38+0{,}5 \end{pmatrix}$$

$F^* = 0{,}612$ est à comparer à $F^{\alpha}(q, n-k-1) = F^{0{,}05}_{2,10} = 4{,}10$, le F^* empirique est inférieur au F lu dans la table, on accepte l'hypothèse H0. Les données ne sont pas incompatibles avec la possibilité que les coefficients a_1 et a_2 soient simultanément et respectivement égaux à 1 et – 0,5.

4) L'intervalle de confiance de la variance de l'erreur à un seuil $(1-\alpha)\% = 95\%$ $(\alpha = 0{,}05)$ est calculé à partir de la formule pour 10 degrés de liberté :

$$IC = \left[\frac{(n-k-1)\,\widehat{\sigma}^2_{\varepsilon}}{\chi^2_{0{,}025}}\ ;\ \frac{(n-k-1)\,\widehat{\sigma}^2_{\varepsilon}}{\chi^2_{0{,}975}}\right] = \left[\frac{10 \times 6{,}745}{20{,}48}\ ;\ \frac{10 \times 6{,}745}{3{,}25}\right]$$

Soit $3{,}30 \leqslant \sigma^2_{\varepsilon} \leqslant 20{,}75$. La variance vraie (mais inconnue) σ^2_{ε} de l'erreur a 95 % de chance de se situer à l'intérieur de cet intervalle.

5) Le calcul de h_i et des résidus standardisés

Tableau 3 – Valeur des leviers h_i et des résidus standardisés e_i^S

	Résidus		
i	e_i	h_i	e_i^S
1	– 0,8408	0,2790	– 0,3813
2	1,6068	0,2966	0,7377
3	– 3,1800	0,3091	– 1,4732
4	1,6055	0,3248	0,7523
5	– 3,6973	0,2609	– 1,6559
6	1,1220	0,1825	0,4778
7	– 1,2015	0,5327	– 0,6768
8	0,1426	0,2025	0,0615
9	4,4880	0,1804	1,9088
10	– 2,7622	0,1442	– 1,1497
11	1,0830	0,3066	0,5008
12	– 0,8994	0,2115	– 0,3900
13	2,2946	0,4086	1,1489
14	0,2387	0,3605	0,1149

Le seuil du levier est égal à $2\frac{k+1}{n} = 2\frac{4}{14} = 0{,}57$, aucune valeur n'est supérieure à 0,57, nous ne détectons pas de point de levier (ou de point d'influence).

Les résidus studentisés sont tous dans l'intervalle $\pm\, t_{10}^{0,025} = \pm\, 2{,}228$, nous ne détectons pas de valeur anormale.

4. Analyse de la variance et test de signification globale d'une régression

Nous reprenons l'équation fondamentale d'analyse de la variance :

$$\sum_t (y_t - \overline{y})^2 = \sum_t (\widehat{y}_t - \overline{y})^2 + \sum_t e_t^2$$

La régression est jugée significative si la variabilité expliquée est significativement différente de 0.

$$F^* = \frac{\sum_t (\widehat{y}_t - \overline{y})^2 / k}{\sum_t e_t^2 / (n-k-1)} = \frac{R^2/k}{(1-R^2)/(n-k-1)}$$

Source de variation	Somme des carrés	Degré de liberté	Carrés moyens
$x_1, x_2, \ldots, x_k$	$SCE = \sum_t (\widehat{y}_t - \overline{y})^2$	k	SCE/k
Résidu	$SCR = \sum_t e_t^2$	$n-k-1$	$SCR/(n-k-1)$
Total	$SCT = \sum_t (y_t - \overline{y})^2$	$n-1$	

L'hypothèse de normalité des erreurs implique que sous l'hypothèse H0, F^* suit une loi de Fisher (rapport de deux chi-deux). Nous comparons donc ce F^* calculé au F théorique à k et $(n-k-1)$ degrés de liberté : si $F^* > F$ nous rejetons l'hypothèse H0, le modèle est globalement explicatif.

Dans la pratique, ce test est effectué immédiatement grâce à la connaissance du coefficient de détermination R^2

4.1 Autres tests à partir du tableau d'analyse de la variance

1) Introduction d'une ou de plusieurs variables explicatives supplémentaires

L'ajout d'un bloc supplémentaire de variables explicatives améliore-t-il significativement la qualité de l'ajustement ?

2) Stabilité des coefficients du modèle dans le temps (test de CHOW)

Peut-on considérer le modèle comme étant stable sur la totalité de la période, ou bien doit-on considérer deux sous-périodes distinctes d'estimation (changement structurel du modèle) ?

La spécification du modèle est la même, mais les valeurs estimées des coefficients pour les deux échantillons sont différentes.

4) Augmentation de la taille de l'échantillon servant à estimer le modèle

Lorsque la taille de l'échantillon augmente (le nombre d'observations à disposition est plus important), le modèle reste-t-il stable ? Ce test se ramène au test de Chow de stabilité des coefficients sur deux sous-périodes.

Exercice 3:

En reprenant les données de l'exercice précédent

$$
\begin{aligned}
y_t &= 32{,}89 + 0{,}80\, x_{1t} - 0{,}38\, x_{2t} - 0{,}03\, x_{3t} + e_t \\
&\quad (11{,}66) \quad (0{,}29) \quad (0{,}15) \quad (0{,}05) \\
R^2 &= 0{,}702 \\
n &= 14 \\
(\cdot) &= \text{écart type des coefficients}
\end{aligned}
$$

on demande de tester les hypothèses suivantes.

1) L'ajout des variables explicatives x_2 et x_3 améliore-t-il significativement la qualité de l'estimation par rapport à x_1 seul ?

2) Peut-on considérer le modèle (à trois variables explicatives) comme stable sur l'ensemble de la période, ou doit-on procéder à deux estimations, l'une de la période 1 à 7, et l'autre de la période 8 à 14 ?

3) Un économiste suggère que dans ce modèle $a_1 = 1$ et $a_2 = a_3$, qu'en pensez-vous ?

Solution

Nous pouvons tout d'abord appliquer le test de Fisher

$$F^* = \frac{R^2/k}{(1-R^2)/(n-k-1)} = \frac{0,702/3}{(1-0,702)/10} = 7,878 > F_{3,10}^{0,05} = 3,71$$

Nous rejetons l'hypothèse H0 de nullité de tous les coefficients, la régression est globalement significative.

1) Test d'ajout de variables

Étape 1 : calcul de la variabilité totale, expliquée et résiduelle sur le modèle complet.

Les résultats calculés précédemment lors de l'exercice 1 nous ont donné :

$$SCT = \sum_t (y_t - \overline{y})^2 = 226,86\,;\ SCE = \sum_t (\widehat{y}_t - \overline{y})^2 = 159,41\,;$$

$$SCR = e'\ e = 67,45$$

Étape 2 : calcul de la variabilité totale, expliquée et résiduelle sur le modèle à une seule variable explicative x_1 .

Le modèle estimé est le suivant :

$$\begin{aligned} y_t &= 1{,}011 x_{1,t} + 11{,}57 + e_t \\ &\quad (0{,}281) \\ n &= 14 \\ R^2 &= 0{,}52 \\ (.) &= \text{Ecart type} \\ \widehat{\sigma}_\varepsilon &= 3{,}0165 \end{aligned}$$

Nous calculons d'abord

$$SCR^1 = e'\, e = \text{d.d.l.} \times \widehat{\sigma}_\varepsilon^2 = 12 \times 3{,}0165^2 = 109{,}20$$

puis à partir du coefficient de détermination R^2, nous déduisons :

$$SCT^1 = 226{,}86 \text{ et } SCE^1 = 117{,}65$$

Le test d'hypothèses est le suivant :

H0 : $a_2 = a_3 = 0$

H1 : il existe au moins un des deux coefficients non nul.

Étape 3 : tableau d'analyse de la variance.

Tableau d'analyse de la variance pour tester l'ajout d'un bloc de variables explicatives

Source de variation	Somme des carrés	Degré de liberté	Carrés moyens
x_1	$SCE^1 = 117{,}65$	1	117,65
x_1, x_2, x_3	$SCE = 159{,}41$	3	53,14
Résidu	$SCR = 67{,}45$	10	6,74
Total	$SCT = 226{,}85$	13	

Étape 4 : calcul du Fisher empirique.

$$F^* = \frac{(SCE - SCE^1)/(k - k')}{SCR/(n - k - 1)} = \frac{41{,}67/(3-1)}{67{,}45/10} = 3{,}09 < F^{0,05}_{2,10} = 4{,}10$$

Ou encore : F^* : $\dfrac{(SCR^1 - SCR)/(k - k')}{SCR/(n - k - 1)} = \dfrac{(109{,}2 - 67{,}45)/2}{67{,}45/10} = 3{,}09$

Avec $k =$ nombre de variables explicatives du modèle complet et $k' =$ nombre de variables explicatives du modèle sans l'ajout du bloc de variables. Nous acceptons l'hypothèse H0, il n'y a donc pas de différence significative entre les deux variances expliquées, l'ajout des variables explicatives x_2 et x_3 n'améliore pas de manière significative – au seuil de 5 % – le pouvoir explicatif du modèle.

2) Le modèle est-il stable sur la totalité de la période ?

Soit le modèle estimé sur une seule période :

$$y_t = \widehat{a}_1\, x_{1t} + \widehat{a}_2\, x_{2t} + \widehat{a}_3\, x_{3t} + \widehat{a}_0 + e_t \ \text{ pour } t = 1, \ldots, 14$$

ou le modèle estimé sur deux sous-périodes :

$$y_t = \widehat{a}_1^1\, x_{1t} + \widehat{a}_2^1\, x_{2t} + \widehat{a}_3^1\, x_{3t} + \widehat{a}_0^1 + e_t \ \text{ pour } \ t = 1, \ldots, 7$$
$$y_t = \widehat{a}_1^2\, x_{1t} + \widehat{a}_2^2\, x_{2t} + \widehat{a}_3^2\, x_{3t} + \widehat{a}_0^2 + e_t \ \text{ pour } \ t = 8, \ldots, 14$$

Le test d'hypothèses jointes est alors le suivant :

$$\text{H0} : \begin{pmatrix} a_1 = a_1^1 = a_1^2 \\ a_2 = a_2^1 = a_2^2 \\ a_3 = a_3^1 = a_3^2 \\ a_0 = a_0^1 = a_0^2 \end{pmatrix}$$

Ce test de stabilité des coefficients (test de Chow) se ramène à la question suivante : existe-t-il une différence significative entre la somme des carrés des résidus (SCR) de l'ensemble de la période et l'addition de la somme des carrés des résidus calculée à partir des deux sous-périodes $(SCR^1 + SCR^2)$?

Étape 1 : estimation du modèle sur chacune des deux sous-périodes et calcul des sommes des carrés de résidus.

sous-période 1 : données de 1 à 7

$$\begin{aligned} y_t &= 0{,}774x_{1,t} - 0{,}293x_{2,t} - 0{,}012x_{3,t} + 25{,}27 + e_t \\ &\quad (0{,}53) \qquad (0{,}31) \qquad (0{,}10) \\ n &= 7 \\ R^2 &= 0{,}692 \\ (.) &= \text{Ecart type} \\ \widehat{\sigma}_\varepsilon &= 3{,}01759 \end{aligned}$$

Nous pouvons en déduire comme précédemment :

$$SCT^1 = 88{,}85\,;\ SCE^1 = 61{,}54\,;\ SCR^1 = 27{,}31$$

sous-période 2 : données de 8 à 14

$$\begin{aligned} y_t &= 1{,}228x_{1,t} - 0{,}620x_{2,t} - 0{,}184x_{3,t} + 62{,}63 + e_t \\ &\quad (0{,}69) \qquad (0{,}52) \qquad (0{,}15) \\ n &= 7 \\ R^2 &= 0{,}543 \\ (.) &= \text{Ecart type} \\ \widehat{\sigma}_\varepsilon &= 2{,}6281 \end{aligned}$$

D'où $SCT^2 = 45{,}43\,;\ SCE^2 = 24{,}70\,;\ SCR^2 = 20{,}73.$

Étape 2 : calcul du Fisher empirique.

En prenant au dénominateur la plus faible des sommes des carrés (soit $SCR^1 + SCR^2$), le Fisher empirique est égal à :

$$F^* = \frac{[SCR - (SCR^1 + SCR^2)]/ddl_n}{(SCR^1 + SCR^2)/ddl_d}$$

avec $ddl_n = (n - k - 1) - [(n_1 - k - 1) + (n_2 - k - 1)] = k + 1 = 4$

car $n = n_1 + n_2$

$$ddl_d = (n_1 - k - 1) + (n_2 - k - 1) = n - 2(k + 1) = 6$$

d'où

$$F^* = \frac{[(67{,}45 - (27{,}31 + 20{,}73))]/4}{(27{,}31 + 20{,}73)/6} = \frac{4{,}852}{8{,}00} = 0{,}606 < F_{4;6}^{0,05} = 4{,}53$$

L'hypothèse H0 est acceptée, les coefficients sont significativement stables sur l'ensemble de la période.

3) Test de $a_1 = 1$ et $a_2 = a_3$

Si cette hypothèse est vérifiée, le modèle :

$$y_t = a_0 + a_1\, x_{1t} + a_2\, x_{2t} + a_3\, x_{3t} + \varepsilon_t$$

peut s'écrire :

$$y_t = a_0 + 1\, x_{1t} + a_2\, x_{2t} + a_2\, x_{3t} + \varepsilon_t$$

ou encore :

$$\begin{aligned} y_t - x_{1t} &= a_0 + a_2(x_{2t} + x_{3t}) + \varepsilon_t \\ z_t &= a_0 + a_2\, v_t + \varepsilon_t \end{aligned}$$

Il convient de constituer la nouvelle variable à expliquer, z_t, et la nouvelle variable explicative v_t, puis d'effectuer la régression de z_t sur v_t.

L'estimation des deux $(k' + 1)$ coefficients du modèle conduit aux résultats suivants :

$$\begin{aligned} z_t &= -0{,}0111 v_t + 13{,}74 + e_t \\ &\quad (0{,}051) \\ n &= 14 \\ R^2 &= 0{,}0389 \\ (.) &= \text{Ecart type} \\ \widehat{\sigma}_\varepsilon &= 3{,}0109 \end{aligned}$$

Variables transformées sous l'hypothèse de vérification des contraintes

t	$z_t = y_t - x_{1t}$	$v_t = x_{2t} + x_{3t}$
1	10	166
2	13	175
3	7	197
4	10	192
5	7	171
6	11	197
7	13	164
8	14	180
9	16	169
10	8	201
11	15	193
12	12	203
13	13	209
14	14	209

Nous pouvons en déduire :

$$SCT^1 = 109{,}21\,;\ SCE^1 = 0{,}425\,;\ SCR^1 = 108{,}78$$

$$SCT^1 = 109{,}21\,;\ SCE^1 = 0{,}425\,;\ SCR^1 = 108{,}78$$

L'hypothèse à tester est donc :

H0 : les restrictions sont toutes vérifiées $(SCR^1 = SCR)$.

H1 : il existe au moins une restriction non vérifiée $(SCR^1 \neq SCR)$.

Le Fisher empirique est donné par :

$$F^* = \frac{(SCR^1 - SCR)/ddl_n}{SCR/(n-k-1)} = \frac{(108{,}78 - 67{,}45)/2}{67{,}45/10} = 3{,}06 < F_{2,10}^{0,05} = 4{,}10$$

avec $ddl_n = (n-k'-1) - (n-k-1) = k - k' = 2$.

L'hypothèse H0 est acceptée, les contraintes envisagées sur les coefficients sont compatibles avec les données.

4. La prévision à l'aide du modèle linéaire général et la régression récursive

A. Prédiction conditionnelle

Le modèle général estimé est le suivant :

$$y_t = \widehat{a}_0 + \widehat{a}_1 x_{1t} + \widehat{a}_2 x_{2t} + \ldots + \widehat{a}_k x_{kt} + e_t$$

La prévision pour la donnée $t + h$ (respectivement $i + h$ pour les modèles en coupe instantanée) est la suivante :

$$\widehat{y}_{t+h} = \widehat{a}_0 + \widehat{a}_1 x_{1t+h} + \widehat{a}_2 x_{2t+h} + \ldots + \widehat{a}_k x_{kt+h}$$

L'erreur de prévision est donnée par :

$$e_{t+h} = y_{t+h} - \widehat{y}_{t+h}$$

Considérant que les hypothèses du modèle linéaire général sont vérifiées, la prévision $\widehat{y}_{t+h}$ est sans biais.

B. Fiabilité de la prévision et intervalle de prévision

L'erreur de prévision calculée en t à l'horizon h peut s'écrire aussi :

$$e_{t+h} = y_{t+h} - \widehat{y}_{t+h} = X'_{t+h}(a - \widehat{a}) + \varepsilon_{t+h}$$

La variance de l'erreur de prévision est donc égale à :

$$\boxed{\sigma^2_{et+h} = \sigma^2_{\varepsilon}[X'_{t+h}(X'\ X)^{-1}\ X_{t+h} + 1]}$$

Avec $X_{t+h} = \begin{bmatrix} 1 \\ x_{1t+h} \\ x_{2t+h} \\ \dots \\ x_{kt+h} \end{bmatrix}$ vecteur des valeurs prévues des variables explicatives.

$$\boxed{y_{t+h} = \widehat{y}_{t+h} \pm t^{\alpha/2}_{n-k-1}\sqrt{\widehat{\sigma}^2_{\varepsilon}[X'_{t+h}(X'\ X)^{-1}\ X_{t+h} + 1]}}$$

Exercice 4:

Prévision à partir du modèle linéaire

En reprenant les données de l'exercice 1 (tableau 1), on demande :

1) d'estimer le modèle à deux variables explicatives :

$y_t = a_0 + a_1 x_{1t} + a_2 x_{2t} + \varepsilon_t$ (puisque nous avons montré que la variable x_3 n'est pas significative) ;

2) de calculer une prévision et son intervalle à 95 % pour les périodes 15 et 16, sachant que :

$$x_{1\,15} = 3\,;\; x_{1\,16} = 6 \;\text{ et }\; x_{2\,15} = 24\,;\; x_{2\,16} = 38$$

Solution

1) L'estimation du modèle à deux variables explicatives conduit aux résultats suivants :

$$\begin{aligned} y_t &= 25{,}84 + 0{,}715x_{1,t} - 0{,}328x_{2,t} + e_t \\ &\qquad\quad (0{,}26) \qquad (0{,}13) \\ n &= 14 \\ R^2 &= 0{,}687 \\ (.) &= \text{Ecart type} \\ \widehat{\sigma}_\varepsilon &= 2{,}538 \end{aligned}$$

Nous remarquons les t de Student supérieurs à 2,201, les coefficients a_1 et a_2 sont significativement différents de 0.

2) La prévision pour la période 15 est calculée à partir du modèle estimé :

$$\widehat{y}_{15} = 25{,}84 + 0{,}71\ x_{1\,15} - 0{,}33\ x_{2\,15} = 25{,}84 + 0{,}71 \times 3 - 0{,}33 \times 24$$

$$\widehat{y}_{15} = 20{,}25$$

De même, pour la période 16, on obtient :

$$\widehat{y}_{16} = 25{,}84 + 0{,}71\ x_{1\,16} - 0{,}33\ x_{2\,16} = 25{,}84 + 0{,}71 \times 6 - 0{,}33 \times 38$$

$$\widehat{y}_{16} = 17{,}26$$

Les écarts types de l'erreur de prévision sont donnés par

$\widehat{\sigma}^2_{e15} = \widehat{\sigma}^2_{\varepsilon}[X'_{15}(X'\ X)^{-1}\ X_{15} + 1]$. Nous devons calculer $(X'\ X)^{-1}$, les autres éléments étant connus.

$$X'\ X = \begin{bmatrix} 14 & 85 & 532 \\ 85 & 631 & 3\,126 \\ 532 & 3\,126 & 20\,666 \end{bmatrix} \rightarrow (X'\ X)^{-1} =$$

$$= \begin{bmatrix} 5,707687 & -0,16341 & -0,12221 \\ -0,16341 & 0,011001 & 0,002542 \\ -0,12221 & 0,002542 & 0,002809 \end{bmatrix}$$

$$\widehat{\sigma}^2_{e\,15} = (2,538)^2 \left[(1\ 3\ 24) \begin{bmatrix} 5,707687 & -0,16341 & -0,1222 \\ -0,16341 & 0,011001 & 0,002542 \\ -0,12221 & 0,002542 & 0,0022809 \end{bmatrix} \begin{bmatrix} 1 \\ 3 \\ 24 \end{bmatrix} + 1 \right]$$

$$\widehat{\sigma}^2_{e15} = 6,44 \cdot [0,94 + 1] = 12,53 \rightarrow \widehat{\sigma}_{e15} = 3,54$$

De même, nous pouvons déterminer :

$$\widehat{\sigma}^2_{e16} = 6,44 \cdot [0,071 + 1] = 6,90 \rightarrow \widehat{\sigma}_{e16} = 2,62$$

Les intervalles de prévision peuvent être calculés par

$$y_{t+h} = \widehat{y}_{t+h} \pm t_{n-k-1}^{\alpha/2} \sqrt{\widehat{\sigma}_{\varepsilon}^{2}[X'_{t+h}(X'\ X)^{-1}\ X_{t+h} + 1]}$$

$$y_{15} = \widehat{y}_{15} \pm t_{14-2-1}^{0,025} \cdot \widehat{\sigma}_{e15} = 20{,}05 \pm 2{,}201 \times 3{,}54$$

$IC_{15}^{0,05} = [12{,}26\,;\ 27{,}84]$, la prévision pour la période 15 a 95 % de chances de se situer dans cet intervalle et la prévision de la période 16 :

$$IC_{16}^{0,05} = [11{,}49\,;\ 23{,}03]$$