TESTS D'HYPOTHÈSES LINÉAIRES SUR UN MODÈLE DE RÉGRESSION

par

Georges ROTTIER (1)

SOMMAIRE

(1) Professeur associé à l'Université Paris I, Panthéon Sorbonne.

INTRODUCTION

Cette note présente une formulation simple, mais générale, du test d'une hypothèse linéaire quelconque sur le vecteur des paramètres d'un modèle de régression. Cette catégorie de tests s'introduit couramment, non seulement dans les problèmes de régression proprement dits, mais plus encore dans les modèles d'analyse de la variance et de la covariance, dont on sait qu'ils peuvent se ramener à des modèles de régression. L'exposé évite l'utilisation du théorème de Cochran sur la décomposition des formes quadratiques.

Nous nous placerons dans l'espace des observations. On sait que si l'on a n observations de la variable expliquée y et des variables explicatives x_k, cet espace a n dimensions, les n observations de chaque variable y étant représentées par un vecteur. On admet que l'espace des observations est euclidien, afin de pouvoir donner un sens aux notions d'orthogonalité et de distance.

1. LE MODÈLE

Nous avons une variable endogène y, $p-1$ vraies variables explicatives $x_1 \ldots x_{p-1}$, et une variable auxiliaire x_p, égale à 1 pour toutes les observations et dont le coefficient est le terme constant du modèle. Pour n observations, le modèle s'écrit :

$$\underset{(n,\,1)}{\mathbf{y}} = \underset{(n,\,p)}{\mathbf{X}} \; \underset{(p,\,1)}{\mathbf{a}} + \underset{(n,\,1)}{\mathbf{u}} \tag{1}$$

Nous admettrons que le modèle (1) satisfait à toutes les hypothèses classiques des moindres carrés : les x_j sont mesurés sans erreur, aucune contrainte « a priori » n'existe sur le vecteur $\mathbf{a}$, la matrice $\mathbf{X}$ est de rang p, et le résidu $\mathbf{u}$ est un vecteur qui suit une loi normale à n dimensions de moyenne nulle et de matrice des variances et covariances $V(\mathbf{u}) = \sigma^2 \mathbf{I}_n$, ou, en résumé:

$$\mathbf{u} \rightsquigarrow N(o, \sigma^2 \mathbf{I}_n) \tag{2}$$

On sait que, sous ces hypothèses, le meilleur estimateur sans biais et linéaire en $\mathbf{y}$ de $\mathbf{a}$ (ou estimateur de Gauss-Markov) s'écrit :

$$\hat{\mathbf{a}} = (\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'\mathbf{y} \tag{3}$$

Si $\hat{\mathbf{y}}$ est l'explication linéaire de $\mathbf{y}$ en termes des p vecteurs colonnes de la matrice $\mathbf{X}$, et $\hat{\mathbf{u}}$ le vecteur des écarts entre $\mathbf{y}$ et $\hat{\mathbf{y}}$, on peut écrire :

$$\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'\mathbf{y} \tag{4}$$

$$\mathbf{y} = \hat{\mathbf{y}} + \hat{\mathbf{u}} \tag{5}$$

Posons :

$$\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}' \tag{6}$$

$$\mathbf{M} = \mathbf{I}_n - \mathbf{P} \tag{7}$$

On peut alors écrire :

$$\hat{\mathbf{y}} = \mathbf{P}\mathbf{y} \tag{8}$$

$$\hat{\mathbf{u}} = (\mathbf{I}_n - \mathbf{P})\mathbf{y} = \mathbf{M}\mathbf{y} \tag{9}$$

2. PROPRIÉTÉS DES MATRICES P ET M

On vérifie par calcul direct que $\mathbf{P}$ et $\mathbf{M}$ sont symétriques et idempotentes :

$$\mathbf{P} = \mathbf{P}' = \mathbf{P}^2 \tag{10}$$

$$\mathbf{M} = \mathbf{M}' = \mathbf{M}^2 \tag{11}$$

On sait que le rang d'une matrice idempotente est égal à sa trace, or :

$$\text{tr } \mathbf{P} = \text{tr } \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}^1 = \text{tr } \mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \text{tr } \mathbf{I}_p = p \tag{12}$$

$\mathbf{P}$ est donc de rang p. On déduit immédiatement de (7) et (12) que le rang de $\mathbf{M}$ est égal au rang de $\mathbf{I}_n - \mathbf{I}_p$, c'est-à-dire à $n - p$.

La matrice $\mathbf{P}$ définit une projection orthogonale sur la variété linéaire des $\mathbf{X}$: soit en effet $\mathbf{v}$ un vecteur orthogonal à cette variété ; il sera tel que :

$$\mathbf{X}'\mathbf{v} = 0 \tag{13}$$

L'image $\mathbf{v}_1$ de $\mathbf{v}$ par l'application de matrice $\mathbf{P}$ s'écrira :

$$\mathbf{v}_1 = \mathbf{P}\mathbf{v} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'\mathbf{v} = 0 \tag{14}$$

Le produit des matrices $\mathbf{M}$ et $\mathbf{P}$ est nul :

$$\mathbf{M}\mathbf{P} = (\mathbf{I}_n - \mathbf{P})\mathbf{P} = \mathbf{P} - \mathbf{P}^2 = \mathbf{P} - \mathbf{P} = 0 \tag{15}$$

Il en résulte que si $\mathbf{v}$ est un vecteur quelconque de $\mathbb{R}^n$, ses images $\mathbf{v}_1$ et $\mathbf{v}_2$ par les applications de matrices $\mathbf{P}$ et $\mathbf{M}$ sont orthogonales.

En effet :

$$\mathbf{v}_1 = \mathbf{P}\mathbf{v} \qquad \mathbf{v}_2 = \mathbf{M}\mathbf{v} \tag{16}$$

d'où le produit scalaire de $\mathbf{v}_1$ et $\mathbf{v}_2$:

$$\mathbf{v}_1'\mathbf{v}_2 = \mathbf{v}'\mathbf{P}'\mathbf{M}\mathbf{v} = \mathbf{v}'\mathbf{P}\mathbf{M}\mathbf{v} = 0 \tag{17}$$

Les matrices **P** et **M** définissent donc des projections orthogonales d'un vecteur **v** quelconque de $\mathbb{R}^n$ sur deux sous-espaces complémentaires et orthogonaux de dimensions respectives p et $n - p$.

Comme les matrices **P** et **M** sont symétriques et idempotentes, elles peuvent être diagonalisées par une matrice orthogonale qui définit une rotation d'axes orthonormés dans $\mathbb{R}^n$. Montrons que la même matrice orthogonale **W** (telle donc que $\mathbf{W}' = \mathbf{W}^{-1}$) diagonalise **P** et **M** :

La matrice **P** est diagonalisée par la relation :

$$\mathbf{W}'\mathbf{P}\mathbf{W} = \mathbf{J}_p \tag{18}$$

puisque les valeurs propres d'une matrice idempotente (n,n) de rang p sont p fois 1 et $(n-p)$ fois 0. Or,

(19)

$$\mathbf{W}'\mathbf{M}\mathbf{W} = \mathbf{W}'(\mathbf{I}_n - \mathbf{P})\mathbf{W} = \mathbf{W}'\mathbf{W} - \mathbf{W}'\mathbf{P}\mathbf{W} = \mathbf{W}^{-1}\mathbf{W} - \mathbf{J}_p = \mathbf{J}_{n-p}$$

en désignant par $\mathbf{J}_p$ et $\mathbf{J}_{n-p}$ les matrices carrées de format (n, n) qui ont respectivement p fois et $n - p$ fois l'unité comme termes de leur diagonale principale, et dont tous les autres éléments sont nuls.

La rotation d'axes orthonormés définie par la matrice **W** applique donc les p premières coordonnées de **y** sur la variété linéaire des **X** et les $n - p$ autres coordonnées sur la variété complémentaire qui lui est orthogonale.

3. ANALYSE DE LA VARIANCE DE y

D'après l'hypothèse (2),

$$V(\mathbf{y}) = \mathrm{E}\,[(\mathbf{y} - \mathbf{X}\mathbf{a})(\mathbf{y} - \mathbf{X}\mathbf{a})'] = \sigma^2\mathbf{I}_n \tag{20}$$

En outre, d'après (8)

$$\hat{\mathbf{y}} = \mathbf{P}\mathbf{y} = \mathbf{P}(\mathbf{X}\mathbf{a} + \mathbf{u}) \quad \text{et} \quad \mathrm{E}\,(\hat{\mathbf{y}}) = \mathbf{P}\mathbf{X}\mathbf{a} \tag{21}$$

Donc :

$$\hat{\mathbf{y}} - \mathrm{E}\,(\hat{\mathbf{y}}) = \mathbf{P}\mathbf{u} \tag{22}$$

$$V(\hat{\mathbf{y}}) = \mathrm{E}\,\{[\hat{\mathbf{y}} - \mathrm{E}\,(\hat{\mathbf{y}})]\,[\hat{\mathbf{y}} - \mathrm{E}\,(\hat{\mathbf{y}})]'\} = \mathbf{P}\mathbf{u}\mathbf{u}'\mathbf{P}' = \sigma^2\mathbf{P}\mathbf{P}' = \sigma^2\mathbf{P} \tag{23}$$

Dans le nouveau système d'axes :

$$V(\hat{\mathbf{y}}) = \sigma^2\mathbf{J}_p \tag{24}$$

et

$$V(\hat{\mathbf{u}}) = \sigma^2\mathbf{J}_{n-p} \tag{25}$$

Comme $\hat{\mathbf{y}}$ et $\hat{\mathbf{u}}$ appartiennent à deux sous-espaces complémentaires orthogonaux, on peut écrire par blocs :

$$V(\mathbf{y}) = \begin{pmatrix} V(\hat{\mathbf{y}}) & 0 \\ 0 & V(\hat{\mathbf{u}}) \end{pmatrix} = \sigma^2 \left(\begin{array}{c|c} \mathbf{I}_p & 0 \\ \hline 0 & \mathbf{I}_{n-p} \end{array}\right) = \sigma^2 \mathbf{I}_n \tag{26}$$

En tenant compte de l'hypothèse (2), on peut écrire, dans le nouveau système d'axes :

$$\hat{\mathbf{y}} \rightsquigarrow N(\mathbf{X}\mathbf{a},\ \sigma^2 \mathbf{I}_p) \tag{27}$$

$$\hat{\mathbf{u}} \rightsquigarrow N(0,\ \sigma^2 \mathbf{I}_{n-p}) \tag{28}$$

On en déduit que :

$$\hat{\mathbf{u}}'\hat{\mathbf{u}} \rightsquigarrow \sigma^2 \chi^2_{n-p} \tag{29}$$

et, *si et seulement si*, $\mathbf{a} = 0$,

$$\hat{\mathbf{y}}'\hat{\mathbf{y}} \rightsquigarrow \sigma^2 \chi^2_p \tag{30}$$

Les deux χ^2 ainsi définis sont indépendants en probabilité, puisque $\hat{\mathbf{y}}$ et $\hat{\mathbf{u}}$ sont vecteurs aléatoires normaux dont la covariance est nulle.

Une autre démonstration consiste à poser :

$$\hat{\mathbf{u}}'\hat{\mathbf{u}} = \mathbf{u}'\mathbf{M}\mathbf{u} \tag{31}$$

qui s'écrit dans le nouveau système d'axes :

$$\mathbf{u}'\mathbf{I}_{n-p}\mathbf{u} = (n-p)u^2 \tag{32}$$

et à remarquer que u est une variable normale centrée de variance σ^2 (1).

De même, si $\mathbf{a} = 0$,

$$\hat{\mathbf{y}}'\hat{\mathbf{y}} = \mathbf{u}'\mathbf{I}_p\mathbf{u} = pu^2 \text{ (1)} \tag{33}$$

Si $\mathbf{a} = 0$, on peut établir le tableau suivant d'analyse de la variance :

Variance	degrés de liberté	Distribution
$\hat{\mathbf{y}}'\hat{\mathbf{y}}$	p	$\sigma^2\chi^2_p$
$\hat{\mathbf{u}}'\hat{\mathbf{u}}$	$n-p$	$\sigma^2\chi^2_{n-p}$
$\mathbf{y}'\mathbf{y}$	n	$\sigma^2\chi^2_n$

Si $\mathbf{a} = 0$, on peut donc écrire :

$$\frac{\hat{\mathbf{y}}'\hat{\mathbf{y}}}{p} \cdot \frac{n-p}{\hat{\mathbf{u}}'\hat{\mathbf{u}}} = \frac{\chi_p}{p} \cdot \frac{n-p}{\chi^2_{n-p}} = \mathbf{F}_{p,n-p} \tag{34}$$

(1) Les propriétés que nous venons d'établir ne valent que dans le cas des moindres carrés ordinaires, c'est-à-dire quand $V(\mathbf{u}) = \sigma^2\mathbf{I}_n$. Dans le cas des moindres carrés généralisés, où $V(\mathbf{u}) = \sigma^2\Omega, \Omega$ étant une matrice définie positive quelconque, l'étude est beaucoup plus délicate. (Cf. E. MALINVAUD, **Méthodes statistiques de l'économétrie**, 2e édition, 1969, Dunod, Paris, Chap. 5.)

(34) permet d'établir le test de $\mathbf{a} = 0$. Ce test est cependant d'un faible intérêt pratique, puisque la dernière composante de $\mathbf{a}$ est le terme constant du modèle. Ce qui est intéressant, en fait, c'est de tester la nullité des coefficients des vraies variables exogènes, ou d'un sous-ensemble de ces coefficients. Nous retrouverons ces tests comme cas particulier du test d'une hypothèse linéaire générale que nous allons maintenant établir.

4. TEST D'UNE HYPOTHÈSE LINÉAIRE GÉNÉRALE SUR a

Une hypothèse linéaire sur $\mathbf{a}$ peut s'écrire :

$$\mathbf{Ca} = 0 \tag{35}$$

où $\mathbf{C}$ est une matrice de format rp, avec $r < p$. (35) traduit l'hypothèse que $\mathbf{Xa}$ est orthogonal aux r vecteurs lignes de $\mathbf{C}$. Si $\mathbf{C}$ est de rang r, ce que nous admettrons dans la suite, (35) implique que $\mathbf{Xa}$ appartient à un sous-espace à $p - r$ dimensions de la variété linéaire à p dimensions définie par $\mathbf{X}$.

Sous les hypothèses posées dans le chapitre 1, la distribution de $\mathbf{C}\hat{\mathbf{a}}$ peut être facilement établie : $\mathbf{C}\hat{\mathbf{a}}$ suit une loi normale de moyenne $\mathbf{Ca}$ et de matrice des variances et covariances [1] :

$$V(\mathbf{Ca}) = \mathbf{C}V(\hat{\mathbf{a}})\mathbf{C}' = \sigma^2\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}' \tag{36}$$

La densité de probabilité de $\mathbf{C}\hat{\mathbf{a}}$ s'écrit donc :

$$p(\mathbf{C}\hat{\mathbf{a}}) = (2\pi)^{-\frac{2}{r}} \left|\sigma^2\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}'\right|^{-\frac{1}{2}} \tag{37}$$
$$\exp\left\{-\frac{1}{2}[(\hat{\mathbf{a}} - \mathbf{a})'\mathbf{C}'[\sigma^2\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}']^{-1}\mathbf{C}(\hat{\mathbf{a}} - \mathbf{a})]\right\}$$

Dénotons par Q_c la forme quadratique qui figure entre crochets dans l'exponentielle. En tenant compte de ce que :

$$\hat{\mathbf{a}} - \mathbf{a} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u} \tag{38}$$

on peut écrire :

$$Q_c = \frac{1}{\sigma^2}\{\mathbf{u}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}'[\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}']^{-1}\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}\} \tag{39}$$

Posons :

$$\mathbf{N} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}'[\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}']^{-1}\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \tag{40}$$

d'où :

$$\sigma^2 Q_c = \mathbf{u}'\mathbf{N}\mathbf{u} \tag{41}$$

On vérifie par calcul direct que $\mathbf{N}$ est symétrique et idempotente. Son rang est donc égal à sa trace, or :

$$\text{tr } \mathbf{N} = \text{tr } \mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}'[\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}']^{-1} = \text{tr } \mathbf{I}r = r \tag{42}$$

---

(1) Il suffit d'écrire : $V(\mathbf{C}\hat{\mathbf{a}}) = E[\mathbf{C}(\hat{\mathbf{a}} - \mathbf{a})(\hat{\mathbf{a}} - \mathbf{a})'\mathbf{C}']$

En effet, la trace d'un produit est inchangée si on permute l'ordre des termes, à condition que les termes permutés restent conformes, ce qui est le cas ici.

La matrice **C** permet de décomposer l'espace à p dimensions des **X** en deux sous-espaces complémentaires : E_{p-r}, ensemble des vecteurs **Xa** tels que $\mathbf{Ca} = 0$ et E_r.

Considérons la matrice $\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, de format rn et de rang r. Ses r vecteurs lignes forment une base de E_r. Montrons que la matrice **N** définit une projection orthogonale sur E_r. Si un vecteur **z** est orthogonal à E_r, il sera tel que :

$$\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{z} = 0 \tag{43}$$

qui implique :

$$\mathbf{Nz} = 0 \tag{44}$$

L'image par **N** d'un vecteur orthogonal à E_r est donc bien nulle.

Si nous considérons la matrice **M** définie par la relation (7), le calcul direct montre que :

$$\mathbf{MN} = 0 \tag{45}$$

M et **N** projettent sur des variétés linéaires orthogonales. Ce résultat était évident, puisque **M** projette sur un espace orthogonal à l'espace des **X** et **N** sur un sous-espace des **X**.

Si **u** est le résidu du modèle (1), les deux vecteurs aléatoires :

$$\hat{\mathbf{u}} = \mathbf{Mu} \tag{46}$$

$$\hat{\mathbf{v}} = \mathbf{Nu} \tag{47}$$

appartiennent à des sous-espaces orthogonaux. Comme ils suivent l'un et l'autre une loi normale centrée, ils sont indépendants en probabilité.

Comme **N** est idempotente, elle peut être diagonalisée par une matrice orthogonale **S**, généralement différente de **W** :

$$\mathbf{S}'\mathbf{NS} = \mathbf{I}_r \tag{48}$$

Par raisonnement identique à celui qui nous a permis d'établir les lois de $\hat{\mathbf{y}}'\hat{\mathbf{y}}$ et de $\hat{\mathbf{u}}'\hat{\mathbf{u}}$, on peut écrire :

$$\sigma^2 Q_c = \mathbf{u}'\mathbf{Nu} \rightsquigarrow \sigma^2\chi^2_r \tag{49}$$

la distribution de $\sigma^2 Q_c$ étant indépendante de celle de $\hat{\mathbf{u}}'\hat{\mathbf{u}}$.

En général, on ne peut pas calculer $\sigma^2 Q_c$, qui dépend du paramètre inconnu **a** :

$$\sigma^2 Q_c = (\hat{\mathbf{a}} - \mathbf{a})\mathbf{C}'[\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}']^{-1}\mathbf{C}\ (\hat{\mathbf{a}} - \mathbf{a}) \tag{50}$$

Si, par contre,

$$\mathbf{Ca} = 0 \tag{51}$$

ce qui est justement l'hypothèse que l'on veut tester, $\sigma^2 Q_c$ s'écrit à partir de l'échantillon :

$$\sigma^2 Q_c = \hat{\mathbf{a}}'\mathbf{C}'[\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}']^{-1}\mathbf{C}\hat{\mathbf{a}} \tag{52}$$

ou, en remplaçant $\hat{\mathbf{a}}$ par sa valeur $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$:

$$\begin{aligned}\sigma^2 Q_c &= \mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}'[\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}']\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{y}'\mathbf{N}\mathbf{y} \\ &= \mathbf{y}'\mathbf{N}'\mathbf{N}\mathbf{y} = (\mathbf{N}\mathbf{y})'\mathbf{N}\mathbf{y}\end{aligned} \tag{53}$$

Les deux vecteurs :

$$\hat{\mathbf{u}} = \mathbf{M}\mathbf{y} \tag{54}$$

et

$$\hat{\mathbf{v}} = \mathbf{N}\mathbf{y} \tag{55}$$

sont la projection de $\mathbf{y}$ sur deux espaces orthogonaux :

— l'espace à $n - p$ dimensions des écarts $\hat{\mathbf{u}}$,

— le sous-espace à r dimensions de l'espace des $\mathbf{X}$ défini par la matrice $\mathbf{C}$.

On peut donc écrire :

$$\hat{\mathbf{v}}'\hat{\mathbf{v}} + \hat{\mathbf{u}}'\hat{\mathbf{u}} = \hat{\mathbf{w}}'\hat{\mathbf{w}} \tag{56}$$

si nous désignons par $\hat{\mathbf{w}}$ l'écart de $\mathbf{y}$ au sous-espace à $p - r$ dimensions qui contient $\mathbf{X}\mathbf{a}$ sous l'hypothèse $\mathbf{C}\mathbf{a} = 0$.

Le test de $\mathbf{C}\mathbf{a} = 0$ s'écrit finalement :

$$\frac{n-p}{r}\frac{\sigma^2 Q_c}{\mathbf{u}'\mathbf{M}\mathbf{u}} = \frac{\mathbf{u}'\mathbf{N}\mathbf{u}}{\mathbf{u}'\mathbf{M}\mathbf{u}} \times \frac{n-p}{r} = \frac{\sigma^2\chi_r^2}{\sigma^2\chi_{n-p}^2}\frac{n-p}{r} = F_{r,n-p} \tag{57}$$

Nous avons établi le test d'une hypothèse linéaire et homogène sur $\mathbf{a}$. Le test d'une hypothèse plus générale est facile à établir par un changement d'origine. Supposons, pour prendre un exemple simple, que nous voulions tester l'égalité du vecteur des r premières composantes de $\mathbf{a}$ à un vecteur $\mathbf{a}_0$ donné a priori. Partitionnons le vecteur $\mathbf{a}$ en deux vecteurs, $\mathbf{a}_1$, formé des r premières composantes de $\mathbf{a}_1$ et $\mathbf{a}_2$, formé des $p - r$ composantes restantes. Le modèle (1) s'écrit :

$$\underset{(n,\,p)}{\mathbf{y}} = \underset{(n,\,r)}{\mathbf{X}_1}\,\underset{(r,\,1)}{\mathbf{a}_1} + \underset{(n,\,p-r)}{\mathbf{X}_2}\,\underset{(p-r,\,1)}{\mathbf{a}_2} + \underset{(n,\,1)}{\mathbf{u}} \tag{58}$$

Le vecteur $\mathbf{a}_0$ étant connu, on peut écrire :

$$\mathbf{y} - \mathbf{X}_1\mathbf{a}_0 = \mathbf{X}_1(\mathbf{a}_1 - \mathbf{a}_0) + \mathbf{X}_2\mathbf{a}_2 + \mathbf{u} \tag{59}$$

Il suffit de tester, de la façon établie précédemment :

$$\mathbf{a}_1 - \mathbf{a}_0 = 0 \tag{60}$$

5. INTERPRÉTATION GÉOMÉTRIQUE

La forme quadratique $\sigma_2 Q_c$, donnée par la formule (53) peut paraître laborieuse à calculer. Mais le vecteur $\hat{\mathbf{v}}$ défini en (55) a une interprétation géométrique simple, qui aide à le calculer.

Nous appelerons « ajustement selon le modèle » la projection $\hat{\mathbf{y}}$ de $\mathbf{y}$ sur la variété linéaire des $\mathbf{X}$. De même, appelons « ajustement sous l'hypothèse » l'explication linéaire de $\mathbf{y}$ par les $\mathbf{X}$ si l'hypothèse $\mathbf{Ca} = 0$ est vérifiée. Cet ajustement est représenté par le vecteur $\mathbf{y}^*$, projection orthogonale de $\mathbf{y}$ sur la variété linéaire à $p - r$ dimensions, appartenant à la variété des $\mathbf{X}$ et orthogonale aux r vecteurs lignes de $\mathbf{C}$.

$\mathbf{y}^*$ et $\hat{\mathbf{v}}$ sont les projections de $\mathbf{y}$ sur deux sous-espaces complémentaires et orthogonaux de l'espace des $\mathbf{X}$. On peut écrire :

$$\hat{\mathbf{y}} = \mathbf{y}^* + \hat{\mathbf{v}} \tag{61}$$

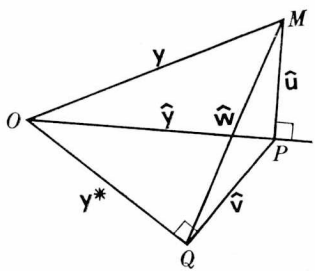
$$\hat{\mathbf{y}}'\hat{\mathbf{y}} = \mathbf{y}^{*\prime}\mathbf{y}^* + \hat{\mathbf{v}}'\hat{\mathbf{v}} \tag{62}$$

$\hat{\mathbf{v}}$ s'interprète donc simplement comme la différence entre l'explication linéaire $\hat{\mathbf{y}}$ de $\mathbf{y}$ dans le modèle général et l'explication linéaire $\mathbf{y}^*$ de $\mathbf{y}$ sous l'hypothèse $\mathbf{Ca} = 0$. Dans la plupart des applications concrètes, la matrice $\mathbf{C}$ est telle que $\mathbf{y}^*$ peut être calculé sans difficulté.

Pour donner un support intuitif à l'interprétation géométrique que nous venons d'établir, nous pouvons construire une figure dans le cas (qui ne peut évidemment avoir qu'un caractère d'illustration) où :

$$n = 3 \qquad p = 2 \qquad r = 1$$

$\overrightarrow{OM} = \mathbf{y}$
$\overrightarrow{OP} = \hat{\mathbf{y}}$
$\overrightarrow{OQ} = \mathbf{y}^*$
$\overrightarrow{MP} = \hat{\mathbf{u}}$
$\overrightarrow{MQ} = \hat{\mathbf{w}}$
$\overrightarrow{PQ} = \hat{\mathbf{v}}$



Les triangles OMP et OPQ sont rectangles respectivement en P et Q, d'où :

$$MQ^2 = MP^2 + PQ^2$$

$$OP^2 = OQ^2 + PQ^2$$

6. EXEMPLES

a) **Test de l'absence d'effet des $p - 1$ vraies variables exogènes dans un modèle de régression**

Modèle :

$$\mathbf{y} = \mathbf{Xa} + \mathbf{u} \tag{1}$$

Hypothèse :

(63) $$a_j = 0 \forall j \neq p$$

donc $\mathbf{y}^*$ s'obtient en ajustant le modèle très simple

(64) $$\mathbf{y} = a_p \mathbf{i} + \mathbf{w}$$

L'estimation $\hat{a}_p$ de a_p est égale à $\bar{y}$ et

(65) $$\mathbf{y}^* = \bar{y}\,\mathbf{i}$$

où $\mathbf{i}$ est le vecteur de l'espace des $\mathbf{X}$ dont toutes les composantes sont égales à l'unité. En effet,

(66) $$\hat{a}_p = (\mathbf{i}'\mathbf{i})^{-1}\mathbf{i}'\mathbf{y} = \frac{1}{n}\sum y = \bar{y}$$

Le test s'écrit :

(67) $$\frac{(\hat{\mathbf{y}} - y\mathbf{i})'(\hat{\mathbf{y}} - \bar{y}\mathbf{i})}{p-1} \cdot \frac{n-p}{\hat{\mathbf{u}}'\hat{\mathbf{u}}} = F_{p-1,\ n-p}$$

Or, un résultat classique de la théorie des régressions permet d'écrire, si R est le coefficient de corrélation multiple :

(68) $$\frac{(\hat{\mathbf{y}} - \bar{y}\mathbf{i})'\,(\hat{\mathbf{y}} - \bar{y}\mathbf{i})}{(\mathbf{y} - \bar{y}\mathbf{i})'\,(\mathbf{y} - \bar{y}\mathbf{i})} = R^2$$

(69) $$\frac{\mathbf{u}'\hat{\mathbf{u}}}{(\mathbf{y} - y\mathbf{i})'(y - y\mathbf{i})} = 1 - R^2$$

(67) prend alors la forme très simple :

(70) $$\frac{R^2}{1-R^2}\,\frac{n-p}{p-1} = F_{p-1,\ n-p}$$

b) **Test de l'absence d'interaction dans un modèle d'analyse de variance à deux facteurs**

Les données sont classées dans un tableau à p lignes et q colonnes. On a n observations en tout. Nous noterons j la ligne, k la colonne et i une observation dans la case jk. Le modèle avec interaction s'écrit :

(71) $$y_{ijk} = m + a_j + b_k + c_{jk} + u_{ijk}$$

où les u_{ijk} sont des résidus qui suivent une loi normale de moyenne nulle, de variance σ^2 indépendante de i, j, k, et sans autocorrélation.

Le modèle sans interaction s'écrit :

(72) $$y_{ijk} = m + a_j + b_k + w_{ijk}$$

où les w_{ijk} ont les mêmes propriétés que les u_{ijk}.

On sait que les estimations de m, a_j et b_k sont les mêmes dans le modèle (71) et le modèle (72). On peut donc écrire :

$$\hat{y}_{jk} = \hat{m} + \hat{a}_j + b_k + \hat{c}_{jk} \tag{73}$$

$$y^*_{jk} = \hat{m} + \hat{a}_j + \hat{b}_k \tag{74}$$

donc
$$\hat{\mathbf{v}} = \hat{\mathbf{y}} - \mathbf{y}^* = \hat{\mathbf{c}} \tag{75}$$

avec (nous ne donnons pas les démonstrations)

$$\hat{c}_{jk} = y_{.jk} - y... - \hat{a}_j - \hat{b}_k \tag{76}$$

$\hat{a}_j$ et $\hat{b}_k$ étant, dans le cas général, solutions du système à $p + q$ équations

$$\begin{cases} \hat{a}_j + \sum_k \dfrac{n_{jk}}{n_{j.}} \hat{b}_k = y_{.j.} - y... \\ \hat{b}_k + \sum_j \dfrac{n_{jk}}{n_{.k}} \hat{a}_j = y_{.k.} - y... \end{cases} \tag{77}$$

où n_{jk}, $n_{j.}$ et $n_{.k}$ désignent respectivement le nombre d'observations dans la case jk dans la ligne j et dans la colonne k.

Dans le cas où les facteurs sont orthogonaux, (77) s'écrit plus simplement :

$$\begin{cases} \hat{a}_j = y_{.j.} - y... \\ \hat{b}_k = y_{..k} - y... \end{cases} \tag{78}$$

On sait que les termes d'interaction sont liés par $p + q - 1$ relations linéaires indépendantes.

Le vecteur $\hat{\mathbf{v}} = \hat{\mathbf{c}}$ a donc $(p-1)(q-1)$ dimensions. Le vecteur des écarts $\hat{\mathbf{u}}$ du modèle (71) a $n - pq$ dimensions. Le test d'absence d'interaction s'étudie à partir de :

$$\frac{\sum_j \sum_k c_{jk}^2}{(p-1)(q-1)} \frac{n-pq}{\hat{\mathbf{u}}'\hat{\mathbf{u}}} = F_{(p-1)(q-1),\ n-pq} \tag{79}$$

où $\hat{\mathbf{u}}'\hat{\mathbf{u}}$ est la variance résiduelle dans l'estimation du modèle (71).

CONCLUSION

De précédents articles de cette revue (1) ont montré que des notions élémentaires d'algèbre linéaire permettaient un exposé simple et général des principales méthodes d'analyse des données. Nous avons montré ici que ces mêmes notions se prêtaient à un exposé à la fois élémentaire et rigoureux de test de l'hypothèse linéaire générale dont la présentation classique est plus laborieuse. Ce résultat a principalement un intérêt pédagogique, en permettant à des lecteurs dont la formation en statistique mathématique est élémentaire de déborder le raisonnement approché par analogie qui est encore souvent présenté. Dans les utilisations concrètes, les formules générales que nous avons établies seraient d'application maladroite. Le calcul direct des éléments des deux vecteurs $\hat{u}$ et $\hat{v}$ est toujours plus rapide. La seule difficulté que peuvent rencontrer des débutants réside en la détermination de la dimension de ces deux vecteurs, qui fournit les nombres de degrés de liberté à prendre en considération.

---

(1) Cf. C. DENIAU et L. LEBART : **Introduction à l'analyse des données,** n° 3 et 4/1969.

BIBLIOGRAPHIE

PAGNY (F.) — **La stratégie des produits dans l'entreprise,** Dunod.

Les complexités de la vie industrielle, d'une part, l'incertitude de la demande des consommateurs, d'autre part, orientent de plus en plus l'analyse économique vers le « produit nouveau » dont les dirigeants d'entreprise savent par avance qu'ils auront finalement une vie assez brève.

L'ouvrage de Françoise PAGNY étudie les différentes données permettant d'appréhender l'influence de la production et de la consommation sous l'angle des caractéristiques des produits et de leurs relations (substitution ou complémentarité dans la gamme en particulier) ce qui détermine le cycle de vie.

De ce choix du produit nouveau, découlent les politiques commerciales de production et d'investissement dans lequel le cycle de vie devient une variable que le calcul économique peut alors étudier.

Jean TABOULET

BERRY (Brian J. L.). — **Géographie des marchés et du commerce de détail.** Traduction de Bernard Marchand. Collection U 2, Armand Colin.

Si un certain nombre de praticiens des études de marché ou de chalandise d'un point de vente s'interrogent sur la validité des instruments utilisés pour passer des données globales relevant de la macro-économie à des données détaillées caractéristiques de la micro-économie, cet ouvrage devrait leur donner les moyens de dissiper leurs scrupules de nature méthodologique.

En effet, « la nouvelle géographie » a mis au point la théorie des places centrales ou théorie de la localisation, de la nature de l'espacement des groupes d'activité. Ceux-ci résultent d'une division du travail permettant de différencier le bloc (ou groupes d'immeuble), le quartier et le centre commercial et par conséquent de hiérarchiser leur importance par rapport aux consommateurs ainsi qu'à la nature de leurs dépenses.

L'auteur peut ensuite, par l'examen de situations géographiques dans Iowa, à Chicago où il est Professeur, puis par des modèles, expliquer le rôle des places centrales en économies complexes et les variations de leurs hiérarchies. C'est ainsi qu'on peut alors parler pour les magasins de la localisation 100 % pour déterminer la valeur de leur implantation par rapport au quartier central des affaires.

Un peu avant la seconde guerre mondiale, sont apparus les nouveaux centres commerciaux de périphérie liés à la décadence graduelle de l'habitat en centre ville aux États-Unis, dont les causes sont très clairement expliquées. Les migrations de populations venues du Sud à faibles revenus, une dégradation dans l'entretien des immeubles, entraînent un rétrécissement du commerce de détail et provoquent l'évasion des commerces les plus solides vers l'extérieur.

A mesure que le nombre des nouveaux centres augmente, leurs spécialisations s'accentuent ce qui permet de retrouver les hiérarchies commerciales en fonction de leur importance.

Plus récemment, l'application de « la nouvelle géographie » encore appelée géographie du marketing, a permis de mettre au point des outils efficaces d'évaluation qui sont utilisés pour l'implantation des centres commerciaux et la planification des villes, qu'il s'agisse de rénovations ou de créations.

Jean TABOULET

L'HARDY (P.). — **Structure de l'épargne et du patrimoine des ménages en 1966** (Collections de l'INSEE M 13).

Après une première publication de ses résultats (Collections de l'INSEE M 6), l'enquête INSEE 1967 sur l'épargne des ménages salariés ou inactifs fournit la matière d'un nouveau dossier présenté par Philippe L'HARDY.

Ce dossier constituera un document de travail extrêmement précieux pour le chercheur : il comporte en effet une énorme masse de données brutes sous la forme de tableaux qui complètent ceux de la publication précédente en indiquant : taux de possession des divers actifs étudiés, répartition des ménages possesseurs, répartition des montants des actifs, etc... suivant les trois critères habituels (revenu, âge et catégorie socio-professionnelle). De plus, le dossier comporte aussi un véritable guide pour l'exploitation de cette masse de données : il facilite l'interprétation des trois critères d'étude (dont les plus influents sont l'âge et le revenu) et il présente cinq exemples d'utilisation possibles. De ces exemples, nous en retiendrons deux, du point de vue de l'intérêt méthodologique :

— la mise en évidence de l'influence spécifique sur la valeur du logement des ménages, de l'appartenance à la catégorie « cadre supérieur », une fois tenu compte de l'influence du revenu : cette mise en évidence s'opère par la reconstitution, à partir des revenus des cadres supérieurs, d'une distribution théorique des valeurs du logement et par la comparaison de cette distribution reconstituée avec la distribution observée.

— l'étude de la concentration des valeurs mobilières montrant que « ce qui distingue « actions » et « emprunts et obligations » n'est pas une différence d'homogénéité entre les répartitions dans la population classée selon le montant de l'avoir détenu, mais essentiellement le degré de liaison de ce montant avec le revenu ».

Le dossier présente toutefois des lacunes qui sont dues aux lacunes de l'enquête : la valeur explicative de la catégorie socio-professionnelle est assez limitée dans une population comprenant uniquement des salariés et des inactifs ; le fait que l'enquête ne fournisse qu'une « coupe instantanée » interdit toute distinction entre l'effet cycle de vie et l'effet génération qui se retrouvent tous deux dans l'influence de l'âge. Le défaut le plus grave, à notre sens, réside dans l'impossibilité de prendre en compte le critère fortune totale (en raison de « l'omission dans l'enquête de différentes parties du patrimoine » et à cause de la sous-estimation de certains postes) ; cette lacune réduit sensiblement le champ des recherches possibles qui reste cependant très large: la parution, dans le courant de l'année 1972, d'études portant sur différents thèmes est annoncée et attendue avec grand intérêt.

Patrice LANCO

Le directeur de la publication : P. BORDAS.

Dépôt légal : 3e trimestre 1972. Numéro 7405. Imprimé en France.

Imprimerie Nouvelle, Orléans. — N° 6598.

CONSOMMATION (ANNALES DU C. R. E. D. O. C.)

1967

N° 1. — Une étude économétrique de la demande de viande. — La consommation des Français en 1965. — Intégration des méthodes d'approche psycho-sociologiques à l'étude de l'épargne.

N° 2. — Un indicateur de la morbidité appliqué aux données d'une enquête sur la consommation médicale. — La diffusion des services collectifs : phénomène économique ou social ? — Les travaux de préparation du V[e] Plan et l'élaboration d'un modèle national de fonctionnement du marché du logement. — Les conditions de vie des familles.

N° 3. — L'épargne des exploitants agricoles. — Structure et équilibre du marché du textile. — Les dépenses touristiques.

N° 4. — L'appareil commercial et les circuits de distribution. — Le développement de la radiologie.

1968

N° 1. — Étude critique de méthodes d'enquête. — Étude sur l'offre et la demande de créance.

N° 2. — Théorie et politique de l'épargne. — Un modèle prévisionnel de la demande de logements. — L'évolution de la consommation de viande.

N° 3. — La consommation et la demande de monnaie. — Valeur prédictive des intentions d'achats au niveau du ménage pris individuellement.

N° 4. — Quelques éléments sur le comportement des propriétaires vis-à-vis du prix du logement acheté et de la mise de fonds versée. — Facteurs « irrationnels » de l'offre d'épargne (recherches allemandes).

1969

N° 1. — L'offre de monnaie par les banques commerciales. — L'économie des services de soins médicaux en France. — L'évolution de la consommation de produits laitiers de 1950 à 1966.

N° 2. — L'économie des services de soins médicaux en France. — La formation de l'épargne liquide (l'exemple du Crédit Mutuel). — Consommation individuelle et consommation collective. — Étude sur la demande en logement des ménages.

N° 3. — Les prix de détail en France par rapport aux autres pays de la Communauté. — La consommation des ménages en France et en Hongrie. — Introduction à l'analyse des données.

N° 4. — Durée d'observation et précision dans les enquêtes de consommation. — Un essai de classification de titres boursiers fondée sur l'analyse factorielle. — Introduction à l'analyse des données.

1970

N° 1. — La fréquentation des équipements collectifs. — La supériorité de la gestion collective de l'épargne mobilière : analyse méthodologique et application aux SICAV. — Le comportement des exploitants agricoles en Eure-et-Loir et en Ille-et-Vilaine.

N° 2-3. — L'Évolution de la consommation des ménages de 1959 à 1968.

N° 4. — Les services médicaux en Suède et en France. — Proposition pour une méthodologie de l'étude de la redistribution. — La consommation des boissons dans quelques pays d'Europe.

1971

N° 1. — Les familles devant l'éducation des enfants. — Nouvelle évaluation de la fortune des ménages (1959-1967). — Budget-temps et choix d'activité.

N° 2. — Enquête sur les loisirs et mode de vie du personnel de la Régie Nationale des Usines Renault. — Étude des effets différentiels des impôts sur la consommation. — La morphologie sociale des communes urbaines.

N° 3. — La consommation élargie. — Étude économique de l'activité des médecins. — Possibilités et difficultés de la régulation des problèmes d'environnement et de nuisance par entente spontanée entre les intéressés.

N° 4. — Nature et prix des soins médicaux en ville. — Quelques résultats de l'étude des bilans de petites et moyennes entreprises.

1972

N° 1. — Enquête sur les loisirs et mode de vie du personnel de la Régie Nationale des Usines Renault. — Les choix de consommation et les budgets des ménages. — Placement et Investissement. — Les budgets familiaux dans les régions de la C.E.E.

SOMMAIRE DES PROCHAINS NUMÉROS

Le système d'Indicateurs du VI[e] Plan. Recherche de projections cohérentes pour des variables interdépendantes. L'arbitrage entre salaire et temps libre : une application de l'analyse d'indifférence. Consommation des ménages, séries chronologiques.

sommaire

CENTRE DE RECHERCHES ET DE DOCUMENTATION SUR LA CONSOMMATION

45, boulevard de la Gare, PARIS-13e

Tél. POR. 97-59

1972 n° 2

Avril Juin