CAHIERS DU BURO

J. BRENOT
P. CAZES
N. LACOURLY

Pratique de la régression : qualité et protection

Cahiers du Bureau universitaire de recherche opérationnelle. Série Recherche, tome 23 (1975), p. 3-84

<http://www.numdam.org/item?id=BURO_1975__23__3_0>

PRATIQUE DE LA RÉGRESSION : QUALITÉ ET PROTECTION

J. BRENOT *, P. CAZES **, N. LACOURLY *

Cette monographie n'est pas un cours.

On y insiste sur certains aspects de la théorie ou de la pratique de la régression. Certains points sont seulement soulignés, d'autres omis. Ainsi, les résultats classiques de la théorie des moindres carrés sont supposés connus.

Pour les démonstrations, le lecteur est souvent invité à consulter une référence.

Les auteurs remercient J.P. PAGES dont la contribution à la fois critique et fructueuse dans une ambiance amicale leur a été précieuse.

(*) Laboratoire de Statistique et d'Analyse Quantitative. Commissariat à l'Energie Atomique.

(**) Laboratoire de Statistique et ISUP. Université Pierre et Marie Curie (Paris VI).

TABLE DES MATIÈRES

Pages

INTRODUCTION 07

I. MODELE LINEAIRE ET REGRESSION LINEAIRE : GENERALITES

1.1 INTRODUCTION – NOTATIONS 09

1.2 SOLUTION DES MOINDRES CARRES 12

1.3 CHOIX DE LA METRIQUE N DANS F – THEOREME DE GAUSS-MARKOV 14

1.3.1. Théorème fondamental sur les formes quadratiques 14
1.3.2. Inégalités entre formes quadratiques semi-définies positives 14
1.3.3. Le théorème de GAUSS-MARKOV 15

1.3.3.1. Le théorème 16
1.3.3.2. Les conséquences du théorème 17
1.3.3.3. Rappel : forme classique du théorème de GAUSS-MARKOV 19
1.3.3.4 Est-il possible de choisir $N = \Gamma^{-1}$? 20

1.4 DISTINGUO ENTRE "MODELE LINEAIRE" ET "REGRESSION LINEAIRE" 20

1.5 QUALITE DE LA SOLUTION $\hat{y}$ 21

1.5.1. Optique : Régression entre variables aléatoires 21
1.5.2. Optique : Modèle linéaire 22

CONCLUSION 26

II. EXAMEN CRITIQUE DU MODELE : ANALYSE DES RESIDUS

INTRODUCTION 27

2.1 VISUALISATION DES RESIDUS 28

2.2 UTILISATION DE TESTS NON-PARAMETRIQUES 28

2.3 CONTROLE DE LA "NON-CORRELATION DES ERREURS" .. 29

Pages

2.3.1. Introduction : présentation du problème 29
2.3.2. Conséquences dues à la corrélation des erreurs 30
2.3.3. Procédure de Durbin et Watson 31
2.3.4. La méthode "BLUS" 34
2.3.5. Un exemple 37
2.3.6. Conclusion 44

III. PROTECTION DE LA REGRESSION : SELECTION DE VARIABLES

3.1 Le PAS à PAS 45
3.1.1. Examen de toutes les régressions 45
3.1.2. Elimination "descendante" des variables 45
3.1.3. Introduction "ascendante" des variables 46
3.1.3.1. La régression ascendante 46
3.1.3.2. La régression progressive ("Stepwise") 47
3.1.3.3. Un autre critère pour "l'introduction ascendante": le critère C_q 47
3.1.4. Conclusion 49

3.2 REGRESSION SUR VARIABLES ORTHOGONALES 50
3.2.1. Exemple : "Etude des bénéfices de l'instruction" 51
3.2.2. Stratégie 52

IV. PROTECTION DE LA REGRESSION PAR UTILISATION DE CONTRAINTES

4.1 INTRODUCTION 55
4.2 REGRESSION SOUS CONTRAINTES LINEAIRES 55
4.2.1. Le problème 55
4.2.2. La solution 57

4.3 La "RIDGE REGRESSION" 57
4.3.1. Le problème de la quasi-colinéarité des variables explicatives 57
4.3.2. La solution 59

4.4 REGRESSION SUR VARIABLES ENTACHEES D'ERREURS .. 61
4.4.1. Le problème 61
4.4.2. La solution 63

Pages

4.5 EXEMPLE D'APPLICATION DE LA REGRESSION SOUS CONTRAINTES 64

ANNEXE 70

V. REGRESSION PAR BOULES ET REGRESSION PAR ANALYSE DES CORRESPONDANCES

5.1 INTRODUCTION 76

5.2 REGRESSION PAR BOULES (ou par VOISINAGES) 76

5.3 REGRESSION PAR ANALYSE DES CORRESPONDANCES ... 77

BIBLIOGRAPHIE 81

OPTIMISATION ET GRAPHES – APPLICATIONS A L'ARCHITECTURE – par Patrick LEROY 85

INTRODUCTION

Un modèle de "régression linéaire" est utilisé pour traiter des problèmes de nature parfois très diverse.

La diversité existe :

• **au niveau des variables considérées** :

– La "variable à expliquer" peut être soumise à des aléas qui interviendront soit globalement sur l'ensemble des observations, soit de façon différente sur chacune des observations ;

– Les "variables explicatives" peuvent être "certaines" ou aléatoires ; elles peuvent être saisies directement ou par l'intermédiaire d'indicateurs ; elles peuvent être très corrélées entre elles ou, au contraire, orthogonales ; elles peuvent résulter du codage de variables qualitatives (analyse de covariance) ;

• **au niveau des objectifs**

– soit, il s'agit de *prévoir* un phénomène et le modèle sera considéré comme un bon modèle de prévision, si, ajusté sur le passé, il se révèle précis (bonne reconstruction de la variable à expliquer) et stable (les estimations des paramètres du modèle évoluent peu avec le temps par exemple) ;

– soit, il s'agit de *quantifier l'effet d'une variable* dans un modèle qu'il est justifié de poser car il exploite la nature linéaire des liaisons entre les variables ; le modèle sera bon s'il est précis et fiable (les estimations des paramètres dépendent peu des "aléas d'échantillonnage") ;

– soit, il s'agit de mieux comprendre un phénomène (telle variable a-t-elle une influence sur telle autre ?) ; le modèle posé, dit modèle "explicatif", n'a d'intérêt que s'il renseigne sur une *dépendance ;* ici, plus que la précision du modèle, c'est la contribution "significative" de certaines des variables explicatives à la reconstruction de la variable à expliquer qui intéresse l'analyste ;

– soit, il s'agit de résoudre un système surabondant d'équations linéaires ;

etc.

Il est naturel de tenir compte de toute l'information disponible pour estimer les paramètres du modèle ; cette information peut être introduite à différents niveaux :

– au niveau des coefficients sous forme de contraintes ;

– au niveau du critère par le choix d'un critère autre que le critère classique des moindres carrés par exemple.

Introduire cette information revient à associer au critère des moindres carrés classique un ensemble de "garde-fous" qui permettront dans certains cas d'obtenir des coefficients interprétables, même si le nombre d'observations est relativement petit par rapport au nombre de variables explicatives ou si ces variables sont fortement corrélées entre elles.

L'objet de cette monographie est de présenter quelques techniques où les soucis précédents interviennent.

Le premier chapitre a pour but d'introduire aux notations et rappelle les résultats classiques de la théorie ; on insiste sur le théorème de GAUSS-MARKOV dont on donne une version totalement algébrique. La lecture de ce chapitre peut être omise par le lecteur désireux des seules techniques ou pour lequel les notions "modèle linéaire" et "régression linéaire" ne prêtent pas à confusion.

Apprécier la précision d'un modèle et la stabilité des estimations de ses paramètres nécessite l'emploi des résidus et leur analyse. L'analyse graphique des résidus, leur utilisation pour contrôler un corps d'hypothèses posé a priori font l'objet du chapitre deuxième.

Quand le nombre de variables explicatives est grand par rapport au nombre d'observations ou quand les variables explicatives sont très corrélées, la précision du modèle est souvent illusoire. Il faudra donc se protéger lors de l'application de la technique des moindres carrés.

Le troisième chapitre rappelle qu'il est possible de se protéger simplement par sélection de variables. On présente les procédures de pas à pas et la régression sur variables orthogonales.

Dans le quatrième chapitre, on expose quelques techniques de protection basées sur l'introduction de contraintes : contraintes linéaires, contrainte sur la norme du vecteur des coefficients de régression, contrainte sur la variance résiduelle de la régression.

Dans le dernier chapitre, on proprose deux méthodes pragmatiques pour la mise en évidence de liaisons fonctionnelles. La première est basée sur l'extension du rapport de corrélation, la seconde sur l'Analyse des Correspondances.

I – MODÈLE LINÉAIRE ET RÉGRESSION LINÉAIRE : GÉNÉRALITÉS

1.1. INTRODUCTION – NOTATIONS

Dans ce chapitre 1, on abordera l'étude du Modèle Linéaire, principalement sous l'aspect algébrique de l'Analyse de Données.

Dans le cadre du Modèle Linéaire, on est en présence :

– d'une variable quantitative y, qu'on appelle la "variable à expliquer",

– d'un ensemble de p variables $\{x^j / j = 1, \ldots, p\}$ appelées "variables explicatives".

Chacune de ces variables prenant n valeurs, on considère le modèle :

$$\boxed{\underline{y} = \sum_{j=1}^{p} \beta_j \underline{x}^j + \underline{\varepsilon} = X'\underline{\beta} + \underline{\varepsilon}}$$

où :

$\underline{y} \in R^n$ est le vecteur des n valeurs y_i prises par la variable y,

$\underline{x}^j \in R^n$ est le vecteur des n valeurs x_i^j prises par la variable $x^j (j = 1, p)$

$X' = X'\ (n \times p) = (\underline{x}^1\ \underline{x}^2 \ldots \underline{x}^p)$

$\underline{\beta} \in R^p$ est le vecteur des coefficients du modèle qu'il s'agit d'estimer et $\underline{\varepsilon}$ est le vecteur des erreurs associées au modèle.

Par convention, on note $\underline{x}'$ le vecteur transposé d'un vecteur $\underline{x}$, et A' l'application transposée d'une application linéaire A.

Si le modèle posé se révèle "bon" (les erreurs sont petites et les estimations des coefficients sont "stables"), il pourra être utilisé par exemple pour une prévision de la variable y.

Ajuster au mieux un vecteur de la forme $X'\underline{b}$, où $\underline{b}' = (b_1, b_2, \ldots, b_p)$, au vecteur $\underline{y}$ à expliquer est le premier objectif du traitement mathématique.

Le vecteur $X'\underline{b}$ obtenu servira comme approximation du vecteur $\underline{g} = X'\underline{\beta}$ qui est inconnu.

On a noté X,

$$X = \begin{pmatrix} x_1^1 & x_2^1 & \dots & x_n^1 \\ x_1^2 & x_2^2 & \dots & x_n^2 \\ \dots & \dots & \dots & \dots \\ x_1^p & x_2^p & \dots & x_n^p \end{pmatrix} = X\ (p \times n)$$

le tableau de données quantitatives relatives aux variables explicatives.

Pour ce tableau, deux représentations peuvent être considérées :

– une représentation du nuage $\mathfrak{N}$ des n "points-individus" dans l'espace vectoriel $E = R^p$

$$\mathfrak{N} = \{\underline{x}_i / i = 1, \dots, n\} \quad \text{avec} \quad \underline{x}_i \in R^p \text{ de composantes } (x_i^1, \dots, x_i^p)$$

E est "l'espace des individus".

– une représentation du nuage $\mathfrak{N}$ des p "points-caractères" dans l'espace vectoriel $F = R^n$

$$\mathfrak{N} = \{\underline{x}^j / j = 1, \dots, p\} \quad \text{avec} \quad \underline{x}^j \in R^n$$

F est "l'espace des caractères".

Si on choisit comme bases respectives de E et F, les bases canoniques

$$B_1 = \{\underline{e}_j / j = 1, \dots, p\} \quad \text{et} \quad B_2 = \{\underline{f}_i / i = 1, \dots, n\}$$

ainsi que les bases duales

$$B_1^* = \{\underline{e}_j^* / j = 1, \dots, p\} \quad \text{et} \quad B_2^* = \{\underline{f}_i^* / i = 1, \dots, n\}$$

dans les espaces duaux E^* et F^* de E et F, le tableau X peut être considéré comme la matrice associée à l'application (notée aussi X) de F^* dans E définie par :

$$X(\underline{f}_i^*) = \underline{x}_i$$

De même l'application transposée X' de X, application de E^* dans F, est définie par :

$$X'(\underline{e}_j^*) = \underline{x}^j$$

On peut dire que les points $\underline{x}^j$ et $\underline{e}_j^*$ $(j = 1, 2, \dots p)$, représentent les caractères tandis que les points $\underline{x}_i$ et $\underline{f}_i^*$ représentent les individus $(i = 1, 2, \dots n)$.

Soit N, une métrique euclidienne sur l'espace F :

$$N : F \longrightarrow F^* ,$$

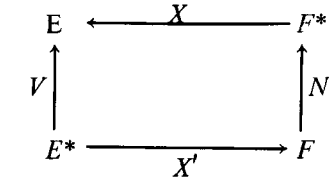
N définit une distance entre les caractères(*). Il est alors naturel de choisir sur E^*, une forme quadratique V conservant les distances entre les points-caractères $\underline{x}^j$, images par l'application X' des points $\underline{e}_j^*$; ceci revient à poser :

$$V(\underline{e}_j^*, \underline{e}_k^*) = N(\underline{x}^j, \underline{x}^k) = N(X'\underline{e}_j^*, X'\underline{e}_k^*)$$

pour $j = 1, 2, \ldots p$ et $k = 1, 2, \ldots p$

soit $$V = X . N . X'$$

On a finalement le schéma de dualité suivant (cf. C 3 E (11))



$\underline{\beta}, \underline{b}$ $\qquad\qquad$ $\underline{y}, X'\underline{\beta}, X'\underline{b}$

Le carré de la norme d'un vecteur $\underline{x}$, $N(\underline{x}, \underline{x})$ sera indifféremment noté $\|\underline{x}\|_N^2$ ou $N(\underline{x})$.

Remarques :

1. rang (V) = rang (X')
2. $\underline{\beta}$ et $\underline{b}$ sont des éléments de E^*
3. Si $N = D_p = \begin{pmatrix} p_1 & & 0 \\ & p_2 & \\ 0 & & p_n \end{pmatrix}$ où p_i est le poids affecté à

l'observation i avec $p_i \geqslant 0$ et $1 = \sum_{i=1}^{n} p_i$,

et si le centre de gravité de $\mathfrak{M} = \{\underline{x}_i / i = 1, n\}$ est l'origine de $E = R^p$, c'est-à-dire si $\sum_{i=1}^{n} p_i \underline{x}_i = \underline{0}$

(*) Se donner une métrique euclidienne sur F revient à se donner :

- une forme bilinéaire symétrique définie positive sur F (application de $F \times F$ dans R^+) ;
- ou bien son isomorphisme associé (application linéaire bijective de F sur F^*).

alors, $V = X D_p X'$ est la forme quadratique d'inertie (f.q.i) du nuage $\mathfrak{N}$.

Si on considère $\mathfrak{N}$ comme un échantillon de taille n du p-uple $(x^1, x^2, \ldots, x^p)$, $X D_p X'$ est la matrice de variance-covariance empirique de la variable p-dimensionnelle $(x^1, x^2, \ldots, x^p)$

$$V = X D_p X' = \text{var}((x^1, x^2, \ldots, x^p))$$

1.2 SOLUTION DES MOINDRES CARRES

Pour le modèle $\underline{y} = X'(\underline{\beta}) + \underline{e}$, le premier objectif est de reconstruire "au mieux" le vecteur $\underline{y}$ à l'aide d'un vecteur de la forme $X'(\underline{b})$.

Si on se place dans l'espace des caractères $F = R^n$, cela revient à chercher le point $\hat{\underline{y}} \in W$ le plus proche possible de $\underline{y}$ au sens de la métrique N définie sur F, W étant la variété linéaire engendrée par les vecteurs $\underline{x}^j$.

$$W = \{\underline{z} = X'(\underline{b}) \quad \text{où} \quad \underline{b} \in E^*\} \subset R^n$$

Le point $\hat{\underline{y}}$ solution est le point de W vérifiant :

$$\|\underline{y} - \hat{\underline{y}}\|_N^2 = \text{Inf}\{\|\underline{y} - \underline{z}\|_N^2 \quad \text{où} \quad \underline{z} \in W\}$$

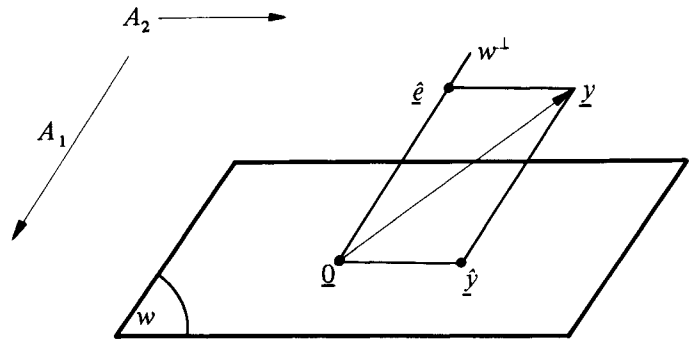
$\hat{\underline{y}}$ est donc la projection N – orthogonale de $\underline{y}$ sur W.

Dans la décomposition en somme directe :

$$F = W \oplus W^{\perp} \tag{1}$$

où $W^{\perp}$ est le sous-espace supplémentaire N-orthogonal à W dans F, on a la représentation *unique* :

$$\underline{y} = \hat{\underline{y}} + \hat{\underline{e}} \quad \text{avec} \quad N(\hat{\underline{y}}, \hat{\underline{e}}) = 0 .$$



Si A_1 et A_2 sont les projecteurs associés à la décomposition (1) :

$$\underline{\hat{y}} = A_1(\underline{y}) \qquad \text{et} \qquad \underline{\hat{e}} = A_2(\underline{y})$$

A_1 et A_2 ont les propriétés suivantes :

- $A_1 + A_2 = I_n$ (application identité dans F) ;
- A_1 et A_2 sont idempotents : $A_i \cdot A_i = A_i$ $(i = 1, 2)$
- A_1 et A_2 sont "N-symétriques" : $(N \cdot A_i)' = N \cdot A_i = A_i' \cdot N$ $(i = 1, 2)$

Toute forme linéaire $\underline{\hat{b}} \in E^*$, vérifiant :

$$X'(\underline{\hat{b}}) = \sum_{j=1}^{p} \hat{b}_j \, \underline{x}^j = \underline{\hat{y}} \quad \text{est une solution au problème posé}$$

$\mathfrak{C}_{\underline{\hat{b}}} = \{\underline{b} \in E^* / X'(\underline{b}) = X'(\underline{\hat{b}})\}$ est l'ensemble des solutions : c'est la "classe d'équivalence" de représentant $\underline{\hat{b}}$, sous-espace affine du noyau de l'application X'.

Les solutions $\underline{b}$, éléments de la classe d'équivalence $\mathfrak{C}_{\underline{\hat{b}}}$, vérifient les équations normales :

$$\boxed{X \cdot N(\underline{y}) = X \cdot N \cdot X' \, (\underline{b})} \tag{2}$$

Un tel vecteur $\underline{b}$ sera appelé vecteur des coefficients de régression.

Si l'ensemble des vecteurs $\underline{x}^j$ forme une base de W, l'application X' est injective (le rang de X' est égal à p) et son noyau se réduit à l'élément $\underline{0}$.

Alors l'ensemble des solutions ne comporte qu'un élément $\underline{\hat{b}}$:

$$\mathfrak{C}_{\underline{\hat{b}}} = \{\underline{\hat{b}}\}$$

avec

$$\boxed{\underline{\hat{b}} = (X \cdot N \cdot X')^{-1} \cdot X \cdot N(\underline{y})}$$

Voici l'expression du projecteur A_1 :

$$A_1 = X' \cdot (X \cdot N \cdot X')^{-1} \cdot X \cdot N$$

Remarque :

Si X' n'est pas injective, il suffit, dans l'expression précédente, de remplacer $(X \cdot N \cdot X')^{-1}$ par une inverse généralisée $(X \cdot N \cdot X')^{-}$ de l'application $X \cdot N \cdot X'$, pour obtenir l'expression du projecteur A_1 (40).

1.3 CHOIX DE LA METRIQUE N DANS F-THEOREME DE GAUSS-MARKOV

Le théorème de Gauss-Markov est le théorème fondamental dans l'étude du Modèle Linéaire. C'est ce théorème qui précise le choix de la métrique N dans F.II sera présenté ici dans une optique purement algébrique, l'énoncé classique apparaissant comme un corollaire.

1.3.1 Théorème fondamental sur les formes quadratiques

Soit la décomposition du vectoriel F en deux sous-espaces vectoriels supplémentaires W_1 et W_2 :

$$F = W_1 \oplus W_2$$

Dans cette décomposition, un nuage $\mathfrak{N}$ de points de F se projette en un nuage $\mathfrak{N}_1$ sur W_1 et en un nuage $\mathfrak{N}_2$ sur W_2 .

On note Γ , Γ_1 , Γ_2 les formes quadratiques d'inertie associées respectivement aux trois nuages $\mathfrak{N}$, $\mathfrak{N}_1$, $\mathfrak{N}_2$ et Γ est supposée régulière, alors :

Théorème :

Les deux propriétés :

a) W_1 est Γ^{-1}-orthogonal à W_2

b) $\Gamma = \Gamma_1 + \Gamma_2$

sont équivalentes.

Le théorème dû à J.P. PAGES est démontré dans CAZES (8 ; I-14).

Remarque : si Γ n'est pas régulière, il suffit de remplacer Γ^{-1} par une inverse généralisée linéaire quelconque Γ^{-} .

1.3.2 Inégalités entre formes quadratiques définies ou semi-définies positives

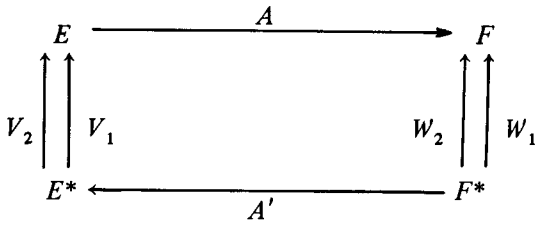
Soient V_1 et V_2 deux formes quadratiques définies ou semi-définies positives (ou leurs matrices associées).

Définition : on dira que V_1 est inférieure à V_2 ($V_1 \leqslant V_2$) si la forme quadratique $V_2 - V_1$ (ou la matrice associée) est semi-définie positive.

L'ordre entre deux f.q.i. reste-t-il inchangé lorsqu'on transforme les nuages de points auxquels elles sont associées par une application linéaire ?

C'est l'objet du théorème qui va suivre.

On considère dans l'espace vectoriel E deux nuage $\mathfrak{M}_1$ et $\mathfrak{M}_2$ de f.q.i. associées V_1 et V_2. Soit A une application linéaire de E dans F. Aux nuages $\mathfrak{N}_1 = A(\mathfrak{M}_1)$ et $\mathfrak{N}_2 = A(\mathfrak{M}_2)$ sont associées les f.q.i. W_1 et W_2.



$$W_1 = A\ V_1\ A'$$
$$W_2 = A\ V_2\ A'.$$

Théorème :

Si l'application A est injective, les propriétés :

a) $V_1 \leqslant V_2$

b) $W_1 \leqslant W_2$

sont équivalentes.

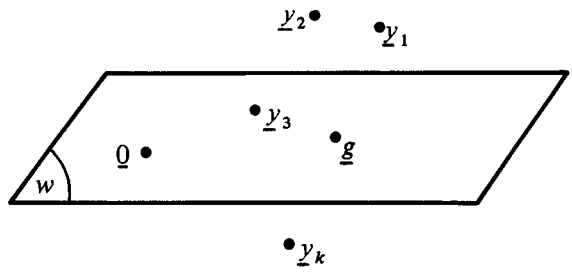
La démonstration est laissée au lecteur.

Remarque : quelle que soit l'application A, on a toujours a) ⇒ b).

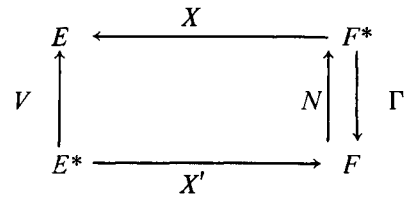
1.3.3 Théorème de Gauss-Markov

Dans l'étude du Modèle Linéaire, à l'ensemble des valeurs $\{x_i^j / j = 1,p\}$ des p variables explicatives correspond en réalité non pas une valeur de y, y_i, mais tout un ensemble de valeurs possibles que l'on pourrait appréhender si on répétait les observations pour le même p-uple $\{x_i^j / j = 1\ , p\}$

A une seule observation des n p-uples $\{x_i^j / j = 1\ , p\}$ correspond le vecteur $\underline{y}$. Si on répétait les observations k fois, on aurait un ensemble de k vecteurs $\{\underline{y}_l / l = 1,\ k\}$ contenu dans F. Si le modèle linéaire posé est "bon", le centre de gravité $\underline{g}$ (on a attribué à chaque $\underline{y}_l$ un poids) se trouve dans F proche du sous-espace vectoriel $W = X'(E^*)$. Si $\underline{g} \in W$, c'est-à-dire si $\underline{g} = X'\underline{\beta}$, le modèle est "sans biais". C'est ce qui sera supposé désormais.



Au nuage $\mathcal{N} = \{\underline{y}_l / l = 1\ , k\}$ de centre de gravité $\underline{g}$ est associé la forme quadratique d'inertie Γ supposée définie. Γ apparaît dans le schéma :



On a la décomposition

$$F = W \oplus W^{\perp} \tag{1}$$

où $W^{\perp}$ est le supplémentaire N-orthogonal de W dans F. A cette décomposition correspond la décomposition du nuage $\mathfrak{N}$ en deux sous-nuages $\mathfrak{N}_1$ et $\mathfrak{N}_2$ de f.q.i. Δ_1 et Δ_2 .

1.3.3.1 Le théorème

Théorème

La métrique N sur F qui rend Δ_1 minimum est la métrique définie par Γ^{-1}

Démonstration :

Si on note Γ_1 et Γ_2 les f.q.i. Δ_1 et Δ_2 dans le cas particulier où $N = \Gamma^{-1}$, d'après le théorème fondamental du paragraphe 1.3.1 :

$$\Gamma = \Gamma_1 + \Gamma_2$$

Soit A le projecteur sur W associé à une décomposition "quelconque" (1) de F ; on a :

$$\Delta_1 = A\,\Gamma\,A' = A\,\Gamma_1\,A' + A\,\Gamma_2\,A'$$

Or la f.q.i. Γ_1 étant associée à un nuage contenu dans W ,

$$\Gamma_1 = A\,\Gamma_1\,A'$$

d'où

$$\Delta_1 = \Gamma_1 + A\,\Gamma_2\,A'$$

$A\,\Gamma_2\,A'$ est une forme quadratique semi-définie positive donc,

$$\Gamma_1 \leqslant \Delta_1 \qquad \text{(c.q.f.d.)}$$

Remarque 1

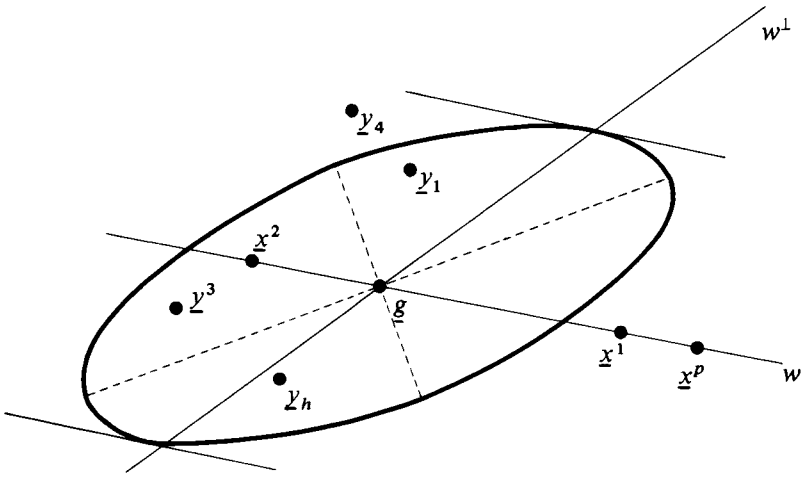
Si Γ n'est pas régulière, il suffit de remplacer dans le théorème Γ^{-1} par une inverse généralisée linéaire quelconque Γ^{-}.

Remarque 2

Du point de vue géométrique, que signifie choisir $N = \Gamma^{-1}$? On se place dans le cas particulier : $F = R^2$ et dim W = 1

Comment projeter le nuage $\mathcal{N}$ sur W de façon à ce que le nuage des $\underline{\hat{y}}_l$ soit le moins dispersé possible ?

Γ = f.q.i. du nuage $\mathcal{N}$ (matrice (2,2))



Si la dispersion de $\mathcal{N}$ est suffisamment "elliptique" dans le plan, l'axe de projection tel que la dispersion sur W soit minimale est défini par la direction "conjuguée" $W^{\perp}$ de W par rapport aux ellipses dans F dont l'équation est de la forme $\Gamma^{-1}(\underline{y} - \underline{g}) = c$

$W^{\perp}$ est le supplémentaire Γ^{-1}-orthogonal de W dans F.

1.3.3.2 Les conséquences du théorème

Conséquence si X' est injective

A chaque vecteur $\underline{y}_l \in F$ correspond le vecteur de coefficients de regression $\underline{b}_l$ vérifiant :

$$\underline{\hat{y}}_l = A\underline{y}_l = X'(\underline{b}_l) \qquad \text{et donc} \qquad \underline{b}_l = (XNX')^{-1} XN\underline{y}_l$$

Au nuage $\{\underline{y}_l / l = 1, 2, \ldots, k\}$ de f.q.i. Γ, correspond donc dans E^* le nuage des vecteurs de régression $\{\underline{b}_l \ / \ l = 1, 2, \ldots, k\}$ dont la f.q.i. (ou matrice de variance des coefficients de régression) traduit la dispersion dans E^*. Cette forme quadratique d'inertie vaut :

$$(XNX')^{-1}\ XN\ \Gamma\ NX'\ (XNX')^{-1}\ .$$

Le modèle étant sans biais, le centre de gravité des $\underline{y}_l$ est $\underline{g}$, et le centre de gravité des vecteurs de régression est le vecteur $\underline{\beta}$ des coefficients du modèle.

D'après le théorème de GAUSS-MARKOV et le théorème énoncé en 1.3.2. *la dispersion des vecteurs de régression est minimum quand on choisit pour métrique N dans F la métrique* Γ^{-1}.

La f.q.i. qu'on notera var $(\underline{\hat{b}})$ vaut alors : $(X\,\Gamma^{-1}\,X')^{-1}$

Conséquence si X' est quelconque

Si X' n'est pas injective, il existe une infinité de forme linéaires $\underline{b}$ vérifiant $\underline{\hat{y}} = X'(\underline{b})$ (cf. 1.2).

On ne peut alors parler du nuage des formes linéaires $\underline{b}_l$ et la phrase "la dispersion des coefficients de régression est minimum quand on choisit pour métrique N dans F la métrique Γ^{-1}" n'a pas de sens.

Si la forme linéaire $\underline{b}$ vérifiant $\underline{\hat{y}} = X'(\underline{b})$ n'est pas définie de façon unique, il n'en est pas de même des images de $\underline{b}$ par certaines applications linéaires dites "estimables".

Définition : l'application linéaire $s : E^* \xrightarrow{s} G$ est dite *estimable* s'il existe une application linéaire $h : F \xrightarrow{h} G$ telle que : $s = h\ .\ X'$.

Définition équivalente : l'application $s : E^* \longrightarrow G$ est dite estimable s'il existe une application linéaire $h : F \longrightarrow G$ telle que $h(\underline{g}) = s(\underline{\beta})$.

Si s est une application linéaire estimable, pour tout $\underline{b}$ vérifiant

$$\underline{\hat{y}} = X'\underline{b}\ ,\ \text{l'égalité :}$$

$$s(\underline{b}) = h\ .\ X'(\underline{b}) = h(\underline{\hat{y}}) = h\ .\ A(\underline{y})$$

montre en effet que le vecteur $s(\underline{b})$ est défini de façon unique.

Aux applications linéaires estimables s on peut donc associer les nuages de points $\{s(\underline{b}_l)/l = 1, \ldots, k\}$.

Compte tenu de la remarque faite après le théorème rencontré en 1.3.2., on peut alors énoncer le corollaire suivant du théorème de GAUSS-MARKOV :

Corollaire : Pour toute application linéaire estimable s, la forme quadratique d'inertie du nuage $\{s(\underline{b}_l)/l = 1, \ldots, k\}$ est minimum quand on choisit pour métrique N dans F la métrique Γ^{-1} ; elle s'écrit alors :

$$s\,(X\,\Gamma^{-1}\ X')^{-}\,s'$$

Remarques :

– s estimable $\Longleftrightarrow$ $\mathrm{Ker}\,(X') \subset \mathrm{Ker}\,(s)$

– si dim $G = 1$, c'est-à-dire si $G = R$, aux formes linéaires s et h sont associés respectivement les vecteurs $\underline{s} \in E^{**} = E$ et $\underline{h} \in F^*$; on a alors :

s estimable $\Longleftrightarrow$ il existe $\underline{h} \in F^*$ tel que $\underline{s} = X(\underline{h})$

– si l'application linéaire X' est injective, toute application linéaire s est estimable.

1.3.3.3 Rappel : forme classique du théorème de GAUSS-MARKOV

L'introduction d'une structure probabiliste est nécessaire dès qu'on a une infinité de réalisations possibles pour le vecteur $\underline{y}$.

Pour le Modèle Linéaire, $\underline{y} = \sum_{j=1}^{p} \underline{x}^j . \beta_j + \underline{e} = X'\underline{\beta} + \underline{e}$

On considère :

– $\underline{\beta}$ comme un vecteur de paramètres inconnus ;

– les $\underline{x}^j$ comme des vecteurs connus ;

– $\underline{y}$ comme une variable aléatoire n-dimensionnelle dont on possède une réalisation encore notée $\underline{y}$;

– $\underline{e}$ comme une variable aléatoire n-dimensionnelle appelée "écart" ou "erreur".

Les hypothèses sur les distributions de $\underline{y}$ et de $\underline{e}$ sont les suivantes :

Hypothèse 1. $E(\underline{y}) = X'\underline{\beta} \Longleftrightarrow E(\underline{e}) = \underline{0}$, c'est-à-dire le modèle est sans biais.

Hypothèse 2. Var $(\underline{y})$ = Var $(\underline{e})$ = Γ.

Hypothèse 3. $\underline{y}$ suit une loi normale $LG_n(X'\underline{\beta} ; \Gamma)$ ce qui est équivalent à : $\underline{e}$ suit une loi normale $LG_n(\underline{0} ; \Gamma)$.

L'hypothèse 3 est souvent posée car elle permet de construire des régions de confiance au niveau des β_j et de faire des tests d'hypothèses. On doit souligner que, dans de nombreux cas, cette hypothèse n'est pas irréaliste : il suffit que le théorème central limite puisse s'appliquer à $\underline{e}$.

L'hypothèse 3 n'est pas nécessaire pour le théorème de Gauss-Markov.

On note :

$\underline{\hat{y}}$ la projection Γ^{-1} orthogonale de $\underline{y}$ sur W ;

$\underline{\hat{b}}$ un vecteur de régression qui vérifie $\underline{\hat{y}} = X'\underline{\hat{b}}$;

s une application estimable quelconque.

Alors $s(\underline{\hat{b}})$ est unique, linéaire par rapport à $\underline{y}$ et d'espérance $s(\underline{\beta})$.

Théorème de GAUSS-MARKOV

$s(\hat{\underline{b}})$ est de matrice variance minimale dans la classe des estimateurs de $s(\underline{\beta})$ qui sont sans biais et linéaires en $\underline{y}$.

Comme de plus, $s(\hat{\underline{b}})$ est unique, $s(\hat{\underline{b}})$ est l'estimateur efficace de $s(\underline{\beta})$ pour cette classe d'estimateurs.

Classiquement, $\hat{\underline{b}}$ est appelé "le meilleur estimateur linéaire sans biais" ou encore l'estimateur "B.L.U." (Best Linear Unbiaised).

On va énoncer un résultat, conséquence du Théorème de GAUSS-MARKOV, qui sera utilisé ultérieurement.

Corollaire : Parmi les estimateurs $\underline{b}$ de $\underline{\beta}$, linéaires par rapport à $\underline{y}$ et sans biais, si X' est injective, $\hat{\underline{b}}$ est l'unique estimateur qui minimise $E(\|\underline{b} - \underline{\beta}\|_M^2)$, quelque soit la métrique M choisie dans E^*.

Démonstration dans CAZES (8).

On considérera désormais que Γ est régulière,

X' est injective,

et on emploiera plutôt le langage probabiliste.

1.3.3.4 Est-il possible de choisir $N = \Gamma^{-1}$?

En général, on ne connaît pas la forme de Γ ; on ne possède en effet qu'une réalisation $\underline{y}$ du n-uple alors qu'il en faudrait au moins trois pour un début d'appréciation des variabilités "intra-observation" et des liaisons "inter-observations".

Aussi, est-on amené à faire des hypothèses très simplificatrices sur Γ. On peut supposer, par exemple, que Γ soit égale à $\sigma^2 . I_n$; alors, la solution "classique" des moindres carrés : $\hat{\underline{b}} = (X X')^{-1} X \underline{y}$ est le BLU dans la mesure où les "observations" $y_1, y_2, \ldots, y_n$ sont effectivement non corrélées et de même variance.

1.4 DISTINGUO ENTRE "MODELE LINEAIRE" ET "REGRESSION LINEAIRE"

Supposons qu'on ait un n-échantillon *centré* de la variable $(p + 1)$ – dimensionnelle $(y, x^1, x^2, \ldots, x^p)$ et qu'on désire à partir de ce n-échantillon rechercher la meilleure approximation linéaire de y en fonction des x^i, $(i = 1, 2, \ldots p)$. On a donc $(p + 1)$ points $\underline{y}, \underline{x}^1, \underline{x}^2, \ldots, \underline{x}^p$ dans $F = R^n$.

Avec F muni de la métrique N, ceci revient à projeter $\underline{y}$ sur

$$W = \{\underline{z} \in F \mid \underline{z} = \sum_{i=1}^{p} b_i \underline{x}^i \quad \text{où} \quad \underline{b} \in E^*\}$$

et la solution obtenue est formellement identique à celle du Modèle Linéaire.

Pour retrouver la meilleure approximation linéaire de y, on travaille comme dans le cas de deux variables à l'aide de la covariance ou de la corrélation : la meilleure approximation linéaire de y en fonction des x^i, $\sum_{i=1}^{p} b_i x^i$, est obtenue quand la variance empirique de y-$\Sigma\, b_i x^i$ est minimale .

Ceci revient à munir F, considéré comme l'espace L^2 des variables aléatoires de carré intégrable, de la distance en moyenne quadratique D_p (c.f. 1.1 remarque 3).

Théoriquement, dans le cadre du "modèle linéaire" on considérera le tableau X' comme non aléatoire et $\underline{y}$ comme un 1-échantillon d'une variable n-dimensionnelle de matrice variance Γ ; la métrique à choisir est alors : $N = \Gamma^{-1}$.

Dans l'optique "régression linéaire", le tableau des données $(\underline{y}\ \underline{x}^1 \ldots \underline{x}^p)$ sera considéré comme un n-échantillon du $(p+1)$ -uple $(y, x^1, x^2, \ldots, x^p)$ et la métrique à choisir est naturellement D_p.

Pratiquement, il n'est pas toujours aisé de savoir si le problème qu'on étudie relève de l'une ou l'autre notion ; il en est ainsi dans tous les cas où il n'existe pas pour le problème étudié de modèle établi a priori.

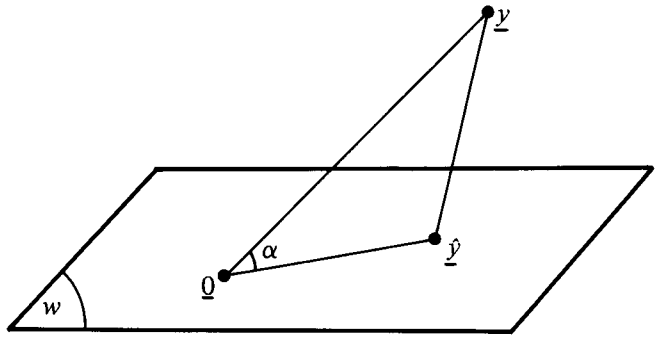
1.5 QUALITE DE LA SOLUTION $\hat{\underline{y}}$

La qualité de la solution est jugée différemment selon qu'on traite d'une Régression entre variables aléatoires ou d'un Modèle Linéaire.

1.5.1 Optique régression entre variables aléatoires

L'objectif étant d'obtenir la combinaison linéaire de variables $x^1, x^2, \ldots, x^p$ qui soit la plus proche de y, on mesurera la qualité de la régression par le rapport R :

$$R = \frac{\|\hat{\underline{y}}\|_{D_p}}{\|\underline{y}\|_{D_p}}$$



$F = W \oplus W^{\perp} : \underline{y} = \underline{\hat{y}} + \underline{\hat{e}}$

Si les axes D_p-orthogonaux sont représentés orthogonaux, R n'est autre que le cosinus de l'angle "$\underline{y}\,0\,\underline{\hat{y}}$" : $R = \cos \alpha$.

R est analogue à un coefficient de corrélation linéaire :

$- R = 1 \iff \underline{y} \in W \iff \underline{\hat{e}} = \underline{0}$

$- R = 0 \iff \underline{y} \perp_{D_p} W \iff \underline{y} = \underline{\hat{e}}$

Si le n-échantillon du $(p + 1)$ -uple $(y, x^1, x^2, \ldots, x^p)$ est centré, alors $R = r(y, \{x^1, x^2, \ldots, x^p\}) = r(y, \hat{y})$ est le coefficient de corrélation multiple entre y et $\{x^1, x^2, \ldots, x^p\}$, c'est-à-dire le coefficient de corrélation linéaire entre y et la combinaison linéaire $\hat{y}$.

1.5.2 Optique modèle linéaire

L'objectif étant d'approcher au plus près du centre de gravité $\underline{g} = E(\underline{y}) \in W$, le choix de la métrique Γ^{-1} concourt à réaliser cet objectif. La dispersion du nuage des solutions $\underline{\hat{y}}$ est minimale ou de manière équivalente

$$\text{var}(\underline{\hat{b}}) = (X\,\Gamma^{-1}\,X')^{-1}$$

est minimale. La qualité du modèle se juge alors en regardant si les variances des estimations $\text{var}(\hat{b}_i)$ sont faibles relativement aux $\hat{b}_i$.

Dans la pratique, on ne connaît pas Γ et on fera souvent l'hypothèse que Γ est de la forme $\sigma^2 I_n$. On améliorera ce premier modèle, soit en ôtant des variables explicatives soit en prenant pour Γ une forme plus compliquée si ces choix conduisent globalement à diminuer les variances des $\hat{b}_i$.

La qualité de la solution $\underline{\hat{y}}$ pourrait être mesurée par sa distance à $E(\underline{y})$, c'est-à-dire par $\|\underline{\hat{y}} - E(\underline{y})\|^2_{\Gamma^{-1}}$ mais on ne connaît pas $E(\underline{y})$.

On connaît cependant la valeur moyenne de cette distance (qui n'est rien d'autre que l'inertie par rapport au centre de gravité $\underline{g}$ du nuage des $\underline{y}_l$ dans le cas fini) ; en effet :

$$F = W^{\perp} \oplus W \qquad \text{au sens} \qquad \text{de } \Gamma^{-1}$$

$$\underline{y} - E(\underline{y}) = \underline{y} - \hat{\underline{y}} + \hat{\underline{y}} - E(\underline{y})$$

$$\|\underline{y} - E(\underline{y})\|^2_{\Gamma^{-1}} = \|\underline{y} - \hat{\underline{y}}\|^2_{\Gamma^{-1}} + \|\hat{\underline{y}} - E(\underline{y})\|^2_{\Gamma^{-1}} \quad \text{Théorème de Pythagore.}$$

$$E(\|\hat{\underline{y}} - E(\underline{y})\|^2_{\Gamma^{-1}}) = \text{trace}\,(\Gamma^{-1} \,.\, \text{var}\,(\hat{\underline{y}})) = \text{trace}\,(\Gamma^{-1} X' \,\text{var}\,(\hat{\underline{b}})\, X)$$

$$= \text{trace}\,(\Gamma^{-1} X'(X\,\Gamma^{-1} X')^{-1} X) = p$$

p est la dimension du sous-espace vectoriel W dont une base est

$$\{\underline{x}^1, \underline{x}^2, \ldots, \underline{x}^p\} .$$

Par un calcul analogue sur les autres termes de la somme, on obtient :

$$E(\|\underline{y} - E(\underline{y})\|^2_{\Gamma^{-1}}) = n$$

et

$$E(\|\underline{y} - \hat{\underline{y}}\|^2_{\Gamma^{-1}}) = n - p .$$

On a ainsi une première possibilité de vérifier le modèle :

comparer : $\dfrac{\|\underline{y} - \hat{\underline{y}}\|^2_{\Gamma^{-1}}}{n - p} = \dfrac{\|\hat{\underline{e}}\|^2_{\Gamma^{-1}}}{n - p}$ à sa valeur moyenne 1.

En pratique, ce rapport est rarement proche de 1 ; en effet, le modèle est parfois biaisé et la forme de Γ est très souvent inconnue.

Dans le cas où $\Gamma = \sigma^2 I_n$ avec σ^2 inconnu, le rapport ci-dessus fournit une estimation de σ^2.

On pourrait songer au rapport :

$$F = \frac{\|\hat{\underline{y}} - \underline{g}\|^2_{\Gamma^{-1}}/p}{\|\hat{\underline{e}}\|^2_{\Gamma^{-1}}/n - p} = \frac{n - p}{p} \cdot \frac{\|\hat{\underline{y}} - \underline{g}\|^2_{\Gamma^{-1}}}{\|\hat{\underline{e}}\|^2_{\Gamma^{-1}}}$$

dont la valeur moyenne vaut 1 sous l'hypothèse de normalité, mais qui ne peut être calculé si on ne connaît pas $\underline{g} = X'\underline{\beta}$.

Par contre, le rapport :

$$F = \frac{n - p}{p} \cdot \frac{\|\hat{\underline{y}}\|^2_{\Gamma^{-1}}}{\|\underline{y} - \hat{\underline{y}}\|^2_{\Gamma^{-1}}} \qquad \begin{array}{l} \text{devrait être proche de 1 si } \underline{g} = \underline{0} \\ \text{et très élevé si } \underline{g} \neq \underline{0} \end{array}$$

d'où la possibilité de juger la "signification" du modèle, c'est-à-dire d'apprécier si $\underline{g}$ est significativement différent de $\underline{0}$ (cas où le rapport est très élevé).

Faire l'hypothèse que $\underline{\varepsilon}$ suit une $LG_n(\underline{0}\ ;\ \Gamma)$ fournit une règle de décision statistique ; en effet, avec cette hypothèse supplémentaire, F suit une loi de FISHER-SNEDECOR à $(p\ ,\ n-p)$ degrés de liberté et on peut effectuer le test de l'hypothèse : $\underline{g} = \underline{0}$ contre celle : $\underline{g} \neq \underline{0}$.

Plus généralement, voici comment on mesure la "signification" d'un groupe de caractères explicatifs.

– soit W_p la variété linéaire engendrée par la totalité des p caractères explicatifs, notée précédemment W.

$\underline{\hat{y}}_p$, projection Γ^{-1}-orthogonale de $\underline{y}$ sur W_p, est "l'estimation" de $\underline{g}$ obtenue par le modèle "complet" ;

– soit W_q la variété linéaire engendrée par q caractères explicatifs seulement pris parmi les p

$$W_q \subset W_p$$

$\underline{\hat{y}}_q$, projection Γ^{-1}-orthogonale de $\underline{y}$ sur W_q, est "l'estimation" de $\underline{g}$ obtenue par le modèle à q caractères explicatifs.

$\underline{\hat{y}}_p$ a p degrés de liberté

$\underline{\hat{y}}_q$ a q degrés de liberté

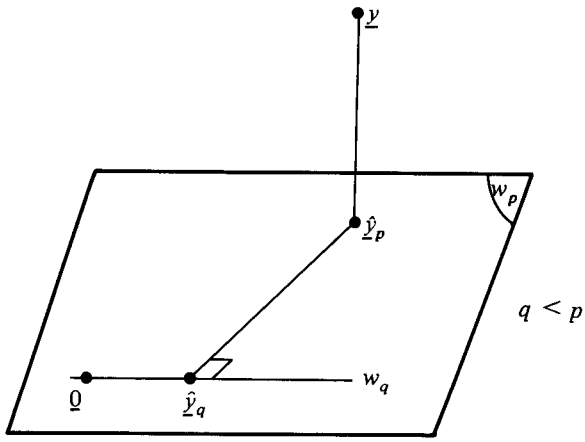
$\underline{\hat{y}}_p - \underline{\hat{y}}_q$ a p-q degrés de liberté

$\underline{y}$ a n degrés de liberté

Remarque

$\underline{\hat{y}}_q$ est aussi la projection Γ^{-1}-orthogonale de $\underline{\hat{y}}_p$ sur W_q

$$\|\underline{y}\|^2 = \|\underline{\hat{y}}_q\|^2 + \|\underline{\hat{y}}_p - \underline{\hat{y}}_q\|^2 + \|\underline{y} - \underline{\hat{y}}_p\|^2 \qquad \text{(Pythagore)}$$



On s'interroge sur l'intérêt d'inclure les p-q caractères dans le modèle.

$$\text{L'indice } F = \frac{\|\underline{\hat{y}}_p - \underline{\hat{y}}_q\|^2 / _{p-q}}{\|\underline{y} - \underline{\hat{y}}_p\|^2 / _{n-p}}$$

est utilisé pour mesurer le "pouvoir explicatif" des p-q caractères restants alors que les q premiers sont utilisés.

Il est d'autant plus intéressant de faire entrer les p-q caractères dans le modèle que F est fort.

Ici encore, moyennant les hypothèses de normalité, on peut, par référence à la table "du F" de FISHER-SNEDECOR, juger si les p-q caractères encore inutilisés ont un pouvoir explicatif lorsqu'on a inclus les q premiers dans le modèle.

Deux cas particuliers :

a) $q = p - 1$: un caractère explicatif, $\underline{x}^p$ par exemple, n'est pas utilisé ; F mesure le pouvoir explicatif apporté par le caractère $\underline{x}^p$, alors que les les caractères $\underline{x}^1, \underline{x}^2, \ldots, \underline{x}^{p-1}$ sont utilisés.

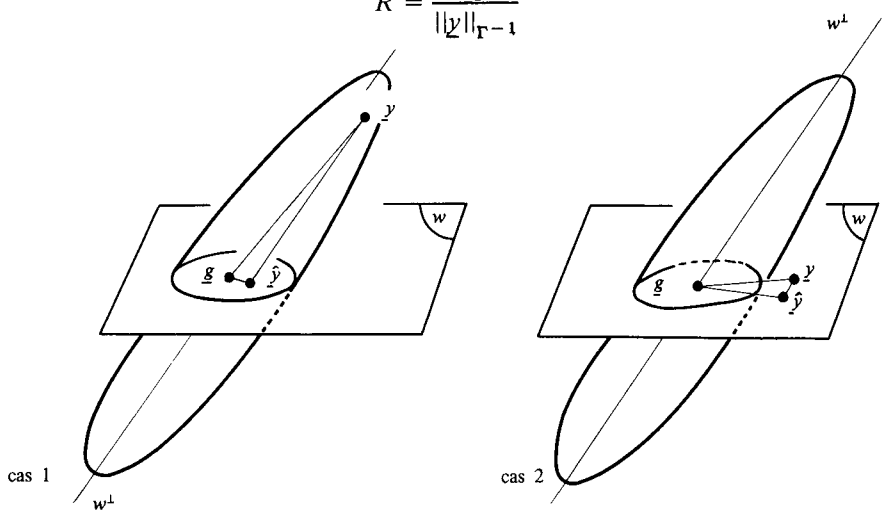
b) $q = 0 \iff W_q = \{\underline{0}\} \Rightarrow \underline{\hat{y}}_q = \underline{0}$

On retrouve l'indice $F = \dfrac{n-p}{p} \dfrac{\|\underline{\hat{y}}\|^2}{\|\underline{y} - \underline{\hat{y}}\|^2}$

donné ci-dessus pour juger si le modèle est "significatif".

A propos du rapport :

$$R = \frac{\|\underline{\hat{y}}\|_{\Gamma^{-1}}}{\|\underline{y}\|_{\Gamma^{-1}}}$$



On voit bien maintenant

– comment le rapport R peut être mauvais et la solution obtenue $\underline{\hat{y}}$ satisfaisante (cas 1) ;

– comment le rapport R peut être bon et la solution $\underline{\hat{y}}$ mauvaise (cas 2).

Heureusement, si le modèle posé est bon (c'est-à-dire que $E(\underline{y}) = \underline{g} \in W$ et que l'hypothèse $\Gamma = \sigma^2 I_n$ est acceptable, σ^2 étant suffisamment petit) on aura en général $\underline{y}$ proche du sous-espace W et proche de $\underline{g}$, ce qui justifie le réflexe usuel de consulter le rapport R pour apprécier la qualité de la solution dans le cadre du Modèle Général Linéaire.

CONCLUSION

Ces précisions ayant été apportées et toute confusion n'étant plus à craindre, on utilisera le mot "régression" pour signifier les résultats d'un problème relevant soit du "Modèle Linéaire", soit de la "Régression entre variables aléatoires".

II – EXAMEN CRITIQUE DU MODÈLE LINÉAIRE : ANALYSE DES RÉSIDUS

INTRODUCTION

Les rapports R et F définis en 1.5 sont des indices globaux de qualité de la représentation. Pour se rendre compte dans le détail de l'approximation faite, il faut analyser les résidus,

$$\hat{e}_i = y_i - \hat{y}_i \quad (i = 1 , n)$$

Qu'espère-t-on des résidus ? Ils doivent être du même ordre de grandeur et ne pas présenter des tendances qui peuvent être considérées comme graves, à savoir par exemple : résidus tous négatifs pour les faibles y_i puis tous positifs ; résidus dont la valeur absolue croît avec y_i ou inversement.

Leur analyse conduit souvent à remettre en cause le modèle.

Le meilleur procédé et le plus commode consiste à visualiser les résidus.

Cette analyse graphique doit précéder tout test d'hypothèse construit à partir des résidus. En effet, un test qu'il soit paramétrique ou non est spécifique d'un corps d'hypothèses et n'utilise donc qu'en partie l'information apportée par les résidus ; ainsi, par exemple, il est inutile de savoir qu'on ne rejette pas l'hypothèse "e_i suit une LG $(0 , \sigma)$ " en appliquant aux résidus le test du χ^2 (chi-deux) par exemple, si l'analyse des séquences remet en cause le modèle linéaire posé.

De plus, le test ne fait en général que confirmer ce qui transparaît dans un graphique approprié.

On insiste ici sur le contrôle de la "non-corrélation des erreurs" l'idée d'obtenir les résidus BLUS étant issue de réflexions sur ce thème. Volontairement, le contrôle de l'hypothèse classique "e_i suit une LG $(0 , \sigma)$" n'est pas abordé ici, ce contrôle étant couramment pratiqué. On trouvera cependant :

– une bonne comparaison de tests possibles dans SHAPIRO et WILK (42, 43) ;

– la comparaison de certains tests lorsqu'ils sont construits à partir des résidus "ordinaires" ou à partir des résidus BLUS dans HUANG et BOLCH (24).

Tout aussi volontairement, le test de BARTLETT "homocédasticité – hétérocédasticité" est omis. Le test est présenté dans KENDALL et STUART vol. 2 p. 234 et p. 244 (26), appliqué aux résidus "ordinaires" ou BLUS dans RAMSEY (38, 39).

2.1 VISUALISATION DES RESIDUS

On rappelle ici rapidement quelques types d'analyse graphique des résidus.

– Histogramme des $\hat{e}_i$

Calcul de $\sum_{i=1}^{n} \hat{e}_i$

– Graphique $(i\ ,\ \hat{e}_i)$ $(i = 1\ ,\ n)$

– Graphique $(\hat{e}_i, \hat{e}_{i+1})$ $(i = 1\ ,\ n - 1)$

– Graphique $(\hat{e}_i, y_i)$ $(i = 1\ ,\ n)$

– Graphiques "résidus – variable explicative"

et la liste n'est pas exhaustive !

Pour plus d'information, on consultera C 3 E (11), DRAPER and SMITH (13) ou ANSCOMBE and TUKEY (3) (2).

Si on se réfère à une structure probabiliste, on attend des résidus, dans le cas où le modèle posé est correct, qu'ils se comportent sensiblement comme des erreurs observées. Les graphiques permettent de vérifier la compatibilité avec les *hypothèses usuelles :*

(1) : $E(e_i) = 0$

(2) : $V(y_i) = V(e_i) = \sigma^2$, cov $(e_i, e_j) = 0$.

(3) : e_i suit une $LG(0\ ,\ \sigma)$.

2.2 UTILISATION DE TESTS NON-PARAMETRIQUES

Les résidus, s'ils ne révèlent pas de tendances, devraient se succéder au hasard. On utilise alors un test de séquences ou un test de signes.

On va étudier plus particulièrement comment contrôler l'hypothèse de non-corrélation des erreurs.

2.3 CONTROLE DE LA "NON CORRELATION DES ERREURS"

2.3.1 Introduction : Présentation du problème

On emploie trop souvent la technique des moindres carrés pour estimer les paramètres β_j, sans se préoccuper de savoir si les hypothèses usuelles sur $\underline{\varepsilon}$ sont satisfaites. Or, si l'hypothèse (1) ou (2) ne l'est pas, il faudra utiliser une autre méthode :

– quand la matrice Γ de covariance des erreurs est en réalité $\sigma^2 D$ (où D est une matrice diagonale positive), on devra utiliser la méthode des moindres carrés pondérés ;

– quand les erreurs sont corrélées, le problème est plus compliqué. Si les coefficients de la matrice Γ sont connus a priori, une transformation linéaire de toutes les variables permet alors de se ramener à un modèle vérifiant les hypothèses (1) et (2). Si au contraire, ces coefficients ne sont pas connus, et si l'hypothèse (2) a été rejetée à l'aide d'un test sur les résidus, on pourra supposer que la matrice Γ se présente sous une forme suffisamment simple, pour que l'estimation de ses coefficients soit possible à partir des résidus. On est alors ramené au cas des coefficients connus ; on doit souligner qu'une telle procédure doit être utilisée avec précautions.

C'est dans le cadre de l'analyse des séries chronologiques par des techniques de régression, que les soucis précédents ont été nettement exprimés. En effet, dans de telles données, les observations successives sont souvent corrélées ; *alors l'estimateur usuel des moindres carrés, bien que sans biais, n'est plus de variance minimale.*

Les inconvénients en ont été étudiés de façon plus détaillée par COCHRANE et ORCUTT (12).

Dans la pratique, on ne connaît pas les erreurs e_i aussi, dans un test de "non corrélation des erreurs", on devra nécessairement utiliser les résidus $\hat{e}_i$.

Les distributions de statistiques usuelles :

– le rapport de Von NEUMANN (35),

– les coefficients d'autocorrélation d'ANDERSON (1), ont été établies sous les hypothèses de non corrélation (2) et de normalité (3).

Pour un test, ces statistiques, calculées avec les résidus $\hat{\underline{e}}_i$, ne peuvent être utilisées puisque les résidus sont nécessairement corrélés par construction, que les erreurs e_i soient corrélées ou non.

Pour le rapport de Von NEUMANN calculé avec $\hat{\underline{e}}$, DURBIN et WATSON ont proposé une procédure de test de l'hypothèse "d'indépendance" qui possède l'inconvénient de ne pas toujours conclure. La procédure précédente a été améliorée par THEIL et NAGAR (46) pour un cas particulier.

Compte tenu des cas où la procédure de DURBIN et WATSON n'est pas satisfaisante (soit qu'elle ne permette pas de conclure, soit que le rapport de Von NEUMANN ne soit pas approprié en fonction de l'hypothèse alternative choisie), THEIL a été conduit à remplacer le vecteur des résidus $\underline{\hat{e}}$ obtenu à partir du critère des moindres carrés par un vecteur $\underline{e}^*$ vérifiant les mêmes hypothèses (1) et (2) que le vecteur des erreurs $\underline{e}$. C'est ce vecteur $\underline{e}^*$ qui est appelé le vecteur BLUS par THEIL.

Dans les paragraphes ci-dessous, les points abordés dans cette introduction sont développés.

2.3.2 Conséquences dues à la corrélation des erreurs

Avec les hypothèses :

$$(H_0) \begin{cases} (1)\ E(\underline{e}) = \underline{0} \\ (2)\ \text{var}\ (\underline{e}) = \text{var}\ (\underline{y}) = \sigma^2 I\,, \end{cases}$$

la solution des moindres carrés $\underline{\hat{b}}$,

$$\underline{\hat{b}} = (X \,.\, I \,.\, X')^{-1} \,.\, X \,.\, I\ (\underline{y})$$

possède les propriétés :

- $E(\underline{\hat{b}}) = \underline{\beta}$, $\underline{\hat{b}}$ est sans biais ;
- $\text{var}\ (\underline{\hat{b}}) = \sigma^2\ (X \,.\, I \,.\, X')^{-1}$

En outre, parmi les estimateurs linéaires en $\underline{y}$ et sans biais, $\underline{\hat{b}}$ est de matrice variance minimale (théorème de GAUSS-MARKOV – § 1).

Avec les hypothèses :

$$(H'_0) \begin{cases} (1)\ E(\underline{e}) = \underline{0} \\ (2)\ \text{var}\ (\underline{e}) = \text{var}\ (\underline{y}) = \Gamma \end{cases}$$

la solution des moindres carrés $\underline{\tilde{b}}$ de matrice variance minimale est :

$$\underline{\tilde{b}} = (X \,.\, \Gamma^{-1} \,.\, X')^{-1} \,.\, X \,.\, \Gamma^{-1}\ (\underline{y})$$

On a les propriétés :

- $E(\underline{\tilde{b}}) = \underline{\beta}$
- $\text{var}\ (\underline{\tilde{b}}) = (X \,.\, \Gamma^{-1} \,.\, X')^{-1}$

On constate donc que sous les hypothèses (H'_0), $\underline{\tilde{b}}$ est un estimateur sans biais mais ce n'est pas l'estimateur efficace : la variance de l'estimateur $\hat{b}_j$ sera en général plus grande que celle de $\tilde{b}_j$ et l'estimation de σ^2 à l'aide des résidus $(\underline{y} - X'(\underline{\hat{b}}))$ sera inexacte.

2.3.3 Procédure de DURBIN et WATSON

Pour vérifier l'hypothèse de "non corrélation des erreurs" on peut considérer les erreurs $e_1, \ldots, e_n$ comme un processus dont on veut vérifier la nullité des coefficients d'autocorrélation.

2.3.3.1 Quelques statistiques

On peut utiliser l'une des statistiques suivantes :

– *Le rapport de Von NEUMANN* (35) :

$$Q = \frac{\sum_{i=2}^{n} (e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

Cette statistique est particulièrement adaptée pour le test de l'hypothèse nulle de "non corrélation" contre l'hypothèse alternative d'erreurs obéissant à un processus d'autocorrélation de MARKOV du premier degré ;

– *les coefficients d'autocorrélation* (ANDERSON (1)) :

$$R_k = \frac{\sum_{i=1}^{n-k} e_i \, e_{i+k}}{\sum_{i=1}^{n} e_i^2}$$

R_k est le coefficient d'autocorrélation de rang k.

On remarque que ces statistiques, dont les distributions ont été établies sous les hypothèses (1) (2) (3) peuvent s'écrire comme rapport de deux formes quadratiques réelles :

$$\frac{H(\underline{e})}{\|\underline{e}\|^2}$$

Le calcul de leurs distributions nécessite la diagonalisation des matrices associées aux formes quadratiques.

D'une part, la distribution de Q a été donnée par HART (20) pour certaines valeurs de n, et d'autre part les valeurs de R_k, associées aux risques 5 % et 1 %, ont été données par ANDERSON pour différentes valeurs de n et k.

Mais les erreurs e_i étant inconnues, on est amené naturellement à utiliser ces statistiques avec les résidus $\hat{e}_i$. Or, on rappelle que (§ 1) :

$$\underline{\hat{e}} = \underline{y} - X'(\underline{\hat{b}}) = \underline{y} - A\,(\underline{y}) = (I - A)\,(\underline{y}) = P\,(\underline{y})\,,$$

A étant le projecteur I-orthogonal sur le sous-espace vectoriel W engendré par les vecteurs $\underline{x}^1, \ldots, \underline{x}^p$ de R^n et P le projecteur sur l'orthogonal de W.

Etant donné les propriétés du projecteur P :

$$P^2 = P = P',$$

$$\text{var}\,(\underline{\hat{e}}) = P \cdot \text{var}\,(\underline{y}) \cdot P' = \sigma^2\, P \cdot \mathrm{I} \cdot P' = \sigma^2\, P.$$

Ainsi les résidus $\hat{e}_i$ sont nécessairement corrélés ; l'hypothèse (2) étant en défaut, on ne peut employer les distributions des statistiques précédentes. Mais DURBIN et WATSON ont apporté une solution partielle à ce problème.

2.3-3.2 Le résultat de DURBIN et WATSON

DURBIN et WATSON ont encadré, sous l'hypothèse nulle, le rapport $\dfrac{\mathrm{H}(\underline{\hat{e}})}{\|\underline{\hat{e}}\|^2}$ par deux statistiques dont les distributions sont connues et du type précédent (14 (1)).

Nous allons préciser la procédure du test dans le cas suivant :

– soient les hypothèses :

$$(H_0)\begin{cases} E\,(\underline{e}) = \underline{0} \\ \text{var}\,(\underline{e}) = \sigma^2\,\mathrm{I} \\ \underline{e} \text{ suit une loi normale.} \end{cases}$$

$$(H_1)\begin{cases} E\,(\underline{e}) = \underline{0} \\ \text{var}\,(\underline{e}) = \sigma^2\,\Gamma_0 \quad \text{où} \quad \Gamma_0 = \begin{pmatrix} 1 & \rho & \rho^2 \ldots \rho^{n-1} \\ \rho & 1 & \rho \ \ldots \ \rho^{n-2} \\ \rho^{n-1} & \ldots & 1 \end{pmatrix} \text{ et } \sigma^2 = \dfrac{\alpha^2}{1-\rho^2} \\ \underline{e} \text{ suit une loi normale} \end{cases}$$

L'hypothèse (H_1) spécifie une des hypothèses les plus simples que l'on puisse opposer à l'hypothèse nulle ; elle exprime que les erreurs obéissent à un processus stationnaire de MARKOV du premier degré donné par :

$$e_i = \rho\, e_{i-1} + \eta_i$$

où $|\rho| < 1$

et η_i est une variable aléatoire normale telle que $\begin{cases} E(\eta_i) = 0 \\ \text{var}(\eta_i) = \alpha^2 , \forall\ i > 1 \\ \text{cov}(\eta_i, \eta_j) = 0 , \forall\ j \neq i \end{cases}$

ρ est le coefficient d'autocorrélation de rang 1 .

Sous l'hypothèse (H_1) le corrélogramme des erreurs décroît de façon exponentielle avec le rang.

La statistique de Von NEUMANN, appliquée aux résidus :

$$Q = \frac{\sum_{i=2}^{n} (\hat{e}_i - \hat{e}_{i-1})^2}{\sum_{i=1}^{n} \hat{e}_i^2}$$

est particulièrement pertinente pour l'alternative (H_1) .

Ce rapport est compris entre les valeurs 0 et 4. Si on connaît le signe de ρ, on choisira un test unilatéral ; pour une autocorrélation positive, le rapport Q est relativement petit (inférieur à 2) tandis que pour une autocorrélation négative, il est relativement grand. Si on ignore le signe de ρ, on choisira un test bilatéral.

Le fait d'avoir encadré Q par deux statistiques Q_L et Q_U de distributions connues empêche quelquefois de conclure. Ainsi, dans le cas d'une autocorrélation positive, pour un risque α donné, il existe deux bornes $Q_L(\alpha)$ et $Q_U(\alpha)$:

- si $Q < Q_L(\alpha)$, on rejetera (H_0)
- si $Q > Q_U(\alpha)$, on conservera (H_0)
- si $Q_L(\alpha) < Q < Q_U(\alpha)$, on ne peut conclure .

Les bornes $Q_L(\alpha)$ et $Q_U(\alpha)$ ont été tabulées par Durbin et Watson pour le risque α de 5 % et diverses valeurs de n et du nombre de variables explicatives p.

Quelques exemples sont traités par Durbin et Watson (14 (2)) ; mais pour les cas où on ne peut pas conclure, on sera conduit à faire des approximations.

2.3-3.3 Procédures approximatives

Pour le cas où on a $Q_L(\alpha) < Q < Q_U(\alpha)$, Durbin et Watson ont proposé (14 (2), § 4) un test suffisamment précis dès que le nombre de degrés de

liberté $(n - p)$ est relativement grand $(n - p > 40)$. La méthode consiste à transformer le rapport Q de façon à ce que son intervalle de variation soit l'intervalle [0,1] puis à ajuster une distribution β de même moyenne et de même variance. Theil et Nagar (46) ont appliqué la même méthode lorsque pour chaque variable explicative, les différences premières et secondes sont petites relativement à l'étendue de la variable elle-même.

La complexité des procédures précédentes étant due au fait que les résidus sont corrélés, Theil a abandonné l'idée d'améliorer ces procédures, préférant revenir sur le bien fondé du critère des moindres carrés. Le critère qu'il propose (44) et (45) a pour but d'exiger que les résidus soient "sphériques" (ce terme sera défini plus loin). Alors, pour de tels résidus, les statistiques données au § 2.3-3.1 sont immédiatement applicables.

2.3-4 La méthode "BLUS"

Le vecteur des résidus $\underline{\hat{e}}$ est un vecteur aléatoire de dimension n appartenant au sous espace supplémentaire orthogonal de W :

$$\underline{\hat{e}} = P(\underline{y})$$

et

$$\text{var}\,(\underline{\hat{e}}) = \sigma^2 P$$

où P est le rang $n - p$.

Etant donné un sous vecteur $\underline{\hat{e}}_1$ de $\underline{\hat{e}}$ de dimension $n - p$, il existe au moins une application linéaire $L : R^{n-p} \longrightarrow R^{n-p}$, telle que le vecteur

$$\underline{e}_1^* = L(\underline{\hat{e}}_1), \quad \text{de dimension } (n - p)$$

soit de variance

$$\sigma^2 \, I_{n-p} .$$

La matrice de variance de $\underline{\hat{e}}$ étant d'ordre n mais de rang $n - p$ ($n - p$ est le nombre de degrés de liberté de $\underline{\hat{e}}$), ce n'est qu'à partir de $n - p$ variables de $\underline{\hat{e}}$ qu'on construit un vecteur aléatoire $\underline{e}_1^*$ qui soit de matrice variance $\sigma^2 \, I_{n-p}$.

Le choix des p variables qui sont abandonnées dépend de l'alternative du test que l'on a en vue. Après une réindexation éventuelle des lignes des vecteurs et des matrices introduits dans le modèle, on néglige, par convention, les p premières variables de $\underline{\hat{e}}$.

2.3-4.1 La classe des solutions

On rappelle qu'étant donné la métrique euclidienne "classique" I_n choisie dans l'espace F de dimension n :

$$\underline{\hat{e}} = P\,\underline{y} = P\,\underline{e} \qquad \text{(notations matricielles)}$$

avec $$P^2 = P = P' = I_n - X'(XX')^{-1} X$$

et $$\text{var}(\underline{\hat{e}}) = \sigma^2 P$$

Considérons les partitions suivantes :

$\underline{e} = \begin{pmatrix} \underline{e}_0 \\ \underline{e}_1 \end{pmatrix}$ $\begin{matrix} p \times 1 \\ (n-p) \times 1 \end{matrix}$ $\underline{\hat{e}} = \begin{pmatrix} \underline{\hat{e}}_0 \\ \underline{\hat{e}}_1 \end{pmatrix}$ $\begin{matrix} p \times 1 \\ (n-p) \times 1 \end{matrix}$ ($\underline{\hat{e}}_0$ représente les p variables abandonnées de $\underline{\hat{e}}$).

$X' = \begin{pmatrix} X'_0 \\ X'_1 \end{pmatrix}$ où $\begin{matrix} X'_0 : p \times p \\ X'_1 : (n-p) \times p \end{matrix}$

$P = \begin{pmatrix} P_0 \\ P_1 \end{pmatrix} = \begin{pmatrix} P_{00} & P_{01} \\ P_{10} & P_{11} \end{pmatrix}$ où $\begin{matrix} P_0 = (P_{00}\ P_{01}) \\ P_1 = (P_{10}\ P_{11}) \end{matrix}$

avec

$$P_{10} = -X'_1 (XX')^{-1} X_0 \quad \text{et} \quad P_{11} = I_{n-p} - X'_1 (XX')^{-1} X_1$$

On a alors :

$$\underline{\hat{e}}_1 = P_1 \underline{y} = P_1 \underline{e} \quad \text{et} \quad \text{var}(\underline{\hat{e}}_1) = \sigma^2 P_{11}$$

Si X_0 est régulière (et on choisit les p variables à abandonner de telle sorte que X_0 le soit), alors P_{11} est aussi régulière, et on a :

$$P_{11}^{-1} = I_{n-p} + X'_1 (X_0 X'_0)^{-1} X_1$$

Soit $$\mathcal{C}_0 = \{C : R^{n-p} \longrightarrow R^{n-p} / P_{11} = C'C\}$$

L'ensemble $\mathcal{C}_0$ est non vide puisque P_{11} est une matrice réelle symétrique définie positive, et les matrices qui y appartiennent sont régulières.

Soit $\underline{e}_1^* = C'^{-1}(\underline{\hat{e}}_1)$ où $C \in \mathcal{C}_0$, alors $\underline{e}_1^*$ est une solution :

- $E(\underline{e}_1^*) = \underline{0}$
- $\text{var}(\underline{e}_1^*) = \sigma^2 C'^{-1} P_{11} C^{-1} = \sigma^2 I_{n-p}$

Soit $S = \{\underline{e}_1^* \in R^{n-p} / \underline{e}_1^* = C'^{-1} \underline{\hat{e}}_1$ où $C \in \mathcal{C}_0\}$; tout élément de S est une solution ; on peut définir aussi S de la façon suivante :

$$S = \{\underline{\tilde{u}} = U \underline{e}_1^* / U \in \mathcal{O} \quad \text{et} \quad \underline{e}_1^* = C'^{-1} \underline{\hat{e}}_1 \quad \text{avec} \quad C \in \mathcal{C}_0\},$$

$\mathcal{O}$ étant l'ensemble des matrices orthogonales d'ordre $n - p$.

Etant donné $\underline{e}_1^*$ appartenant à S, la variance du vecteur aléatoire $(\underline{e}_1 - \underline{e}_1^*)$ vaut,

$$\text{var}(\underline{e}_1 - \underline{e}_1^*) = \sigma^2 (2 I_{n-p} - C' - C)$$

Si $\qquad \underline{\tilde{u}} = U \underline{e}_1^*, \qquad$ où $\qquad U U' = I_{n-p}$,

alors $\qquad \text{var}(\underline{e}_1 - \underline{\tilde{u}}) = \sigma^2 (2 I_{n-p} - C' U' - U C)$

Proposition : tout vecteur aléatoire $\underline{\tilde{u}}$ de dimension $n - p$, linéaire en $\underline{y}$, de moyenne nulle et de matrice variance $\sigma^2 I_{n-p}$, appartient à l'ensemble S.

Le résultat est démontré dans LACOURLY (28, p. 83).

2.3-4.2 La solution optimale

Dans l'ensemble des solutions S, il existe une solution unique $\underline{e}_1^*$ optimale au sens du critère suivant :

$$\text{"var } (\underline{e}_1 - \underline{e}_1^*) \quad \text{minimum*}$$

$\underline{e}_1^*$, que Theil appelle le vecteur "BLUS", s'obtient à partir de la "racine carrée positive symétrique" B de P_{11}.

$$P_{11} = Q D_\lambda^2 Q'$$

où D_λ^2 est la matrice diagonale des valeurs propres de P_{11} ; ces valeurs propres sont toutes positives ; Q est la matrice des vecteurs propres associés ($Q Q' = Q' Q = I_{n-p}$).

$$P_{11} = Q D_\lambda Q' Q D_\lambda Q' = (Q D_\lambda Q')^2$$

où D_λ est la matrice diagonale des racines carrées, choisies positives, des valeurs propres de P_{11}.

$B = Q D_\lambda Q'$ est symétrique définie positive ; elle est unique : c'est la racine carrée positive symétrique de P_{11}.

$$\underline{e}_1^* = B^{-1} \underline{e}_1 \text{ est la solution optimale}$$

Le résultat est démontré dans THEIL (45, p. 248).

Remarque – La procédure de recherche du vecteur BLUS que nous avons utilisée est équivalente à employer des multiplicateurs de Lagrange avec le critère "$E(\|\underline{e}_1 - \underline{e}_1^*\|^2)$ minimum" sous les contraintes : $\underline{e}_1^*$ est linéaire en $\underline{y}$, sans biais et de variance $\sigma^2 I_{n-p}$.

2.3-4.3 Discussion

Le choix des p variables $\underline{\hat{e}}_i$ qui ne serviront pas à la construction du vecteur BLUS doit tenir compte :

– de la contrainte X_0' régulière ;

– de l'hypothèse alternative choisie pour le test de "non corrélation des erreurs".

Ainsi pour une série chronologique, si l'hypothèse alternative retenue est un processus stationnaire de Markov de premier degré, il est préférable de négliger les p premières ou les p dernières lignes de la matrice X' et des vecteurs utilisés dans le modèle ; ce test est alors bâti sur $n - p$ observations successives. Le rapport de Von Neumann calculé avec le vecteur BLUS permet de rejeter ou non l'hypothèse nulle.

Pour une hypothèse alternative qui spécifie que la variance de l'erreur augmente avec son numéro d'ordre, il est naturel de négliger p lignes associées à des observations de rang proche du rang médian. Mais dans ce cas, la somme des carrés des $(n - p)/2$ premières coordonnées du vecteur BLUS rapportée à la somme des carrés des $(n - p)/2$ dernières a pour distribution un F de Fisher, sous l'hypothèse nulle. Mais il semble excessif d'utiliser dans ce cas une procédure aussi complexe : en effet, le simple graphique des résidus avec la variable à expliquer suffit pour répondre au problème.

La construction du vecteur BLUS nécessite le calcul des valeurs propres et des vecteurs propres de la matrice régulière P_{11} d'ordre $n - p$. Du point de vue numérique, cette procédure devient lourde dès que le nombre de degré de liberté $n - p$ est grand ($n - p > 50$). Cependant, Theil a montré qu'il suffit d'extraire les valeurs propres et les vecteurs propres de P_{00}, qui sont au nombre de p, pour construire le vecteur BLUS (45). La procédure est alors numériquement beaucoup plus performante. (Le programme peut être obtenu auprès du "Center Mathematical Studies in Business and Economics" The University of Chicago. Chicago Illinois 60637).

2.3.5 Un exemple

Par commodité, on a repris les données de A.R. PREST (37) utilisées par DURBIN et WATSON (14 (2)).

Les variables sont les suivantes :

y = logarithme de la consommation de spiritueux par tête

x^2 = logarithme du revenu par tête

x^3 = logarithme du prix actualisé des spiritueux.

Ces trois séries chronologiques annuelles portent sur la période 1870-1938 en Grande-Bretagne.

On retient le modèle linéaire :

$$y_i = b_i + b_2\, x_i^2 + b_3\, x_i^3 + e_i \quad (i = 1{,}69)$$

où

b_1 est une constante

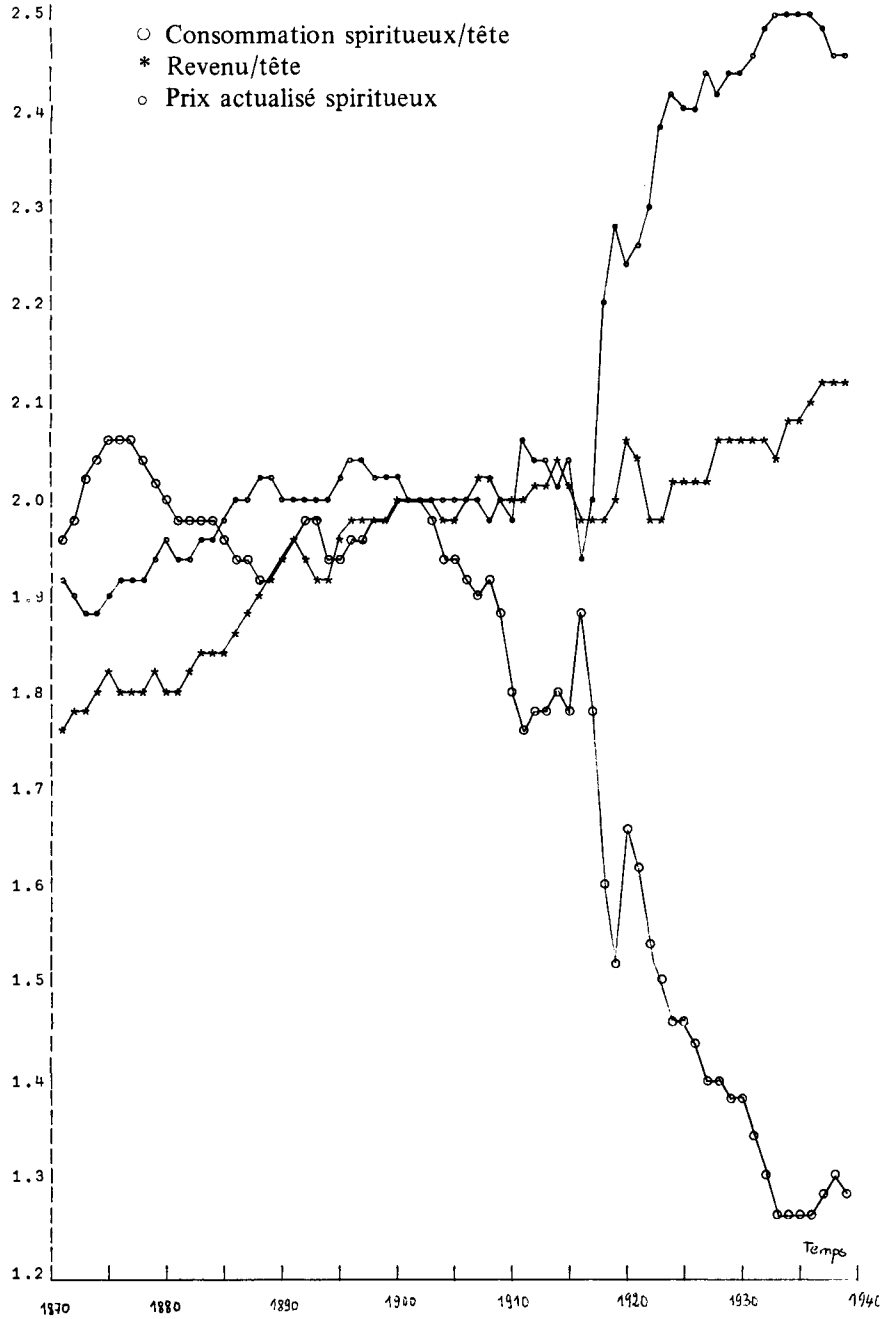

○ Consommation spiritueux/tête
* Revenu/tête
o Prix actualisé spiritueux
2.5
2.4
2.3
2.2
2.1
2.0
1.9
1.8
1.7
1.6
1.5
1.4
1.3
1.2
1870
1880
1890
1900
1910
1920
1930
1940
Temps

Figure 1

b_2 est l'élasticité du revenu

b_3 l'élasticité du prix

et on fait les hypothèses (1), (2), (3) (cf. 2.1) pour $\underline{e}$.

Représentation graphique des chroniques

Pour les chiffres, on se reportera à (14 (2)).

Matrice des corrélations

Temps	1			
y	− 0,90	1		
x^2	0,93	− 0,74	1	
x^3	0,88	− 0,98	0,74	1
	Temps	y	x^2	x^3

La régression

On donne pour chaque coefficient son estimation et l'écart-type estimé associé :

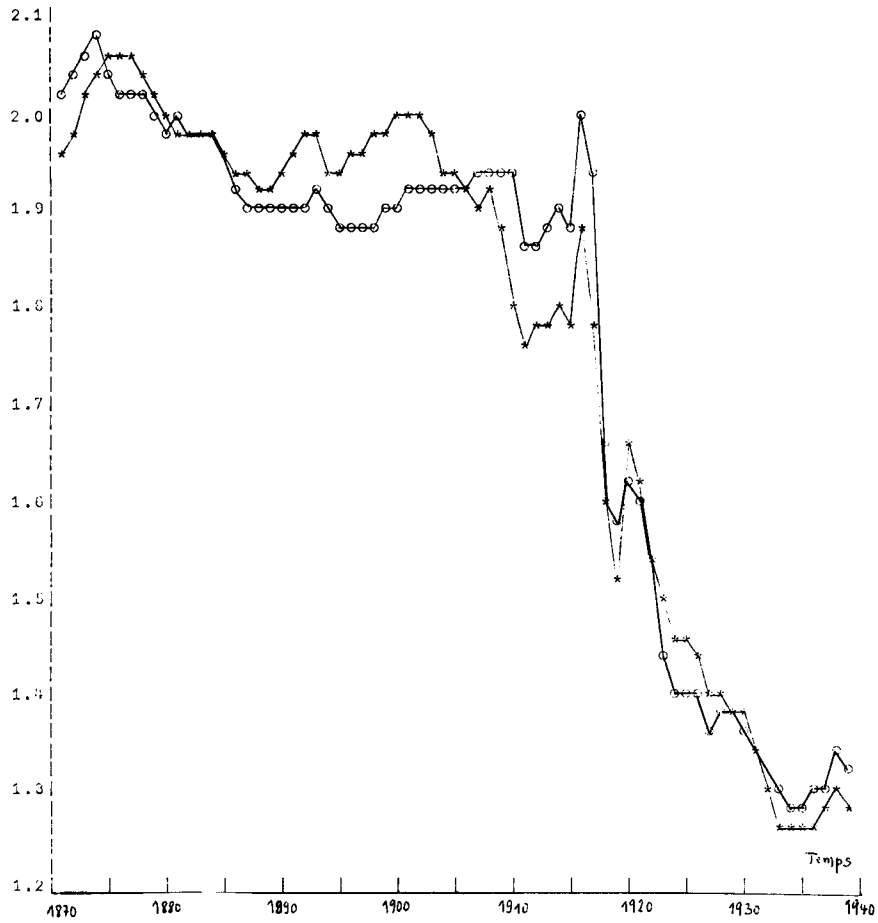
$\hat{b}_1$ = constante = 4,411 $\quad S_1 = 0{,}107$

$\hat{b}_2 = 0{,}030 \quad S_2 = 0{,}054$

$\hat{b}_3 = -1{,}275 \quad S_3 = 0{,}036$

$\hat{b}_2$ estimation de l'élasticité du revenu n'est pas significativement différente de 0.

Le modèle fournit une très bonne approximation de la consommation ($R^2 = 0{,}95$).



* Consommation/tête observée
○ Consommation/tête avec le modèle

Figure 2

Analyse des résidus

Les résidus présentent cependant des tendances qui vont conduire à redéfinir le modèle.

Deux causes peuvent expliquer ces tendances : le modèle est biaisé ou bien Γ a été supposée à tort valoir $\sigma^2 I_n$. On retiendra la seconde cause.

Utilisation du rapport de VON-NEUMAN

La très faible valeur du rapport :

$$Q = \frac{\sum_{i=2}^{69} (\hat{e}_i - \hat{e}_{i-1})^2}{\sum_{i=1}^{69} (\hat{e}_i)^2} = 0{,}233$$

après consultation de la table donnée par DURBIN et WATSON dans (14 (2)), conduit à rejeter l'hypothèse de "non-corrélation des erreurs" avec un risque < 1 %.

Les résidus BLUS

Dans ce cas, où la réponse a été fournie en utilisant la table de DURBIN et WATSON, le calcul des résidus BLUS ne s'avère pas nécessaire. Ils sont donnés ici pour l'exemple :

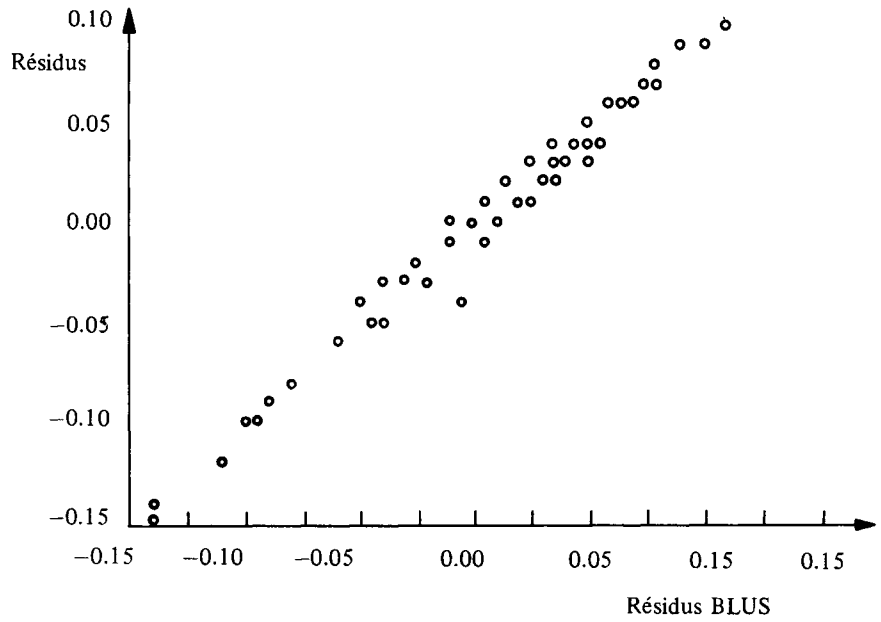
Le calcul imposant l'abandon de trois années, on a choisi les trois premières.

Les résidus BLUS $\{e_i^* \mid i = 4{,}69\}$ sont très proches des résidus initiaux $\{\hat{e}_i \mid i = 4{,}69\}$, ainsi qu'on peut le constater sur le graphique suivant :

Le rapport de VON-NEUMANN, calculé avec les résidus BLUS,

$$Q = \frac{\sum_{i=5}^{69} (e_i^* - e_{i-1}^*)^2}{\sum_{i=4}^{69} (e_i^*)^2} = 0{,}230$$

conduit encore à rejeter l'hypothèse de "non corrélation des erreurs" mais après consultation de la table donnée par HART (20).



On choisit un nouveau modèle :

Pour préciser une hypothèse alternative, on effectue des graphiques $\{(\hat{e}_i, \hat{e}_{i+1}) \mid i = 1,68\}$,$\{\{(\hat{e}_i, \hat{e}_{i+2}) \mid i = 1,67\}$ etc. La procédure est complétée par le calcul des premiers coefficients d'autocorrélation de la série des résidus, coefficients qui sont tous compris entre 0,8 et 0,9.

On va considérer que les erreurs $\{e_i \mid i = 1,69\}$ obéissent à un processus stationnaire de MARKOV du premier degré :

$$(A) \begin{cases} E(\underline{e}) = \underline{0} \\ \text{var}(\underline{e}) = \dfrac{\alpha^2}{1-\rho^2} \begin{pmatrix} 1 & \rho & \rho^2 & \dots & \rho^{68} \\ \rho & & & & \\ \cdot & & & & \\ \cdot & & & & \\ \rho^{68} & \dots & \dots & \dots & 1 \end{pmatrix} = \Gamma \\ \underline{e} \text{ suit une } LG_N\ (\underline{0}\ ;\ \Gamma) \end{cases}$$

ce qui est identique à :

$$(B)\begin{cases} e_i = \rho \cdot e_{i-1} + \eta_i \\ \quad \text{avec } \rho \text{ coefficient d'autocorrélation de rang 1} \\ \text{et} \quad E(\eta_i) = 0 \\ \quad V(\eta_i) = \alpha^2 \quad \forall\, i > 1 \\ \quad \operatorname{cov}(\eta_i , \eta_j) = 0 \quad \text{si} \quad j \neq i \\ \quad \eta_i \text{ suit une } LG(0 , \alpha) \end{cases}$$

Si on utilise le corps d'hypothèses (A),

Γ, matrice symétrique définie positive, peut être mise sous la forme $T.T'$ (au moyen d'une décomposition de DOOLITTLE-CHOLESKI par exemple).

On pose :

$$\underline{z} = T^{-1} . \underline{y} \qquad A = T^{-1} . X' \qquad \text{et} \qquad \underline{\epsilon} = T^{-1}\underline{e}$$

$$\underline{y} = X'\underline{b} + \underline{e} \quad \text{multiplié à gauche par } T^{-1}$$

devient :

$$\underline{z} = A\,\underline{b} + \underline{\epsilon} \qquad \text{avec} \quad \begin{cases} E(\underline{\epsilon}) = \underline{0} \\ V(\underline{\epsilon}) = I_n \\ \underline{\epsilon} \in LG_n(\underline{0} ; I_n) \end{cases}$$

Si on utilise le corps d'hypothèses (B), il est judicieux de réécrire :

$$y_i - \rho\, y_{i-1} = b_1(1 - \rho) + b_2(x_i^2 - \rho\, x_{i-1}^2) + b_3(x_i^3 - \rho\, x_{i-1}^3) + \eta_i$$

$$\text{avec} \quad \begin{cases} E(\eta_i) = 0 \\ V(\eta_i) = \alpha^2 \qquad \forall\, i = 2,69 \\ \operatorname{cov}(\eta_i , \eta_j) = 0 \qquad \text{si} \qquad j \neq i \\ \eta_i \in LG\ (0 , \alpha) \end{cases}$$

Les estimations obtenues peuvent différer selon la procédure choisie. Cette différence dépend de la valeur de ρ.

Il s'agit d'estimer ρ : on le fera à partir des résidus

$$Q = \frac{\sum_{i=2}^{69} (\hat{e}_i - \hat{e}_{i-1})^2}{\sum_{i=1}^{69} (\hat{e}_i)^2}$$ peut être considéré comme un estimateur

de $$\frac{V(e_i - e_{i-1})}{V(e_i)} = 2(1 - \rho)$$

On aura : $$\hat{\rho} = 1 - \frac{1}{2} Q = 1 - \frac{0{,}233}{2} = 0{,}8835$$

L'emploi de $\hat{\rho}$, calculé à partir de la première régression, conduira en moyenne à des valeurs de S_1, S_2, S_3 légèrement plus faibles que celles qu'on obtiendrait avec ρ.

Solution des moindres carrés pour le nouveau modèle

	$\hat{b}_1$	$\hat{b}_2$	$\hat{b}_3$	S_1	S_2	S_3
Procédure (*A*)	3,941	0,0445	−1,0708	0,163	0,02593	0,0716
Procédure (*B*)	3,953	0,0443	−1,0750	0,160	0,02616	0,0758

On constate des différences minimes entre les deux procédures. b_2 est estimé avec une meilleure précision que dans le cas précédent. Les résidus obtenus dans ce modèle ne présentent pas de tendances "suspectes".

Poser un nouveau modèle répond principalement au souci d'améliorer la fiabilité des estimations des coefficients, ce qui est naturel dans une optique de prévision. Mais avec ces nouvelles estimations, l'approximation de la consommation n'est pas meilleure que dans le premier cas. Ceci nous laisse supposer que le modèle considéré ici doit être légèrement biaisé.

2.3.6. Conclusion

L'utilisation de résidus BLUS n'est pas limitée au seul contrôle de la non-corrélation des erreurs ; on aurait pu d'abord présenter la méthode pour obtenir de tels résidus puis les utiliser pour ce problème particulier.

Le volume important de calcul nécessaire à l'obtention des résidus BLUS ou de tests plus récents (7) valorise le test de DURBIN et WATSON qui est construit à partir des résidus "ordinaires", et qui est simple à calculer. Ses auteurs ont en 1971 dans (14 (3)) contrôlé la procédure publiée en 1950 (qualité des tables et des approximations proposées). Cette procédure reste la plus simple et la plus efficace.

III – PROTECTION DE LA RÉGRESSION : SÉLECTION DE VARIABLES

3.1 LE PAS A PAS

Le problème est le suivant :

ayant calculé la régression $\underline{\hat{y}} = \hat{b}_1\underline{x}^1 + \ldots + \hat{b}_p\underline{x}^p$, on se demande si on n'aurait pas obtenu une régression "presque aussi bonne" en ne prenant qu'une partie seulement des p variables explicatives.

On se propose en somme de prédire y "au moindre coût".

On mesurera la qualité d'une régression par le rapport :

$$R = \frac{\|\underline{\hat{y}}\|_N}{\|\underline{y}\|_N} .$$

Sur "la qualité de la solution $\underline{\hat{y}}$", on consultera le chapitre 1.5

Tout d'abord, il est évident que *parmi les $2^p - 1$ régressions qu'il est possible de calculer, la meilleure au sens de R est celle qui contient les p variables.*

Trois stratégies sont possibles pour diminuer le nombre de variables explicatives en détériorant le moins possible la qualité de la régression.

3.1.1 Examen de toutes les régressions

Comme il y a $2^p - 1$ régressions à calculer, la méthode devient vite impraticable. Cependant, on pourra consulter (34, 41).

3.1.2 Elimination "descendante" des variables

– on part de la régression complète (p variables)

– on enlève la variable qui provoque la plus faible diminution du rapport F (voir cas particulier a. du § 1.5.2).

– parmi les $p - 1$ variables restantes, on enlève celle qui provoque la plus faible diminution de F.

et ainsi de suite . . .

On arrêtera le processus d'élimination en choisissant la régression qui précède l'élimination d'un caractère "significatif" au sens de F.

Opinion : On pratiquera de la sorte si on veut la régression complète. Cependant des ennuis se présentent si les $\underline{x}^j$ sont colinéaires (c'est-à-dire : si dim $(W) < p$) et alors XNX' n'est pas inversible.

3.1.3 Introduction "ascendante" des variables

3.1.3.1 La régression ascendante

On cherche la meilleure régression à une variable : la première variable sélectionnée sera évidemment celle dont le coefficient de corrélation linéaire avec $\underline{y}$ est le plus élevé ;

$$r(\underline{y}, \underline{x}^1) \qquad \text{est maximum en valeur absolue.}$$

On note $\hat{\underline{y}}(\underline{x}^1)$ la régression de $\underline{y}$ sur $\underline{x}^1$; de même $\hat{\underline{x}}^2(\underline{x}^1)$ est la régressiont de $\underline{x}^2$ sur $\underline{x}^1$.

On garde $\underline{x}^1$ et on ajoute $\underline{x}^2$ si le coefficient de corrélation linéaire

$$r(\underline{y} - \hat{\underline{y}}(\underline{x}^1), \underline{x}^2 - \hat{\underline{x}}^2(\underline{x}^1)) \qquad \text{est maximum en valeur absolue.}$$

On reconnaît ici le coefficient de corrélation partielle entre $\underline{y}$ et $\underline{x}^2$ sachant $\underline{x}^1$.

La procédure précédente est encore équivalente à rechercher $\underline{x}^2$ tel que :

$$R_{\underline{y}|\underline{x}^1,\underline{x}^2} \qquad \text{soit maximum}$$

où $R_{\underline{y}|\underline{x}^1,\underline{x}^2}$ est le coefficient de corrélation multiple de $\underline{y}$ en fonction de $\underline{x}^1$ et $\underline{x}^2$, ce qui est encore équivalent à rechercher $\underline{x}^2$ tel que le rapport F (cf. 1.5.2)

$$F = \frac{\|\hat{\underline{y}}(\underline{x}^1, \underline{x}^2) - \hat{\underline{y}}(\underline{x}^1)\|^2 / 1}{\|\underline{y} - \hat{\underline{y}}(\underline{x}^1, \underline{x}^2)\|^2 / n-2} = \frac{(n-2)(R^2_{\underline{y}|\underline{x}^1,\underline{x}^2} - R^2_{\underline{y}|\underline{x}^1})}{1 - R^2_{\underline{y}|\underline{x}^1,\underline{x}^2}}$$

soit maximum,

où $\hat{\underline{y}}(\underline{x}^1, \underline{x}^2)$ est la régression de $\underline{y}$ sur le plan $(\underline{x}^1, \underline{x}^2)$.

On garde $\underline{x}^1$ et $\underline{x}^2$ et on ajoute $\underline{x}^3$ si . . .

On arrêtera la procédure en choisissant la régression qui précède l'introduction d'un caractère "non significatif" au sens de F.

Remarque. La procédure précédente revient dans $F = R^n$ muni de la métrique $N = D_p$ à construire à partir des $\underline{x}^i$, suivant un certain ordre, une base orthogonale du sous-espace vectoriel W des combinaisons linéaires des $\underline{x}^i$ par le

procédé d'Hilbert-Schmidt ; si W_{i-1} désigne le sous-espace vectoriel engendré par $\{\underline{x}^1, \underline{x}^2, \ldots \underline{x}^{i-1}\}$, on sélectionne à l'étape i la variable $\underline{x}^i$, si $\underline{y}$ fait un angle minimum avec le sous-espace W_i engendré par $\{\underline{x}^1, \underline{x}^2 \ldots \underline{x}^{i-1}, \underline{x}^i\}$.

Opinion : Cette procédure est économique : elle évite en effet de travailler avec plus de caractères qu'il n'est nécessaire à chaque étape. Cependant, on ne tente pas de savoir en quoi l'introduction d'une variable modifie le rôle de celles précédemment introduites.

Ce souci est pris en compte dans la régression progressive.

3.1.3.2 La régression progressive (Stepwise regression)

C'est une régression ascendante modifiée de la façon suivante : après chaque introduction de variable, et avant d'en introduire une nouvelle, on enlève s'il y a lieu la ou les variables qui ne sont pas "significatives" et on arrête le processus d'introduction quand la dernière variable introduite n'est pas "significative" ; "significatif" est sous-entendu chaque fois " au sens de F'' (cas particulier a, paragraphe 1.5.2).

L'algorithme de la régression progressive est dû à EFROYMSON (15).

Les procédures précédentes qui ont été brièvement exposées sont détaillées dans DRAPER and SMITH (13, chapitre 6).

3.1.3.3 Un autre critère pour l'introduction "ascendante" : le critère C_q

On introduit la structure probabiliste.

Dans l'espace des caractères $F = R^n$, muni d'une métrique euclidienne N, on considère la distance :

$D^2 = ||\underline{y}^* - E(\underline{y})||_N^2$; c'est le carré de la norme du vecteur $\underline{y}^* - E(\underline{Y})$ avec $E(\underline{y}) = X'\underline{\beta}$ et $\underline{y}^* = X'(\underline{b}^*)$ où $\underline{b}^*$ est un estimateur du vecteur des paramètres $\underline{\beta}$.

$$E(D^2) = E(||\underline{y}^* - E(\underline{y})||_N^2) = E(||\underline{y}^* - E(\underline{y}^*)||_N^2) + ||E(\underline{y}^*) - E(\underline{y})||_N^2$$

La quantité $E(D^2)$ peut servir de critère dans une régression pas à pas (18, 21).

En effet, en ajoutant des variables explicatives dans le modèle, on diminue, en général, le biais des estimations, mais on risque d'augmenter les variances des estimations et l'erreur totale moyenne $E(D^2)$, car on a pu ajouter des variables liées aux variables initialement utilisées dans le modèle.

On pose : $\Gamma = \text{var}(\underline{y}) = \sigma^2 \cdot \Gamma_c$ et on choisit $N = \Gamma_c^{-1}$

alors :

$$E(\|\underline{y}^* - E(\underline{y}^*)\|_N^2) = q\,\sigma^2$$

où q est le nombre de variables explicatives introduites dans le modèle : la constante est incluse dans ce nombre et $q \leqslant p$ (voir le paragraphe 1.5.2).

Si $\underline{b}^*$ est un estimateur sans biais de $\underline{\beta}$, $\underline{y}^* = X'(\underline{b}^*)$ est sans biais pour $X'\underline{\beta}$ et $E(D^2) = q\,\sigma^2$.

Il s'agit d'estimer $\dfrac{E(D^2)}{\sigma^2}$. Avec une bonne estimation de σ^2, s^2, MALLOWS (32) propose la quantité C_q :

$$C_q = \frac{1}{s^2} \cdot (\|\underline{y} - \underline{y}^*\|_N^2 - (n-q)\,s^2 + q\,s^2)$$

où le biais $\|E(\underline{y}) - E(\underline{y}^*)\|_N^2$ est estimé par $\|\underline{y} - \underline{y}^*\|_N^2 - (n-q)s^2$ où $E(\|\underline{y}^* - E(\underline{y}^*)\|_N^2)$ est estimé par $q\,s^2$.

En général s^2 est obtenu à partir des résidus de la régression complète.

Si l'estimateur $\underline{y}^*$ est sans biais, C_q est voisin de q.

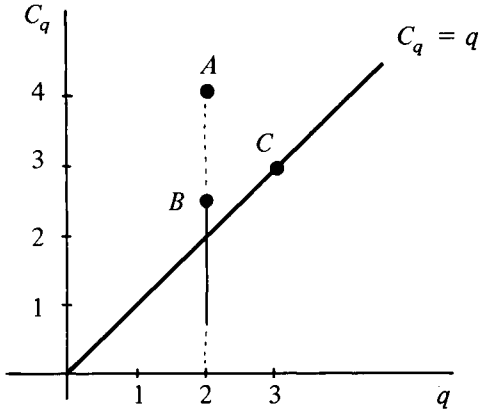
GORMAN et TOMAN donnent une méthode graphique pour comparer des régressions à l'aide de la quantité C_q (18). Ils fournissent un exemple avec 0, 1 ou 2 variables explicatives ; les valeurs de C_q sont données dans le tableau suivant :

Variables	q	C_q
$y = b_0$	1	84,6
$y = b_0 + b_1x^1$	2	4,1
$y = b_0 + b_2x^2$	2	2,5
$y = b_0 + b_1x^1 + b_2x^2$	3	3,0

Les points $(q\,,\,C_q)$ sont portés sur un graphique : tous ces points sont au voisinage de la droite $C_q = q$, sauf si les estimations sont très biaisées auquel cas les points sont au-dessus de la droite.

Ainsi, le modèle avec les deux variables explicatives x^1 et x^2 (le point C) fournit des estimations non biaisées, mais le modèle avec seulement la variable x^2 est "meilleur" au sens du critère "C_q" (le point B est plus bas que le point C) car bien que les estimations soient biaisées, leurs variances sont plus faibles

ainsi que "l'erreur totale". Dans le cadre prévisionnel, on aura donc intérêt à utiliser la "meilleure" régression au sens du critère "C_q minimum".



Remarque :

Pour la métrique euclidienne classique $N = I_n$

$$E(D^2) = \sum_{i=1}^{n} \text{var}(y_i^*) + \sum_{i=1}^{n} (E(y_i^*) - E(y_i))^2$$

$$E(D^2) = \sum \text{(variances)} + \sum \text{(biais)}^2$$

c'est-à-dire $E(D^2)$ est composé de deux termes : l'un est la somme des variances des estimations, l'autre la somme des carrés des biais de ces estimations y_i^* par rapport à $E(y_i)$.

$E(D^2)$ représente "l'erreur totale" moyenne faite en estimant $E(\underline{y})$ par $\underline{y}^*$.

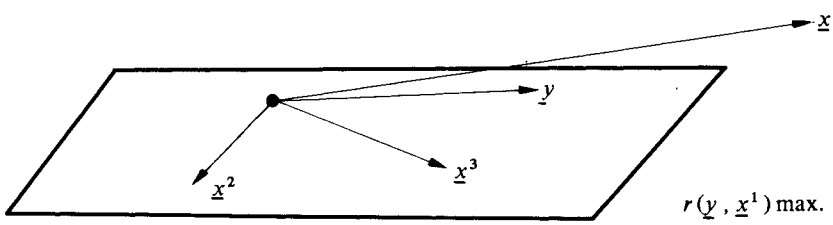
3.1.4 Conclusion

Il est à noter que ces procédures de sélection ne garantissent pas l'obtention de la meilleure régression à chaque stade, c'est-à-dire de celle qui, pour un nombre fixé de variables explicatives, fournit le coefficient de corrélation multiple le plus élevé.

On peut le constater sur l'exemple très simple suivant.

Dans l'espace des caractères F, les variables $\underline{y}$, $\underline{x}^2$, $\underline{x}^3$ sont coplanaires et $\underline{x}^1$, n'appartenant pas au plan, est telle que son coefficient de corrélation linéaire avec $\underline{y}$ est le plus fort.

Graphiquement, on a :



La meilleure régression avec une variable est celle de $\underline{y}$ sur $\underline{x}^1$. La meilleure régression avec deux variables est celle de $\underline{y}$ sur $\underline{x}^2$ et $\underline{x}^3$.

On trouve dans le tableau suivant, les choix successifs des différentes méthodes :

élimination "descendante"	introduction "ascendante"	régression "progressive"
(x^1, x^2, x^3)	x^1	x^1
(x^2, x^3)	(x^1, x^2) ou (x^1, x^3)	(x^1, x^2) ou (x^1, x^3)
x^2 ou x^3	(x^1, x^2, x^3)	(x^1, x^2, x^3)
		(x^2, x^3)

Les différences dépendent du seuil de signification choisi pour le rapport F et deviennent plus nombreuses quand le nombre de variables augmente.

3.2 REGRESSION SUR VARIABLES ORTHOGONALES

Dans cette méthode, on cherche à éliminer les inconvénients résultant de la colinéarité des variables explicatives, comme dans la "ridge régression" (cf. paragraphe 4.3) ; la discussion ne se fera pas au niveau du critère mais au niveau du choix des variables explicatives.

L'objectif ici est de quantifier l'effet d'une ou plusieurs variables explicatives ; considérons le modèle :

$$y = b_1 x^1 + b_2 x^2 + \ldots + b_p x^p + e \qquad (1)$$

$b_2\ x^2$ peut être appelé "effet" de la variable x^2, si les effets des différentes variables explicatives x^j sont additifs.

Pour que cette condition soit vérifiée, il faut que les variables x^j soient orthogonales.

Cette condition est rarement réalisée, les variables étant sélectionnées a priori sans tenir compte de leurs corrélations éventuelles. Compte tenu de cette remarque, pour avoir une idée à la fois des effets et des liaisons entre les variables explicatives, on cherchera à utiliser un ensemble de variables orthogonales "équivalent" à l'ensemble des variables explicatives $\{x^j | j = 1, p\}$ (47).

3.2.1 Exemple : étude des "bénéfices" de l'instruction

L'augmentation très rapide des dépenses pour l'enseignement dans tous les pays a conduit à analyser le taux de rendement de l'éducation. Une des études concerne les "bénéfices" apportés par l'instruction. Ces bénéfices étant décrits à l'aide du revenu individuel, un modèle linéaire a été posé pour tenter de mettre en évidence l'influence du facteur éducation sur le revenu [16, p. 18].

Soit le modèle :

$$y = \Sigma\, b_j x^j + e$$

où la variable à expliquer y représente le revenu d'un individu et les variables explicatives x^j sont réparties en quatre catégories :

– facteurs d'éducation : nombre d'années d'études, diplômes etc.

– facteurs professionnels : catégorie socio-professionnelle, expérience acquise dans le métier, situation de marché etc.

– facteurs du milieu : catégorie socio-professionnelle du père, lieu de résidence pendant les études, niveau de revenu des parents, race, religion etc.

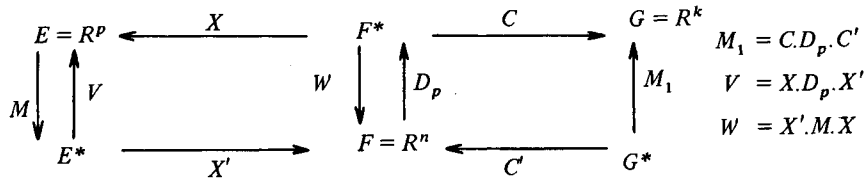
– qualités naturelles : intelligence, motivation, volonté d'arriver etc.

L'objectif est de mesurer à partir de ce modèle linéaire la part de l'éducation expliquant le revenu. Mais les indicateurs du milieu social, du milieu professionnel et des qualités naturelles et ceux de l'éducation sont corrélés entre eux. Les effets des différentes variables ne sont donc pas additifs et chaque coefficient b_j associé au modèle ne décrit qu'un effet partiel de la variable x^j.

L'influence du niveau d'instruction sur le revenu étant le but de l'étude, on pourrait chercher à éviter l'inconvénient de la colinéarité en contrôlant

toutes les autres variables. Mais cela suppose un très gros échantillon, ce qui en pratique est rarement réalisé. La régression sur variables orthogonales, qui ne fournira pas exactement une mesure de l'effet du facteur éducation, est pourtant une technique qui aidera peut-être à mettre en évidence certaines causalités.

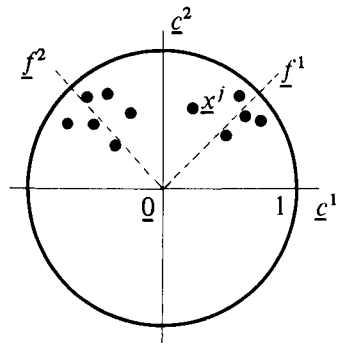
3.2.2 Stratégie



La métrique diagonale $N = D_p$ est choisie dans l'espace des caractères F et la métrique M est choisie dans E.

L'analyse en composantes principales effectuée sur les variables x^j permet d'associer à l'ensemble des variables explicatives un système équivalent de k variables orthogonales. En effet, si on note $\{\underline{c}^1, \underline{c}^2, \ldots, \underline{c}^k\}$ les k composantes principales associées aux valeurs propres non nulles de $V \,.\, M$, le sous espace $X'(E^*)$ engendré par les vecteurs $\underline{x}^j$ admet pour base orthogonale le système $\{\underline{c}^1, \ldots, \underline{c}^k\}$.

Ayant exprimé les variables $\underline{x}^j$ dans la base des composantes principales, on peut, à l'aide de graphiques, se rendre compte des dépendances linéaires entre les variables explicatives : si toutes les variables sont centrées et si les deux plus grandes valeurs propres expliquent une part importante de la variabilité totale, on peut faire un graphique analogue à celui de la figure ci-dessous par exemple.



Dans ce graphique, les variables explicatives, qui ont été normées à 1 (variables réduites) et projetées sur le plan engendré par les deux premières composantes principales, apparaissent comme des points à l'intérieur du cercle de centre $\underline{0}$ et de rayon 1 :

– les variables proches du cercle peuvent être considérées comme des combinaisons linéaires de $\underline{c}^1$ et $\underline{c}^2$;

– deux variables proches, et proches du cercle sont presque colinéaires ;

– deux variables orthogonales, combinaisons linéaires de $\underline{c}^1$ et $\underline{c}^2$, sont aux extrémités d'un arc de 90°.

Le modèle

$$\underline{y} = C'(\underline{a}) + \underline{e} \tag{2}$$

où $$C' = (\underline{c}^1, \underline{c}^2, \ldots, \underline{c}^k)\,,$$

compte tenu de l'équivalence entre les 2 ensembles

$$\{\underline{x}^1, \ldots, \underline{x}^p\} \quad \text{et} \quad \{\underline{c}^1, \ldots, \underline{c}^k\}\,,$$

peut être considéré comme équivalent au modèle initial (1).

On a en effet :

$$C'(\underline{a}) = X'(\underline{b})$$

Le modèle (2) est sur le plan du calcul, grâce à l'orthogonalité des composantes principales, beaucoup plus facile à mettre en œuvre : il suffit de projeter le vecteur $\underline{y}$ sur chacune des composantes principales $\underline{c}^j$. Dans ce cas, le coefficient de régression associé à la variable c^j n'est autre que :

$$a_j = \frac{D_p(\underline{y}\,,\,\underline{c}^j)}{D_p(\underline{c}^j)}$$

Si l'on ne cherche pas à quantifier les effets des variables x^j, la solution précédente convient. Par contre, si l'on veut préciser l'effet d'une variable particulière x^j, il est préférable d'utiliser pour le modèle un système orthogonal

$$\{\underline{f}^1, \underline{f}^2, \ldots, \underline{f}^k\} \qquad \text{avec}$$

$\underline{f}^1 = \underline{x}^j$, par exemple, obtenu par rotation à partir du système des composantes principales (cf. figure ci-dessus).

$$\underline{y} = \sum_{j=1}^{k} d_j \underline{f}^j + \underline{e} \tag{3}$$

Le terme $d_1 \underline{f}^1$ représente alors les effets conjugués de la variable x^j et partiellement des variables corrélées à x^j ; si les corrélations sont positives, $d_1 \underline{f}^1$ représente l'effet maximum de la variable $\underline{x}^j$.

Il est évident que l'on n'obtient pas ainsi exactement l'effet cherché mais simplement des éléments qui permettront peut-être pour une analyse ultérieure de démêler l'écheveau des causalités imbriquées.

Remarque : Les rotations peuvent être effectuées de façon automatique en utilisant des critères comme le "varimax" de KAISER (25) ou le "quartimax" de FERGUSON (17).

IV – PROTECTION DE LA RÉGRESSION PAR UTILISATION DE CONTRAINTES

4.1 INTRODUCTION

Si le nombre de variables explicatives est grand relativement au nombre d'observations, ou si les variables explicatives sont très corrélées, la régression n'a souvent pas de sens.

Pour protéger la régression, on peut alors imposer des contraintes, résultant de connaissances supplémentaires sur les données.

Trois types de contraintes sont exposés ici :

– *contraintes de type linéaire,* si pour avoir un sens, les coefficients de régression doivent être compris entre deux bornes connues (par exemple être positifs) (cf. 4.2)

– *contraintes de type non linéaire* où l'on impose au vecteur des coefficients de régression de ne pas avoir une norme trop élevée, de façon à éviter une reconstitution illusoire de la variable à expliquer par sommation de termes de valeurs absolues élevées et de signes contraires (cf. 4.3).

– *contraintes de type non linéaire,* quand on a des variables entachées d'erreurs, la variance résiduelle devant être supérieure à la variance d'erreur (cf. 4.4).

Le lecteur se rendra compte aisément dans les chapitres suivants que les trois cas présentés ont une formalisation mathématique et une interprétation géométrique communes ; ces cas ne sont donc, du point de vue mathématique, que trois variantes d'un même problème. Mais il serait erroné de les considérer comme identiques car ils diffèrent quant à l'esprit.

4.2 REGRESSION SOUS CONTRAINTES LINEAIRES

4.2.1 Le problème

Dans le cadre du Modèle Linéaire, les coefficients à estimer ont en général un sens, et de ce fait, on est souvent conduit à imposer des contraintes

sur ces coefficients. Dans ce cas, l'estimation par la méthode des moindres carrés n'est pas toujours satisfaisante ; aussi est-on amené à utiliser des méthodes de programmation quadratique.

Supposons que pour avoir un sens, les coefficients du modèle β_j doivent être compris entre deux bornes α_j et γ_j, et que les valeurs de la variable à expliquer y (et donc des valeurs approchées y_i^+ de chaque observation y_i) soient comprises entre deux bornes c et d.

On a donc le modèle :

$$\underline{y} = X'(\underline{\beta}) + \underline{e} \tag{1}$$

où $\underline{\beta} \in E^* = (R^p)^*$ est le vecteur des coefficients à estimer

$$X' = (\underline{x}^1 \ldots \underline{x}^p) \qquad \text{avec} \qquad \underline{x}^j \in F = R^n$$

avec

$$\left.\begin{array}{ll} \forall j : 1 \leqslant j \leqslant p & \alpha_j \leqslant \beta_j \leqslant \gamma_j \\ \forall i : 1 \leqslant i \leqslant n & c \leqslant (X'\underline{\beta})_i = \sum_{j=1}^{p} \beta_j x_i^j \leqslant d \end{array}\right\} \tag{2}$$

Les inégalités (2) s'écrivent de façon vectorielle

$$\left.\begin{array}{l} \underline{\alpha} \leqslant \underline{\beta} \leqslant \underline{\gamma} \\ c\underline{j}_n \leqslant X'(\underline{\beta}) \leqslant d\underline{j}_n \end{array}\right\} \tag{3}$$

$\underline{\alpha}$ (respectivement $\underline{\gamma}$) désignant le vecteur de E^* de composantes α_i (resp. γ_i), $\underline{j}_n$ désignant le vecteur de F dont toutes les composantes valent 1, l'inégalité $\underline{z} \leqslant \underline{t}$ entre deux vecteurs de R^q étant équivalente aux q inégalités entre les coordonnées $z_i \leqslant t_i$, $i = 1, q$.

Soit

$$W = \{\underline{z} = X'(\underline{b}) \mid \underline{b} \in E^*\}$$

$$C = \{\underline{b} \mid \underline{b} \in E^* : \underline{\alpha} \leqslant \underline{b} \leqslant \underline{\gamma} \; ; \; c\underline{j}_n \leqslant X'(\underline{b}) \leqslant d\underline{j}_n\}$$

$$D = X'(C) = \{\underline{z} = X'(\underline{b}) \mid \underline{b} \in E^* : \underline{\alpha} \leqslant \underline{b} \leqslant \underline{\gamma} \; ; \; c\underline{j}_n \leqslant \underline{z} \leqslant d\underline{j}_n\}$$

W est le sous-espace vectoriel engendré par les $\{\underline{x}^j \mid j = 1, p\}$;

D est le convexe fermé de W des vecteurs $X'(\underline{b})$, où $\underline{b}$ appartient au convexe fermé C.

$F = R^n$ étant muni de la métrique N, on désire minimiser $\| \underline{y} - X'(\underline{b}) \|_N^2$, sous la contrainte $\underline{b} \in C$, c'est-à-dire $X'(\underline{b}) \in D$.

4.2.2 La solution

La solution $\underline{b}^+$ est donc telle que $\underline{y}^+ = X'(\underline{b}^+)$ est la projection orthogonale de $\underline{y}$ sur D.

Si $\hat{\underline{y}}$ désigne la solution des moindres carrés sans contraintes, c'est-à-dire la projection Γ^{-1}-orthogonale de $\underline{y}$ sur W, alors :

ou bien $\hat{\underline{y}} \in D$ et donc $\underline{y}^+ = \hat{\underline{y}}$

ou bien $\hat{\underline{y}} \in D$ auquel cas $\underline{y}^+ \neq \hat{\underline{y}}$; $\underline{y}^+$ appartenant à la frontière de D s'obtient en projetant $\hat{\underline{y}}$ sur la frontière de D (application du théorème des trois perpendiculaires).

Dans ce dernier cas, on doit donc chercher à minimiser :

$$||\underline{y} - X'(\underline{b})||^2_{N=\Gamma^{-1}}$$

qui est une fonction quadratique des composantes de $\underline{b}$ avec $\underline{b} \in C$; c'est un problème classique de programmation quadratique que l'on pourra résoudre par un processus itératif, comme la méthode du gradient réduit de WOLFE, ou des méthodes dérivées (cf. CAZES – TURPIN (9)).

Pour la recherche de coefficients de régression positifs, l'algorithme D'ESOPO a été programmé par P. CAZES.

4.3 LA "RIDGE REGRESSION"

Dans tout ce paragraphe, on est dans le cadre du Modèle Linéaire ; le choix de N est Γ^{-1}.

4.3.1 Le problème de la quasi-colinéarité des variables explicatives

Considérons dans $E^* = (R^p)^*$ muni de la métrique $M = T I_p T'$, la variable aléatoire :

$$L^2 = ||\underline{b} - \underline{\beta}||^2_M$$

où $\underline{b}$ est un estimateur quelconque de $\underline{\beta}$.

On a alors :

$$E(L^2) = E(||\underline{b} - \underline{\beta}||^2_M) = E(||\underline{b} - E(\underline{b})||^2_M) + ||E(\underline{b}) - \underline{\beta}||^2_M$$

$$= \gamma_1(\underline{b}) + \gamma_2(\underline{b})$$

où :

$$- \gamma_1(\underline{b}) = E(\|\underline{b} - E(b)\|_M^2) = E(\|\underline{b}\|_M^2) - \|E(\underline{b})\|_M^2 = \text{trace } M.V(\underline{b})$$

représente le terme de variance (égal à la somme des variances des b_i si $M = I_p$), $V(\underline{b})$ désignant la matrice variance de $\underline{b}$.

$- \gamma_2(\underline{b}) = \|E(\underline{b}) - \underline{\beta}\|_M^2$ représente le biais qui est nul si $\underline{b}$ est sans biais.

Quand on se restreint à la classe des estimateurs sans biais qui sont linéaires en $\underline{y}$, l'estimateur des moindres carrés $\hat{\underline{b}}$ est l'unique estimateur (si X' est injective) rendant minimale $E(L^2)$ quel que soit M (cf. corollaire de Gauss-Markov § 1.3.3.3)

Pour cet estimateur, on a :

$$E(\hat{L}^2) = \text{trace } (M\,V(\hat{\underline{b}})) = \text{trace } (M\,(X\,\Gamma^{-1}X')^{-1})$$

$$= \text{trace } (T'(X\,\Gamma^{-1}X')^{-1}\,T)$$

Si les variables explicatives sont quasiment colinéaires, X' restant toujours injective, $X\,\Gamma^{-1}X'$ a des valeurs propres petites, et la trace de $T'(X\,\Gamma^{-1}X')^{-1}\,T$ est grande ; $E(\hat{L}^2)$ est donc grand.

En particulier, si $M = I_p$, on a :

$$E(\hat{L}^2) = E(\|\hat{\underline{b}}\|^2) - \|E(\hat{\underline{b}})\|^2 = \sum_{i=1}^{p} \text{var } \hat{b}_i$$

$E(\|\hat{\underline{b}}\|^2)$ est alors élevée ; de plus, certains des $\hat{b}_i$ ont des variances fortes (instabilité des $\hat{b}_i$).

C'est bien ce qu'on constate en pratique : le vecteur de régression est de norme élevée et les coefficients de régression ont de fortes variances.

L'estimateur obtenu $\hat{\underline{b}}$ est alors mauvais : le modèle, pour prévoir y à partir des variables x^j est illusoire et valable uniquement pour les données considérées ; en effet, comme on l'a déjà signalé, les composantes y_i de $\underline{y}$ sont expliquées à l'aide d'une somme dont les termes sont de module élevé (bien plus grand que $|y_i|$), de signes contraires et se retranchant.

Un estimateur $\underline{b}^+$ situé au voisinage de $\hat{\underline{b}}$ ne changera pratiquement pas la somme résiduelle des carrés $\|\underline{y} - X'\underline{b}\|_N^2$ qui passe par un minimum pour $\underline{b} = \hat{\underline{b}}$ et $\underline{b}^+$ pourra être plus satisfaisant que $\hat{\underline{b}}$ dans le sens ou $E(L^{+2})$, valeur de $E(L^2)$ pour $\underline{b}^+$, sera plus petite que $E(\hat{L}^2)$.

Ceci implique (du moins pour un modèle gaussien) que :

1/ $\underline{b}^+$ est biaisé, puisque $\hat{b}$ rend minimum $E(L^2)$ dans la classe des estimateurs sans biais de $\underline{\beta}$;

2/ $\gamma_1(\underline{b}^+) \leqslant \gamma_1(\hat{\underline{b}})$ car $E(L^{+2}) = \gamma_1(\underline{b}^+) + \gamma_2(\underline{b}^+) < E(\hat{L}^2) = \gamma_1(\hat{\underline{b}})$

d'où l'on déduit que $\mathrm{Var}(\hat{\underline{b}}) - \mathrm{Var}(\underline{b}^+)$ est une matrice définie ou semi-définie positive ; ceci assure une stabilité plus grande des estimateurs b_i^+ de β_i.

Le critère $E(L^2)$ minimum ne pourra toutefois être appliqué directement pour déterminer $\underline{b}^+$; en effet, $E(L^{+2})$ dépend $\underline{\beta}$ qui est inconnu, alors que $E(\hat{L}^2)$ ne dépendait pas de $\underline{\beta}$.

Pour pouvoir trouver un estimateur $\underline{b}^+$ meilleur que $\hat{\underline{b}}$ au sens de $E(L^2)$ et qui modifie peu la somme résiduelle des carrés, on cherchera un estimateur biaisé de $\underline{\beta}$ proche de $\hat{\underline{b}}$ par la procédure exposée ci-dessous.

Puisque

$$E(\hat{\underline{b}}) = \underline{\beta} \quad \text{et} \quad \|\underline{y}^+ - \hat{\underline{y}}\|_N^2 = \|X'\underline{b}^+ - X'\hat{\underline{b}}\|_N^2 = \|\underline{b}^+ - \hat{\underline{b}}\|_{XNX'}^2$$

on mesurera le biais entre $\underline{b}^+$ et $\underline{\beta}$ par $\|\underline{y}^+ - \hat{\underline{y}}\|_N^2$

On cherche alors parmi tous les estimateurs $\underline{b}^+$ tels que $\|\hat{\underline{y}} - \underline{y}^+\|_N = \epsilon$ celui qui rend la norme de $\underline{b}^+$, $\|\underline{b}^+\|_M$ minimum, critère assez naturel d'après les considérations faites plus haut.

4.3.2 La solution

Introduisant le multiplicateur de Lagrange 1/k, on rendra minimum

$$\|\underline{b}^+\|_M^2 + \frac{1}{k}(\|\hat{\underline{y}} - \underline{y}^+\|_N^2 - \epsilon^2) = \|\underline{b}^+\|_M^2 + \frac{1}{k}(\|X'\hat{\underline{b}} - X'\underline{b}^+\|_N^2 - \epsilon^2)$$

d'où l'on déduit par dérivation :

$$M\underline{b}^+ - \frac{1}{k}XN(X'\hat{\underline{b}} - X'\underline{b}^+) = \underline{0}$$

soit :

$$\boxed{(XNX' + kM)\underline{b}^+ = XNX'\hat{\underline{b}} = XN\underline{y}} \qquad (4)$$

k étant déterminé de telle sorte que $\|\hat{\underline{y}} - \underline{y}^+\|_N^2 = \epsilon^2$

Remarque : Si on s'impose la norme de $\underline{b}^+$, et si on cherche à minimiser $\|\underline{y} - \underline{y}^+\|_N$, ce qui revient à minimiser $\|\hat{\underline{y}} - \underline{y}^+\|_N$ on obtient encore l'équation (4).

En fait, il est plus logique d'imposer $\|\underline{b}^+\|_M^2 \leqslant a^2$ que d'imposer le biais ; alors :

ou bien $\|\hat{\underline{b}}\|_M^2 \leqslant a^2$, et la solution des moindres carrés est valable ;

ou bien $\|\hat{\underline{b}}\|_M^2 > a^2$, auquel cas $\underline{b}^+$ sera tel que $\|\underline{b}^+\|_M^2 = a^2$, et on cherchera sous cette condition le vecteur $\underline{b}^+$ qui minimise

$$\|\underline{y}^+ - \hat{\underline{y}}\|_N^2 = \|\underline{b}^+ - \hat{\underline{b}}\|_{XNX'}^2 .$$

Géométriquement :

posant $M_1 = XNX'$ et supposant X' injective, dans $E^* = (R^p)^*$ muni de la métrique M_1, on cherche le point $\underline{b}^+$ le plus proche de $\hat{\underline{b}}$, $\underline{b}^+$ appartenant à l'éllipsoïde C de centre $\underline{0}$, et d'équation $\|\underline{b}^+\|_M^2 = a^2$; $\underline{b}^+$ est donc la projection de $\hat{\underline{b}}$ sur C.

La recherche de $\underline{b}^+$ peut se faire à l'aide de techniques de linéarisation. En fait, il est plus simple et plus rapide d'employer un processus itératif, en appliquant la formule (4) pour diverses valeurs de k (0, 0.1, 0.2, ...) et en comparant pour chaque valeur de k, la norme du vecteur $\underline{b}^+(k)$ ainsi obtenu à a. Ce processus est justifié par le résultat suivant.

Posant $\underline{b}^+(k) = \underline{b}_k = (XNX' + kM)^{-1} XN\underline{y}$ et $f(k) = \|\underline{b}_k\|_M^2$, on montre, en se basant sur le spectre de $XNX'M^{-1}$, qu'il existe une seule valeur k_0 de k telle que $f(k_0) = \|\underline{b}_{k_0}\|_M^2 = a^2$ et que $g(k) = \|\underline{b}_k - \hat{\underline{b}}\|_{XNX'}^2$ est minimum ; de plus k_0 est positif (cf. Annexe).

La façon d'opérer précédente est également intéressante si on ne sait pas a priori quelle limitation donner à $\|\underline{b}^+\|_M$; on appliquera alors la formule (4) pour $k = 0, 0.1, 0.2, \ldots$, et on s'arrêtera quand tous les coefficients de régression se seront stabilisés.

Cette règle empirique peut tout aussi bien se justifier si on revient au critère statistique utilisé, à savoir : diminuer $E(L^2)$ avec un estimateur biaisé $\underline{b}^+$ de $\underline{\beta}$ qui soit proche de $\hat{\underline{b}}$. $E(L^2(k))$ s'interprète comme "l'erreur totale" moyenne faite en prenant $\underline{b}^+(k)$ comme estimation de $\underline{\beta}$

$$E(L^2(k)) = E(\|\underline{b}^+(k) - \underline{\beta}\|_M^2) = \gamma_1(\underline{b}^+(k)) + \gamma_2(\underline{b}^+(k))$$

Dans le cas où $M = I_p$, HOERL et KENNARD (22, 23) ont montré, en utilisant le spectre de XNX', qu'au voisinage de 0 :

- $E(L^2(k))$ est une fonction décroissante de k pour $k > 0$;
- $E(L^2(k))$ passe par un minimum en k_1 ou k_1 est > 0 ;
- $E(L^2(k_1)) < E(L^2(0)) = E(\hat{L}^2) = \text{trace}\,(X\Gamma^{-1}X')^{-1}$.

La généralisation au cas où M est une métrique quelconque est immédiate.

Du point de vue de la métrique M, il semble logique, soit de prendre la métrique unité, soit la métrique diagonale dont la matrice a les mêmes éléments diagonaux que XNX'.

4.4 REGRESSION SUR VARIABLES ENTACHEES D'ERREURS

4.4.1 Le problème

On suppose ici que, du fait des conditions d'observation, chaque variable observée (variables explicatives $x^1, \ldots, x^p$, ou variable à expliquer y) s'écrit comme somme de deux variables : celle qu'on aurait aimé observer plus une erreur d'observation.

On a donc :

$$y = y^* + r^0$$

$$x^j = x^{j*} + r^j \qquad (j = 1, \ldots, p)$$

où :

$y^*, x^{1*}, \ldots, x^{p*}$ sont les variables "vraies", considérées comme des variables aléatoires centrées ;

$y, x^1, \ldots, x^p$ sont les variables aléatoires observées ;

$r^0, r^1, \ldots, r^p$ sont les distorsions aléatoires centrées, de variances respectivement égales à $s_0^2, s_1^2 \ldots, s_p^2$.

On suppose que ces distorsions sont indépendantes entre elles, et indépendantes des variables $y^*, y^{1*}, \ldots, x^{p*}$

On désigne par Q la matrice de variance des erreurs $r^1, \ldots, r^p$

$$Q = \begin{pmatrix} s_1^2 & & 0 \\ & \ddots & \\ 0 & & s_p^2 \end{pmatrix}$$

Les valeurs observées ne portant que sur y et sur les x^i, on cherchera la combinaison linéaire $\sum_{i=1}^{p} b_i x^i$ telle que $e = y - \sum_{i=1}^{p} b_i x^i$ soit de variance minimale ; or :

$$e = y - \Sigma\, b_i x^i = y^* - \Sigma\, b_i x^{i*} + r^0 - \Sigma\, b_i r^i = e^* + r^0 - \Sigma\, b_i r^i$$

où e^* est égal à $y^* - \Sigma\, b_i x^{i*}$

et du fait des hypothèses d'indépendance sur les distorsions, on a :

$$\text{var } e = \text{var } e^* + \text{var } (r^0 - \Sigma\, b_i r^i)$$

$$\text{var } e \geqslant \text{var } (r^0 - \Sigma\, b_i r^i) = s_0^2 + \Sigma\, b_i^2 s_i^2 = s_0^2 + \|\underline{b}\|_Q^2$$

La variance résiduelle est supérieure à la variance d'erreur : $s_0^2 + \|\underline{b}\|_Q^2$

Remarques

1/ Il ne s'agit pas ici de rechercher les coefficients $b_1 \ldots b_p$ de la régression de y^* sur les x^{i*} en utilisant ce qui est observé, à savoir les variables $y, x^1, \ldots x^p$; en effet, dans ce cas, les coefficients $b_1, \ldots, b_p$ apparaîtraient comme caractéristiques d'une liaison qui ne dépendrait pas des erreurs d'observation.

2/ Il ne s'agit pas non plus de rechercher les coefficients $b_1, \ldots b_p$ de la régression de y^* sur les x^i dans le but de se servir de l'équation obtenue pour prédire y^* à partir de ce qui est observable, à savoir les $x^1, \ldots, x^p$.

Les deux problèmes présentés ci-dessus ont déjà été longuement étudiés et leurs solutions ne sont pas simples en général. On consultera à ce sujet les deux articles de KENDALL "Regression, Structure and Functional Relationship" Part. 1 et Part. 2 (27).

On est donc ici plus modeste puisqu'on ne travaille que sur les variables observées.

Supposons que l'on ait un n-échantillon $(\underline{y}, \underline{x}^1, \ldots, \underline{x}^p) = (\underline{y}, X')$ du $(p+1)$ - uple $(y, x^1, \ldots, x^p)$.

Soit $p_i \left(1 \leqslant i \leqslant n\,,\ \Sigma\, p_i = 1\,,\ \text{en général } p_i = \frac{1}{n}\right)$ la masse affectée à l'observation "i". La métrique $N = D_p$ est alors choisie dans l'espace des observations $F = R^n$:

$$N = D_p = \begin{pmatrix} p_1 & & 0 \\ & \ddots & \\ 0 & & p_n \end{pmatrix}$$

L'échantillon est centré :

$$\sum_{i=1}^{n} p_i y_i = 0 = \sum_{i=1}^{n} p_i x_i^j \quad \forall j = 1, 2, \ldots, p$$

La méthode des moindres carrés revient à chercher la combinaison linéaire $\sum_{j=1}^{p} b_j \underline{x}^j$ telle que $\underline{e} = \underline{y} - \Sigma\, b_j \underline{x}^j = \underline{y} - X'\underline{b}$ soit de variance minimale (il s'agit bien ici de la variance empirique).

Pour que cette solution ait un sens, il faut que cette variance résiduelle soit supérieure à la variance d'erreur.

4.4.2 La solution

On recherchera donc la combinaison linéaire $y^+ = \sum_{j=1}^{p} b_j^+ x^j$ telle que :

$$\text{var}\,(y - \Sigma\, b_j^+ x^j) = \|y - \Sigma\, b_j^+ x^j\|_N^2 \tag{5}$$

soit minimum sous la contrainte,

$$\text{var}\,(y - \Sigma\, b_j^+) \geqslant s_0^2 + \Sigma\, (b_j^+)^2\, s_j^2 \ ,$$

c'est-à-dire :

$$\|y - \Sigma\, b_j^+ x^j\|_N^2 \geqslant s_0^2 + \|b^+\|_Q^2 \tag{6}$$

– ou bien la solution $\hat{y} = \Sigma\, b_j x^j$ des moindres carrés vérifie (6) et alors $y^+ = \hat{y}$

– bien cette solution $\hat{y}$ ne vérifie pas (6), auquel cas le minimum de (5), sous la contrainte (6) est tel que :

$$\|y - \Sigma\, b_j^+ x^j\|_N^2 = \|b^+\|_Q^2 + s_0^2 \tag{7}$$

Il revient au même de minimiser $\|b^+\|_Q^2$ sous la contrainte (7).

Géométriquement :

Si on appelle C la quadrique dont l'équation est donnée dans $E^* = (R^p)^*$ par (7), et si on munit E^* de la métrique $M = Q$, on est ramené à chercher

le point de C le plus proche de l'origine, c'est-à-dire à projeter $\underline{0}$ sur C, ce que l'on peut réaliser par un algorithme itératif (cf. CAZES (10), BENZECRI (5)) Cet algorithme a été programmé.

D'un point de vue numérique, pour minimiser

$||\underline{y} - \underline{y}^+||_N^2 = ||\underline{y} - X'\underline{b}^+||_N^2$ sous la contrainte (7), on introduit le multiplicateur de Lagrange λ, et on minimisera :

$$||\underline{y} - \underline{y}^+||_N^2 + \lambda\,(||\underline{b}^+||_Q^2 + s_0^2 - ||\underline{y} - \underline{y}^+||_N^2)$$

soit encore :

$$(1 - \lambda)||\underline{y} - X'\underline{b}^+||_N^2 + \lambda\,(||\underline{b}^+||_Q^2 + s_0^2)$$

d'où l'on déduit par dérivation matricielle :

$$-(1 - \lambda)\,XN\,(\underline{y} - X'\underline{b}^+) + \lambda Q\underline{b}^+ = \underline{0}$$

$$(XNX' + \frac{\lambda}{1 - \lambda}\,Q)\,\underline{b}^+ = XN\underline{y} \qquad (8)$$

λ étant déterminé de telle sorte que l'équation (7) soit satisfaite.

On retrouve la "ridge-regression" avec $k = \dfrac{\lambda}{1 - \lambda}$ et $M = Q$.

On montre qu'il existe une seule valeur k_0 de k, rendant minimum $||\underline{y} - X'\underline{b}^+||_N^2$ sous la contrainte (7), où $\underline{b}^+$ est la solution de l'équation (8) ; de plus, cette valeur k_0 est positive (cf. Annexe).

On pourra donc employer un processus itératif basé sur la formule (8) pour calculer $\underline{b}^+$, en comparant pour chaque valeur de k la variance résiduelle et la variance d'erreur.

4.5 – EXEMPLE D'APPLICATION DE LA REGRESSION SOUS CONTRAINTE

Estimation d'une courbe granulométrique

Le problème est le suivant : on désire estimer à partir d'une méthode photoélectrique la courbe granulométrique d'un nuage de particules en suspension dans un gaz quelconque ; cela revient, si l'on a divisé l'intervalle (d, d') des diamètres que peut avoir une particule du nuage en p classes

$I_1 = (d, d_1), I_2 = (d_1, d_2) \ldots, I_p = (d_{p-1}, d')$, à rechercher les proportions $\{\beta_j | j = 1, p\}$, des particules du nuage pour les différentes classes de diamètre $\{I_j | j = 1, p\}$.

Pour faire cette estimation, on se sert des propriétés de diffusion du gaz considéré (dont les particules sont de diamètre compris entre 0 et 0,4 μ): toute particule de ce gaz placée dans un champ lumineux émet un rayonnement aléatoire ; le nombre de photons émis par cette particule dans un intervalle de temps donné θ choisi comme unité ($\theta = 50\mu s$), suit une loi de Poisson de paramètre α fonction du diamètre de cette particule.

Par étalonnage, on connaît la valeur α_j de α associée à chaque classe de de diamètre $I_j (1 \leqslant j \leqslant p)$.

La probabilité qu'une particule du nuage appartenant à I_j donne lieu à une émission de i photons est donc :

$$x_i^j = e^{-\alpha_j} (\alpha_j)^i / i !$$

et la probabilité qu'une particule du nuage émette i photons est donc :

$$p_i = \Sigma \{x_i^j \beta_j | j = 1, p\}$$

Faisant passer une à une les particules du nuage dans un champ lumineux, pendant un intervalle de temps T grand par rapport à θ, et comptant à l'aide d'un appareillage physique approprié le nombre de particules N_i du nuage ayant émis i photons lors de leur traversée du champ lumineux, on obtient un histogramme expérimental associé à la loi des p_i ; si n désigne le nombre de classes de cet histogramme, et N le nombre total de particules qui sont passées dans le champ :

$$N = \Sigma \{N_i | i = 1, n\}$$

$y_i = \dfrac{N_i}{N}$ est une estimation de p_i (l'estimation donnée par la méthode du maximum de vraisemblance)

$$y_i \cong p_i = \Sigma \{x_i^j \beta_j | j = 1, p\} \quad 1 \leqslant i \leqslant n$$

$$y_i = \Sigma \{x_i^j \beta_j | j = 1, p\} + e_i \quad 1 \leqslant i \leqslant n$$

avec

$$\begin{cases} \beta_j \geqslant 0 \\ \Sigma \{\beta_j | j = 1, p\} = 1 \end{cases}$$

pour que le modèle précédent ait un sens.

Nous donnons ci-dessous les résultats relatifs à deux histogrammes l'un à 16 classes, l'autre à 12 classes (n vaut soit 16, soit 12), les classes de diamètres étant au nombre de 8 ($p = 8$).

Dans chacun des deux cas, avec la métrique usuelle, nous avons effectué :

– La régression classique

– La régression classique en imposant à la somme des coefficients de régression de valoir 1.

– La ridge régression, en imposant aux coefficients de régression d'être de somme 1.

– La régression sous contraintes linéaires : coefficients de régression positifs et de somme 1.

– La régression sous contraintes linéaires : coefficients de régression positifs.

La somme des estimations $\hat{b}_j$ des β_j obtenue par la régression classique valant 1,0001 (histogramme à 16 classes) et 0,998 (histogrammes à 12 classes), les résultats de cette régression sont identiques à ceux de la régression où l'on impose à cette somme de valoir 1 ; nous ne donnerons donc que les résultats relatifs à cette dernière régression.

Les résultats sont présentés dans les tableaux 1, 2, 3, 4 ; on a également fait figurer dans le tableau 2 les résultats fournis par la ridge régression classique (où l'on n'impose pas à la somme des b_j^+ de valoir 1) dans le cas de l'histogramme à 16 classes.

On voit que la régression classique donne des résultats aberrants, certains coefficients de régression étant négatifs, ou de module supérieur à 1 (cf. tableaux 3 et 4) ; ceci est essentiellement dû au fait que les variables explicatives sont très corrélées. C'est la raison pour laquelle, en plus de la régression où l'on impose aux valeurs b_j^+ d'être positives, on a appliqué la ridge régression de façon à limiter la valeur de la norme de $\underline{b}^+$.

Pour la ridge régression, nous avons choisi comme métrique M dans $(R^p)^*$ la métrique diagonale ayant mêmes éléments diagonaux que la matrice $XNX' = XX'$ et nous avons fait varier k (cf. 4.3) de 0 à 0,95 avec un pas de 0,05.

A partir d'une valeur k_0 de k (du moins quand on impose à la somme des coefficients de régression de valoir 1 : $k_0 = 0{,}3$ pour la première régression, $k_0 = 0{,}35$ pour la seconde), toutes les estimations des coefficients de régression deviennent positives. Pour cette valeur k_0 de k, les résultats de la ridge régression ($\Sigma\, b_j^+ = 1$) sont voisins de ceux fournis par la régression sous contraintes linéaires (coefficients de régression positifs ; coefficients de régression positifs et de somme 1) (cf. tableaux 1 et 3).

Les tableaux 2 et 4 résument les résultats fournis par la ridge régression pour différentes valeurs de k ; on a également représenté sur le tableau 2, comme on l'a déjà signalé, les résultats de la ridge régression où l'on n'impose pas à la somme des coefficients de régression de valoir 1. On voit que dans ce cas, dès que $k \geqslant 0{,}05$, les valeurs négatives des b_j^+ deviennent faibles en module ; et pour $k = 0{,}15$ (où $\|\underline{b}^+\|_M^2 = 0{,}05$), les résultats obtenus sont voisins de ceux fournis par la ridge régression avec $\Sigma\, b_j^+ = 1$, pour la même valeur 0,05 de $\|\underline{b}^+\|_M^2$ (qui correspond à $k = k_0 = 0{,}30$).

En ce qui concerne la seconde régression (cf. tableau 4), dès que k est supérieur ou égal à 0,05, toutes les estimations b_j^+ des β_j deviennent inférieures en module à 1, le carré de la norme de $\underline{b}^+$ passant de 58,95 ($k = 0$), à une valeur plus petite ou égale à 0,06 ($k > 0{,}05$).

Quand k augmente, dans les deux cas étudiés ($n = 16$ et $n = 12$), les valeurs b_j^+ peu stables dans la ridge régression correspondent aux b_j^+ annulées par les contraintes de positivité.

En fait, la somme pour $j \geqslant j_0$ des b_j^+ obtenus par la ridge régression, $\Sigma\{b_j^+ \mid j = j_0 , p\}$, ($j_0 = 5$ dans la première régression, $j_0 = 6$ dans la seconde), varie peu quand k varie, cette somme étant voisine de la même somme calculée sur les b_j^+ fournis par la régression sous contrainte de positivité ($b_j^+ > 0$ ou $b_j^+ > 0$ et de somme 1) ; dans le dernier cas les b_j^+ étant nuls pour $j > j_0$, cette somme se réduit à $b_{j_0}^+$.

D'un point de vue physique, cela revient à dire que ce que l'on estime avec précision est le pourcentage de particules de diamètre supérieur ou égal à d_{j_0-1} .

L'approximation réalisée par ces différents types de régression est excellente, puisque le résidu S^2 (égal à $\|\underline{y} - \underline{y}^+\|^2 / \|\underline{y}\|^2$, si $\underline{y}^+$ désigne l'approximation de $\underline{y}$) est toujours inférieur à 0,1 sauf dans le cas de la ridge régression où l'on n'impose pas $\Sigma\, b_j^+ = 1$ (cas de la première régression, où $n = 16$), S^2 devenant supérieur à 0,1 pour $k = 0{,}75$, et atteignant la valeur 0,136 pour $k = 0{,}95$.

En conclusion, on voit que sur les deux cas étudiés (et en particulier dans le second), la ridge régression, et la régression sous contraintes de positivité des coefficients protégent bien la régression, et donnent des résultats équivalents, la régression sous contraintes de positivité semblant s'imposer dans ces deux exemples, du fait de la nature du modèle.

Tableau n° 1

j	$\hat{b}_j$	$\hat{\sigma}_j$	b_j^+	b_j^+	b_j^+
1	0,087	0,055	0,076	0,080	0,064
2	0,104	0,083	0,104	0,138	0,141
3	0,131	0,040	0,125	0,105	0,110
4	0,190	0,021	0,212	0,200	0,212
5	0,572	0,033	0,404	0,477	0,490
6	−0,201	0,088	0,058	0	0
7	0,161	0,128	0,001	0	0
8	−0,044	0,07	0,020	0	0
			0,483		
Σb_j	1		1	1	1,017
S^2	0,0041		0,026	0,0098	0,0095
Modèle ou Technique	Moindres carrés classiques $\Sigma \hat{b}_j = 1$		ridge régression $\Sigma b_j^+ = 1$ $\|\|\underline{b}^+\|\|^2 \leqslant 0{,}05$ $(k = 0{,}3)$	contraintes $b_j^+ > 0$ $\Sigma b_j^+ = 1$	contraintes $b_j^+ > 0$

Comparaison des résultats obtenus entre la régression classique, la ridge régression, et la régression sous contraintes de positivité ($b_j^+ > 0$) dans le cas de l'histogramme à 16 classes.

$$S^2 = \frac{||\underline{y} - \underline{y}^+||^2}{||\underline{y}||^2}$$

avec $\underline{y}^+$ approximation de $\underline{y}$ pour le modèle considéré.

$\hat{\sigma}_j$ estimation de l'écart type de $\hat{b}_j$.

Tableau n° 2

j	$(b_j^+)_0 = \hat{b}_j$	$(b_j^+)_{0,05}$	$(b_j^+)_{0,30}$	$(b_j^+)_{0,95}$	$(b_j^+)_{0,05}$	$(b_j^+)_{0,15}$	$(b_j^+)_{0,95}$
1	0,087	0,080	0,076	0,066	0,080	0,077	0,059
2	0,104	0,112	0,104	0,092	0,111	0,106	0,080
3	0,131	0,125	0,125	0,123	0,122	0,118	0,096
4	0,190	0,204	0,212	0,203	0,198	0,195	0,143
5	0,572	0,503	0,404	0,319	0,496	0,432	0,246
6	−0,201	−0,046	0,058	0,120	−0,049	0,004	0,059
7	0,161	−0,009	0,001	0,046	−0,012	−0,023	0,002
8	−0,044	0,031	0,020	0,031	0,027	0,017	0,004
$\Sigma (b_j^+)_k$	1	1	1	1	0,973	0,926	0,677
k	0	0,05	0,30	0,95	0,05	0,15	0,95
$\|\|\underline{b}_k^+\|\|^2$	0,08	0,06	0,05	0,04	0,06	0,05	0,03
S^2	0,0041	0,0062	0,026	0,082	0,0067	0,016	0,136
	Ridge régression avec $\Sigma (b_j^+)_k = 1$				Ridge régression classique		

0,483 (rows 4–8, $(b_j^+)_{0,30}$) ; 0,516 (rows 4–8, $(b_j^+)_{0,95}$)

Etude en fonction de k des résultats donnés par la ridge régression dans le cas de l'histogramme à 16 classes.

Tableau n° 3

j	$\hat{b}_j$	$\hat{\sigma}_j$	b_j^+	b_j^+	b_j^+
1	5,095	5,28	0,052	0,012	0,077
2	−6,041	6,38	0,059	0,079	0,0003
3	1,246	1,18	0,102	0,129	0,161
4	0,050	0,10	0,141	0,110	0,106
5	0,240	0,036	0,280	0,291	0,288
6	0,671	0,084	0,308	0,379	0,374
7	0,560	0,178	0,056	0	0
8	−0,299	0,130	0,002	0	0
Σb_j	1		1	1	1,0063
S^2	0,0034		0,032	0,0194	0,0191
Modèle ou technique	Moindres carrés classiques $\Sigma b_j = 1$		ridge régression $\Sigma b_j^+ = 1$ $\|\|\underline{b}^+\|\|^2 \leqslant 0,046$ $(k = 0,35)$	contraintes $b_j^+ > 0$ $\Sigma b_j^+ = 1$	contraintes $b_j^+ > 0$

0,366 (rows 6–8, b_j^+)

Comparaison des résultats obtenus entre la régression classique, la ridge régression, et la régression sous contraintes linéaires ($b_j^+ > 0$) (histogramme à 12 classes)

Tableau n° 4

j	$(b_j^+)_0 = \hat{b}_j$	$(b_j^+)_{0,05}$	$(b_j^+)_{0,35}$	$(b_j^+)_{0,95}$
1	5,095	0,048	0,052	0,051
2	−6,041	0,056	0,059	0,057
3	1,246	0,119	0,102	0,095
4	0,050	0,132	0,141	0,144
5	0,240	0,271	0,280	0,267
6	0,671	0,427	0,308	0,256
7	−0,560	−0,033	0,056	0,088
8	0,299	−0,02	0,002	0,042
			0,366	0,386
$\Sigma\ (b_j^+)_k$	1	1	1	1
k	0	0,05	0,35	0,95
$\|\|\underline{b}_k^+\|\|^2$	58,95	0,066	0,046	0,041
S^2	0,0034	0,013	0,032	0,061

Etude en fonction de k des résultats donnés par la ridge régression dans le cas de l'histogramme à 12 classes.

ANNEXE

Etude en fonction de k de :

$$f(k) = \|\underline{b}_k\|_M^2 \quad (f(0) > a^2)$$

$$g(k) = \|\underline{b}_k - \hat{\underline{b}}\|_{M_1 = XNX'}^2$$

$$l(k) = s_0^2 + \|\underline{b}_k\|_M^2 - \|\underline{y} - X'\underline{b}_k\|_N^2 \quad (\text{avec } l(0) > 0)$$

sachant que :

$$(XNX' + kM)\underline{b}_k = XN\underline{y} \ ; \ (\underline{b}_0 = \hat{\underline{b}})$$

M étant une forme quadratique définie positive peut se décomposer sous la forme :

$$M = T\,T' \tag{1}$$

On a alors :

$$XNX' + kM = XNX' + kTT'$$
$$= T(T^{-1}XNX'(T')^{-1} + k)T' \qquad (2)$$
$$= T(M_2 + kI)T'$$

en posant

$$M_2 = T^{-1}XNX'(T')^{-1} \qquad (3)$$

qui est une forme quadratique définie positive (si XNX' est régulière, ce que nous supposerons).

On a donc d'après les relations (1) à (3), et du fait que $XNX'\underline{\hat{b}} = XN\underline{y}$:

$$\underline{b}_k = T'^{-1}(M_2 + kI)^{-1}T^{-1}XNX'\underline{\hat{b}}$$
$$= T'^{-1}(M_2 + kI)^{-1}M_2T'\underline{\hat{b}} \qquad (4)$$
$$= T'^{-1}(I + kM_2^{-1})^{-1}T'\underline{\hat{b}}$$

Expression de $f(k)$

De (4) l'on déduit

$$f(k) = ||\underline{b}_k||^2_{M=TT'} = ||(I + kM_2^{-1})^{-1}T'\underline{\hat{b}}||^2_I \qquad (5)$$

si $\{\underline{w}_i | i = 1, p\}$ désigne un système de vecteurs propres orthonormés (au sens usuel, i.e. pour la métrique unité) de M_2, $\underline{w}_i$ étant relatif à la valeur propre λ_i (nous supposerons les valeurs propres rangées par ordre décroissant i.e. $\lambda_1 \geqslant \lambda_2 \geqslant \lambda_3 \ldots \lambda_p > 0$).

$$M_2\underline{w}_i = \lambda_i\underline{w}_i$$
$$< \underline{w}_i, \underline{w}_j > = \delta_i^j = \begin{cases} 1 & \text{si } i = j \\ 0 & \text{sinon} \end{cases}$$

si l'on pose :

$$\alpha_i = <T'\underline{\hat{b}}, \underline{w}_i>_I = <\underline{\hat{b}}, T\underline{w}_i>_I \qquad (5 \text{ bis})^{(*)}$$

on a

$$f(k) = \Sigma\left\{\frac{\alpha_i^2}{\left(1 + \frac{k}{\lambda_i}\right)^2} \mid i = 1, p\right\} = \Sigma\left\{\left(\frac{\lambda_i}{\lambda_i + k}\right)^2 \alpha_i^2 \mid i = 1, p\right\} \qquad (6)$$

(*) $\underline{w}_i$ étant vecteur propre de $M_2 = T^{-1}XNX'(T')^{-1}$, $T\underline{w}_i$ est vecteur propre (relatif à la même valeur propre) de $XNX'(T')^{-1}T^{-1} = XNX'M^{-1}$.

Expression de $g(k) = \|\underline{b}_k - \hat{\underline{b}}\|^2_{M_1 = XNX'}$

d'après (4), l'on a :

$$\underline{b}_k - \hat{\underline{b}} = T'^{-1} \left[(I + kM_2^{-1})^{-1} - I\right] T' \hat{\underline{b}} \tag{7}$$

on en déduit, puisque, d'après (3)

$$M_1 = XNX' = TM_2T' \tag{8}$$

que

$$\begin{aligned} g(k) &= \|\underline{b}_k - \hat{\underline{b}}\|^2_{M_1 = XNX'} \\ &= \|T'^{-1} \left[(I + kM_2^{-1})^{-1} - I\right] T' \hat{\underline{b}}\|^2_{TM_2T'} \\ &= \|\left[(I + kM_2^{-1})^{-1} - I\right] T' \hat{\underline{b}}\|^2_{M_2} \\ &= \|T' \hat{\underline{b}}\|^2_A \end{aligned} \tag{9}$$

avec

$$A = ((I + kM_2^{-1})^{-1} - I)\, M_2 ((I + kM_2^{-1})^{-1} - I) \tag{10}$$

A a mêmes vecteurs propres que M_2, la valeur propre $\lambda_i(A)$ associée a $\underline{w}_i$ s'écrivant :

$$\lambda_i(A) = \left(\frac{1}{1 + \dfrac{k}{\lambda_i}} - 1\right)^2 \lambda_i = k^2 \frac{\lambda_i}{(\lambda_i + k)^2} \tag{11}$$

on a donc d'après (5 bis), (9) et (11) :

$$g(k) = k^2 \sum \left\{ \frac{\lambda_i \alpha_i^2}{(\lambda_i + k)^2} \mid i = 1, p \right\} \tag{12}$$

Expression de $l(k)$

Nous poserons :

$$\gamma = s_0^2 - \|\underline{y} - X' \hat{\underline{b}}\|^2_N \tag{13}$$

Ayant

$$\|\underline{y} - X' \underline{b}_k\|^2_N = \|\underline{y} - X' \hat{\underline{b}}\|^2_N + \|\underline{b}_k - \hat{\underline{b}}\|^2_N \tag{14}$$

on a :

$$l(k) = \gamma + f(k) - g(k)$$

soit d'après (6) et (12) :

$$l(k) = \gamma + \Sigma \left(\frac{\lambda_i}{\lambda_i + k}\right)^2 \alpha_i^2 - k^2 \Sigma \frac{\lambda_i \alpha_i^2}{(\lambda_i + k)^2} \qquad (15)$$

$$= \gamma + \Sigma \left\{ \frac{\lambda_i \alpha_i^2 (\lambda_i - k^2)}{(\lambda_i + k)^2} \mid i = 1, p \right\}$$

Etude de $f(k)$ (cas de la ridge régression)

Dès que k est plus grand que $-\lambda_p$, λ_p étant rappelons le la plus petite valeur propre de M_2, $f(k)$ est d'après (6) une fonction décroissante de k. Si donc $f(0) = \|\hat{\underline{b}}\|_M^2 > a^2$, ils existe une seule valeur positive k_0 de k telle que $f(k_0) = a^2$.

Par contre il peut exister une ou plusieurs valeurs négatives k_0' de k telles que $f(k_0') = a^2$ (k_0' étant bien sûr plus petit que $-\lambda_p$) .

Nous allons voir que le minimum de $\|\underline{b}^+ - \hat{\underline{b}}\|^2_{M_1 = XNX'}$ sous la contrainte $\|\underline{b}^+\|_M^2 = a^2$ (on suppose bien sûr que $\|\hat{\underline{b}}\|_M^2 > a^2$) , i.e. le minimum de

$$R = \|\underline{b}^+ - \hat{\underline{b}}\|^2_{M_1 = XNX'} + k(\|\underline{b}^+\|_M^2 - a^2)$$

sous la contrainte précédente est réalisé pour $\underline{b}^+ = \underline{b}_{k_0}$.

En effet, on a

$$\left.\begin{aligned} \partial R/\partial \underline{b}^+ &= 2(XNX' + kM)\,\underline{b}^+ - 2\,XNX'\,\hat{\underline{b}} \\ \partial^2 R/_{(\partial \underline{b}^+)^2} &= 2(XNX' + kM) = 2T(M_2 + kI)\,T' \end{aligned}\right\} \qquad (16)$$

avec $$\|\underline{b}^+\|_M^2 = a^2$$

d'où l'on déduit les résultats déjà vus dans l'étude de la ridge régression :

$$\underline{b}^+ = \underline{b}_k \qquad \text{avec} \qquad f(k) = a^2$$

Comme $\partial^2 R/(\partial \underline{b}^+)^2$ n'est définie positive que pour $k > -\lambda_p$, la seule valeur de k qui convient est k_0.

C.Q.F.D.

Etude de $l(k)$ en fonction de k : cas de la régression sur variables entachées d'erreurs, où $M = Q$ (c.f. figure 1)

De (15) on déduit :

$$\frac{dl}{dk} = -2(k+1)\,\Sigma\left\{\frac{\lambda_i \alpha_i^2}{(\lambda_i + k)^3} \mid i = 1, p\right\}$$

Dés que k est supérieur à $\sup(-1, -\lambda_p)$, $\dfrac{dl}{dk}$ est négatif, donc l est une fonction décroissante de k ; comme $l(0) > 0$, il existe donc une valeur positive unique k_0 de k telle que $l(k_0) = 0$.

Il existe au moins une autre valeur k_0' de k négative telle que $l(k_0') = 0$ $(k_0' < \sup(-1, -\lambda_p))$.

Nous allons montrer que seul $\underline{b}^+ = \underline{b}_{k_0}$ assure le minimum de

$$\|\underline{b}^+ - \hat{\underline{b}}\|^2_{M_1 = XNX'} \qquad \text{sous la contrainte}$$

$$s_0^2 + \|\underline{b}^+\|^2_M - \|\underline{y} - X'\hat{\underline{b}}\|^2_N - \|\hat{\underline{b}} - \underline{b}^+\|^2_{M_1 = XNX'} = 0 \qquad (17)$$

En effet minimiser $\|\underline{b}^+ - \hat{\underline{b}}\|^2_{M_1 = XNX'}$ sous la contrainte (17) revient à minimiser :

$$U = \|\underline{b}^+ - \hat{\underline{b}}\|^2_{M_1 = XNX'} + \lambda(\gamma + \|\underline{b}^+\|^2_M - \|\hat{\underline{b}} - \underline{b}^+\|^2_{M_1 = XNX'})$$

sous la contrainte (17)

U s'écrit encore :

$$U = (1 - \lambda)\,\|\underline{b}^+ - \hat{\underline{b}}\|^2_{M_1 = XNX'} + \lambda(\gamma + \|\underline{b}^+\|^2_M)$$

d'où l'on déduit :

$$\partial U/_{\partial \underline{b}^+} = 2[(1 - \lambda)\,XNX'\,(\underline{b}^+ - \hat{\underline{b}}) + \lambda M\,\underline{b}^+] = \underline{0}$$

$$\partial^2 U/_{(\partial \underline{b}^+)^2} = 2\,[(1 - \lambda)\,XNX' + \lambda M]$$

Posant $k = \dfrac{\lambda}{1 - \lambda}$, on retrouve les résultats déjà vus dans la régression sur variables entachées d'erreurs.

$$\underline{b}^+ = \underline{b}_k \qquad \text{avec} \qquad l(k) = 0$$

De plus, on a

$$\partial^2 U/(\partial \underline{b}^+)^2 = \frac{2\,T(M_2 + kI)\,T'}{(k+1)}$$

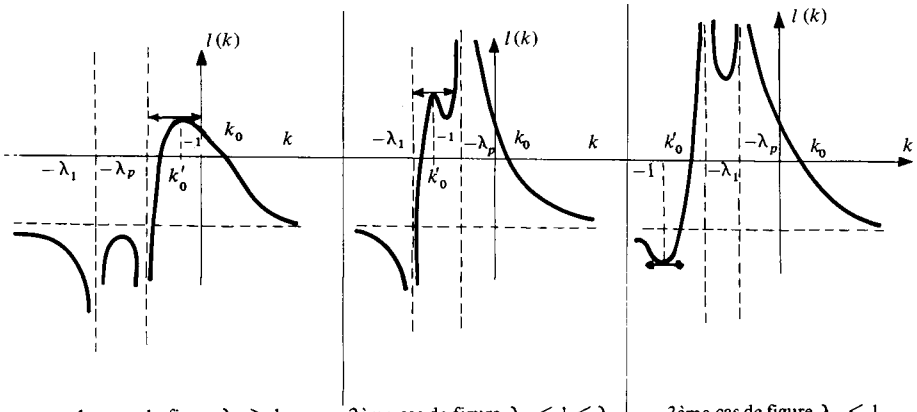
qui n'est définie positive que si k est plus grand que $\sup(-1, -\lambda_p)$ (ce qui est le cas pour k_0, mais pas pour k'_0), ou si k est plus petit que $\inf(-1, -\lambda_1)$.

Comme $\forall k < -\lambda_1$, on a (cf. figure 2) :

$$g(k) = \|\underline{b}_k - \hat{\underline{b}}\|^2_{M_1} = k^2 \sum \frac{\lambda_i \alpha_i^2}{(\lambda_i + k)^2} > \sum \lambda_i \alpha_i^2 > g(k_0) = \|\underline{b}_{k_0} - \hat{\underline{b}}\|^2_{M_1}$$

(puisque $k_0 > 0$), on en déduit que le minimum de $\|\underline{b}^+ - \hat{\underline{b}}\|^2_{M_1}$ sous la contrainte (17) est obtenu pour $\underline{b}^+ = \underline{b}_{k_0}$

C.Q.F.D.

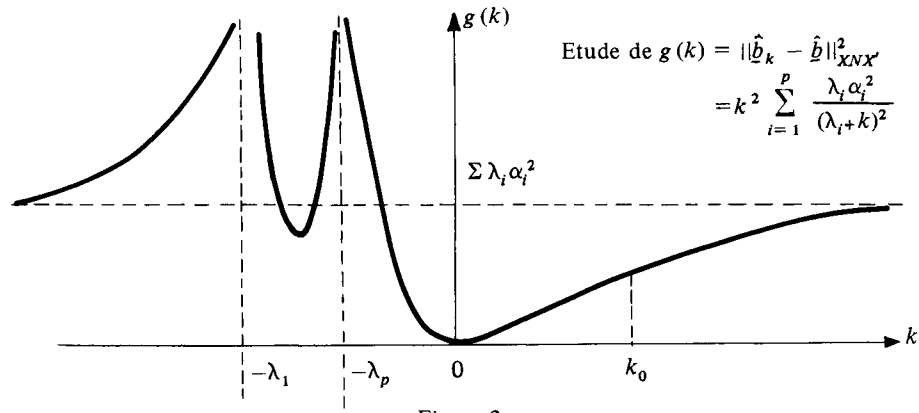


1er cas de figure $\lambda_p > 1$ | 2ème cas de figure $\lambda_p < 1 < \lambda_1$ | 3ème cas de figure $\lambda_1 < 1$

Etude de $l(k) = \gamma + \sum_{i=1}^{p} \frac{\lambda_i \alpha_i^2 (\lambda_i - k^2)}{(\lambda_i + k)^2}$

(avec $l(0) > 0$)

Figure 1



Etude de $g(k) = \|\hat{\underline{b}}_k - \hat{\underline{b}}\|^2_{XNX'}$

$= k^2 \sum_{i=1}^{p} \frac{\lambda_i \alpha_i^2}{(\lambda_i + k)^2}$

Figure 2

V – RÉGRESSION PAR BOULES ET RÉGRESSION PAR L'ANALYSE DES CORRESPONDANCES

Dans ce paragraphe, on veut expliquer une variable y en fonction des variables $\{x^j \mid j = 1, p\}$ à partir d'un n-échantillon de ces $(p + 1)$ variables.

5.1. – INTRODUCTION

La méthode de régression par boules, BENZECRI (6(2)), revient à traiter dans le cas empirique le problème de la régression fonctionnelle, c'est-à-dire la recherche de la meilleure approximation de y par les $\{x^j | j = 1, p\}$ (approximation qui n'est pas en général du type linéaire).

Dans la régression par Analyse des Correspondances (6), pour expliquer y en fonction des x^j, on met en correspondance y avec l'ensemble des x^j, après avoir rendu toutes les variables qualitatives par découpage en classes ; puis on étudie les liaisons entre y et les x^j par l'analyse de la correspondance obtenue.

Ces deux méthodes font appel aux techniques d'Analyse des Données, et nécessitent donc un échantillon dont l'effectif doit être assez élevé relativement au nombre de variables.

5.2. – REGRESSION PAR BOULES (OU PAR VOISINAGES)

Aux n réalisations $(y_j, x_j^1, \ldots, x_j^p)$ de $(y, x^1, \ldots, x^p)$ on peut faire correspondre un nuage $\mathfrak{N}$ de n points $M_1, \ldots, M_j, \ldots, M_n$ de l'espace $E = R^p$, le point M_j étant le point de coordonnées $x_j^1, \ldots, x_j^p$. A chaque point M_j est donc affectée une valeur y_j de y (il y a une probabilité nulle si la loi de $(x^1, \ldots, x^p)$ est absolument continue d'avoir deux points M_j et $M_{j'}$ confondus).

Dans l'espace E, que nous supposerons muni de la métrique Q (ou dans l'espace condensé des premiers axes factoriels du nuage $\mathfrak{M}$ des points M_j affectés des masses p_j), pour prédire y_j, on cherche les r (r fixé) points $M_{j'}$ de $\mathfrak{M}$ les plus proches de M_j. La moyenne des valeurs $y_{j'}$ associées et l'écart-type (empirique) de ces valeurs fournissent respectivement une approximation y_j^* de y_j et la précision de cette approximation.

Le coefficient de corrélation (empirique) entre les y_j^* et les y_j donne une idée de la qualité globale de l'approximation réalisée en remplaçant les y_j par les y_j^*, et on pourra l'appeler rapport de corrélation multiple empirique de y par rapport à $\underline{x}$ (i.e. par rapport à $(x^1, x^2, \ldots, x^p)$), puisque la méthode précédente revient en quelque sorte à prendre la moyenne de y quand $\underline{x}$ est fixé et égal à $\underline{x}_0$ (ou se trouve dans un voisinage de $\underline{x}_0$).

Si l'on a une observation supplémentaire s pour laquelle les valeurs $x_s^1, \ldots, x_s^p$ de $(x^1, \ldots, x^p)$ sont connues, mais dont on ne connaît pas la valeur y_s de y, on pourra faire correspondre à s un point M_s de E. La recherche des r points M_j de $\mathfrak{M}$ les plus proches de M_s (on dira le "voisinage" de M_s) permet, en faisant comme précédemment moyenne et écart-type des r valeurs y_j associées, de prédire y_s ainsi que la précision de cette estimation.

Remarque : Au lieu de prendre pour "voisinage" d'un point M de E les r points de $\mathfrak{M}$ les plus proches de M, pour estimer la valeur de y en M, on peut aussi prendre les points de $\mathfrak{M}$ qui sont à l'intérieur de la boule de centre M et de rayon R donné. Pour un programme de régression par boules, on consultera LEBEAUX (49).

5.3. – REGRESSION PAR ANALYSE DES CORRESPONDANCES

Supposons que l'on veuille expliquer toujours à partir d'un n échantillon une variable y qualitative ou quantitative en fonction de p variables qualitatives ou quantitatives.

On rend toutes les variables qualitatives en divisant l'intervalle de variation des variables quantitatives en classes, en général 4 ou 5, sauf pour la variable à expliquer y, que l'on divise (si elle est quantitative) en $r = 8$ ou 10 classes, ou même plus si l'on a un nombre suffisant d'observations.

Pour une variable quantitative donnée, dont l'histogramme est assez régulier, et ne comporte pas de trous, on choisit en général les limites des classes de façon à avoir un découpage en classes d'égal effectif.

Soient :

$-I = \{1, 2, \ldots, r\}$ l'ensemble des classes de y, rangées de façon croissante (si cette variable est quantitative ou semi-quantitative).

$-J_l = \{x_1^l, \ldots, x_{n_l}^l\}$ l'ensemble des classes de x^l, rangées de façon croissante.

$-J = U\{J_l \mid l = 1, p\}$

On construit sur $I \times J$ le tableau de correspondance suivant :

$$\forall i \in I, \forall j \in J_l \subset J :$$

$k(i, j)$ = nombre d'observations appartenant à la classe i de y et à la classe j de x^l, et l'on effectue l'analyse des correspondances du tableau k, puis l'on fait la représentation simultanée de I et J sur les premiers axes factoriels.

De cette représentation, l'on déduit les liaisons entre y et les x^l. Nous supposerons maintenant que y est quantitatif.

S'il y a une liaison entre y et les x^l, la représentation des modalités $1, 2, \ldots, r$ de y dans le plan des deux premiers axes factoriels se fait en général sur une parabole d'axe, le second axe factoriel, les classes $1, 2, \ldots, i, \ldots, r$ de y étant rangées dans l'ordre en projection sur le premier axe factoriel. On a donc, en orientant convenablement cet axe :

$$c_1 \leqslant c_2 \leqslant \ldots \leqslant c_r$$

où c_i désigne la projection de la classe i de y sur le premier axe factoriel.

Cet axe est donc l'axe caractérisant la croissance de y.

La projection des modalités des variables explicatives sur cet axe permettra donc de caractériser l'influence de ces variables sur y.

En particulier, si une variable quantitative x^l est telle que :

$$d_1^l \leqslant d_2^l \leqslant \ldots \leqslant d_{n_l}^l \tag{1}$$

où d_j^l désigne la projection de la $j^{\text{ème}}$ modalité de x^l sur l'axe factoriel 1, x^l est liée positivement à y, tandis que si :

$$d_1^l \geqslant d_2^l \geqslant \ldots \geqslant d_{n_l}^l \tag{2}$$

x^l est liée négativement à y.

Dans ces cas, la représentation des modalités de x^l se fait en général comme pour y suivant une parabole d'axe, le deuxième axe factoriel.

L'examen approfondi du plan 1-2 permet :

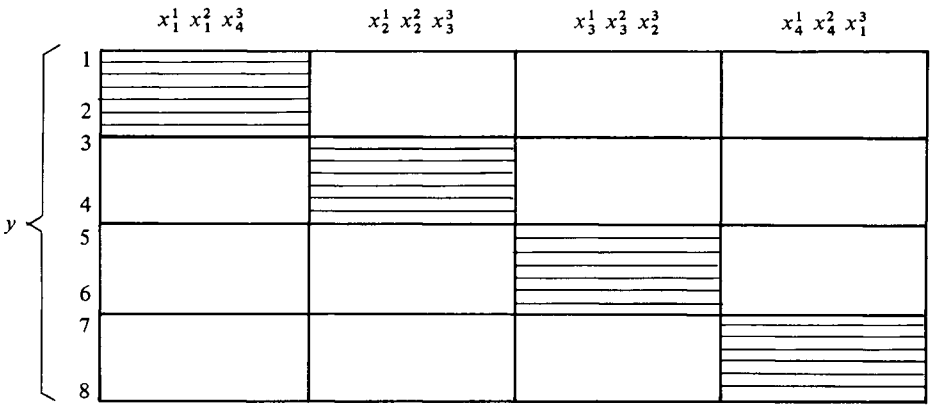
- de nuancer les résultats précédents,
- de voir l'influence des variables qualitatives,
- de voir l'influence des variables quantitatives,

dont les modalités ne se projettent pas dans l'ordre suivant le premier axe factoriel (i.e. pour lesquelles ni (1), ni (2) n'est vérifiée).

Le fait d'obtenir une parabole pour la disposition des modalités de y correspond à ce que l'on appelle classiquement l'effet Guttman, et revient à dire que l'on peut disposer le tableau k (par permutation des lignes et des colonnes de ce tableau) sous une forme diagonale.

Supposons par exemple que l'on ait trois variables explicatives quantitatives x^1, x^2, x^3, divisées respectivement en quatre classes, et que y (divisé en 8 classes) soit lié positivement à x^1 et x^2, et négativement à x^3.

Le tableau k pourra se mettre sous la forme :



où x_j^l désigne toujours la $j^{\text{ème}}$ classe ($1 \leqslant j \leqslant 4$) de la variable x^l ($1 \leqslant l \leqslant 3$).

Les hachures indiquent les régions du tableau ayant une masse importante.

On obtient alors en projection dans le plan 1-2 un graphique tel que celui qui est représenté sur la figure page suivante :

Les classes $\{x_j^l \mid j = 1, 4\}$ de $x^l (1 \leqslant l \leqslant 3)$ sont supposées rangées de façon croissante, de même que les classes 1 à 8 de y.

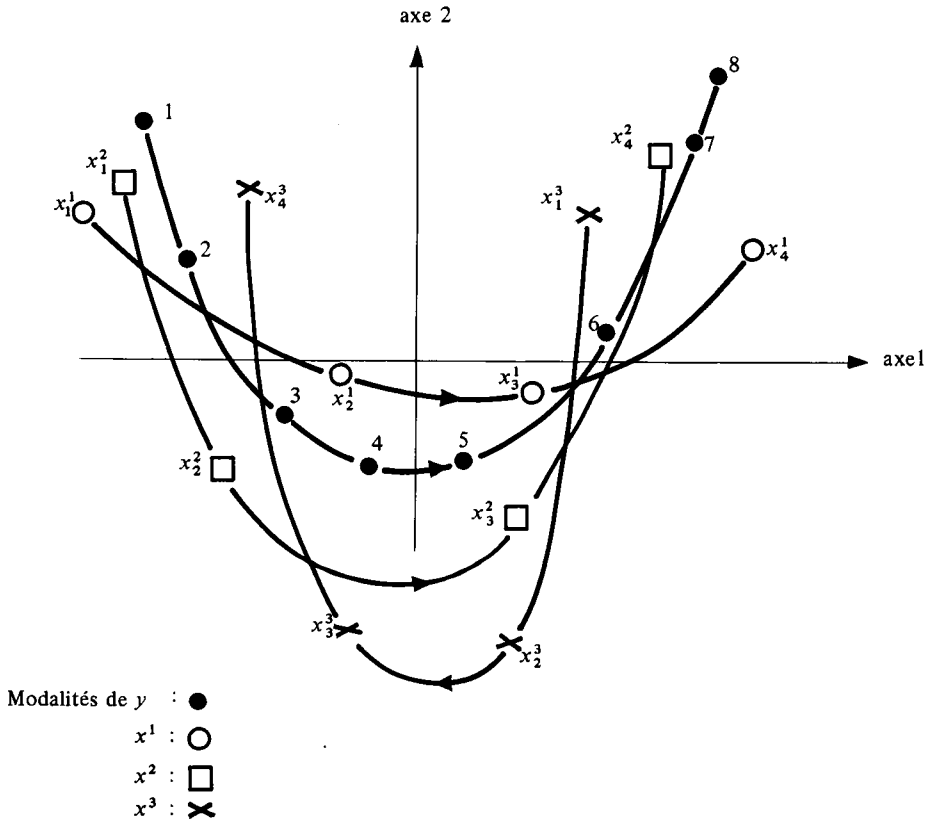
Notons que sur ce graphique, on peut placer toute observation 0 ($1 \leqslant 0 \leqslant n$) : il suffit d'adjoindre au tableau k la ligne supplémentaire que nous appellerons également 0, et telle que :

$$\forall j \in J_l \subset J : k(0, j) = 1$$

si l'observation 0 tombe dans la classe j de x^l

$$k(0, j) = 0 \text{ sinon}$$

puis de projeter 0 sur le plan 1-2.



On pourra de la même façon placer toute observation supplémentaire s (pour laquelle y est inconnue, tandis que les x^l sont connues) sur le graphique précédent, et l'on pourra donc faire une régression par boules pour prédire la valeur de y pour s.

Cette méthode de régression par l'analyse des correspondances, qui permet de visualiser les observations, est en quelque sorte une régression fonctionnelle empirique permettant de voir le lien entre y et les x^l, et de prédire y à partir des x^l.

BIBLIOGRAPHIE

A.M.S. Annals of Mathematical Statistics

J.R.S.S. Journal of the Royal Statistical Society

J.A.S.A. Journal of the American Statistical Association

(1) R. ANDERSON – Distribution of the serial correlation coefficient. *A.M.S.*, 1942, *13*, 1-13.

(2) F. ANSCOMBE – Examination of residuals. *Proceedings of the 4th Berkely Symposium on Mathematics, Statistics and Probability*, 1961 , *1*, 1-36.

(3) F. ANSCOMBE et J. TUKEY – The examination and analysis of residuals. *Technometrics*, 1963, *5* (2), 141-160.

(4) R. BARGMANN – Theory of least squares. Blaksburg, Virginia Polytechnic Institute, 1961.

(5) J.P. BENZECRI

La régression.

Laboratoire de Statistiques Mathématiques (Université de PARIS VI), Paris 1970.

(6) J.P. BENZECRI

(1) L'analyse des données, T1 et T2, Dunod, Paris, 1974.

(2) Méthodes statistiques de la taxinomie 1ère partie. Laboratoire de Statistiques Mathématiques (Université Paris VI), Paris 1974.

(7) I. BERENBLUT et G. WEBB – A new test for autocorrelated errors in the linear regression model. *J.R.S.S.*, B (1973) *35*, p. 33-50.

(8) P. CAZES – Regression – Loi normale et modèle linéaire (1972) Polycopié. ISUP – PARIS.

(9) P. CAZES et P.Y. TURPIN – 1971, *R. Stat. Appl.*, *19* (4), p. 23-44.

(10) P. CAZES – Protection de la régression par utilisation de contraintes linéaires et non linéaires (1975) (à paraître dans *R. Stat. Appl.*).

(11) C.3 E
Cours d'analyse de données multidimensionnelles, Paris 1971.

(12) D. COCHRANE et G.H. ORCUTT – Application of least squares regression to relationships containing autocorrelated error terms. *J.A.S.A.*, 1949, *44*, 32-59.

(13) N. DRAPER et H. SMITH – Applied Regression Analysis, Wiley, New-York 1966.

(14) J. DURBIN et G. WATSON – Testing for serial correlation in least squares regression :

(1) *Biometrika*, 1950, *37*, 409-428 ;
(2) *Biometrika*, 1951, *38*, 159-178 ;
(3) *Biometrika*, 1971, *58*, 1-21.

(15) M. EFROYMSON – Multiple regression analysis dans : RALSTON-WILF Mathematical methods for digital computers, Wiley, New-York, 1962.

(16) EICHER – L'éducation comme investissement : la fin des illusions ? L'économie de l'éducation. *Revue de l'économie politique*, 1973.

(17) G. FERGUSON – The concept of parsimony in factor analysis. *Psychometrika*, 1954, *19* (4), 281-290.

(18) J. GORMAN et R. TOMAN – Selection of variables for fitting equations to data. *Technometrics*, 1966, *8* (1), 27-51.

(19) F. GRAYBILL – An introduction to linear statistical models. Mac Graw-Hill, New-York, 1961.

(20) B.I. HART – Tabulation of probabilities for the ratio of the mean square successive difference to the variance. *A.M.S.*, 1942, *13*, 207-214.

(21) R.R. HOCKING et R.N. LESLIE – Selection of the best subset in regression analysis. *Technometrics* 1967, *9* (4), 531-540.

(22) A. HOERL et R. KENNARD – Ridge regression : biased estimation for non orthogonal problems. *Technometrics*, 1970, *12* (1), 55-67.

(23) A. HOERL et R. KENNARD – Ridge regression : application to non-orthogonal problems. *Technometrics*, 1970, *12* (2), 69-82.

(24) C. HUANG et B. BOLCH – On the testing of regression disturbances for normality. *J.A.S.A.*, 1974, *69*, p. 330-335.

(25) M.F. KAISER – The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 1958, *23* (3), 187-200.

(26) M.G. KENDALL et A. STUART – The advanced theory of statistics, Griffin Londres 1966.

(27) M.G. KENDALL – Regression, Structure and Functional Relationship Part 1 – *Biometrika,* 1951, *38,* 11-25 ;
Part 2 – *Biometrika,* 1952, *39,* 96-108.

(28) N. LACOURLY – Thèse de 3ème cycle (Université PARIS VI), 1974.

(29) D. LAWLEY et A. MAXWELL – Factor analysis as a statistical method. London, 1963.

(30) D. LAWLEY et A. MAXWELL – Regression and factor analysis *Biometrika,* n° *60,* 1973.

(31) E. MALINVAUD – Méthodes statistiques de l'économétrie Dunod, 1964.

(32) C.L. MALLOWS – Choosing variables in a linear regression : a graphical aid. Presented at C.R.M. of I.M.S. Manhattan – Kansas, 1964.

(33) D. MARQUARDT – Generalized inverses, ridge regression, biased linear estimation and non linear estimation. *Technometrics,* vol. 12, n° 3, 1970, 591-612.

(34) J.A. MORGAN et J.F. TATAR – Calculation of the residual sum of squares for all possible regressions. *Technometrics* 1972, *14* (2), 317-326.

(35) J. VON NEUMANN – Distribution of the ratio of the mean square successive difference to the variance. *A.M.S.,* 1941, *12,* 367-395.

(36) P. POPE, J. WEBSTER – The use of an F-statistic in stepwise-regression procedures. *Technometrics,* 1972, *14* (2), 327-340.

(37) A.R. PREST – *Rev. Ec. Stat.* 1949, *31,* 33.

(38) J. RAMSEY – Tests for specification errors in classical linear least squares regression analysis. *J.R.S.S.* 1969, B, *31,* p. 350-371.

(39) J. RAMSEY et R. GILBERT – Some small samples properties of tests for specification errors. *J.A.S.A.* 1972, *67,* p. 180-186.

(40) C.R. RAO. et S.K. MITRA – Generalized inverse of matrices and its application. Wiley, New-York, 1971.

(41) M. SCHATZOFF, R. TSAO, S. FIENBERG – Efficient calculation of all possible regressions *Technometrics,* 1968, vol. *10,* n° 4, 769-779.

(42) S. SHAPIRO et M. WILK – An analysis of variance test for normality (complete samples) *Biometrika* 1965, *52,* p. 591-612.

(43) S. SHAPIRO, M. WILK, H. CHEN – A comparative study of various tests for normality. *J.A.S.A.* 1968, *63,* 1343-1372.

(44) H. THEIL – The analysis of disturbances in regression analysis. *J.A.S.A.,* 1965, *60,* 1067-1079.

(45) H. THEIL – A Simplification of the BLUS procedure for analyzing regression disturbances. *J.A.S.A.*, 1968, *63*, 242-251.

(46) H. THEIL et A. NAGAR – Testing the independance of regression disturbances. *J.A.S.A.* 1961, *56*, 793-806.

(47) R. TOMASSONE – Une méthode d'investigation : la regression orthogonale *Ann. Sci. Forest.* 1967, *24* (3), 233-258.

(48) H. WOLD – On least-squares regression with auto-correlated variables and residuals. *Bull. Inst. Int. Statist.* 1950, *32*.

(49) M.O. LEBEAUX – Programmes de régression et de classification utilisant la notion de voisinage. Thèse 3ème cycle Université Paris VI, 1974.