

REVUE DE STATISTIQUE APPLIQUÉE

CATHERINE PARDOUX

Problèmes statistiques posés par un modèle de régression étudiant l'évolution d'une démographie scolaire

Revue de statistique appliquée, tome 25, n° 4 (1977), p. 79-91

http://www.numdam.org/item?id=RSA_1977__25_4_79_0

© Société française de statistique, 1977, tous droits réservés.

L'accès aux archives de la revue « *Revue de statistique appliquée* » (<http://www.sfds.asso.fr/publicat/rsa.htm>) implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
<http://www.numdam.org/>

PROBLÈMES STATISTIQUES POSÉS PAR UN MODÈLE DE RÉGRESSION ÉTUDIANT L'ÉVOLUTION D'UNE DÉMOGRAPHIE SCOLAIRE

Catherine PARDOUX

Laboratoire de statistique - Université Paris IX - Dauphine

I – INTRODUCTION

De nombreux modèles mathématiques ont été élaborés depuis une quinzaine d'années pour la planification de l'enseignement (cf. [8]). Mais aucun d'entre eux n'étudie les flux scolaires en fonction de l'âge, facteur pourtant reconnu comme étant extrêmement important.

Le modèle de régression qui va être exposé a pour but d'expliquer les flux scolaires en fonction de l'âge dans le premier cycle de l'enseignement secondaire public français.

Nous allons établir ce modèle, puis l'étudier. Nous verrons alors que son étude statistique a soulevé beaucoup de problèmes.

Pour un exposé plus complet des données utilisées et des résultats obtenus, on pourra se reporter à [12].

II – POSITION DU PROBLEME

Le Ministère de l'Education [13] donne la répartition par âge et les origines scolaires non réparties par âge dans chacune des années d'études du premier cycle de l'enseignement secondaire public. On dispose donc pour chaque année d'études "j" ($j = 1$ pour la sixième, $j = 4$ pour la troisième) et pour chaque année scolaire "t" (t variant de t_1 à t_5 : $t = t_1$ en 1968-69, $t = t_5$ en 1972-73) d'un tableau croisé –retard-âge, origine– dont on connaît seulement les marges :

origine retard- âge	Redoublants	Promus	Venant du privé ou de fin d'études	Total
- 1 et -	$R_{-1,j}(t)$	$P_{-1,j}(t)$	$E_{-1,j}(t)$	$N_{-1,j}(t)$
0	$R_{0,j}(t)$	$P_{0,j}(t)$	$E_{0,j}(t)$	$N_{0,j}(t)$
1	$R_{1,j}(t)$	$P_{1,j}(t)$	$E_{1,j}(t)$	$N_{1,j}(t)$
2 et +	$R_{2,j}(t)$	$P_{2,j}(t)$	$E_{2,j}(t)$	$N_{2,j}(t)$
Total	$R_{.,j}(t)$	$P_{.,j}(t)$	$E_{.,j}(t)$	$N_{.,j}(t)$

L'indice retard-âge "i" s'interprète ainsi :

i = - 1 pour les élèves ayant un "an d'avance", ,

i = 2 pour les élèves qui ont deux "ans de retard".

Le problème est donc d'estimer pour chaque retard-âge "i" les $R_{i,j}(t)$, $P_{i,j}(t)$ et $E_{i,j}(t)$.

Le modèle va en fait permettre d'estimer les valeurs moyennes des coefficients de promotion et de redoublement suivants :

$$a_{i,j-1}(t-1) = \frac{P_{i,j}(t)}{N_{i,j-1}(t-1)}$$

et :

$$b_{i-1,j}(t-1) = \frac{R_{i,j}(t)}{N_{i-1,j}(t-1)}$$

III – FORMALISATION DU MODELE

III.1 - Les hypothèses

Avant d'établir un modèle, nous avons fait les six hypothèses suivantes :

H₁ : Un élève qui n'est pas nouveau dans le cycle ne peut pas suivre une année d'études sans avoir suivi l'année d'études précédente.

H₂ : Un élève n'abandonne pas son année d'études en cours d'année scolaire.

H₃ : Un élève ne peut avoir plus d'une année d'avance dans une année d'études : on a donc : $R_{-1,j}(t) = 0$ pour $j = 1$ à 4.

H₄ : Un élève qui a deux ans de retard ou plus ne peut pas redoubler son année d'études ; ce qui implique : $b_{2,j}(t) = 0$ pour $j = 1$ à 4.

H₅ : Il y a indépendance entre l'orientation des élèves à l'issue d'une année d'études et leur situation dans cette année d'études (nouveau ou redoublant) : hypothèse dite d'"homogénéité".

H_6 : On suppose que pour tous i, j et t :

$$E_{i,j}(t) = \frac{E_{i,j}(t)}{N_{i,j}(t)} N_{i,j}(t).$$

– Les deux premières hypothèses sont généralement admises pour les études de déroulement des scolarités.

– Les deux suivantes ont été faites à la suite de la prise en considération de résultats d'enquêtes.

La cinquième hypothèse limite l'interprétation des résultats, ce qui la rend fondamentale pour le modèle. Elle implique que le modèle permettra d'estimer les distributions par âge des promus et des redoublants, mais ne permettra pas d'étudier les "cursus scolaires" qui feront l'objet d'un autre modèle (cf. [12], VI-3).

– La sixième hypothèse a été faite pour simplifier le modèle après avoir constaté que les entrées intermédiaires représentaient en moyenne 5 % de l'effectif total d'une année d'études. Elle rend les $E_{i,j}(t)$ exogènes au modèle.

Ces hypothèses impliquent que l'on peut estimer pour tous "j" et "t" les nombres de promus de retard-âge "-1" : $P_{-1,j}(t)$, et les coefficients de promotion correspondants : $a_{-1,j-1}(t-1)$ (cf. [12], I-3).

Il reste donc à estimer les nombres de redoublants et de promus pour les autres retards-âge.

III.2 – Le modèle

On dispose pour ces estimations de 16 équations par année scolaire ; en effet, pour $j = 1$ à $4'$ et $i = 0, 1, 2$, on a :

$$\begin{cases} N_{i,j}(t) = R_{i,j}(t) + P_{i,j}(t) + E_{i,j}(t) \\ R_{i,j}(t) = R_{0,j}(t) + R_{1,j}(t) + R_{2,j}(t) \end{cases}$$

Puisque les $E_{i,j}(t)$ sont exogènes au modèle, posons :

$$Y_{i,j}(t) = N_{i,j}(t) - E_{i,j}(t) ;$$

à partir des définitions posées pour les coefficients de promotion et de redoublement, on obtient le modèle suivant :

$$\begin{cases} Y_{i,t}(t) = a_{i,j-1}(t-1) N_{i,j-1}(t-1) + b_{i-1,j}(t-1) N_{i-1,j}(t-1) \\ R_{i,j}(t) = b_{-1,j}(t-1) N_{-1,j}(t-1) + b_{0,j}(t-1) N_{0,j}(t-1) + \\ \quad + b_{1,j}(t-1) N_{1,j}(t-1) \end{cases}$$

On dispose pour les variables exogènes et endogènes de ce modèle de cinq observations qui ne peuvent pas être considérées comme fournies sans erreurs – puisqu'elles sont agrégées à partir des renseignements des nombreux établissements scolaires concernés.

Le modèle n'est donc pas un modèle déterministe, car les valeurs exactes des paramètres ne peuvent pas correspondre par le modèle aux valeurs connues des variables.

D'autre part, les coefficients de transition (i.e. de redoublement et de promotion) varient au cours du temps d'après les informations fournies par le Ministère de l'Éducation. On va être ainsi amené à les traiter d'une manière bien particulière.

III.3 – Les coefficients et leur caractère aléatoire

Puisque les coefficients de transition ne peuvent pas être tenus pour constants dans le temps, on a estimé qu'il serait intéressant de leur affecter un caractère aléatoire. Mais on ne dispose pas d'informations suffisantes pour établir des hypothèses a priori sur leurs distributions. Ceci nous a ainsi amenés à juger qu'on ne pouvait considérer que leurs valeurs moyennes au cours de la période étudiée.

On les a donc décomposés en somme de leurs valeurs moyennes au cours de la période considérée et d'un écart :

$$a_{i,j}(t) = \alpha_{i,j} + \mu_{i,j}(t)$$

$$b_{i,j}(t) = \beta_{i,j} + \nu_{i,j}(t).$$

Cette décomposition entraîne que les quantités $\mu_{i,j}$ et $\nu_{i,j}$ sont centrées.

On va ainsi avoir un modèle de régression dont les écarts ne sont pas des variables aléatoires de variances correspondant aux erreurs d'échantillonnage, l'échantillon étant énorme, mais sont fonctions des variables explicatives.

Du fait que les $a_{i,j}(t)$ et $b_{i,j}(t)$ sont des proportions qui doivent évidemment satisfaire l'inégalité : $a_{i,j}(t) + b_{i,j}(t) \leq 1$; on a les contraintes suivantes pour les $\alpha_{i,j}$ et $\beta_{i,j}$:

- $\alpha_{i,j} \geq 0, \beta_{i,j} \geq 0$ pour tous i et j
- $\alpha_{i,j} + \beta_{i,j} \leq 1$ pour $i = 0, 1$ et $j = 1$ à 3
- $\alpha_{2,j} \leq 1$ pour tout j
- $\alpha_{i,0} \leq 1$ pour tout i
- $\beta_{i,4} \leq 1$ pour $i \neq 2$
- $\beta_{-1,j} \leq 1 - \hat{\alpha}_{-1,j}$ pour tout j ,

si on désigne par $\hat{\alpha}_{-1,j}$ la moyenne arithmétique des estimations $\hat{\alpha}_{-1,j}(t)$ au cours de la période étudiée.

III.4. – Ecriture matricielle du modèle

On peut écrire le modèle de la façon suivante si on tient compte des décompositions des coefficients de transition ; on a pour $i = 0, 1, 2$ et $j = 1$ à 4 :

$$Y_{i,j}(t) = \alpha_{i,j-1}N_{i,j-1}(t-1) + \beta_{i-1,j}N_{i-1,j}(t-1) + \mu_{i,j-1}(t-1)N_{i,j-1}(t-1) + \nu_{i-1,j}(t-1)N_{i-1,j}(t-1)$$

$$R_{.,j}(t) = \beta_{-1,j}N_{-1,j}(t-1) + \beta_{0,j}N_{0,j}(t-1) + \beta_{1,j}N_{1,j}(t-1) + \nu_{-1,j}(t-1)N_{-1,j}(t-1) + \nu_{0,j}(t-1)N_{0,j}(t-1) + \nu_{1,j}(t-1)N_{1,j}(t-1)$$

On va maintenant l'écrire sous forme matricielle. Pour cela désignons par n le nombre d'années scolaires considérées et posons :

$$X = \begin{pmatrix} X_{1,1} & 0 & 0 & 0 \\ 0 & X_{1,2} & 0 & 0 \\ 0 & 0 & X_{1,3} & 0 \\ 0 & 0 & 0 & X_{1,4} \\ X_{2,1} & 0 & 0 & 0 \\ 0 & X_{2,2} & 0 & 0 \\ 0 & 0 & X_{2,3} & 0 \\ 0 & 0 & 0 & X_{2,4} \end{pmatrix} \text{ avec } \begin{cases} X_{1,j} = \begin{pmatrix} \dot{N}_{0,j-1} & \dot{N}_{-1,j} & 0 & 0 & 0 & 0 \\ 0 & 0 & \dot{N}_{1,j-1} & \dot{N}_{0,j} & 0 & 0 \end{pmatrix} \\ X_{2,j} = \begin{pmatrix} 0 & \dot{N}_{-1,j} & 0 & \dot{N}_{0,j} & 0 & \dot{N}_{1,j} \end{pmatrix} \\ \dot{N}_{i,j} = (N_{i,j}(t-1)) \end{cases}$$

$$\tilde{X} = \begin{pmatrix} \tilde{X}_{1,1} & 0 & 0 & 0 \\ 0 & \tilde{X}_{1,2} & 0 & 0 \\ 0 & 0 & \tilde{X}_{1,3} & 0 \\ 0 & 0 & 0 & \tilde{X}_{1,4} \\ \tilde{X}_{2,1} & 0 & 0 & 0 \\ 0 & \tilde{X}_{2,2} & 0 & 0 \\ 0 & 0 & \tilde{X}_{2,3} & 0 \\ 0 & 0 & 0 & \tilde{X}_{2,4} \end{pmatrix} \text{ avec } \begin{cases} \tilde{X}_{1,j} = \begin{pmatrix} \tilde{N}_{0,j-1} & \tilde{N}_{-1,j} & 0 & 0 & 0 & 0 \\ 0 & 0 & \tilde{N}_{1,j-1} & \tilde{N}_{0,j} & 0 & 0 \\ 0 & 0 & 0 & 0 & \tilde{N}_{2,j-1} & \tilde{N}_{1,j} \end{pmatrix} \\ \tilde{X}_{2,j} = \begin{pmatrix} 0 & \tilde{N}_{-1,j} & 0 & \tilde{N}_{0,j} & 0 & \tilde{N}_{1,j} \end{pmatrix} \\ \tilde{N}_{i,j} = \begin{pmatrix} & & & & 0 \\ & & & & \vdots \\ & & & & N_{i,j}(t-1) \\ 0 & & & & \end{pmatrix} \end{cases}$$

Ces matrices X et \tilde{X} sont donc les matrices où sont rangées convenablement les observations des variables prédéterminées du modèle. Définissons maintenant le vecteur \underline{Y} des observations des variables endogènes, le vecteur \underline{p} des paramètres à estimer et le vecteur \underline{u} :

$$\underline{Y} = \begin{pmatrix} \underline{Y}_1 \\ \underline{Y}_2 \\ \underline{Y}_3 \\ \underline{Y}_4 \\ \underline{R}_{.,1} \\ \underline{R}_{.,2} \\ \underline{R}_{.,3} \\ \underline{R}_{.,4} \end{pmatrix} \text{ avec } \begin{cases} \underline{Y}_j = \begin{pmatrix} \underline{Y}_{0,j} \\ \underline{Y}_{1,j} \\ \underline{Y}_{2,j} \end{pmatrix} \text{ et } : \underline{Y}_{i,j} = (Y_{i,j}(t)) \\ \underline{R}_{.,j} = (R_{i,j}(t)) \end{cases}$$

$$\underline{p} = \begin{pmatrix} \underline{p}_1 \\ \underline{p}_2 \\ \underline{p}_3 \\ \underline{p}_4 \end{pmatrix} \text{ avec } : \underline{p}_j = \begin{pmatrix} \alpha_{0,j-1} \\ \beta_{1,j} \\ \alpha_{1,j-1} \\ \beta_{0,j} \\ \alpha_{2,j-1} \\ \beta_{1,j} \end{pmatrix}$$

$$\underline{u} = \begin{pmatrix} \underline{u}_1 \\ \underline{u}_2 \\ \underline{u}_3 \\ \underline{u}_4 \end{pmatrix} \text{ avec } : \underline{u}_j = \begin{pmatrix} \underline{\mu}_{0,j-1} \\ \underline{\nu}_{-1,j} \\ \underline{\mu}_{1,j-1} \\ \underline{\nu}_{0,j} \\ \underline{\mu}_{2,j-1} \\ \underline{\nu}_{1,j} \end{pmatrix} \text{ et } : \begin{cases} \underline{\mu}_{i,j} = (\mu_{i,j}(t-1)) \\ n \times 1 \\ \underline{\nu}_{i,j} = (\nu_{i,j}(t-1)) \\ n \times 1 \end{cases}$$

Toutes ces définitions étant posées, le modèle s'écrit :

$$\underline{Y} = X\underline{p} + \tilde{X}\underline{u}.$$

IV – ETUDE STATISTIQUE DU MODELE

IV.1 – La méthode d'estimation utilisée et les problèmes statistiques

Si on pose : $\underline{e} = \tilde{X}\underline{u}$, on obtient pour décrire les flux scolaires le modèle de régression linéaire suivant écrit sous sa forme la plus classique :

$$\underline{Y} = X\underline{p} + \underline{e}$$

$$\text{avec } : \begin{cases} E^X(\underline{e}) = \tilde{X} E(\underline{u}) = \underline{0} \\ E^X(\underline{e}^t \underline{e}) = \Sigma = \tilde{X} \Omega^t \tilde{X} \end{cases}$$

si on désigne par Ω la matrice des variances-covariances des coefficients de transition.

Ce modèle est autorégressif et ses écarts sont fonctions des variables explicatives. Ces deux caractéristiques vont rendre son étude statistique délicate.

Les études de T.C. Lee, G.G. Judge et A. Zellner (cf. [9], [10]) nous ont amenés à considérer que la méthode des moindres carrés était la mieux adaptée à ce modèle.

On ne peut pas estimer valablement la matrice Ω des variances-covariances des coefficients de transition puisqu'on ne dispose que de cinq observations sur les taux globaux de transition - les taux globaux étant les taux par année d'études et non par âge.

On ne peut donc pas employer la méthode des moindres carrés généralisée.

Nous avons donc utilisé la méthode des moindres carrés ordinaire qui donne pour estimateur de \underline{p} :

$$\hat{\underline{p}} = ({}^t\mathbf{X}\mathbf{X})^{-1} {}^t\mathbf{X}\mathbf{Y} ;$$

cet estimateur existe puisqu'on peut aisément vérifier que \mathbf{X} est de plein rang (cf. [12], III-3).

Il est linéaire en \mathbf{Y} , sans biais si on raisonne conditionnellement aux variables explicatives, mais ce n'est pas le "meilleur" relativement au critère de la plus petite variance.

On doit toujours raisonner conditionnellement aux variables explicatives pour étudier les propriétés de $\hat{\underline{p}}$. Mais on peut remarquer que le modèle étant autorégressif, il est très délicat d'établir des propriétés conditionnellement à des variables qui ont été endogènes à la période précédente. L'espérance et la variance d'un estimateur n'ont en effet pas beaucoup de sens dans ce cas.

D'autre part, on a tenu compte de l'évolution du modèle dans le temps en considérant qu'on ne pouvait estimer que des coefficients moyens. Ceci implique que la régression n'est envisageable que pour un petit nombre d'années. On constatera en effet au § V.2 que les solutions de la régression évoluent quand on retire une année, puis deux à la période considérée. On en conclut alors que l'étude des propriétés asymptotiques de l'estimateur des paramètres du modèle des flux scolaires est sans intérêt.

On a été ainsi amené à considérer que les problèmes de ce type ne relevaient pas des méthodes d'estimation, mais des méthodes d'ajustement.

IV.2 – La protection de la régression

Si on interprète géométriquement ce problème d'ajustement dans l'espace E – soit \mathcal{R}^{16n} – des "observations" muni de la norme euclidienne, la solution des moindres carrés ordinaire correspond au point $\hat{\underline{Y}} = \mathbf{X}\hat{\underline{p}}$ le plus proche de \underline{Y} .

$\hat{\underline{Y}}$ est donc la projection de \underline{Y} sur le sous-espace vectoriel H engendré par les 24 vecteurs colonnes de \mathbf{X} . Ces vecteurs sont linéairement indépendants puisqu'on a vu que la matrice \mathbf{X} est de plein rang. On en déduit alors que $\hat{\underline{p}}$ est unique.

Mais si on examine la matrice \mathbf{X} , on peut remarquer que d'après sa constitution générale, ses vecteurs colonnes linéairement indépendants sont soumis à des variations de même nature significatives d'un certain nombre de liaisons.

Les variables explicatives qui leur correspondent sont donc corrélées. J.P. Benzecri (cf. [2]) montre que dans ce cas, la régression risque d'être de mauvaise qualité si on ne la "protège" pas. En effet, si m vecteurs ($m \leq 24$) parmi les vecteurs colonnes de \mathbf{X} satisfont approximativement à une relation linéaire, alors du fait des erreurs, H est mal défini. Il n'est guère mieux déterminé que comme sous-espace vectoriel arbitraire de dimension 24 contenant un des m vecteurs et les $(24 - m)$ autres vecteurs. Il se peut alors que \underline{Y} paraisse contenu dans H sans que ce soit le cas et on a alors une décomposition de \underline{Y} sur les 24 vecteurs complètement dépourvue de sens avec en particulier, des coefficients qui peuvent être élevés et de signes opposés. Si on impose aux coefficients de la régression d'être

positifs, toutes les variables intervenant dans le modèle étant positives, on élimine donc ce type de combinaisons linéaires de variables explicatives positives se soustrayant entre elles parce qu'affectées de coefficients de signes opposés.

On a donc "protégé" la régression en prenant en compte les contraintes nécessaires pour que les coefficients aient un sens (cf. § III.3).

IV.3. – Une propriété intéressante de la solution du problème

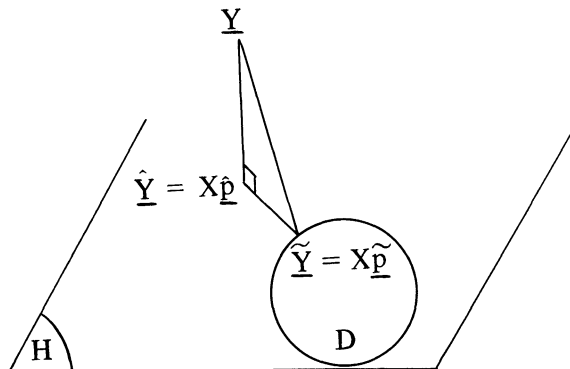
On a ainsi obtenu le problème suivant : minimiser la forme quadratique convexe : $Q(\underline{p}) = \|\underline{Y} - X\underline{p}\|^2$ sur l'ensemble convexe fermé C ainsi décrit :

$$C = \left\{ \begin{array}{l} \underline{p} = {}^t(p_1, p_2, \dots, p_{24}) \text{ tels que : } p_i \geq 0 \quad \forall i, \\ 1 - p_i \geq 0 \quad \text{pour } i = 1, 3, 5, 11, 17, 22, 23, 24, \\ 1 - p_4 - p_7 \geq 0, 1 - p_6 - p_9 \geq 0, 1 - p_{10} - p_{13} \geq 0, \\ 1 - p_{12} - p_{15} \geq 0, 1 - p_{16} - p_{19} \geq 0, 1 - p_{18} - p_{21} \geq 0, \\ 1 - p_2 \geq \hat{\alpha}_{-1,1}, 1 - p_8 \geq \hat{\alpha}_{-1,2}, 1 - p_{14} \geq \hat{\alpha}_{-1,3}, 1 - p_{20} \geq \hat{\alpha}_{-1,4} \end{array} \right\}$$

Si on appelle D le sous-ensemble convexe fermé de E suivant :

$$D = \{ \underline{Z} \in E \mid \underline{Z} = X\underline{p} \text{ avec } \underline{p} \in C \},$$

on montre aisément que la projection de \underline{Y} sur D est unique et correspond à la projection de $\hat{\underline{Y}}$ sur la frontière de D si \underline{Y} n'appartient pas à D (cf. [5], II.2).



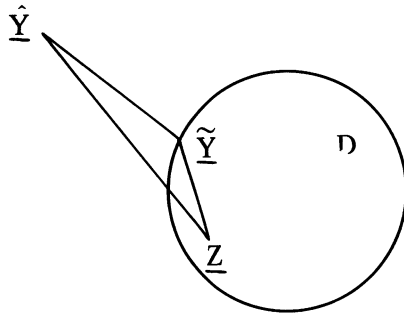
Une propriété intéressante –due à la convexité de C– de la solution $\tilde{\underline{p}}$ a été démontrée :

Si $\underline{p} \in C$, $X\underline{p}$ est plus proche de $\tilde{\underline{Y}} = X\tilde{\underline{p}}$ que de $\hat{\underline{Y}} = X\hat{\underline{p}}$ au sens de la métrique euclidienne, ou encore \underline{p} est plus proche de $\tilde{\underline{p}}$ que de $\hat{\underline{p}}$ au sens de la métrique induite par tXX ;

en effet, pour tout $\underline{Z} \in D$, $\tilde{\underline{Y}}$ étant la projection de $\hat{\underline{Y}}$ sur D, on a la relation classique :

$$\langle \hat{\underline{Y}} - \tilde{\underline{Y}}, \underline{Z} - \tilde{\underline{Y}} \rangle \leq 0, \quad (1)$$

relation qui est encore valable si on remplace \hat{Y} par Y puisque \tilde{Y} est la projection de Y sur D ;



(1) s'écrit encore :

$$\langle \hat{Y} - Z + Z - \tilde{Y}, Z - \tilde{Y} \rangle \leq 0,$$

en développant, on obtient :

$$\langle \hat{Y} - Z, Z - \tilde{Y} \rangle + \|Z - \tilde{Y}\|^2 \leq 0,$$

ce qui entraîne en appliquant l'inégalité de Cauchy-Schwarz :

$$\|Z - \tilde{Y}\|^2 \leq \langle \hat{Y} - Z, \tilde{Y} - Z \rangle \leq \|\hat{Y} - Z\| \|\tilde{Y} - Z\|,$$

et si Z est différent de \tilde{Y} , on a :

$$\|Z - \tilde{Y}\| \leq \|Z - \hat{Y}\|.$$

V – ETUDE DE LA VALIDITE DES RESULTATS

La résolution de ce modèle a été faite à l'aide du programme de régression à contraintes positives écrit par P. CAZES et adapté pour tenir compte des contraintes linéaires autres que les contraintes de positivité. La méthode programmation quadratique utilisée par ce programme est la méthode de D'ESOPO (cf. [4], IV.4.2).

V.1 – Mesure de la qualité de la régression

On mesure habituellement la "qualité de l'approximation" de Y par \tilde{Y} par le rapport :

$$R^2 = \frac{\|\tilde{Y}\|^2}{\|Y\|^2},$$

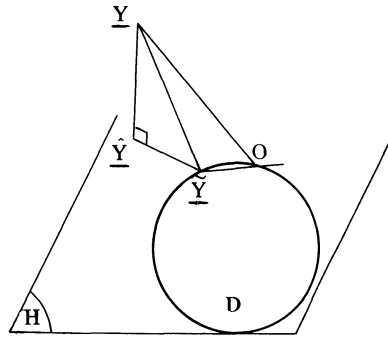
et on calcule le "résidu" :

$$S^2 = \frac{\|Y - \tilde{Y}\|^2}{\|Y\|^2}.$$

On peut remarquer que le numérateur de S^2 est égal au minimum sur C de la forme quadratique Q .

Lorsque \tilde{Y} est égal à \hat{Y} , R et S représentent respectivement le cosinus et le sinus de l'angle $(YQ\hat{Y})$ et on a alors :

$$R^2 + S^2 = 1.$$



Plus généralement, lorsque D contient l'origine – ce qui est le cas ici puisque l'origine appartient à la frontière de D – on a : $R^2 + S^2 \leq 1$; en effet :

$$\|Y\|^2 = \|Y - \tilde{Y}\|^2 + \|\tilde{Y}\|^2 + 2 \langle Y - \tilde{Y}, \tilde{Y} \rangle,$$

en divisant par $\|Y\|^2$, on obtient :

$$R^2 + S^2 = 1 - 2 \frac{\langle Y - \tilde{Y}, \tilde{Y} \rangle}{\|Y\|^2} ;$$

Comme la relation (1) du paragraphe précédent est toujours valable si on remplace \hat{Y} par \underline{Y} et comme $\underline{Z} = \underline{Q}$ appartient à D , on a :

$$\langle Y - \tilde{Y}, \tilde{Y} \rangle \geq 0,$$

l'égalité étant réalisée lorsque $\tilde{Y} = \hat{Y}$.

Les rapports R^2 et S^2 ne permettent pas systématiquement de conclure quant à la validité des résultats :

– R^2 peut être voisin de 1 sans pour cela que le modèle soit bon : ceci peut être le cas lorsque le nombre d'observations est peu différent du nombre de variables ;

– ou R^2 peut être voisin de 0 bien que le modèle soit adapté : ceci est le cas lorsque du fait des fluctuations d'échantillonnage, \underline{Y} se trouve éloigné du sous-espace vectoriel H .

Mais, lorsque le modèle posé est bon, on a en général \underline{Y} proche du sous-espace vectoriel H , ce qui justifie le fait d'apprécier la qualité de la solution trouvée par le rapport R^2 bien que ce soit S^2 qu'on minimise.

V.2. – Interprétation et critique des résultats

Examinons les valeurs de R^2 et S^2 pour les trois régressions sous contraintes effectuées :

n	1967-68 à 71-72 5	1968-69 à 71-72 4	1969-70 à 71-72 3
R^2	0,994	0,995	0,996
S^2	0,004	0,002	0,001
$R^2 + S^2$	0,998	0,997	0,997

On constate que pour chacune de ces régressions, R^2 est toujours très voisin de 1 et S^2 très voisin de 0. Mais les valeurs du résidu S^2 sont très peu différentes pour les régressions sans contraintes et avec contraintes bien que \tilde{p} soit fort éloigné de \hat{p} .

On en a conclu que la forme quadratique $Q(\underline{p})$ devait admettre un minimum très plat en \hat{p} . Il y a donc vraisemblablement d'autres solutions auxquelles correspondent des résidus de faible valeur.

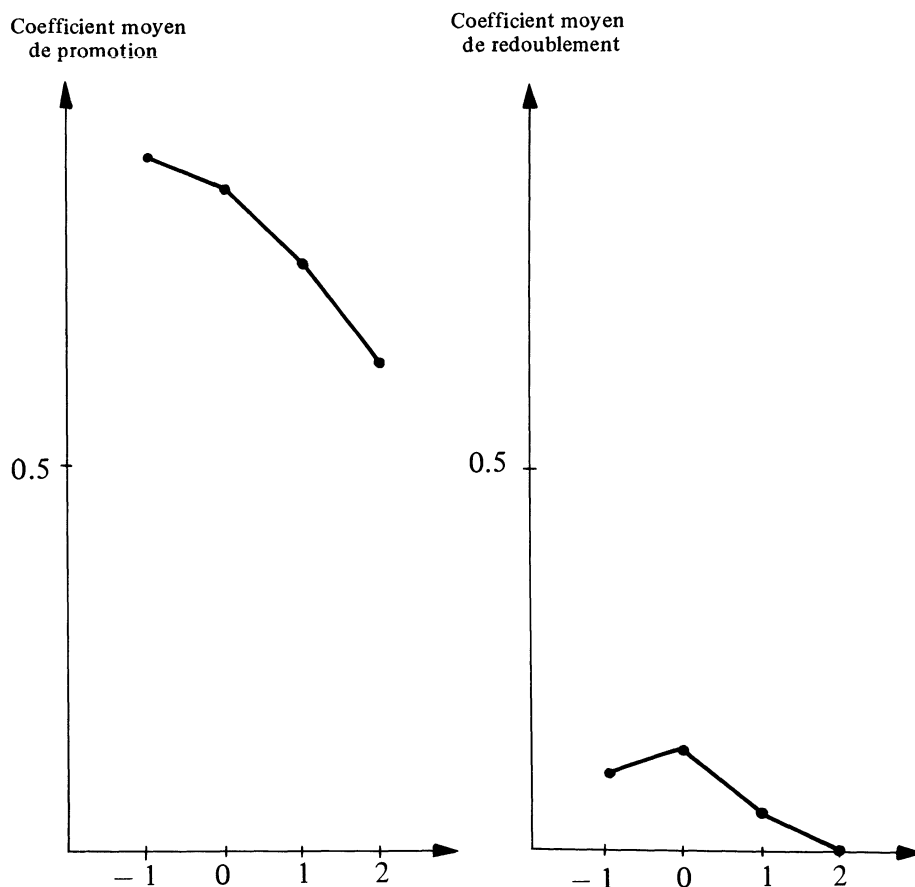
Bien que la cohérence des valeurs trouvées ainsi que la protection imposée à la régression nous permettent de considérer les résultats comme crédibles, on ne tiendra les solutions trouvées que pour approximatives à cause d'une part, des erreurs de mesure et d'autre part, d'une erreur de spécification sur le modèle due à l'hypothèse H_6 sur la répartition des entrées intermédiaires.

Nous n'allons examiner qu'une partie des résultats – pour les autres résultats, on se référera à [12].

Le tableau ci-dessous donne les estimations des valeurs moyennes des coefficients de redoublement et de promotion à l'issue de la quatrième au cours des trois périodes considérées ; les $\tilde{\alpha}_{-1,3}$ sont les moyennes arithmétiques des $\hat{\alpha}_{-1,3}(t)$ estimés directement (cf. § III.1) et les autres coefficients sont les solutions des régressions sous contraintes :

n	5	4	3
$\tilde{\alpha}_{-1,3}$	0,89	0,90	0,91
$\tilde{\beta}_{-1,3}$	0,11	0,10	0,09
$\tilde{\alpha}_{0,3}$	0,85	0,86	0,86
$\tilde{\beta}_{0,3}$	0,15	0,13	0,13
$\tilde{\alpha}_{1,3}$	0,75	0,77	0,77
$\tilde{\beta}_{1,3}$	0,05	0,05	0,05
$\tilde{\alpha}_{2,3}$	0,72	0,64	0,63
$\tilde{\beta}_{2,3}$	0	0	0

En traçant pour la régression correspondant à $n = 4$ les graphes des coefficients de promotion et de redoublement en fonction du retard-âge, on obtient les représentations suivantes :



On constate que conformément aux résultats des enquêtes (cf. [12], annexe III) : le coefficient de promotion décroît avec le retard-âge tandis que le coefficient de redoublement croît du retard-âge "-1" au retard-âge "0" pour décroître ensuite.

VI – CONCLUSION

Le modèle des flux scolaires pose beaucoup de problèmes statistiques, d'une part parce qu'on ne peut pas le classer parmi les modèles déterministes, d'autre part parce qu'il n'est pas un modèle autorégressif classique.

Les propriétés des modèles autorégressifs ont été jusqu'à présent peu étudiées. Des résultats généraux sur ces modèles devraient donc permettre de mieux appréhender les problèmes –assez usuels– du genre de ceux qui ont été soulevés dans cette étude.

BIBLIOGRAPHIE

- [1] ARAGON Y. — “Sur l'estimation des probabilités de transition d'une chaîne de Markov à partir de données marginales”, Thèse de 3ème cycle, Toulouse, 1971 ; même titre, *R.S.A.* 20 (2), 1972.
- [2] BENZECRI J.P. — “La régression”, Laboratoire de Statistique Mathématique, Paris VI, 1970.
- [3] BRENOT J., CAZES P., LACOURLY N. — “Pratique de la régression : qualité et protection”, *Cahiers du Bureau Universitaire de Recherche Opérationnelle* n° 23, Paris 1975.
- [4] CAZES P. — “Application de l'analyse des données au traitement de problèmes géologiques”, Thèse de 3ème cycle, Paris 1970.
- [5] CAZES P. — “Protection de la régression par utilisation de contraintes linéaires et non linéaires”, *R.S.A.* 23 (3), 1975.
- [6] DRAPER N., SMITH H. — “Applied regression analysis”, Wiley, New-York, 1966.
- [7] FOURGEAUD C., FUCHS A. — “Statistique”, Dunod, Paris, 1967.
- [8] JOHNSTONE J.N. — “Mathematical models developed for use in educational planning : a review”, *Review of educational research*, 44, 2, 1974.
- [9] LEE T.C., JUDGE G.G. — “Estimation of transition probabilities in a nonstationary finite Markov chain”, *Metroeconomica*, 34 (2), 1972.
- [10] LEE T.C., JUDGE G.G., ZELLNER A. — “Estimating the parameters of the Markov probability model from aggregate time series data” ; North Holland Publishing Company, Amsterdam, 1970.
- [11] MALINVAUD E. — “Méthodes statistiques de l'économétrie”, Dunod, 1964,
- [12] PARDOUX C. — “Problèmes statistiques posés par un modèle de régression étudiant l'évolution d'une démographie scolaire”, Thèse de 3ème cycle, Paris VI, 1977.
- [13] Tableaux de l'Education Nationale : Editions 1966 à 1974.