Statistique et analyse des données

Michel Bonneu

Une démarche méthodologique en modélisation statistique

Statistique et analyse des données, tome 12, n° 1-2 (1987), p. 28-45
<http://www.numdam.org/item?id=SAD_1987__12_1-2_28_0>

Statistique et Analyse des Données
1987 - Vol. 12 n° 1 et 2 - pp. 28-45

UNE DEMARCHE METHODOLOGIQUE EN MODELISATION STATISTIQUE

Michel BONNEU

Laboratoire de Statistique et Probabilités
U.A. CNRS 745
Université Paul Sabatier - TOULOUSE

Résumé : *Un problème de choix de modèles est illustré sur des données médicales : l'objectif est d'expliquer l'évolution de la masse osseuse chez une femme par des modèles de régression linéaire gaussienne, utilisant des variables quantitatives (ou leurs transformations) et un facteur qualitatif. Le critère de choix de modèles est une fonction de risque de prédiction dont l'estimation dans le cas d'un plan expérimental déséquilibré, est une statistique qui généralise le Cp-Mallows* [6]. *Le traitement informatique est réalisé par des macros en langage de commandes GLIM* [2].

Abstract : *A problem of model choice is illustrated on medical data : the aim is to explain women mineral bone loss by means of gaussian linear regression models which use quantitative covariates (or their transformations) and a qualitative factor. The criteria of model choice is a prediction risk function the estimation of which, in an unbalanced experimental design, is a statistic which generalizes the Cp-Mallows.The computer processing is done by macros in GLIM commande language.*

Mots clés : *fonction de risque de prédiction - Cp Mallows -*

Indices de classification STMA : *07-010, 07-070.*

Manuscrit reçu le 13.7.87, révisé le 26.10.87

0 - INTRODUCTION

En régression linéaire multiple, il existe une abondante littérature sur la sélection de variables explicatives. Des règles de décision pour choisir les "meilleurs" sous ensembles de variables explicatives ont été développées soit pour mettre au point des statistiques basées sur les moindres carrés quand le modèle de régression linéaire est défini a priori (carré moyen résiduel, statistique de Fisher, coefficient de corrélation multiple,...), soit pour définir des critères qui permettent de choisir un sous ensemble de variables explicatives quand le modèle de régression n'est pas défini a priori : critère d'Akaike [1], critère Cp-Mallows, erreur quadratique moyenne de prédiction, critère de validation croisée,...

Dans le cas de modèles de régression utilisés pour la prédiction, les variables explicatives retenues dans le modèle sont celles qui rendent minimum la fonction de risque de prédiction définie dans le paragraphe 1. Une estimation de cette fonction de risque est proposée dans le cas linéaire gaussien ainsi que sa mise en oeuvre.

Dans le paragraphe 2, les données sont présentées et décrites par des résumés et graphiques statistiques élémentaires.

La mise en oeuvre de l'estimation de la fonction de risque est réalisée dans le paragraphe 2.3, pour définir un premier modèle de référence. Dans le paragraphe 2.4, on décrit la stratégie de recherche exploratoire des meilleurs modèles, basée sur le risque de prédiction et des modèles, parmi les meilleurs, sont proposés pour la prédiction de la masse osseuse chez la femme.

1 - FONCTION DE RISQUE DE PREDICTION

1.1 - Définition

La définition est formulée dans le cas où les données expérimentales sont structurées de la façon suivante :

- le vecteur aléatoire Y (nx1) à expliquer, des observations y_{ij}.
- k conditions expérimentales avec un nombre inégal de

répétitions n_i tel que $n = \sum_{i=1}^{k} n_i$.

- la matrice X_p (kxp) de p variables explicatives de rang p, relative aux k conditions expérimentales.

Cet ensemble de p variables explicatives n'est pas supposé avoir été choisi a priori et peut contenir toute transformation ou combinaison de variables explicatives, jugées intéressantes.

- l'objectif est de prédire le vecteur Z de k variables aléatoires futures correspondant aux k conditions expérimentales, à partir d'un modèle faisant appel au plus petit nombre possible de variables explicatives. Le vecteur aléatoire Z est indépendant de Y.

La fonction de risque de prédiction associée au modèle courant est définie par analogie au critère d'Akaike, par la fonction de perte associée qui est égale à 2 fois le logarithme du rapport de la vraisemblance du modèle exact sur la vraisemblance estimée du modèle courant.

Le modèle exact est le modèle qui contient autant de paramètres inconnus que de conditions expérimentales.

La fonction de risque de prédiction a été explicitée dans le cas de modèles linéaires généralisés (M. BONNEU,1986 [3]).

L'estimation de ce risque de prédiction est toujours possible soit à partir d'un estimateur simulé du bootstrap, soit par un calcul asymptotique. Elle a été mise en oeuvre dans des modèles généralisant le modèle logit (M. BONNEU,1987 [4]).

Par contre, dans le cas linaire gaussien, un calcul analytique exact permet de déterminer un estimateur sans biais de variance minimale de la fonction de risque.

1.2 - Estimation dans le cas linéaire gaussien

Le modèle exact, dont le vecteur moyenne m(kx1) est inconnu, est défini par :

$\left| \begin{array}{l} Y = Am + \varepsilon \text{ avec } \varepsilon \sim N_n(0,\sigma^2 I_n) \quad \text{pour les données expérimentales.} \\ Z = m + \varepsilon \text{ avec } \varepsilon \sim N_k(0,\sigma^2 I_k) \quad \text{pour les variables à prédire.} \end{array} \right.$

La matrice A (nxk) du plan expérimental est telle que le vecteur $\bar{Y}(kx1)$ des moyennes arithmétiques $\bar{y}_i$ par condition expérimentale est :

$\bar{Y} = L^{-1}A^TY$, où L(kxk) est la matrice diagonale des répétitions n_i. La fonction de risque de prédiction associée au modèle courant dont le sous ensemble S de s variables explicatives est défini par la matrice $X_S(kxs)$, s'écrit :

$$r_S = \frac{1}{\sigma^2} E \|m - \Pi_S \bar{Y}\|^2 \text{ avec } \Pi_S = X_S\left(X_S^T X_S\right)^{-1} X_S^T$$

Le risque r_S s'explicite en fonction de l'erreur quadratique moyenne de prédiction MSEP $(\hat{Z}_S) = E\|Z-\Pi_S\bar{Y}\|^2$, définie pour des pondérations égales à 1, le cas où les pondérations sont différentes étant largement étudié par O.BUNKE, B.DROGE(1984) [5], dans un cadre analogue :

$$r_S = \frac{1}{\sigma^2} \text{MSEP}(\hat{Z}_S) - k$$

Compte tenu de la relation déduite de l'indépendance des v.a. $\bar{Y}$ et Z :

$$E\|Z-\Pi_S\bar{Y}\|^2 = E\|Z-m\|^2 + E\|\Pi_S\bar{Y}-m\|^2 = k\sigma^2 + E\|\Pi_S\bar{Y}-m\|^2$$

Propriété : *Dans le cas où n>k+2, l'estimateur de variance minimale dans la classe des estimateurs sans biais de r_S est :*

$$\mathcal{R}_S = \frac{1}{\hat{\sigma}^2} \|\bar{Y}-\Pi_S\bar{Y}\|^2 + 2\text{tr}\,\Pi_S L^{-1} - \text{tr}\,L^{-1}$$

l'estimateur $\hat{\sigma}^2$ de la variance des résidus σ^2 est :

$$\hat{\sigma}^2 = \frac{1}{n-k-2} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_i\right)^2$$

Remarques :

1° Dans l'estimation de $\frac{1}{\sigma^2}$, on divise par n-k-2 au lieu de n-k, pour avoir un estimateur sans biais (cf [3]).

2° L'expression de $\mathcal{R}_S$ met en évidence la nécessité pour un bon modèle d'un équilibre entre les 2 termes :

- $\|\bar{y} - \Pi_S \bar{y}\|^2$ qui mesure l'adéquation des données au modèle et qui diminue quand on augmente le nombre de variables explicatives.

- $\text{tr}\ \Pi_S L^{-1}$ qui diminue quand on diminue le nombre de variables explicatives.

3° Dans le cas où n=k, il n'y a pas de répétition et il n'est pas possible de déterminer un estimateur optimal de r_S quand σ^2 est inconnu.

L'estimation de σ^2 pose un problème et il faut la déterminer à partir d'un modèle de référence (M_0) défini par un sous ensemble S_0 de variables explicatives. Le paramètre σ^2 sera estimé par le carré moyen résiduel du modèle (M_0) :

$$\tilde{\sigma}^2 = \frac{D_{S_0}}{n - \text{card } S_0}$$

où $D_{S_0} = \|Y - \Pi_{S_0} Y\|$ déviance du modèle (M_0) qui n'est autre, dans le cas linéaire gaussien, que la somme des carrés résiduelle.

La statistique $\mathcal{R}_S$ a une expression simplifiée quand il n'y a pas de répétition, qui est analogue à la statistique de Mallows, à l'estimation de σ près :

$$C_S = \frac{D_S}{\tilde{\sigma}^2} + 2 \text{ card } S - n$$

1.3 - Stratégie de choix de modèles

Dans le cas linéaire gaussien le choix de modèles se ramène à trouver le sous ensemble S qui minimise le rique de prédiction r_S.

On considère un ensemble de modèles à explorer qui est défini par le spécialiste et le statisticien : par exemple le spécialiste apporte sa connaissance du problème a priori, alors que le statisticien propose dans un premier temps des modèles, à partir des résultats d'une première analyse descriptive des données. Cet ensemble de modèles s'enrichit au cours de l'exploration et de la validité de certains résultats.

La stratégie diffère selon les 2 cas suivants :

1.3.1 - Les répétitions sont en nombre suffisant (n>k+2)

Pour chaque modèle courant (M_s), on calcule la statistique de prédiction $\mathcal{R}_s$, au moyen de macros élaborés dans le logiciel GLIM.

Le modèle (M_s) est acceptable si la valeur de $\mathcal{R}_s$ est faible, comparativement à d'autres modèles et si l'analyse des estimations des coefficients de régression et l'analyse des résidus sont satisfaisantes.

On obtient ainsi un ensemble des meilleurs modèles acceptables dans lequel le spécialiste peut choisir celui dont l'interprétation est la plus simple.

1.3.2 - Il n'y a pas de répétition (n=k)

Le problème est de définir un modèle de référence qui permette d'estimer σ^2. On peut trouver des éléments de réponse en étudiant l'influence d'un petit nombre de variables importantes sur la variable à expliquer : on regroupe alors les données pour obtenir de nouvelles conditions expérimentales avec des répétitions. Il est alors possible de calculer la statistique $\mathcal{R}_s$ et de suivre la stratégie décrite au paragraphe 1.3.1, pour déterminer la prise en compte éventuelle d'une variable ou de sa transformée dans le modèle de référence (M_0).

Le modèle de référence (M_0) étant déterminé, l'estimation de σ^2 est alors réalisée. On reprend alors les étapes de la stratégie décrite au paragraphe 1.3.1, en calculant la statistique de Mallows pour estimer le risque de prédiction.

2 - APPLICATION A DES DONNES REELLES

2.1 - Origine des données

Les données ont été recueillies auprès de femmes en consultation au service d'Endocrinologie du C.H.U. Toulouse-Purpan du docteur C.RIBOT.

Après élimination d'individus pour lesquels des valeurs étaient aberrantes ou manquantes, ou bien d'individus

pathologiques (individus sous traitement hormonal, fumeurs,...), un échantillon de 778 femmes a été constitué pour répondre à l'objectif : "Pour mieux prévenir les risques d'ostéoporose, comment peut-on expliquer la perte de masse osseuse chez la femme ?"

L'objectif est de déterminer un modèle qui prédise au mieux la masse osseuse mesurée par la variable BMD (Bone Mass Density) en g/cm^2, notée y.

On dispose de la variable à expliquer BMD et des variables explicatives :

- âge, en ans : x^1
- poids, en kg : x^2
- taille, en cm : x^3
- surpoids : variable quantitative rendue qualitative à 2 modalités : SP^+ = avec surpoids ; SP^- = sans surpoids.
- statut ménopausique : variable qualitative à 5 modalités :

AJ : adulte jeune non ménopausée d'âge inférieur à 45 ans.
MN : ménopause naturelle de plus de 45 ans.
MP : ménopause précoce avant 40 ans.
MPC : ménopause post chirurgicale après 40 ans.
PM : périménopause : encore réglées et plus de 45 ans.

- ancienneté de ménopause, en ans : x^4 (quand il y a lieu).

Dans une étude antérieure (C.RIBOT, F.TREMOLIERES, J.M. POULHES, M.BONNEU, F. GERMAIN, J.P.LOUVET-1987) [7], une analyse de variance dans un plan complet déséquilibré à 2 facteurs contrôlés (statut ménopausique et surpoids) a permis de conclure que la variable surpoids n'avait pas d'effet sur le BMD pour les femmes non ménopausées. Les 2 facteurs ont été regroupés en une seule variable, la variable TYPE ayant 8 modalités :

TYPE1 = AJ ; TYPE2 = MN et SP^- ; TYPE3 = MN et SP^+ ;
TYPE4 = MP et SP^- ; TYPE5 = MPC et SP^- ; TYPE6 = PM ;
TYPE7 = MP et SP^+ ; TYPE8 = MPC et SP^+.

2.2 - Traitement descriptif

Une description statistique des données a été réalisée au moyen de calculs et graphiques classiques :

- calculs de paramètres statistiques, selon le type (cf tableau 1)
- calculs des coefficients de corrélation linéaire (cf

tableau 2)

- bi-plots : BMD en fonction de l'âge, selon le type
  BMD en fonction de l'ancienneté de la ménopause,...
- histogrammes : de la variable BMD, selon le type.
- analyse factorielle : analyse en composantes principales ; analyse factorielle discriminante pour le facteur type.

Cette première analyse descriptive des données est absolument nécessaire et antérieure à toute modélisation. elle permet notamment de préciser la liaison entre certaines variables et dans notre cas, de mettre en évidence l'influence prépondérante des variables type, âge et poids sur la variable BMD.

Variables / Type	âge	poids	taille	anc. de ménopause	BMD	âge de ménopause	effectif
1	33.9 ± 7.9 20 ⟷ 45	56.6 ± 7.8 40 ⟷ 88	161.6 ± 5.8 136 ⟷ 173	/	1.14 ± 0.11 0.84 ⟷ 1.52	/	141
2	58.9 ± 8.5 47 ⟷ 95	55.1 ± 6.6 37 ⟷ 80	158.8 ± 5.6 142 ⟷ 175	8.0 ± 8.4 0 ⟷ 45	0.95 ± 0.13 0.47 ⟷ 1.33	50.9 ± 2.9 41 ⟷ 59	286
3	57.1 ± 5.2 48 ⟷ 72	71.2 ± 11.0 49 ⟷ 103	158.3 ± 6.2 142 ⟷ 170	5.7 ± 4.8 1 ⟷ 20	1.01 ± 0.12 0.79 ⟷ 1.27	51.4 ± 2.7 46 ⟷ 57	69
4	44.8 ± 11.7 29 ⟷ 71	54.2 ± 6.6 38 ⟷ 64	160.4 ± 6.4 148 ⟷ 171	11.0 ± 10.2 0 ⟷ 44	0.93 ± 0.13 0.69 ⟷ 1.24	33.8 ± 5.1 20 ⟷ 40	37
5	51.1 ± 6.2 40 ⟷ 66	55.6 ± 6.2 43 ⟷ 71	159.6 ± 6.7 144 ⟷ 175	5.8 ± 4.8 0 ⟷ 20	0.96 ± 0.13 0.67 ⟷ 1.30	45.3 ± 4.0 40 ⟷ 56	83
6	50.9 ± 3.0 45 ⟷ 57	60.0 ± 10.5 41 ⟷ 95	159.7 ± 5.9 146 ⟷ 172	/	1.09 ± 0.11 0.88 ⟷ 1.41	/	113
7	44.6 ± 7.9 32 ⟷ 59	70.4 ± 10.5 54 ⟷ 90	160.0 ± 6.4 147 ⟷ 174	12.8 ± 8.8 1 ⟷ 31	0.99 ± 0.14 0.78 ⟷ 1.36	31.8 ± 5.6 21 ⟷ 40	16
8	53.6 ± 6.9 41 ⟷ 66	70.0 ± 8.3 52 ⟷ 88	158.0 ± 6.0 143 ⟷ 168	9.0 ± 7.1 0 ⟷ 22	0.99 ± 0.13 0.72 ⟷ 1.25	44.6 ± 3.8 40 ⟷ 58	33
TOTAL	51.9 ± 11.8	58.5 ± 9.7	159.5 ± 6.0	/	1.01 ± 0.14	/	778

TABLEAU 1 : Calcul des paramètres statistiques

m + DS
min ⟷ max

m = moyenne ; min = valeur minimum
DS = déviation standard ; max = valeur maximum

L'analyse descriptive a mis également en évidence que le phénomène de perte de masse osseuse commençait environ vers 25 ans et qu'au delà de 75 ans, la mesure du BMD n'était plus pertinente. Après élimination des femmes agées de moins de 25 ans et de plus de 70 ans, l'échantillon de 716 femmes a été retenu pour déterminer le modèle de prédiction.

AGE	-0.48			
POIDS	0.19	0.03		
TAILLE	0.23	0.23	0.35	
ANCIENNETE	-0.48	0.68	-0.1	-0.1

TABLEAU 2 : Coefficients de corrélation linéaire

2.3 - Modèle de référence

Ne disposant pas d'un plan expérimental avec répétitions pour les 716 femmes considérées, nous avons appliqué la stratégie explicitée dans le paragraphe 1.3.2 pour définir un modèle de référence.

2.3.1 - Quelle transformation de la variable âge choisir ?

Compte tenu de l'intérêt de la variable âge et des résultats de l'analyse descriptive mettant en évidence la décroissance du BMD en fonction de l'âge pour certains types, on se propose de préciser la transformation de l'âge la plus adéquate selon le type. La variable poids non corrélée avec la variable âge est conservé dans le modèle.

Pour le type 2 : pour les 256 femmes du type 2, les données ont été regroupées sous forme d'un plan d'expérience où la condition expérimentale est définie par le couple (âge, poids) avec des répétitions en nombre inégal de la variable BMD :

$$y_{ij} = \mu + a.\ f_{\alpha}(x_i^1) + b.\ x_i^2 + e_{ij} \qquad \begin{cases} i=1 \text{ à } 178 \ ; \ \sum_{i=1}^{178} n_i = 256 \\ j=1 \text{ à } n_i \\ e_{ij}\ i.i.d \sim N(0,\sigma^2) \end{cases}$$

où la transformation f_α est :

$$f_\alpha(x)=x^\alpha \text{ pour } \alpha\in\{1,0,-1,-2,-3,-4\}$$

Le choix des meilleures transformations f_α correspond aux valeurs minimum de la statistique explicitée au paragraphe 1.2 :

$$R = \mathcal{R}+trL^{-1} = \frac{1}{\hat{\sigma}^2}\|\bar{Y}-\Pi\bar{Y}\|^2+2tr\Pi L^{-1}$$

Les résultats relatifs aux différents modèles selon les valeurs de α et selon la nullité du coefficient de régression b du poids, sont donnés dans le tableau 3. Le choix des meilleurs modèles correspond aux plus petites valeurs de la statistique R : la transformation retenue correspond à une valeur α égale à -1,-2,-3 ou -4 et le modèle explicatif doit prendre en compte la variable poids.

Pour avoir un nombre plus important de répétitions, on peut regrouper la variable poids en classe d'étendue 5 kg et considérer les centres de classe pour la variable poids. On obtient une nouvelle valeur R' de la statistique de prédiction. Il est intéressant de comparer la statistique r au Cp-Mallows, explicitée dans le paragraphe 1.2 :

$$C = \mathcal{C}+n = \frac{\text{Déviance}}{\tilde{\sigma}^2} + 2s \text{ pour } \tilde{\sigma}^2 = 0.013$$

On rappelle que s correspond au nombre de variables dans le modèle.

modèle	Statistique R de prédiction			statistique	
	$\|\bar{y}-\hat{y}\|^2$	2 trL^{-1}	R	R'	C. Mallows
âge + poids	1.760	5.163	140.7	42.95	254.1
1/âge+poids	1.746	5.161	139.6	42.6	243.8
1/âge²+poids	1.742	5.160	139.2	42.6	243.5
1/âge³+poids	1.739	5.160	139.0	42.6	243.4
1/âge⁴+poids	1.738	5.161	139.0	42.7	243.4
poids	2.025	3.469	159.3	51.8	270.8
âge	1.860	3.348	146.5	45.4	252.2
1/âge	1.846	3.345	145.4	45.15	251.0
1/âge²	1.842	3.345	145.1	45.13	250.6
1/âge³	1.839	3.345	144.9	45.19	250.5
1/âge⁴	1.839	3.345	144.9	45.34	250.6
Calculs intermédiaires	Nombre de conditions expérimentales k=178 trace de la matrice L^{-1} : $tr\,L^{-1}$=147.1 estimation de la variance σ^2 : $\hat{\sigma}^2$=0.01299			k = 91 trL^{-1} = 527 $\tilde{\sigma}^2$ = 0.013	n = 256 $\hat{\sigma}^2$ = 0.013

Tableau 3 - Statistiques R,R' et C, pour le type 2.

Dans le tableau 3, les résultats comparatifs des statistiques R, R' et C conduisent aux remarques suivantes :

- le classement des modèles suivant R est légèrement différent de celui obtenu à partir de R' ou C.
- pour les 3 critères R,R' ou C, il apparait que le modèle de référence doit prendre en compte la variable poids et que la transformation de l'âge la plus adéquate correspond à α égal à -2 ou -3.

Pour les autres types

Les calculs explicités précédemment ont été de même réalisés pour les types 1,3 et 6, dans la mesure où il y avait des répétitions.

2.3.2 - Définition du modèle (M_0)

Pour la variable x^4, ancienneté de ménopause, on a pu vérifier que le pourcentage moyen de perte osseuse par an pour les premières années d'ancienneté de ménopause était plus grand que pour les suivantes.

On considère alors le modèle de régression le plus général :

$$y_{ij} = g_i + a_i f_i(x^1_{ij}) + b_i x^2_{ij} + c_i x^3_{ij} + d.(\alpha_i).x^4_{ij} + e_{ij} \qquad (1)$$

L'indice i est l'indice du type qui varie de 1 à 8, et l'indice j est relatif à l'individu, qui varie de 1 à $n_i \left(\sum_{i=1}^{8} n_i = 716 \right)$.

Le coefficient $d.(\alpha_i)$ de la variable x^4 est relatif à la perte moyenne osseuse par an sur les x_i premières années d'ancienneté de ménopause pour le type i :

$$d.(\alpha_i) = \begin{cases} d_1(\alpha_i) \text{ si } x^4_{ij} \leqslant \alpha_i \\ d_2(\alpha_i) \text{ si } x^4_{ij} > \alpha_i \end{cases}$$

Ce coefficient n'intervient pas pour les types 1 et 6, car $x^4_{1j} = x^4_{6j} = 0$.

Le paramètre inconnu α_i est un entier qui peut prendre une valeur dans l'ensemble $\{-1,0,1,\ldots,10\}$. Le cas où $\alpha_i=-1$, correspond au cas d'un pourcentage moyen de perte osseuse par an identique quelle que soit l'ancienneté de la ménopause.

Les erreurs e_{ij} sont indépendantes de loi normale centrée de variance σ^2.

L'ensemble des transformations f_i est :

$$\left\{f_\alpha(x)=x^\alpha/\alpha\in\{1,0,-1,-2,-3,-4\}\right\}$$

Compte tenu des résultats sur la variable x^4 relatifs à la comparaison de pentes de régression, la valeur $x_i=8$ a été retenue pour tous les types, afin de définir le modèle (M_0).

Le modèle de référence (M_0) est donc défini comme un cas particulier du modèle (1) pour :

$$f_i(x) = \begin{cases} 0 \text{ si } i=1 \\ x \text{ si } i=6 \\ 1/x^3 \text{ sinon} \end{cases} \qquad \text{et } \alpha_i=8 \text{ pour } i\in\{2,3,4,5,7,8\},\ 0 \text{ sinon.}$$

Ce modèle contient 33 paramètres inconnus. Il n'est pas acceptable car il a trop de paramètres et certains coefficients sont mal estimés. Néanmoins il nous donne une estimation de σ^2, qui comme on peut le vérifier dans les résultats du paragraphe suivant, est acceptable (cf. tableau 4).

Pour une déviance de 8.035 correspondant à 683 degrés de liberté, l'estimation de σ^2 est $\tilde{\sigma}^2=0.01176$.

2.4 - Recherche exploratoire des meilleurs modèles

Les modèles à explorer sont, soit des sous modèles de (M_0) obtenus en regroupant ou en éliminant certains coefficients, soit des modèles définis pour d'autres transformations f_i et des valeurs α_i autres que celles du modèle (M_0).

Pour chaque modèle, on a calculé la déviance, le carré moyen résiduel et le Cp-Mallows qui est l'estimation du risque de prédiction :

$$C_S = \frac{\text{Déviance}}{0.01176} + 2\text{card}S$$

La valeur de C_S constitue le critère majeur de choix des meilleurs modèles. Néanmoins, pour certains modèles, des tests de signification des coefficients de régression et de leurs différences ont été réalisés, afin de trouver les hypothèses sur les coefficients les plus adéquates. La prise en compte de ces hypothèses permet d'enrichir ou de réduire l'ensemble des modèles à explorer.

2.4.1 - Hypothèses sur les coefficients

Les principales hypothèses amenant les modifications intéressantes de la valeur C_8 par rapport à la valeur C_0=749 du modèle (M_0) sont :

- pour le coefficient g_i , relatif au type :
  $H_g=\{g_i=g\}$ $H_{gp_1} = \{g_1=g_6 ; g_2=g_4 ; g_3=g_7=g_8 ; g_5\}$
  $H_{gp}=\{g_1=g_6 ; g_2=g_4=g_5 ; g_3=g_8=g_7\}$

- pour le coefficient a_i, relatif à l'âge :
  $H_a = \{a_1=0 \ ; \ a_i=a \text{ sinon}\}$
  $H_x = \{a_1=0 ; a_2=a_5 ; a_3=a_4=a_7=a_8 ; a_6\}$
  $H_{x1} = \{a_1=0 ; a_2 ; a_3=a_7=a_8 ; a_4 ; a_6\}$

- pour le coefficient b_i, relatif au poids :
  $H_b = \{b_i=b\}$ $H_z = \{b_5=0 ; b_i=b \text{ sinon}\}$

- pour le coefficient C_i, relatif à la taille :
  $H_C = \{C_i=C\}$ $H_t = \{C_3=C_7=C_8=0 \ ; \ C_1=C_2=C_4=C_5=C_6\}$

- pour le coefficient $d.(\alpha_i)$, relatif à l'ancienneté de ménopause :
  $H_{d8} = \{d_1(8) = d_2(8)\}$ $H_d^{\alpha} = \{\alpha_i=\alpha\}$
  $H_d^i = \{d.(\alpha_i)=d_i\}$

2.4.2 - Analyse des résultats

Dans les tableaux 4 et 5, figurent les résultats correspondant à l'exploration de quelques modèles intermédiaires. Parmi les sous modèles de H_0 considérés, les deux meilleurs sont :

- $M_0 \cap H_{gp} \cap H_{x1} \cap H_z \cap H_t$: modèle à 12 paramètres obtenu pour la valeur C_3=721.5 (RMS=0.01165).
- $M_0 \cap H_{gp1} \cap H_{x1} \cap H_z \cap H_t$: modèle à 13 paramètres obtenu pour la valeur C_5=722 (RMS=0.01164).

Les estimations des coefficients de régression et leurs tests de signification sont données au tableau 6, pour ces 2 modèles. Une analyse des résidus permet d'accepter l'hypothèse de normalité des erreurs, pour ces 2 modèles.

Modèle M_S	deviance	d.d.ℓ.	RMS en 10^{-2}	C_S	Modèle $M_S \subset M_0$	déviance	d.d.ℓ	RMS en 10^{-2}	C_S
M_0	8.04	683	1.176	749	$H_b \cap H_c \cap H_{gp} \cap H_x$	8.31	706	1.177	727
$M_0 \cap H_g$	8.10	690	1.174	741	$H_b \cap H_{gp} \cap H_x$	8.17	699	1.169	729
$M_0 \cap H_a$	8.31	689	1.21	761	$H_c \cap H_{gp} \cap H_x$	8.21	699	1.174	732
$M_0 \cap (a_1=0)$	8.33	690	1.21	760	$H_{gp} \cap H_x \cap H_b \cap (c_1=0)$	8.39	707	1.186	731
$M_0 \cap H_b$	8.10	690	1.174	741	$H_{gp} \cap H_x \cap H_c \cap (b_1=0)$	8.46	707	1.196	737
$M_0 \cap (b_i=0)$	8.26	691	1.195	752	H_z	8.10	690	1.173	740
$M_0 \cap H_2$	8.12	690	1.177	743	H_t	8.09	690	1.172	740
$M_0 \cap c_i=0$	8.20	691	1.186	747	$H_z \cap H_t$	8.14	697	1.168	730
$M_0 \cap H_{d8}$	8.03	684	1.181	751	$H_{gp} \cap H_x \cap H_z \cap H_t$	8.28	706	1.172	724
$M_0 \cap (d(8)=0)$	8.14	689	1.188	754	$H_{gp} \cap H_z \cap H_t$	8.18	702	1.166	724
$M_0 \cap H_{gp}$	8.07	688	1.174	743	$H_{gp} \cap H_{x1} \cap H_z \cap H_t$	8.20	704	1.165	721.5
$M_0 \cap H_x$	8.07	687	1.174	744	$H_{gp1} \cap H_{x1} \cap H_z \cap H_t$	8.19	703	1.164	722
$M_0 \cap H_b \cap H_c$	8.18	697	1.173	733	H_{x1}	8.07	685	1.177	748
$M_0 \cap H_{gp} \cap H_x$	8.10	692	1.171	737	$H_{gp1} \cap H_{x1} \cap H_t \cap H_b$	8.21	703	1.167	724
$M_0 \cap H_b \cap H_c \cap H_z$	8.75	704	1.24	768	$H_{gp1} \cap H_{x1} \cap H_z \cap H_c$	8.20	703	1.167	723,5
$M_0 \cap H_b \cap H_c \cap H_{gp}$	8.23	702	1.172	728	$H_{gp1} \cap H_{x1} \cap H_b \cap H_c$	8.22	703	1.169	725
$M_0 \cap H_b \cap H_c \cap H_x$	8.20	701	1.169	727	$H_{gp} \cap H_{x1} \cap H_z \cap H_t \cap H_{d8}$	8.26	705	1.172	724

TABLEAU 4 - Déviance, RMS (carré moyen résiduel) et C_s-Mallows pour des sous-modèles de M_0.

Parmi les modèles condidérés soit pour une autre transformation f_i, soit pour une autre hypothèse sur α_i, nous avons obtenu des modèles meilleurs que les précédents uniquement pour certaines transformations précisées dans le tableau 5. Le meilleur de ces modèles, noté M^*, est obtenu pour la valeur C_s=720.5 (RMS=0.01163). L'analyse des estimations des coefficients ainsi que des résidus est satisfaisante (cf. tableau 6). D'autres sous modèles de M^* ont été explorés sans apporter d'amélioration au critère C_s.

Le choix définitif incombe au spécialiste qui choisira, parmi les meilleurs modèles acceptables sélectionnés par les

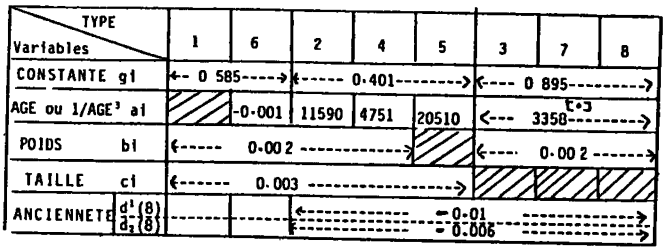
critères statistiques (valeur de C_S inférieure à 722, coefficients significatifs, normalité des résidus), celui pour lequel l'interprétation des coefficients est la plus simple.

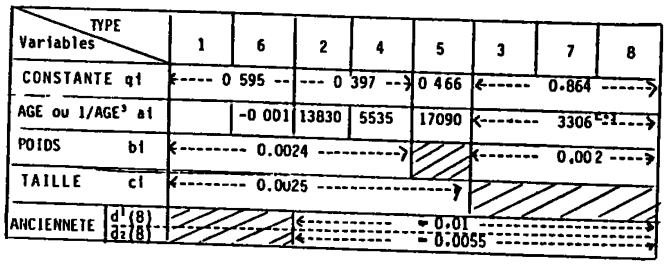
modèles $M_S^{(1)}$	déviance	d.d.l	RMS en 10^{-2}	C_S	modèles $M_S^{(2)}$	déviance	d.d.l	RMS en 10^{-2}	C_S
$H_1 \cap H_d^8$	8.20	704	1.165	721,5	$f_i(x)=\begin{cases}0 \text{ si } i=1\\ x \text{ si } i=3,4,7,8\\ 1/x^2 \text{ sinon}\end{cases}$	8.20	704	1.165	721
$H_1 \cap H_d^1$	8.24	704	1.170	724					
$H_1 \cap H_d^2$	8.26	704	1.173	726.5	$f_i(x)=\begin{cases}0 \text{ si } i=1\\ x \text{ si } i=3,6,7,8\\ 1/x^2 \text{ sinon}\end{cases}$	8.19	704	1.163	720.5*
$H_1 \cap H_d^2$	8.25	704	1.172	726					
$H_1 \cap H_d^4$	8.26	704	1.173	726	$f_i(x)=\begin{cases}0 \text{ si } i=1\\ x \text{ si } i=3,6,7,8\\ 1/x^2 \text{ sinon}\end{cases}$	8.19	704	1.164	721
$H_1 \cap H_d^5$	8.23	704	1.168	723.5					
$H_1 \cap H_d^6$	8.23	706	1.169	724	$f_i(x)=\begin{cases}0 \text{ si } i=1\\ x \text{ si } i=6\\ 1/x^2 \text{ sinon}\end{cases}$	8.19	704	1.164	721
$H_1 \cap H_d^7$	8.23	704	1.169	724					
$H_1 \cap H_d^9$	8.25	704	1.170	725	$f_i(x)=\begin{cases}0 \text{ si } i=1\\ x \text{ si } i=6\\ 1/x \text{ sinon}\end{cases}$	8.19	704	1.164	721
$H_1 \cap H_d^{10}$	8.22	704	1.167	723					
$H_1 \cap H_{d8}$	8.26	705	1.172	724.5					
H_d^1	8.23	700	1.176	732					

TABLEAU 5 - Déviance, RMS et C_S-Mallows pour les modèles suivants :

$M_S^{(1)}$ définis par $H_{gp} \cap H_{x1} \cap H_z \cap H_t$ et $f_i(x)=\begin{cases}0 \text{ si } i=1\\ x \text{ si } i=6\\ 1/x^2 \text{ sinon}\end{cases}$

$M_S^{(2)}$ définis par $H_{gp} \cap H_{x1} \cap H_z \cap H_t \cap H_d^8$ d'autres transformations.





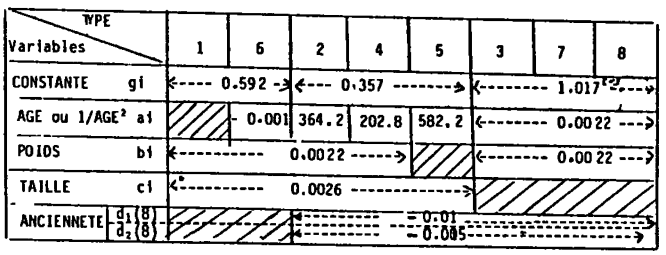
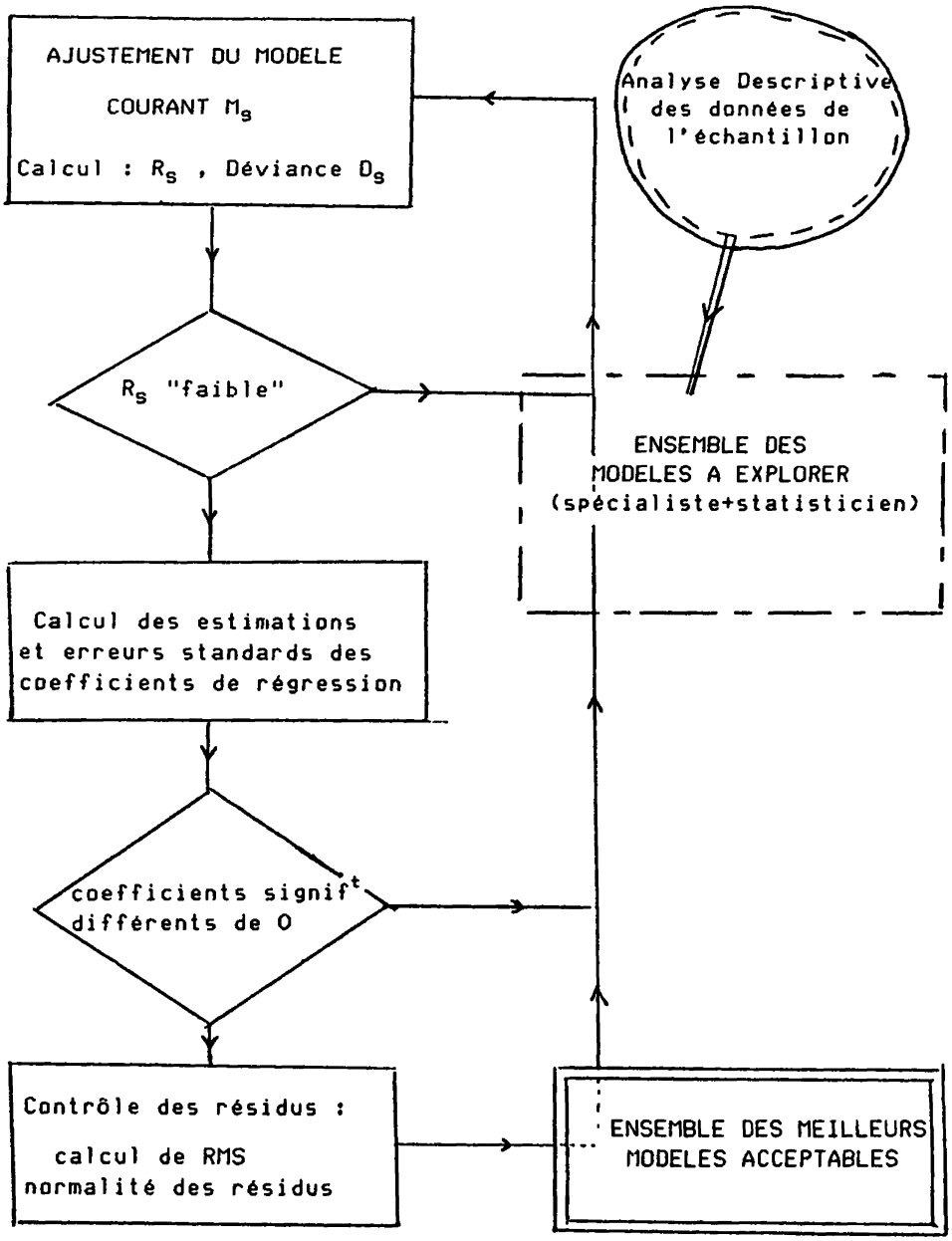


TABLEAU 6 - Estimations des paramètres pour 3 modèles acceptables.

Tous les tests de signification des coefficients sont significatifs ($P<0.01$) sauf ceux marqués du signe [.] ($P<0.1$).

CONCLUSION

Dans un problème de choix de modèles en vue de la prédiction, quand le modèle de régression n'est pas défini a priori et qu'il contient beaucoup de paramètres inconnus possibles, la démarche méthodologique basée sur l'estimation R_s du risque de prédiction peut être résumée par le schéma suivant :


AJUSTEMENT DU MODELE
COURANT M_s
Calcul : R_s , Déviance D_s
Analyse Descriptive des données de l'échantillon
R_s "faible"
ENSEMBLE DES MODELES A EXPLORER (spécialiste+statisticien)
Calcul des estimations et erreurs standards des coefficients de régression
coefficients signif^t différents de 0
Contrôle des résidus :
calcul de RMS
normalité des résidus
ENSEMBLE DES MEILLEURS MODELES ACCEPTABLES

REFERENCES BIBLIOGRAPHIQUES

[1] AKAIKE H. : "Information theory and an extension of the maximum likelihood principle". Proceedings of the 2nd International Symposium of Information theory 1973, 267-281.

[2] BAKER R.J. et NELDER J.A. : "The GLIM system Generalized linear Interactive Modelling". Rothamsted experimental station, Harpenden Hertforshire ALD ZJQ G.Britain, 1978.

[3] BONNEU M : "Choix de modèles linéaires généralisés en vue de la prédiction". Thèse de 3ème cycle, 1986 - Laboratoire de Statistique et Probabilités - Université Paul Sabatier TOULOUSE.

[4] BONNEU M. (à paraitre) : "Model choice for prediction". Statistics,1988,19,3.

[5] BUNKE O. et DROGE B. : "Estimators of the mean squared error of prediction in linear regression". Technometrics, 1984,26,145-156.

[6] MALLOWS C.L. : "Some comments on C_p". Technometrics, 1973,15,661-675.

[7] RIBOT C, TREMOLLIERES F., POULHES J.M., BONNEU M., GERMAIN F., LOUVET J.P. (à paraître) : "Obesity and Postmenopausal boneloss : the influence of overweight on vertebral density and bone turnover in post menopausal women". Bone - U.S.A..