REVUE DE STATISTIQUE APPLIQUÉE

H. C. HAMAKER

De l'analyse à régression multiple

Revue de statistique appliquée, tome 10, n° 1 (1962), p. 23-48

<http://www.numdam.org/item?id=RSA_1962__10_1_23_0>

DE L'ANALYSE A RÉGRESSION MULTIPLE (1)

H. C. HAMAKER

Laboratoires de recherche Philips – EINDHOVEN (Pays-Bas)

RESUME

Les sommes des carrés associés aux variables indépendantes dans une équation à régression multiple dépendent de l'ordre dans lequel ces variables sont introduites. Deux méthodes ont été proposées dans les écrits sur la question pour pallier cet inconvénient : la sélection progressive ou l'élimination régressive.

Avec la sélection progressive les variables indépendantes sont introduites par stades successifs. L'ordre n'est pas prédéterminé mais à chaque stade la variable prise en suivant est celle qui produit la plus forte réduction du total résiduel des carrés de la variable dépendante.

Avec l'élimination régressive par contre, nous commençons par l'équation de régression complète et en éliminons les variables indépendantes dans l'ordre dans lequel elles produisent les plus faibles augmentations du total résiduel des carrés.

Cet article décrit un plan de calcul simple et pratique qui peut s'appliquer aux deux processus. Pour la sélection progressive, nous commençons par la matrice des produits, et pour l'élimination régressive nous travaillons à partir de la matrice inverse.

De plus, ces techniques sont appliquées à divers exemples pratiques afin de démontrer les résultats auxquels elles aboutissent et les écueils que l'on peut rencontrer.

I - INTRODUCTION

Avec la régression multiple telle qu'on l'applique habituellement, les variables indépendantes sont introduites selon un ordre fixé à l'avance : X_1, X_2, X_3, ... Pour estimer l'importance relative des variables successives, nous calculons les sommes des carrés dus à X_1, X_2 après X_1, X_3 après X_1 et X_2, etc, et nous les comparons alors au carré résiduel moyen.

Cette technique a le gros inconvénient de faire dépendre les produits des carrés associés à X_1, X_2, X_3 de l'ordre dans lequel ces variables sont prises, et comme nous le verrons il peut fort bien arriver que notre opinion sur l'importance relative des variables indépendantes soit radicalement faussée

---

(1) Communication présentée au Séminaire sur les Applications Industrielles de la Statistique - Paris, 4 et 5 Septembre 1961.

lorsqu'on les interchange. Le but de cet article est la mise au point et l'application d'une méthode qui élimine cet inconvénient, au moins dans une large mesure.

Le principe de base est simple. Si par exemple nous avons quatre variables indépendantes, X_1, X_2, X_3, X_4, nous les essayons d'abord chacune en première position et vérifions laquelle d'entre elles possède la plus grosse somme de carrés. Cette variable est alors maintenue en premier dans notre équation de régression et toutes les composantes qui lui sont proportionnelles sont alors éliminées de Y et de toutes les autres variables indépendantes. Nous procédons alors de la même manière avec les résidus ; nous cherchons, parmi les variables qui restent celle qui a la plus forte somme de carrés. Puis nous continuons, en introduisant à chaque stade la variable indépendante qui, parmi celles qui restent, donne la plus grosse somme de carrés. Nous dénommerons ceci la méthode de *sélection progressive* (forward selection) des variables indépendantes par ordre *d'importance décroissante*.

Cette idée n'a rien de nouveau, et le présent article ne prétend donc pas à l'originalité. La première relation qui nous soit connue remonte à un article de Horst (1934). Quelques articles plus récents sont notés dans la liste finale de références. La plupart de ces articles ne comportent qu'un exemple numérique à l'appui de l'argumentation théorique. Cette argumentation elle-même porte sur la question de décider par des critères statistiques combien et lesquelles de variables indépendantes doivent être retenues dans l'équation de régression.

Ci-après, nous n'aborderons pas ce problème. A notre avis, du moins dans les applications industrielles, la question de la détermination des variables indépendantes à retenir et de leur nombre est un problème plus technique que statistique. Nous pouvons fort bien exclure une ou plusieurs variables indépendantes dotées de signification statistique, parce qu'elles ne contribuent que fort peu à améliorer ou à faire cadrer la variable dépendante observée avec la valeur calculée par équation de régression, et parce que nous considérons que la simplicité de cette équation est bien plus importante que la signification statistique. De même, nous pourrons préférer la variable X_3 bien qu'elle ait un peu moins de signification que X_1 ou X_2, parce que l'expérience passée nous aura appris que X_3 est la variable la plus importante techniquement.

Un autre problème qui ne retiendra pas notre attention consiste à savoir quel est le processus de calcul le plus commode. La technique de base est la réduction systématique de la matrice des produits croisés telle qu'elle est expliquée au tableau 2.1. Cette technique est la même que celle décrite au chapitre 5 du livre de Dubois (1957). En conjonction avec elle, nous avons utilisé le plan de calculs de la méthode Doolittle tel qu'il est proposé par Wishart et Metakides (1956) que nous avons trouvé à notre goût. Au fond, tous les programmes de calcul d'équations de régression multiple reposent sur les mêmes opérations de base bien que l'ordre d'exécution en soit variable. Si nous préférions la méthode de la racine carrée (Dwyer 1945) ou l'utilisation de la matrice des corrélations et du coefficient de corrélation multiple, au lieu de la matrice des sommes des produits croisés, les plans de calculs correspondants pourraient facilement être modifiés de façon à y inclure la sélection progressive des variables.

La sélection progressive n'aboutit pas nécessairement au meilleur choix. Par exemple, si nous essayons toutes les C_6^3 combinaisons possibles pouvant

être formées à partir de 6 variables indépendantes, nous pouvons trouver parmi celles-ci des combinaisons d'une plus faible somme résiduelle des carrés que le jeu des 3 résultant de la sélection progressive. Ceci risque d'arriver en particulier lorsque quelques-unes des variables indépendantes sont en étroite corrélation. Un exemple est discuté au § 7.

L'alternative est la technique *d'élimination régressive* (backward elimination) des variables indépendantes. Nous débutons avec l'équation de régression complète contenant toutes les variables indépendantes ; puis nous annulons ces variables une à une dans l'ordre où elles produisent les plus faibles augmentations de la somme résiduelle des carrés. Il se trouve que la même technique de calcul que celle de la sélection progressive peut servir, à condition de travailler avec l'inverse de la matrice des sommes des produits. Ceci ne sera pas discuté en détail mais illustré par un exemple au § 8.

Cette idée d'élimination régressive n'est pas neuve non plus. Elle nous a été suggérée pour la première fois il y a plusieurs années par notre collègue G.C. Nielen qui s'en était servi au Collège Agricole de Wageningèn. Les références écrites à cette méthode sont rares et espacées ; la méthode est mentionnée et utilisée par Canning (1959) qui nous reporte à deux textes de conférences inédits (Effroysom (1955), Bartov et Laird (1958)).

Certains de ces principes ont également été incorporés aux programmes modernes de régression des calculatrices électroniques. L'un d'eux que nous trouvons particulièrement attrayant est brièvement décrit au § 9. Une comparaison avec l'utilisation de polynomes orthogonaux dans le cas d'une seule variable indépendante équidistante est discutée au § 10.

Dans l'ensemble, nous considérons donc que la sélection d'un jeu de variables indépendantes à retenir dans une équation de régression multiple est un problème qui ne peut se résoudre avec l'aide des seules considérations statistiques. Les argumentations technique et statistique doivent s'épauler. Les principes de sélection progressive et d'élimination régressive des variables peuvent utilement servir de guides mais ne fournissent pas une solution à toute épreuve. L'expérience est indispensable et le principal propos de cet article est de fournir un ensemble pratique et illustré d'exemples faisant ressortir quelques difficultés auxquelles on peut se heurter dans la pratique.

II - TECHNIQUE DE CALCUL POUR LA SELECTION PROGRESSIVE

Evidemment, la technique de calcul est différente de la coutume usuelle. La méthode sera illustrée par un exemple emprunté à un article de Kramer (Clyde Y. Kramer : Simplified computations for multiple regression. Industrial Quality Control 13, n° 8, Fev. 57 : 8 à 11). Les sommes des produits pour son problème sont données au stade I du tableau 2.1. Ce sont les sommes de produits corrigées. Kramer ne donne pas les sommes des produits brutes. On peut aussi bien partir des sommes brutes comme il sera démontré au § 3.

Dans son article, Kramer utilise des variables transformées $x_i = 10^{mi} X_i$, les exposants m_i ayant été déterminés, selon une méthode proposée par Duncan et Kenney, de façon que toutes les sommes de carrés soient comprises entre 0.1 et 10. On peut alors garder le même nombre de décimales tout au long des calculs. Kramer utilise 7 décimales et nous n'en avons conservé que 4, ce qui suffit à notre but actuel.

Dans tout cet article nous utiliserons les symboles X_1, X_2, ... Y pour les variables originales et x_1, x_2, ... y pour les variables codées.

De plus nous introduisons la variable supplémentaire $x_0 = 1$ pour tenir compte du terme constant de l'équation de régression. Au tableau 2.1 nous commençons par $x_{1.0}$, $x_{2.0}$, ..., $y_{.0}$, ce qui signifie que nous utilisons des variables codées et qu'à partir de chacune le terme constant correspondant à x_0 (la moyenne) a déjà été compensé.

Si nous considérons maintenant les équations de régression

$$y = b_{1.0}\, x_{1.0} \; ; \; y = b_{2.0}\, x_{2.0}, \quad \text{etc.}$$

la somme des carrés due à la régression est :

pour $x_{1.0}$: $$SS_{1.0} = \frac{(\Sigma x_{1.0}\, y_{.0})^2}{\Sigma x_{1.0}^2} = \frac{(-0,5266)^2}{0,8239} = 0,366$$

pour $x_{2.0}$: $$SS_{2.0} = \frac{(\Sigma x_{1.0}\, y)^2}{\Sigma x_{2.0}^2} = \frac{(0,7405)^2}{1,7021} = 0,3225,$$

et ainsi de suite. Ces sommes de carrés ont été inscrites à la dernière colonne du tableau 2.1 Stade I. Il en ressort que c'est $x_{3.0}$ qui a la plus grosse somme de carrés, et pour cette raison, après la constante x_0, nous prendrons x_3 comme variable indépendante suivante dans notre équation de régression.

Nous éliminons donc $x_{3.0}$ de $y_{.0}$ et des variables indépendantes c'est-à-dire que nous remplaçons

$$y_{.0} \text{ par } y_{.03} = y_{.0} - b_{y3.0}\, x_{3.0} \quad ,$$

$$x_{1.0} \text{ par } x_{1.03} = x_{1.0} - b_{13.0}\, x_{3.0} \quad ,$$

etc, ou $b_{y3.0}$, $b_{13.0}$, ... sont les coefficients de régression obtenus par la méthode des moindres carrés. Nous avons alors :

$$\Sigma y_{.03}^2 = \Sigma y_{.0}^2 - \frac{(\Sigma x_{3.0}\, y_{.0})^2}{\Sigma x_{3.0}^2} \quad ,$$

$$\Sigma x_{1.03}^2 = \Sigma x_{1.0}^2 - \frac{(\Sigma x_{3.0}\, x_{1.0})^2}{\Sigma x_{3.0}^2} \quad ,$$

$$\Sigma x_{1.03}\, x_{2.03} = \Sigma x_{1.0}\, x_{2.0} - \frac{\Sigma x_{3.0}\, x_{1.0}\, \Sigma x_{3.0}\, x_{2.0}}{\Sigma x_{3.0}^2} \quad ,$$

etc, si bien que les sommes des produits après élimination de $x_{3.0}$ peuvent être calculées très simplement. A cette fin nous complétons la colonne et la rangée pour $x_{3.0}$ au Stade I et nous obtenons

$$\Sigma x_{1.03}^2 = 0,8239 - \frac{(-0,8636)^2}{2,2139} = 0,4870,$$

$$\Sigma x_{1.03}\, x_{2.03} = -1,1310 - \frac{-0,8636 \times 1,3775}{2,2139} = -0,5937,$$

Tableau 2.1

Sélection progressive des variables indépendantes ;
données d'après C.Y. Kramer (1957)

	$x_{1.0}$	$x_{2.0}$	$x_{3.0}$	$x_{4.0}$	$y_{.0}$	Controle	ΔS_{yy}	
$x_{1.0}$	0,8239	-1,1310	-0,8636	+0,0607	-0,5266	-1,6366	0,3366	Stade I
$x_{2.0}$		1,7021	1,4775	-0,1499	0,7405	2,5392	0,3225	Sommes des produits originaux
→ $x_{3.0}$	-0,8636	1,3775	2,2139	0,0986	0,9112	3,7376	0,3750	56 degrés de liberté
$s_{4.0}$			0,0986	0,8691	0,0462	0,9247	0,0216	
$y_{.0}$			0,9112		0,5098	1,6811		

	$x_{1.03}$	$x_{2.03}$	$x_{4.03}$	$y_{.03}$	Controle	ΔS_{yy}	
→ $x_{1.03}$	0,4870	-0,5937	+0,0992	-0,1712	-0,1787	0,0602	Stade II
$x_{2.03}$	-0,5937	0,8450	-0,2112	+0,1735	0,2136	0,0356	Sommes des produits après élimination de x_3
$x_{4.03}$	+0,0992		0,8647	+0,0056	0,7583	0,0000	
$y_{.03}$	-0,1712			0,1348	0,1427		

	$x_{2.031}$	$x_{4.031}$	$y_{.031}$	Controle	ΔS_{yy}	
→ $x_{2.031}$	0,1212	-0,0903	-0,0352	-0,0043	0,0102	Stade III
$x_{4.031}$	-0,0903	0,8445	0,0405	0,7947	0,0000	Sommes des produits après élimination de x_3, x_1
$y_{.031}$	-0,0352		0,0746	0,0799		

	$x_{4.0312}$	$y_{.0312}$	Controle	ΔS_{yy}	
→ $x_{4.0312}$	0,7772	0,0143	0,7915	0,0003	Stade IV
$y_{.0312}$		0,0644	0,0787		Sommes des produits après élimination de x_3, x_1, x_2

	$y_{.03124}$	
$y_{.03124}$	0,0641	0,0641

Stade V
Résidus après élimination de x_3, x_1, x_2, x_4

Carré moyen résiduel
0,0641/52 = 0,0012

etc. Ces nouvelles sommes de produits constituent le stade II du tableau 2.1 ; les calculs sont de nature simple et peuvent être achevés en une opération sur une machine à calculer électrique de bureau.

Nous pouvons maintenant nous occuper du second stade exactement de la même manière. La somme de carrés correspondant à x_1 d'après x_3 est donnée par

$$SS_{1.03} = \frac{(\Sigma x_{1.03}\, y_{.03})^2}{\Sigma x_{1.03}^2} = \frac{(-0,1712)^2}{0,4870} = 0,0602.$$

Cette valeur ainsi que $SS_{2.03}$ et $SS_{4.03}$ est inscrite à la dernière colonne du Stade II. Nous voyons qu'en seconde position x a la plus forte somme de carrés et nous entreprenons d'éliminer cette variable en passant du Stade II au Stade III exactement de la même manière que pour passer de I à II.

Le contrôle des calculs est effectué d'une façon bien connue. Les données inscrites dans la colonne "Contrôle" sont les sommes des rangées dans la matrice complète des sommes des produits. Ainsi dans le stade II par exemple on a

$$-0.5937 + 0.8450 - 0.2112 + 0.1735 = 0.2136,$$
$$0.0992 - 0.2112 + 0.8647 + 0.0056 = 0.7583,$$

et ainsi de suite.

Si, en allant d'un stade au suivant, cette colonne "contrôle" est réduite par les mêmes opérations que les autres colonnes du tableau, des contrôles réduits doivent être égaux aux sommes de rangées de la matrice réduite.

Nous constatons ainsi qu'à chaque stade nous obtenons la somme de carrés maximale en introduisant les variables indépendantes dans l'ordre x_3, x_1, x_2, x_4. Le tableau 2.2 contient une vue d'ensemble des sommes de carrés associées à la variable à chaque stade de calcul et présentées dans l'ordre donné.

Tableau 2.2

Somme des carrés pour x_3, x_1, x_2 et x_4 au stade successif

Stade	I	II	III	IV
	Somme des carrés			
x_3	0,3750			
x_1	0,3366	0,0602		
x_2	0,3225	0,0356	0,0102	
x_4	0,0216	0,0000	0,0003	0,0003
Carré moyen résiduel : 0,0012 ; 52 d. de l.				

Un tableau de ce genre contient des renseignements révélateurs. Nous constatons par exemple qu'au Stade I, x_3, x_1 et x_2 ont tous trois des sommes de carrés assez élevées, mais que celles de x_1 et x_2 sont considérablement

réduites au Stade II. Ceci ne peut être dû qu'au fait que x_1 et x_2 sont tous deux en corrélation étroite avec x_3 ; une bonne part de la variabilité de y peut s'expliquer par chacune de ces variables.

La somme totale des carrés pour les quatre variables indépendantes est :

$$0,3750 + 0,0602 + 0,0102 + 0,0003 = 0,4457,$$

dont le montant

$$0,3750 + 0,0602 = 0,4352$$

peut être impliqué à x_3 et x_1 seuls. Evidemment, x_4 est sans importance et peut être laissé de côté. On peut se demander si x_2 devrait faire partie ou non d'une équation de régression.

Comparée aux carrés moyens résiduels, la plus petite somme de carrés associée à x_2 au tableau 2.2 (0,0102) a une signification statistique au niveau de 1 %. Ceci peut nous autoriser à conserver x_2 comme variable essentielle de l'équation de régression.

Par contre, si nous incluons x_2, l'écart-type résiduel est de 0,035, et si nous le laissons de côté elle est de 0,037. D'un point de vue technique le gain de précision très faible ne peut pas justifier les complications qu'entraîne un terme supplémentaire dans l'équation de régression. Tout dépend du but fixé pour cette équation. A notre avis une question comme celle-ci doit être tranchée en fonction de considérations techniques plutôt que statistiques. Kramer ne donne aucun renseignement sur l'aspect technique dont dépendent ses données.

Evidemment, les calculs exécutés au tableau 2.3 sont très voisins de ceux de la méthode progressive Doolittle habituelle. C'est afin de le démontrer que nous représentons au tableau III les résultats de cette méthode selon le plan de calculs proposé par Wishart et Metadikes (Biometrika 1953) que nous avons trouvé extrêmement pratique. Nous avons légèrement modifié leur programme en ce sens que, en accord avec ce que propose Kramer, y est compris dans les calculs comme la dernière variable indépendante.

Dans ce programme, le plan est décomposé en deux moitiés, la partie droite contenant la technique progressive Doolittle et celle de gauche le calcul des coefficients de régression.

Les calculs progressifs Doolittle sont reportés in extenso : sous les sommes de produits inscrites dans les rangées 1, 3, 7, 12 et 18 nous trouvons les termes de correction successifs, la solution progressive Doolittle étant obtenue dans les rangées 1, 2, ; 5, 6 ; 10, 11 ; 16, 17 ; 23. Quant aux données des rangées 1, 5, 10, 16 et 23, elles peuvent être relevées immédiatement dans les rangées indiquées par des flèches au tableau 2.1 si nous les reclassons dans l'ordre x_3, x_1, x_2, x_4. Ainsi, au tableau 2.1 nous avons déjà obtenu la solution progressive. La différence est que les corrections successives apportées au tableau 2.3 aux sommes de produits conjointement ne sont introduites qu'une par une au tableau 2.1. Le léger surcroît de travail ainsi causé représente le prix de l'avantage qui nous est offert de pouvoir choisir chemin faisant l'ordre des variables indépendantes.

Le calcul des coefficients de régression est parallèle à celui de la solution progressive Doolittle, et si nous combinons les deux sur une même page

Tableau 2.3

Exemple de Kramer : Plan de calcul de Wishart et Mitakides

Rang	$x_{3.0}$	$x_{1.0}$	$x_{2.0}$	$x_{4.0}$	$y_{.0}$	Contrôle
1	+2,2139	-0,8636	+1,3775	+0,0986	+0,9112	+3,7376
2	-1	+0,3901	-0,6222	-0,0445	-0,4116	-1,6882
3		+0,8239	-1,1310	+0,0607	-0,5266	-1,6366
4	+0,3901	-0,3369	+0,5373	+0,0385	+0,3554	+1,4580
5	+0,3901	+0,4870	-0,5937	+0,0992	-0,1712	-0,1786
6		-1	+1,2191	-0,2037	+0,3515	+0,3669
7			+1,7021	-0,1499	+0,7405	+2,5392
8	-0,6222		-0,8571	-0,0613	-0,5670	-2,3255
9	+0,4756	+1,2191	-0,7238	+0,1209	-0,2087	-0,2179
10	-0,1466	+1,2191	+0,1212	-0,0903	-0,0352	-0,0042
11			-1	+0,7450	+0,2904	+0,0347
12				-0,8691	+0,0462	+0,9247
13	-0,0445			-0,0044	-0,0406	-0,1665
14	-0,0795	-0,2037		-0,0202	+0,0348	+0,0364
15	-0,1091	+0,9083	0,7450	-0,0673	-0,0262	-0,0031
16	-0,2332	+0,7046	0,7450	+0,772	+0,0143	+0,7915
17				-1	-0,0184	-1,0184
18					+0,5098	+1,6811
19	-0,4116				-0,3750	-1,5383
20	+0,1371	+0,3515			-0,0602	-0,0629
21	-0,0426	+0,3541	+0,2904		-0,0102	-0,0012
22	+0,0043	-0,0130	-0,0137	-0,0184	-0,0003	-0,0146
23	-0,3128	+0,6926	+0,2767	-0,0184	+0,0641	+0,0641

comme au tableau 2.2 les calculs s'en trouvent simplifiés. Pour les détails, le lecteur est prié de se reporter à l'article de Wishart et Metakides.

Les coefficients de régression sont donnés à la rangée du bas, les signes étant inversés. Ainsi nous avons :

$$y_{.0} = 0,3128\ x_{3.0} - 0,6926\ x_{1.0} - 0,2767\ x_{2.0} + 0,0184\ x_{4.0} \qquad (2.1)$$

si nous incluons les quatre variables.

Si nous voulions laisser tomber x_4 parce qu'il est sans importance, nous n'avons qu'à annuler les termes de correction de la rangée 22 pour obtenir

$$y_{.0} = 0,3171\ x_{3.0} - 0,7056\ x_{1.0} - 0,2904\ x_{2.0}, \qquad (2.2)$$

où

$$0,3171 = -(-0,4116 + 0,1371 - 0,0426), \text{ etc.}$$

De même nous avons

$$y_{.0} = 0,2745\ x_{3.0} - 0,3515\ x_{1.0} \qquad (2.3)$$

si nous n'incluons que ces deux variables.

En plus du plan du tableau 2 Wishart et Metakides donnent aussi un programme simple pour le calcul de la matrice de covariance des coefficients de régression. Nous n'en parlerons pas ici.

III - UN AUTRE EXEMPLE : L'ESTIMATION DU RENDEMENT EN ESSENCE DES PETROLES BRUTS -

On peut se demander si cela représente un réel avantage d'introduire les variables indépendantes par sélection progressive. Il ne saurait en être décidé sur la base de l'unique exemple traité ci-dessus. Nous prendrons donc en considération dans ce paragraphe et les suivants quelques autres exemples empruntés à des communications pour illustrer la conclusion à laquelle la méthode peut conduire. Ce faisant, nous ne reproduirons que la matrice des sommes de produits dont nous sommes partie et les conclusions auxquelles nous sommes amenés, en laissant de côté les calculs numériques intermédiaires.

Nous commençons par un problème qui a servi à Hader et Grandage en 1956 pour illustrer la régression multiple.

Le tableau 3.1 A donne des renseignements sur les variables mises en jeu et le tableau 3.1 B sur la matrice de base de sommes de produits pour les variables codées (voir tableaux 3.1 A et B).

Dans le problème de Kramer nous partions des sommes de produits corrigées. Pourtant, cela n'est pas nécessaire et dans le présent exemple nous partons des sommes brutes, en incluant le terme constant dans l'équation de régression par une variable $X_0 = 1$.

L'ordre d'importance décroissante est :

$$x_4,\ x_3,\ x_1,\ x_2,\ x_0$$

et le tableau 3.2 donne les sommes des carrés pour les stades successifs .

Un trait remarquable de cet exemple est que dans l'ordre d'importance le terme constant x_0 vient en dernier et est alors parfaitement négligeable. Habituellement le besoin d'un terme constant dans l'équation de régression n'est jamais mis en question, comme si c'était une condition à priori. Par exemple, dans leur analyse de variance, Hader et Grandage ne donnent que les sommes des carrés de x_1 à x_4 et ils laissent x_0 complètement de côté

Tableau 3.1

Données de base pour le problème du pétrole brut

A - Variables et codage

Variables indépendantes	Variables codées
Constante = $X_0 = 1$	$x_0 = 10^{-1} X_0$
Densité pétrole brut = X_1	$x_1 = 10^{-2} X_1$
Pression vapeut pétrole brut = X_2	$x_2 = 10^{-1} X_2$
Point 10 % du pétrole brut (ASTM) = X_3	$x_3 = 10^{-3} X_3$
Point final essence = X_4	$x_4 = 10^{-3} X_4$
Variable dépendante	
% essence extraite = Y	$y = 10^{-2} Y$

B - Sommes de produits pour les variables codées

	x_0	x_1	x_2	x_3	x_4	y
x_0	0,320000	1,256000	1,338000	0,772800	1,063700	0,629100
x_1		5,028250	5,535680	2,987321	4,131897	2,515359
x_3			7,722200	2,954970	4,274600	2,964880
x_4				1,910002	2,599887	1,479966
y		n = 32			3,680003	2,254180
						1,593179

Tableau 3.2

Sommes des carrés des stades successifs
pour le problème du tableau 3.1

Stade	I	II	III	IV	V
	Somme des carrés				
x_4	1,38079				
x_3	1,14675	0,17316			
x_1	1,25830	0,00063	0,02446		
x_2	1,13834	0,04353	0,02251	0,00106	
x_0	1,23677	0,03642	0,02216	0,00016	0,00023
Carré moyen résiduel : 0,00050 ; 27 d. de l.					

En fait, nous constatons qu'un terme constant n'est pas nécessaire. Les variations de y peuvent très bien s'expliquer par un plan en x_1, x_2, x_3, x_4 passant par l'origine. Ceci peut constituer une caractéristique importante qui apporterait un soutien considérable à l'équation obtenue ou à son explication. Un des traits intéressants de notre méthode d'analyse est sans conteste, de révéler qu'un terme constant est superflu dans l'équation de régression.

On peut aussi conclure que dans notre équation de régression finale nous devrions inclure x_4, x_3 et soit x_0, x_1 ou x_2. Au troisième stade x_1 possède la plus forte somme de carrés mais celles de x_0 et x_2 n'en sont pas loin. Nous devrions inclure une de ces trois variables et abandonner les deux autres ; ensemble, elles correspondront à une somme de carrés de 0,003 au plus et ceci est sans signification statistique et sans importance technique.

Pour l'équation de régression sur x_4, x_3 et x_1 nous avons obtenu :

$$y = 1,572\ x_4 - 1,793\ x_3 + 0,274\ x_1.$$

IV - UN PROBLEME DE FONDERIE -

Ce problème, dû à Moriceau (1954), consistait à estimer le temps nécessaire à la production d'une pièce de fonderie, d'après des données concernant ses dimensions, sa forme et la complexité de sa fabrication.

D'après enquête préliminaire, il a été déduit qu'une description satisfaisante de la relation recherchée serait exprimée par une équation de la forme $T = A^a\ B^b\ C^c\ \ldots$, où T = temps, A, B, C, ... = caractéristiques numériques de fonderie et a, b, c, ... = paramètres inconnus. Il en découlait un problème de régression linéaire entre log T et les logarithmes des caractéristiques de fonderie. Finalement 6 variables indépendantes ont été incluses dans l'étude selon la liste du tableau 4.1A ; le tableau 4.1B donne la matrice des sommes de produits corrigées(1).

Pour le problème de Moriceau, l'ordre d'importance décroissante et la somme des carrés aux stades successifs sont présentés au tableau 4.1.

Moriceau ne donne que les sommes corrigées des produits ce qui revient à dire que le terme constant X_0 a déjà été éliminé. Nous ne pouvons donc pas l'inclure dans notre recherche.

La contribution de $x_{6.0}$ et de $x_{3.0}$ est nettement sans importance et ces deux variables peuvent être négligées. Moriceau a abouti à la même conclusion par une méthode bien plus recherchée. En fait, il a pris les 6 variables indépendantes une à une, en paires, par jeux de trois, etc., il a calculé les 63 équations de régressions concevables ainsi créées et choisi finalement celle qui donnait le plus petit carré moyen résiduel. Nous sommes satisfaits de voir que la technique de sélection progressive aboutit directement à la même réponse.

---

(1) La matrice telle qu'elle est donnée à la page 71 de l'article de Moriceau comporte un certain nombre d'erreurs typographiques. Par exemple la somme de produits $X_1\ X_2$ qui devrait être -6,574113 est inscrite comme -.574113, et la somme X_3 qui devrait être 24,831676 est notée comme 94,831676. Heureusement, Moriceau donne également les colonnes de contrôle, et certains articles de la matrice principale se retrouvent dans d'autres calculs portés à la même page. Ceci nous a permis d'effectuer les corrections nécessaires.

Tableau 4.1

Données de base pour le problème de Moriceau concernant le temps nécessaire à la fabrication d'une pièce de fonderie

A - Variables et codage

Variables indépendantes	
X_1 : log. du poids de la pièce	$x_1 = 10^{-1} X_1$
X_2 : log (vol. du moule/poids de la pièce)	$x_2 = X_2$
X_3 : log (nombre de levées/poids de la pièce)	$x_3 = 10^{-1} X_3$
X_4 : log (nombre d'attaques/poids de la pièce)	$x_4 = 10^{-1} X_4$
X_5 : log (indice de disposition du moule/poids de la pièce)	$x_5 = 10^{-1} X_5$
X_6 : log (indice de modèle/poids de la pièce)	$x_6 = 10^{-1} X_6$
Variable dépendante	
Y : log. du temps de fabrication de la pièce	$y = Y$

B - Sommes de produits corrigées pour les variables codées

	$x_{1.0}$	$x_{2.0}$	$x_{3.0}$	$x_{4.0}$	$x_{5.0}$	$x_{6.0}$	$y_{.0}$
$x_{1.0}$	0,349101	-0,657411	-0,284878	-0,236666	0,053903	-0,333877	-1,310196
$x_{2.0}$		5.391991	0,510139	0,516592	-0,379458	0,589493	3,247153
$x_{3.0}$			0,248317	0,194005	-0,010429	0,273543	1,077847
$x_{4.0}$				0,188029	-0,046791	0,222650	0,846185
$x_{5.0}$					1.051746	-0,024542	0,144762
$x_{6.0}$		n = 90				0,362518	1,234280
$y_{.0}$							6,493102

Tableau 4.2

Ordre d'importance décroissante et somme des carrés pour le problème de fonderie de Moriceau

Stade	I	II	III	IV	V	VI
Somme des carrés						
$x_{1.0}$	4,9172					
$x_{2.0}$	1,9555	0,1464				
$x_{5.0}$	0,0299	0,1154	0,1555			
$x_{4.0}$	3,8081	0,0640	0,1161	0,1074		
$x_{6.0}$	4,2024	0,0082	0,0030	0,0104	0,0171	
$x_{3.0}$	4,6785	0,0048	0,0118	0,0001	0,0012	0,0013
Carré moyen résiduel : 0,0138 ; 83 d. de l.						

Comme dans l'exemple de Kramer, on voit que $x_{4.0}$, $x_{6.0}$ et $x_{3.0}$ ont toutes, au premier stade, des sommes de carrés élevées qui sont réduites à de faibles résidus quand on a tenu compte de $x_{1.0}$. Ces trois variables doivent être toutes en étroite corrélation avec x_1.

Le cas de $x_{5.0}$ est également intéressant. La somme de ses carrés part d'une valeur faible mais augmente régulièrement. Au stade II, $x_{5.0}$ est encore moins important que $x_{2.0}$, mais au Stade III la somme des carrés pour $x_{5.0}$ en arrive à dépasser celle de $x_{2.0}$ au stade précédent. Ce résultat n'est pas contradictoire, il signifie simplement que $x_{5.0}$ est particulièrement efficace lorsqu'on l'introduit en conjonction avec $x_{1.0}$ et $x_{2.0}$. Nous reviendrons sur ce point au § 6.

A nouveau se pose la question de savoir quelle équation de régression nous devrions finalement adopter. $x_{1.0}$ est responsable tout seul d'une somme de carrés de 4,9172 ; ensuite $x_{2.0}$, $x_{5.0}$, et $x_{4.0}$ sont tous à peu près de même importance, mais ensemble ils justifient une somme additionnelle de carrés ne dépassant pas 0,4903. L'écart-type résiduel est de 0,14 si nous utilisons $x_{1.0}$ seul et de 0,12 si nous employons une équation avec $x_{1.0}$, $x_{2.0}$, $x_{3.0}$, et $x_{4.0}$. Il se pourrait bien qu'à des fins techniques une équation basée sur $x_{1.0}$ tout seul soit suffisamment exacte. Nous pouvons prédire le temps nécessaire à la fabrication d'une pièce de fonderie avec assez de précision d'après le poids de la pièce ; les autres indices ne contiennent qu'une petite proportion de renseignements complémentaires.

V - UN PROBLEME DE FABRICATION CHIMIQUE -

Considérons maintenant quelques données résultant d'une expérience sur un catalyseur empruntée à De Baun et Schneider (1957).

Comme l'article n'est pas facile à se procurer, les données originales sont présentées au tableau 5.1 ainsi que la matrice des sommes de produits brutes y compris les termes quadratiques des variables indépendantes.

Le plan de l'expérience n'était pas pleinement équilibré, et de ce fait les variables indépendantes ne sont pas orthogonales. Deux variables dépendantes ont été observées : la durée de vie et le degré d'activité du catalyseur. L'un des problèmes était de faire cadrer une surface quadratique.

Si nous considérons Y_1 et Y_2 séparément et que nous introduisons les variables par sélection progressive, nous obtenons les résultats du tableau 5.2. Nous n'avons pas donné toutes les sommes de carrés pour les stades successifs, mais seulement les plus fortes valeurs correspondant aux variables portées en haut des colonnes successives des tableaux 2.2, 3.2 et 4.2.

Dans les deux cas ces carrés moyens indiquent qu'au moins deux et peut-être trois des variables indépendantes peuvent être écartées ; mais ce ne sont pas les mêmes dans les deux cas. Y_1 peut très bien être décrit par une équation en X_2, X_1^2 et X_2^2, alors que pour Y_2 une équation en X_1X_2, X_2 et peut-être X_1^2 suffirait. Si nous désirions retenir les mêmes variables pour Y_1 que pour Y_2, X_1 semblerait être le seul auquel on puisse renoncer sans risque.

D'habitude, dans un problème comme celui-ci, on essaie les termes linéaires et les termes quadratiques séparément en jeu. Nous nous demandons si cette politique est vraiment avisée. Par exemple dans le cas de Y_2, X_2^2 est nettement sans importance et X_1^2 n'a pas une signification évidente. Si nous essayons X_1^2, X_2^2 et X_1X_2 comme un seul jeu avec trois degrés de

Tableau 5.1

Données originales et matrice des sommes de produits brutes pour le problème de catalyseur de De Baun et Schneider

A - Données originales

X_1	X_2	Y_1	Y_2	
2	0	17,8	67,9	
0	2	10,6	63,1	X_1 : temps
2	2	3,5	20,6	X_2 : p
0	0	27,9	79,8	
0	0	32,8	76,4	Y_1 : durée de vie d'un catalyseur
0	-1	29,6	88,1	Y_2 : degré d'activité d'un catalyseur
-1	0	26,7	68,7	
-1	-1	29,8	79,3	Toutes données codées
1	-1	29,5	76,5	
1	0	31,5	68,3	

B - Matrice des sommes de produits brutes

	X_0	X_1	X_2	X_1^2	X_2^2	$X_1 X_2$	Y_1	Y_2
X_0	10	4	1	12	11	4	239,7	688,7
X_1		12	4	16	8	6	47,1	173,8
X_{2_2}			11	6	13	8	-60,7	-76,5
X_{1_2}				36	18	16	202,7	646,8
X_2					35	16	145,3	578,7
X_1X_2						18	14,3	85,2
Y_1, Y_2		n = 10					6635,29	50508,31

liberté, ceci peut nous induire à ne conserver les deux premiers termes que parce que le dernier a une forte signification. Cela compliquerait inutilement nos équations. De même, l'essai de X_1 et de X_2 ensemble nous obligerait à conserver X_1 qui, au tableau 5.2, est sans importance tant pour Y_1 que pour Y_2.

VI - LA NECESSITE DE CALCULS COMPLETS

Si nous introduisons les variables indépendantes par sélection progressive, nous pouvons être enclins à arrêter, disons au troisième stade, parce qu'aucune des variables qui restent ne semble plus provoquer de réduction de la somme des carrés digne de considération. Ceci peut toutefois être trompeur ; il peut toujours arriver que des variables paraissant sans importance à un stade deviennent très importantes à un stade suivant. Il peut être utile d'illustrer ce fait par un exemple extrême représenté au tableau 6.1.

Tableau 5.2

Ordre d'importance décroissante et carrés moyens pour le problème du tableau 5.1

Y_1			Y_2		
	Carré moyen	d.d.l.		Carré moyen	d.d.l.
X_0	5745	1	X_0	47430	1
X_2	656	1	X_1X_2	2205	1
X_1^2	117	1	X_2	443	1
X_2^2	73	1	X_1^2	199	1
X_1X_2	5	1	X_1	13	1
X_1	4	1	X_2^2	1	1
résiduel	9	4	résiduel	54	4

Nous constaterons, d'après le tableau 6.1B qu'en seconde position tant x_1 que x_2 donnent des sommes de carrés faibles alors qu'en troisième position les deux sommes des carrés sont beaucoup plus élevées ; ces deux variables ensemble sont responsables dans une large mesure de la variabilité de y que chacune ne pourrait expliquer séparément.

L'exemple se rapporte au jeu de points tracé à la figure 6.1, X_2 étant égal à X_1^2. Si nous prenons X_1 seul, nous essayons de tracer une ligne droite, et si nous prenons $X_2 = X_1^2$ seul nous essayons de tracer une parabole ayant son sommet sur l'axe Y. Il devient de suite évident qu'en aucun des deux cas on n'aboutira à une équation satisfaisante. Mais en employant à la fois

Tableau 6.1

Un exemple typique

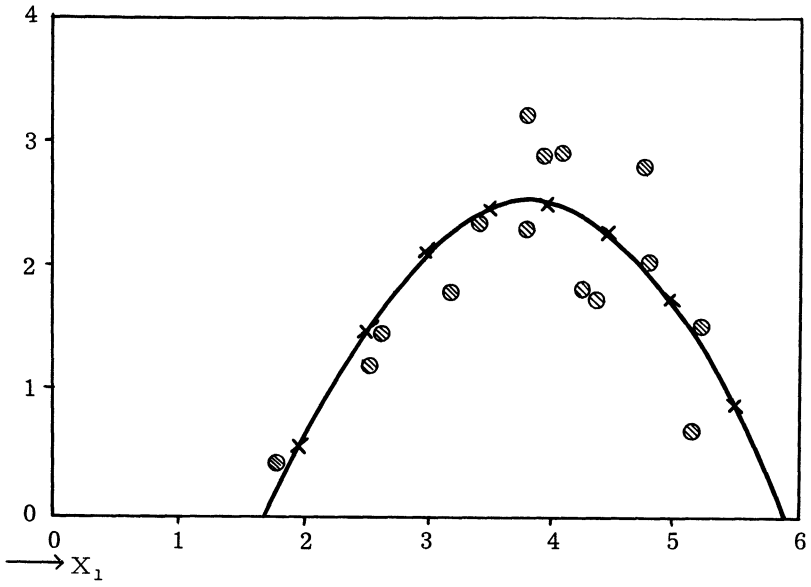
A - Sommes de produits brutes pour les données codées

	x_0	x_1	x_2	y
x_0	0,16000	0,62900	0,26275	0,31100
x_1		2,62750	1,14586	1,25630
x_2			0,51550	0,52596
y				0,69910
$x_0 = 10^{-1} X_0$, $x_1 = 10^{-1} X_1$, $x_2 = 10^{-2} X_1^2$, $y = 10^{-1} Y$				

B - Sommes des carrés pour x_1 et x_2 après élimination de x_0

	Somme des carrés		Somme des carrés
$x_{1.0}$	0,00733	$x_{2.0}$	0,00276
$x_{2.01}$	0,05429	$x_{1.02}$	0,05886
	Carré moyen résiduel : 0,00253 ; 13 d.d.l.		

Figure 6.1 - Illustration de l'exemple du tableau 6.1.



X_1 et $X_2 = X_1^2$ nous pouvons obtenir une parabole comme celle de la figure et qui cadre bien.

Evidemment, la situation illustrée par le tableau 6.1 et la figure 6.1 est un cas limite que l'on ne rencontrera pas souvent en pratique. Mais il est là pour nous inciter à la prudence.

Sur la base du tableau 3.2 par exemple, nous pourrions être enclin à mettre un terme à notre analyse au stade IV parce que les carrés moyens pour x_2 et x_0 sont sans portée à ce stade. Il est toutefois concevable qu'après avoir éliminé x_2 au stade IV nous trouvions soudain une signification à x_0 au stade V, puisque x_2 et x_0 ensemble en ont une. Il est donc toujours à conseiller de poursuivre l'analyse jusqu'à ce que l'on ait tenu compte de toutes les variables indépendantes.

Les résultats que nous venons de discuter expliquent également le comportement des sommes des carrés pour x_5 au tableau 4.2, sur lequel nous avons déjà attiré l'attention.

VII - UN EXEMPLE DANS LEQUEL LA SELECTION PROGRESSIVE N'ABOUTIT PAS A UN JEU OPTIMAL.

Il nous est fourni par Hald (1952). Nous n'en reproduirons pas les données en détail puisque son livre se trouve partout.

Son exemple se rapporte à la chaleur mise en jeu lors du durcissement de ciment comme fonction de sa composition. Les quatre variables indépendantes sont :

x_1 : pourcentage en poids de 3 $CaO.Al_2O_3$,
x_2 : " " 3 $CaO.Si\ O_2$,

x_3 : pourcentage en poids de 4 $CaO.Al_2O_3.Fe_2O_3$,

x_4 : " " 2 $CaO.Si\ O_2$.

Hald a calculé toutes les équations de régression possibles avec une, deux, trois et quatre de ces variables indépendantes. On constate d'après ses données que la sélection progressive introduirait successivement x_0, x_4 et x_1 en tant que trois premières variables avec une somme résiduelle de carrés $S_{yy.014} = 74,76$, alors que les variables x_0, x_1 et x_2 donnent une somme résiduelle de carrés de $S_{yy.012} = 57,91$.

Il faut sans doute s'attendre à ce que des difficultés de cet ordre surviennent lorsque les variables indépendantes sont en étroite corrélation. Du moins c'était le cas dans l'exemple de Hald. Les quatre variables sont les pourcentages en poids de la composition d'un ciment ; bien que leur addition ne donne pas exactement 100 %, nous constatons d'après les données originales fournies par Hald que la somme : $x_1 + x_2 + x_3 + x_4$ ne varie que de 95 à 99, si bien que n'importe laquelle de ces quatre variables est presque fonction des trois autres. En de telles conditions, la construction d'une équation de régression constitue toujours un problème délicat.

VIII - ELIMINATION REGRESSIVE

Comme nous l'avons déjà expliqué dans l'introduction, l'élimination régressive est une alternative à la sélection progressive. Nous commençons par l'équation de régression complète et nous abandonnons les variables indépendantes une par une selon l'ordre dans lequel elles donnent le moins d'accroissement à la somme résiduelle des carrés.

A cette fin, nous pouvons suivre exactement le même plan de calcul qu'au tableau 2.1 à condition de partir de la matrice des produits inversée. On peut en faire la preuve au moyen des formules données aux pages 364 et 365 de l'article de Wishart et Metakides (1953). Nous n'entrerons pas ici dans le détail théorique, mais nous illustrerons le procédé par un exemple numérique - (Tableau 8.1).

Tableau 8.1

Elimination régressive s'appliquant à l'exemple du tableau 2.1. Les flèches indiquent les variables successivement éliminées x_4, x_2, x_1, x_3.

Stade I (inverse de la matrice du tableau 2.1)

	$\xi_{1.234y}$	$\xi_{2.134y}$	$\xi_{3.124y}$	$\xi_{4.123y}$ ↓	$\eta_{.1234}$	ΔS_{yy}^{-1}
$\xi_{1.234y}$	+22,3679	+13,6816	- 4,2476	+ 0,7074	+10,7600	5,1760
$\xi_{2.134y}$		+10,1343	- 2,7731	+ 0,8790	+ 4,2889	1,8150
$\xi_{3.124y}$			+ 2,5335	- 0,2102	- 4,8688	9,3567
→ $\xi_{4.123y}$				+ 1,2919	- 0,2873	0,0639
$\eta_{.1234}$					+15,5746	
$S_{yy.01234} = (15,5746)^{-1} = 0,0642$						

Tableau 8.1 (suite)

Stade II

	$\xi_{1.23y}$	$\xi_{2.13y}$	$\xi_{3.12y}$	$\eta_{.123}$	ΔS_{yy}^{-1}
$\xi_{1.23y}$	+21,9806	+13,2003	- 4,1326	+10,9175	5,4225
$\xi_{2.13y}$		+ 9,5363	- 2,6301	+ 4,4843	2,1086
$\xi_{3.12y}$			+ 2,4993	+ 4,9155	9,6675
$\eta_{.123}$				+15,5108	
$S_{yy.0123} = (15,5108)^{-1} = 0,0645$					

Stade III

	$\xi_{1.3y}$	$\xi_{3.1y}$	$\eta_{.13}$	ΔS_{yy}^{-1}
$\xi_{1.3y}$	+ 3,7086	- 0,4920	+ 4,7105	5,9830
$\xi_{3.1y}$		+ 1,7740	- 3,6788	7,6288
$\eta_{.13}$			+13,4022	
$S_{yy.013} = (13,0422)^{-1} = 0,0746$				

Stade IV

	$\xi_{3.y}$	$\eta_{.3}$	ΔS_{yy}^{-1}
$\xi_{3.y}$	+ 1,7088	- 3,0539	5,4578
$\eta_{.3}$		+ 7,4192	
$S_{yy.03} = (7,4192)^{-1} = 0,1348$			

Stade V

	η	
η	+ 1,9614	
$S_{yy.0} = (1,9614)^{-1} = 0,5098$		

La matrice portée au tableau 8.1 comme Stade I est l'inverse de la matrice de produits du tableau 2.1, en haut. Cette matrice inverse a été calculée sur une calculatrice électronique à six décimales et arrondie par la suite à 4 décimales.

Il convient de remarquer que la variable dépendante est incluse dans la matrice des sommes de produits et dans son inverse. Au tableau 8.1 l'élément de la colonne marquée $\xi_{2.134y}$ et le rang marqué $\xi_{4.123y}$ est l'élément de la matrice inverse qui correspond à la somme de produits de $x_{2.0}$ et $x_{4.0}$. Au tableau 2.1, etc. Cette notation indiquait que nous avions inversé la matrice de somme de produits contenant x_1, x_2, x_3, x_4 et y. De même la notation $\xi_{2.13y}$ au stade II signifie que nous avons inversé la matrice contenant x_1, x_2,

x_3 et y, et c'est la matrice obtenue du tableau 2.1 par abandon de la colonne et du rang correspondant à x_4. Nous tenons tout le temps compte de matrices basées sur les sommes de produit corrigées ; c'est-à-dire qu'on a déjà tenu compte du terme constant x_0. Au tableau 2.1 ceci était clairement indiqué par la notation, mais au tableau 8.1 nous n'avons pas vu comment indiquer aisément la position exceptionnelle de x_0 sans amener une notation encore plus complexe.

On peut voir d'après les formules données par Wishart et Metakides (1953, pp. 364 et 365) que les éléments inscrits sous $(\eta_{.1234})^2$, $(\eta_{.123})^2$ dans les matrices inverses du tableau 8.1 son égaux à 1 divisé par les sommes résiduelles de carrés pour y. On a par exemple :

$$S_{yy.01234} = (15,5746)^{-1} = 0,0642$$

et de même :

$$S_{yy.0123} = (15,5108)^{-1} = 0,0645$$

contre les valeurs de 0,0641 et de 0,0644 rencontrées aux derniers stades du tableau 2.1. Ici $S_{yy.0123}$ représente la somme résiduelle de carrés de y lorsque l'équation de régression contient x_0, x_1, x_2 et x_3, etc.

Il s'ensuit également d'après la communication de Wishart et Metakides que l'élimination regressive peut être effectuée au moyen d'une réduction systématique de la matrice inverse exactement de la même manière que la réduction opérée au tableau 2.1 pour la sélection progressive. Nous calculons d'abord la dernière colonne du stade I, qui est la colonne η portée au carré et divisée par les éléments diagonaux. Par exemple :

$$\Delta s_{yy}^{-1} = 1,8150 = (4,2889)^2/10,1343 \quad ,$$

est la réduction à la somme inverse résiduelle de carrés qui aura lieu si nous éliminons la variable x_2 de l'équation de régression. Puisque nous souhaitons maintenant éliminer en premier la variable indépendante la moins importante, nous devons sortir la variable pour laquelle ΔS_{yy}^{-1} a la plus faible valeur. C'est x_4, et nous réduisons ensuite la matrice inverse en éliminant x_4 exactement comme nous avons réduit la matrice originale au tableau 2.1 en introduisant x_3.

En procédant de la même manière, nous constatons que les variables indépendantes doivent être éliminées dans l'ordre x_4, x_2, x_1, x_3. C'est l'ordre inverse de celui dans lequel elles étaient introduites par sélection progressive au tableau 2.1, si bien que les deux procédés dans ce cas aboutissent exactement à la même conclusion. Les sommes résiduelles de carrés déduites du tableau 8.1 coïncident donc avec celles du tableau 2.1 comme on le voit au tableau 8.2

Nous avons également essayé l'élimination régressive en quelques autres cas. Dans le problème de raffinage (§ 3) et dans le problème de fonderie (§ 4) l'élimination régressive entraîne le même ordre des variables indépendantes que la sélection progressive et nous n'avons pas de raison d'élucider les conclusions atteintes aux paragraphes cités. Dans l'exemple de Hald (§ 7), nous constatons une différence représentée au tableau 8.3.

Nous constatons que pour une équation de régression à trois variables indépendantes les deux procédés conduisent au même choix, qu'avec deux variables l'élimination régressive donne le meilleur résultat, mais que si

Tableau 8.2

Sommes résiduelles de carrés pour y selon l'élimination régressive

	Somme résiduelle de carrés	
	d'après le tableau 8.1	d'après le tableau 2.1
aucune	$(15,5746)^{-1} = 0,0642$	0,0641
x_4	$(5,5108)^{-1} = 0,0645$	0,0644
x_4, x_2	$(13,4022)^{-1} = 0,0746$	0,0746
x_4, x_2, x_1	$(7,4192)^{-1} = 0,1348$	0,1348
x_4, x_2, x_1, x_3	$(1,9614)^{-1} = 0,5098$	0,5098

nous ne retenons qu'une variable c'est la sélection progressive qui est la meilleure. Il semblerait donc qu'aucun des deux procédés ne conduise infailliblement au meilleur choix possible. On pourrait être enclin à supposer que si les deux fournissent la même séquence, cette séquence est unique et amène à la meilleure sélection de jeux de variables indépendantes. Nous n'avons pas été en mesure de faire la preuve du vrai ou du faux de cette supposition ; la théorie requise ne paraît pas facile.

Tableau 8.3

Application de la sélection progressive et de l'élimination régressive aux données de Hald sur le ciment (§ 7)

Sélection progressive		Elimination régressive	
Variables introduites	Somme résiduelle des carrés	Variables conservées	Somme résiduelle des carrés
aucune	2715,96	aucune	2715,96
x_4	883,86	x_2	906,34
x_1, x_4	74,76	x_1, x_2	57,91
x_1, x_2, x_4	47,97	x_1, x_2, x_4	47,97
toutes	47,92	toutes	47,92

A quel point le problème peut être embarrassant m'est rendu évident par une remarque privée du Dr. P.G. Moore(1). Il avait opéré les méthodes de sélection progressive et d'élimination regressive dans une situation où il s'agissait de construire une équation de régression à trois variables à choisir parmi un nombre total de 12 variables indépendantes. Les deux méthodes aboutissaient à des solutions différentes. Alors il avait fait calculer toutes les $C_{12}^3 = 220$ équations à trois variables concevables. Celle avec le vrai minimum de la somme des carrés résiduelle était encore différente des deux équations antérieures. Toutefois on était arrivé à trois équations à trois

---

(1) Statisticien chez A.E. Reed Ltd., West Malling, Angleterre.

variables expliquant à peu près la même fraction de la variabilité de la variable dépendante.

IX - PROGRAMMES DE CALCULATRICES -

Le principe de la sélection progressive a été appliqué dans quelques programmes de calculatrices. Nous n'avons pas l'intention d'en discuter ici en détail mais désirons simplement en mentionner un qui nous a semblé attrayant. Il a été mis au point chez Arthur D. Little, Inc., Boston, U.S.A.

La caractéristique principale de ce programme est d'opérer pas à pas et de permettre de procéder de plusieurs façons différentes après chacun de ces pas.

Par exemple, en partant de la matrice des sommes de produits nous pouvons d'abord demander à la calculatrice laquelle des variables indépendantes prises à part procure la réduction la plus forte de la somme résiduelle des carrés. Dans le cas du tableau 2.1 la réponse est :

$$x_3,\ S_{yy} = 0,5098,\ \Delta S_{yy} = 0,3750,\ S_{yy.3} = 0,1348$$

c'est-à-dire que la variable demandée est x_3 ; la somme résiduelle des carrés est $S_{yy} = 0,5098$ et sera réduite par 0,3750 à 0,1348 ; quelle est votre décision suivante ?

En nous appuyant sur ce renseignement, nous pouvons alors décider de conserver x_3 et donner ordre à la calculatrice de l'éliminer de la matrice et de passer au stade II du tableau 2.1. Ou en alternative nous pouvons remettre cette décision et demander d'abord laquelle des variables indépendantes prise à part suit x_3 par ordre d'importance.

La réponse à cette dernière question serait :

$$x_1,\ S_{yy} = 0,5098,\ \Delta S_{yy} = 0,3366,\ S_{yy.1} = 0,1732$$

Nous pouvons maintenant choisir : puisque x_1 produit également une réduction draconienne de la somme des carrés, nous pouvons décider de conserver x_1 en premier, parce que nous croyons pour des raisons techniques que x_1 est en réalité plus essentiel que x_3. A partir d'ici, nous pouvons maintenant décider de conserver soit x_1, soit x_3 et de passer au stade II , ou encore nous pouvons préférer voir d'abord laquelle des variables est de troisième importance.

Après avoir procédé au stade II, on peut recommencer le même processus de sélection.

Ou, si nous le souhaitons, nous pourrions aussi ordonner à la calculatrice d'introduire immédiatement les variables indépendantes par sélection progressive ou selon un ordre prédéterminé. L'avantage d'un tel programme est que nous pouvons chercher notre chemin à tâtons et bâtir l'équation de régression par degrés sur la base de l'expérience recueillie aux degrés précédents. Inutile de préciser que l'on ne peut appliquer un programme de cette sorte que sur une calculatrice relativement lente qui imprime une à une les réponses aux différentes questions. Avec une calculatrice rapide on perdrait trop de temps à lire les réponses et à prendre des décisions aux stades intermédiaires.

X - EMPLOI DES POLYNOMES ORTHOGONAUX -

Lorsque nous avons une variable indépendante X à intervalles équidistants et que nous voulons exprimer la variable dépendante Y comme fonction polynôme de X

$$Y = b_0 + b_1 X + b_2 X^2 + b_3 X^3 + \ldots \qquad (10.1)$$

l'analyse peut s'effectuer commodément au moyen de polynômes orthogonaux. Il faut toutefois les appliquer avec prudence comme le montre l'exemple suivant.

Une barre de métal est supportée en deux points (P_1, P_2 à la figure 10.1) et chargée en son centre d'un poids W. Au moyen d'un micromètre à cadran, nous mesurons la flexion d en fonction de la longueur L de la barre et du poids W.

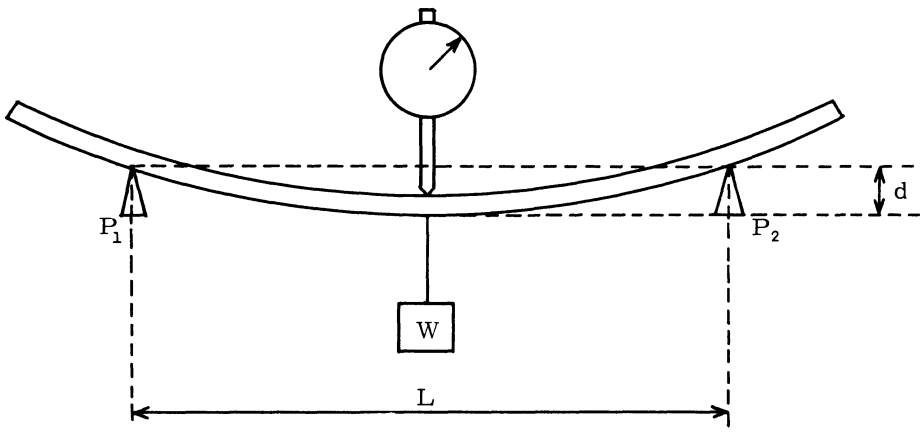


Figure 10.1

C'est un exercice élémentaire de physique expérimentale.

Il nous a été donné de recueillir les résultats de cette expérience pour les employer dans la démonstration de l'application de méthodes statistiques d'analyse à un tel cas. Pour l'instant nous ne parlerons que de quelques résultats obtenus avec un poids constant et enregistrés au tableau 10.1A.

L'analyse de ces données par polygônes orthogonaux donne les résultats du tableau 10.1B. Nous voyons que ξ_0, ξ_1 et ξ_2 sont tous trois significatifs et que ξ_3 et ξ_4 ne le sont pas, la conclusion étant que la flexion d peut être représentée convenablement comme fonction quadratique de la longueur L.

D'un point de vue purement statistique, il se peut que ce résultat soit pleinement satisfaisant. Du point de vue physique il ne l'est pas ! Car d'après la théorie de l'élasticité la flexion devrait varier comme L^3, alors que nous constatons au tableau 10.1B que le polynôme au troisième degré en L est insignifiant. De plus, cet exemple n'est pas un cas isolé ; nous avons plusieurs fois répété l'expérience avec des barres différentes, mais toujours

avec le même résultat. Le carré moyen pour ξ_3 était habituellement plus élevé qu'au tableau 10.1B mais n'avait jamais de signification statistique. C'est là depuis longtemps un des paradoxes non résolus de nos cartons.

Tableau 10.1

Flexion d'une barre de métal de longueur L
sous une charge constante de 400 grammes

A - Données originales

L :	55	60	65	70	75	80	85	90	cm
d :	1,165	1,518	1,948	2,428	2,965	3,610	4,242	5,010	mm

B - Analyse de variance par polygônes orthogonaux

		S.C. (mm^2)	degrés de l.	Var. (mm^2)
Terme constant	0	65,4711	1	65,4711
Linéaire	1	12,6270	1	12,6270
Quadratique	2	0,1701	1	0,1701
Cubique	3	0,0000	1	0,0000
Quartique	4	0,0001	1	0,0001
Résiduel	5	0,0011	3	0,0004

C - Sommes des carrés par importance décroissante

Source	Somme des carrés (mm^2)				
	Stade I	II	III	IV	V
L^3	78,2580				
L^4	77,1641	0,0079			
L^2	77,1902	0,0066	0,0023		
L	73,0492	0,0050	0,0023	0,0000	
1	65,4711	0,0058	0,0023	0,0000	0,0000
Carré moyen résiduel = 0,0004 ; degrés de l. = 3					

Ces remarques sont particulièrement pertinentes si nous avons une raison de supposer que la ligne de régression passe par l'origine, si bien que l'équation de régression polynomiale n'aurait pas besoin d'un terme constant. En ce cas les polynômes orthogonaux types ne conviennent pas car ils contiennent la constante de par leur essence. On en trouve l'exemple dans un article de Leech et Healy (1959) dans une étude sur les taux de croissance. Ils ont employé des polynômes orthogonaux types et ont dû mettre au point une technique spéciale et assez compliquée pour justifier que la courbe de croissance doive passer par l'origine.

Tableau 10.2

Valeurs de d calculées d'après les équations (10.2), (10.3) et (10.4) comparées à l'observation originale.

L =	d (mm) observé	d calculé d'après éq. (10.2)	diff.	éq. (10.3)	diff.	éq. (10.4)	diff.
55	1,165	1,164	1	1,162	3	1,158	7
60	1,518	1,522	- 5	1,523	- 5	1,503	15
65	1,948	1,943	5	1,944	4	1,911	37
70	2,428	2,428	0	2,427	1	2,367	41
75	2,965	2,976	-11	2,973	- 8	2,936	29
80	3,610	3,588	22	3,584	26	3,563	47
85	4,242	4,263	-21	4,260	-18	4,273	31
90	5,010	5,003	7	5,001	9	5,073	63

Les différences entre les valeurs observées et calculées ont été enregistrées en 10^{-3} mm.

En ces conditions, il vaudrait mieux employer les polynômes orthogonaux sans terme constant tels qu'ils ont été dérivés et tabulés par Sibuya et Haga (1959). Sinon nous pourrions construire un jeu de polynômes orthogonaux en prenant les puissances de X dans l'ordre X^1, X^0, X^2, X^3, etc, ou X^1, X^2, X^0, X^3, X^4, ... Alors, à partir du troisième ou quatrième polynôme ils seraient identiques aux polynômes orthogonaux types, et seuls les deux ou trois premiers devraient être révisés. De plus, en employant ces suites, le polynôme introduisant X^0 peut servir à vérifier si la courbe de régression passe par l'origine ou non.

Une autre question est de savoir quelle est la valeur d'un test de signification dans les situations prises en considération dans ce chapitre.

Cependant, cette anomalie disparaît lorsque nous renonçons aux polynômes orthogonaux pour introduire 1, L, L^2, L^3, L^4 dans l'ordre de leur importance décroissante. Pour l'instant, (voir tableau 10.1C) L^3 ressort comme la plus importante des variables indépendantes, et une fois que l'on en a tenu compte, le reste des sommes de carrés, bien qu'elles aient encore une signification statistique, sont environ 10 000 fois moindres que celle qui se rapporte à L^3. Elles sembleraient devoir être imputées à des imperfections mineures de l'expérience. Les données sont en accord presque parfait avec la théorie physique.

Le présent exemple comporte un avertissement contre l'usage inconsidéré des polynômes orthogonaux.

Les données en question peuvent être aussi bien décrites comme suit :

$$d = (143,350 - 0,74896\ L + 0,12728\ L^2)\quad 10^{-2}\ \text{mm} \qquad (10.2)$$

formule résultant des polynômes orthogonaux que par

$$d = (-1,6659\ L^2 + 0,12059\ L^3 - 0,000372\ L^4)\quad 10^{-4}\ \text{mm} \qquad (10.3)$$

obtenu en achevant l'analyse du tableau 10.16, alors qu'enfin

$$d = 0,69586 \quad 10^{-5} \; L^3 \text{ mm} \qquad (10.4)$$

bien que moins parfait, soit encore une très bonne approximation. Ceci est illustré au tableau 10.2 où les observations originales sont comparées aux valeurs données par chacune de ces fonctions. Les équations 10.2 et 10.3 donnent des résultats presque identiques ; 10.4 présente de légères différences systématiques, mais si notre but était de vérifier la théorie physique, il est probable qu'on les négligerait comme résultant de petites imperfections.

Il faut se rappeler qu'en construisant des polynômes orthogonaux on introduit la variable indépendante successivement par ordre de puissance croissante de X : 1, X, X^2, X^3, ...

Bien que, en raison de l'orthogonalité, les carrés moyens associés à ces polynômes ne dépendent plus de l'ordre dans lequel nous les prenons, l'ordre dans lequel nous prenons les puissances de X est encore primordial. Ainsi le carré moyen pour le polynôme du second degré ξ_2 est en réalité le carré moyen de X^2 lorsqu'on a déjà tenu compte de X^0 et de X^1, etc. Si nous ne visons qu'à une description des observations nous pouvons n'avoir besoin que des polynômes orthogonaux et de l'équation qui en résulte (10.2), mais si nous désirons vérifier une relation fonctionnelle, les polynômes orthogonaux peuvent nous dérouter. Nous devrions prendre la relation donnée pour base de notre analyse.

Pour la barre qui se plie nous constatons que $d = b \; L^3$ ne décrit pas complètement nos observations, et que l'adaptation est nettement meilleure si nous ajoutons également un terme à L^4 (Tableau 10.1C). Le physicien répondrait qu'il ne s'attend pas à ce que son expérience soit absolument parfaite et que le terme L^3 à lui seul approche d'assez près pour prouver que la théorie de l'élasticité est correcte. Le degré auquel les données expérimentales devraient coïncider avec une équation théorique pour prouver ou démentir la théorie ne fait pas partie, à notre avis, des questions auxquelles on peut répondre en se basant sur une argumentation statistique.

REFERENCES

[1] - ANDERSON, H.E. and FRUCHTER, B. (1960) - Some multiple correlation and predictor selection methods. Psychometrika (1960) *25*, 59-76, refs.

[2] - BARTOO, J.D. and LAIRD, D. (1958) - A program for applying the principle of parsimony in multiple regression. Paper presented to the Association of Computing Machinery at Urbana, III. 1958.

[3] - CANNING, F.L. (1959) - Estimating load requirements in a job shop. J. Ind. Eng. (1959) *10*, 447-479.

[4] - DE BAUN, R.M. and SCHNEIDER, M. (1957) - Some examples of multivariate analysis. Transaction of the Middle Atlantic Conference of the A.S.Q.C. (1957), 19-25.

[5] - DUBOIS, P.H. (1956) - Multivariate correlational analysis.- Harper, New-York, 1956.

[6] - DUNCAN and KENNEY - On the solution of normal equations and related topics. Edwards Brothers, Ann. Arbor (1948).

[7] - DWYER, P.S. (1945) - The square root method and its use in correlation and regression J.A.S.A. (1945) *40*, 493-503.

[8] - EFFROYMSON, M.A. (1955) - Stepwise procedure for calculation of multiple regression. Paper delivered at the Gordon Research Conference on Statistics in 1955.

[9] - HADER, R.J. and GRANDAGE, A.H.E. (1956) - Simple and multiple regression analysis in "Experimental designs in industry", Proceedings of a symposium held in 1956, edited by V. Chew and published by Wiley and Sons, New York, 1958.

[10] - HALD, A. (1952) - Statistical theory with engineering applications, pages 647-650. Wiley and Sons, New York, 1952.

[11] - HORST, P. (1934) - Item analysis by the method of successive residuals. Jl. of Experimental Education (1934), *2*, 254-263.

[12] - KRAMER, C.Y. (1957) - Simplified computations for multiple regression. Industrial Quality Control (1957) *13*, Nr. 8, 8-11.

[13] - LAYTON, W.L. (1951) - The relationship between the method of successive résiduals and the method of exhaustion - Psychometrika (1951) *16*, 51-56.

[14] - LEECH, F.B. and HEALY, M.J.R. (1959) - The analysis of experiments on growth rates. Biometrics (1959) *15*, 98-106.

[15] - MORICEAU, J. (1954) - Une méthode statistique d'évaluation des temps opérationnels. Revue de Statistique Appliquée (1954) *2*, n° 3, 57-74.

[16] - PEACH - The abbreviated Doolittle method for matrix inversion. Institute of Statistics, Mimeo series 28, Raleigh N.C.

[17] - SIBUYA, M. and HAGA, T. (1959) - Orthogonal polynomials without constant term. Annals of the Institute of Statistical Mathematics, Tokyo (1959), *10*, 209-222.

[18] - SUMMERFIELD, A. and LUBIN (1951) - A square root method of selecting a minimum set of variables in multiple regression.
I - The method, Psychometrika (1951) *16*, 271-283.
II - A worked exemplae, Psychometrika (1951) 16, 425-437.

[19] - WISHART, J. and METAKIDES, Th. (1953) - Orthogonal polynomial fitting, Biometrika (1953) *40*, 361-369.