QUELQUES PIÈGES CACHÉS DES MÉTHODES DE SÉLECTION DE VARIABLES EN RÉGRESSION LINÉAIRE MULTIPLE

par

Jean-François Dubois

mémoire présenté au Département de mathématiques et d'informatique
en vue de l'obtention du grade de maître ès sciences (M.Sc.)

FACULTÉ DES SCIENCES
UNIVERSITÉ DE SHERBROOKE

Sherbrooke, Québec, Canada, janvier 2000

Le 3 mai 2000 , le jury suivant a accepté ce mémoire dans sa version finale.
date

Président-rapporteur: M. François Dubeau
Département de mathématiques et d'informatique

Membre: M. Bernard Colin
Département de mathématiques et d'informatique

Membre: M. Ernest Monga
Département de mathématiques et d'informatique

SOMMAIRE

La régression linéaire est une méthode d'analyse des données parmi les plus anciennes. On attribue à K. F. Gauss (vers 1809) le mérite d'avoir développé la méthode des moindres carrés afin d'ajuster une équation qui est linéaire du point vue de ses paramètres. Des développements impressionnants sont survenus au cours des 150 années qui ont suivies pour finalement stagner vers 1959 puisque la majorité des idées nouvelles nécessitait l'utilisation d'un ordinateur très performant pour éclore. Le perfectionnement des ordinateurs combiné à la réduction des coûts d'utilisation a permis à des méthodes telles que la sélection d'un sous-ensemble de variables, dont il sera question dans ce mémoire, de voir le jour.

Dans le présent document, nous jetterons un regard à la fois descriptif et critique à l'égard des méthodes de sélection développées principalement au cours des trente dernières années. Nous étudierons un certain nombre de critères et de méthodes de sélection construites de façon à identifier un sous-ensemble de variables de qualité acceptable que l'on qualifiera de "meilleur" sous-ensemble de variables.

REMERCIEMENTS

Je tiens à remercier mon directeur de recherche Monsieur Ernest Monga pour la confiance qu'il a eue en moi en acceptant de diriger mon travail ainsi que pour le sujet intéressant qu'il m'a proposé. J'ai apprécié sa compétence et ses qualités humaines. Je le remercie également pour le support financier qu'il m'a accordé au cours de ma première année.

J'aimerais également remercier ma famille qui m'a grandement encouragé à finaliser la rédaction de ce travail dans des délais raisonnables malgré un emploi à temps plein à grande distance de l'Université de Sherbrooke.

TABLE DES MATIÈRES

INTRODUCTION

Le but premier de ce travail est d'étudier certaines méthodes statistiques permettant de réduire le nombre de variables dans un modèle de régression linéaire multiple. Par l'entremise de ce document, nous présenterons non seulement ces méthodes d'un point de vue descriptif, mais nous nous attarderons également à exposer certains de leurs inconvénients. Nous verrons, en particulier, que derrière ces méthodes se cachent des pièges potentiels qui guettent toute personne n'y étant pas sensibilisée. Le problème de la sélection de sous-ensembles de variables est abordé au deuxième et au troisième chapitre alors que le premier chapitre est réservé à la théorie concernant la régression linéaire multiple.

Le premier chapitre constitue un résumé de la théorie nécessaire à la bonne compréhension des chapitres subséquents. On présentera tout d'abord le modèle de régression linéaire multiple sous forme d'équation mathématique puis sous forme matricielle. Au cours de ce chapitre, on exploitera aussi la théorie matricielle afin de démontrer de façon esthétique certains résultats importants en régression linéaire multiple. Les formes quadratiques y seront omniprésentes.

Le second chapitre traite des critères de sélection du "meilleur" sous-ensemble de variables. À l'origine de ces critères de sélection se trouvent des statistiques qui, une fois interprétées correctement, permettent d'identifier un sous-ensemble optimal. On déduira non seulement les statistiques mais on s'intéressera aussi aux inconvénients potentiels de chacun des critères considérés. On remarquera, entre autres, que certains des critères présentés sont affectés par "l'inclinaison" de la surface de régression ou par la dispersion des valeurs des variables. Au cours de ce chapitre, nous avons exploité certaines remarques faites par HAHN [22] pour mettre en relief quelques inconvénients se rapportant au coefficient de détermination. La majorité des inconvénients énumérés sont accompagnés d'exemples que nous avons construits; nous les avons voulus simples d'une part mais surtout percutants, dans la mesure du possible. Chaque critère étudié est également brièvement discuté afin de bien saisir les implications de son utilisation. Ce chapitre présente également la façon d'utiliser les divers critères sur un ensemble de données.

Pour ce qui est du troisième chapitre, on s'intéresse aux procédures de sélection automatisée, i.e. à des algorithmes qui, une fois programmés, permettent d'identifier le "meilleur" sous-ensemble de variables. Tout comme au deuxième chapitre, on s'intéressera

non seulement à l'aspect descriptif mais également aux inconvénients. Ceci nous permettra de réaliser l'impact potentiel d'éliminer complètement un sous-ensemble à partir du moment où il contient une ou des variables qu'une procédure automatisée juge inadéquates. On reprendra aussi le même ensemble de données qu'au chapitre précédent pour montrer de façon concrète les diverses étapes de sélection de chacune des procédures de sélection successives.

CHAPITRE 1

LE MODÈLE DE RÉGRESSION LINÉAIRE MULTIPLE

1.1 Introduction

La modélisation des données est un contexte d'application de la statistique très utile pour établir des relations entre deux groupes de variables. La méthode probablement la plus répandue en ce sens est la régression linéaire multiple qui permet de mettre en relation un groupe composé d'un certain nombre (fini) de variables dites explicatives avec un autre groupe formé d'une seule variable dite expliquée.

1.1.1 Présentation du modèle

Afin de construire le modèle, on considère k - 1 variables indépendantes (x_i, $i = 1,\ldots,k-1$) à valeurs fixées (donc non aléatoires) et une variable dépendante Y que l'on prendra soin de désigner par une lettre majuscule pour insister sur son caractère aléatoire. Dans ce travail, nous supposerons que l'équation de régression décrivant les deux groupes de variables comporte toujours un terme constant. Le modèle de régression linéaire multiple avec terme constant est donné par l'équation suivante:

$$Y = \beta_0 + \sum_{j=1}^{k-1} \beta_j x_j + \varepsilon \tag{1.1}$$

où $\beta_0, \beta_1, \ldots, \beta_{k-1}$ sont les paramètres inconnus du modèle et ε est une perturbation aléatoire dont l'espérance est nulle. L'ajout de cette perturbation aléatoire est motivé par le fait qu'on a alors $E(Y) = \beta_0 + \sum_{j=1}^{k-1} \beta_j x_j$. Ce modèle reflète mieux la réalité qu'un modèle sans erreur (modèle (1.1) où $\varepsilon = 0$).

En effet, imaginons par exemple, un modèle à une seule variable explicative où certaines observations de cette variable prennent des valeurs identiques mais où la variable expliquée ne prend que des valeurs distinctes. En tentant d'expliquer ces données à l'aide d'une droite, il est impossible que celle-ci passe par tous les points d'où l'intérêt d'admettre un terme

d'erreur. Pour tout modèle où $\varepsilon \neq 0$, on pourra considérer plusieurs distributions de probabilité sur la perturbation aléatoire.

Supposons que nous ayons n observations indépendantes $Y_1, Y_2, \ldots, Y_n$ de la variable dépendante Y, alors chaque observation est décrite par le modèle :

$$Y_i = \beta_0 + \sum_{j=1}^{k-1} \beta_j x_{ij} + \varepsilon_i , \quad i = 1, \ldots, n . \tag{1.2}$$

S'il était possible de contrôler tous les t facteurs qui peuvent expliquer la variable dépendante nous considérerions alors le modèle "exact" :

$$Y_i = \beta_0 + \sum_{j=1}^{k-1} \beta_j x_{ij} + \sum_{j=k}^{t} \beta_j x_{ij} , \quad i = 1, \ldots, n . \tag{1.3}$$

Afin de rendre le modèle de régression opérationnel, il est indispensable d'estimer les coefficients $\beta_0, \beta_1, \ldots, \beta_{k-1}$ appelés : coefficients de régression. En ce sens, remarquons qu'une des conséquences de la présence des perturbations aléatoires est de faire de la variable expliquée une variable aléatoire. Ainsi, le problème de la détermination des coefficients de régression se ramène au problème de l'estimation statistique de paramètres inconnus.

La procédure la plus courante pour estimer les paramètres d'un modèle linéaire est la méthode des moindres carrés. Supposons que $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_{k-1}$ soient les estimateurs de $\beta_0, \beta_1, \ldots, \beta_{k-1}$ et posons $\hat{Y}_i = \hat{\beta}_0 + \sum_{j=1}^{k-1} \hat{\beta}_j x_{ij}$, $i = 1, \ldots, n$. Il est alors clair que $\hat{Y}_i$ est un estimateur de $E(Y_i)$ et l'erreur commise par le modèle est donnée par $Y_i - \hat{Y}_i$. La méthode des moindres carrés recommande de choisir les valeurs de $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_{k-1}$ qui rendent minimum la somme des carrés des erreurs, à savoir $\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$.

1.1.2 Représentation matricielle du modèle

Afin de faciliter bon nombre de calculs nous aurons recours à la représentation matricielle du modèle considéré.

Posons:

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1,k-1} \\ 1 & x_{21} & \cdots & x_{2,k-1} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1} & \cdots & x_{n,k-1} \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{k-1} \end{pmatrix}, \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Ainsi le modèle (1.2) peut s'écrire sous la forme $Y = X\beta + \varepsilon$.

Pour arriver à nos fins, il suffira de minimiser la quantité $\varepsilon'\varepsilon$, où ε' est la tranposée du vecteur ε.

Puisque $Y' X \beta$ est un scalaire, on a:

$$\varepsilon'\varepsilon = (Y-X\beta)'(Y-X\beta) = Y'Y - 2\beta' X'Y + \beta'(X'X)\beta .$$

Ainsi,

$$\frac{\partial(\varepsilon'\varepsilon)}{\partial\beta} = 0 \Leftrightarrow -2X'Y + 2X'X\beta = 0 \Leftrightarrow X'X\beta = X'Y .$$

Donc, si $X'X$ est non singulière,

$$\hat{\beta} = (X'X)^{-1}X'Y .$$

De plus, on établit sans peine que la quantité :

$$(Y-X\beta)'(Y-X\beta) = (Y-X\hat{\beta})'(Y-X\hat{\beta}) + (\hat{\beta}-\beta)' X'X (\hat{\beta}-\beta) ,$$

atteint son minimum pour $\beta = \hat{\beta}$.

Lorsqu'il n'y a qu'une seule variable explicative, la régression linéaire sera dite simple. Pour simplifier certains concepts, nous aurons souvent recours à ce modèle. Il devient alors intéressant d'évaluer $\hat{\beta}$ dans ce cas particulier.

On a:

$$Y=\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \quad , \quad X=\begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \quad , \quad \beta=\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \quad , \quad \varepsilon=\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Alors,

$$X^tX=\begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix} \quad \text{et donc} \quad (X^tX)^{-1}=\frac{1}{n\sum_{i=1}^n x_i^2-\left(\sum_{i=1}^n x_i\right)^2}\begin{pmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{pmatrix}.$$

De plus,

$$X^tY=\begin{pmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n x_iY_i \end{pmatrix}.$$

Ainsi,

$$\hat{\beta}=(X^tX)^{-1}X^tY=\frac{1}{n\sum_{i=1}^n(x_i-\bar{x})^2}\begin{pmatrix} \left(\sum_{i=1}^n Y_i\right)\left(\sum_{i=1}^n x_i^2\right)-\left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n x_iY_i\right) \\ n\sum_{i=1}^n x_iY_i-\left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n Y_i\right) \end{pmatrix}$$

$$=\begin{pmatrix} \bar{Y}-\hat{\beta}_1\bar{x} \\ \dfrac{\sum_{i=1}^n(x_i-\bar{x})(Y_i-\bar{Y})}{\sum_{i=1}^n(x_i-\bar{x})^2} \end{pmatrix},$$

c'est-à-dire : $\hat{\beta}_0=\bar{Y}-\hat{\beta}_1\bar{x}$ et $\hat{\beta}_1=\dfrac{\sum_{i=1}^n(x_i-\bar{x})(Y_i-\bar{Y})}{\sum_{i=1}^n(x_i-\bar{x})^2}$.

1.1.3 Conventions

Voici maintenant quelques conventions qui prévaudront pour tout le présent document à moins d'avis contraire. Nous utiliserons l'expression "modèle à k paramètres" lorsque nous nous référerons au modèle (1.2). Dans ce cas, l'équation de régression sera construite à l'aide d'une combinaison linéaire de k -1 variables indépendantes et d'un terme d'ordonnée à l'origine. Remarquons que l'équation peut également être déterminée à partir de k -1 transformations de variables (linéaire ou non) puisque la linéarité du modèle n'intervient qu'au travers des paramètres et non des variables. Cependant le sujet des transformations ne sera pas discuté dans ce mémoire de sorte que les modèles seront construits à partir des variables à leur état original. Lorsque nous voudrons insister sur le fait qu'un modèle est construit à partir de toutes les variables on utilisera le lettre "p " pour désigner le nombre de paramètres. De plus, on dira que l'on construit alors le modèle complet. Notons que l'appelation "modèle complet" est un abus de language en quelque sorte puisque, comme nous l'avons remarqué précédemment, on ne peut considérer toutes les variables. Ainsi, par "toutes les variables" on entend "toutes les variables ayant été mesurées".

Tout au long de ce travail on supposera que le nombre d'observations excède le nombre de variables de sorte que les quantités n-p et n-k soient strictement positives. On supposera également la non singularité de la matrice $X'X$ puisque l'on s'intéressera à son inverse. De toutes façons, dans la plupart des situations pratiques, si $n > p$, la matrice $X'X$ sera habituellement non singulière. Bien qu'il existe des notions d'inverses généralisés en présence de singularité, nous ne nous étenderons pas sur le sujet. Le lecteur intéressé pourra consulter la section 1.5 de GRAYBILL [20].

Afin d'estimer les paramètres de l'équation de régression aucune hypothèse sur les distributions des erreurs (ε_i, $i = 1,\ldots,n$) n'est requise. Cependant, dans la suite, et particulièrement lorsque nous voudrons faire de l'inférence, une distribution devra être spécifiée. La loi normale semble toute indiquée pour remplir ce rôle puisqu'en pratique pour des tailles d'échantillon assez grandes cette hypothèse est habituellement vérifiée. La densité de cette loi ainsi que celle de plusieurs autres lois sont présentées dans l'annexe 1. À l'instar des aspects antérieurs, nous nous placerons de nouveau dans le contexte "le plus courant" et nous supposerons que les perturbations aléatoires $\varepsilon_1,\ldots,\varepsilon_n$ sont indépendantes, de loi normale, d'espérance nulle et de variance finie. Nous noterons ceci $\varepsilon_i \sim N(0\ ,\sigma^2)$ où le symbole " $\sim$ " signifie "distribué suivant une loi". Des résultats élémentaires de probabilité permettent de conclure que le vecteur ε suit une loi multinormale d'espérance $E(\varepsilon) = 0_n$ (où

0_n désigne un vecteur de longueur n formé de 0) et de matrice de variances-covariances $\mathrm{Var}(\varepsilon) = \sigma^2 I_n$ (où I_n est la matrice identité de dimension $n \times n$). Ce comportement sera décrit par $\varepsilon \sim N_n(0_n, \sigma^2 I_n)$. À partir du modèle $Y = X\beta + \varepsilon$, il en découle immédiatement que $Y \sim N_n(X\beta, \sigma^2 I_n)$.

1.2 Résultats en régression linéaire

Certains résultats importants en régression linéaire seront exposés dans la section qui suit. Cette section, ne se voulant pas exhaustive, présente des résultats en insistant sur ceux qui interviendront dans les chapitres subséquents.

1.2.1 Analyse de la variance

Certaines quantités jouent un rôle primordial dans la théorie de la régression linéaire avec terme constant. Chaque quantité sera désignée par l'abréviation "la plus répandue" dans la littérature en langue anglaise. Les deux premières lettres de chaque abréviation seront SS pour désigner la somme des carrés (sum of squares).

$SSR = \sum_{i=1}^{n} \left(\hat{Y}_i - \overline{Y}\right)^2$ = somme des carrés due à la régression,

$SSE = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$ = somme des carrés due à l'erreur,

$SST = \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2$ = somme totale des carrés.

On vérifie sans peine que $SST = SSR + SSE$.

On résume habituellement cette décomposition de l'écart quadratique à l'aide d'un tableau couramment appelé "tableau ANOVA".

Voici le tableau ANOVA élémentaire :

Tableau 1.1: ANOVA

Source de variation	Degrés de liberté	Sommes des carrés	Somme pondérée des carrés
Régression	$k-1$	SSR	$MSR = SSR/(k-1)$
Erreur	$n-k$	SSE	$MSE = SSE/(n-k)$
Total	$n-1$	SST	$MST = SST/(n-1)$

On peut décomposer davantage la variabilité en ce qui à trait à la somme des carrés due à l'erreur pour tenir compte de ce qu'il sera convenu d'appeler des "répétitions". Afin de définir cette notion, notons d'abord que pour chacune des observations, les variables explicatives prennent diverses valeurs. Il peut parfois arriver que pour une ou plusieurs observations données, les valeurs de ces variables soient tout à fait identiques (sans nécessairement que la variable expliquée prenne la même valeur). Dans de tels cas, on affirmera qu'il existe des répétitions.

On débute en plaçant dans le même bloc les valeurs de la variable expliquée pour lesquelles les valeurs des variables explicatives sont identiques. On aura donc:

$Y_{11}, Y_{12}, \ldots, Y_{1n_1}$ dans le premier bloc contenant n_1 observations,
$Y_{21}, Y_{22}, \ldots, Y_{2n_2}$ dans le 2e bloc contenant n_2 observations,
$\vdots$
$Y_{I1}, Y_{I2}, \ldots, Y_{Ink}$ dans le Ie bloc contenant n_k observations.

Si on dénote par $\overline{Y}_i = \left(\sum_{j=1}^{n_i} Y_{ij} \right) \Big/ n_i$ la moyenne dans le ie bloc, alors, en remarquant que l'estimateur est le même pour chaque observation dans un bloc donné ($\hat{Y}_{ij} = \hat{Y}_i$), on a que :

$$Y_{ij} - \hat{Y}_{ij} = (Y_{ij} - \overline{Y}_i) + (\overline{Y}_i - \hat{Y}_{ij}) = (Y_{ij} - \overline{Y}_i) + (\overline{Y}_i - \hat{Y}_i).$$

Alors,

$$\begin{aligned} SSE &= \sum_{i=1}^{I} \sum_{j=1}^{n_i} (Y_{ij} - \hat{Y}_{ij})^2 \\ &= \sum_{i=1}^{I} \sum_{j=1}^{n_i} \{(Y_{ij} - \overline{Y}_i) + (\overline{Y}_i - \hat{Y}_i)\}^2 \\ &= \sum_{i=1}^{I} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_i)^2 + \sum_{i=1}^{I} \sum_{j=1}^{n_i} (\overline{Y}_i - \hat{Y}_i)^2 \quad \text{(puisque le terme issu du produit croisé s'annule).} \end{aligned}$$

On a donc $SSE = SSPE + SSLF$ où:

$SSPE = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y}_i)^2$= somme des carrés des erreurs pures (pure error sum of squares),

$SSLF = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(\overline{Y}_i-\hat{Y}_i)^2 = \sum_{i=1}^{I} n_i(\hat{Y}_i-\overline{Y}_i)^2$= somme des carrés due au manque d'ajustement (lack of fit sum of squares)

On peut alors présenter le tableau ANOVA sous la forme suivante:

Tableau 1.2: ANOVA avec décomposition de la variation due à l'erreur

Source de variation	Degrés de liberté	Sommes des carrés	Somme pondérée des carrés
Régression	$k-1$	SSR	$MSR = SSR/(k-1)$
Manque d'ajustement	$I-k$	$SSLF$	$MSLF = SSLF/(I-k)$
Erreur pure	$n-I$	$SSPE$	$MSPE = SSPE/(n-I)$
	$n-k$	SSE	
Total	$n-1$	SST	$MST = SST/(n-1)$

Une autre forme intéressante est celle permettant de comparer un modèle à k paramètres au même modèle auquel on enlève r variables (r paramètres parmi $\beta_1,\beta_2,\ldots,\beta_{k-1}$).

On utilisera:

SSR_{k-r} = somme des carrés due à la régression dans le modèle réduit à k-r paramètres,

$SSR_r^* = SSR - SSR_{k-r}$ = somme des carrés dans la portion omise.

On est alors amené a considérer le tableau ANOVA suivant :

Tableau 1.3: ANOVA avec décomposition de la variation due à la régression

Source de variation	Degrés de liberté	Sommes des carrés	Somme pondérée des carrés
Variation dans le modèle réduit (à k-r paramètres)	$k-r-1$	SSR_{k-r}	$MSR_{k-r} = SSR_{k-r}/(k-r-1)$
Variation dans la portion omise	r	SSR_r^*	$MSR_r^* = SSR_r^*/r$
	$k-1$	SSR	
Erreur	$n-k$	SSE	$MSE = SSE/(n-k)$
Total	$n-1$	SST	$MST = SST/(n-1)$

1.2.2 Formes quadratiques

Par commodité, certaines propriétés bien connues sur les matrices sont énoncées dans l'annexe 2. Dans certaines des démonstrations des résultats de cette section on utilise ces propriétés qui sont numérotées de (i) à (vi).

Définition 1.1

Soit $Y = (Y_1, Y_2, ..., Y_n)^t$ un vecteur de $\mathbb{R}^n$ et $A = (a_{ij})_{n\times n}$ une matrice carrée d'ordre n. La forme quadratique Q associée à A est la fonction de $\mathbb{R}^n$ dans $\mathbb{R}$ définie par $Q(Y) = Y^t A\, Y = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij} Y_i Y_j$.

On peut supposer sans perte de généralité que A est symétrique (i.e. $A^t = A$) car $Y^t A\, Y = Y^t B\, Y$ où $B = \frac{1}{2}(A^t + A)$ est une matrice symétrique.

Proposition 1.2

Les quantités SSR et SSE telles qu'introduites précédemment sont en fait des formes quadratiques. On a :

$SSR = Y^t\left[X\,(X^tX)^{-1}X^t - \frac{1}{n}1_n\,1_n^t\right]Y$ et

$SSE = Y^t\left[I_n - X\,(X^tX)^{-1}X^t\right]Y$,

où $1_n = (1,...,1)^t$ est le vecteur de longueur n dont toutes les composantes sont égales à 1.

Démonstration :

$$SSR = \sum_{i=1}^{n}\left(\hat{Y}_i - \overline{Y}\right)^2$$

$$= \sum_{i=1}^{n}\hat{Y}_i^2 - 2\overline{Y}\sum_{i=1}^{n}\hat{Y}_i + \sum_{i=1}^{n}\overline{Y}^2$$

$$= \sum_{i=1}^{n}\hat{Y}_i^2 - n\,\overline{Y}^2 \quad (\text{car } \sum_{i=1}^{n}\hat{Y}_i = n\overline{Y}, \text{ en vertu du fait que } 0 = \sum_{i=1}^{n}\varepsilon_i = \sum_{i=1}^{n}(Y_i - \hat{Y}_i) = n\overline{Y} - \sum_{i=1}^{n}\hat{Y}_i)$$

$$= \hat{Y}^t\hat{Y} - n(\tfrac{1}{n}1_n{}^t Y)^t(\tfrac{1}{n}1_n{}^t Y)$$

$$= (\hat{\beta}^t X^t)(X\,\hat{\beta}) - \tfrac{1}{n}Y^t 1_n\, 1_n{}^t Y$$

$$= Y^t X\,(X^tX)^{-1}X^tX\,(X^tX)^{-1}X^tY - \tfrac{1}{n}Y^t 1_n\, 1_n{}^t Y$$

$$= Y^t\left[X\,(X^tX)^{-1}X^t - \tfrac{1}{n}1_n\, 1_n{}^t\right]Y\,.$$

$$SSE = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$$

$$= (Y - X\hat{\beta})^t(Y - X\hat{\beta})$$

$$= Y^tY - Y^tX\,\hat{\beta} - \hat{\beta}^tX^tY + \hat{\beta}^tX^tX\,\hat{\beta}$$

$$= Y^tY - \hat{\beta}^tX^tY - \hat{\beta}^tX^tY + \hat{\beta}^tX^tX\,[(X^tX)^{-1}X^tY] \quad (\text{puisque } Y^tX\,\hat{\beta} \text{ est un scalaire})$$

$$= Y^tY - \hat{\beta}^tX^tY$$

$$= Y^tY - Y^tX\,(X^tX)^{-1}X^tY$$

$$= Y^t\left[I_n - X\,(X^tX)^{-1}X^t\right]Y\,.$$

□

<u>Proposition 1.3</u>

Soit Y un vecteur aléatoire de dimension $n \times 1$. Supposons que $\mathrm{E}(Y) = \mu$ et que $\mathrm{Cov}(Y) = \mathrm{E}\left[(Y-\mu)\,(Y-\mu)^t\right] = \mathrm{E}\,(Y\,Y^t) - \mu\,\mu^t = \Sigma$. Alors $\mathrm{E}(Y^t A\,Y) = \mathrm{tr}(A\Sigma) + \mu^t A\,\mu$.

Démonstration :

$$\mathrm{E}\,(Y^t A\,Y) = \mathrm{E}\left[\mathrm{tr}(Y^t A\,Y)\right]$$

$$= \mathrm{E}\left[\mathrm{tr}\,(A\,Y\,Y^t)\right] \qquad \text{(propriété (ii))}$$

$$= \mathrm{tr}\left[\mathrm{E}\,(A\,Y\,Y^t)\right]$$

$$= \mathrm{tr}\left[A\,\left(\mathrm{E}\,(Y\,Y^t)\right)\right]$$

$$= \mathrm{tr}\left[A\,(\Sigma + \mu\,\mu^t)\right]$$

$$= \mathrm{tr}(A\Sigma) + \mathrm{tr}(A\,\mu\,\mu^t)$$

$$= \mathrm{tr}(A\Sigma) + \mathrm{tr}(\mu^t A\,\mu) \qquad \text{(propriété (ii))}$$

$$= \mathrm{tr}(A\Sigma) + \mu^t A\,\mu$$

□

1.2.3 Distributions des formes quadratiques

Il est bien connu que si les variables aléatoires Y_i, $i = 1,\ldots,r$, sont indépendantes et identiquement distribuées de loi $N(0,1)$ alors $\sum_{i=1}^{r} Y_i^2 \sim \chi^2(r)$ où $\chi^2(r)$ est la loi d'une variable distribuée suivant une loi du khi-deux avec r degrés de liberté. Sous forme vectorielle, on dira que si $Y \sim N_r(0_r, I_r)$ alors $Y'Y \sim \chi^2(r)$.

Considérons maintenant la distribution de $Y'Y$ quand $Y \sim N_r(\mu_r, I_r)$. Une démonstration du théorème qui suit est présentée dans GRAYBILL [1976, p.125].

Théorème 1.4
Si $Y \sim N_r(\mu_r, I_r)$, alors $Y'Y \sim \chi^2(r,\lambda)$ où $\chi^2(r,\lambda)$ est la loi d'une variable distribuée suivant une loi du khi-deux avec r degrés de liberté et paramètre de décentralité $\lambda = \mu'\mu / 2$.
Théorème 1.5
Soit $Y \sim N_r(\mu,\Sigma)$. Si $A\Sigma$ est idempotente alors, pour tout μ, $Y'A\,Y \sim \chi^2(r,\lambda)$ où r désigne le rang de A et $\lambda = \mu' A\,\mu / 2$.

Démonstration :
Supposons que $A\Sigma$ soit idempotente et posons $\Sigma = C'C$ où C est non singulière.

Alors, puisque $C\,A\,C'C\,A\,C' = C\,(A\,\Sigma\,A\,\Sigma)C^{-1} = C\,(A\,\Sigma)C^{-1} = C\,A\,C'$, la matrice $B = C\,A\,C'$ est idempotente et, en raison de la propriété (i), est de rang $r = \mathrm{rg}(A)$.

En vertu de la propriété (vi), il existe une matrice orthogonale P de dimension $n \times n$ telle que
$$P'B\,P = P'C\,A\,C'P = \begin{bmatrix} I_r & 0_{r\times(n-r)} \\ 0_{(n-r)\times r} & 0_{(n-r)\times(n-r)} \end{bmatrix}.$$

Définissons $Z = P'(C')^{-1}Y$.
Alors, $\mathrm{E}(Z) = P'(C')^{-1}\mu$ et $\mathrm{Cov}(Z) = [P'(C')^{-1}]\ \mathrm{Cov}(Y)\ [P'(C')^{-1}]' = I_n$.

Partitionnons le vecteur aléatoire Z de sorte que $Z' = (Z_1', Z_2')$ où Z_1 est un vecteur de dimension $r \times 1$.

Alors, $Z_1 \sim N_r(\theta_r, I_r)$ où $\theta_r = (1_r, 0_{n-r}) P^t (C^t)^{-1} \mu$. Ici, $(1_r, 0_{n-r})$ représente un vecteur de dimension $1 \times n$ où 1_r est un vecteur de format $1 \times r$ composé de 1 et 0_{n-r} est un vecteur de dimension $1 \times (n-r)$ composé de 0.

Puisque P est orthogonale on a $Y = C^t (P^t)^{-1} Z = C^t P Z$ et donc :

$$Y^t A\ Y = (C^t P Z)^t A\ (C^t P Z) = Z^t (P^t C\ A\ C^t P) Z = (Z_1^t, Z_2^t) \begin{bmatrix} I_r & 0_{r \times (n-r)} \\ 0_{(n-r) \times r} & 0_{(n-r) \times (n-r)} \end{bmatrix} \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix}.$$

Ainsi $Y^t A\ Y = Z_1^t Z_1$ qui, par le théorème précédent, est distribué suivant une loi $\chi^2(r, \lambda)$ où

$$\begin{aligned} \lambda = \theta^t \theta\ / 2 &= (P^t (C^t)^{-1} \mu)^t (1_r, 0_{n-r})^t (1_r, 0_{n-r})\ (P^t (C^t)^{-1} \mu)\ / 2 \\ &= (P^t (C^t)^{-1} \mu)^t P^t B\ P\ (P^t (C^t)^{-1} \mu)\ / 2 \\ &= \mu^t C^{-1} (P\ P^t) B\ (P\ P^t) (C^t)^{-1} \mu\ / 2 \\ &= \mu^t C^{-1} B\ (C^t)^{-1} \mu\ / 2 \qquad \text{(propriété (iv))} \\ &= \mu^t C^{-1} C\ A\ C^t (C^t)^{-1} \mu\ / 2 \\ &= \mu^t A\ \mu\ / 2 \end{aligned}$$

□

Corollaire 1.6

Si $Y \sim N_n(X\beta, \sigma^2 I_n)$, alors $SSR / \sigma^2 \sim \chi^2(k-1, (X\beta)^t (I_n - \frac{1}{n} 1_n 1_n^t) (X\beta) / 2\sigma^2)$ et $SSE / \sigma^2 \sim \chi^2(n-k)$.

Démonstration :

Soit $\Sigma = \sigma^2 I_n$ et $SSR / \sigma^2 = Y^t A\ Y$ où $A = (X\ (X^t X)^{-1} X^t - \frac{1}{n} 1_n\ 1_n^t) / \sigma^2$.

D'après le théorème précédent, il suffit de montrer que $A\Sigma$ est idempotente, que $\mathrm{rg}(A) = k-1$ et que le paramètre de décentralité est égal à $(X\beta)^t (I_n - \frac{1}{n} 1_n 1_n^t) (X\beta)) / 2\sigma^2$.

Remarquons tout d'abord que $1_n^t 1_n = n$.

De plus, puisque $X\ (X^t X)^{-1} X^t X = X$ et que la première colonne de X est 1_n, on a : $X\ (X^t X)^{-1} X^t 1_n = 1_n$.

De la même façon, $X^t X\ (X^t X)^{-1} X^t = X^t$ entraîne que $1_n^t X\ (X^t X)^{-1} X^t = 1_n^t$.

Ainsi, $A\Sigma A\Sigma = (X\ (X^tX)^{-1}X^t - \frac{1}{n}1_n\ 1_n^t)(X\ (X^tX)^{-1}X^t - \frac{1}{n}1_n\ 1_n^t)$

$= X\ (X^tX)^{-1}X^tX\ (X^tX)^{-1}X^t + \frac{1}{n^2}1_n\ 1_n^t 1_n\ 1_n^t$

$\quad - \frac{1}{n}X\ (X^tX)^{-1}X^t 1_n\ 1_n^t - \frac{1}{n}1_n\ 1_n^t X\ (X^tX)^{-1}X^t$

$= A\Sigma.$

Donc $A\Sigma$ est idempotente.

Remarquons maintenant que, en vertu de la propriété (i), $\mathrm{rg}(A) = \mathrm{rg}(A\Sigma)$.

De plus, comme $A\Sigma$ est idempotente alors $\mathrm{rg}(A\Sigma) = \mathrm{tr}(A\Sigma)$ en raison de la propriété (iii).

Ainsi, $\mathrm{rg}(A) = \mathrm{tr}(X\ (X^tX)^{-1}X^t - \frac{1}{n}1_n\ 1_n^t)$

$= \mathrm{tr}(X^tX\ (X^tX)^{-1}) - \mathrm{tr}(\frac{1}{n}1_n\ 1_n^t)$ (propriété (ii))

$= \mathrm{tr}(I_k) - 1$

$= k - 1.$

Enfin, $\lambda = (X\beta)^t A\ (X\beta)\ /\ 2 = (\beta^t X^t)\left[(X\ (X^tX)^{-1}X^t - \frac{1}{n}1_n\ 1_n^t)/\sigma^2\right](X\beta)\ /\ 2$

$= (X\beta)^t(I_n - \frac{1}{n}1_n\ 1_n^t)(X\beta)\ /\ 2\sigma^2.$

Ainsi, par le théorème précédent, SSR/σ^2 suit une loi du khi-deux décentrée (à $k-1$ degrés de liberté) et puisqu'on ne possède aucune information particulière sur β ou sur X, $\lambda \neq 0$.

D'autre part, on a $SSE/\sigma^2 = Y^t B\ Y$ où $B = (I_n - X\ (X^tX)^{-1}X^t)/\sigma^2$.
Soit $\Sigma = \sigma^2 I_n$ et montrons maintenant que $B\Sigma$ est idempotente.

$B\Sigma B\Sigma = (I_n - X\ (X^tX)^{-1}X^t)(I_n - X\ (X^tX)^{-1}X^t)$

$= I_n + X\ (X^tX)^{-1}X^tX\ (X^tX)^{-1}X^t - X\ (X^tX)^{-1}X^t - X\ (X^tX)^{-1}X^t$

$= B\Sigma.$

De plus, comme précédemment, on utilise successivement les propriétés (i), (iii), puis (i) à nouveau et on a:

$$\mathrm{rg}(B) = \mathrm{rg}(B\Sigma) = \mathrm{tr}(B\Sigma) = \mathrm{tr}(I_n - X\ (X^tX)^{-1}X^t) = \mathrm{tr}(I_n) - \mathrm{tr}(X\ (X^tX)^{-1}X^t) = n - k.$$

Enfin, sans aucune hypothèse ou information particulière autant sur β que sur X, on utilise le fait que $X^t(I_n - X\ (X^tX)^{-1}X^t) = 0$ pour déduire que :

$$\lambda = (X\beta)^t B\ (X\beta)\ /\ 2 = (\beta^t X^t)\left[(I_n - X\ (X^t X\ )^{-1} X^t)/\sigma^2\right](X\beta)\ /\ 2 = 0.$$

Ainsi, par le théorème précédent, SSE / σ^2 suit une loi du khi-deux centrée à n-k degrés de liberté. □

Théorème 1.7

Soit $Y \sim N_n(\mu, \Sigma)$ et soient A et B deux matrices de dimension $n \times n$ telles que $A\Sigma B = 0$. Alors, pour toutes valeurs de μ, $Y^t A\ Y$ et $Y^t B\ Y$ sont des variables aléatoires indépendantes.

Démonstration :

Posons $\Sigma = C^t C$ où C n'est pas singulière.

Comme $A\Sigma B = 0$, alors $A\ C^t\ C\ B\ = 0$ et donc $(C\ A\ C^t)(C\ B\ C^t) = 0$.

Posons $Q = C\ A\ C^t$ et $T = C\ B\ C^t$.

En vertu de la propriété (v), comme les formes quadratiques Q et T sont symétriques (puisque A et B le sont) et que QT l'est également (car $QT = 0$), il existe une matrice orthogonale P telle que $P^t\ Q\ P$ et $P^t\ T\ P$ sont diagonales.

De plus, puisque $QT = 0$, les éléments non nuls sur les diagonales de $P^t\ Q\ P$ et $P^t\ T\ P$ doivent apparaître à des positions distinctes, autrement on a :

$$P^t\ Q\ P = \begin{bmatrix} D_1 & 0 \\ 0 & 0 \end{bmatrix} \text{ et } P^t\ T\ P = \begin{bmatrix} 0 & 0 \\ 0 & D_2 \end{bmatrix}.$$

Soit $Z = P^t (C^t)^{-1} Y$ un vecteur de $\mathbb{R}^n$.

Partitionnons Z de sorte que $Z^t = (Z_1{}^t, Z_2{}^t)$ où Z_1 et Z_2 sont de dimension respective n_1 et $n - n_1$.

Ainsi,

$$Y^t A\ Y = (C^t P Z)^t A\ (C^t P Z) = Z^t (P^t C\ A\ C^t P\ ) Z = (Z_1{}^t, Z_2{}^t) \begin{bmatrix} D_1 & 0 \\ 0 & 0 \end{bmatrix} \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} = Z_1^t D_1 Z_1.$$

De façon semblable, $Y^t B\ Y = Z_2^t D_2 Z_2$.

Ainsi, $Y^t A\ Y$ ne dépend que des n_1 premiers éléments de Z et $Y^t B\ Y$ ne dépend que des $n - n_1$ derniers éléments de Z.

Puisque $Z \sim N_n(P^t(C^t)^{-1}\mu\ , I_n)$, Z_1 et Z_2 sont indépendants, donc il en va de même pour $Y^t A\ Y$ et $Y^t B\ Y$. □

Corollaire 1.8
Soit $Y \sim N_n(X\beta, \sigma^2 I_n)$. Alors SSR/σ^2 et SSE/σ^2 sont des variables aléatoires indépendantes.

Démonstration :
Soit $\Sigma = \sigma^2 I_n$.

On a $SSR/\sigma^2 = Y^t A\ Y$ où $A = (X\ (X^tX)^{-1}X^t - \frac{1}{n}1_n\ 1_n^t)/\sigma^2$ et $SSE/\sigma^2 = Y^t B\ Y$ où $B = (I_n - X\ (X^tX)^{-1}X^t)/\sigma^2$.

D'après le théorème précédent, il suffit de montrer que $A\Sigma B = 0$.

Puisque $1_n^t\ X\ (X^tX)^{-1}X^t = 1_n^t$ (1_n étant la première colonne de X) on a:

$$\begin{aligned}(X\ (X^tX)^{-1}X^t - \tfrac{1}{n}1_n\ 1_n^t)(I_n - X\ (X^tX)^{-1}X^t) \\ = X\ (X^tX)^{-1}X^t + \tfrac{1}{n}1_n\ 1_n^t X\ (X^tX)^{-1}X^t - \tfrac{1}{n}1_n\ 1_n^t - X\ (X^tX)^{-1}X^t X\ (X^tX)^{-1}X^t \\ = 0.\end{aligned}$$ □

Le théorème suivant, énoncé sans démonstration, porte sur le quotient de deux lois indépendantes du khi-deux.

Théorème 1.9
Si $Z_1 \sim \chi^2(r_1,\lambda)$ et $Z_2 \sim \chi^2(r_2)$ où Z_1 et Z_2 sont des variables aléatoires indépendantes, alors $F = \dfrac{Z_1/r_1}{Z_2/r_2} \sim F(r_1,r_2,\lambda)$ où $F(r_1,r_2,\lambda)$ est la loi d'une variable distribuée suivant une loi de Fisher avec r_1 et r_2 degrés de liberté et paramètre de décentralité λ.

Corollaire 1.10

Soit un modèle de régression linéaire à k paramètres incluant un terme constant β_0. Alors, sous l'hypothèse nulle $H_0: \beta_1 = \beta_2 = \ldots = \beta_{k-1} = 0$, $F = \dfrac{SSR/(k-1)}{SSE/(n-k)} \sim F(k-1, n-k)$.

Démonstration :

D'après les corollaires 1.6 et 1.8, $SSR / \sigma^2 \sim \chi^2(k-1, (X\beta)^t (I_n - \frac{1}{n} 1_n 1_n^t) (X\beta) / 2\sigma^2)$ indépendamment de $SSE / \sigma^2 \sim \chi^2(n-k)$.

D'après le théorème précédent, $\dfrac{SSR/(k-1)}{SSE/(n-k)} \sim F(k-1, n-k, \lambda)$.

Il suffit donc de montrer que, sous H_0, $\lambda = (X\beta)^t (I_n - \frac{1}{n} 1_n 1_n^t) (X\beta) / 2\sigma^2 = 0$.

Décomposons la matrice X sous la forme $X = (1_n, \tilde{X})$ et le vecteur β sous la forme $\beta = \begin{pmatrix} \beta_0 \\ \tilde{\beta} \end{pmatrix}$.

Si $\tilde{\beta} = 0_{k-1}$ alors $X\beta = (1_n, \tilde{X}) \begin{pmatrix} \beta_0 \\ 0_{k-1} \end{pmatrix} = 1_n \beta_0$.

Ainsi, puisque $1_n^t 1_n = n$, il vient :

$\lambda = (1_n \beta_0)^t (I_n - \frac{1}{n} 1_n 1_n^t) (1_n \beta_0) / 2\sigma^2 = (\beta_0^t 1_n^t 1_n \beta_0 - \frac{1}{n} \beta_0^t 1_n 1_n^t 1_n 1_n^t \beta_0) / 2\sigma^2 = 0$.

D'où $\dfrac{SSR/(k-1)}{SSE/(n-k)} \sim F(k-1, n-k)$. □

1.2.4 Ajout de présupposés pour tester des hypothèses

Si l'on admet que les perturbations aléatoires $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ sont indépendantes et identiquement distribuées de loi $N(0, \sigma^2)$ (supposition que l'on se doit de vérifier en pratique), on peut tester certaines hypothèses. Les hypothèses les plus intéressantes lorsque l'on modélise des données sont celles qui nous assurent, à un niveau donné, de la "qualité" du modèle utilisé. Rappelons qu'un test de niveau $\alpha \in [0,1]$ fait tout simplement allusion au cas où l'expérimentateur est prêt à se tromper avec une probabilité α lorsque l'hypothèse nulle est vraie.

Un test primordial en pratique est celui qui valide la pertinence de l'apport des variables explicatives dans l'explication de la variable dépendante. On dit alors que l'on teste si la régression est statistiquement significative. L'hypothèse nulle (notée H_0) consiste à se demander si le modèle constant est suffisant à expliquer la variable dépendante. En d'autres termes, on confronte :

$$H_0: \beta_1 = \beta_2 = \ldots = \beta_{k-1} = 0$$

à H_1: les β_i, $i = 1,..,k-1$, ne sont pas tous nuls.

Pour ce, on se réfère au Tableau 1.1 et on considère le quotient des quantités MSR et MSE qui, sous H_0, suit une loi de Fisher $F(k-1,n-k)$ (voir le corollaire 1.10). Soit $F_\alpha(k-1,n-k)$ le quantile d'ordre $1-\alpha$ de cette loi, alors si $F = \frac{MSR}{MSE} > F_\alpha(k-1,n-k)$ on rejette l'hypothèse nulle; on en vient donc à la conclusion que la régression est statistiquement significative. Notons que pour un nombre donné de paramètres, moins on dispose d'observations, plus il est difficile de vérifier si la régression est statistiquement significative.

Par exemple, supposons que l'on mesure 10 variables et dans un premier temps supposons que nous disposions que de 12 observations. Alors, pour que la régression soit significative au niveau $\alpha = 0.05$ la valeur expérimentale de la statistique F doit être supérieure à 241.9 (à savoir $F_{0.05}(10,1)$) ou de façon identique $(SSR/SSE) > 2419$. Au même niveau, avec 71 observations, il suffit que la statistique excède 1.99, i.e. $(SSR/SSE) > 0.3317$.

Le test précédent n'est valide, cependant, que si le manque d'ajustement n'est pas significatif. À ce sujet, DRAPER & SMITH [14] mentionnent qu'il faut alors utiliser le rapport $MSLF/MSPE$ (voir Tableau 1.2) qui, sous H_0 (manque d'ajustement), suit une loi de Fisher $F(I-k,n-I)$. Notons que s'il n'y a pas de manque d'ajustement, alors $MSE = SSE/(n-k) = (SSLF + SSPE)/(n-k)$ (souvent appelé s^2) est un estimateur sans biais de σ^2. Si le manque d'ajustement ne peut être testé, l'utilisation de s^2 comme estimateur de σ^2 implique l'hypothèse que le modèle est adéquat. Si le modèle n'est pas adéquat, s^2 sera habituellement trop grande puisque c'est une variable aléatoire avec une moyenne supérieure à σ^2. Cependant, il est possible que s^2 soit trop petite du à la fluctuation dans l'échantillonage. Cette fois, si $F = \frac{MSLF}{MSPE} < F_\alpha(I-k,n-I)$, on ne rejette pas l'hypothèse nulle et donc on en conclut que le manque d'ajustement n'est pas significatif.

Une fois que l'on s'est assuré que le modèle ne se ramène pas au modèle constant (et que le manque d'ajustement n'est pas significatif), on peut être tenté de vouloir simplifier notre modèle original sans encourir une perte trop considérable en ce qui a trait à l'explication de la variable dépendante. Pour ce, on voudra vérifier si, en enlevant certaines variables "non significatives", la somme des carrés des erreurs n'augmente pas trop. Soient les modèles suivants:

(1) $Y_i = \beta_0 + \sum_{i=1}^{k-r-1} \beta_i x_i + \sum_{i=k-r}^{k-1} \beta_i x_i + \varepsilon_i$ (modèle à k paramètres)

(2) $Y_i = \beta_0 + \sum_{i=1}^{k-r-1} \beta_i x_i + \varepsilon_i$ (modèle à k-r paramètres)

Le modèle (1) est celui de départ et le modèle (2) est ce même modèle auquel on retire r variables données (ce qui revient à annuler r paramètres parmi $\beta_{k-r},\beta_{k-r+1},\ldots,\beta_{k-1}$). Notons qu'ici, $\beta_{k-r},\beta_{k-r+1},\ldots,\beta_{k-1}$ désignent r paramètres parmi $\beta_1,\beta_2,\ldots,\beta_{k-1}$ sans nécessairement que ce soient les r derniers.

Pour ce test d'hypothèse, on considérera l'hypothèse nulle $H_0:\beta_{k-r} = \beta_{k-r+1} = \ldots = \beta_{k-1} = 0$ et en se référant au Tableau 1.3, on considérera le quotient MSR_r^*/MSE. On peut remarquer que si on enlève r = k-1 paramètres à un modèle à k paramètres, la statistique $F = \dfrac{MSR_r^*}{MSE} = \dfrac{(SSR_k - SSR_{k-r})/r}{MSE}$ peut servir à tester si la régression est statistiquement significative. En effet, puisque $SSR_1 = 0$ (car $\hat{y}_i = \bar{y}$, $i = 1,\ldots,n$, dans le modèle constant) on retrouve la statistique $F = \dfrac{SSR_k/r}{MSE} = \dfrac{MSR}{MSE}$.

Plutôt que d'utiliser la statistique $F = \dfrac{(SSR_k - SSR_{k-r})/r}{MSE}$, il est plus habituel de considérer la statistique équivalente $F = \dfrac{(SSE_{k-r} - SSE_k)/r}{MSE}$ (l'équivalence découlant du fait que $SSE_{k-r} + SSR_{k-r} = SST = SSE_k + SSR_k$). Cette forme rend plus évidente l'exploitation de la différence entre la somme des carrés des erreurs entre les modèles à k-r et k paramètres; si cette différence est faible on s'attend à accepter l'hypothèse nulle. Puisque, sous H_0, cette dernière statistique F suit une loi de Fisher $F(r,n-k)$ on rejettera l'hypothèse nulle si $F > F_\alpha(r,n-k)$. Ici, rejeter l'hypothèse nulle signifie, qu'au niveau considéré, le modèle réduit (à k-r paramètres) est suffisant afin d'expliquer la variable dépendante.

1.2.5 Intervalles de confiance et de prédiction

a) Intervalles de confiance de la valeur moyenne de Y

La valeur $\hat{Y} = \hat{\beta}_0 + \sum_{i=1}^{k-1} \hat{\beta}_i x_i$ est une estimation de $E(Y) = \beta_0 + \sum_{i=1}^{k-1} \beta_i x_i$.

Supposons que l'on désire un intervalle de confiance de la combinaison de paramètres $a'\beta$ où $a = (1, a_1, \ldots, a_{k-1})'$. Un estimateur ponctuel sans biais de $a'\beta$ est $\hat{Y} = a'\hat{\beta}$

De plus, comme $\mathrm{Cov}(Y) = \sigma^2 I_n$,

$$\mathrm{Cov}(\hat{\beta}) = \mathrm{Cov}[(X'X)^{-1}X'Y] = [(X'X)^{-1}X'] \mathrm{Cov}(Y) [X (X'X)^{-1}] = \sigma^2 (X'X)^{-1}.$$

On a alors:

$$\mathrm{Var}(\hat{Y}) = \mathrm{Cov}(\hat{Y}) = \mathrm{Cov}(a'\hat{\beta}) = a' \mathrm{Cov}(\hat{\beta}) a = \sigma^2 [a'(X'X)^{-1} a].$$

En pratique, on estimera σ^2 par $s^2 = SSE / (n-k)$ de sorte qu'un intervalle de confiance au niveau $100(1-\alpha)\%$ pour la valeur moyenne de Y sera donnée par:

$\hat{Y} \pm s \cdot t_{\alpha/2}(n-k) \cdot \sqrt{a'(X'X)^{-1}a}$ où $t_{\alpha/2}(n-k)$ représente le quantile d'ordre $1-\alpha/2$ de la loi de Student à n-k degrés de liberté.

b) Intervalle de prédiction pour Y

Tout d'abord, il est important de réaliser que Y est une variable aléatoire et non un paramètre et prédire sa valeur n'a rien à voir avec la prédiction d'un paramètre (ou d'une combinaison linéaire de paramètres). Afin de prédire une valeur de Y dans le futur, on utilisera le même estimateur que pour $E(Y)$ à savoir $\hat{Y}$ (que l'on notera $\hat{Y}_{n+1}$). La notation $\hat{Y}_{n+1}$ sert à mieux réaliser que cette observation future s'ajoute aux n observations actuelles. L'erreur que l'on fait en prédisant une valeur Y_{n+1} par $\hat{Y}_{n+1}$ est donnée par $\varepsilon_{n+1} = Y_{n+1} - \hat{Y}_{n+1}$.

On a alors :

$$E(\varepsilon_{n+1}) = E(Y_{n+1}) - E(\hat{Y}_{n+1}) = 0$$

et

$$Var(\varepsilon_{n+1}) = Var(Y_{n+1}) + Var(\hat{Y}_{n+1}) - 2\,Cov(Y_{n+1}, \hat{Y}_{n+1}).$$

Puisque Y_{n+1} est dans le futur et que $\hat{Y}_{n+1}$ est fonction des observations actuelles, les deux variables sont indépendantes et donc $Cov(Y_{n+1}, \hat{Y}_{n+1}) = 0$. Ainsi,

$$Var(\varepsilon_{n+1}) = Var(Y_{n+1}) + Var(\hat{Y}_{n+1}) = \sigma^2 + \sigma^2[\, a^t (\, X^t X\,)^{-1} a\,] = \sigma^2[\,1 + a^t (\, X^t X\,)^{-1} a\,].$$

Encore une fois, on estimera σ^2 par $s^2 = SSE\,/\,(n-k)$ de sorte qu'un intervalle de prédiction au niveau $100(1-\alpha)\%$ pour Y est donné par:

$\hat{Y} \pm s \cdot t_{\alpha/2}(n-k) \cdot \sqrt{1 + a^t (\, X^t X\,)^{-1} a}$ où $t_{\alpha/2}(n-k)$ représente le quantile d'ordre $1-\alpha/2$ de la loi de Student à n-k degrés de liberté.

On peut noter que toutes choses étant égales par ailleurs, un intervalle de prédiction sera plus long (et donc moins précis) qu'un intervalle de confiance.

CHAPITRE 2

CRITÈRES DE SÉLECTION DU "MEILLEUR" SOUS-ENSEMBLE DE VARIABLES

Plusieurs raisons justifient la réduction du nombre de variables; par exemple, MILLER [38] en énumère quelques-unes, sans ordre particulier. Notons que ces objectifs ne sont pas complètement compatibles.

- Estimer ou prédire à moindres coûts en réduisant le nombre de variables pour lesquelles on collecte des données.
- Prédire avec précision en éliminant des variables n'apportant peu ou pas d'information.
- Décrire plus simplement les données (de façon à faciliter une interprétation éventuelle).
- Estimer les coefficients de la régression avec de petites erreurs quadratiques moyennes (particulièrement lorsque certaines variables indépendantes sont grandement corrélées).

Même si certaines raisons peuvent justifier le retrait de variables, il faut quand même être conscient de l'implication de cette modification sur le modèle initial. Voici quelques points dont il faut être conscient avant d'éliminer des variables :

- La sélection de variables redéfinit le modèle original. Ceci peut faire en sorte qu'on ne réponde plus à la question initiale à partir de laquelle le choix des variables à recueillir a été établi.
- La sélection de variables optimise l'ajustement du modèle en fonction de l'échantillon utilisé ce qui a pour effet de surévaluer presqu'assurément la précision des résultats. Il y a donc un risque très élevé de tirer profit des caractéristiques de l'échantillon résultant du hasard.
- La sélection de variables automatisée ne conservera pas conjointement les variables définies en groupes. On peut penser, entre autres, à une variable catégorique (n choix de

réponses) résumée en $n-1$ variables dichotomiques ("dummy variables"). Il incombera donc à l'analyste des données de conserver ces variables ensemble.

Dans le but de rechercher un sous-ensemble "optimal", certains critères peuvent être considérés. Nous présenterons trois critères parmi les plus utilisés: à savoir le coefficient de détermination, le coefficient de détermination corrigé et le coefficient C_p de Mallows. On ne peut espérer qu'un critère donné fournisse en tout temps le "meilleur" sous-ensemble; c'est pourquoi il en existe toute une panoplie. En comptant tous les critères suggérés au cours des années on peut estimer ce nombre à plus de cinquante. Il semble plus important de bien maîtriser les critères principaux. Dans ce chapitre, on examinera certaines propriétés de ces critères et on exposera également quelques uns de leurs inconvénients. Enfin, on présentera quelques liens à établir entre ces critères et on mentionnera dans quelles circonstances un critère est plus approprié qu'un autre.

2.1 Coefficient de détermination (R^2)

Le coefficient de détermination (ou R^2) représente la proportion de la variation totale par rapport à la moyenne $\overline{Y}$ qui est expliquée par la régression. Clairement, les bornes du critère sont 0 et 1. Une valeur de 1, pour une régression donnée, signifie que celle-ci explique toute la variabilité des différentes valeurs prises par la variable dépendante Y. À l'opposé, une valeur de 0 indique qu'aucune variabilité n'est expliquée.

2.1.1 Définition

Sans aucune information sur les variables indépendantes, la meilleure prédiction pour une valeur future de la variable dépendante Y est $\overline{Y}$, la moyenne des n observations dont on dispose. On appellera ces n valeurs de Y les valeurs actuelles (pour établir un contraste avec les valeurs futures). Pour une valeur actuelle donnée, Y_i, l'erreur commise en utilisant $\overline{Y}$ comme prédicteur est simplement $Y_i - \overline{Y}$. Une mesure de la variabilité totale sera donnée par $SST = \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2$.

En exploitant l'information sur les variables indépendantes, le meilleur estimateur d'une valeur future pour Y est $\hat{Y}$. Cette fois, une mesure de la variation sur Y qui subsiste après l'ajustement de l'équation de régression est donnée par $SSE = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$.

La proportion de la variabilité qui n'est pas expliquée par la régression est donnée par SSE/SST. Ainsi, la proportion de la variation totale (par rapport à $\overline{Y}$) qui est expliquée par la régression est donnée par :

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2} \quad . \tag{1}$$

Une autre façon de saisir le critère consiste à remarquer que R^2 représente le carré du coefficient de corrélation entre le vecteur de la variable expliquée et le vecteur des estimations de celle-ci i.e.:

$$R^2 = \frac{\left[\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)\left(\hat{Y}_i - \overline{\hat{Y}}\right)\right]^2}{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2 \sum_{i=1}^{n}\left(\hat{Y}_i - \overline{\hat{Y}}\right)^2} \quad . \tag{2}$$

Proposition 2.1

Dans le cadre de la régression linéaire avec terme constant, les quantités (1) et (2) sont égales.

Démonstration :

Posons $H = X\left(X'X\right)^{-1}X'$ et remarquons que $H' = H$ et que $H^2 = H$.

On a:

$\hat{Y} = X\hat{\beta} = X\left(X'X\right)^{-1}X'Y = H\,Y$ et $\varepsilon = Y - \hat{Y} = Y - HY = (I - H)Y$.

De plus, on remarque que :

$$\sum_{i=1}^{n}\hat{Y}_i\varepsilon_i = \hat{Y}'\varepsilon = (HY)'(I-H)Y = Y'[H'(I - H)]Y = Y'[H\,(I - H)]Y = Y'[H - H^2]Y = 0$$

En vertu du fait que $\sum_{i=1}^{n}\varepsilon_i = \sum_{i=1}^{n}(Y_i - \hat{Y}_i) = 0$ on a :

$$\sum_{i=1}^{n}\varepsilon_i(\hat{Y}_i-\overline{Y})=\sum_{i=1}^{n}\hat{Y}_i\varepsilon_i-\overline{Y}\sum_{i=1}^{n}\varepsilon_i=0\ .$$

En utilisant le même résultat, on peut affirmer également que $\hat{Y}_i-\overline{\hat{Y}}=\hat{Y}_i-\overline{Y}$.
Ainsi,

$$\sum_{i=1}^{n}(Y_i-\overline{Y})(\hat{Y}_i-\overline{\hat{Y}})=\sum_{i=1}^{n}(Y_i-\overline{Y})(\hat{Y}_i-\overline{Y})=\sum_{i=1}^{n}\Big((Y_i-\hat{Y}_i)+(\hat{Y}_i-\overline{Y})\Big)\Big(\hat{Y}_i-\overline{Y}\Big)=\sum_{i=1}^{n}\varepsilon_i(\hat{Y}_i-\overline{Y})+\sum_{i=1}^{n}(\hat{Y}_i-\overline{Y})^2\ .$$

D'où, $\sum_{i=1}^{n}(Y_i-\overline{Y})(\hat{Y}_i-\overline{\hat{Y}})=\sum_{i=1}^{n}(\hat{Y}_i-\overline{Y})^2$.

Donc, $$\frac{\left[\sum_{i=1}^{n}(Y_i-\overline{Y})(\hat{Y}_i-\overline{\hat{Y}})\right]^2}{\sum_{i=1}^{n}(Y_i-\overline{Y})^2\sum_{i=1}^{n}(\hat{Y}_i-\overline{\hat{Y}})^2}=\frac{\sum_{i=1}^{n}(\hat{Y}_i-\overline{Y})^2}{\sum_{i=1}^{n}(Y_i-\overline{Y})^2}=\frac{SSR}{SST}=1-\frac{SSE}{SST}\ .$$ □

La forme (1) est la plus répandue, probablement parce que celle-ci suggère d'extraire *SSE* et *SST* d'un tableau d'analyse de la variance.

Lorsque l'on voudra insister sur le fait qu'on utilise un modèle à k paramètres, on désignera le coefficient de détermination par :

$$R_k^2=1-\frac{SSE_k}{SST}$$

Rappelons qu'au sens des moindres carrés, on vise a minimiser SSE_k. Ainsi, cette dernière expression, où SSE_k est affecté d'un signe négatif, suggère de choisir un modèle ayant un coefficient de détermination le plus élevé possible, soit des valeurs près de 1. Ceci va de pair avec l'idée voulant qu'une "bonne" régression explique le plus possible la variabilité des différentes valeurs prises par la variable dépendante Y.

Notons que si l'approximation est parfaite, i.e. lorsque $Y_i=\hat{Y}_i,\ 1\le i\le n$, on aura $R^2=1$. Cette situation est peu probable en pratique. À l'opposé, si $Y_i=\overline{Y},\ 1\le i\le n$ (quand $\beta_1=\beta_2=\ldots=\beta_{k-1}=0$) alors $R_k^2=0$.

2.1.2 Propriétés

Voici quelques propriétés importantes du coefficient de détermination. Certains résultats ultérieurs seront établis à partir de ces propriétés.

Proposition 2.2
Soit $Y \sim N_n(X\beta, \sigma^2 I_n)$. Alors sous l'hypothèse $H_0: \beta_1 = \beta_2 = \ldots = \beta_{k-1} = 0$ on a:

i) $R_k^2 \sim \beta eta(\frac{k-1}{2}, \frac{n-k}{2})$, où $\beta eta(\frac{k-1}{2}, \frac{n-k}{2})$ est la loi d'une variable distribuée suivant une loi Beta de paramètres $\frac{k-1}{2}$ et $\frac{n-k}{2}$.

ii) $E(R_k^2) = \frac{k-1}{n-1}$.

Démonstration :

i) En vertu des corollaires 1.6 et 1.8, les quantités $\frac{SSR_k}{\sigma^2}$ et $\frac{SSE_k}{\sigma^2}$ sont indépendantes et de lois respectives $\chi^2(k-1)$ et $\chi^2(n-k)$.

Posons donc $U = \frac{SSR_k}{\sigma^2}$ et $V = \frac{SSE_k}{\sigma^2}$ et montrons que

$$\frac{U}{U+V} = \frac{SSR_k}{SSR_k + SSE_k} = 1 - \frac{SSE_k}{SST} = R_k^2 \sim \beta eta(\tfrac{k-1}{2}, \tfrac{n-k}{2}) .$$

En posant $Z = \frac{U}{U+V}$ et $T = V$, on établit facilement que le jacobien de la transformation inverse donne $\frac{t}{(1-z)^2}$.

Par indépendance, la densité conjointe de U et V se ramène au produit des deux densités de lois du khi-deux. On prendra soin d'écrire celle-ci en fonction de z et t .

Pour évaluer la distribution marginale de Z, on effectuera l'intégration de la densité conjointe (multipliée par le jacobien de la transformation inverse) sur toutes les valeurs de t .

On reconnaîtra alors sans peine la densité d'une loi $\beta eta(\frac{k-1}{2}, \frac{n-k}{2})$.

(ii) Ce résultat s'établit de façon triviale en exploitant le fait que si $W \sim \beta eta(\alpha,\beta)$ alors $E(W) = \frac{\alpha}{\alpha+\beta}$. □

La partie (i) de la proposition ci-haut nous fait réaliser que l'on peut tester si une régression est statistiquement significative à partir de la valeur observée du coefficient R^2 (dans la mesure où on dispose de tables de la loi βeta). Vu la plus grande disponibilité des tables de Fisher, il semble intéressant d'établir, sous $H_0: \beta_1 = \beta_2 = ... = \beta_{k-1} = 0$, une relation entre les statistiques F et R^2.

On a:

$$F = \frac{MSR}{MSE} = \frac{(SST - SSE)/(k-1)}{SSE/(n-k)} = \left(\frac{n-k}{k-1}\right)\left(\frac{SST}{SSE} - 1\right) = \left(\frac{n-k}{k-1}\right)\left(\frac{1}{1-R^2} - 1\right)$$

d'où :

$$R^2 = \frac{(k-1)F}{(k-1)F + (n-k)}.$$

On remarque sans peine que cette dernière quantité est une fonction croissante en F.

Ainsi, sous $H_0: \beta_1 = \beta_2 = ... = \beta_{k-1} = 0$, la régression est statistiquement significative au niveau α si

$$R^2 > \frac{(k-1)F_\alpha(k-1, n-k)}{(k-1)F_\alpha(k-1, n-k) + (n-k)}.$$

Proposition 2.3
Si on enlève r variables (r paramètres parmi $\beta_1,...,\beta_{k-1}$) à un modèle comportant k paramètres, alors $R_k^2 > R_{k-r}^2$.

Démonstration :
Il est clair que $SSE_k < SSE_{k-r}$ puisqu'en retirant des variables, celles-ci font alors partie du terme d'erreur et contribuent à l'augmentation de ce dernier.

Ainsi, on a les implications suivantes :

$$SSE_k < SSE_{k-r} \Leftrightarrow \frac{SSE_k}{SST} < \frac{SSE_{k-r}}{SST}$$

$$\Leftrightarrow 1 - \frac{SSE_k}{SST} > 1 - \frac{SSE_{k-r}}{SST}$$

$$\Leftrightarrow R_k^2 > R_{k-r}^2$$ □

Nous venons de voir que le coefficient de détermination diminue systématiquement lorsque l'on retire des variables. Ainsi, le modèle complet possède toujours la valeur de R^2 la plus élevée. Le jugement entre alors en cause si l'on veut sélectionner un bon sous-ensemble de variables. Un sous-ensemble de variables intéressant en est donc un qui possède une valeur du R^2 pas trop éloignée de celle du modèle complet, ce qui signifie une perte minime quant au pourcentage d'explication.

Comme le mentionne HOCKING [29], on peut également tracer le graphique de la plus grande valeur de R_k^2 en fonction de k. En reliant les points entre eux, on constatera évidemment la croissance de la courbe. Une pente abrupte entre deux points consécutifs indique un gain important au niveau du R^2. Pour décider du nombre de paramètres à retenir, on étudiera la courbe de la droite vers la gauche en débutant en $k = p$ (le modèle complet). On visera à déceler à partir de quelle valeur de k on peut observer une pente descendant particulièrement abruptement. Cette valeur de k correspond alors au nombre minimum de paramètres a retenir sans encourir une perte trop importante au niveau du R^2.

Remarquons que, dans un cas où le modèle possédant le R^2 le plus élevé parmi ceux à $p-1$ paramètres était inacceptable, la pente abrupte apparaîtrait immédiatement de façon à ne garder que le modèle complet. Notons également que parfois la valeur de k recherchée est difficilement discernable car le graphique dépend de l'échelle utilisée.

2.1.3 Inconvénients

a) Inconvénient relié à un petit nombre d'observations

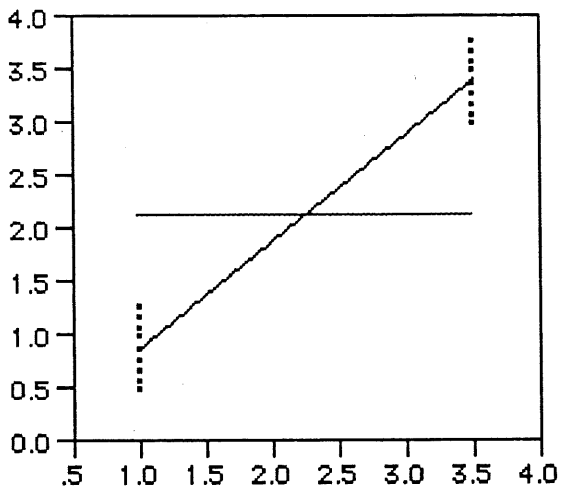
En utilisant l'égalité $E(R_k^2)=(k-1)/(n-1)$, CROCKER [11] fait remarquer qu'on peut obtenir des valeurs du coefficient de détermination aussi près de 1 que l'on veut. Pour ce faire, il suffit de choisir des valeurs de k près de n (la taille de l'échantillon). La valeur 1 représentant la borne supérieure du critère, $E(R_k^2) \approx 1$ signifie que tous les modèles ayant k paramètres ont sensiblement tous la valeur 1 comme valeur de R_k^2. Dans une telle situation, la pertinence de l'utilisation du critère du R^2 comme mesure de l'explication des variations de Y doit être remise en question.

En effet, d'une part, si pour chaque sous ensemble de variables la valeur du R^2 est élevée, il devient difficile, sinon impossible, de choisir le meilleur sous ensemble; d'autre part, on peut établir qu'alors les estimateurs des paramètres du modèle ne seront que très peu affectés par les variations des variables explicatives X_i, $i = 1,\ldots,p-1$ en ce sens où ils donneront presque toujours lieu à des valeurs "toutes égales". L'augmentation du nombre de paramètres (et donc de variables) produit une croissance artificielle du R^2 vers 1. Cependant, cette croissance a un effet négatif sur les autres aspects de la régression linéaire, en particulier sur l'estimation des paramètres.

b) Inconvénients reliés aux variations de la somme totale des carrés

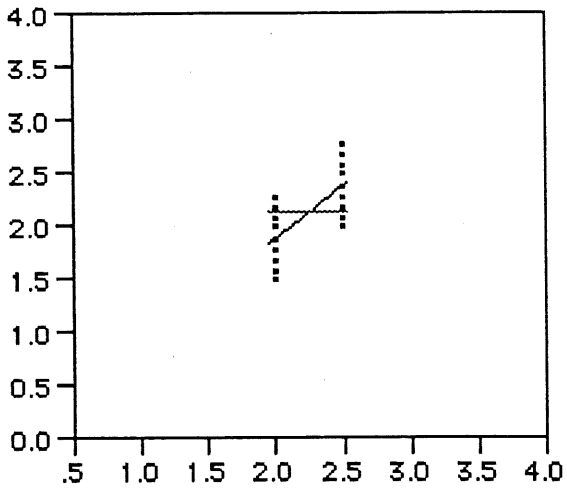
On peut remarquer que le R^2 est influencé par la dispersion des valeurs des variables indépendantes. En fait, moins les valeurs sont dispersées, i.e. moins il y a d'écart entre les valeurs, moins la valeur du R^2 est grande. Illustrons ce phénomène à l'aide d'une régression linéaire simple. Pour ce, on comparera deux ensembles (nuages) ayant le même nombre de points et la même valeur de SSE. Pour faciliter la comparaison, les deux ensembles auront la même moyenne pour la variable dépendante (notée $\bar{y}$) et seront tels qu'ils produisent exactement la même droite de régression.

Considérons les graphes suivants (les données apparaissent à l'annexe 3) :



$R^2 = 0.959$

Figure (a)



$R^2 = 0.484$

Figure (b)

Dans les deux cas, la droite oblique désigne la droite de régression dont l'équation est $\hat{Y} = -0.1 + x$ et la droite horizontale représente l'équation de la droite moyenne, i.e. la droite $\overline{Y} = 2.15$. En examinant les graphes, on voit que, dans la Figure (a), la valeur de $SST = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$ sera élevée puisque les valeurs y_i sont éloignées de $\overline{y}$. Dans la Figure (b), la proximité des valeurs y_i par rapport à $\overline{y}$ fait en sorte que SST donne une petite valeur. On comprend bien que pour des valeurs identiques de la somme des carrés due à l'erreur (SSE), la quantité $R^2 = 1 - \dfrac{SSE}{SST}$ fournit des valeurs plus faibles lorsque les valeurs de la variable dépendante sont voisines.

En fait, pour ce qui est de la régression linéaire simple, on peut se convaincre de l'impact de la dispersion de la variable indépendante en écrivant le coefficient de détermination en fonction de la mesure de dispersion $\sum_{i=1}^{n}(x_i - \overline{x})^2$. On notera cette quantité SST_X.

On considère donc, pour $1 \le i \le n$, le modèle $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$.

Ainsi, on a $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ et $\overline{Y} = \hat{\beta}_0 + \hat{\beta}_1 \overline{x}$.

Ceci entraîne que $\hat{Y}_i - \overline{Y} = (\hat{\beta}_0 + \hat{\beta}_1 x_i) - (\hat{\beta}_0 + \hat{\beta}_1 \overline{x}) = \hat{\beta}_1 (x_i - \overline{x})$, $1 \le i \le n$.

On a donc $SSR = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 = \hat{\beta}_1^{\,2}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \hat{\beta}_1^{\,2} SST_X$.

Ce qui nous permet de conclure que:

$$R^2 = 1 - \frac{SSE}{SST} = \frac{SSR}{SSR + SSE} = \frac{\hat{\beta}_1^{\,2} SST_X}{\hat{\beta}_1^{\,2} SST_X + SSE} = \frac{SST_X}{SST_X + d} \quad \text{où } d = SSE/\hat{\beta}_1^{\,2}.$$

On constate qu'en comparant deux cas où $\hat{\beta}_1$ et SSE sont identiques alors, peu importe β_0, seule la mesure de dispersion SST_X joue un rôle. Pour comprendre l'effet de la dispersion, examinons les équivalences qui suivent. Pour un cas comme dans la Figure (a) on indexera la lettre a tandis qu'un cas se rapportant à la Figure (b) sera indexé par b.

$$SST_{Xa} > SST_{Xb} \Leftrightarrow \frac{d}{SST_{Xb}} > \frac{d}{SST_{Xa}}$$

$$\Leftrightarrow 1 + \frac{d}{SST_{Xb}} > 1 + \frac{d}{SST_{Xa}}$$

$$\Leftrightarrow \frac{SST_{Xb} + d}{SST_{Xb}} > \frac{SST_{Xa} + d}{SST_{Xa}}$$

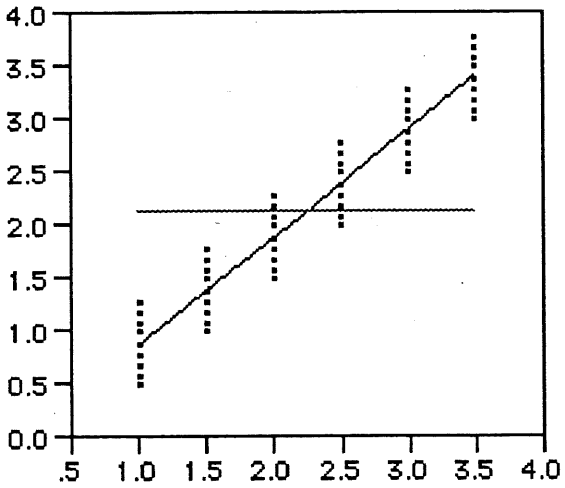
$$\Leftrightarrow \frac{SST_{Xa}}{SST_{Xa} + d} > \frac{SST_{Xb}}{SST_{Xb} + d}$$

$$\Leftrightarrow R_a^2 > R_b^2.$$

On remarque donc qu'une dispersion moindre entraîne une plus petite valeur du coefficient de détermination. Puisque l'on compare des ensembles ayant le même nombre d'observations, notons que comparer SST_{Xa} et SST_{Xb} revient à la comparaison des variances expérimentales $\frac{SST_{Xa}}{n-1}$ et $\frac{SST_{Xb}}{n-1}$.

On peut également reprendre l'idée précédente en raisonnant en termes d'étendue des variables indépendantes.

Considérons la figure suivante (tout comme précédemment les données apparaissent à l'annexe 3; il en sera de même pour les autres cas de figure de cette section):



$R^2 = 0.916$

Figure (c)

Supposons que nous n'ayons accès qu'aux 18 observations associées aux valeurs centrales, i.e. les 9 valeurs en x=2.0 et les 9 valeurs en x=2.5. On se retrouve alors exactement dans le cas de la Figure (b) et on observe alors une diminution assez importante au niveau de la valeur du R^2 qui est due à l'étendue, maintenant limitée, de la variable indépendante. La valeur de SSE est plus élevée d'un certain facteur γ dans un cas où l'étendue est grande comme dans la Figure (c) comparativement à un cas où on n'aurait accès qu'aux observations centrales (d'étendue limitée) comme dans la Figure (b). Il est évident dans le cas présent que γ=3. Par de simples manipulations, on montre que dans un cas comme celui-ci (où $SSE_c = \gamma SSE_b$), si $SST_{Xc} > \gamma SST_{Xb}$ on a $R_c^2 > R_b^2$. Sans effectuer de calculs, il est clair qu'en comparant les Figures (c) et (b), la valeur de SST_{Xc} est nettement supérieure à trois fois SST_{Xb}. L'idée d'étudier les répercussions de l'étendue a été amené par HAHN [22].

Une autre situation influençant la valeur du coefficient de détermination est l'inclinaison de la surface de régression. Cette observation à été notée par George Furnival et publiée par BARRETT [2]. La valeur du R^2 augmente avec "l'inclinaison" de la surface de régression. Illustrons ce phénomène, encore une fois, à l'aide d'une régression linéaire simple. Cet exemple est à la fois suffisamment simple pour effectuer les calculs sans avoir recours à un outil informatique et suffisamment général pour effectuer une infinité de comparaisons entre des ensembles d'observations. Considérons des ensembles ayant le même nombre

d'observations et la même valeur de SSE. Pour faciliter les comparaisons ceux-ci seront tels que $\bar{x} = \bar{y} = 0$.

Exemple

Soit trois observations associées aux valeurs (-2, -2tan θ), (1, 1+tan θ) et (1, -1+tan θ) où θ est un angle quelconque qui est supérieur ou égal à 0° et inférieur à 90°.

Alors $(\bar{x}, \bar{Y}) = (0,0)$.

De plus, $\hat{\beta}_1 = \dfrac{\sum_{i=1}^{3}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{3}(x_i - \bar{x})^2} = \dfrac{\sum_{i=1}^{3} x_i Y_i}{\sum_{i=1}^{3} x_i^{\,2}} = \tan\theta$ et $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x} = 0$.

Ainsi, $\hat{Y} = \tan\theta \cdot x$.

D'où,

$$SSE = \sum_{i=1}^{3}\left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{3}\left(Y_i - \tan\theta \cdot x_i\right)^2 = 2$$

et $R^2 = 1 - \dfrac{SSE}{SST} = 1 - \dfrac{2}{\sum_{i=1}^{3}(y_i - \bar{y})^2} = 1 - \dfrac{2}{\sum_{i=1}^{3} y_i^{\,2}} = 1 - \dfrac{2}{6\tan^2\theta + 2} = 1 - \dfrac{1}{3\tan^2\theta + 1}$.

On est donc en présence d'un cas où en faisant varier $\theta \in [0, \frac{\pi}{2}[$, chaque ensemble d'observations a le même centre de gravité ($(\bar{x}, \bar{Y}) = (0,0)$) et la même valeur de SSE (à savoir 2). De plus, on remarque que la droite de régression, $\hat{Y} = \tan\theta \cdot x$, admet tan θ (une fonction croissante en θ) comme pente, passe par l'origine et que θ correspond à l'angle formé entre cette droite et l'axe des abscisses.

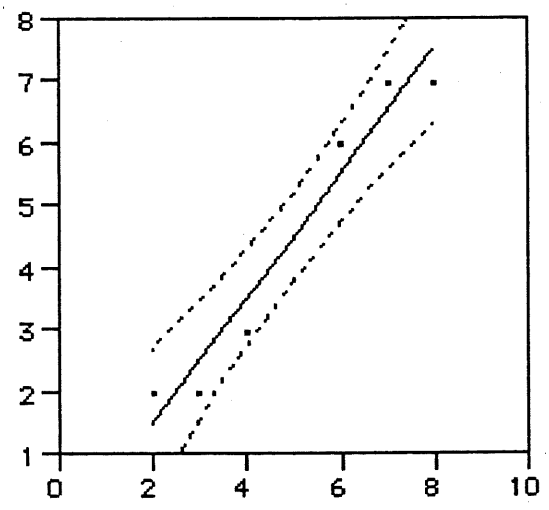
Voici un tableau présentant quelques valeurs:

θ	tan θ	R^2
0°	0	0
30°	$\sqrt{3}/3 = 0,577$	0,5
45°	1	0,75
60°	$\sqrt{3} = 1,732$	0,9
89°	57,29	0,9999

On remarque qu'en augmentant la valeur de θ, on accentue la pente de la droite de régression; ceci a pour effet d'éloigner les valeurs Y_i de $\overline{Y}$ et donc d'augmenter SST. Notons qu'ici, la fonction correspondant à SST, soit $6\tan^2\theta+2$, est croissante en θ. Ainsi, SST augmente de pair avec θ ce qui explique, à SSE constante, l'augmentation du coefficient de détermination.

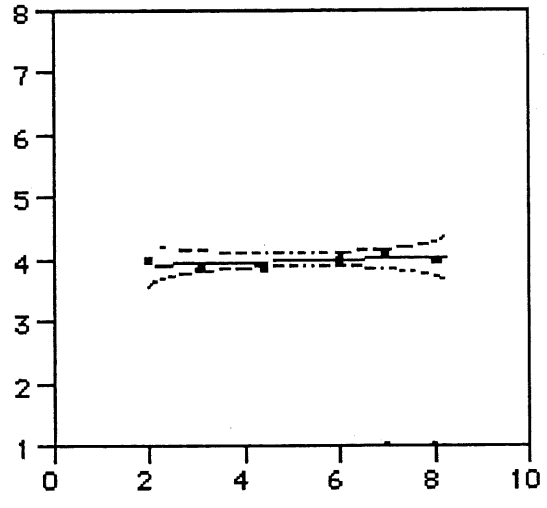
Un autre inconvénient du coefficient de détermination est qu'il n'est pas un bon indicateur de la qualité de l'équation de régression en ce qui à trait à la véritable valeur moyenne de la variable expliquée (en utilisant un intervalle de confiance) ou lorsque l'on veut prédire une valeur future (en utilisant un intervalle de prédiction). En fait, il est possible qu'un ensemble d'observations pour lequel la valeur du R^2 est élevée fournisse de moins bons intervalles de confiance qu'un autre ensemble d'observations avec une valeur moindre du R^2 (il en va de même pour les intervalles de prédiction).

Voici un exemple d'intervalles de confiance à $1-\alpha=95\%$ pour y . Pour faciliter la comparaison, la variable indépendante prend exactement les mêmes valeurs dans les deux cas de sorte que seule la valeur de $s=\sqrt{SSE/(n-k)}$ influence la largeur des intervalles de confiance.



R^2=0.949

Figure (d)



R^2=0.321

Figure (e)

On constate uniquement par comparaison visuelle que l'intervalle de confiance est plus large dans la Figure (d) que dans la Figure (e). Plus formellement, la largeur de l'intervalle donnée par $2 \cdot s \cdot t_{\alpha/2}(n-k) \cdot \sqrt{a^t (X^t X)^{-1} a}$ (voir section 1.2.5 (a)) varie de 1.39 à 2.38 dans la Figure (d) et de 0.19 à 0.32 dans la Figure (e). En considérant plutôt des intervalles de prédiction à 95%, la largeur de ceux-ci, donnée par $2 \cdot s \cdot t_{\alpha/2}(n-k) \cdot \sqrt{1 + a^t (X^t X)^{-1} a}$ (voir section 1.2.5 (b)) varie de 3.67 à 4.15 dans la Figure (d) et de 0.49 à 0.56 dans la Figure (e).

Cet exemple reflète une situation où, dans la Figure (d), la variabilité non expliquée par la régression est relativement grande (ce qui contribue à élargir les intervalles de confiance et de prédiction) mais est toutefois relativement faible comparée à la variabilité totale. De cette façon la valeur de $R^2 = 1 - \frac{SSE}{SST}$ est élevée bien que les intervalles soient larges. La Figure (e) reflète la situation contraire, à savoir SSE petit mais relativement grand par rapport à SST. Notons que l'exemple considéré exploite l'argument concernant l'inclinaison de la surface de régression voulant que la valeur de SST augmente avec l'inclinaison de la surface. On remarque que l'inclinaison est élevée dans la Figure (d) ce qui n'est pas le cas dans la Figure (e). De façon évidente, on peut exploiter également l'argument de dispersion des variables dépendantes pour faire varier la somme totale des carrés.

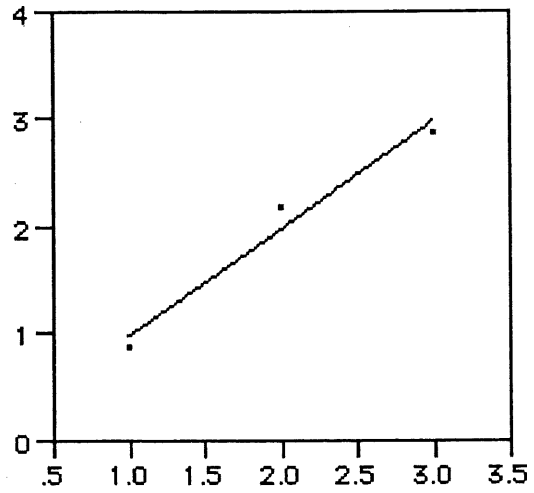
c) Inconvénients reliés aux "répétitions"

DRAPER & SMITH [14] ont remarqué qu'en présence de répétitions, sauf dans des cas très rares, le coefficient de détermination ne peut prendre la valeur 1. Ce résultat, en soit, n'est pas très étonnant puisqu'on observe que très rarement la valeur 1 en pratique, mais ceci constitue néanmoins une limite théorique qui peut être contournée, comme on le verra sous peu, en redéfinissant un nouveau coefficient de détermination.

Puisque $R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{(SSPE + SSLF)}{SST}$, on aura $R^2 = 1$ que si $SSPE = SSLF = 0$. Donc, en supposant qu'il y a au moins une répétition, i.e. qu'il existe au moins un indice i tel que $n_i > 1$, cette situation ne survient que si on a simultanément des variables expliquées ayant les mêmes valeurs dans les blocs de répétitions ($Y_{ij} = \overline{Y}_i$) et un modèle parfaitement ajusté aux moyennes ($\hat{Y}_i = \overline{Y}_i$). On en déduit que sauf dans des cas très rares on a $R^2 < 1$ lorsqu'il y a des répétitions (en fait lorsqu'il y a de l'erreur pure).

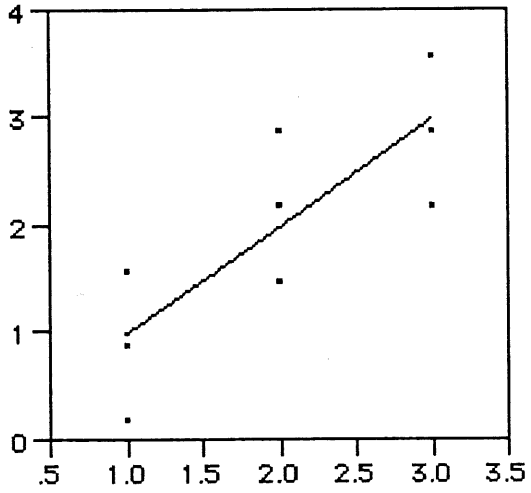
Dans un autre ordre d'idée, notons qu'en général, lorsque l'on augmente le nombre d'observations, on réussit à mieux cerner l'ensemble des points diminuant ainsi les erreurs individuelles et de ce fait la somme des carrés des erreurs (SSE). Le coefficient R^2 se voit donc augmenté. Comme le mentionne DRAPER [15], dans le cas de répétitions, il est possible qu'en augmentant le nombre d'observations, on diminue le R^2.

Examinons les graphes suivants:



R^2=0.971

Figure (f)



R^2=0.658

Figure (g)

Les deux ensembles d'observations considérés donnent tous deux lieu à la droite de régression $y = x$. On remarque que dans la Figure (g), le R^2 est moins élevé, bien que le nombre d'observations soit plus grand. Cet exemple illustre une situation où la valeur de SSE est amplifiée à cause de l'erreur pure. Dans la Figure (f), on a $SSPE = 0$ puisqu'il n'y a aucune répétition (ce qui n'est pas le cas dans la Figure (g)). Ainsi, puisque $SSE = SSPE + SSLF$ est plus élevé dans la Figure (b), le R^2 associé s'en voit diminué.

Le cas de la Figure (g) peut être généralisé. Afin de conserver la même droite de régression, on considérera des ensembles d'observations avec un nombre identique de répétitions en chaque valeur différente de x et comme dans la Figure (g) les répétitions doivent être symétriques par rapport aux observations initiales (1, 0.9), (2, 2.2) et (3, 2.9). Supposons que l'on utilise les indices f et rep pour identifier respectivement l'ensemble des observations de la Figure (f) et un autre ensemble d'observations avec des répétitions, on a alors :

$SSPE_f = 0$ et donc $SSLF_f = SSE_f$,

$SSR_{rep} = n_i SSR_f$,

$SSLF_{rep} = n_i SSLF_f$.

Ainsi,

$$SSPE_{rep} > 0 \Leftrightarrow n_i[SSLF_f + SSR_f] + SSPE_{rep} > n_i SSR_f\left[\frac{SSLF_f + SSR_f}{SSR_f}\right]$$

$$\Leftrightarrow \frac{SSR_f}{SSLF_f + SSR_f} > \frac{n_i SSR_f}{n_i[SSLF_f + SSR_f] + SSPE_{rep}}$$

$$\Leftrightarrow \frac{SSR_f}{SSLF_f + SSR_f} > \frac{SSR_{rep}}{SSLF_{rep} + SSPE_{rep} + SSR_{rep}}$$

$$\Leftrightarrow \frac{SSR_f}{SST_f} > \frac{SSR_{rep}}{SST_{rep}}$$

$$\Leftrightarrow R_f^2 > R_{rep}^2$$

Vue la décomposition de l'erreur en deux parties, à savoir l'erreur pure et l'erreur due au manque d'ajustement, la qualité de l'ajustement d'une équation de régression peut être examinée de façon plus efficace en considérant les I résidus moyens plutôt que les $n = \sum_{i=1}^{I} n_i$ résidus individuels. Dans le cas de répétitions, le terme important dans l'erreur est vraiment $SSLF = \sum_{i=1}^{I} n_i(\hat{Y}_i - \overline{Y}_i)^2$ qui considère n_i fois les résidus moyens dans chaque bloc. C'est pourquoi CHANG & AFIFI [8] suggèrent une modification du coefficient de détermination afin de tenir compte des répétitions. Ainsi, plutôt que de considérer $R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{SSLF + SSPE}{SSR + (SSLF + SSPE)}$, ils proposent $R_M^2 = 1 - \frac{SSLF}{SSR + SSLF}$. On remarque que s'il n'y a pas de répétitions, à savoir $SSPE = 0$, alors $R_M^2 = R^2$. De plus, si le modèle est parfaitement ajusté aux moyennes, i.e. $SSLF = 0$, alors R_M^2 peut atteindre la valeur 1 comme borne supérieure et, tout comme le coefficient de détermination conventionnel, si $SSR = 0$, alors R_M^2 atteint la valeur 0 comme borne inférieure. On peut remarquer que dans l'exemple considéré, comme $SSR_g = n_i SSR_f$ et que $SSLF_g = n_i SSLF_f$, alors $R_M^2 = R^2$. Comme précédemment, l'indice g se rapporte à la Figure (g).

d) Inconvénient relié aux valeurs extrêmes

Le coefficient de détermination est sensible aux valeurs extrêmes puisqu'il est fonction de la somme des carrés des erreurs. D'ailleurs, la régression par la méthode des moindres carrés est elle-même influencée par les valeurs extrêmes. Notons que dans de tels cas, il vaut mieux avoir recours à des méthodes dites robustes; ces méthodes ne feront pas l'objet du présent document mais le lecteur intéressé peut se référer à HUBER [31] ou à MONTGOMERY & PECK [39]. Évidemment, en changeant de méthode de modélisation, le coefficient de détermination tel que défini devient inadéquat. Si on utilise une méthode robuste, il devient désirable de modifier le coefficient de détermination afin de le rendre robuste à son tour. KVALSETH [33] suggère la statistique $1-\left[\frac{med\{|Y_i-\hat{Y}_i|\}}{med\{|Y_i-\overline{Y}|\}}\right]^2$ qui est construite de la façon suivante: la moyenne arithmétique $SSE/n=\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2/n$ est remplacée par la médiane $med\{(Y_i-\hat{Y}_i)^2\}$ et $SST/n=\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2/n$ par $med\{(Y_i-\overline{Y})^2\}$. Ces deux médianes sont respectivement exactement égales à $\left[med\{|Y_i-\hat{Y}_i|\}\right]^2$ et $\left[med\{|Y_i-\overline{Y}|\}\right]^2$ quand n est impair et approximativement égales aux deux mêmes quantités quand n est pair. On peut donc remplacer $1-\frac{SSE}{SST}=1-\frac{SSE/n}{SST/n}$ par $1-\left[\frac{med\{|Y_i-\hat{Y}_i|\}}{med\{|Y_i-\overline{Y}|\}}\right]^2$.

2.1.4 Discussion

Nous venons de voir des situations où le coefficient de détermination n'est pas nécessairement un bon indicateur de la qualité d'une régression. Remarquons que les exemples considérés portaient sur des ensembles d'observations différents ayant une somme totale des carrés différente. Cependant, lorsque l'on étudie divers sous-ensembles de variables pour un ensemble donné, la somme totale des carrés est constante et seule la somme des carrés des erreurs varie. En fait, en pratique, le R^2 est utilisé comme une mesure relative et non comme une mesure absolue. Les exemples considérés servent à réaliser que même en observant une valeur très faible, voire nulle, du R^2 pour le modèle complet ou pour un sous-modèle de celui-ci, il ne faut pas nécessairement en conclure que l'équation de régression résume mal les données. Par exemple, en régression linéaire simple, si $\hat{\beta}_1=0$ alors $R^2=0$ et ce, même si les données sont arbitrairement près de la droite estimée (un cas semblable

fournirait une très petite valeur de la quantité SSE). Inversement, une équation "mal ajustée" peut fournir à tort une valeur élevée du R^2.

Il semble raisonnable de n'avoir recours au coefficient de détermination que pour vérifier la qualité de l'ajustement d'une équation de régression. Utiliser ce critère à d'autres fins (pour prédire, par exemple) est fortement déconseillé.

Rappelons qu'à la lumière des exemples précédents il faut se méfier du terme SST qui peut injustement contribuer à l'augmentation ou la diminution du R^2.

On peut penser obtenir de "meilleures" conclusions en utilisant SSE au lieu de R^2 comme critère de sélection. En effet, il est évident que le SSE n'est pas influencé par la dispersion des variables dépendantes ni par l'inclinaison de la surface de régression. Cependant, l'atout non négligeable du coefficient de détermination est son interprétation comme un pourcentage d'explication; ceci nous fait mieux réaliser l'impact du retrait de variables. Le SSE quant à lui n'est pas borné supérieurement (si ce n'est que par SST) et bien qu'il possède une borne inférieure (à savoir 0), il est difficile de concevoir une notion de proximité vers celle-ci.

Notons finalement que, comme nous l'avons vu précédemment, le R^2 diminue systématiquement lorsque l'on retire une ou des variables. La comparaison de deux ensembles de variables de tailles différentes, que l'on appellera comparaison interclasses, devient donc difficile puisque les ensembles de plus grandes tailles sont "privilégiés". Ceci constitue une raison suffisante pour définir un autre critère qui, lui, nous permettra d'effectuer des comparaisons interclasses.

2.2 Coefficient de détermination corrigé ($\overline{R^2}$)

Devant le problème des comparaisons interclasses, EZEKIEL [18] a suggéré de remplacer $R^2 = 1 - \frac{SSE}{SST}$ par $\overline{R^2} = 1 - \frac{MSE}{MST}$. Notons que bien qu'en enlevant r variables à un modèle à k paramètres on ait $SSE_k < SSE_{k-r}$, il est possible que $[n-(k-r)]SSE_k > (n-k)SSE_{k-r}$, i.e. $MSE_k > MSE_{k-r}$ et donc $\overline{R^2_{k-r}} > \overline{R^2_k}$.

2.2.1 Définition

En considérant un modèle à k paramètres (incluant β_0) on a montré que sous l'hypothèse $H_0 : \beta_1 = \beta_2 = ... = \beta_{k-1} = 0$, $E(R_k^2) = (k-1)/(n-1)$.

Ainsi, la première étape pour corriger R_k^2 consiste à considérer $R_k^2 - \frac{k-1}{n-1}$.

Notons que si R_k^2 vaut 1 alors cette dernière quantité ne vaut pas 1 mais plutôt $1 - \frac{k-1}{n-1}$, i.e. $\frac{n-k}{n-1}$ quantité qui est inférieure à 1.

Ainsi, afin de définir un R_k^2 corrigé qui prend la valeur 1 si R_k^2 vaut 1, on doit multiplier $R_k^2 - \frac{k-1}{n-1}$ par le facteur $\frac{n-1}{n-k}$.

On obtient donc une formule permettant d'évaluer ce qu'il sera convenu d'appeler le R^2 corrigé, et noté $\overline{R^2}$, qui est donné par:

$$\overline{R_k^2} = \left(R_k^2 - \frac{k-1}{n-1}\right)\left(\frac{n-1}{n-k}\right).$$

On rencontre plus fréquemment cette formule sous une forme équivalente soit:

$$\overline{R_k^2} = \left(R_k^2 - \frac{k-1}{n-1}\right)\left(\frac{n-1}{n-k}\right) = 1 - \left(\frac{n-1}{n-k}\right)(1 - R_k^2) = 1 - \left(\frac{n-1}{n-k}\right)\frac{SSE_k}{SST}.$$

Le facteur $\frac{n-1}{n-k}$ est appelé facteur d'ajustement et correspond au rapport des degrés de liberté de SST et SSE_k. On peut donc également présenter $\overline{R_k^2}$ sous la forme $1 - \frac{MSE_k}{MST}$.

2.2.2 Propriété

Sachant que $0 \le R_k^2 \le 1$, on en déduit les équivalences suivantes:

$$0 \le 1 - R_k^2 \le 1 \iff 0 \le \left(\frac{n-1}{n-k}\right)(1 - R_k^2) \le \frac{n-1}{n-k}$$

$$\Leftrightarrow \frac{-(n-1)}{n-k} \leq -\left(\frac{n-1}{n-k}\right)(1-R_k^2) \leq 0$$

$$\Leftrightarrow 1-\frac{n-1}{n-k} \leq 1-\left(\frac{n-1}{n-k}\right)(1-R_k^2) \leq 1$$

$$\Leftrightarrow \frac{1-k}{n-k} \leq \bar{R}_k^2 \leq 1$$

2.2.3 Inconvénients

Le coefficient de détermination corrigé possède évidemment les mêmes limites que le coefficient de détermination pour ce qui est de l'influence de SST. Dans certaines situations, il est intéressant de modifier ce critère tout comme on l'a fait avec le coefficient de détermination. Les modifications suggérées ne sont ni plus ni moins que l'ajout du facteur d'ajustement ou d'une adaptation de celui-ci. Ainsi, en ce qui a trait au problème des répétitions, CHANG & AFIFI [8] suggèrent d'utiliser le critère modifié $\overline{R_M^2} = 1-\left(\frac{I-1}{I-k}\right)\left(\frac{SSLF}{SSR+SSLF}\right)$. Pour ce qui est des valeurs extrêmes, KVALSETH [33] propose de remplacer $\overline{R^2}$ par $1-\left(\frac{n-1}{n-k}\right)\left[\frac{med\{|Y_i-\hat{Y}_i|\}}{med\{|Y_i-\bar{Y}|\}}\right]^2$.

2.2.4 Discussion

Le coefficient de détermination corrigé n'apporte pas tellement plus d'information que le coefficient de détermination, mais il a l'avantage de faciliter les comparaisons interclasses. De façon parallèle au critère précédent, on peut penser utiliser le critère du MSE au lieu du $\overline{R^2}$. Cependant, on se retrouve de nouveau dans l'impasse de se demander s'il est plus avantageux d'utiliser un critère qui ne sera pas influencé par l'inclinaison de la surface de régression mais, à partir duquel, on ne peut facilement définir une notion de proximité vers la borne inférieure.

2.3 C_p de Mallows

Le critère du C_p de Mallows a été défini par Colin L. Mallows et publié initialement par GORMAN & TOMAN [21]. Le C_p (C_k si on considère un modèle à k paramètres) repose sur le prncipe de la minimisation simultanée du biais et de la variance des valeurs prédites par l'entremise de l'erreur quadratique moyenne calculée sur chacune des n observations d'un modèle comportant k paramètres. Celle-ci est notée MSE_{ik} (mean square error).

2.3.1 Définition

On pose $\begin{cases} \mathrm{E}[\hat{Y}_{ik}] = \eta_{ik}, & 1 \le i \le n \\ \mathrm{E}[Y_{ik}] = \upsilon_{ik}, & 1 \le i \le n \end{cases}$.

Puisque pour une variable aléatoire Z quelconque $Var(Z) = E(Z^2) - [E(Z)]^2$ et que dans le cas présent υ_{ik} est un vecteur de constantes on a :

$$MSE_{ik} = \mathrm{E}[(\hat{Y}_{ik} - \upsilon_{ik})^2] = \mathrm{Var}(\hat{Y}_{ik} - \upsilon_{ik}) + [\mathrm{E}(\hat{Y}_{ik} - \upsilon_{ik})]^2 = \mathrm{Var}(\hat{Y}_{ik}) + [\eta_{ik} - \upsilon_{ik}]^2 .$$

Étant donné que l'on veut minimiser MSE_{ik} pour la variable Y, on minimise plutôt la somme des MSE_{ik}. Ainsi, en définissant la somme des carrés des biais comme $SSB = \sum_{i=1}^{n} (\upsilon_{ik} - \eta_{ik})^2$ on a alors: $\sum_{i=1}^{n} MSE_{ik} = \sum_{i=1}^{n} \mathrm{Var}(\hat{Y}_{ik}) + SSB$.

Dans la suite, nous allons considérer une version "réduite" de cette quantité à savoir:

$$\Gamma_k = \frac{1}{\sigma^2}\left[\sum_{i=1}^{n} \mathrm{Var}(\hat{Y}_{ik}) + SSB\right].$$

Définissons $H_k = X\ (X^t X)^{-1} X^t$ où X est de dimension $n \times k$ et posons $\mathrm{E}(Y) = X\beta = \mu$ et $\mathrm{Cov}(Y) = \sigma^2 I_n$.

On a alors :

$$\sum_{i=1}^{n} \mathrm{Var}(\hat{Y}_{ik}) = \mathrm{tr}[\mathrm{Cov}(\hat{Y}_k)] = \mathrm{tr}[\mathrm{Cov}(H_k Y)] = \mathrm{tr}[H_k \mathrm{Cov}(Y) H_k^t] = \mathrm{tr}[\sigma^2 H_k H_k^t] = \mathrm{tr}[\sigma^2 H_k],$$

et donc $\sum_{i=1}^{n} \mathrm{Var}(\hat{Y}_{ik}) = k\sigma^2$.

De plus,

$$\begin{aligned} \mathrm{E}(SSE_k) &= \mathrm{E}\,(\,Y'[I_n - H_k]Y\,) \\ &= \sigma^2 \mathrm{tr}(I_n - H_k) + \mu'(I_n - H_k)\mu \quad \text{(en vertu de la proposition 1.3)} \\ &= \sigma^2(n-k) + \mu'\left[(I_n - H_k)'(I_n - H_k)\right]\mu \\ &= \sigma^2(n-k) + (\mu - H_k\,\mu)'(\mu - H_k\,\mu)\ . \end{aligned}$$

Mais puisque $H_k\,\mu = X\,(X'X)^{-1}X'\,\mathrm{E}(Y) = \mathrm{E}[X\,(X'X)^{-1}X'Y\,] = \mathrm{E}(X\hat{\beta}) = \mathrm{E}(\hat{Y})$, on en déduit que $SSB = \mathrm{E}[SSE_k] - (n-k)\sigma^2$

Ainsi: $\Gamma_k = \frac{1}{\sigma^2}\left[k\sigma^2 + \mathrm{E}[SSE_k] - \sigma^2(n-k)\right] = \frac{\mathrm{E}[SSE_k]}{\sigma^2} - (n-2k)$.

Puisqu'on ne peut évaluer directement Γ_k, on est amené à considérer la statistique C_k qui est un estimateur de Γ_k.

La statistique C_k est définie de la façon suivante :

$$C_k = \frac{SSE_k}{s^2} - (n-2k) \text{ où } s^2 \text{ est un estimateur de } \sigma^2. \qquad (3)$$

Un bon estimateur s^2 de σ^2 est donné par l'erreur quadratique moyenne du modèle complet.

Ainsi on a: $C_k = (n-p)\frac{SSE_k}{SSE_p} - (n-2k)$. (4)

Si l'équation possède un biais négligeable, i.e. $SSB = \mathrm{E}[SSE_k] - (n-k)\sigma^2 \approx 0$, alors SSE_k estime bien $(n-k)\sigma^2$ et donc, à partir de (3), on a :

$$C_k \approx \frac{(n-k)s^2}{s^2} - (n-2k) = k\,.$$

On en déduit qu'à toute équation possédant un faible biais correspond une valeur de C_k voisine de k, tandis qu'un biais important entraînera une valeur de C_k beaucoup plus élevée que k.

D'ailleurs, la statistique possède la propriété voulant que le modèle sans biais qu'est le modèle complet (quand k=p) donne $C_p = p$ (simple substitution dans (4)).

2.3.2 Propriété

Puisque $SSE_k \geq SSE_p$ on a les équivalences suivantes:

$$\begin{aligned}\frac{SSE_k}{SSE_p} \geq 1 &\Leftrightarrow (n-p)\frac{SSE_k}{SSE_p} \geq n-p \\ &\Leftrightarrow (n-p)\frac{SSE_k}{SSE_p} - (n-2k) \geq 2k-p \\ &\Leftrightarrow C_k \geq 2k-p\ .\end{aligned}$$

2.3.3 Inconvénients

$$C_k = \frac{SSE_k}{\hat{\sigma}^2} - (n-2k) = \frac{SSE_k}{\hat{\sigma}^2} - (n-k) + k = (n-k)\left[\frac{SSE_k/(n-k)}{\hat{\sigma}^2} - 1\right] + k\ .$$

On réalise que la quantité entre crochets devrait idéalement être près de 0. On retrouve donc l'idée voulant que l'on doit choisir des équations telles que $C_k \approx k$. Cependant, la quantité entre crochets peut être grandement amplifiée du au facteur $(n-k)$. Comme le suggère HOCKING [27], le critère du C_p est davantage recommandé pour des valeurs élevées de k .

Ce critère est également sensible aux répétitions. Afin de définir une statistique qui tient compte de ce problème, CHANG & AFIFI [8] suggèrent de remplacer $C_k = \frac{SSE_k}{s^2} - (n-2k)$ par $C_k' = \frac{SSLF_k}{s^2} - (I-2k)$ où s^2 est un estimateur sans biais de σ^2 donné par $s^2 = SSPE_k/n-I$.

Tout comme les critères précédents, le coefficient C_p est fondé sur l'estimation par la méthode des moindres carrés ce qui le rend sensible aux valeurs extrêmes. Le lecteur intéressé pourra consulter RONCHETTI & STAUDTE [43] où est introduite une version robuste du coefficient C_p.

2.3.4 Discussion

Le coefficient C_p est reconnu comme étant un critère relativement satisfaisant puisqu'il permet d'éviter, dans la mesure du possible, deux types d'erreurs lorsque vient le temps de choisir un sous-ensemble de variables. La première de ces erreurs est l'inclusion de variables non pertinentes et la seconde est l'exclusion de variables ayant une réelle influence sur la variable expliquée.

L'inclusion de variables non pertinentes ne biaise pas les résultats pour ce qui est des autres variables indépendantes mais a toutefois un impact sur celles-ci. Principalement, la variance des variables prédites se voit augmentée. On peut ajouter également que ces variables "inutiles" nuisent à la simplicité du modèle ce qui complique une interprétation éventuelle. En plus, ces variables additionnelles peuvent réduire la précision des tests de signification statistique.

D'un autre côté, l'omission de variables pertinentes peut sérieusement biaiser les résultats et compromettre une interprétation de ceux-ci. Dans le cas le plus simple où les variables exclues ne sont pas corrélées avec les variables incluses, la précision des prédictions se voit réduite. Cependant, en présence de corrélation, plus celle-ci est grande, plus le biais augmente. Ceci peut s'expliquer par le fait que les effets estimés des variables incluses proviennent non seulement de leur propre effet mais également de celui partagé avec les variables omises.

On en vient à la conclusion qu'un critère exploitant un compromis entre les deux types d'erreurs, par le minimisation simultanée du biais et de la variance des valeurs prédites, est définitivement intéressant.

À la lumière de la limite mentionnée ci-haut, il est préférable d'utiliser ce critère lorsque l'on dispose d'un nombre élevé de variables. Idéalement, d'après les arguments précédents, on cherche une équation ayant valeur de C_k qui n'est pas trop loin de k. Si le choix n'est pas

univoque, le jugement personnel entre en cause dépendant de la préférence de l'utilisateur. Une approche consiste à choisir une équation biaisée (habituellement un sous-ensemble avec peu de variables) qui ne représente pas nécessairement bien les données due à une grande valeur de SSE_k (et donc $C_k > k$) mais dont la valeur de C_k (issue de la somme de la variance et du biais des valeurs prédites) est la plus petite de toutes les valeurs peu importe le nombre de variables. Une autre approche consiste à choisir une équation avec plus de paramètres qui représente bien les données (à savoir $C_k \approx k$) mais dont la valeur de C_k n'est pas la plus petite (peu importe le nombre de variables).

Pour la deuxième approche, une façon visuelle de repérer les meilleurs ensembles de variables est d'identifier les diverses valeurs de C_k sur un graphique de C_k en fonction de k. En traçant la droite correspondant à $C_k = k$, tout point près de cette droite correspond à un bon sous-ensemble. Une autre façon de procéder consiste à tracer le graphique de $C_k - k$ en fonction de k et de repérer les points près de la droite horizontale $C_k - k = 0$.

2.4 Relations entre MSE_k, $\overline{R_k^2}$, C_k et F

$$\overline{R_k^2} < \overline{R_{k-r}^2} \overset{(1)}{\Leftrightarrow} 1 - \frac{SSE_k/n-k}{TSS/n-1} < 1 - \frac{SSE_{k-r}/[n-(k-r)]}{TSS/(n-1)}$$

$$\overset{(2)}{\Leftrightarrow} \frac{SSE_{k-r}/[n-(k-r)]}{TSS/(n-1)} < \frac{SSE_k/(n-k)}{TSS/(n-1)}$$

$$\overset{(3)}{\Leftrightarrow} SSE_{k-r}/[n-(k-r)] < SSE_k/(n-k)$$

$$\overset{(4)}{\Leftrightarrow} (n-k)\frac{SSE_{k-r}}{SSE_k} < n-(k-r)$$

$$\overset{(5)}{\Leftrightarrow} (n-k)\frac{SSE_{k-r}}{SSE_k} - [n-2(k-r)] < k-r$$

$$\overset{(6)}{\Leftrightarrow} C_{k-r} < k-r \ .$$

Notons que l'équivalence (3) revient à $MSE_{k-r} < MSE_k$.

À partir de l'équation (3), on a également les relations suivantes:

$$\frac{SSE_{k-r}}{n-(k-r)} < \frac{SSE_k}{n-k} \Leftrightarrow \frac{SSE_{k-r}}{n-(k-r)} - \frac{SSE_k}{n-(k-r)} < \frac{SSE_k}{n-k} - \frac{SSE_k}{n-(k-r)}$$

$$\Leftrightarrow \frac{(SSE_{k-r} - SSE_k)}{n-(k-r)} < \left(\frac{r}{(n-k)[n-(k-r)]}\right) SSE_k$$

$$\Leftrightarrow \frac{(SSE_{k-r} - SSE_k)/r}{SSE_k/(n-k)} < 1$$

$$\Leftrightarrow F < 1 .$$

Dans les équations ci-dessus, les équivalences demeurent inchangées si l'on subsitue partout le symbole d'inégalité (<) par le symbole d'égalité (=) ou par l'inégalité inverse (>).

À l'occasion, on peut constater que $MSE_{k-r} < MSE_k$. Dans ce cas, inutile de se surprendre si l'on observe que $\overline{R_k^2} < \overline{R_{k-r}^2}$, $C_{k-r} < k-r$ ou $F < 1$.

SEBER [45] procède autrement pour constater la similarité entre $\overline{R_k^2}$ et C_k. Il fait remarquer qu'en considérant $C_k = (n-p)\frac{SSE_k}{SSE_p} - (n-2k)$ et $\overline{R_k^2} = 1 - \left(\frac{n-1}{n-k}\right)\frac{SSE_k}{SST}$ on a:

$$1 + \frac{(C_k - k)}{n-k} = \left(\frac{n-p}{n-k}\right)\frac{SSE_k}{SSE_p} = \frac{1-\overline{R_k^2}}{1-\overline{R_p^2}} .$$

On en conclut que si $n \geq k$, en remarquant que $1-\overline{R_p^2}$ est seulement un "facteur d'échelle", alors $C_k - k$ est pratiquement équivalent à $1-\overline{R_k^2}$. En fait, les deux quantités fournissent une mesure de l'amplitude du biais.

2.5 Exemple d'utilisation des critères sur les données de Hald

Les critères vus précédemment seront maintenant mis à l'essai sur des données réelles. L'ensemble de données choisi est tiré du livre de HALD [1952, p.647] et portent sur la chaleur dégagée au cours de la solidification du ciment en fonction des pourcentages de quatre constituants. Les données ainsi que la matrice des corrélations sont reproduites à l'annexe 4. Ces données peuvent être qualifiées de données "classiques" puisqu'elles ont été étudiées par bon nombre d'auteurs de renom qui les ont utilisées pour démontrer certaines difficultés qui peuvent survenir en régression linéaire multiple.

Bien que les données soient réelles, très peu d'auteurs ont pris la peine de mettre en garde leurs lecteurs quant au nombre restreint d'observations, soit 13 observations pour 5 variables, et n'ont pas pris non plus la peine de mettre les données en contexte. Ceci relègue l'exercice de sélection de variables sur ces données à un exercice presque purement académique.

La citation qui suit en est une de Ronald D. Snee extraite de la discussion sur l'article de HOCKING [28]. Elle montre bien son étonnement face à l'utilisation par certains auteurs d'ensemble de données sans tenir compte du contexte. Cette citation fait également référence à un autre ensemble de données historiques dont il ne sera pas question dans ce travail.

> " The extensive use (frequently with no reference to the physical context of the problem) of the 10-factor example mentioned earlier is consistent with my impression that there are many developers of regression methodology who do not analyze many data sets. Hald's 13-observation cement data is another good example (Draper and Smith 1981). Both of these data sets have been analyzed by more authors than there are observations in the data set. The cement data also have the distinction of being a mixture problem for which a constant term model and variable selection are not appropriate. Only Daniel and Wood (1980) have recognized this important feature of the data."

En examinant la matrice des corrélations, on s'aperçoit qu'il existe des corrélations élevées entre certaines variables explicatives. La corrélation entre X1 et X3 est de -0,824 et celle entre X2 et X4 de -0,973. On doit donc s'attendre à ce qu'une seule variable du premier groupe combinée à une seule variable du deuxième groupe explique la majeure partie de la variation.

Critère du coefficient de détermination

Voici le tableau des valeurs du coefficient de détermination les plus élevées pour les sous-ensembles de tailles différentes. $f(.)$ sert à énumérer les variables qui interviennent dans le modèle.

Variables dans l'équation	Coefficient de détermination
f(X4)	0,67454
f(X1, X2)	0,97868
f(X1, X4)	0,97247
f(X1, X2, X4)	0,98234
f(X1, X2, X3, X4)	0,98238

À la lumière de ce tableau, on remarque qu'un modèle à deux variables suffit à expliquer presque toute la variation. Conserver trois ou quatre variables ne ferait qu'alourdir le modèle en le rendant plus difficile d'interprétation. On peut également noter qu'il existe une très légère différence entre les R^2 du meilleur modèle à trois variables et celui à quatre variables. Ceci peut s'expliquer par le fait que les variables indépendantes représentent les pourcentages d'un mélange d'ingrédients et que la somme des valeurs de ces variables est pratiquement constante, variant entre 95 et 99.

Pour ce qui est des deux modèles à deux variables, aucun ne semble se démarquer; certains choisiront le modèle faisant intervenir X1 et X2 puisque le R^2 est plus élevé, alors que d'autres affirmeront que le modèle qui est fonction de X1 et X4 est meilleur puisque le meilleur modèle à une variable fait intervenir X4. Bien sûr, mis à part les arguments statistiques, la préférence de l'un ou l'autre des sous-ensembles peut être déterminée à partir d'une bonne connaissance des caractéristiques des composantes du ciment par un expert dans le domaine.

Critère du coefficient de détermination corrigé

Voici le tableau des valeurs du coefficient de détermination corrigé les plus élevées pour les sous-ensembles de tailles différentes:

Variables dans l'équation	Coefficient de détermination corrigé
f(X4)	0,64495
f(X1, X2)	0,97441
f(X1, X4)	0,96697
f(X1, X2, X4)	0,97645
f(X1, X2, X3, X4)	0,97356

Encore une fois, on remarque qu'un modèle à deux variables est suffisant. Rappelons qu'une des propriétés intéressantes de ce critère est de permettre les comparaisons interclasses sans que les modèles avec un plus grand nombre de variables soient systématiquement meilleurs. Ici, on remarque que le $\overline{R^2}$ est plus élevé pour le modèle utilisant X1 et X2 que pour le modèle complet.

Critère du C_p de Mallows

Voici le tableau des plus petites valeurs du C_p de Mallows pour certains sous-ensembles de tailles différentes. Les ensembles non listés affichent des valeurs supérieures à 20.

Variables dans l'équation	C_k	k
f(X4)	138,7	2
f(X1, X2)	2,7	3
f(X1, X4)	5,5	3
f(X1, X2, X3)	3,0	4
f(X1, X2, X4)	3,0	4
f(X1, X3, X4)	3,5	4
f(X2, X3, X4)	7,3	4
f(X1, X2, X3, X4)	5,0	5

Puisque l'on recherche des sous-ensembles comportant un minimum de variables où $C_k \approx k$, i.e. où C_k est à peu près égal au nombre de paramètres (nombre de variables + 1), le choix tout indiqué est le modèle faisant intervenir X1 et X2.

2.6 Autres critères

La plupart des critères autres que ceux traités dans ce chapitre constituent des adaptations de ceux-ci : des versions robustes, des versions permettant des choix à l'intérieur de domaines, etc. Cette section présente deux critères qui amènent des idées différentes, il s'agit du critère γ_m et du critère PRESS. Le premier exploite une avenue menant à des décisions "pratique" plutôt que "théorique" et le deuxième exploite une technique du type "ré-échantillonnage".

2.6.1 Le critère γ_m

Le critère γ_m, que nous désignerons par γ_k afin d'uniformiser les notations, a été développé par Box et Wetz en 1973. Ce critère va plus loin que le concept de régression "statistiquement significative" en apportant l'idée de régression "utile en pratique".

Soit $Y = \eta + \varepsilon = X\beta + Z\Psi + \varepsilon$.

Ici, $Z\Psi$ représente des effets que l'on souhaite éliminer dans la variation des données (moyenne, variables en blocs, tendances reliées au temps, ...). Plus spécifiquement, Ψ est un vecteur de paramètres dits de nuisance. Comme précédemment, on suppose que ε est d'espérence nulle et de variance $\sigma^2 I$.

Soient $\tilde{\eta}_i$, la i^e observation du vecteur $\tilde{\eta} = Z\Psi$ et η_i, la i^e observation du vecteur $\eta = X\beta + \tilde{\eta}$.

Les changements sur les valeurs η_i calculés sur toutes les observations expérimentales peuvent être mesurés par :

$$\sum_{i=1}^{n}\left(\eta_i-\tilde{\eta}_i\right)^2\Big/n\ . \qquad (5)$$

L'estimation par la méthode des moindres carrés de $\eta_i-\tilde{\eta}_i$ est $\hat{Y}_i-\tilde{Y}_i$.

Afin de calculer efficacement les variances des $\hat{Y}_i-\tilde{Y}_i$, $i=1,\ldots,n$, on considère le vecteur $\hat{Y}-\tilde{Y}=X\hat{\beta}=HY$ où $H=X\,(X'X)^{-1}X'$. La matrice de variances-covariances est donnée par :

$$\begin{aligned}
\mathrm{E}\left\{[HY-\mathrm{E}(HY)]\,[HY-\mathrm{E}(HY)]'\right\} &= \mathrm{E}\left[(HY)\,(HY)'\right]-\mathrm{E}(HY)\,\mathrm{E}\left[(HY)'\right]\\
&= \mathrm{E}\left[H\,Y\,Y'H'\right]-H\,\mathrm{E}(Y)\,\mathrm{E}(Y')\,H'\\
&= H\left[\mathrm{E}(Y\,Y')\;-\;\mathrm{E}(Y)\,\mathrm{E}(Y')\right]H'\\
&= H\left[I\,\sigma^2\right]H'\\
&= H\,\sigma^2\ .
\end{aligned}$$

Ainsi, $\mathrm{Var}(\hat{Y}_i-\tilde{Y}_i)$ est donnée par le i^e élément de la diagonale de $H\,\sigma^2$. Une mesure de la qualité de l'estimation des différences $\eta_i-\tilde{\eta}_i$ est donnée par la moyenne de ces variances soit

$$\begin{aligned}
\mathrm{tr}(H\,\sigma^2)/n &= (\sigma^2/n)\mathrm{tr}\left([X\,(X'X)^{-1}]X'\right)\\
&= (\sigma^2/n)\mathrm{tr}\left(X\,[(X'X)^{-1}X']\right)\\
&= (\sigma^2/n)\mathrm{tr}(I_k)\\
&= k\sigma^2/n\ . \qquad (6)
\end{aligned}$$

Une mesure de la comparaison de l'ampleur des changements des $\eta_i-\tilde{\eta}_i$ par rapport à leurs erreurs d'estimation est donnée par la racine carrée du rapport entre les équations (5) et (6) soit

$$\gamma_k=\left\{\sum_{i=1}^{n}\left(\eta_i-\tilde{\eta}_i\right)^2\Big/k\sigma^2\right\}^{1/2}\ .$$

On s'intéressera à la plus petite valeur de γ_k pour laquelle l'équation est "utile en pratique" plutôt que seulement "statistiquement significative". Celle-ci est, jusqu'à un certain point, arbitraire, de la même façon que le choix d'un niveau de signification l'est. BOX & WETZ [7] ont montré que si γ_0 désigne le niveau minimal acceptable pour γ_k, il faut calculer

une certaine valeur F_0, dépendante de γ_o, à savoir $F_0 \approx (1+\gamma_0)^2 F_\alpha(\nu_o, \nu_r)$ où ν_r est le nombre de degrés de liberté résiduels et où ν_o est la partie entière de $k(1+\gamma_0^2)^2 / (1+2\gamma_0^2)$. Si la statistique F habituelle excède cette valeur F_0, on pourra accepter que la valeur de γ_k est suffisamment grande pour que l'ajustement de l'équation de régression soit utile en pratique. Selon DRAPER [14], si la valeur observée de F est trois ou quatre fois plus élevée que la valeur critique de F, la "signification" pratique est presque sûrement atteinte et est clairement atteinte si la valeur expérimentale est cinq ou six fois plus élevée que la valeur critique. Cette conclusion est contestée par DARLINGTON [13] qui affirme que pour de grands échantillons, de taille 100 000 par exemple, une corrélation simple de seulement 0,02 excède de 10 fois la valeur critique de F.

2.6.2 Le critère PRESS

Le critère PRESS a été proposé par ALLEN [1] et exploite l'idée de calculer une somme de carrés afin de choisir une équation fournissant les "meilleures" prédictions. Dans un monde sans contrainte, on considérerait deux échantillons (indépendants), un premier pour calculer les paramètres du modèle et l'autre pour valider la qualité des prédictions générées par le modèle. Pour ce, il suffirait d'étudier, pour chacune des observations de l'échantillon de validation, les écarts entre les Y_i observés et ceux calculés par le modèle. Comme, dans la majorité des cas, on ne peut s'offrir le "luxe" d'un deuxième échantillon, le critère PRESS offre une alternative intéressante.

Soit k le nombre de paramètres dans le modèle (incluant β_0) et n le nombre total d'observations. On débute en ignorant la première observation et on établit toutes les régressions possibles sur les *n-1* observations restantes. On trouve alors $\hat{Y}_{1k}$ un prédicteur de Y_1 (pour chacune des régressions). On poursuit en calculant $\hat{Y}_{2k}$ un prédicteur de Y_2 en ignorant la 2^e observation et en calculant encore une fois toutes les régressions à l'aide des *n-1* observations restantes. On répète le processus jusqu'à ce que l'on ait tous les prédicteurs $\hat{Y}_{1k},\ldots,\hat{Y}_{nk}$. On pourra alors évaluer le somme de carrés suivante :

$$\text{PRESS} = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_{ik}\right)^2 .$$

On cherchera évidemment une équation avec un minimum de paramètres pour laquelle la valeur de PRESS est la plus petite possible.

La principale critique qui puisse être formulée concernant ce critère est la quantité importante de calculs en cause. On peut également s'objecter à l'utilisation du même échantillon pour l'estimation et la validation. Il reste que le critère demeure un bon compromis si on ne peut se permettre un deuxième échantillon.

2.7 Correspondance entre les objectifs et les critères

Une des grandes richesses de la régression est qu'elle peut servir à atteindre divers objectifs. Dépendamment de l'usage que l'on veut en faire, on ne recherchera pas les mêmes qualités dans les équations de régression. Il devient donc évident qu'employer un critère non adapté à l'usage souhaité est l'une des principales sources d'erreurs.

D'un côté nous avons un certain nombre d'objectifs et de l'autre des critères qui ont été développés sans vraiment répondre totalement à l'un ou plusieurs de ces objectifs. En 1976, HOCKING [29] affirmait qu'il n'existe toujours pas de relations solides entre les objectifs et les critères. En 1992, TOMASSONE et al [46] réitèrent cette affirmation.

Afin d'en arriver à établir des correspondances raisonnables, commençons d'abord par énoncer certains objectifs fréquemment visés en régression linéaire multiple.

- *la description* i.e. utiliser une équation de régression dans le but de préciser des relations entre les variables indépendantes et d'analyser leur action sur la variable dépendante.

- *l'estimation et la prédiction* i.e. estimer la valeur moyenne correspondant à une observation donnée ou prédire le valeur d'une observation.

- *l'extrapolation* i.e. prolonger les résultats hors du domaine des données employées pour calculer les éléments de la régression.

Reprenons maintenant ces trois objectifs et tentons de leur associer les critères les mieux adaptés.

a) Description

Pour cet usage, il y a habituellement conflit entre le désir d'obtenir une description à la fois précise et la plus simple possible. En effet, la précision demande l'inclusion d'un grand nombre de variables, ce qui est incompatible avec la simplicité.

Au sens des moindres carrés, les équations avec de petites sommes des carrés des erreurs sont toutes indiquées (critère du SSE). Par conséquent, on peut également penser au coefficient de détermination qui quoique plus facile à interpréter possède certains inconvénients. Rappelons qu'une discussion concernant ces deux critères apparaît à la section 2.1.4 de ce mémoire.

b) Estimation et prédiction

Afin d'estimer et de prédire efficacement, une petite erreur quadratique moyenne est recherchée. D'ailleurs le terme $s = \sqrt{MSE}$ apparaît dans la formule du calcul de l'intervalle de confiance pour la valeur moyenne de Y tout comme dans celle du calcul de l'intervalle de prédiction pour Y.

En pensant au critère du MSE, on est amené à considérer le critère du coefficient de détermination corrigé (voir discussion à la section 2.2.4). Le critère du C_p de Mallows sert également bien cet objectif. Un autre critère est spécifiquement conçu pour prédire efficacement : le critère PRESS.

Nous savons qu'il est possible que $MSE_{p-r} \leq MSE$, i.e. qu'un sous-ensemble possède une plus petite erreur quadratique moyenne que le modèle complet. Un tel sous-ensemble se porte bien à l'estimation ou à la prédiction.

c) Extrapolation

Lorsqu'il est question d'extrapolation, la notion de danger est omniprésente. Bien que rien n'empêche une telle pratique, l'utilisateur doit être conscient des risques qu'il encourt.

D'abord, le modèle, dont on estime les paramètres, peut ne plus être valable hors du domaine des données ayant servi à calculer les éléments de la régression. Ensuite, en supposant qu'il l'est encore, une prédiction valable dans les limites du domaine peut être totalement inefficace à l'extérieur. MASON et al. [37] montrent que la quasi dégénérescence de la matrice X mène à de tels résultats.

Les mêmes critères que pour la prédiction peuvent être utilisés ici, à savoir les critères MSE, $\overline{R^2}$, C_p de Mallows et PRESS. En plus d'étudier ces critères, HOCKING [29] recommande de considérer des sous-ensembles où $F \leq \frac{1}{r}$ i.e. tels que $\left[\frac{(n-p)+r}{(n-p)+1}\right] MSE_{p-r} \leq MSE.$

À la lumière de cette section, il apparaît important d'utiliser des critères adaptés aux objectifs visés. Résumons les recommandations:

- Pour décrire des relations, les critères du SSE et du R^2 sont suggérés.

- Pour estimer la valeur moyenne correspondant à une observation donnée ou prédire la valeur d'une observation, on peut utiliser les critères MSE, $\overline{R^2}$, C_p de Mallows et PRESS. Des équations où $MSE_{p-r} \leq MSE$ sont fortement suggérées.

- Dans le but d'extrapoler, on peut utiliser les quatre mêmes critères. Étant donné la nature plus sensible de cet exercice, une condition plus restrictive est recommandée à savoir $\left[\frac{(n-p)+r}{(n-p)+1}\right] MSE_{p-r} \leq MSE.$

CHAPITRE 3

PROCÉDURES DE SÉLECTION SUCCESSIVES

L'étude des "meilleurs" sous-ensembles de variables peut s'avérer, de prime abord, une tâche très ardue. Une façon de procéder est de produire la liste de tous les sous-ensembles possibles et d'effectuer chacune des régressions. Il restera alors à faire un choix éclairé en utilisant les critères exposés au chapitre 2.

Le problème de cette approche est que si l'on considère le modèle complet à p-1 variables, chaque variable pouvant ou non appartenir au modèle, il existe 2^{p-1} sous-ensembles possibles. En considérant un modèle comportant 10 variables explicatives, on serait donc amené à étudier plus de mille sous-ensembles! De plus, bien que les ordinateurs se soient constamment perfectionnés au cours des années, le temps nécessaire pour générer 2^{p-1} régressions peut encore, de nos jours, être relativement long. Dans la suite, nous appellerons cette durée des calculs le "temps-machine".

Les procédures de sélection successives ont donc été créées en vue d'alléger ce fardeau. En effet, ces méthodes trouvent le "meilleur" sous-ensemble de variables pour toutes les tailles de sous-ensembles possibles.

Nous étudierons trois méthodes de sélection successives. Les deux premières sont appelées respectivement la sélection progressive ("stepwise regression") et la sélection ascendante ("forward selection"). Ces deux méthodes considèrent d'abord le modèle constant, à chaque étape, elles essaient d'ajouter une variable "significative" jusqu'à ce que l'on atteigne un critère d'arrêt donné. La seule différence entre ces procédures est que l'algorithme de sélection progressive prévoit des tests supplémentaires visant à juger si une variable entrée à une étape antérieure peut éventuellement être retirée si elle est devenue "non significative" par la suite. La troisième méthode s'appelle la sélection descendante ("backward elimination"). Celle-ci considère d'abord le modèle complet et tente, à chaque étape, de retirer une variable "non significative" jusqu'à ce qu'un critère d'arrêt, fixé à l'avance, soit atteint.

On établit qu'une variable faisant partie d'un ensemble est significative de la façon suivante: on trouve la valeur du paramètre $\hat{\beta}_i$ associé à la variable ainsi que son écart-type $s(\hat{\beta}_i)$. On calcule la statistique $t = \frac{\hat{\beta}_i}{s(\hat{\beta}_i)}$ qui est distribuée suivant une loi de Student à $n-k$ degrés de liberté. On s'intéresse alors au test $H_0: \beta_i = 0$ contre $H_1: \beta_i \neq 0$. On acceptera l'hypothèse H_0 que si $|t| \leq t_{\alpha/2}(n-k)$ où le terme de droite représente le quantile d'ordre $1-\alpha/2$ de la loi de Student. La majorité des logiciels utilisent la statistique $F = t^2$ qui est distribuée suivant une loi de Fisher $F(1, n-k)$. En respectant les quantiles, on peut établir que $F_\alpha(1, \nu) = t_{\alpha/2}(\nu)$.

Une notion qui intervient dans le choix des variables est le niveau critique ("p-value"). Le niveau critique correspond au plus petit niveau pour lequel on rejetterait l'hypothèse H_0 (nullité d'un paramètre donné) et ce, à partir des observations dont on dispose. Ainsi, si le niveau critique excède 5%, voire 10%, on aura toutes les raisons d'accepter l'hypothèse H_0 sur la nullité d'un paramètre.

3.1 Sélection progressive et ascendante

3.1.1 Description

Comme la sélection ascendante n'est rien d'autre qu'un cas particulier, nous décrirons la sélection progressive afin de comprendre le principe de façon générale.

La sélection progressive nécessite l'attribution de deux niveaux. Le premier niveau, $\alpha_{\text{entrée}}$, déterminera si une variable doit entrer ou non dans l'équation et le deuxième niveau, α_{conserve}, déterminera si on conserve ou si on retire une variable une fois entrée. En général, on choisit $\alpha_{\text{entrée}}$ égal à α_{conserve}. Bien que l'on puisse choisir α_{conserve} plus grand que $\alpha_{\text{entrée}}$, la relation inverse n'est pas recommandée car cela rendrait trop probable la suppression d'une variable une fois entrée, ce qui pourrait entraîner un effet de cycle.

En second lieu, on considère le modèle constant. Si ce modèle n'est pas déjà acceptable, on ajoutera des variables, une à une, à partir de leur degré de signification (par l'entremise des niveaux critiques) à chaque pas. Deux méthodes sont utilisées pour choisir les variables :

celle des coefficients de corrélation partiels et celle des tests-t (ou tests-F). La méthode des coefficients de corrélation consiste à calculer, dans un premier temps, le carré des corrélations entre les variables explicatives et la variable expliquée et à considérer la plus grande parmi celles-ci. On effectue une régression à l'aide de cette seule variable expliquée et on compare le niveau critique du test-t (ou du test-F) à $\alpha_{\text{entrée}}$. On conservera la variable à moins que le niveau critique n'excède $\alpha_{\text{entrée}}$ puisqu'alors la procédure serait terminée et on garderait le modèle constant $Y = \overline{Y}$. Si la variable est conservée, on passe à la deuxième étape qui consiste à calculer le carré du coefficient de corrélation partiel appelé le coefficient de détermination partiel. On parle de partiel car on ne calculera le coefficient de détermination que pour les variables qui ne sont pas encore dans le modèle. Si $X_1, X_2, \ldots, X_g$ sont déjà dans le modèle et que $X_{g+1}, X_{g+2}, \ldots, X_{p-1}$ n'y sont pas alors:

$$R^2(X_i | X_1, X_2, \ldots, X_g) = \frac{SSE_{X_1,X_2,\ldots,X_g} - SSE_{X_1,X_2,\ldots,X_g,X_i}}{SSE_{X_1,X_2,\ldots,X_g}}, \quad i = g+1, \ldots, p-1 .$$

Encore une fois, si la variable expliquée ayant le coefficient de corrélation partiel le plus élevé possède un niveau critique inférieur au seuil $\alpha_{\text{entrée}}$, elle entre dans le modèle. Sinon, la procédure se termine avec le modèle à une variable. On vérifie maintenant si la première variable entrée a toujours sa place dans le modèle. Pour ce, on compare son niveau critique à α_{conserve}. On retirera la variable que si le niveau critique excède α_{conserve}. On poursuit la procédure jusqu'à ce qu'aucune variable ne soit suffisamment significative pour entrer dans le modèle.

Pour ce qui est de la méthode des test-t (ou test-F), la différence réside dans le fait que l'on dresse une liste des modèles candidats avant d'en choisir un. Ainsi, la procédure débute en listant les valeurs des statistiques |t| (ou F) de tous les modèles à une variable et consiste à choisir la variable pour laquelle la statistique est la plus élevée. Cette stratégie s'explique par le fait que l'on recherche des variables où on rejettera $H_0: \beta_i = 0$, i.e. quand $|t| > t_{\alpha/2}$. On prendra également soin de vérifier si celle-ci est significative au niveau $\alpha_{\text{entrée}}$, faute de quoi la procédure se termine.

On dresse ensuite la liste de tous les modèles à deux variables contenant la première variable sélectionnée et encore une fois, on choisit celle qui offre la plus grande valeur de t (ou F) tout en étant significative au niveau $\alpha_{\text{entrée}}$. Par la suite, tout comme dans la procédure de sélection progressive, on vérifie que la première variable est toujours

significative au niveau $\alpha_{conserve}$, faute de quoi on la retire. Reste à poursuivre la procédure jusqu'à ce qu'aucune variable ne soit suffisamment significative pour entrer dans le modèle.

La sélection ascendante, quant à elle, n'est rien d'autre qu'une sélection progressive où on ne vérifie pas si une variable devient non significative suite à l'ajout d'autres variables. On est alors assuré qu'une variable entrée ne peut ressortir (ce qui peut être désirable dans certains domaines d'application).

3.1.2 Inconvénients

Un des inconvénients de ces méthodes est qu'elles ont tendance à fournir des résultats discutables en présence d'un grand nombre de variables; il arrive souvent que ces procédures se terminent hâtivement ignorant ainsi des variables cruciales.

Il est relativement facile de fabriquer des exemples, même de petites tailles, montrant une limite potentielle de ces méthodes; tout est relié au fait que la sélection de variables se fait une à une. Il se peut que deux variables réunies fournissent une excellente corrélation avec la variable expliquée mais qu'individuellement elles n'apportent que peu de contribution à l'explication. Un exemple tiré de MILLER [38] exploite cette idée. On considère les données "artificielles" suivantes:

Y	X1	X2	X3
-2	1000	1002	0
-1	-1000	-999	-1
1	-1000	-1001	1
2	1000	998	0

La variable Y est exactement égale à X1-X2 mais Y est orthogonale à X1 et est presque orthogonale à X2. La procédure ascendante choisira X3 comme première variable. La prochaine variable est X2 qui n'offre qu'une diminution très négligeable de la somme des carrés des erreurs et ne sera pas prise en compte pour cette raison. On notera que X1 et X2 réunies offrent une valeur parfaite du coefficient de détermination mais sont ignorées.

Le fait que X1 et X2 puissent bien expliquer conjointement la variable Y même si individuellement il n'en est rien peut être détecté par un examinateur remarquant une corrélation importante entre ces deux variables explicatives. Cependant, il est possible, tout comme MANTEL [36], de fabriquer des exemples où des indices d'un tel comportement est presque impossible à déterminer à partir de l'étude de la corrélation simple.

Supposons que pour un certain k les variables $X_1, X_2, \ldots, X_{k-1}$ sont mutuellement indépendantes entre elles tout comme avec la variable Y. À l'opposé, les variables $X_{k+1}, X_{k+2}, \ldots, X_{p-1}$ sont individuellement presque parfaitement corrélées avec la variable Y et leur corrélation multiple est également presque parfaite avec Y. Pour une certaine variable X_k, il est possible que la meilleure régression multiple sur les $p-1$ variables ne fasse intervenir que les k premières variables. Ceci peut survenir même si la corrélation entre X_k et Y est arbitrairement petite mais non nulle et que les corrélations de X_k avec $X_1, X_2, \ldots, X_{k-1}$ sont faibles, et ce, particulièrement lorsque k est grand. Pour ces fins, il suffit de définir $X_k = c_0 Y + c_1 X_1 + \ldots + c_{k-1} X_{k-1}$ (où les c_i, $i = 0, \ldots, k-1$, sont non nuls).

Supposons que σ_0^2 désigne la variance non conditionnelle de Y et que σ_i^2 désigne la variance de la variable X_i, $i = 1, \ldots, k-1$. La corrélation de X_k avec Y est donnée par $c_0 \sigma_0 / (\sum_{i=0}^{k-1} c_i^2 \sigma_i^2)^{\frac{1}{2}}$ qui peut être définie arbitrairement petite en valeur absolue en choisissant un c_0 arbitrairement près de 0 tout en fixant les autres c_i et les σ_i, $i = 1, \ldots, k-1$. La corrélation de X_k avec un quelconque X_i, $i = 1, \ldots, k-1$, est donnée par $c_i \sigma_i / (\sum_{i=0}^{k-1} c_i^2 \sigma_i^2)^{\frac{1}{2}}$. Pour c_0 arbitrairement petit et avec des c_i, $i = 1, \ldots, k-1$, égaux et avec des σ_i^2, $i = 1, \ldots, k-1$, également égaux entre eux on obtient des carrés de coefficients de corrélation approchant la valeur $1/k-1$.

Une procédure ascendante détecterait "à tort" les variables $X_{k+1}, X_{k+2}, \ldots, X_{p-1}$ puisque les variables $X_1, X_2, \ldots, X_k$ fournissent une corrélation parfaite.

3.2 Sélection descendante

3.2.1 Description

Cette procédure, débutant avec le modèle complet, nécessite l'introduction d'un seuil qui permettra de déterminer quelles variables peuvent être retirées sans trop compromettre la qualité de l'équation de régression. On calculera les valeurs des statistiques |t| (ou F) et on enlèvera la variable pour laquelle la valeur de cette statistique est la moins élevée, et ce, seulement si le niveau critique excède le seuil fixé à l'avance (habituellement 5% ou 10%). La procédure se terminera quand tous les niveaux critiques associés aux paramètres, à une étape donnée, seront inférieurs au seuil fixé.

3.2.2 Inconvénients

Contrairement aux deux méthodes précédentes, des résultats discutables auront tendance à survenir à partir d'un petit ensemble de variables. Cependant, cette méthode se comportera relativement bien en présence de beaucoup de variables.

Cette méthode ne permet pas à une variable une fois exclue de réintégrer l'équation ultérieurement. Il se peut que la présence d'une variable dans l'équation finale puisse contribuer à une réduction substantielle de la somme des carrés des erreurs mais qu'elle soit irrémédiablement exclue dès les premières étapes de la procédure. Un exemple conceptuel d'une telle situation est donnée par BEALE [3]. Il suppose qu'une composition chimique affecte la variable dépendante, disons le "rendement". Cette composition est formée de huit produits chimiques dont la somme représente des valeurs se situant entre 99.9 et 100% du total dans tous les cas. Le reste, sous forme de poussière, a un effet que l'on supposera négligeable. Si on ajuste une équation de régression avec terme constant sur ces données, alors un des huit éléments sera éliminé puisque, si la poussière est vraiment sans importance, une équation sur sept éléments est à peu près équivalente à une équation sur huit éléments. Une pure question de chance déterminera l'élément exclu. Cependant, l'élément retiré est possiblement le seul à vraiment avoir un effet sur la variable expliquée.

Un autre inconvénient de cette méthode (qui était pratiquement vue comme une objection à celle-ci il y a plusieurs années) est la grande quantité de temps-machine requis. En effet,

rappelons que la procédure considère d'abord le modèle complet. Ainsi, si celui-ci comporte un nombre relativement élevé de variables, le temps requis à chaque étape (où on n'enlève qu'une variable) est non négligeable.

3.3 Comparaisons entre les méthodes

Les procédures de sélection semblent venir simplifier la vie de l'analyste des données mais encore faut-il que les résultats soient comparables à ceux que l'on pourrait obtenir en considérant les meilleurs sous-ensembles issus de l'étude des 2^{p-1} possibilités.

BERK [4] a étudié les relations existant entre la sélection ascendante, descendante et globale ("all subsets"). La sélection globale, qui considère les 2^{p-1} possibilités, trouve le meilleur sous-ensemble pour toutes les tailles possibles en utilisant un critère de sélection donné. La plupart du temps, le critère de coefficient de détermination maximal est utilisé.

Afin d'éviter de faire intervenir des critères d'arrêt, la sélection ascendante sera menée jusqu'au modèle complet et la sélection descendante jusqu'au modèle constant. On dira que deux procédures sont en accord si, pour toutes les tailles de sous-ensembles possibles, la composition de ces sous-ensembles sont identiques. Nous supposerons que les trois méthodes choisissent leurs sous-ensembles en se basant sur le critère du coefficient de détermination maximal à chaque étape.

Théorème

La sélection ascendante est en accord avec la sélection globale pour toutes les tailles de sous-ensembles si et seulement si la sélection descendante est en accord avec la sélection globale pour toutes les tailles de sous-ensembles.

Démonstration :

Supposons que les sélections ascendantes et globales soient en accord pour toutes les tailles de sous-ensembles. Alors il faut nécessairement que l'ensemble à k+1 variables qui est choisi contienne les variables de l'ensemble sélectionné comportant k variables puisque la sélection ascendante produit toujours des sous-ensembles de la sorte. Dans le suite, on désignera cette caractéristique par l'appellation "sous-ensembles emboîtés".

La sélection descendante choisit, à chaque étape, le sous-ensemble ayant le coefficient de détermination le plus élevé sous la contrainte que les sous-ensembles doivent être emboîtés. Comme les sous-ensembles ayant le coefficient R^2 maximal sont justement emboîtés (pour chacune des tailles possibles) la procédure descendante choisit, à chaque étape, exactement les mêmes ensembles puisqu'elle utilise le même critère de sélection. On en conclût que la sélection descendante est en accord avec la sélection globale si la sélection ascendante l'est. Un raisonnement identique permet d'établir la relation inverse. □

Ce théorème nous apprend qu'une procédure est en désaccord avec la sélection globale si et seulement si l'autre procédure l'est également. Il en découle que si les procédures ascendantes et descendantes sont en désaccord alors aucune d'entre-elles n'est en accord avec la sélection globale.

La question de savoir si un accord entre les sélections ascendantes et descendantes entraîne un accord commun avec la sélection globale a été soulevée par HAMAKER [25]. Il n'a offert aucune preuve de cette affirmation. Plusieurs années plus tard, BERK [4] a montré que cette conjecture était fausse.

Supposons qu'il y ait quatre variables explicatives X_1, X_2, X_3 et X_4. Après quelques manipulations, on découvre que les meilleurs sous-ensembles par la méthode de sélection globale sont respectivement (X_1), (X_2, X_3) et (X_1, X_2, X_4); le sous-ensemble (X_1, X_2) constituant la deuxième meilleure paire. Dans un tel cas, les procédures ascendantes et descendantes choisiraient les sous-ensembles (X_1), (X_1, X_2) et (X_1, X_2, X_4) créant un désaccord avec la sélection globale.

Pour ce, nous définissons $X_1, \ldots, X_4$ de sorte que X_1 soit la mieux corrélée à Y bien qu'elle explique piètrement cette dernière. Les variables X_2 et X_3 seront définies de sorte à expliquer peu individuellement mais beaucoup une fois réunies, tandis que les variables X_1 et X_2 constitueront la deuxième meilleure paire. La variable X_4, quant à elle, sera créée de sorte à faire de (X_1, X_2, X_4) le meilleur ensemble à trois variables.

Supposons que $Z_1, \ldots, Z_5$ soient des variables orthonormales, i.e. mutuellement orthogonales (non-corrélées) avec moyenne 0 et variance 1.

Soient $0 < \delta < \gamma < \eta < 1$ et définissons

$Y = Z_5$,

$X_1 = Z_1 + \eta Z_5$,

$X_2 = Z_2 - Z_3 + \eta Z_5$,

$X_3 = -Z_2 + Z_3 + \gamma Z_4 + \eta Z_5$,

$X_4 = -Z_1 - Z_2 + Z_3 + \delta Z_4$.

Si A désigne la matrice des coefficients des variables $Z_1,\ldots,Z_5$, alors la matrice de variance-covariance de (Y, X_1, X_2, X_3, X_4) sera donnée par $A\ \mathrm{Cov}(Z)\ A' = A\ A'$.

Soient $\delta = 0.015$, $\gamma = 0.02$ et $\eta = 0.1$. Les coefficients de détermination pour les ensembles à une variable (carré des coefficients de corrélation entre Y et X_i $i = 1,\ldots,4$) sont donnés par:

$$R^2_{X_1} = \frac{\eta^2}{1+\eta^2} = 0.009901,$$

$$R^2_{X_2} = \frac{\eta^2}{2+\eta^2} = 0.0049751,$$

$$R^2_{X_3} = \frac{\eta^2}{2+\gamma^2+\eta^2} = 0.0049741 \text{ et}$$

$$R^2_{X_4} = 0$$

On a:

$$R^2 = 1 - \frac{SSE}{SST} = 1 - SSE \quad (\text{car } SST = Y'Y = Z_5'Z_5 = 1).$$

On peut alors vérifier que le coefficient R^2 maximal pour les ensembles à deux variables est donné par

$$R^2_{X_2,X_3} = \frac{\eta^2(8+\gamma^2)}{\eta^2(8+\gamma^2)+2\gamma^2} = 0.9900995, \text{ suivi de}$$

$$R^2_{X_1,X_2} = \frac{3\eta^2}{3\eta^2+2} = 0.0147783.$$

Parmi les triplets, le R^2 maximal est donné par

$$R^2_{X_1,X_2,X_4} = \frac{\eta^2(8+3\delta^2)}{\eta^2(8+3\delta^2)+2\delta^2} = 0.9944069, \text{ suivi de}$$

$$R^2_{X_1,X_2,X_3} = \frac{\eta^2(8+3\gamma^2)}{\eta^2(8+3\gamma^2)+2\delta^2} = 0.9901005.$$

Ici, le choix de l'ensemble impliquant X_1 et X_2 comme modèle à deux variables par les procédures ascendantes et descendantes apparaît insensé. En gardant en tête que cet exemple ne reflète pas nécessairement une situation qui risque de survenir fréquemment avec des données réelles, il porte quand même à réflexion.

3.4 Exemple d'utilisation des procédures sur les données de Hald

Nous allons maintenant appliquer les procédures de sélection successives sur l'ensemble de données de Hald. Rappelons que celui-ci est reproduit à l'annexe 4.

On peut retrouver à l'annexe 5, les étapes détaillées des trois procédures appliquées sur ces données. Dans un premier temps, nous allons suivre en détail la procédure de sélection progressive qui utilise les niveaux $\alpha_{\text{entrée}} = \alpha_{\text{conserve}} = 0,1$. Par la suite, les résultats des deux autres méthodes seront brièvement commentés.

La procédure de sélection progressive débute par l'ajustement des quatre modèles à une variable. C'est le modèle impliquant la variable X4 qui produit la plus grande valeur de la statistique $F = t^2 = \left(\frac{\hat{\beta}_i}{s(\hat{\beta}_i)}\right)^2$ avec 22,80. Au niveau $\alpha_{\text{entrée}} = 0,1$, cette variable est significative puisque F=22,80 > 3,23=$F_{0,1}(1,11)$. On combine alors X4 avec chacune des trois variables restantes. La statistique F calculée à partir de l'estimateur du paramètre associé à la variable X1 produit la valeur la plus élevée. Puisque F=108,22 > 3,28=$F_{0,1}(1,10)$, X1 est significative au niveau $\alpha_{\text{entrée}}$. De plus, au niveau $\alpha_{\text{conserve}} = 0,1$,

X4 demeure dans le modèle puisque 159,30 > 3,28. À la troisième étape, X2 est ajoutée. Elle est significative puisque F=5,03 > 3,36= $F_{0,1}(1,9)$. On remarque que X1 demeure significative, ce n'est cependant pas le cas pour X4 qui, avec une valeur de F=1,86 doit quitter le modèle. On reste alors avec le modèle à trois paramètres impliquant X1 et X2. À l'étape 4, on estime à nouveau les paramètres associés à ces variables et on établit que les variables sont significatives. Par la suite, la procédure tente d'ajouter la variable X3 qui, tout comme X4, produit une valeur de la statistique F telle que le niveau critique excède $\alpha_{\text{entrée}}$. Le modèle retenu est donc $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X1 + \hat{\beta}_2 X2$.

Pour ce qui est de la procédure de sélection ascendante menée au niveau 0,1, les trois premières étapes sont identiques à la procédure de sélection progressive. Comme cette méthode ne peut rejeter de variables, elle tente d'ajouter X3 aux trois variables déjà dans le modèle. Cependant, celle-ci n'est pas significative au niveau 0,1. Le modèle retenu est donc $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X1 + \hat{\beta}_2 X2 + \hat{\beta}_4 X4$.

Pour ce qui a trait à la procédure de sélection descendante, le déroulement est très simple à suivre. Au niveau 0,1, on enlève à chaque étape la variable générant le niveau critique le plus élevé tout en étant supérieur à 0,1. On enlève donc X3, puis X4. On demeure avec X1 et X2 pour lesquelles on rejette l'hypothèse nulle $H_0: \hat{\beta}_i = 0,\ i = 1, 2$. Ici, tout comme pour la procédure de sélection progressive, le modèle retenu contient X1 et X2.

Rappelons que c'est le même modèle que celui qui avait été choisi à la section 2.5. À partir de cet ensemble de petite taille, on peut vérifier la validité des conclusions résultant des méthodes de sélection automatisées présentées dans ce chapitre. Cependant, sur des ensembles comportant beaucoup de variables, l'analyste doit se fier, jusqu'à un certain point, aux résultats issus des procédures de sélection successives.

3.5 Autres méthodes

Nous savons maintenant que les procédures présentées dans ce chapitre ne trouvent pas nécessairement le "meilleur" sous-ensemble; souvent celui-ci sera toutefois raisonnable. Certaines alternatives ont été proposées et deux d'entre elles sont brièvement exposées dans DRAPER & SMITH [14].

Une de ces alternatives consiste à effectuer une sélection progressive en spécifiant certains niveaux (valeurs de α) permettant de contrôler l'ajout ou le retrait des variables. Lorsque la procédure se termine, il suffit de noter le nombre de variables dans l'équation finale, disons q , et d'effectuer toutes les régressions utilisant ces q variables. Il restera à choisir le "meilleur" sous-ensemble déterminé à partir des critères de sélection comme ceux présentés au chapitre 2.

Cette approche demande un peu plus de temps-machine qu'une simple sélection progressive mais permet de faire un choix éclairé dans l'éventualité où il y aurait deux modèles de qualité presque équivalente comme c'est le cas avec les modèles à deux variables de l'ensemble de données de Hald.

Une autre procédure suggérée consiste à utiliser une procédure de sélection progressive avec des niveaux très peu restrictifs (i.e. des valeurs élevées pour α). Ceci forcera la procédure à ajouter un certain nombre de variables qui ne l'aurait pas été avec les niveaux "habituels". Nous pourrons donc étudier de nouvelles variables, ce qui peut mener à un modèle différent.

Cette procédure semble raisonnable lorsque certaines variables sont grandement corrélées entre elles quoique dans de tels cas il vaut mieux éviter complètement les procédures automatisées.

3.6 Des méthodes de sélection successives pour traiter la colinéarité ?

On parlera de multicolinéarité, de colinéarité multiple ou simplement de colinéarité lorsqu'une dépendance linéaire existe entre un sous-ensemble de variables. La dépendance peut être parfaite ou presque parfaite. Bien que la colinéarité soit un problème avec les données et non pas, nécessairement, un problème avec le modèle, une approche visant à résoudre ce problème consiste à considérer un modèle différent. On peut penser qu'après mûre réflexion au niveau des concepts, certaines variables peuvent servir à en représenter d'autres.

Vues sous cet angle, des corrélations élevées entre les variables explicatives indiqueront un haut niveau de fiabilité; un point visiblement positif. Pour ce, un analyste des données

pourra choisir de prendre certaines variables hautement corrélées entre elles et de les traiter comme des indicateurs généraux lors de l'analyse.

Une des approches visant à traiter les problèmes de colinéarité consiste à utiliser les procédures de sélection successives de façon à réduire le nombre de variables tout en ayant un ensemble moins corrélé qu'initialement.

Selon FOX [19], ces méthodes sont fréquemment utilisées de façon abusive par certains chercheurs qui veulent interpréter l'ordre d'entrée des variables dans l'équation comme un indice de leur "importance". On sait bien que ceci est faux car dans une sélection progressive, par exemple, une variable entrée tôt dans la procédure peut éventuellement être retirée. On peut penser également à un cas où deux variables explicatives grandement corrélées entre elles ont toutes les deux une corrélation élevée et presqu'identique avec la variable expliquée. Une seule de ces variables entrera dans l'équation puisque l'autre ne peut apporter que très peu d'information additionnelle. Il suffirait alors d'une petite modification aux données, ou encore d'un nouvel échantillon, et l'autre variable pourrait être celle qui est choisie.

CONCLUSION

Ce mémoire a fait l'objet d'une étude à la fois descriptive et critique des méthodes de sélection du "meilleur" sous ensemble de variables. La régression linéaire étant un sujet des plus vastes, nous avons dû établir un certain nombre de conventions permettant de restreindre quelque peu le champ d'étude avant d'arriver aux conclusions énoncées.

Néanmoins, il nous semble important de revenir sur certains points que nous avons choisis de ne pas aborder. En particulier, nous faisions remarquer qu'une équation de régression peut être construite non seulement sur les variables indépendantes mais également sur des transformations (linéaires ou non) de celles-ci. Notons que le sujet des transformations, à lui seul, est traité dans plusieurs livres généraux sur la régression sans compter tous les articles qui y sont consacrés.

Sommairement, l'idée derrière ces transformations est d'améliorer l'ajustement d'un modèle aux données, ce qui constitue définitivement un atout. Cependant, si le modèle doit être interprété, ces modifications rendent très ardue toute tentative à cet égard. Par exemple, imaginons un modèle sans transformation qui nous permet d'affirmer qu'une augmentation du poids, disons X_1, accentue le risque de maladies cardio-vasculaires, disons Y (en gardant les autres variables fixes). Ce modèle se compare avantageusement à un autre, mieux ajusté, qui affirme qu'une augmentation de $\log(X_1^3)$ accentue Y.

Nous avons également exploité uniquement la méthode des moindres carrés qui ne constitue pas une méthode robuste à l'égard des valeurs extrêmes. Nous n'avons qu'effleuré le sujet de la robustesse en ne présentant que certains critères de sélection robuste. Ce sujet est encore plus vaste que le précédent puisqu'on retrouve des ouvrages complets sur ce thème.

En terminant, nous espérons que ce travail permettra aux utilisateurs de la régression linéaire de profiter des nombreux exemples de ce mémoire pour éviter les pièges cachés des méthodes de sélection de variables. Profitons également de l'occasion pour rappeler qu'il vaut mieux passer plus de temps à examiner attentivement les données pour déceler des problèmes potentiels que de se fier uniquement aux résultats des méthodes de sélection en espérant qu'elles le feront à notre place!

ANNEXE 1

- Les densités des lois sont indexées par X
- $\Gamma(\alpha)=\int_0^{\infty} e^{-x}x^{\alpha-1}dx$ $(\Gamma(\alpha)=(\alpha-1)!$ si $\alpha>0$ est un entier)

Loi de X

Normale (*)
$N(\mu,\sigma^2)$
$\mu\in\mathbb{R},\ \sigma\geq 0$

$$N_X(\mu,\sigma^2)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}} \quad \text{où } x\in\mathbb{R}$$

Khi-deux (**)
$\chi^2(r,\lambda)$
$r\geq 1,\ \lambda\geq 0$

$$\chi^2_X(r,\lambda)=\sum_{k=0}^{\infty}\left(\frac{e^{-\lambda}\lambda^k}{k!}\right)\frac{x^{(r+2k-2)/2}e^{-x/2}}{\Gamma\left(\frac{r+2k}{2}\right)2^{k+r/2}} \quad \text{où } x\in(0,\infty)$$

Fisher (**)
$F(r_1,r_2,\lambda)$
$r_1,r_2\geq 1,\ \lambda\geq 0$

$$F_X(r_1,r_2,\lambda)=\sum_{k=0}^{\infty}\left(\frac{e^{-\lambda}\lambda^k}{k!}\right)\frac{\Gamma\left(\frac{2k+r_1+r_2}{2}\right)\left(\frac{r_1}{r_2}\right)^{(r_1+2k)/2}x^{(r_1+2k-2)/2}}{\Gamma\left(\frac{r_2}{2}\right)\Gamma\left(\frac{2k+r_1}{2}\right)\left(1+\frac{r_1}{r_2}x\right)^{(r_1+r_2+2k)/2}} \quad \text{où } x\in(0,\infty)$$

Student (centrée)
$t(r)$
$r\geq 1$

$$t_X(r)=\frac{\Gamma\left(\frac{r+1}{2}\right)}{\Gamma\left(\frac{r}{2}\right)\sqrt{r\pi}}\left(1+\frac{x^2}{r}\right)^{-(r+1)/2} \quad \text{où } x\in\mathbb{R}$$

Beta (centrée)
$Beta(\alpha,\beta)$
$\alpha,\beta>0$

$$Beta_X(\alpha,\beta)=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1} \quad \text{où } 0<x<1$$

* = loi centrée si $\mu=0$

** = loi centrée si $\lambda=k=0$ où on pose $0^0=1$

ANNEXE 2

Nous rappelons ici quelques propriétés très connues sur les matrices.

(i) Rang d'une matrice

Le rang d'une matrice A correspond au nombre de lignes, ou de colonnes, linéairement indépendantes et est noté $\mathrm{rg}(A)$.

* Si A et C sont non singulières alors $\mathrm{rg}(ABC) = \mathrm{rg}(B)$, pour toute matrice B tel que le produit ABC soit bien défini.

(ii) Trace d'une matrice

La trace d'une matrice carrée A est définie comme étant la somme des éléments sur la diagonale de A et est notée $\mathrm{tr}(A)$.

* $\mathrm{tr}(AB) = \mathrm{tr}(BA)$. (si les produits sont bien définis)

(iii) Matrice idempotente

Une matrice carrée A est dite idempotente si $AA = A$.

* Si A est idempotente alors $\mathrm{tr}(A) = \mathrm{rg}(A)$.

(iv) Matrice orthogonale

Une matrice carrée A est dite orthogonale si et seulement si A est inversible et $A^{-1} = A'$.

* Si A est orthogonale alors $A'A = A\,A' = I$.

(v) Paires de matrices symétriques

* Si A et B sont symétriques alors il existe une matrice orthogonale V telle que $V'A\,V$ et $V'B\,V$ sont diagonales si et seulement si AB est symétrique.

(vi) Diagonalisation

* Si A est de dimension $m \times n$ et de rang r alors il existe des matrices non singulières P et Q telles que $PAQ = \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix}$ où I_r désigne la matrice identité de dimension $r \times r$.

ANNEXE 3

Nous fournissons ici les données utilisées pour les exemples présentés dans la section 2.1.3

Tableau des données de la Figure (a)

X	Y
1,0	0,5
1,0	0,6
1,0	0,7
1,0	0,8
1,0	0,9
1,0	1,0
1,0	1,1
1,0	1,2
1,0	1,3
3,5	3,0
3,5	3,1
3,5	3,2
3,5	3,3
3,5	3,4
3,5	3,5
3,5	3,6
3,5	3,7
3,5	3,8

Tableau des données de la Figure (b)

X	Y
2,0	1,5
2,0	1,6
2,0	1,7
2,0	1,8
2,0	1,9
2,0	2,0
2,0	2,1
2,0	2,2
2,0	2,3
2,5	2,0
2,5	2,1
2,5	2,2
2,5	2,3
2,5	2,4
2,5	2,5
2,5	2,6
2,5	2,7
2,5	2,8

Tableau des données de la Figure (c)

X	Y
1,0	0,5
1,0	0,6
1,0	0,7
1,0	0,8
1,0	0,9
1,0	1,0
1,0	1,1
1,0	1,2
1,0	1,3
1,5	1,0
1,5	1,1
1,5	1,2
1,5	1,3
1,5	1,4
1,5	1,5
1,5	1,6
1,5	1,7
1,5	1,8
2,0	1,5
2,0	1,6
2,0	1,7
2,0	1,8
2,0	1,9
2,0	2,0
2,0	2,1
2,0	2,2
2,0	2,3
2,5	2,0
2,5	2,1
2,5	2,2
2,5	2,3
2,5	2,4
2,5	2,5
2,5	2,6
2,5	2,7
2,5	2,8
3,0	2,5
3,0	2,6
3,0	2,7
3,0	2,8
3,0	2,9
3,0	3,0
3,0	3,1
3,0	3,2
3,0	3,3
3,5	3,0
3,5	3,1
3,5	3,2
3,5	3,3
3,5	3,4
3,5	3,5
3,5	3,6
3,5	3,7
3,5	3,8

Tableau des données de la Figure (d)

X	Y
2,0	2,0
3,0	2,0
4,0	3,0
6,0	6,0
7,0	7,0
8,0	7,0

Tableau des données de la Figure (e)

X	Y
2,0	4,0
3,0	3,9
4,0	3,9
6,0	4,1
7,0	4,1
8,0	4,0

Tableau des données de la Figure (f)

X	Y
1,0	0,9
2,0	2,2
3,0	2,9

Tableau des données de la Figure (g)

X	Y
1,0	0,2
1,0	0,9
1,0	1,6
2,0	1,5
2,0	2,2
2,0	2,9
3,0	2,2
3,0	2,9
3,0	3,6

ANNEXE 4

Données de Hald

Y	X1	X2	X3	X4
78,5	7	26	6	60
74,3	1	29	15	52
104,3	11	56	8	20
87,6	11	31	8	47
95,9	7	52	6	33
109,2	11	55	9	22
102,7	3	71	17	6
72,5	1	31	22	44
93,1	2	54	18	22
115,9	21	47	4	26
83,8	1	40	23	34
113,3	11	66	9	12
109,4	10	68	8	12

Matrices des corrélations

	X1	X2	X3	X4	Y
X1	1				
X2	0,2286	1			
X3	**-0,8241**	-0,1392	1		
X4	-0,2454	**-0,9730**	0,0295	1	
Y	0,7307	0,8163	-0,5347	-0,8213	1

ANNEXE 5

Stepwise Procedure for Dependent Variable Y

Step 1 Variable X4 Entered R-square = 0.67454196 C(p) =138.73083349

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	1	1831.89616002	1831.89616002	22.80	0.0006
Error	11	883.86691690	80.35153790		
Total	12	2715.76307692			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	117.56793118	5.26220651	40108.47690796	499.16	0.0001
X4	-0.73816181	0.15459600	1831.89616002	22.80	0.0006

Bounds on condition number: 1, 1

Step 2 Variable X1 Entered R-square = 0.97247105 C(p) = 5.49585082

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	2	2641.00096477	1320.50048238	176.63	0.0001
Error	10	74.76211216	7.47621122		
Total	12	2715.76307692			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	103.09738164	2.12398361	17614.67006622	2356.10	0.0001
X1	1.43995828	0.13841664	809.10480474	108.22	0.0001
X4	-0.61395363	0.04864455	1190.92463664	159.30	0.0001

Bounds on condition number: 1.064105, 4.256421

Step 3 Variable X2 Entered R-square = 0.98233545 C(p) = 3.01823347

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	3	2667.79034752	889.26344917	166.83	0.0001
Error	9	47.97272940	5.33030327		
Total	12	2715.76307692			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	71.64830697	14.14239348	136.81003409	25.67	0.0007
X1	1.45193796	0.11699759	820.90740153	154.01	0.0001
X2	0.41610976	0.18561049	26.78938276	5.03	0.0517
X4	-0.23654022	0.17328779	9.93175378	1.86	0.2054

Bounds on condition number: 18.94008, 116.3601

Step 4 Variable X4 Removed R-square = 0.97867837 C(p) = 2.67824160

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	2	2657.85859375	1328.92929687	229.50	0.0001
Error	10	57.90448318	5.79044832		
Total	12	2715.76307692			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	52.57734888	2.28617433	3062.60415609	528.91	0.0001
X1	1.46830574	0.12130092	848.43186034	146.52	0.0001
X2	0.66225049	0.04585472	1207.78226562	208.58	0.0001

Bounds on condition number: 1.055129, 4.220516

All variables left in the model are significant at the 0.1000 level.
No other variable met the 0.1000 significance level for entry into the model.

Summary of Stepwise Procedure for Dependent Variable Y

Step	Variable Entered	Variable Removed	Number In	Partial R**2	Model R**2	C(p)	F	Prob>F
1	X4		1	0.6745	0.6745	138.7308	22.7985	0.0006
2	X1		2	0.2979	0.9725	5.4959	108.2239	0.0001
3	X2		3	0.0099	0.9823	3.0182	5.0259	0.0517
4		X4	2	0.0037	0.9787	2.6782	1.8633	0.2054

Forward Selection Procedure for Dependent Variable Y

Step 1 Variable X4 Entered R-square = 0.67454196 C(p) =138.73083349

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	1	1831.89616002	1831.89616002	22.80	0.0006
Error	11	883.86691690	80.35153790		
Total	12	2715.76307692			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	117.56793118	5.26220651	40108.47690796	499.16	0.0001
X4	-0.73816181	0.15459600	1831.89616002	22.80	0.0006

Bounds on condition number: 1, 1

Step 2 Variable X1 Entered R-square = 0.97247105 C(p) = 5.49585082

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	2	2641.00096477	1320.50048238	176.63	0.0001
Error	10	74.76211216	7.47621122		
Total	12	2715.76307692			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	103.09738164	2.12398361	17614.67006622	2356.10	0.0001
X1	1.43995828	0.13841664	809.10480474	108.22	0.0001
X4	-0.61395363	0.04864455	1190.92463664	159.30	0.0001

Bounds on condition number: 1.064105, 4.256421

Step 3 Variable X2 Entered R-square = 0.98233545 C(p) = 3.01823347

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	3	2667.79034752	889.26344917	166.83	0.0001
Error	9	47.97272940	5.33030327		
Total	12	2715.76307692			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	71.64830697	14.14239348	136.81003409	25.67	0.0007
X1	1.45193796	0.11699759	820.90740153	154.01	0.0001
X2	0.41610976	0.18561049	26.78938276	5.03	0.0517
X4	-0.23654022	0.17328779	9.93175378	1.86	0.2054

Bounds on condition number: 18.94008, 116.3601

No other variable met the 0.1000 significance level for entry into the model.

Summary of Forward Selection Procedure for Dependent Variable Y

Step	Variable Entered	Number In	Partial R**2	Model R**2	C(p)	F	Prob>F
1	X4	1	0.6745	0.6745	138.7308	22.7985	0.0006
2	X1	2	0.2979	0.9725	5.4959	108.2239	0.0001
3	X2	3	0.0099	0.9823	3.0182	5.0259	0.0517

Backward Elimination Procedure for Dependent Variable Y

Step 0 All Variables Entered R-square = 0.98237562 C(p) = 5.00000000

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	4	2667.89943757	666.97485939	111.48	0.0001
Error	8	47.86363935	5.98295492		
Total	12	2715.76307692			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	62.40536930	70.07095921	4.74551686	0.79	0.3991
X1	1.55110265	0.74476987	25.95091138	4.34	0.0708
X2	0.51016758	0.72378800	2.97247824	0.50	0.5009
X3	0.10190940	0.75470905	0.10909005	0.02	0.8959
X4	-0.14406103	0.70905206	0.24697472	0.04	0.8441

Bounds on condition number: 282.5129, 2489.203

Step 1 Variable X3 Removed R-square = 0.98233545 C(p) = 3.01823347

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	3	2667.79034752	889.26344917	166.83	0.0001
Error	9	47.97272940	5.33030327		
Total	12	2715.76307692			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	71.64830697	14.14239348	136.81003409	25.67	0.0007
X1	1.45193796	0.11699759	820.90740153	154.01	0.0001
X2	0.41610976	0.18561049	26.78938276	5.03	0.0517
X4	-0.23654022	0.17328779	9.93175378	1.86	0.2054

Bounds on condition number: 18.94008, 116.3601

Step 2 Variable X4 Removed R-square = 0.97867837 C(p) = 2.67824160

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	2	2657.85859375	1328.92929687	229.50	0.0001
Error	10	57.90448318	5.79044832		
Total	12	2715.76307692			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	52.57734888	2.28617433	3062.60415610	528.91	0.0001
X1	1.46830574	0.12130092	848.43186034	146.52	0.0001
X2	0.66225049	0.04585472	1207.78226562	208.58	0.0001

Bounds on condition number: 1.055129, 4.220516

All variables left in the model are significant at the 0.1000 level.

Summary of Backward Elimination Procedure for Dependent Variable Y

Step	Variable Removed	Number In	Partial R**2	Model R**2	C(p)	F	Prob>F
1	X3	3	0.0000	0.9823	3.0182	0.0182	0.8959
2	X4	2	0.0037	0.9787	2.6782	1.8633	0.2054

BIBLIOGRAPHIE

[1] ALLEN (D.M.), The Prediction Sum of Squares as a Criterion for Selecting Predictor Variables, Technical Report No. 23 (University of Kentucky, Department of Statistics, 1971).

[2] BARRETT (J. P.), The Coefficient of Determination - Some Limitations, The Amer. Statistician, 28, (1974) pp. 19-20.

[3] BEALE (E. M. L.), Note on Procedures for Variable Selection in Multiple Regression, Technometrics, 12, (1970) pp. 909-914.

[4] BERK (K. N.), Comparing Subset Regression Procedures, Technometrics, 20, (1978) pp. 1-6.

[5] BOWERMAN (B. L.) et O'CONNELL (R. T.), Linear Statistical Models: An Applied Approach, 2nd ed. (PWS-Kent, Boston, 1990).

[6] BOX (G. E. P.), Use and Abuse of Regression, Technometrics, 8, (1966) pp. 625-629.

[7] BOX (G.E.P.) et WETZ (J.), Criteria for Judging Adequacy of Estimation by an Approximating Response, Technical Report No. 9 (University of Wisconsin, Statistics Department, 1973).

[8] CHANG (P. C.) et AFIFI (A. A.), Goodness-of-Fit Statistics for General Linear Regression Equations in the Presence of Replicated Responses, The Amer. Statistician, 41, (1987) pp.195-199.

[9] COX (D. R.), Regression Methods, J. R. Statist. Soc. A, 131, (1968) pp. 265-279.

[10] COX (D. R.) et SNELL (E. J.), The Choice of Variables in Observational Studies, Appl. Statist., 23, (1974) pp. 51-59.

[11] CROCKER (D. C.), Some Interpretations of the Multiple Correlation Coefficient, The Amer. Statistician, 26, (1972) pp. 31-33.

[12] DANIEL (C.) et WOOD (F.), Fitting Equations to Data: Computer Analysis of Multivariate Data, 2nd ed. (Wiley, New York, 1980).

[13] DARLINGTON (R.B.), Book Review, Appl. Psychol. Meas., 6, (1982) pp. 245-246.

[14] DRAPER (N.) et SMITH (H.), Applied Regression Analysis, 2nd ed. (Wiley, New York, 1981).

[15] DRAPER (N. R.), The Box-Wetz Criterion Versus R^2, J. R. Statist. Soc. A, 147, (1984) pp. 100-103.

[16] DRAPER (N. R.), Corrections - The Box-Wetz Criterion Versus R^2, J. R. Statist. Soc. A, 148, (1985) p. 357.

[17] EDWARDS (J. B.), The Relation Between the F-Test and $\overline{R}^2$, The Amer. Statistician, 23, (1969) p. 28.

[18] EZEKIEL (M.), Methods of Correlation Analysis (Wiley, New York, 1930).

[19] FOX (J.), Applied Regression Analysis, Linear Models and Related Methods (Sage, Thousand Oaks, Ca., 1997).

[20] GRAYBILL (F. A.), Theory and Application of the Linear Model (Duxbury Press, North Scituate, Mass., 1976).

[21] GORMAN (J. W.) et TOMAN (R. J.), Selection of Variables for Fitting Equations to Data, Technometrics, 8, (1966) pp. 27-51.

[22] HAHN (G. J.), The Coefficient of Determination Exposed!, Chemtech, 3, (1973) pp. 609-612.

[23] HAIR (J. F. Jr.) et al., Multivariate Data Analysis With Readings, 4th ed. (Prentice-Hall, Englewood Cliffs, N. J., 1995).

[24] HALD (A.), Statistical Theory with Engineering Applications (Wiley, New York, 1952)

[25] HAMAKER (H. C.), On Multiple Regression Analysis, Statistica Neerlandica, 16, (1962) pp. 31-56.

[26] HEALY (M. J. R.), The Use of R^2 as a Measure of Goodness of Fit, J. R. Statist. Soc. A, 147, (1984) pp. 608-609.

[27] HOCKING (R. R.), Criteria for Selection of a Subset Regression: Which One Should Be Used? Technometrics, 14, (1972) pp. 967-971.

[28] HOCKING (R. R.), Developments in Linear Regression Methodology: 1959-1982, Technometrics, 25, (1983) pp. 219-249.

[29] HOCKING (R. R.), The Analysis and Selection of Variables in Linear Regression, Biometrics, 32, (1976) pp. 1-49.

[30] HOCKING (R. R.), The Analysis of Linear Models, (Brooks/Cole, Monterey, Ca., 1985).

[31] HUBER (P. J.), Robust Statistics (Wiley, New York, 1981).

[32] KENNARD (R. W.), A Note on the C_p Statistic, Technometrics, 13, (1971) pp. 899-900.

[33] KVALSETH (T. O.), Cautionary Note About R^2, The Amer. Statistician, 39, (1985) pp. 279-285.

[34] MALLOWS (C. L.), More Comments on C_p, Technometrics, 37, (1995) pp. 362-372.

[35] MALLOWS (C. L.), Some Comments on C_p, Technometrics, 15, (1973) pp. 661-675.

[36] MANTEL (N.), Why Stepdown Procedures in Variable Selection, Technometrics, 12, (1970) pp. 621-625.

[37] MASON (R.L.) et al., Regression Analysis and Problems of Multicolinearity, Comm. in Statist., 4, (1975) pp. 277-292.

[38] MILLER (A. J.), Selection of Subsets of Regression Variables, J. R. Statist. Soc. A, 147, (1984) pp. 389-425.

[39] MONTGOMERY (D. C.) et PECK (E. A.), Introduction to Linear Regression Analysis (Wiley, New York, 1982).

[40] NETER (J.), WASSERMAN (W.) et KUTNER (M. H.), Applied Linear Statistical Models: Regression Analysis of Variance and Experimental Designs, 3rd ed. (Irwin, Homewood, Il., 1990).

[41] RAO (C. R.), Linear Models: Least Squares and Alternatives (Springer, New York, 1995).

[42] RAO (C. R.), Linear Statistical Inference and its Applications, (Wiley, New York, 1965).

[43] RONCHETTI (E.) et STAUDTE (R.G.), *A Robust Version of Mallow's C_p*, J. Amer. Statist. Assoc., 89, (1994) pp. 550-559.

[44] SEARLE (S. R.), Linear Models (Wiley, New York, 1971).

[45] SEBER (G. A. F.), Linear Regression Analysis (Wiley, New York, 1977).

[46] TOMASSONE (R.) et al., La régression: nouveaux regards sur une ancienne méthode statistique, 2nd ed. (Masson, Paris, 1992).