Régression linéaire multiple ou modèle gaussien

Résumé

Introductions au modèle linéaire et modèle linéaire général. Retour au plan du cours.

1 Introduction

Le modèle de régression linéaire multiple est l'outil statistique le plus habituellement mis en œuvre pour l'étude de données multidimensionnelles. Cas particulier de modèle linéaire, il constitue la généralisation naturelle de la régression simple.

2 Modèle

Une variable quantitative Y dite *à expliquer* (ou encore, réponse, exogène, dépendante) est mise en relation avec p variables quantitatives $X^1, \ldots, X^p$ dites *explicatives* (ou encore de contrôle, endogènes, indépendantes, régresseurs).

Les données sont supposées provenir de l'observation d'un échantillon statistique de taille n ($n > p+1$) de $\mathbb{R}^{(p+1)}$:

$$(x_i^1, \ldots, x_i^j, \ldots, x_i^p, y_i) \quad i = 1, \ldots, n.$$

L'écriture du *modèle linéaire* dans cette situation conduit à supposer que l'espérance de Y appartient au sous-espace de $\mathbb{R}^n$ engendré par $\{\mathbf{1}, X^1, \ldots, X^p\}$ où $\mathbf{1}$ désigne le vecteur de $\mathbb{R}^n$ constitué de "1" . C'est-à-dire que les $(p+1)$ variables aléatoires vérifient :

$$y_i = \beta_0 + \beta_1 x_i^1 + \beta_2 x_i^2 + \cdots + \beta_p x_i^p + u_i \quad i = 1, 2, \ldots, n$$

avec les hypothèses suivantes :

1. Les u_i sont des termes d'erreur, d'une variable U, non observés, indépendants et identiquement distribués ; $E(u_i) = 0, Var(U) = \sigma_u^2 \mathbf{I}$.
2. Les termes x^j sont supposés déterministes (facteurs contrôlés) **ou bien** l'erreur U est indépendante de la distribution conjointe de $X^1, \ldots, X^p$. On écrit dans ce dernier cas que :
$$\mathbb{E}(Y|X^1, \ldots, X^p) = \beta_0 + \beta_1 X^1 + \beta_2 X^2 + \cdots + \beta_p X^p \text{ et } \mathrm{Var}(Y|X^1, \ldots, X^p) = \sigma_u^2.$$
3. Les paramètres inconnus $\beta_0, \ldots, \beta_p$ sont supposés constants.
4. En option, pour l'étude spécifique des lois des estimateurs, une quatrième hypothèse considère la normalité de la variable d'erreur U ($\mathcal{N}(0, \sigma_u^2 \mathbf{I})$). Les u_i sont alors i.i.d. de loi $\mathcal{N}(0, \sigma_u^2)$.

Les données sont rangées dans une matrice $\mathbf{X}(n \times (p+1))$ de terme général x_i^j, dont la première colonne contient le vecteur $\mathbf{1}$ ($x_0^i = 1$), et dans un vecteur $\mathbf{Y}$ de terme général y_i. En notant les vecteurs $\mathbf{u} = [u_1 \cdots u_p]'$ et $\boldsymbol{\beta} = [\beta_0 \beta_1 \cdots \beta_p]'$, le modèle s'écrit matriciellement :

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}.$$

3 Estimation

Conditionnellement à la connaissance des valeurs des X^j, les paramètres inconnus du modèle : le vecteur β et σ_u^2 (paramètre de nuisance), sont estimés par minimisation du critère des moindres carrés (M.C.) ou encore, en supposant (iv), par maximisation de la vraisemblance (M.V.). Les estimateurs ont alors les mêmes expressions, l'hypothèse de normalité et l'utilisation de la vraisemblance conférant à ces derniers des propriétés complémentaires.

3.1 Estimation par M.C.

L'expression à minimiser sur $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$ s'écrit :

$$\begin{aligned} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i^1 - \beta_2 x_i^2 - \cdots - \beta_p x_i^p)^2 &= \|\mathbf{y} - \mathbf{X}\beta\|^2 \\ &= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \mathbf{y}'\mathbf{y} - 2\boldsymbol{\beta}'\mathbf{X}'\mathbf{y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}. \end{aligned}$$

Par dérivation matricielle de la dernière équation on obtient les *"équations normales"* :

$$\mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{X}\boldsymbol{\beta} = 0$$

dont la solution correspond bien à un minimum car la matrice hessienne $2\mathbf{X}'\mathbf{X}$ est semi définie-positive.

Nous faisons l'hypothèse supplémentaire que la matrice $\mathbf{X}'\mathbf{X}$ est inversible, c'est-à-dire que la matrice $\mathbf{X}$ est de rang $(p+1)$ et donc qu'il n'existe pas de colinéarité entre ses colonnes. En pratique, si cette hypothèse n'est pas vérifiée, il suffit de supprimer des colonnes de $\mathbf{X}$ et donc des variables du modèle. Des diagnostics de colinéarité et des aides au choix des variables seront explicités plus loin.

Alors, l'estimation des paramètres β_j est donnée par :

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

et les valeurs ajustées (ou estimées, prédites) de $\mathbf{y}$ ont pour expression :

$$\widehat{\mathbf{y}} = \mathbf{X}\mathbf{b} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}\mathbf{y}$$

où $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ est appelée "*hat matrix*" ; elle met un chapeau à $\mathbf{y}$. Géométriquement, c'est la matrice de projection orthogonale dans $\mathbb{R}^n$ sur le sous-espace Vect($\mathbf{X}$) engendré par les vecteurs colonnes de $\mathbf{X}$.

On note

$$\mathbf{e} = \mathbf{y} - \widehat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\mathbf{b} = (\mathbf{I} - \mathbf{H})\mathbf{y}$$

le vecteur des résidus ; c'est la projection de $\mathbf{y}$ sur le sous-espace orthogonal de Vect($\mathbf{X}$) dans $\mathbb{R}^n$.

3.2 Propriétés

Les estimateurs des M.C. $b_0, b_1, \ldots, b_p$ sont des estimateurs sans biais : $E(\mathbf{b}) = \boldsymbol{\beta}$, et, parmi les estimateurs sans biais fonctions linéaires des y_i, ils sont de variance minimum (propriété de Gauss-Markov) ; ils sont donc "BLUE" : *best linear unbiaised estimators*. Sous hypothèse de normalité, les estimateurs du M.V., qui coïncident avec ceux des moindres carrés, sont uniformément meilleurs ; ils sont efficaces c'est-à-dire que leur matrice de covariance atteint la borne inférieure de Cramer-Rao.

On montre que la matrice de covariance des estimateurs se met sous la forme

$$E[(\mathbf{b} - \boldsymbol{\beta})(\mathbf{b} - \boldsymbol{\beta})'] = \sigma_u^2(\mathbf{X}'\mathbf{X})^{-1},$$

celle des prédicteurs est

$$E[(\widehat{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta})(\widehat{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta})'] = \sigma_u^2\mathbf{H}$$

et celle des estimateurs des résidus est

$$E[(\mathbf{e} - \mathbf{u})((\mathbf{e} - \mathbf{u}))'] = \sigma_u^2(\mathbf{I} - \mathbf{H})$$

tandis qu'un estimateur sans biais de σ_u^2 est fourni par :

$$s^2 = \frac{\|e\|^2}{n-p-1} = \frac{\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2}{n-p-1} = \frac{\text{SSE}}{n-p-1}.$$

Ainsi, les termes $s^2 h_i^i$ sont des estimations des variances des prédicteurs $\widehat{y}_i$.

3.3 Sommes des carrés

SSE est la somme des carrés des résidus (*sum of squared errors*),

$$\text{SSE} = \|\mathbf{y} - \widehat{\mathbf{y}}\|^2 = \|\mathbf{e}\|^2.$$

On définit également la somme totale des carrés (*total sum of squares*) par

$$\text{SST} = \|\mathbf{y} - \overline{y}\mathbf{1}\|^2 = \mathbf{y}'\mathbf{y} - n\overline{y}^2$$

et la somme des carrés de la régression (*regression sum of squares*) par

$$\text{SSR} = \|\widehat{\mathbf{y}} - \overline{y}\mathbf{1}\|^2 = \widehat{\mathbf{y}}'\widehat{\mathbf{y}} - n\overline{y}^2 = \mathbf{y}'\mathbf{H}\mathbf{y} - n\overline{y}^2 = \mathbf{b}'\mathbf{X}'\mathbf{y} - n\overline{y}^2.$$

On vérifie alors : SST = SSR + SSE.

3.4 Coefficient de détermination

On appelle *coefficient de détermination* le rapport

$$R^2 = \frac{\text{SSR}}{\text{SST}}$$

qui est donc la part de variation de Y expliquée par le modèle de régression. Géométriquement, c'est un rapport de carrés de longueur de deux vecteurs.

C'est donc le cosinus carré de l'angle entre ces vecteurs : $\mathbf{y}$ et sa projection $\widehat{\mathbf{y}}$ sur Vect($\mathbf{X}$).

Attention, dans le cas extrême où $n = (p + 1)$, c'est-à-dire si le nombre de variables explicatives est grand comparativement au nombre d'observations, $R^2 = 1$. Ou encore, il est géométriquement facile de voir que l'ajout de variables explicatives ne peut que faire croître le coefficient de détermination.

La quantité R est appelée *coefficient de corrélation multiple* entre Y et les variables explicatives, c'est le coefficient de corrélation usuel entre $\mathbf{y}$ et sa prédiction (ou projection) $\widehat{\mathbf{y}}$.

4 Inférences dans le cas gaussien

En principe, l'hypothèse optionnelle (iv) de normalité des erreurs est nécessaire pour cette section. En pratique, des résultats asymptotiques, donc valides pour de grands échantillons, ainsi que des études de simulation, montrent que cette hypothèse n'est pas celle dont la violation est la plus pénalisante pour la fiabilité des modèles.

4.1 Inférence sur les coefficients

Pour chaque coefficient β_j on montre que la statistique

$$\frac{b_j - \beta_j}{\sigma_{b_j}}$$

où $\sigma^2_{b_j}$, variance de b_j est le $j$ième terme diagonal de la matrice $s^2(\mathbf{X}'\mathbf{X})^{-1}$, suit une loi de Student à $(n - p - 1)$ degrés de liberté. Cette statistique est donc utilisée pour tester une hypothèse $H_0 : \beta_j = a$ ou pour construire un intervalle de confiance de niveau $100(1 - \alpha)\%$:

$$b_j \pm t_{\alpha/2;(n-p-1)}\sigma_{b_j}.$$

Attention, cette statistique concerne un coefficient et ne permet pas d'inférer conjointement (cf. §3.4) sur d'autres coefficients car ils sont corrélés entre eux ; de plus elle dépend des absences ou présences des autres variables X^k dans le modèle. Par exemple, dans le cas particulier de deux variables X^1 et X^2 très corrélées, chaque variable, en l'absence de l'autre, peut apparaître avec un coefficient significativement différent de 0 ; mais, si les deux sont présentes dans le modèle, elles peuvent chacune apparaître avec des coefficients insignifiants.

De façon plus générale, si $\mathbf{c}$ désigne un vecteur non nul de $(p+1)$ constantes réelles, il est possible de tester la valeur d'une combinaison linéaire $\mathbf{c}'\mathbf{b}$ des paramètres en considérant l'hypothèse nulle $H_0 : \mathbf{c}'\mathbf{b} = a$; a connu. Sous H_0, la statistique

$$\frac{\mathbf{c}'\mathbf{b} - a}{(s^2\mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c})^{1/2}}$$

suit une loi de Student à $(n - p - 1)$ degrés de liberté.

4.2 Inférence sur le modèle

Le modèle peut être testé globalement. Sous l'hypothèse nulle $H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$, la statistique

$$\frac{\text{SSR}/p}{\text{SSE}/(n - p - 1)} = \frac{\text{MSR}}{\text{MSE}}$$

suit une loi de Fisher avec p et $(n - p - 1)$ degrés de liberté. Les résultats sont habituellement présentés dans un tableau *"d'analyse de la variance"* sous la forme suivante :

Source de variation	*d.d.l.*	*Somme des carrés*	*Variance*	*F*
Régression	p	SSR	MSR=SSR/p	MSR/MSE
Erreur	$n - p - 1$	SSE	MSE=SSE/$(n - p - 1)$	
Total	$n - 1$	SST		

4.3 Inférence sur un modèle réduit

Le test précédent amène à rejeter H_0 dès que l'une des variables X^j est liée à Y. Il est donc d'un intérêt limité. Il est souvent plus utile de tester un modèle réduit c'est-à-dire dans lequel certains coefficients sont nuls (à l'exception du terme constant) contre le modèle complet avec toutes les variables. En ayant éventuellement réordonné les variables, on considère l'hypothèse nulle $H_0 : \beta_1 = \beta_2 = \ldots = \beta_q = 0, q < p$.

Notons respectivement SSR_q, SSE_q, R_q^2 les sommes de carrés et le coefficient de détermination du modèle réduit à $(p-q)$ variables. Sous H_0, la statistique

$$\frac{(\text{SSR}-\text{SSR}_q)/q}{\text{SSE}/(n-p-1)} = \frac{(R^2-R_q^2)/q}{(1-R^2)/(n-p-1)}$$

suit une loi de Fisher à q et $(n-p-1)$ degrés de liberté.

Dans le cas particulier où $q=1$ ($\beta_j=0$), la F-statistique est alors le carré de la t-statistique de l'inférence sur un paramètre et conduit donc au même test.

4.4 Ellipsoïde de confiance

Les estimateurs des coefficients β_j étant corrélés, la recherche d'une région de confiance de niveau $100(1-\alpha)\%$ pour tous les coefficients conduit à considérer l'ellipsoïde décrit par

$$(\mathbf{b}-\boldsymbol{\beta})'\mathbf{X}'\mathbf{X}(\mathbf{b}-\boldsymbol{\beta}) \sim (p+1)s^2 F_{\alpha;p+1,(n-p-1)}.$$

Plus généralement, un ellipsoïde de confiance conjoint à q combinaisons linéaires $\mathbf{T}\boldsymbol{\beta}$ est donné par

$$(\mathbf{Tb}-\mathbf{T}\boldsymbol{\beta})'[\mathbf{T}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{T}']^{-1}(\mathbf{Tb}-\mathbf{T}\boldsymbol{\beta}) \leq qs^2 F_{\alpha;q,(n-p-1)}$$

où $\mathbf{T}(q\times(p+1))$ est une matrice de rang q de constantes fixées.

En application, étant donnés une matrice $\mathbf{T}$ et un vecteur $\mathbf{a}$, un test de l'hypothèse H_0 : $\mathbf{T}\beta=\mathbf{a}$ est obtenu en considérant la statistique

$$(\mathbf{Tb}-\mathbf{a})'[\mathbf{T}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{T}']^{-1}(\mathbf{Tb}-\mathbf{a})/qs^2$$

qui suit sous H_0 une loi de Fisher à q et $(n-p-1)$ degrés de liberté.

4.5 Prévision

Connaissant les valeurs des variables X^j pour une nouvelle observation : $\mathbf{x}_0'=[x_0^1,x_0^2,\ldots,x_0^p]$ appartenant au domaine dans lequel l'hypothèse de linéarité reste valide, une prévision, notée $\widehat{y}_0$ de Y ou $E(Y)$ est donnée par :

$$\widehat{y}_0 = b_0 + b_1 x_0^1 + \cdots + b_p x_0^p.$$

Les intervalles de confiance des prévisions de Y et $E(Y)$, pour *une* valeur $\mathbf{x}_0\in\mathbb{R}^p$ et en posant $\mathbf{v}_0=(1|bmx_0')'\in\mathbb{R}^{p+1}$, sont respectivement

$$\begin{aligned}\widehat{y}_0 &\pm t_{\alpha/2;(n-p-1)}s(1+\mathbf{v}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{v}_0)^{1/2},\\ \widehat{y}_0 &\pm t_{\alpha/2;(n-p-1)}s(\mathbf{v}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{v}_0)^{1/2}.\end{aligned}$$

Enfin, un intervalle de confiance de niveau $100(1-\alpha)\%$ recouvrant globalement la surface de régression est donné par

$$\widehat{y}_0 \pm [(p+1)F_{\alpha;(p+1),(n-p-1)}]^{1/2}s(\mathbf{v}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{v}_0)^{1/2}.$$

Il peut être utilisé pour définir un intervalle conjoint à plusieurs prédictions.

5 Sélection de variables, choix de modèle

De façon un peu schématique, on peut associer la pratique de la modélisation statistique à trois objectifs qui peuvent éventuellement être poursuivis en complémentarité.

Descriptif : Il vise à rechercher de façon exploratoire les liaisons entre Y et d'autres variables, potentiellement explicatives, X^j qui peuvent être nombreuses afin, par exemple d'en sélectionner un sous-ensemble. À cette stratégie, à laquelle peuvent contribuer des Analyses en Composantes Principales, correspond des algorithmes de recherche (pas à pas) moins performants mais économiques en temps de calcul si p est grand.

Attention, si n est petit, et la recherche suffisamment longue avec beaucoup de variables explicatives, il sera toujours possible de trouver un "bon" modèle expliquant y ; c'est l'effet *data mining* dans les modèles économétriques.

Explicatif : Le deuxième objectif est sous-tendu par une connaissance *a priori* du domaine concerné et dont des résultats théoriques peuvent vouloir être confirmés, infirmés ou précisés par l'estimation des paramètres. Dans ce cas, les résultats inférentiels précédents permettent de construire *le* bon test conduisant à la prise de décision recherchée. Utilisées hors de ce contexte, les statistiques de test n'ont plus alors qu'une valeur indicative au même titre que d'autres critères plus empiriques.

Prédictif : Dans le troisième cas, l'accent est mis sur la qualité des estimateurs et des prédicteurs qui doivent, par exemple, minimiser une erreur

quadratique moyenne. Ceci conduit à rechercher des modèles *parcimonieux* c'est-à-dire avec un nombre volontairement restreint de variables explicatives. Le "meilleur" modèle ainsi obtenu peut donner des estimateurs légèrement biaisés au profit d'un compromis pour une variance plus faible. Un bon modèle n'est donc plus celui qui explique le mieux les données au sens d'une déviance (SSE) minimale (ou d'un R^2 max) au prix d'un nombre important de variables pouvant introduire des colinéarités. Le bon modèle est celui qui conduit aux prédictions les plus fiables.

5.1 Critères

De nombreux critères de choix de modèle sont présentés dans la littérature sur la régression linéaire multiple. Citons le critère d'information d'Akaïke (AIC), celui bayésien de Sawa (BIC), l'erreur quadratique moyenne de prédiction (cas gaussien)... Ils sont équivalents lorsque le nombre de variables à sélectionner, ou niveau du modèle, est fixé. Le choix du critère est déterminant lorsqu'il s'agit de comparer des modèles de niveaux différents. Certains critères se ramènent, dans le cas gaussien, à l'utilisation d'une expression pénalisée de la fonction de vraisemblance afin de favoriser des modèles parcimonieux. En pratique, les plus utilisés ou ceux généralement fournis par les logiciels sont les suivants.

5.1.1 Statistique du F de Fisher

Ce critère, justifié dans le cas explicatif est aussi utilisé à titre indicatif pour comparer des séquences de modèles emboîtés. La statistique partielle de Fisher est

$$\frac{(\text{SSR} - \text{SSR}_q)/q}{\text{SSE}/(n-p-1)} = \frac{(R^2 - R_q^2)}{(1-R^2)} \frac{n-p-1}{q}$$

dans laquelle l'indice q désigne les expressions concernant le modèle réduit avec $(p-q)$ variables explicatives. On considère alors que si l'accroissement $(R^2 - R_q^2)$ est suffisamment grand :

$$R^2 - R_q^2 > \frac{q(1-R^2)}{(n-p-1)} F_{\alpha;q,(n-p-1)},$$

l'ajout des q variables au modèle est justifié.

5.1.2 R^2 et R^2 ajusté

Le coefficient de détermination $R^2 = 1 - \text{SSE/SST}$, directement lié à la déviance (SSE) est aussi un indice de qualité mais qui a la propriété d'être monotone croissant en fonction du nombre de variables. Il ne peut donc servir qu'à comparer deux modèles de même niveau c'est-à-dire avec le même nombre de variables.

En revanche, le R^2 *ajusté* :

$$R'^2 = 1 - \frac{n-1}{n-p-1}(1-R^2) = 1 - \frac{\text{SSE}/(n-p-1)}{\text{SST}/(n-1)}.$$

dans lequel le rapport SSE/SST est remplacé par un rapport des estimations sans biais des quantités σ_u^2 et σ_y^2 introduit une pénalisation liée au nombre de paramètres à estimer.

Ce coefficient s'exprime encore par

$$1 - \frac{(n-1)\text{MSE}}{\text{SST}}$$

ainsi dans la comparaison de deux modèles partageant la même SST, on observe que $R'^2 > R'^2_j$ si et seulement si MSE<MSE$_j$; MSE et MSE$_j$ désignant respectivement l'erreur quadratique moyenne du modèle complet et celle d'un modèle à j variables explicatives. Maximiser le R^2 ajusté revient donc à minimiser l'erreur quadratique moyenne.

5.1.3 C_p de Mallows

Une erreur quadratique moyenne s'écrit comme la somme d'une variance et du carré d'un biais. L'erreur quadratique moyenne de prédiction sécrit ainsi :

$$\text{MSE}(\widehat{y}_i) = \text{Var}(\widehat{y}_i) + [\text{Biais}(\widehat{y}_i)]^2$$

puis après sommation et réduction :

$$\frac{1}{\sigma_u^2}\sum_{i=1}^{n}\text{MSE}(\widehat{y}_i) = \frac{1}{\sigma_u^2}\sum_{i=1}^{n}\text{Var}(\widehat{y}_i) + \frac{1}{\sigma_u^2}\sum_{i=1}^{n}[\text{Biais}(\widehat{y}_i)]^2.$$

En supposant que les estimations du modèle complet sont sans biais et en utilisant des estimateurs de $Var(\widehat{y}_i)$ et σ_u^2, l'expression de l'erreur quadratique

moyenne totale standardisée (ou réduite) pour un modèle à q variables explicatives s'écrit :

$$C_p = (n-q-1)\frac{\text{MSE}_q}{\text{MSE}} - [n-2(q+1)]$$

et définit la valeur du C_p de Mallow pour les q variables considérées. Il est alors d'usage de rechercher un modèle qui minimise le C_p tout en fournissant une valeur inférieure et proche de $(q+1)$. Ceci revient à considérer que le "vrai" modèle complet est moins fiable qu'un modèle réduit donc biaisé mais d'estimation plus précise.

5.1.4 PRESS de Allen

On désigne par $\widehat{y}_{(i)}$ la prédiction de y_i calculée sans tenir compte de la $i$ème observation $(y_i, x_i^1, \ldots, x_i^p)$, la somme des erreurs quadratiques de prédiction (PRESS) est définie par

$$\text{PRESS} = \sum_{i=1}^{n} (y_i - \widehat{y}_{(i)})^2$$

et permet de comparer les capacités prédictives de deux modèles.

5.2 Algorithmes de sélection

Lorsque p est grand, il n'est pas raisonnable de penser explorer les 2^p modèles possibles afin de sélectionner le "meilleur" au sens de l'un des critères ci-dessus. Différentes stratégies sont donc proposées qui doivent être choisies en fonction de l'objectif recherché et des moyens de calcul disponibles ! Trois types d'algorithmes sont résumés ci-dessous par ordre croissant de temps de calcul nécessaire c'est-à-dire par nombre croissant de modèles considérés parmi les 2^p et donc par capacité croissante d'optimalité. On donne pour chaque algorithme l'option `selection` à utiliser dans la procédure `REG` de SAS.

5.2.1 Pas à pas

Sélection (`forward`) À chaque pas, une variable est ajoutée au modèle. C'est celle dont la valeur p ("prob value")associée à la statistique partielle du test de Fisher qui compare les deux modèles est minimum. La procédure s'arrête lorsque toutes les variables sont introduites ou lorsque p reste plus grande qu'une valeur seuil fixée par défaut à $0,50$.

Élimination (`backward`) L'algorithme démarre cette fois du modèle complet. À chaque étape, la variable associée à la plus grande valeur p est éliminée du modèle. La procédure s'arrête lorsque les variables restant dans le modèle ont des valeurs p plus petites qu'un seuil fixé par défaut à $0,10$.

Mixte (`stepwise`) Cet algorithme introduit une étape d'élimination de variable après chaque étape de sélection afin de retirer du modèle d'éventuels variables qui seraient devenues moins indispensables du fait de la présence de celles nouvellement introduites.

5.2.2 Par échange

Maximisation de R^2 (`maxr`) Cet algorithme tente de trouver le meilleur modèle pour chaque niveau c'est-à-dire pour chaque nombre de variables explicatives. À chaque niveau il commence par sélectionner une variable complémentaire qui rend l'accroissement de R^2 maximum. Puis il regarde tous les échanges possibles entre une variable présente dans le modèle et une extérieure et exécute celui qui fournit l'accroissement maximum ; ceci est itéré tant que le R^2 croit.

Minimisation de R^2 (`minr`) Il s'agit du même algorithme que le précédent sauf que la procédure d'échange fait appel au couple de variables associé au plus petit accroissement du R^2. L'objectif est ainsi d'explorer plus de modèles que dans le cas précédent et donc, éventuellement, de tomber sur un meilleur optimum.

Remarque Pour tous ces algorithmes de sélection ou d'échange, il est important de compléter les comparaisons des différentes solutions retenues à l'aide de critères globaux (C_p ou PRESS).

5.2.3 Global

L'algorithme de Furnival et Wilson est utilisé pour comparer tous les modèles possibles en cherchant à optimiser l'un des critères : R^2, R^2 ajusté, ou C_p de Mallow (`rsquare, adjrsq, cp`) choisi par l'utilisateur. Par souci d'économie, cet algorithme évite de considérer des modèles de certaines

sous-branches de l'arborescence dont on peut savoir a priori qu'ils ne sont pas compétitifs. En général les logiciels exécutant cet algorithme affichent le (`best=1`) ou les meilleurs modèles de chaque niveau.

6 Multi-colinéarité

L'estimation des paramètres ainsi que celle de leur écart-type (standard error) nécessite le calcul explicite de la matrice $(\mathbf{X}'\mathbf{X})^{-1}$. Dans le cas dit *mal conditionné* où le déterminant de la matrice $\mathbf{X}'\mathbf{X}$ n'est que légèrement différent de 0, les résultats conduiront à des estimateurs de variances importantes et même, éventuellement, à des problèmes de précision numérique. Il s'agit donc de diagnostiquer ces situations critiques puis d'y remédier. Dans les cas descriptif ou prédictif on supprime des variables à l'aide des procédures de choix de modèle mais, pour un objectif explicatif nécessitant toutes les variables, d'autres solutions doivent être envisagées : algorithme de résolution des équations normales par transformations orthogonales (procédure `orthoreg` de SAS) sans calcul explicite de l'inverse pour limiter les problèmes numériques, régression biaisée (ridge), régression sur composantes principales.

6.1 Diagnostics

Notons $\widetilde{\mathbf{X}}$ la matrice des données observées, c'est-à-dire $\mathbf{X}$ privée de la première colonne $\mathbf{1}$ et dont on a retranché à chaque ligne le vecteur moyen $\overline{\mathbf{x}} = 1/n \sum_{i=1}^{n} \mathbf{x}_i$, $\mathbf{S}$ la matrice diagonale contenant les écarts-types empiriques des variables X^j et enfin $\mathbf{R}$ la matrice des corrélations :

$$\mathbf{R} = \frac{1}{(n-1)} \mathbf{S}^{-1} \widetilde{\mathbf{X}}' \widetilde{\mathbf{X}} \mathbf{S}^{-1}.$$

6.1.1 Facteur d'inflation de la variance (VIF)

Avec ces notations, la matrice de covariance des estimateurs des coefficients $(\beta_1, \ldots, \beta_p)$ s'écrit :

$$\frac{\sigma_u^2}{n-1} (\widetilde{\mathbf{X}}' \widetilde{\mathbf{X}})^{-1} = \frac{\sigma_u^2}{n-1} \mathbf{S} \mathbf{R}^{-1} \mathbf{S}.$$

On montre alors que chaque élément diagonal s'exprime comme

$$V_j = \frac{1}{1 - R_j^2}$$

où R_j^2 désigne le coefficient de détermination de la régression de la variable X^j sur les autres variables ; R_j est alors un coefficient de corrélation multiple, c'est le cosinus de l'angle dans $\mathbb{R}^n$ entre X^j et le sous-espace vectoriel engendré par les variables $\{X^1, \ldots, X^{j-1}, X^{j+1}, \ldots, X^p\}$. Plus X^j est "linéairement" proche de ces variables et plus R_j est proche de 1 et donc plus la variance de l'estimateur de β_j est élevée ; V_j est appelé *facteur d'inflation de la variance* (VIF). Évidemment, cette variance est minimum lorsque X^j est orthogonal au sous-espace engendré par les autres variables.

Le simple examen de la matrice $\mathbf{R}$ permet de relever des corrélations dangereuses de variables deux à deux mais est insuffisant pour détecter des corrélations plus complexes ou multi-colinéarités. C'est donc l'inverse de cette matrice qu'il faut considérer en calculant les V_j ou encore les valeurs $(1 - R_j^2)$ qui sont appelées *tolérances*.

6.1.2 Conditionnement

On note $\lambda_1, \ldots, \lambda_p$ les valeurs propres de la matrice $\mathbf{R}$ rangées par ordre décroissant. Le déterminant de $\mathbf{R}$ est égal au produit des valeurs propres. Ainsi, des problèmes numériques, ou de variances excessives apparaissent dès que les dernières valeurs propres sont relativement trop petites.

On appelle *indice de conditionnement* le rapport

$$\kappa = \lambda_1 / \lambda_p$$

de la plus grande sur la plus petite valeur propre.

En pratique, si $\kappa < 100$ on considère qu'il n'y a pas de problème. Celui-ci devient sévère pour $\kappa > 1000$. Cet indice de conditionnement donne un aperçu global des problèmes de colinéarité tandis que les VIF, les tolérances ou encore l'étude des vecteurs propres associés au plus petites valeurs propres permettent d'identifier les variables les plus problématiques.

Remarque : Lorsque le modèle est calculé avec un terme constant, la colonne $\mathbf{1}$ joue le rôle d'une variable et peut considérablement augmenter les

problèmes de multi-colinéarité. La matrice $\mathbf{R}$ est alors remplacée par la matrice $\mathbf{T} = \text{diag}(\mathbf{X}'\mathbf{X})^{-1/2}\mathbf{X}'\mathbf{X}\text{diag}(\mathbf{X}'\mathbf{X})^{-1/2}$ dans les discussions précédentes.

6.2 Régression "ridge"

Ayant diagnostiqué un problème mal conditionné mais désirant conserver toutes les variables, il est possible d'améliorer les propriétés numériques et la variance des estimations en considérant un estimateur légèrement biaisé des paramètres. L'estimateur "ridge" introduisant une *régularisation* est donné par

$$\mathbf{b}_R = (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{y},$$

qui a pour effet de décaler de la valeur k toutes les valeurs propres de la matrice à inverser et, plus particulièrement, les plus petites qui reflètent la colinéarité. On montre que l'erreur quadratique moyenne sur l'estimation des paramètres se met sous la forme :

$$\text{MSE}(\mathbf{b}_R) = \sigma_u^2 \sum_{j=1}^{p} \frac{\lambda_j}{(\lambda_j + k)^2} + k^2\boldsymbol{\beta}'(\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\boldsymbol{\beta}.$$

La difficulté est alors de trouver une valeur de k minimisant la quantité ci-dessus. Des méthodes de ré-échantillonnage (jackknife, bootstrap) peuvent être mises en œuvre mais celles-ci sont coûteuses en temps de calcul. Une valeur heuristique de k peut être fixée en considérant le graphique des paramètres en fonction de k. Elle est choisie dans la zone où les valeurs absolues des paramètres commencent à se stabiliser.

6.3 Régression sur composantes principales

L'Analyse en Composantes Principales est, entre autre, la recherche de p variables dites principales qui sont des combinaisons linéaires des variables initiales de variance maximale sous une contrainte d'orthogonalité. En désignant par $\mathbf{V}$ la matrice des vecteurs propres de la matrice des corrélations $\mathbf{R}$ rangés dans l'ordre décroissant des valeurs propres, les valeurs prises par ces variables principales sont obtenues dans la matrice des composantes principales

$$\mathbf{C} = (\widetilde{\mathbf{X}} - \mathbf{1}\overline{\mathbf{x}}')\mathbf{V}.$$

Elles ont chacune pour variance la valeur propre λ_j associée. Le sous-espace engendré par ces variables principales est le même que celui engendré par les variables initiales. Il est donc géométriquement équivalent de régresser Y sur les colonnes de $\mathbf{C}$ que sur celles de $\widetilde{\mathbf{X}}$. Les problèmes de colinéarité sont alors résolus en supprimant les variables principales de plus faibles variances c'est-à-dire associées aux plus petites valeurs propres.

La solution obtenue présente ainsi de meilleures qualités prédictives mais, les coefficients de la régression s'appliquant aux composantes principales, un calcul complémentaire est nécessaire afin d'évaluer et d'interpréter les effets de chacune des variables initiales.

6.4 Modèles curvilinéaires

En cas d'invalidation de l'hypothèse de linéarité, il peut être intéressant de considérer des modèles polynômiaux, très classiques pour décrire des phénomènes physiques, de la forme

$$Y = \beta_0 + \cdots + \beta_j X^j + \cdots + \gamma_{kl} X^k X^l + \cdots + \delta_j X^{j2}$$

qui sont encore appelés *surfaces de réponse*. Ces modèles sont faciles à étudier dans le cadre linéaire, il suffit d'ajouter des nouvelles variables constituées des produits ou des carrés des variables explicatives initiales. Les choix : présence ou non d'une interaction entre deux variables, présence ou non d'un terme quadratique se traitent alors avec les mêmes outils que ceux des choix de variable mais en intégrant une contrainte lors de la lecture des résultats : ne pas considérer des modèles incluant des termes quadratiques dont les composants linéaires auraient été exclus ou encore, ne pas supprimer d'un modèle une variable d'un effet linéaire si elle intervient dans un terme quadratique.

La procédure `rsreg` de SAS est plus particulièrement adaptée aux modèles quadratiques. Elle ne comporte pas de procédure de choix de modèle mais fournit des aides et diagnostics sur l'ajustement de la surface ainsi que sur la recherche des points optimaux.

Attention : Ce type de modèle accroît considérablement les risques de colinéarité, il est peu recommandé de considérer des termes cubiques.

7 Influence, résidus, validation

Avant toute tentative de modélisation complexe, il est impératif d'avoir conduit des analyses uni et bivariées afin d'identifier des problèmes sur les

distributions de chacune des variables : dissymétrie, valeurs atypiques (outliers) ou sur les liaisons des variables prises deux par deux : non-linéarité. Ces préliminaires acquis, des aides ou diagnostics associés à la régression linéaire multiple permettent de détecter des violations d'hypothèses (homoscédasticité, linéarité) ou des points influents dans ce contexte multidimensionnel.

7.1 Effet levier

Comme toute méthode quadratique, l'estimation des paramètres est très sensible à la présence de points extrêmes susceptibles de perturber gravement les résultats. À partir de l'équation de prédiction : $\widehat{\mathbf{y}} = \mathbf{H}\mathbf{y}$ on remarque qu'une observation i est influente si le terme correspondant h_i^i de la diagonale de $\mathbf{H}$ est grand.

On écrit encore :

$$\mathbf{H} = \frac{\mathbf{1}\mathbf{1}'}{n} + \widetilde{\mathbf{X}}(\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}'$$

et

$$h_i^i = \frac{1}{n} + (\mathbf{x}_i - \overline{\mathbf{x}})'(\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}})^{-1}(\mathbf{x}_i - \overline{\mathbf{x}}) = \frac{1}{n} + \sum_{j=1}^{p}\left(\frac{\mathbf{v}'^j(\mathbf{x}_i - \overline{\mathbf{x}})}{\sqrt{\lambda_j}}\right)^2$$

où les $\lambda_j, \mathbf{v}^j$ sont respectivement les valeurs et vecteurs propres de la matrice $\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}$. Ainsi, plus une observation est éloignée du barycentre, et ce dans la direction d'un vecteur propre associé à une petite valeur propre, et plus cette observation a un effet levier important.

7.2 Résidus

Nous désignons comme précédemment par $\mathbf{b}_{(i)}, \widehat{\mathbf{y}}_{(i)}, \mathbf{e}_{(i)}$, et

$$s^2_{(i)} = \frac{\mathbf{e}'_{(i)}\mathbf{e}_{(i)}}{n-p-2}$$

les estimations réalisées sans la $i$ème observation. Les expressions

$$\begin{aligned}
\mathbf{e} &= (\mathbf{I} - \mathbf{H})\mathbf{y}, \\
\mathbf{r} &= \text{diag}[s^2(1-h_i^i)]^{-1/2}\mathbf{e}, \\
\mathbf{t} &= \text{diag}[s^2_{(i)}(1-h_i^i)]^{-1/2}\mathbf{e}
\end{aligned}$$

définissent respectivement les résidus calculés, les résidus *standardisés* (chacun divisé par l'estimation de l'écart-type) et les résidus *studentisés* dans lesquels l'estimation de σ_u^2 ne fait pas intervenir la $i$ème observation.

De trop grands résidus sont aussi des signaux d'alerte. Par exemple, un résidu studentisé de valeur absolue plus grande que 2 peut révéler un problème.

7.3 Mesures d'influence

L'effet levier peut apparaître pour des observations dont les valeurs prises par les variables explicatives sont élevées (observation loin du barycentre $\overline{x}$). De grands résidus signalent plutôt des valeurs atypiques de la variable à expliquer. Les deux diagnostics précédents sont combinés dans des mesures synthétiques proposées par différents auteurs. Les plus utilisées sont

$$D_i = \frac{1}{s^2(p+1)}(\widehat{\mathbf{y}} - \widehat{\mathbf{y}}_{(i)})'(\widehat{\mathbf{y}} - \widehat{\mathbf{y}}_{(i)}) = \left[\frac{h_i^i}{1-h_i^i}\right]\frac{r_i^2}{(p+1)}, \quad (1)$$

$$\text{DFFITS}_i = \frac{1}{s_{(i)}\sqrt{h_i^i}}(\widehat{y}_i - \widehat{y}_{(i)i}) = \left[\frac{h_i^i}{1-h_i^i}\right]^{1/2} t_i. \quad (2)$$

La première, notée *Cook's D* conclut à une influence de l'observation i lorsque la valeur de D_i dépasse 1.

D'autres mesures moins fréquemment utilisées sont proposées dans les logiciels. Certaines considèrent les écarts entre l'estimation d'un paramètre b_i et son estimation sans la $i$ème observation, une autre le rapport des déterminants des matrices de covariance des estimateurs des paramètres calculées avec et sans la $i$ème observation...

7.4 Régression partielle

Un modèle de régression multiple est une technique *linéaire*. Il est raisonnable de s'interroger sur la pertinence du caractère linéaire de la contribution d'une variable explicative à l'ajustement du modèle. Ceci peut être réalisé en considérant une *régression partielle*.

On calcule alors deux régressions :

– la régression de Y sur les variables $X^1, \ldots, X^{j-1}, X^{j+1}, \ldots, X^p$, dans laquelle la $j$ème variable est omise, soit $\mathbf{r}_{y(j)}$ le vecteur des résidus obtenus.

– La régression de X^j sur les variables $X^1, \ldots, X^{j-1}, X^{j+1}, \ldots, X^p$. Soit $\mathbf{r}_{x(j)}$ le vecteur des résidus obtenus.

La comparaison des résidus par un graphe (nuage de points $\mathbf{r}_{y(j)} \times \mathbf{r}_{x(j)}$) permet alors de représenter la nature de la liaison entre X^j et Y *conditionnellement* aux autres variables explicatives du modèle.

7.5 Graphes

Différents graphiques permettent finalement de contrôler le bien fondé des hypothèses de linéarité, d'homoscédasticité, éventuellement de normalité des résidus.

– Le premier considère le nuage de points des résidus studentisés croisés avec les valeurs prédites. Les points doivent être uniformément répartis entre les bornes -2 et $+2$ et ne pas présenter de formes suspectes.
– Le deuxième croise les valeurs observées de Y avec les valeurs prédites. Il illustre le coefficient de détermination R qui est aussi la corrélation linéaire simple entre $\widehat{\mathbf{y}}$ et $\mathbf{y}$. Les points doivent s'aligner autour de la première bissectrice. Il peut être complété par l'intervalle de confiance des y_i ou celui de leurs moyennes.
– La qualité, en terme de linéarité, de l'apport de chaque variable est étudiée par des régressions partielles. Chaque graphe de résidus peut être complété par une estimation fonctionnelle ou régression non-paramétrique (loess, noyau, spline) afin d'en facilité la lecture.
– Le dernier trace la droite de Henri (Normal QQplot) des résidus dont le caractère linéaire de la représentation donne une idée de la normalité de la distribution.

8 Exemple

8.1 Les données

Elles sont extraites de Jobson (1991) et décrivent les résultats comptables de 40 entreprises du Royaume Uni.

Descriptif des 13 variables (en anglais pour éviter des traductions erronées) :

RETCAP	Return on capital employed
WCFTDT	Ratio of working capital flow to total debt
LOGSALE	Log to base 10 of total sales
LOGASST	Log to base 10 of total assets
CURRAT	Current ratio
QUIKRAT	Quick ratio
NFATAST	Ratio of net fixed assets to total assets
FATTOT	Gross sixed assets to total assets
PAYOUT	Payout ratio
WCFTCL	Ratio of working capital flow to total current liabilities
GEARRAT	Gearing ratio (debt-equity ratio)
CAPINT	Capital intensity (ratio of total sales to total assets)
INVTAST	Ratio of total inventories to total assets

8.2 Résultat du modèle complet

La procédure SAS/REG est utilisée dans le programme suivant. La plupart des options sont actives afin de fournir la plupart des résultats même si certains sont redondants ou peu utiles.

```
options linesize=110 pagesize=30 nodate nonumber;
title;
proc reg data=sasuser.ukcompl all;
 model RETCAP = WCFTCL   WCFTDT   GEARRAT   LOGSALE   LOGASST
       NFATAST  CAPINT   FATTOT   INVTAST   PAYOUT   QUIKRAT   CURRAT
       /dw covb  Influence cli clm tol vif collin R P;
output out=resout h=lev p=pred r=res student=resstu ;
run;
```

Les résultats ne sont pas listés de façon exhaustive, les matrices et tableaux trop volumineux et peu significatifs ont été tronqués.

```
Descriptive Statistics
Variables       Sum           Mean      Uncorrected SS        Variance       Std Deviation
INTERCEP         40              1                40                0                   0
WCFTCL        10.29        0.25725            6.4339       0.0970973718       0.3116045118
WCFTDT         9.04          0.226            4.9052       0.0733887179       0.2709035215
...
CURRAT        72.41        1.81025           279.0039       3.7929153205       1.9475408392
RETCAP         5.71        0.14275            1.5233       0.0181589103       0.1347550009

Uncorrected Sums of squares and Crossproducts
USSCP       INTERCEP      WCFTCL      WCFTDT      GEARRAT     LOGSALE     LOGASST     NFATAST
INTERCEP          40       10.29        9.04         12.2       173.7       174.81       13.46
WCFTCL         10.29      6.4339      5.4926       1.5997     40.8722     46.2433      3.5523
WCFTDT          9.04      5.4926      4.9052       1.3972     34.4091     39.8937      2.9568
...
CURRAT         72.41      35.222      33.248      16.3188    265.2051     314.449     20.4126
RETCAP          5.71      2.0009      1.6226       1.5391     26.3636      25.379      1.6199

Correlation
CORR          WCFTCL      WCFTDT      GEARRAT     LOGSALE     LOGASST     NFATAST      CAPINT
WCFTCL        1.0000      0.9620     -0.5520     -0.3100      0.1829      0.0383     -0.2376
WCFTDT        0.9620      1.0000     -0.5611     -0.4533      0.0639     -0.0418     -0.2516
GEARRAT      -0.5520     -0.5611      1.0000      0.2502      0.0387     -0.0668      0.2532
...
```

```
CURRAT      0.7011     0.8205    -0.3309    -0.6406    -0.0460    -0.2698    -0.3530
RETCAP      0.3249     0.2333    -0.1679     0.2948     0.1411    -0.2974     0.3096
```

La matrice des corrélations montre des valeurs élevées, on peut déja s'attendre à des problèmes de colinéarité.

```
Model Crossproducts X'X X'Y Y'Y
X'X         INTERCEP     WCFTCL     WCFTDT    GEARRAT    LOGSALE    LOGASST    NFATAST
INTERCEP          40      10.29       9.04       12.2      173.7     174.81      13.46
WCFTCL         10.29     6.4339     5.4926     1.5997    40.8722    46.2433     3.5523
WCFTDT          9.04     5.4926     4.9052     1.3972    34.4091    39.8937     2.9568
...
X'X Inverse, Parameter Estimates, and SSE
          INTERCEP     WCFTCL     WCFTDT    GEARRAT    LOGSALE    LOGASST    NFATAST
INTERCEP  3.2385537  1.3028641 -1.570579  -0.05877  0.3001809 -0.826512 -0.238509
WCFTCL    1.3028641  7.0714100 -9.955073 -0.54391  -0.007877 -0.292412 -0.233915
WCFTDT    -1.570579  -9.955073  15.968504  1.582975  0.0112826  0.3138925   0.149976
...
Analysis of Variance
                          Sum of          Mean
Source           DF       Squares        Square         F Value       Prob>F
                (1)
Model            12      0.55868 (2)     0.04656 (5)     8.408 (7)     0.0001 (8)
Error            27      0.14951 (3)     0.00554 (6)
C Total          39      0.70820 (4)
  Root MSE       0.07441 (9)     R-square       0.7889 (12)
  Dep Mean       0.14275 (10)    Adj R-sq       0.6951 (13)
  C.V.          52.12940 (11)
```

(1) degrés de liberté de la loi de Fisher du test global
(2) SSR
(3) SSE ou déviance
(4) SST=SSE+SSR
(5) SSR/DF
(6) s^2 =MSE=SSE/DF est l'estimation de σ_u^2
(7) Statistique F du test de Fisher du modèle global
(8) $P(f_{p;n-p-1} > F)$; H_0 est rejetée au niveau α si $P < \alpha$
(9) s =racine de MSE
(10) moyenne empirique de la variable à expliquée
(11) Coefficient de variation $100\times$ (9)/(10)
(12) Coefficient de détermination R^2
(13) Coefficient de détermination ajusté R'^2

```
Parameter Estimates
                    Parameter      Standard    T for H0:                        Variance
Variable  DF       Estimate         Error    Parameter=0 Prob>|T|    Tolerance   Inflation
                       (1)           (2)          (3)        (4)          (5)         (6)
INTERCEP   1       0.188072    0.13391661       1.404     0.1716    .            0.00000000
WCFTCL     1       0.215130    0.19788455       1.087     0.2866    0.03734409  26.77799793
WCFTDT     1       0.305557    0.29736579       1.028     0.3133    0.02187972  45.70441500
GEARRAT    1      -0.040436    0.07677092      -0.527     0.6027    0.45778579   2.18442778
LOGSALE    1       0.118440    0.03611612       3.279     0.0029    0.10629382   9.40788501
LOGASST    1      -0.076960    0.04517414      -1.704     0.0999    0.21200778   4.71680805
NFATAST    1      -0.369977    0.13739742      -2.693     0.0120    0.20214372   4.94697537
CAPINT     1      -0.014138    0.02338316      -0.605     0.5505    0.37587215   2.66047911
FATTOT     1      -0.100986    0.08764238      -1.152     0.2593    0.23929677   4.17891139
INVTAST    1       0.250562    0.18586858       1.348     0.1888    0.13770716   7.26178633
PAYOUT     1      -0.018839    0.01769456      -1.065     0.2965    0.84271960   1.18663431
QUIKRAT    1       0.176709    0.09162882       1.929     0.0644    0.00408524 244.78377222
CURRAT     1      -0.223281    0.08773480      -2.545     0.0170    0.00486336 205.61923071
```

(1) estimations des paramètres (b_j)
(2) écarts-types de ces estimations (s_{b_j})
(3) statistique T du test de Student de H_0 : $b_j = 0$
(4) $P(t_{n-p-1} > T)$; H_0 est rejetée au niveau α si $P < \alpha$
(5) $1 - R^2_{(j)}$
(6) VIF=$1/(1 - R^2_{(j)})$

Ces résultats soulignent les problèmes de colinéarités. De grands "VIF" sont associés à de grands écart-types des estimations des paramètres. D'autre part les nombreux tests de Student non significatifs renforcent l'idée que trop de variables sont présentes dans le modèle.

```
Covariance of Estimates
COVB        INTERCEP     WCFTCL     WCFTDT    GEARRAT    LOGSALE    LOGASST    NFATAST
INTERCEP  0.0179336  0.0072146  -0.008697  -0.000325  0.0016622  -0.004576  -0.001320
WCFTCL    0.0072146  0.039158   -0.055126  -0.003011  -0.000043  -0.00161   -0.00129
WCFTDT    -0.008697  -0.055126  0.0884264  0.0087658  0.0000624  0.0017381  0.0008305
...
```

```
Collinearity Diagnostics
                    Condition
  Eigenvalue          Index

     8.76623        1.00000
     2.22300        1.98580
     0.68583        3.57518
     0.56330        3.94489
     0.31680        5.26036
     0.18140        6.95173
     0.12716        8.30291
     0.08451       10.18479
     0.02761       17.82007
     0.01338       25.59712
     0.00730       34.66338
     0.00223       62.63682
     0.00125       83.83978
```

Valeurs propres de $\mathbf{X'X}$ et indice de conditionnement égal au rapport $\sqrt{\lambda_1/\lambda_j}$. Les grandes

valeurs (> 10) insistent encore sur le mauvais conditionnement de la matrice à inverser.

Obs	Dep Var RETCAP (1)	Predict Value (2)	Std Err Predict (3)	Lower95 Mean (4)	Upper95 Mean (5)	Lower95 Predict (6)	Upper95 Predict (7)	Residual (8)	Std Err Residual (9)	Student Residual (10)
1	0.2600	0.2716	0.053	0.1625	0.3808	0.0839	0.4593	-0.0116	0.052	-0.223
2	0.5700	0.3690	0.039	0.2882	0.4497	0.1962	0.5417	0.2010	0.063	3.183
3	0.0900	0.00897	0.063	-0.1205	0.1385	-0.1912	0.2092	0.0810	0.039	2.055
4	0.3200	0.2335	0.021	0.1903	0.2768	0.0748	0.3922	0.0865	0.071	1.212
5	0.1700	0.1164	0.046	0.0215	0.2113	-0.0634	0.2961	0.0536	0.058	0.920
6	0.2400	0.2542	0.033	0.1864	0.3219	0.0871	0.4212	-0.0142	0.067	-0.213

Obs	-2-1-0 1 2 (11)	Cook's D (12)	Rstudent (13)	Hat Diag H (14)	Cov Ratio (15)	Dffits (15)	INTERCEP Dfbetas (15)	WCFTCL Dfbetas (15)	WCFTDT Dfbetas (15)
1	| |	0.004	-0.2194	0.5109	3.2603	-0.2242	0.0299	0.0632	-0.0911
2	|******	0.302	3.9515	0.2795	0.0050	2.4611	0.9316	-0.3621	0.3705
3	|****	0.832	2.1955	0.7192	0.6375	3.5134	0.5543	2.1916	-2.0241
4	|**	0.010	1.2228	0.0803	0.8585	0.3613	-0.0132	-0.0835	0.1207
5	|*	0.041	0.9175	0.3864	1.7591	0.7280	-0.0386	0.0906	0.0060
6	| |	0.001	-0.2088	0.1969	1.9898	-0.1034	0.0189	-0.0203	0.0243
15	***|	0.150	-1.9223	0.3666	0.4583	-1.4623	-0.2063	0.3056	-0.6231
16	|***	3.471	1.6394	0.9469	8.5643	6.9237	-0.9398	0.2393	-0.2323
17	| |	0.000	0.1401	0.1264	1.8514	0.0533	0.0223	0.0090	-0.0113
20	***|	0.054	-1.9588	0.1677	0.3278	-0.8794	-0.0360	-0.3302	0.4076
21	****|	4.970	-2.2389	0.9367	2.6093	-8.6143	-1.2162	0.1768	-0.1422

...

(1)	variable à expliquer y_i
(2)	valeur ajustée $\widehat{y}_i$
(3)	écart-type de cette estimation$s_{\widehat{y}_i}$
(4)et (5)	Intervalle de confiance pour l'estimation de $E(y_i)$
(6) et (7)	Intervalle de confiance pour l'estimation de y_i
(8)	résidus calculés e_i
(9)	écarts-types de ces estimations
(10)	résidus standardisés (ou studentisés internes) r_i
(11)	repérage graphique des résidus standardisés : $* = 0.5$.
(12)	Distance de Cook
(13)	résidus studentisés (externes) t_i
(14)	Termes diagonaux de la matrice chapeau $\mathbf{H}$
(15)	autres indicateurs d'influence

Seules les observations 16 et 21 seraient à inspecter avec attention.

```
Sum of Residuals                     0
Sum of Squared Residuals        0.1495   (SSE)
Predicted Resid SS (Press)      1.0190   (PRESS)
```

Sélection du modèle

Parmi les trois types d'algorithmes et les différents critères de choix, une des façons les plus efficaces consistent à choisir les options du programme ci-dessous. Tous les modèles (parmi les plus intéressants selon l'algorithme de Furnival et Wilson) sont considérés. Seul le meilleur pour chaque niveau, c'est-à-dire pour chaque valeur p du nombre de variables explicatives sont donnés. Il est alors facile de choisir celui minimisant l'un des critères globaux (C_p ou BIC ou ...).

```
options linesize=110 pagesize=30 nodate nonumber;
title;
proc reg data=sasuser.ukcomp2 ;
model RETCAP = WCFTCL   WCFTDT   GEARRAT   LOGSALE   LOGASST
       NFATAST  CAPINT   FATTOT   INVTAST   PAYOUT   QUIKRAT   CURRAT
       / selection=rsquare cp rsquare bic best=1;
run;
```

```
   N = 40     Regression Models for Dependent Variable: RETCAP
  R-square Adjusted  C(p)     BIC    Variables in Model
In         R-square
 1  0.1055 0.0819 78.3930 -163.26 WCFTCL
 2  0.3406 0.3050 50.3232 -173.72 WCFTDT QUIKRAT
 3  0.6154 0.5833 17.1815 -191.14 WCFTCL NFATAST CURRAT
 4  0.7207 0.6888  5.7146 -199.20 WCFTDT LOGSALE NFATAST CURRAT
 5  0.7317 0.6923  6.3047 -198.05 WCFTDT LOGSALE NFATAST QUIKRAT CURRAT
 6  0.7483 0.7025  6.1878 -197.25 WCFTDT LOGSALE NFATAST INVTAST QUIKRAT CURRAT
 7  0.7600 0.7075  6.6916 -195.77 WCFTDT LOGSALE LOGASST NFATAST FATTOT QUIKRAT CURRAT
 8  0.7692 0.7097  7.5072 -193.87 WCFTDT LOGSALE LOGASST NFATAST FATTOT INVTAST QUIKRAT CURRAT
 9  0.7760 0.7088  8.6415 -191.59 WCFTCL WCFTDT LOGSALE LOGASST NFATAST FATTOT INVTAST QUIKRAT
                                   CURRAT
10  0.7830 0.7082  9.7448 -189.15 WCFTCL WCFTDT LOGSALE LOGASST NFATAST FATTOT INVTAST PAYOUT
                                   QUIKRAT CURRAT
11  0.7867 0.7029 11.2774 -186.40 WCFTCL WCFTDT LOGSALE LOGASST NFATAST CAPINT FATTOT INVTAST
                                   PAYOUT QUIKRAT CURRAT
12  0.7888 0.6950 13.0000 -183.51 WCFTCL WCFTDT GEARRAT LOGSALE LOGASST NFATAST CAPINT FATTOT
                                   INVTAST PAYOUT QUIKRAT CURRAT
```

Dans cet example, C_p et BIC se comportent de la même façon. Avec peu de variables, le modèle est trop biaisé. Ils atteignent un minimum pour un modèle à 4 variables explicatives puis croissent de nouveau selon la première bissectrice. La maximisation du R^2 ajusté conduirait à une solution beaucoup moins parcimonieuse. On note par ailleurs que l'algorithme remplace WCFTCL par WCFTDT. Un algorithme par sélection ne peut pas aboutir à la solution optimale retenue.

Résultats du modèle réduit

```
proc reg data=sasuser.ukcomp1 all;
model RETCAP =  WCFTDT  NFATAST   LOGSALE  CURRAT
       /dw Influence cli clm tol vif collin r p ;
```

```
output out=resout h=lev p=pred r=res student=resstu ;
plot (student. r.)*p.;
plot p.*retcap;
run;
```

```
Analysis of Variance
                          Sum of          Mean
Source           DF      Squares        Square      F Value       Prob>F
Model             4      0.51043       0.12761       22.583       0.0001
Error            35      0.19777       0.00565
C Total          39      0.70820
       Root MSE         0.07517     R-square       0.7207
       Dep Mean         0.14275     Adj R-sq       0.6888
       C.V.            52.65889

Parameter Estimates
             Parameter      Standard    T for H0:                              Variance
Variable  DF  Estimate         Error   Parameter=0    Prob > |T|  Tolerance     Inflation
INTERCEP   1  0.024204    0.07970848       0.304        0.7632     .           0.00000000
WCFTDT     1  0.611885    0.08257125       7.410        0.0001    0.28956358   3.45347296
NFATAST    1 -0.474448    0.07015433      -6.763        0.0001    0.79119995   1.26390301
LOGSALE    1  0.060962    0.01606877       3.794        0.0006    0.54792736   1.82505944
CURRAT     1 -0.068949    0.01321091      -5.219        0.0001    0.21887292   4.56886122

Collinearity Diagnostics
                       Condition Var Prop Var Prop  Var Prop  Var Prop  Var Prop
Number  Eigenvalue       Index  INTERCEP  WCFTDT    NFATAST   LOGSALE   CURRAT
     1     3.86169     1.00000  0.0014   0.0076    0.0098    0.0016    0.0052
     2     0.87647     2.09904  0.0014   0.0608    0.0355    0.0046    0.0427
     3     0.17128     4.74821  0.0206   0.1731    0.5177    0.0170    0.0667
     4     0.07821     7.02670  0.0026   0.7201    0.4369    0.0388    0.5481
     5     0.01235    17.68485  0.9741   0.0384    0.0000    0.9381    0.3373

                      Cook's          Hat Diag  Cov            INTERCEP  WCFTDT  NFATAST
Obs   -2-1-0 1 2          D  Rstudent    H      Ratio  Dffits   Dfbetas Dfbetas  Dfbetas
15  |   ***|      |  0.211  -1.9115  0.2372   0.9096 -1.0659 -0.0240  -0.8161  -0.3075
16  |      |*     |  1.554   0.9919  0.8876   8.9162  2.7871  0.0320  -0.0746   0.1469
17  |      |      |  0.001   0.3866  0.0460   1.1854  0.0849  0.0348  -0.0430   0.0256

Sum of Residuals            0
Sum of Squared Residuals    0.1978  (Par rapport au modèle complet, la déviance augmente
Predicted Resid SS (Press)  0.3529   mais PRESS diminue très sensiblement)
```