NOTES DE STATISTIQUE ET D'INFORMATIQUE

88/1

LES CRITERES DE VALIDATION DES EQUATIONS DE REGRESSION LINEAIRE

R. PALM

Faculté des Sciences Agronomiques
GEMBLOUX
(Belgique)

LES CRITÈRES DE VALIDATION DES ÉQUATIONS DE RÉGRESSION LINÉAIRE

R. PALM[1]

RÉSUMÉ

Cette note présente une série de caractéristiques disponibles dans la littérature et permettant essentiellement la détection de données influentes et des situations de colinéarité, dans le cas de la régression linéaire. Une attention particulière est consacrée aux notions utilisées dans les programmes de régression des logiciels Minitab et SAS. Deux exemples sont examinés, à titre d'illustrations.

SUMMARY

This note briefly reviews the main influence and collinearity diagnostics which have been presented in the literature, in the context of linear regression. A special attention is given to the diagnostics used by the Minitab and SAS softwares. Two illustrative examples are given, too.

1. INTRODUCTION

Au cours des dix dernières années, la régression linéaire a connu d'importants développements théoriques, notamment dans le domaine de la recherche d'observations extrêmes (données anormales ou influentes) et de la détection de la colinéarité.

Différentes caractéristiques ont été proposées pour aider l'utilisateur lors de l'examen des données et de la validation des équations de régression, et les versions récentes des logiciels statistiques permettent le calcul de ces caractéristiques ou, du moins, d'une partie d'entre elles.

[1] Chef de travaux et Maître de conférences à la Faculté des Sciences Agronomiques de l'État, à Gembloux.

L'utilisateur se trouve donc assez automatiquement en face de documents imprimés contenant un ensemble d'informations dont il ne comprend pas toujours très bien la signification.

L'objectif de cette note est de présenter, de façon succincte, une série de notions assez récentes et souvent mal connues des utilisateurs non statisticiens des logiciels de régression. Une attention particulière sera consacrée aux notions utilisées dans les programmes de régression des logiciels Minitab et SAS.

Après quelques informations préliminaires relatives à la présentation des données et au modèle théorique (paragraphe 2), nous consacrerons un paragraphe aux notions classiques d'estimation des paramètres, de variance résiduelle et de coefficient de détermination (paragraphe 3). Nous définirons ensuite les prédictions et différents types de résidus (paragraphe 4). Les mesures de l'influence des individus seront alors présentées au paragraphe 5 et les outils permettant de mettre en évidence la colinéarité au paragraphe 6. Quelques informations concernant les représentations graphiques seront données dans le paragraphe 7. Le paragraphe 8 sera consacré à l'examen de certaines procédures particulières des logiciels Minitab et SAS, et deux exemples seront examinés, à titre d'illustrations, aux paragraphes 9 et 10. Enfin, nous terminerons par quelques conclusions (paragraphe 11).

Des informations complémentaires concernant les notions classiques de régression peuvent être trouvées dans de nombreux ouvrages généraux de statistique et notamment dans les livres de DAGNELIE [1979-1980, 1982]. Pour les notions plus nouvelles, on trouvera des informations complémentaires dans les éditions récentes d'ouvrages généraux consacrés à la régression, tels que ceux de DRAPER et SMITH [1981] et de WEISBERG [1985], et dans les livres très spécialisés de ATKINSON [1986], de BESLEY *et al.* [1980] et de COOK et WEISBERG [1982].

On notera que le vaste et classique problème de la sélection des variables en régression linéaire multiple ne sera pas traité de façon systématique dans cette note. Une synthèse bibliographique a été faite par THOMPSON [1978a, 1978b] et le sujet est bien développé par DRAPER et SMITH [1981]. De même, le problème de l'autocorrélation des résidus ne sera pas abordé. Des informations à ce sujet sont données, notamment, par DRAPER et SMITH [1981] et par PALM [1986].

2. DONNÉES ET MODÈLE THÉORIQUE

Les valeurs observées de la variable dépendante constituent le vecteur y, de dimensions $n \times 1$:

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}.$$

Les valeurs observées des p variables explicatives constituent la matrice $\boldsymbol{X}$, de dimensions $n \times p$:

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}.$$

Quant au modèle théorique, il peut être écrit de la façon suivante pour tout individu caractérisé par un vecteur $\boldsymbol{x}_*$ de variables explicatives:

$$y_* = \beta_1 x_{*1} + \beta_2 x_{*2} + \ldots + \beta_p x_{*p} + \varepsilon_* = \boldsymbol{x}_* \boldsymbol{\beta} + \varepsilon_*,$$

avec:

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix},$$

et:

$$\boldsymbol{x}_* = \begin{bmatrix} x_{*1} & x_{*2} & \cdots & x_{*p} \end{bmatrix}.$$

De plus, et pour autant qu'on envisage la réalisation de tests statistiques ou le calcul de limites de confiance, on considère que les résidus ε_* sont des réalisations indépendantes d'une même variable aléatoire normale, de moyenne nulle et d'écart-type σ.

On constate que ce modèle théorique ne présente pas de terme indépendant de façon explicite. Si on souhaite un terme indépendant, il suffit de considérer qu'une variable explicative est constante et égale à l'unité. Pour tout individu, on a alors, par exemple:

$$x_{*1} = 1.$$

Dans ce cas, la matrice $\boldsymbol{X}$ des observations relatives aux variables explicatives contient également une colonne de valeurs égales à l'unité, par exemple :

$$x_{11} = x_{21} = \ldots = x_{n1} = 1.$$

De façon générale, il ne sera pas nécessaire de dissocier, par la suite, la régression par l'origine et la régression avec terme indépendant, et p désignera toujours le nombre total de paramètres de l'équation.

3. ESTIMATION DES PARAMÈTRES, VARIANCE RÉSIDUELLE ET COEFFICIENT DE DÉTERMINATION

En désignant par $\boldsymbol{e}$ le vecteur, de dimensions $n \times 1$, des n résidus observés:

$$\boldsymbol{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix},$$

on peut écrire la relation matricielle suivante:

$$\boldsymbol{y} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} + \boldsymbol{e},$$

$\widehat{\boldsymbol{\beta}}$ étant le vecteur estimé des paramètres de l'équation de régression.

L'estimation au sens des moindres carrés du vecteur $\boldsymbol{\beta}$ consiste à minimiser la quantité:

$$(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) = \boldsymbol{e}'\boldsymbol{e} = \sum_{i=1}^{n} e_i^2,$$

et on obtient la solution suivante:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}.$$

Dans cette relation, la matrice $\boldsymbol{X}'\boldsymbol{X}$, de dimensions $p \times p$, est la matrice des sommes des carrés et des produits des variables explicatives. En particulier, si tous les éléments de la première colonne de $\boldsymbol{X}$ sont égaux à l'unité, la matrice $\boldsymbol{X}'\boldsymbol{X}$ présente la structure suivante:

$$\boldsymbol{X}'\boldsymbol{X} = \begin{bmatrix} n & \sum_{i=1}^{n} x_{i2} & \cdots & \sum_{i=1}^{n} x_{ip} \\ \sum_{i=1}^{n} x_{i2} & \sum_{i=1}^{n} x_{i2}^2 & \cdots & \sum_{i=1}^{n} x_{i2}x_{ip} \\ \vdots & \vdots & & \vdots \\ \sum_{i=1}^{n} x_{ip} & \sum_{i=1}^{n} x_{ip}x_{i2} & \cdots & \sum_{i=1}^{n} x_{ip}^2 \end{bmatrix}.$$

Le vecteur $\boldsymbol{X}'\boldsymbol{y}$, de dimension $p \times 1$, contient les sommes des produits des valeurs de la variable à expliquer et des différentes variables explicatives. Si tous

les éléments de la première colonne de $\boldsymbol{X}$ sont égaux à l'unité, le premier élément du vecteur $\boldsymbol{X}'\boldsymbol{y}$ est égal à la somme des valeurs de la variable à expliquer.

Enfin, une estimation de la variance résiduelle, $\widehat{\sigma}^2$, est donnée par la relation suivante:

$$\widehat{\sigma}^2 = \frac{\boldsymbol{e}'\boldsymbol{e}}{n-p},$$

et la matrice des variances et covariances du vecteur $\widehat{\boldsymbol{\beta}}$ est donnée par:

$$\mathrm{v}(\widehat{\boldsymbol{\beta}}) = \widehat{\sigma}^2(\boldsymbol{X}'\boldsymbol{X})^{-1}.$$

La quantité $\boldsymbol{e}'\boldsymbol{e}$, qui représente la somme des carrés des écarts des résidus, peut encore s'écrire:

$$\boldsymbol{e}'\boldsymbol{e} = \boldsymbol{y}'\boldsymbol{y} - \widehat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y},$$

$\boldsymbol{y}'\boldsymbol{y}$ étant la somme des carrés des observations y_i et $\widehat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y}$ étant la somme des carrés expliquée par la régression.

Ces différentes sommes de carrés peuvent être présentées dans un tableau d'analyse de la variance (tableau 1).

Tableau 1. Tableau d'analyse de la variance: cas général.

Sources de variation	Degrés de liberté	Sommes des carrés	Carrés moyens
Régression	p	$\widehat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y}$	$CM_{rég}$
Résidus	$n-p$	$\boldsymbol{e}'\boldsymbol{e}$	$\widehat{\sigma}^2$
Totaux	n	$\boldsymbol{y}'\boldsymbol{y}$	

Dans le cas où l'équation de régression contient un terme indépendant, on divise généralement la source de variation "régression" en deux parties. La première somme de carrés, liée à la présence du terme indépendant, possède un degré de liberté et vaut $(\sum_{i=1}^{n} y_i)^2/n$. La seconde somme de carrés, liée à la présence des autres paramètres, possède $p-1$ degrés de liberté et vaut:

$$\widehat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y} - (\sum_{i=1}^{n} y_i)^2/n.$$

Sur base de cette dernière décomposition, on peut établir un tableau d'analyse de la variance légèrement différent du tableau 1 mais plus classique (tableau 2).

Tableau 2. Tableau d'analyse de la variance: régression avec terme indépendant.

Sources de variation	Degrés de liberté	Sommes des carrés des écarts	Carrés moyens
Régression	$p-1$	$\widehat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y} - \left(\sum_{i=1}^{n} y_i\right)^2/n$	$CM_{rég}$
Résidus	$n-p$	$\boldsymbol{e}'\boldsymbol{e}$	$\widehat{\sigma}^2$
Totaux	$n-1$	$\boldsymbol{y}'\boldsymbol{y} - \left(\sum_{i=1}^{n} y_i\right)^2/n$	

D'autres décompositions des p degrés de liberté liés à la régression sont couramment effectuées; elles font notamment appel au concept de régression partielle. Ces décompositions sont utilisées, par exemple, pour tester la signification des coefficients de régression partielle.

La décomposition de la somme des carrés des écarts totale donnée dans le tableau 2 permet de définir le coefficient de détermination:

$$R^2 = \frac{\widehat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y} - \left(\sum_{i=1}^{n} y_i\right)^2/n}{\boldsymbol{y}'\boldsymbol{y} - \left(\sum_{i=1}^{n} y_i\right)^2/n} = 1 - \frac{\boldsymbol{e}'\boldsymbol{e}}{\boldsymbol{y}'\boldsymbol{y} - \left(\sum_{i=1}^{n} y_i\right)^2/n},$$

qui représente la part de la variance de y qui est expliquée par la régression. On peut également tenir compte des nombres de degrés de liberté et définir le coefficient de détermination ajusté:

$$R_a^2 = 1 - \frac{\boldsymbol{e}'\boldsymbol{e}/(n-p)}{\left[\boldsymbol{y}'\boldsymbol{y} - \left(\sum_{i=1}^{n} y_i\right)^2/n\right]/(n-1)} = 1 - (1-R^2)\left(\frac{n-1}{n-p}\right).$$

Ce coefficient de détermination ajusté est une estimation moins biaisée de la valeur théorique correspondante que le coefficient R^2 [DAGNELIE, 1982].

Dans le cas de la régression par l'origine, le coefficient de détermination multiple est généralement défini par la relation:

$$R^2 = 1 - \frac{\boldsymbol{e}'\boldsymbol{e}}{\boldsymbol{y}'\boldsymbol{y}}.$$

Une discussion détaillée des différentes définitions possibles du coefficient de détermination pour les situations autres que le modèle linéaire avec terme indépendant est donnée par KVÅLSETH [1985].

4. PRÉDICTIONS ET RÉSIDUS

Une des utilisations de la régression est la réalisation de prédictions, $\widehat{y}_*$, en fonction de vecteurs $\boldsymbol{x}_*$, de dimension $1 \times p$, de valeurs données des variables explicatives.

La prédiction est obtenue par la relation:

$$\widehat{y}_* = \boldsymbol{x}_* \widehat{\boldsymbol{\beta}},$$

qu'il s'agisse de l'estimation de la moyenne conditionnelle ou de l'estimation d'une valeur individuelle.

Si on pose:

$$h_* = \boldsymbol{x}_* (\boldsymbol{X}'\boldsymbol{X})^{-1} \boldsymbol{x}_*',$$

la variance estimée de la prédiction est égale à:

$$\mathrm{v}(\widehat{y}_*) = \widehat{\sigma}^2 h_*,$$

dans le cas de l'estimation d'une moyenne conditionnelle, et égale à:

$$\mathrm{v}(\widehat{y}_*) = \widehat{\sigma}^2 (1 + h_*),$$

dans le cas de l'estimation d'une valeur individuelle.

Ces variances permettent le calcul des limites de confiance correspondantes:

$$\widehat{y}_* + t_{1-\alpha/2} \sqrt{v(\widehat{y}_*)},$$

$t_{1-\alpha/2}$ étant la valeur de la variable t de STUDENT à $n - p$ degrés de liberté, pour laquelle la fonction de répartition vaut $1 - \alpha/2$.

Les valeurs estimées pour l'ensemble des individus de l'échantillon s'obtiennent par la relation:

$$\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{H}\boldsymbol{y}.$$

La matrice de projection[1] $\boldsymbol{H}$, de dimensions $n \times n$, transforme donc le vecteur des valeurs observées $\boldsymbol{y}$ en un vecteur de valeurs estimées[2] $\widehat{\boldsymbol{y}}$.

[1] En anglais: *hat matrix, projection matrix, prediction matrix.*
[2] En anglais: *y-hat.*

Les éléments diagonaux[3], h_{ii}, de cette matrice de projection sont égaux à:

$$h_{ii} = \boldsymbol{x}_i(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i',$$

$\boldsymbol{x}_i$ étant la $i^{ème}$ ligne de la matrice $\boldsymbol{X}$.

On peut montrer que, dans le cas de la régression avec terme indépendant, ces quantités ont une moyenne égale à p/n et sont toujours telles que:

$$\frac{1}{n} \leq h_{ii} \leq \frac{1}{r},$$

r étant le nombre de lignes de la matrice $\boldsymbol{X}$ qui sont égales à $\boldsymbol{x}_i$, c'est-à-dire aussi le nombre d'individus de l'échantillon pour lesquels les valeurs observées des p variables explicatives sont identiques aux valeurs observées pour le $i^{ème}$ individu. On peut également montrer que, dans le cas de la régression avec terme indépendant, les valeurs h_{ii} sont, à une transformation linéaire près, identiques aux distances de MAHALANOBIS, dans l'espace des $p-1$ variables explicatives $x_2, x_3, \cdots, x_p$, entre le $i^{ème}$ point et le point moyen. Ainsi, pour la régression linéaire simple par exemple:

$$y = \beta_0 + \beta_1 x + \varepsilon,$$

on a:

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{SCE_x}.$$

La variance de l'estimation d'une valeur moyenne, $\widehat{y}_i$, est égale à:

$$\mathrm{v}(\widehat{y}_i) = \widehat{\sigma}^2 h_{ii},$$

et, pour le vecteur des résidus observés, $\boldsymbol{e}$, on peut montrer que sa matrice estimée des variances et covariances est égale à:

$$\mathrm{v}(\boldsymbol{e}) = \widehat{\sigma}^2(\boldsymbol{I} - \boldsymbol{H}).$$

On constate donc que, contrairement aux résidus théoriques ε_i, les résidus estimés n'ont pas une variance constante et sont corrélés. En particulier, le résidu associé à la $i^{ème}$ observation a pour variance estimée:

$$\mathrm{v}(e_i) = \widehat{\sigma}^2(1 - h_{ii}).$$

La variance du résidu est donc d'autant plus faible que h_{ii} est grand, c'est-à-dire qu'il est associé à un individu pour lequel $\boldsymbol{x}_i$ est éloigné de $\bar{\boldsymbol{x}}$, du moins lorsque la régression comporte un terme indépendant. Pour éliminer cette inégalité des variances des résidus observés, on définit des résidus standardisés, de variance constante et égale à l'unité[4]:

$$r_i = \frac{e_i}{\widehat{\sigma}\sqrt{1 - h_{ii}}}.$$

---

[3] En anglais: *leverage, potential.*

[4] En anglais: *internally Studentized residuals, Studentized residuals, standardized residuals.*

Les résidus peuvent aussi être standardisés d'une façon légèrement différente[5]:

$$t_i = \frac{e_i}{\widehat{\sigma}_{(i)}\sqrt{1-h_{ii}}},$$

$\widehat{\sigma}_{(i)}$ étant l'écart-type résiduel estimé à partir de l'équation de régression calculée après élimination de la $i^{ème}$ observation. Sous les conditions d'application définies au paragraphe 2, ces t_i possèdent une distribution de STUDENT à $n-p$ degrés de liberté. Ils peuvent, par conséquent, être utilisés pour la détection de valeurs anormales. Des informations à ce sujet sont données par WEISBERG [1985].

Enfin, on définit encore les résidus de prédiction[6]:

$$e_{(i)} = y_i - \boldsymbol{x}_i\widehat{\boldsymbol{\beta}}_{(i)} = \frac{e_i}{1-h_{ii}},$$

$\widehat{\boldsymbol{\beta}}_{(i)}$ étant le vecteur des coefficients de régression estimés après la suppression de la $i^{ème}$ observation.

On constate que les résidus de prédiction peuvent être obtenus directement à partir des résultats de la régression calculée sur n observations, sans qu'il soit nécessaire de calculer n fois l'équation de régression sur $n-1$ observations. Il en est d'ailleurs de même pour la détermination des variances $\widehat{\sigma}^2_{(i)}$, définies ci-dessus.

Ces résidus de prédiction permettent notamment de définir la somme des carrés des erreurs de prédiction[7], désignée par le symbole *PRESS*:

$$PRESS = \sum_{i=1}^{n} e^2_{(i)}.$$

5. INFLUENCE DES INDIVIDUS DE L'ÉCHANTILLON

Différents paramètres ont été définis dans le but de chiffrer l'influence ou l'importance de chacune des n observations dans l'ajustement du modèle. Ces caractéristiques se basent essentiellement sur la comparaison des résultats obtenus, d'une part, lorsqu'on ajuste le modèle à l'ensemble des données et, d'autre part, lorsqu'on ajuste le modèle après avoir supprimé une observation.

Conformément aux notations déjà utilisées au paragraphe précédent, les caractéristiques obtenues après suppression de l'observation i seront affectées de l'indice i. Ainsi, par exemple, le vecteur des coefficients de régression obtenus après suppression de l'observation i s'écrit:

$$\boldsymbol{\beta}_{(i)} = (\boldsymbol{X}'_{(i)}\boldsymbol{X}_{(i)})^{-1}\boldsymbol{X}'_{(i)}\boldsymbol{y}_{(i)}.$$

---

[5] En anglais: *externally Studentized residuals, Studentized residuals, "cross-validatory" residuals, "jackknife" residuals.*

[6] En anglais: *predicted residuals.*

[7] En anglais: *predicted residual sum of squares.*

Une mesure de l'influence d'une observation sur les valeurs estimées est donnée par la distance de COOK [COOK, 1977], qui est fonction de l'écart entre $\widehat{\boldsymbol{\beta}}$ et $\widehat{\boldsymbol{\beta}}_{(i)}$:

$$D_i = \frac{(\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}})'(\boldsymbol{X}'\boldsymbol{X})(\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}})}{p\widehat{\sigma}^2}.$$

Cette relation peut aussi s'écrire:

$$D_i = \frac{(\widehat{\boldsymbol{y}}_{(i)} - \widehat{\boldsymbol{y}})'(\widehat{\boldsymbol{y}}_{(i)} - \widehat{\boldsymbol{y}})}{p\widehat{\sigma}^2},$$

avec:

$$\widehat{\boldsymbol{y}}_{(i)} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{(i)} \text{ et } \widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}.$$

La distance de COOK est donc aussi directement fonction de la distance entre le vecteur des valeurs prédites, $\widehat{\boldsymbol{y}}$, obtenu quand toutes les observations sont utilisées pour le calcul de la régression, et le vecteur des valeurs prédites, $\widehat{\boldsymbol{y}}_{(i)}$, obtenu lorsque la $i^{ème}$ observation est supprimée pour le calcul de la régression.

On peut montrer que la distance de COOK est encore égale à:

$$D_i = \frac{r_i^2}{p}\frac{h_{ii}}{(1-h_{ii})} = \frac{e_i^2 h_{ii}}{p\widehat{\sigma}^2(1-h_{ii})^2}.$$

Cette relation montre que la distance D_i est une fonction croissante du carré du résidu standardisé et de h_{ii}. Pour une valeur fixée de p et pour une régression avec terme indépendant, D_i sera d'autant plus grand que e_i est grand, en valeur absolue, et que le vecteur $\boldsymbol{x}_i$ est éloigné du vecteur $\bar{\boldsymbol{x}}$. Une valeur élevée de D_i signifie que l'observation i est influente, c'est-à-dire qu'elle a une forte contribution à la régression. Une telle observation a donc une grande influence sur le calcul de $\widehat{\boldsymbol{\beta}}$ et des valeurs ajustées $\widehat{\boldsymbol{y}}$ et sa suppression peut conduire à d'importantes modifications dans les conclusions. Pour déterminer de façon plus précise l'influence de cette observation sur l'estimation des paramètres et sur le calcul des valeurs prédites, on la supprime et on recommence l'analyse.

On sera cependant attentif au fait qu'une donnée influente n'est pas nécessairement une donnée anormale, la valeur élevée de D_i pouvant être due à la valeur élevée de h_{ii}. De même, une donnée anormale, c'est-à-dire à laquelle est associé un résidu e_i important, n'est pas nécessairement influente. Il suffit, en effet, pour cela qu'elle corresponde à une valeur faible de h_{ii}, c'est-à-dire à un vecteur $\boldsymbol{x}_i$ proche du vecteur moyen $\bar{\boldsymbol{x}}$. Ce point sera illustré par un exemple au paragraphe 9.

Sous les hypothèses énoncées au paragraphe 2, les D_i peuvent être comparés à la valeur $F_{1-\alpha}$ relative à la variable F de SNEDECOR à p et $n-p$ degrés de liberté, bien qu'il ne s'agisse pas d'un test statistique rigoureux [OBENCHAIN, 1977]. Sur cette base, WEISBERG [1985] considère qu'une attention particulière doit être accordée aux observations pour lesquelles D_i est supérieur à l'unité.

Une autre mesure de l'influence, représentée généralement par le symbole *DFFITS* ou *DFITS*, a été proposée par BELSLEY*et al.* [1980]:

$$DFFITS_i = \frac{\widehat{y}_i - \widehat{y}_{(i)}}{\widehat{\sigma}_{(i)}\sqrt{h_{ii}}},$$

et, selon ces auteurs, les observations pour lesquelles $DFFITS_i$ a une valeur absolue supérieure à $2\sqrt{p/n}$ devraient être considérées comme influentes.

Cette mesure de l'influence est en fait fort semblable à la distance de COOK. On a en effet:

$$(DFFITS_i)^2 = \frac{(\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}})'(\boldsymbol{X}'\boldsymbol{X})(\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}})}{\widehat{\sigma}^2_{(i)}}.$$

Le carré de $DFFITS_i$ ne diffère donc de la distance de COOK que par un facteur d'échelle et par le remplacement de $\widehat{\sigma}$ par $\widehat{\sigma}_{(i)}$.

Pour mesurer l'influence d'un individu sur la variance des coefficients de régression, BELSLEY *et al.* [1980] proposent l'utilisation du rapport des déterminants des matrices de variances et covariances des vecteurs $\widehat{\boldsymbol{\beta}}_{(i)}$ et $\widehat{\boldsymbol{\beta}}$:

$$COVRATIO_i = \frac{\widehat{\sigma}^2_{(i)} \mid (\boldsymbol{X}'_{(i)}\boldsymbol{X}_{(i)})^{-1} \mid}{\widehat{\sigma}^2 \mid (\boldsymbol{X}'\boldsymbol{X})^{-1} \mid}.$$

Ces auteurs suggèrent qu'on considère comme influentes les données pour lesquelles on a:

$$\mid COVRATIO_i - 1 \mid > 3p/n.$$

Quant à l'influence d'un individu sur chacun des paramètres, elle peut être mesurée par le calcul des quantités suivantes:

$$DFBETAS_{ij} = \frac{\beta_j - \beta_{j(i)}}{\widehat{\sigma}_{(i)}\sqrt{(\boldsymbol{X}'\boldsymbol{X})^{jj}}},$$

$(\boldsymbol{X}'\boldsymbol{X})^{jj}$ étant l'élément (j,j) de la matrice $(\boldsymbol{X}'\boldsymbol{X})^{-1}$. Il s'agit de la différence entre les valeurs du paramètre relatif à la $j^{ème}$ variable lorsque la $i^{ème}$ observation est prise en considération ou non, divisée par une estimation de l'écart-type du paramètre en question. BELSLEY *et al.* [1980] considèrent que les observations pour lesquelles $DFBETAS_{ij}$ est, en valeur absolue, supérieur à $2/\sqrt{n}$ méritent une attention spéciale.

D'autres mesures de l'influence sont encore proposées dans la littérature. Des informations à leur sujet sont données dans les synthèses bibliographiques de CHATTERJEE et HADI [1986] et de HOCKING [1983].

6. ÉTUDE DE LA COLINÉARITÉ

Dans le cas de deux variables explicatives, la colinéarité exacte implique l'existence d'une relation linéaire entre les deux variables. On a alors:

$$c_1 x_{i1} + c_2 x_{i2} = c_0,$$

c_0, c_1 et c_2 étant des constantes. Cette colinéarité exacte correspond à un coefficient de détermination entre x_1 et x_2 égal à l'unité.

Dans le cas de p variables explicatives, on dit qu'il y a colinéarité si:

$$c_1 x_{i1} + c_2 x_{i2} + \cdots + c_p x_{ip} = c_0,$$

c_0, c_1, ..., c_p étant des constantes. Dans ce cas, le coefficient de détermination multiple entre une variable explicative et les $p-1$ autres variables explicatives est égal à l'unité.

La colinéarité stricte rend impossible l'inversion de la matrice $\boldsymbol{X}'\boldsymbol{X}$ et donc aussi le calcul du vecteur $\widehat{\boldsymbol{\beta}}$. Il faut alors supprimer une ou plusieurs variables explicatives.

La colinéarité approximative, qui correspond à un coefficient de détermination multiple entre une variable explicative et les autres variables explicatives proche de 1, conduit à une grande variance des coefficients de régression estimés. En effet, on peut montrer que, pour une équation avec terme indépendant, on a:

$$\mathrm{v}(\widehat{\beta}_j) = \widehat{\sigma}^2 \Big(\frac{1}{1-R_j^2}\Big)\Big(\frac{1}{SCE_{xj}}\Big),$$

SCE_{xj} étant la somme des carrés des écarts relative à la $j^{ème}$ variable et R_j^2 étant le coefficient de détermination multiple entre la variable j et les $p-1$ autres variables explicatives. La quantité:

$$\frac{1}{1-R_j^2},$$

appelée facteur d'inflation de la variance[8] et représentée par le symbole VIF_j, correspond à l'augmentation de variance due à la corrélation entre la variable j et les autres variables explicatives. Le terme $1-R_j^2$ est appelé tolérance[9] et est souvent désigné par le symbole TOL_j.

Dans le cas de la régression avec terme indépendant, le diagnostic de colinéarité peut également s'appuyer sur l'étude des valeurs propres de la matrice de corrélation des $p-1$ variables explicatives $x_2, x_3, \cdots, x_p$. Les rapports de la racine carrée de la valeur propre la plus élevée et de la racine carrée de

[8] En anglais: *variance inflation factor.*

[9] En anglais: *tolerance.*

chacune des autres valeurs propres sont appelés indices de conditionnement[10]. Plus particulièrement, le rapport:

$$k = \frac{\sqrt{l_1}}{\sqrt{l_p}},$$

entre la racine carrée de la plus grande et la racine carrée de la plus petite valeur propre[11], est une mesure de la colinéarité. Si ce rapport est trop élevé, on dit que le problème est mal conditionné[12]. Certains auteurs considèrent que le problème de colinéarité est à prendre en considération dès que $k \geq 30$, mais cette affirmation n'a que peu de justification théorique [WEISBERG, 1985].

Enfin, on peut également calculer la proportion de la variance des coefficients de régression partielle qui est prise en considération par chacune des composantes principales de la matrice des variables explicatives. Le problème de colinéarité existe si une composante à indice de conditionnement élevé contribue fortement à la variance de deux ou de plusieurs coefficients. Ce problème est bien illustré par TOMASSONE *et al.* [1983].

Ce choix de la matrice de corrélation des $p-1$ variables explicatives x_2, x_3, $\cdots$, x_p, comme point de départ pour le calcul des paramètres mesurant l'intensité de la colinéarité n'est pas accepté par tous les auteurs. Ainsi, BELSLEY *et al.* [1980] font toute l'analyse à partir de la matrice $\boldsymbol{X}_0$, de dimensions $p \times n$, qui est la matrice originale $\boldsymbol{X}$ dont les éléments de chaque colonne ont été divisés par la somme des carrés des éléments de la colonne. Chaque colonne de $\boldsymbol{X}_0$ constitue donc un vecteur de moyenne généralement non nulle et de longueur unitaire. Cette matrice $\boldsymbol{X}_0$ est utilisée à la fois pour la régression avec un terme indépendant et pour la régression passant par l'origine.

Une discussion détaillée relative au choix de la matrice de départ peut être trouvée dans l'article de BELSLEY [1984], ainsi que dans les commentaires des différents auteurs concernant cet article. Des informations complémentaires relatives aux diagnostics de colinéarité sont également données par STEWART [1987].

En présence de colinéarité, il y a intérêt à supprimer une ou plusieurs variables explicatives. L'examen du tableau des contributions relatives des composantes principales à la variance des paramètres, auquel nous avons fait allusion ci-dessus, permet de repérer le ou les groupes de variables explicatives dans lesquels une ou plusieurs variables devraient être supprimées.

Le problème de colinéarité sera illustré par un exemple au paragraphe 10.

---

[10] En anglais: *condition indices.*
[11] En anglais: *condition number.*
[12] En anglais: *ill-conditioned.*

7. REPRÉSENTATIONS GRAPHIQUES

De nombreuses représentations graphiques peuvent être réalisées en vue de vérifier la validité des modèles de régression.

On peut tout d'abord établir des histogrammes des résidus ou des résidus standardisés, afin de vérifier, du moins de façon approximative, la symétrie de la distribution et la présence éventuelle de résidus anormaux.

La normalité des résidus peut également être visualisée par le diagramme de dispersion des résidus en fonction des scores normaux[13] [DRAPER et SMITH, 1981; WEISBERG, 1985]. Le score normal relatif au résidu de rang i correspond à l'espérance mathématique de la valeur de rang i lorsqu'on prélève un échantillon de n individus dans une population normale réduite. Dans le cas de résidus issus d'une population normale, le diagramme de dispersion des résidus en fonction des scores normaux est approximativement linéaire et le coefficient de corrélation entre ces deux caractéristiques permet de tester la normalité des résidus [CRYER, 1986; FILLIBEN, 1975].

La représentation graphique des résidus en fonction d'une variable explicative particulière ou de la variable dépendante, ou encore d'une autre variable non encore prise en considération dans le modèle, fournit également une série d'informations concernant l'adéquation du modèle [DRAPER et SMITH, 1981; PALM, 1986; WEISBERG, 1985].

Afin de visualiser l'effet d'une variable explicative x_j sur la variable y après l'élimination des effets des $p-1$ autres variables explicatives, on peut calculer, d'une part, la régression de y en fonction de ces $p-1$ variables explicatives, et, d'autre part, la régression de x_j en fonction de ces mêmes $p-1$ variables explicatives. Les résidus de la première équation sont alors mis en graphique en fonction des résidus de la seconde équation.

Le diagramme de dispersion ainsi obtenu, appelé diagramme de la variable ajoutée[14], donne des informations relatives à la nature de la liaison entre y et x_j. En effet, la pente de la droite de régression qui peut être calculée à partir de ce nuage de points est égale au coefficient de régression partielle de la variable x_j dans le modèle à p variables et la corrélation qui existe entre ces deux séries de résidus est le coefficient de corrélation partielle de y et x_j, après élimination des $p-1$ autres variables explicatives [WEISBERG, 1985].

Une représentation légèrement différente peut également être utilisée [LARSEN et McCLEARY, 1972]. Dans cette représentation graphique[15], on porte en abscisse les valeurs de la variable x_j et en ordonnée les valeurs:

$$e_i + x_{ij}\widehat{\beta}_j,$$

[13] En anglais: *normal probability plot, rankit plot.*

[14] En anglais: *added variable plot, partial regression leverage plot, partial residual plot.*

[15] En anglais: *residual plus component plot, partial residual plot.*

avec:

$$e_i = y_i - x_i\widehat{\beta}.$$

La quantité e_i est donc le résidu observé du $i^{ème}$ individu dans le cas du modèle à p variables et $\widehat{\beta}_j$ est le coefficient de régression partielle relatif à la variable x_j. La pente de la droite de régression relative à ce nuage de points est aussi égale au coefficient de régression partielle $\widehat{\beta}_j$, mais la dispersion des points n'est pas la même que dans le graphique précédent [WEISBERG, 1985].

Des informations complémentaires concernant les représentations graphiques mentionnées ci-dessus et concernant d'autres représentations graphiques sont données par ATKINSON [1986] et par CHATTERJEE et HADI [1986].

8. RÉALISATION DES CALCULS DE RÉGRESSION À L'AIDE DES LOGICIELS MINITAB ET SAS

Pour les problèmes de régression, deux commandes sont disponibles dans le logiciel Minitab. Il s'agit des commandes REGRESS et STEPWISE. La seconde commande est spécialement orientée vers les procédures de sélection de variables et ne donne guère d'informations relatives aux critères de validation des régressions, contrairement à la première.

De même, avec le logiciel SAS, diverses procédures peuvent être employées pour la régression linéaire. En particulier, la procédure STEPWISE offre divers algorithmes de sélection des variables et la procédure RSQUARE permet le calcul du coefficient de détermination multiple de toutes les équations possibles. Mais c'est la procédure REG qui fournit, pour un modèle donné, le plus d'informations relatives à la validation du modèle.

Nous nous limiterons donc à l'examen de la commande REGRESS de Minitab et de la procédure REG de SAS. Des informations complémentaires concernant ces commandes sont données dans les manuels d'utilisation de ces deux logiciels [X, 1985a, 1985b] et, pour le logiciel SAS, dans le document de FREUND et LITTELL [1986].

Le tableau 3 reprend les caractéristiques fournies, en option, par ces deux logiciels, en ce qui concerne les différents types de résidus et les mesures de l'influence des observations.

L'utilisateur devra cependant être attentif au fait que l'impression de ces caractéristiques peut conduire à des documents imprimés volumineux lorsqu'il traite des problèmes pour lesquels il dispose d'un grand nombre d'observations. L'examen de ces documents est alors long et fastidieux et il est préférable, dans ce cas, de se limiter à l'impression des caractéristiques pour une partie des observations seulement.

Cette solution est d'ailleurs proposée d'office par le logiciel Minitab. En effet, indépendamment de la liste complète des caractéristiques, celui-ci fournit un tableau des observations considérées comme inhabituelles. Les observations

Tableau 3. Résidus et mesures de l'influence disponibles par le logiciel Minitab et par le logiciel SAS.

Caractéristiques	Symboles dans les commandes Minitab	Symboles dans les procédures ou dans les documents de sortie SAS
e_i	RESIDS	R ou RESIDUAL
r_i	ST.RES	STUDENT
t_i	TRESIDS	RSTUDENT
$e_{(i)}$	-	PRESS
h_{ii}	HI	H
D_i	COOKD	COOKD
$DFFITS_i$	DFITS	DFFITS
$COVRATIO_i$	-	COVRATIO
$DFBETAS_{ij}$	-	DFBETAS

marquées du signe R sont celles pour lesquelles le résidu standardisé, r_i, est supérieur ou égal, en valeur absolue, à 2 et les observations marquées du signe X sont celles pour lesquelles h_{ii} est supérieur à $3p/n$.

Une telle sélection peut être faite assez facilement dans le cas du logiciel SAS, par exemple par l'utilisation de l'instruction IF, lors d'une étape DATA. Mais des sélections basées sur d'autres paramètres peuvent également être envisagées. On pourrait, par exemple, sélectionner les observations pour lesquelles l'une des trois conditions suivantes est remplie:

$$| r_i | > 2 \ , \ h_{ii} > 3p/n \text{ ou } D_i > 1.$$

En ce qui concerne les diagnostics de colinéarité, Minitab donne, en option, les valeurs de VIF_j et de TOL_j. De plus, lorsque le coefficient de détermination multiple d'une variable explicative x_j avec les $p-1$ autres variables explicatives est supérieur à 0,99 ($VIF > 100$), un message de mise en garde est imprimé. Lorsque ce coefficient dépasse 0,9999 ($VIF > 10.000$), la variable x_j est automatiquement supprimée.

Le logiciel SAS donne également les valeurs de VIF_j et TOL_j, mais il fournit, en outre, les valeurs propres de la matrice des variables explicatives et la contribution de chacune des composantes principales à la variance des coefficients de régression partielle. L'analyse peut être réalisée soit sur la matrice $\boldsymbol{X}_0$ définie au paragraphe 6 (option COLLIN), soit sur la matrice de corrélation des variables explicatives (option COLLINOINT).

Au point de vue des représentations graphiques, SAS donne, en option, pour toutes les variables explicatives, le graphique de la variable ajoutée. D'autres représentations graphiques peuvent évidemment être obtenues, tant par Minitab que par SAS, grâce à des instructions indépendantes de la commande REGRESS ou de la procédure REG.

9. EXEMPLE 1: ÉTUDE DE L'INFLUENCE DES INDIVIDUS

Pour illustrer les notions d'observations influentes et d'observations anormales, nous avons choisi un exemple de régression linéaire simple. Les données sont relatives aux nombres de fruits et aux poids des fruits récoltés (en grammes) sur 100 fraisiers et ont été publiées par DAGNELIE [1981]. On considère la droite de régression du poids des fruits en fonction du nombre de fruits par fraisier, même si le modèle linéaire avec terme indépendant n'est peut-être pas le plus adéquat [DAGNELIE, 1981].

La figure 1 donne le diagramme de dispersion de ces deux variables et le tableau 4 reprend, pour une série de caractéristiques, les cinq valeurs les plus extrêmes et les numéros des observations auxquelles correspondent ces

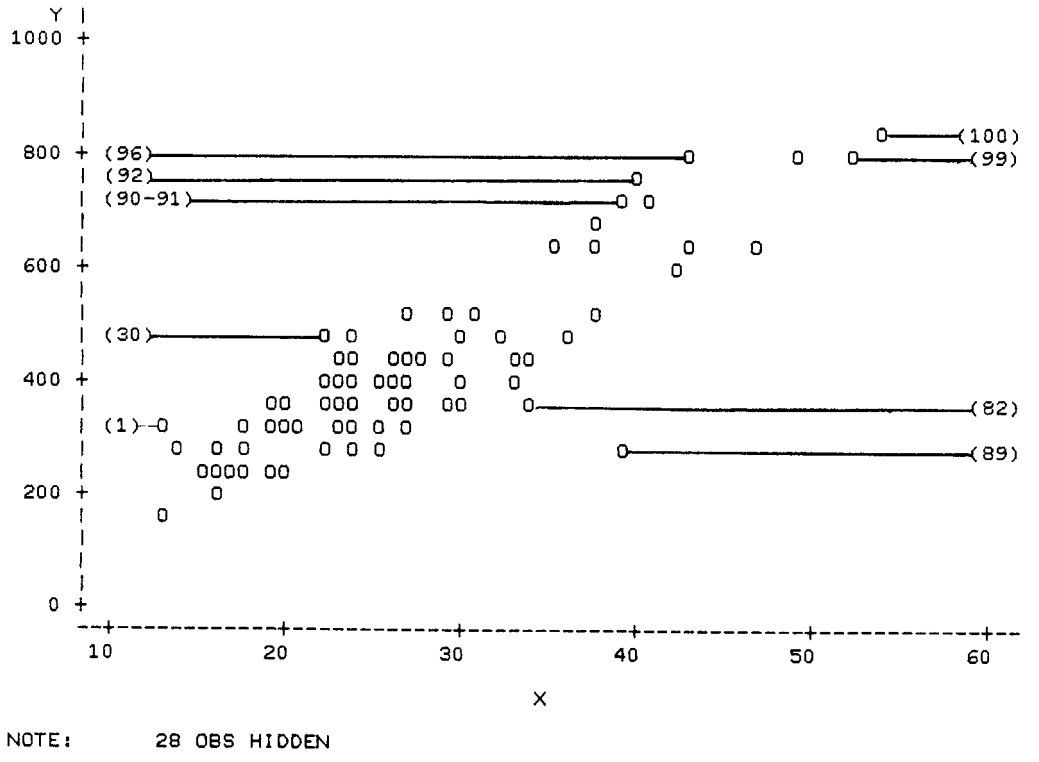

Figure 1. Diagramme de dispersion du poids des fruits et du nombre de fruits observés sur 100 fraisiers (les nombres figurant entre parenthèses correspondent aux numéros des observations reprises dans le tableau 4).

valeurs extrêmes. Ces numéros correspondent aux rangs des observations lorsque les données sont classées par ordre croissant de x et sont donc différents des numéros donnés par DAGNELIE [1981]. Les individus repris dans ce tableau ont également été identifiés sur le diagramme de dispersion.

Tableau 4. Numéros et valeurs des caractéristiques mesurant l'influence des observations pour les cinq individus extrêmes (les observations relatives au paramètre *COVRATIO* sont classées par ordre décroissant des valeurs $| COVRATIO_i - 1 |$).

Rangs des valeurs		1	2	3	4	5
e_i	Numéros	89	82	96	30	92
	Valeurs	−308,3	−168,1	167,3	150,2	138,9
t_i	Numéros	89	96	82	30	92
	Valeurs	−4,93	2,49	−2,48	2,18	2,03
D_i	Numéros	89	96	92	1	90
	Valeurs	0,30	0,14	0,07	0,06	0,06
DFFITS	Numéros	89	96	92	90	1
	Valeurs	−0,86	0,54	0,38	0,35	0,35
COVRATIO	Numéros	89	100	99	82	98
	Valeurs	0,67	1,14	1,13	0,92	1,08
DFBETAS (ordonnée à l'origine)	Numéros	89	96	1	98	92
	Valeurs	0,52	−0,38	0,34	−0,24	−0,24
DFBETAS (x)	Numéros	89	96	92	1	98
	Valeurs	−0,70	0,48	0,31	−0,30	0,29

On constate que l'observation 89 présente la valeur la plus extrême pour toutes les caractéristiques du tableau 4. Il s'agit d'une observation située manifestement en dehors du nuage de points (figure 1) et qui, du point de vue statistique, doit être considérée comme anormale, compte tenu de la valeur du résidu réduit ($t_{89} = -4{,}93$). De plus, cette observation a, par rapport aux autres observations, une influence relativement grande sur le calcul des valeurs estimées de y, sur la matrice des variances et covariances des paramètres et sur la valeur

des paramètres estimés. En effet, pour l'ensemble des données, l'équation de régression obtenue est la suivante:

$$y = 12,0 + 14,85x \quad (\widehat{\sigma}_{y.x} = 70,6),$$

les erreurs-standards des deux paramètres de la relation étant égales à 23,4 et 0,83. Après suppression de l'observation 89, on a:

$$y = 1,0 + 15,38x \quad (\widehat{\sigma}_{y.x} = 63,4),$$

avec, pour erreurs-standards des paramètres, les valeurs 21,2 et 0,75. La suppression de cette donnée entraîne donc une modification assez importante de la régression.

Les observations 96 et 82 présentent toutes deux des résidus standardisés fort proches en valeurs absolues mais ont une influence fort différente. En effet, l'observation 82 est peu influente, alors que l'observation 96 a une influence nettement plus grande sur les résultats de la régression, comme le montrent les caractéristiques D_i, *DFFITS* et *DFBETAS*. Le comportement différent de ces deux observations s'explique par les nombres de fruits correspondants. Pour l'observation 96, le nombre de fruits est de 43, alors que pour l'observation 82, il est de 34. Cette seconde valeur est beaucoup plus proche du nombre moyen de fruits, égal à 27,0, que la première. Ces distances à la moyenne se traduisent par les valeurs de h_{ii} qui sont respectivement égales à 0,045 et 0,017.

On constate également que l'observation 1 a une influence relativement importante, alors que le résidu associé à cette observation n'est pas du tout exceptionnel ($e_1 = 122,0$ et $t_1 = 1,78$). L'influence relative de cette observation est, ici aussi, essentiellement due à la faible valeur du nombre de fruits et donc à son éloignement de la moyenne (13 fruits et $h_{1,1} = 0,037$).

A l'opposé, l'observation 30 a un résidu plus grand ($e_{30} = 150,2$ et $t_{30} = 2,18$), sans que l'observation soit influente, car le nombre de fruits est proche de la moyenne (22 fruits et $h_{30,30} = 0,013$).

Cet exemple montre bien que les concepts de données influentes et de données anormales ne doivent pas être confondus, car une donnée influente n'est pas nécessairement anormale et une donnée anormale n'est pas nécessairement influente.

10. EXEMPLE 2: ÉTUDE DE LA COLINÉARITÉ

Le second exemple a pour but d'illustrer essentiellement les problèmes de colinéarité et la notion de graphique des variables ajoutées. Les données ont trait à l'accroissement en rayon des cinq dernières années déterminé par sondage à la tarière sur 4.784 épicéas. On souhaite mettre cet accroissement en relation avec la circonférence de l'arbre sondé et avec les principales caractéristiques du peuplement dont proviennent les arbres sondés [DAGNELIE *et al.*, 1988]. Pour

améliorer la qualité de la relation, toutes les données ont subi une transformation logarithmique.

Les différentes variables intervenant dans la régression et les symboles correspondants sont les suivants:

LA5DC: logarithme de l'accroissement en rayon;

LC: logarithme de la circonférence de l'arbre sondé;

LCMOY: logarithme de la circonférence moyenne des arbres du peuplement;

LI0: logarithme de l'indice de fertilité du peuplement (hauteur dominante du peuplement ramenée conventionnellement à l'âge de 50 ans);

LHDOM: logarithme de la hauteur dominante du peuplement;

LAGE: logarithme de l'âge du peuplement;

LNT: logarithme du nombre de tiges à l'hectare.

La figure 2 reprend les éléments essentiels du listing obtenu par le logiciel SAS. On constate que toutes les variables sont très hautement significatives, à l'exception de la variable *LHDOM*, qui est non significative et même nuisible, puisque la valeur t_{obs} relative au test de signification du coefficient de régression partielle est inférieure à l'unité. On constate également que les facteurs d'inflation de la variance sont relativement élevés, du moins pour les variables *HDOM* et *LCMOY*.

D'autre part, on remarque que les deux dernières valeurs propres de la matrice des variables explicatives sont du même ordre de grandeur et assez faibles par rapport à la première valeur propre. La racine carrée du rapport de la plus grande et de la plus petite valeur propre reste cependant inférieur à 30, valeur considérée comme maximum admissible (paragraphe 6). La décomposition de la variance des coefficients de régression montre que la variance du coefficient de *LC* est essentiellement liée à la $3^{ème}$ composante, la variance du coefficient de *LCMOY* à la $5^{ème}$ composante, la variance des coefficients de *LI0*, *LHDOM* et *LAGE* à la $6^{ème}$ composante et, enfin, la variance du coefficient de *LNT* à la $4^{ème}$ et à la $5^{ème}$ composantes.

La composante dont la variabilité est la plus faible est donc responsable de la plus grande part de la variance du coefficient de trois variables explicatives (*LI0*, *LHDOM* et *LAGE*), ces variables provoquant un phénomène de multicolinéarité approximative.

Ce résultat n'est, en fait, pas du tout surprenant, quand on sait que l'indice de fertilité des peuplements n'est pas une caractéristique observée sur le terrain, mais est calculé directement à partir de l'âge et de la hauteur dominante, à l'aide d'une relation non linéaire [DAGNELIE *et al.*, 1988]. On remarque cependant que le seul examen de la matrice de corrélation ne permet pas de détecter

DEP VARIABLE: LA5DC

ANALYSIS OF VARIANCE

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PROB>F
MODEL	6	778.06076175	129.67679362	529.432	0.0001
ERROR	4767	1167.60828	0.24493566		
C TOTAL	4773	1945.66904			

ROOT MSE	0.4949097	R-SQUARE	0.3999
DEP MEAN	2.133132	ADJ R-SQ	0.3991
C.V.	23.20108		

PARAMETER ESTIMATES

VARIABLE	DF	PARAMETER ESTIMATE	STANDARD ERROR	T FOR H0: PARAMETER=0	PROB > |T|	TOLERANCE	VARIANCE INFLATION
INTERCEP	1	-2.56943	0.58145896	-4.419	0.0001	.	0
LC	1	1.66183269	0.03307547	50.244	0.0001	0.24419857	4.09502810
LCMOY	1	-1.25986	0.10818935	-11.645	0.0001	0.02918064	34.26929231
LI0	1	0.65540522	0.16509366	3.970	0.0001	0.07853798	12.73269267
LHDOM	1	0.01359315	0.19487803	0.070	0.9444	0.02762873	36.19420935
LAGE	1	-0.490725	0.08825144	-5.561	0.0001	0.04904238	20.39052720
LNT	1	-0.111486	0.03966266	-2.811	0.0050	0.0602534	16.59657372

COLLINEARITY DIAGNOSTICS

NUMBER	EIGENVALUE	CONDITION NUMBER	VAR PROP LC	VAR PROP LCMOY	VAR PROP LI0	VAR PROP LHDOM	VAR PROP LAGE	VAR PROP LNT
1	4.523837	1.000000	0.0098	0.0014	0.0005	0.0012	0.0019	0.0027
2	1.127576	2.002998	0.0006	0.0001	0.0540	0.0017	0.0056	0.0012
3	0.215087	4.586130	0.9088	0.0025	0.0003	0.0082	0.0194	0.0088
4	0.0958309	6.870697	0.0256	0.0144	0.0003	0.0255	0.1741	0.3229
5	0.0213294	14.563455	0.0545	0.9791	0.0019	0.0659	0.0351	0.6015
6	0.0163392	16.639400	0.0007	0.0026	0.9431	0.8975	0.7639	0.0628

Figure 2. Régression de *LA5DC* en fonction de six variables explicatives.

cette colinéarité, car les corrélations simples entre ces trois variables ne sont pas excessives. On a, en effet:

corrélation entre *LI0* et *LHDOM:* 0,59,
corrélation entre *LI0* et *LAGE:* −0,04,
corrélation entre *LAGE* et *LHDOM:* 0,77.

Les variables *LCMOY* et *LNT* ont également toutes deux des coefficients dont les variances sont essentiellement liées à la $5^{ème}$ composante, elle aussi de faible variance. Le lien important qui existe entre ces deux variables se traduit par la valeur élevée du coefficient de corrélation:

corrélation entre *LCMOY* et *LNT:* −0,96.

Pour réduire la colinéarité, il y a donc intérêt à supprimer une variable au sein du premier groupe de trois variables et une variable au sein du deuxième

groupe de deux variables. Compte tenu des valeurs t_{obs}, on peut envisager la suppression des variables *LHDOM* et *LNT*. L'étude de toutes les équations à quatre variables confirme d'ailleurs ce choix.

Les résultats obtenus pour le modèle à quatre variables explicatives (figure 3) montrent que la suppression des deux variables n'a pratiquement pas diminué la valeur du coefficient de détermination multiple. D'autre part, les critères de colinéarité font apparaître une colinéarité approximative, mais moins marquée cependant que dans le modèle à six variables. Cette colinéarité est due aux relations entre les variables *LCMOY*, *LI0* et *LAGE*, et la suppression de la variable *LAGE* peut se justifier, car elle ne provoque, elle aussi, qu'une faible réduction du coefficient R^2, qui, pour le modèle à trois variables, serait égal à 0,391.

La suppression des variables *LHDOM* et *LNT* d'abord et *LAGE* ensuite a pour effet de réduire fortement les facteurs d'inflation et, par conséquent, les écarts-types des paramètres estimés pour les autres variables, comme le montre le tableau 5.

DEP VARIABLE: LA5DC

ANALYSIS OF VARIANCE

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PROB>F
MODEL	4	776.08424393	194.02106098	791.124	0.0001
ERROR	4769	1169.58480	0.24524739		
C TOTAL	4773	1945.66904			

ROOT MSE	0.4952246	R-SQUARE	0.3989
DEP MEAN	2.133132	ADJ R-SQ	0.3984
C.V.	23.21584		

PARAMETER ESTIMATES

VARIABLE	DF	PARAMETER ESTIMATE	STANDARD ERROR	T FOR H0: PARAMETER=0	PROB > |T|	TOLERANCE	VARIANCE INFLATION
INTERCEP	1	-3.30265	0.51961666	-6.356	0.0001	.	0
LC	1	1.66146308	0.03308432	50.219	0.0001	0.24437858	4.09201160
LCMOY	1	-1.03184	0.06923492	-14.903	0.0001	0.0713454	14.01632141
LI0	1	0.59274847	0.06933541	8.549	0.0001	0.44584375	2.24293824
LAGE	1	-0.501766	0.062639	-8.010	0.0001	0.09747158	10.25940099

COLLINEARITY DIAGNOSTICS

NUMBER	EIGENVALUE	CONDITION NUMBER	VAR PROP LC	VAR PROP LCMOY	VAR PROP LI0	VAR PROP LAGE
1	2.752598	1.000000	0.0283	0.0090	0.0049	0.0107
2	1.021621	1.641446	0.0001	0.0000	0.3881	0.0105
3	0.183815	3.869737	0.8862	0.0343	0.0725	0.1141
4	0.0419669	8.098745	0.0854	0.9567	0.5344	0.8647

Figure 3. Régression de *LA5DC* en fonction de quatre variables explicatives.

Tableau 5. Ecarts-types des paramètres des équations à six, quatre et trois variables explicatives.

Nombres de variables explicatives	6	4	3
Ordonnées à l'origine	0,581	0,520	0,364
LC	0,033	0,033	0,033
LCMOY	0,108	0,069	0,038
LI0	0,165	0,069	0,049
LHDOM	0,195	-	-
LAGE	0,088	0,063	-
LNT	0,040	-	-

Pour le modèle à trois variables explicatives, les graphiques de la variable ajoutée sont donnés pour chacune des variables (figures 4 à 6). Ces graphiques montrent clairement l'intérêt des variables *LC* et *LCMOY* dans le modèle. Pour le graphique relatif à *LI0*, la situation est moins nette, mais on devine néanmoins l'existence d'une corrélation positive. Pour ces trois graphiques, les coefficients de corrélation, qui sont en fait des coefficients de corrélation partielle, sont respectivement égaux à 0,41, -0,41 et 0,23.

On notera cependant que ces graphiques, obtenus de façon automatique, ne sont pas tout à fait adéquats. En effet, ils ne font pas clairement apparaître la densité des points dans la partie centrale du nuage de points, l'astérisque représentant deux observations ou plus de deux observations. Compte tenu de l'effectif élevé (n = 4.784), la proportion d'observations "cachées" est donc importante.

11. CONCLUSIONS

Au cours des dix dernières années, une série de nouveaux outils ont été créés pour la validation des modèles de régression et ces outils sont disponibles, du moins partiellement, dans les logiciels récents. Leur utilisation permet incontestablement de mettre en évidence des situations particulières (hypothèses de départ non vérifiées, présence d'individus influents ou anormaux).

Le nombre d'outils disponibles est cependant assez considérable et l'utilisateur occasionnel risque d'être dérouté devant le volume important des documents imprimés que peuvent lui fournir les logiciels modernes. Ces documents risquent, finalement, de compliquer l'interprétation des données au lieu de la simplifier.

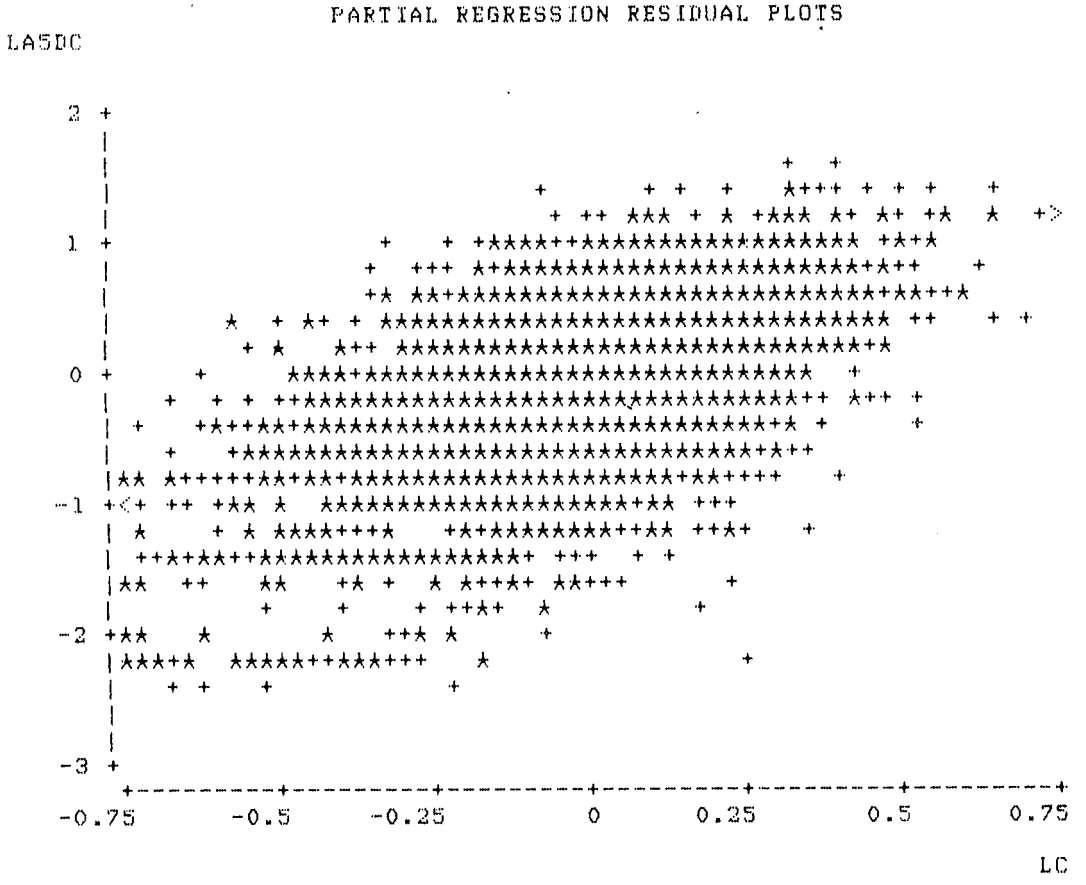

Figure 4. Diagramme des résidus partiels en fonction de LC.

La tâche de l'utilisateur n'est certainement pas facilitée par le manque d'uniformité dans les notations et dans les appellations: des concepts identiques peuvent avoir des dénominations différentes et des dénominations identiques peuvent recouvrir des concepts différents dans la littérature et dans les logiciels statistiques. Il est par conséquent toujours indispensable de vérifier la définition des concepts utilisés.

En pratique, l'utilisateur occasionnel a sans doute intérêt à se limiter à l'analyse d'un petit nombre de caractéristiques. Parmi celles-ci, on pourrait, par exemple, lui conseiller un examen assez systématique des distances de COOK, afin de repérer les observations les plus influentes. Pour ces observations, il analysera ensuite les résidus standardisés, r_i ou t_i, et les valeurs h_{ii}, afin de détecter pourquoi l'observation est influente. Pour la recherche de la colinéarité, l'examen des facteurs d'inflation de la variance (ou des tolérances) et l'analyse du tableau des contributions relatives des composantes principales à la variance des coefficients de régression partielle sont nécessaires.

Après avoir opté pour l'une ou l'autre mesure de l'influence, le problème qui se pose à l'utilisateur est, d'une part, de préciser à partir de quelle valeur de la mesure il doit considérer l'observation comme influente et, d'autre part, de définir l'attitude qu'il doit adopter en présence d'une ou de plusieurs données influentes.

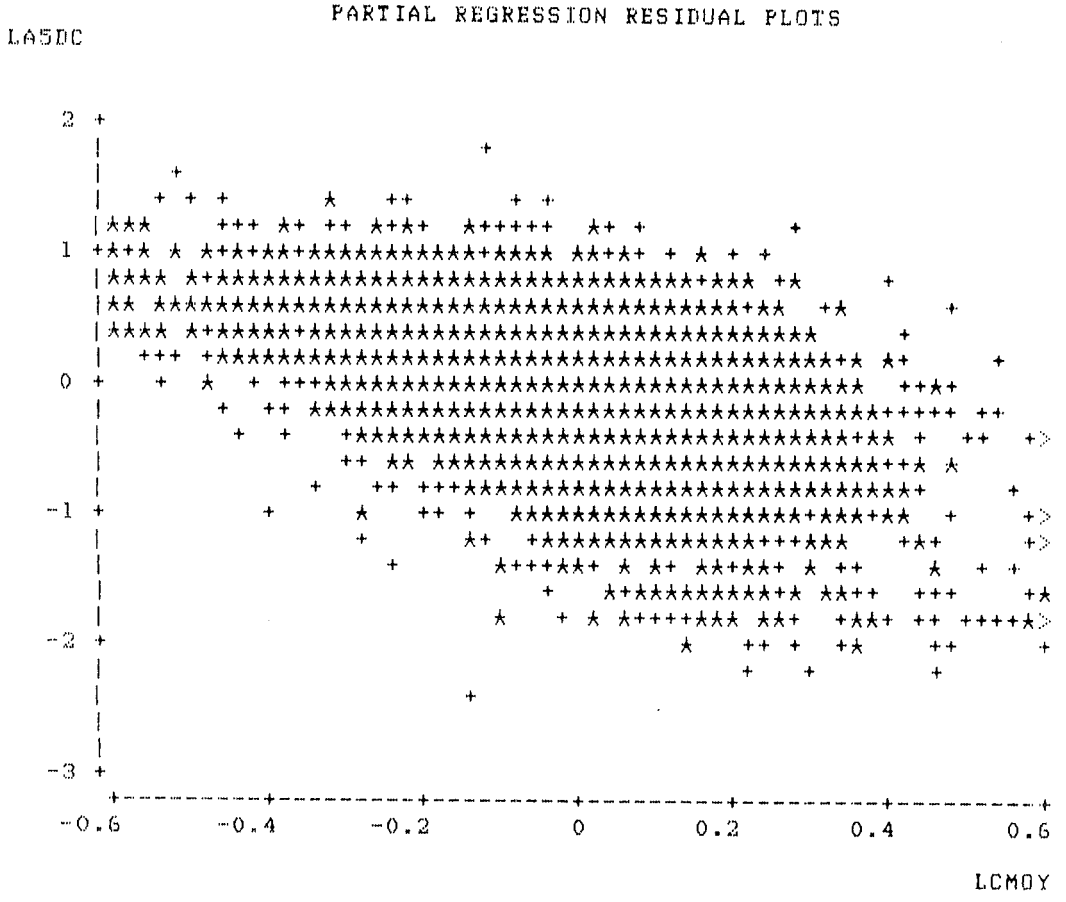

Figure 5. Diagramme des résidus partiels en fonction de *LCMOY*.

Des éléments de réponse à la première question sont donnés dans la littérature et ont été repris au paragraphe 5. Différents auteurs ont, en effet, proposé des valeurs limites pour chacune des mesures de l'influence, au-delà desquelles les données méritent une attention particulière. Les valeurs limites ne reposent cependant sur aucun fondement mathématique ou statistique, mais sont simplement données à titre indicatif.

Quant à l'attitude à adopter en présence de données influentes, aucune directive générale ne peut être proposée et la solution dépendra de la situation considérée. Ainsi, si la donnée est influente à cause de la valeur élevée du résidu, on pourra, par exemple, songer à supprimer cette donnée, dans la mesure où elle peut être considérée comme anormale. Au contraire, si la donnée est influente à cause de la valeur élevée de h_{ii}, deux attitudes opposées peuvent être envisagées. La première consiste à considérer que, dans l'espace des variables explicatives, la donnée est exceptionnelle et doit être supprimée, le modèle de régression étant alors valable pour un espace de variables explicatives plus réduit. La seconde solution consiste, au contraire, à considérer que cette observation est particulièrement intéressante et à diminuer son influence par la collecte d'autres données pour lesquelles le vecteur des variables explicatives est proche de celui de la donnée influente en question.

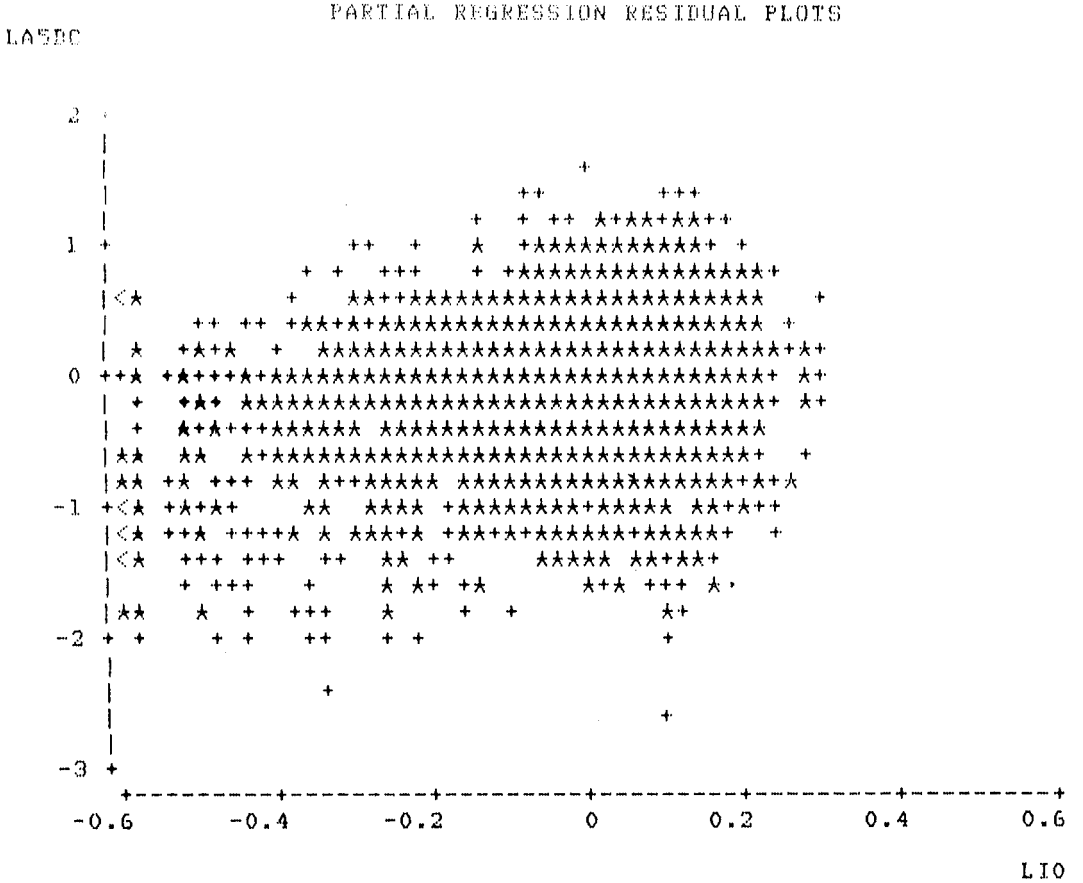

Figure 6. Diagramme des résidus partiels en fonction de *LI0*.

12. BIBLIOGRAPHIE

ATKINSON A.C. [1986]. *Plots, transformations and regression: an introduction to graphical methods of diagnostic regression analysis.* Oxford, Clarendon Press, 282 p.

BELSLEY D.A. [1984]. Demeaning conditioning diagnostics through centering. *Amer. Stat.* 38, 73-93.

BELSLEY D.A., KUH E. et WELSCH R.E. [1980]. *Regression diagnostics: identifying influential data and sources of collinearity.* New York, Wiley, 292 p.

CHATTERJEE S. et HADI A. [1986]. Influential observations, high leverage points and ouliers in linear regression. *Stat. Sci.* 1, 379-416.

COOK R.D. [1977]. Detection of influential observations in linear regression. *Technometrics* 19, 15-18.

COOK R.D. et WEISBERG S. [1982]. *Residuals and influence in regression.* New York, Chapman and Hall, 230 p.

CRYER J.D. [1986]. *Time series analysis.* Boston, Duxbury, 286 p.

DAGNELIE P. [1979-1980]. *Théorie et méthodes statistiques: applications agronomiques* (2 vol.). Gembloux, Presses Agronomiques, 378 + 463 p.

DAGNELIE P. [1981]. *Théorie et méthodes statistiques: exercices.* Gembloux, Presses Agronomiques, 186 p.

DAGNELIE P. [1982]. *Analyse statistique à plusieurs variables.* Gembloux, Presses Agronomiques, 362 p.

DAGNELIE P., PALM R., RONDEUX J. et THILL A. [1988]. *Tables de production relatives à l'épicéa commun* (Picea abies KARST.). Gembloux, Presses Agronomiques, 123 p.

DRAPER N.R. et SMITH H. [1981]. *Applied regression analysis.* New York, Wiley, 709 p.

FILLIBEN J.J. [1975]. The probability plot correlation coefficient test for normality. *Technometrics* 17, 111-117.

FREUND R.F. et LITTELL R.C. [1986]. *SAS system for regression.* Cary, SAS Institute, 164 p.

HOCKING R.R. [1983]. Developments in linear regression methodology: 1959 - 1982. *Technometrics* 25, 219-249.

KVÅLSETH T.O. [1985]. Cautionary note about R^2. *Amer. Stat.* 39, 279-285.

LARSEN W.A. et McCLEARY S.J. [1972]. The use of partial residual plots in regression analysis. *Technometrics* 14, 781-790.

OBENCHAIN R.L. [1977]. Letter to the editor. *Technometrics* 19, 348-351.

PALM R. [1986]. Etude des résidus de régression: principes et application. *Notes Stat. Inform.* (Gembloux) 86/1, 13 p.

STEWART G.W. [1987]. Collinearity and least squares regression. *Stat. Sci.* 2, 68-100.

THOMPSON M. [1978a]. Selection of variables in multiple regression: part I. A review and evaluation. *Int. Stat. Rev.* 46, 1-19.

THOMPSON M. [1978b]. Selection of variables in multiple regression: part II. Chosen procedures, computation and examples. *Int. Stat. Rev.* 46, 129-146.

TOMASSONE R., LESQUOY E. et MILLIER C. [1983]. *La régression: nouveaux regards sur une ancienne méthode statistique.* Paris, Masson, 180 p.

WEISBERG S. [1985]. *Applied linear regression.* New York, Wiley, 324 p.

X [1985a]. *Minitab reference manual.* PA State College, Minitab, 232 p.

X [1985b]. *SAS user's guide: statistics, version 5 edition.* Cary, SAS Institute, 956 p.