

STATISTIQUE ET ANALYSE DES DONNÉES

HENRI CAUSSINUS

Sur l'analyse des résidus dans le modèle linéaire

Statistique et analyse des données, tome 5, n° 3 (1980), p. 29-39

http://www.numdam.org/item?id=SAD_1980__5_3_29_0

© Association pour la statistique et ses utilisations, 1980, tous droits réservés.

L'accès aux archives de la revue « Statistique et analyse des données » implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
<http://www.numdam.org/>

Statistique et Analyse des Données

3 - 1980 pp. 29-39

SUR L'ANALYSE DES RESIDUS
DANS LE MODELE LINEAIRE

Henri CAUSSINUS

Laboratoire de Statistique et Probabilités

E.R.A. - C.N.R.S. n° 591

Université Paul Sabatier - 31062 TOULOUSE Cédex

Résumé : Dans le modèle linéaire gaussien, on modifie le vecteur des résidus par une perturbation aléatoire de sorte que les résidus ainsi modifiés soient "indépendants et de même loi normale centrée". Parmi les statistiques vectorielles ayant cette propriété, le vecteur proposé est, en un certain sens, le plus proche du vecteur des erreurs.

0 - INTRODUCTION.

L'objectif essentiel de cet article est d'introduire une nouvelle sorte de "résidus modifiés" et d'en donner quelques propriétés (§ 3).

Nous avons cru bon toutefois de faire précéder cette présentation d'un assez long paragraphe de généralités et d'historique (§ 2), et cela pour deux raisons. Il était d'abord utile de replacer nos propositions par rapport à la littérature existante pour en faire ressortir les caractéristiques, avantages et inconvénients. Ensuite, il ne semblait guère exister de texte synthétique sur l'ensemble des grands axes de recherche en analyse des résidus. Il ne s'agissait pas pour autant de faire une revue exhaustive de la question, et la bibliographie retenue est bien insuffisante pour y prétendre, mais plutôt de rapprocher des textes qui le sont rarement en nous contentant de ce qui nous paraissait essentiel et en citant plus volontiers les sources que certains développements. Une bibliographie plus abondante sera trouvée par exemple dans [4] et [6] .

Précisons enfin que, s'il est difficile de juger définitivement de l'intérêt pratique des nouveaux résidus modifiés (§ 4), leur présentation nous a paru utile au moins pour préciser certains aspects du modèle linéaire gaussien et plaider pour une formulation géométrique qui les rend plus sensibles.

1 - RAPPELS - NOTATIONS.

Nous considérons le modèle linéaire gaussien classique, avec erreurs indépendantes de même loi $N(0, \sigma^2)$, $\sigma > 0$. Une suite de n observations fournit les coordonnées d'un vecteur de \mathbb{R}^n muni de la métrique euclidienne canonique (\langle, \rangle désignera le produit scalaire et $\|\cdot\|$ la norme). Un changement de métrique dans \mathbb{R}^n permet de traiter de façon analogue tout cas où la matrice des variances et covariances des erreurs est régulière, connue à un facteur près.

On notera $N_E(\mu, A)$ la loi normale sur l'espace euclidien E de moyenne μ ($\mu \in E$) et opérateur de variance A (A est une forme quadratique positive sur E , assimilable à un endomorphisme symétrique positif de E). On désignera par \mathcal{B}_E la tribu des boréliens de E et par I_E l'opérateur identité de E . Pour $E = \mathbb{R}^n$, l'indice \mathbb{R}^n sera partout remplacé par n .

La structure statistique considérée est donc :

$$(1) \quad \mathbb{Q}^n, \mathcal{B}_n, \{N_n(\mu, \sigma^2 I_n) : \mu \in Q, \sigma > 0\}$$

où n est un entier supérieur à 2 et Q un sous-espace vectoriel donné de \mathbb{R}^n de dimension q ($0 < q < n$).

Pour se rapprocher des notations les plus usuelles, on peut supposer que la structure (1) est induite par la v.a. vectorielle Y et assimiler Y à la matrice ${}^t[Y_1, \dots, Y_n]$ où les v.a.r. Y_i sont les applications coordonnées sur (1). Matriciellement, le modèle s'écrit alors :

$$(2) \quad Y = X\beta + U$$

où X est une matrice connue de dimension $n \times q'$ ($q' \geq q$) dont les colonnes engendrent Q , de sorte que $\mu = X\beta$, et U est le vecteur des "erreurs" U_i , $i=1, \dots, n$, supposées indépendantes de même loi $N(0, \sigma^2)$.

π indicé par un sous-espace de \mathbb{R}^n désignera le projecteur orthogonal sur ce sous-espace.

Le meilleur estimateur sans biais de μ , au sens de l'erreur quadratique moyenne, est $\pi_Q(Y)$ (théorème de Gauss-Markov).

Le vecteur des résidus est $R = Y - \pi_Q(Y) = \pi_{Q^\perp}(Y)$. Ses coordonnées dans la base canonique de \mathbb{R}^n seront notées R_i , $i=1, \dots, n$.

Si L est une application linéaire de l'espace euclidien E dans l'espace euclidien F , l'image par L de la loi $N_E(\mu, A)$ est la loi $N_F(L(\mu), LA {}^tL)$. Donc, selon qu'on considère π_Q (resp. π_{Q^\perp}) comme application de \mathbb{R}^n dans lui-même ou de \mathbb{R}^n dans Q (resp. Q^\perp), on aura :

$$(3) \quad \pi_Q(Y) \text{ suit la loi } N_n(\mu, \sigma^2 \pi_Q) \text{ ou } N_Q(\mu, \sigma^2 I_Q)$$

$$(3') \quad \pi_{Q^\perp}(Y) \text{ suit la loi } N_n(0, \sigma^2 \pi_{Q^\perp}) \text{ ou } N_{Q^\perp}(0, \sigma^2 I_{Q^\perp}).$$

(π_Q, I_Q, \dots sont des opérateurs qu'il est inutile ici de représenter par une matrice).

2 - PROBLEMES DE L'ANALYSE DES RESIDUS.

2.1 - L'analyse des résidus

Les résidus R_i sont utilisés en pratique pour étudier la validité du modèle (1).

- Etude descriptive : il s'agit essentiellement de discuter la partie algébrique du modèle,

c'est-à-dire le choix de Q . De simples représentations graphiques sont déjà fort utiles. Cela est traité dans la plupart des ouvrages sur le modèle linéaire, plus particulièrement après l'impulsion donnée par [2].

- Etude inférentielle : la structure (1) est plongée dans une structure plus vaste (1') où (1) apparaît comme une hypothèse nulle. Par exemple, on cherche à tester la normalité des erreurs ou leur indépendance, ou encore à détecter des valeurs aberrantes ou toute autre mauvaise spécification du modèle.

Remarquons cependant que les résultats de l'analyse des résidus ne sauraient être déterminants pour juger les hypothèses du modèle (1) qui sont réellement nécessaires aux pratiques usuelles : test d'une hypothèse linéaire, estimation ensembliste de β . Les conditions de validité concernent en effet la distribution de $(\pi_Q(Y), \|\pi_{Q^\perp}(Y)\|)$ dont les liens avec celle de R dépendent de (1') et peuvent être très lâches. Par exemple, si (1') est la structure $(\mathbb{R}^n, \mathcal{B}_n, \{\mathcal{J}(\mu) : \mu \in Q\})$ où $\mathcal{J}(\mu)$ est la famille des lois isotropes autour de μ , le test de Fisher d'une hypothèse linéaire est toujours valide ([12]) quelle que soit l'issue d'une analyse de normalité basée sur les résidus (à moins que celle-ci ne soit équivalente à l'étude de l'isotropie : voir § 3, Remarque 2).

2.2 - Difficultés de l'étude inférentielle

Tous les problèmes de test sont complexes même pour le simple calcul du niveau (qui se fait sur la structure (1) seule envisagée désormais) car la loi conjointe des R_i n'est pas simple : ils ne sont ni indépendants ni de même variance, laquelle n'est connue qu'à un facteur près. En fait, à σ^2 près, la matrice des variances et covariances est la matrice de π_{Q^\perp} dans la base canonique de \mathbb{R}^n (cf. (3')) c'est-à-dire $I - X(X^t X)^{-1} X^t$; elle est singulière de rang $n - q$ et dépend de X [l'expression indiquée ne vaut que si X est de rang q mais il est toujours aisé de se ramener à ce cas puisque seul le sous-espace Q engendré par les colonnes de X est à considérer].

2.3 - Quelques solutions

On peut schématiquement les classer en deux catégories, selon que l'on affronte directement les difficultés inhérentes à l'usage des R_i ou que l'on modifie ces résidus pour en simplifier l'étude et surtout la rendre indépendante de X .

Pour les statistiques fonctions des R_i , on peut chercher des résultats analytiques exacts ou approchés (n grand) ou assez souvent utiliser une méthode de Monte-Carlo pour dresser des tables de valeurs critiques. Un exemple est celui du test de normalité pour un échantillon de taille n (voir [19]) : en fait on utilise là les résidus R_i dans un modèle linéaire simple où Q est le sous-espace de \mathbb{R}^n engendré par le vecteur $\mathbf{1}$ dont toutes les coordonnées dans la base canonique sont égales à 1.

Pour tester la présence d'une valeur aberrante dans le modèle de régression linéaire simple (Q engendré par $\mathbf{1}$ et un vecteur donné x de \mathbb{R}^n), TIETJEN, MOORE et BECKMAN [23] ont recours à la

simulation pour dresser les tables permettant d'utiliser le test fondé sur $T = \sup_i \frac{R_i}{s_i \|R\|}$

où $s_i^2 = \frac{\text{Var}(R_i)}{\sigma^2}$. Ces tables devraient dépendre de x , mais les auteurs montrent que cette dépendance est très faible et proposent une table "moyenne". Pour la détection de valeurs aberrantes dans le cas général, l'étude de la statistique T se heurte à des difficultés de tabulation : cependant les tables du t de Student permettent de bonnes approximations [13] qui peuvent même être exactes dans certains cas ([16], [17]) ou convenablement corrigées ([18]).

Un exemple où l'on trouve un encadrement des valeurs critiques valable quelle que soit la matrice X , est celui du test de DURBIN et WATSON [8] (dans le cas où la méthode simple conduit à une indécision, le traitement peut être poursuivi par approximation ou recours au calcul exact).

Pour la deuxième voie, on peut d'abord chercher à utiliser une transformation linéaire L afin qu'aux résidus R_i soient substituées $(n-q)$ v.a.r. non corrélées R'_i de même variance σ^2 et normales si le modèle (1) est bon. Il est clair que c'est toujours possible de multiples façons (voir par exemple [14]) mais les R'_i perdent le sens concret des R_i d'où des difficultés d'interprétation, surtout s'il s'agit d'étudier autocorrélation, hétérogénéité des variances ou valeurs aberrantes.

THEIL [20] semble le premier à avoir proposé un remède pertinent en déterminant, parmi les transformations L , celle qui conduit à des R'_i les plus proches respectivement de $(n-q)$ données des U_i (on peut les numéroter pour qu'il s'agisse des $n-q$ premiers) ; la proximité est au sens de $E(\|R^* - U^*\|^2)$ où R^* (resp. U^*) est le vecteur de \mathbb{R}^{n-q} de coordonnées R^*_i (resp. U_i), $i=1, \dots, n-q$, mais on a même une propriété plus forte [21]. Le vecteur R^* est appelé BLUS (Best Linear Unbiased with Scalar covariance matrix).

Le problème essentiel est d'avoir des résidus modifiés dont la distribution ne dépend pas de Q (ou X). C'est pourquoi l'on a essayé ensuite d'éviter de négliger q erreurs en recherchant un vecteur résidu modifié R' à valeurs dans \mathbb{R}^n , de matrice de covariance fixée et minimisant $E(\|R' - U\|^2)$. Ce sont les résidus BLUF (Best Linear Unbiased with Fixed covariance matrix) introduits par ABRAHAMSE et KOERTS [1]. On trouvera dans l'ouvrage récent de DUBBELMAN [6] une présentation détaillée ainsi qu'une discussion approfondie orientée pour l'essentiel vers les tests d'indépendance des erreurs au moyen des résidus BLUS et BLUF.

Dans un ordre d'idée voisin, on peut chercher à modifier les résidus (aussi peu que possible) en les perturbant par un procédé aléatoire adéquat pour obtenir n résidus ajustés non corrélés et de même variance. Le principe de cette démarche est dû à DURBIN [7] qui l'a détaillée pour le cas d'un échantillon mais a aussi indiqué des possibilités de généralisations. TIAO et GUTTMAN ont repris le principe dans un article essentiellement consacré aux valeurs aberrantes ([22], §4) en supposant σ connu. HILDRETH [10] suggère une solution pour σ inconnu qui aboutit aux résidus BAUS (Best Augmented Unbiased with Scalar covariance matrix) particulièrement étudiés par WARD [24].

Le vecteur des résidus BAUS peut s'écrire : $R'' = \pi_{Q^\perp}(Y) + \frac{\|\pi_{Q^\perp}(Y)\|}{\sqrt{n-q}} \pi_Q(V)$

où V est une v.a. de loi $N_n(0, I_n)$ indépendante de Y (ou au moins de $\pi_Q(Y)$ et $\|\pi_{Q^\perp}(Y)\|$). L'opérateur de variance est $\sigma^2 I$, assurant bien des résidus ajustés de même variance et non corrélés, mais la distribution de R'' n'est pas normale.

Nous allons voir maintenant comment poursuivre la démarche pour obtenir la normalité des résidus modifiés, la présentation géométrique adoptée rendant leur élaboration intuitive et leur étude très simple.

3 - NOUVEAUX RESIDUS MODIFIES.

3.1 - Construction des résidus modifiés

On s'appuyera sur le lemme suivant :

Lemme : Si Z est une v.a. à valeurs dans un espace euclidien E de dimension m , dont la norme est notée $\|\cdot\|$, la loi de probabilité de Z est $N_E(0, \sigma^2 I_E)$, $\sigma > 0$, si et seulement si l'on a à la fois :

$$(C) \begin{cases} Z \neq 0 \text{ presque sûrement} \\ \frac{1}{\|Z\|} Z \text{ est uniformément distribué sur la sphère unité de } E. \\ \frac{1}{\sigma^2} \|Z\|^2 \text{ a une loi de } \chi^2 \text{ à } m \text{ d.d.l.} \\ \frac{1}{\|Z\|} Z \text{ et } \|Z\| \text{ sont indépendantes en probabilité.} \end{cases}$$

Une démonstration peut être trouvée dans [9].

Considérons maintenant des v.a. T et w ayant les propriétés suivantes :

- . w est une v.a.r. positive et w^2 a une loi de χ^2 à $(n-q)$ d.d.l.
- . T est à valeurs dans \mathbb{R}^n et de loi $N_n(0, \pi_Q)$ [ou $N_Q(0, I_Q)$].
- . w, T et Y (ou au moins w, T et $\frac{1}{\|R\|} R$) sont indépendantes en probabilité.

Proposition 1 : Le vecteur aléatoire :

$$(4) \quad R_m = T + \frac{w}{\|R\|} R$$

a une loi $N_n(0, I_n)$; autrement dit ses coordonnées dans la base canonique de \mathbb{R}^n sont normales réduites indépendantes.

Démonstration : $R = \pi_{Q^\perp}(Y)$ est de loi $N_{Q^\perp}(0, \sigma^2 I_{Q^\perp})$, donc $\frac{1}{\|R\|} R$ est uniformément distribué sur la sphère unité de Q^\perp . Posant $Z = \frac{w}{\|R\|} R$, les conditions (C) du lemme sont aisées à vérifier avec $E = Q^\perp$ et $\sigma=1$. Donc Z a une loi $N_{Q^\perp}(0, I_{Q^\perp})$ ou $N_n(0, \pi_{Q^\perp})$ en plongeant Q^\perp

dans \mathbb{R}^n . D'autre part T est indépendante de Z et de loi $N_n(0, \pi_Q)$ d'où $T+Z$ est de loi $N_n(0, I_n)$ puisque $\pi_Q + \pi_{Q^\perp} = I_n$ \square

Remarques :

1. Pour obtenir w et T , il suffit de générer, indépendamment de Y , n v.a.r. indépendantes $N(0,1)$. En effet, soit V_i $i=1, \dots, n$ de telles v.a.r. et V le vecteur de \mathbb{R}^n les admettant pour coordonnées dans la base canonique, il est immédiat que $T = \pi_Q(V)$ et $w = \|\pi_{Q^\perp}(V)\|$ ont les propriétés requises.

Or, un programme de calcul pour un modèle linéaire quelconque (y compris par exemple les analyses de variance les plus complexes) fournit justement $\pi_Q(Y)$ et $\|\pi_{Q^\perp}(Y)\|$. Il suffira donc de le faire fonctionner sur les nouvelles données V_i pour obtenir T et w . Compte tenu de cette remarque, on écrira désormais (4) sous la forme :

$$(5) \quad R_m = \pi_Q(V) + \frac{\|\pi_{Q^\perp}(V)\|}{\|R\|} R = \pi_Q(V) + \frac{\|\pi_{Q^\perp}(V)\|}{\|\pi_{Q^\perp}(U)\|} \pi_{Q^\perp}(U)$$

où la loi de V est $N_n(0, I_n)$.

On notera que le nombre de v.a.r. à simuler est en général faible et que les critiques usuelles sur la génération de "nombres au hasard" valent surtout lorsque ceux-ci sont en grande quantité. Reste, bien sûr, le problème de fond concernant l'utilisation de ce type de méthode avec, en particulier, son manque de constance d'une étude à l'autre sur les mêmes données ... nous ne souhaitons pas en discuter ici. Les questions d'efficacité seront abordées plus loin.

2. Pour que la loi de R_m soit $N_n(0, I_n)$, il suffit que la probabilité de $\|\pi_{Q^\perp}(Y)\| = 0$ soit nulle et que $\frac{1}{\|\pi_{Q^\perp}(Y)\|} \pi_{Q^\perp}(Y)$ soit uniformément distribuée sur la sphère unité de Q^\perp , c'est-à-dire que la loi de $\pi_{Q^\perp}(Y)$ soit isotrope dans Q^\perp . C'est réalisé sous la condition suffisante que $Y-\mu$ soit isotrope dans \mathbb{R}^n , condition qui peut être considérée comme un cadre naturel élargi du modèle linéaire (PHILOCHE, [12]), et correspond à la structure (1') évoquée à la fin du paragraphe 2.1. .

3.2 - Quelques propriétés

Pour tenir compte de l'éventuelle utilisation de v.a. "simulées", on considère le produit de la structure (1) et d'un espace de probabilité sur lequel on peut définir des v.a. de loi quelconque indépendantes de Y . Sur cette structure produit, on désigne par \mathcal{J} la famille des statistiques à valeurs dans \mathbb{R}^n de loi $N_n(0, I_n)$.

Proposition 2 : a) $E(\|R_m - u\|^2) = \inf_{S \in \mathcal{J}} E(\|S - u\|^2) = n(\sigma^2 + 1) - 2\sigma\lambda_{n-q}^2$

b) $E(\|R_m - \frac{1}{\sigma} u\|^2) = \inf_{S \in \mathcal{J}} E(\|S - \frac{1}{\sigma} u\|^2) = 2(n - \lambda_{n-q}^2)$

avec $\lambda_{n-q} = \sqrt{2\pi} / \beta(\frac{n-q}{2}, \frac{1}{2})$.

Démonstration : Notons d'abord que la variance de S est arbitrairement fixée égale à I_n . Les statistiques comparées dans la partie b) sont donc plus "homogènes" qu'en a). Les démonstrations étant semblables nous montrerons seulement a). De plus, il est aisé d'obtenir des résultats analogues en remplaçant la variance de S par aI_n , $a > 0$.

On sait que, sur la structure (1), la statistique $(\pi_Q(Y), \|\pi_{Q^\perp}(Y)\|^2)$ est exhaustive et complète ; il est aisé de vérifier que la propriété subsiste sur la structure produit considérée. D'autre part S est libre (sa loi ne dépend pas des paramètres μ et σ). Donc S est indépendante en probabilité de $\pi_Q(U) = \mu + \pi_Q(U)$ et $\|\pi_{Q^\perp}(Y)\|^2 = \|\pi_{Q^\perp}(U)\|^2$ (cf. [3], p. 29).

*l : l
B...*

$$\text{On a } \|S-U\|^2 = \|S\|^2 + \|U\|^2 - 2 \langle S, U \rangle$$

$$\text{d'où } E(\|S-U\|^2) = n + n\sigma^2 - 2 E(\langle \pi_Q(S), \pi_Q(U) \rangle + \langle \pi_{Q^\perp}(S), \pi_{Q^\perp}(U) \rangle)$$

L'indépendance de $\pi_Q(S)$ et $\pi_Q(U)$ entraîne :

$$E(\langle \pi_Q(S), \pi_Q(U) \rangle) = \langle E(\pi_Q(S)), E(\pi_Q(U)) \rangle = 0$$

Puis, d'après l'inégalité de Cauchy-Schwarz et en utilisant l'indépendance de $\|\pi_{Q^\perp}(S)\|$ et $\|\pi_{Q^\perp}(U)\|$:

$$E(\langle \pi_{Q^\perp}(S), \pi_{Q^\perp}(U) \rangle) \leq E(\|\pi_{Q^\perp}(S)\|) \cdot E(\|\pi_{Q^\perp}(U)\|)$$

avec égalité si $\pi_{Q^\perp}(S) = K \pi_{Q^\perp}(U)$ (K v.a. réelle positive) ce qui est le cas pour $S = R_m$ (cf. (5)). La première égalité est donc montrée. De plus, $\|\pi_{Q^\perp}(S)\|^2$ a une loi de χ^2 à $(n-q)$ d.d.l. d'où $E(\|\pi_{Q^\perp}(S)\|) = \lambda_{n-q}$, et de même pour $E(\|\pi_{Q^\perp}(U)\|)$ au facteur σ près. \square

Proposition 3 : $E(\| \frac{1}{\|R_m\|} R_m - \frac{U}{\|U\|} \|^2) = \inf_{S \in \mathcal{F}} E(\| \frac{S}{\|S\|} - \frac{U}{\|U\|} \|^2) = 2(1 - \tau_{n,q}^2)$

$$\text{avec } \tau_{n,q} = \prod_{i=0}^{q-1} \frac{n-q+i}{n-q + \frac{1}{2} + i}$$

Démonstration : Le principe est le même que ci-dessus. On utilise ici l'indépendance de

$$\frac{1}{\|S\|} \pi_Q(S) \text{ et } \frac{1}{\|U\|} \pi_Q(U) = (\|\pi_Q(U)\|^2 + \|\pi_{Q^\perp}(U)\|^2)^{-1/2} \pi_Q(U) \text{ pour obtenir}$$

$$E(\langle \frac{1}{\|S\|} \pi_Q(S), \frac{1}{\|U\|} \pi_Q(U) \rangle) = 0$$

$$\text{On a ensuite } E(\langle \frac{1}{\|S\|} \pi_{Q^\perp}(S), \frac{1}{\|U\|} \pi_{Q^\perp}(U) \rangle) \leq E(\frac{\|\pi_{Q^\perp}(S)\|}{\|S\|}) E(\frac{\|\pi_{Q^\perp}(U)\|}{\|U\|})$$

avec égalité pour $\frac{1}{\|S\|} \pi_{Q^\perp}(S) = H \cdot \frac{1}{\|U\|} \pi_{Q^\perp}(U)$ (H v.a. réelle positive) ce qui est réalisé pour $S = R_m$.

Enfin les v.a.r. $\frac{\|\pi_{Q^\perp}(S)\|^2}{\|S\|^2}$ et $\frac{\|\pi_{Q^\perp}(U)\|^2}{\|U\|^2}$ sont chacune de loi $\beta(n-q, q)$ ce qui

fournit la valeur de $\tau_{n,q}$ \square

Remarques : 1. Cette proposition 3 nous semble plus intéressante que la proposition 2 car, en analyse des résidus, c'est avant tout la direction du vecteur résidu qui importe, abstraction faite de sa norme.

2. Puisque seulement $\frac{1}{\|S\|} S$ est en cause, la condition $S \in \mathcal{J}$ pourrait être remplacée par la condition plus faible "S isotrope" qui assure la même loi pour $\frac{1}{\|S\|} S$. Cependant, en pratique, on souhaite des coordonnées S_i indépendantes de même loi ce qui redonne $S \in \mathcal{J}$ à un facteur multiplicatif près pour la variance dès que $n \geq 3$ (LETAC [11]).

3. Il sera intéressant de comparer $E(\|\frac{1}{\|R_m\|} R_m - \frac{1}{\|U\|} U\|^2)$ à la quantité analogue obtenue pour d'autres types de résidus. Par exemple, pour les résidus ordinaires on trouve $E(\|\frac{1}{\|R\|} R - \frac{1}{\|U\|} U\|^2) = 2(1-\tau_{n,q})$, et pour les résidus BAUS on a

$$E(\|\frac{1}{\|R''\|} R'' - \frac{1}{\|U\|} U\|^2) = 2(1-\tau_{n,q} \cdot \tau_{n,q}^*) \text{ où } \tau_{n,q}^* \text{ est l'espérance mathématique de } \sqrt{\frac{n-q}{n-q+W}},$$

W étant une v.a.r. suivant une loi de χ^2 à q d.d.l. (voir [4] pour démonstrations et compléments). Evidemment, R_m donnant une approximation dans une classe plus restreinte que R ou R'' , il est normal qu'elle soit (légèrement) moins bonne.

4 - DISCUSSION.

Quel que soit le modèle linéaire considéré, les résidus modifiés sont relativement faciles à obtenir et de distribution très simple sous l'hypothèse nulle (validité du modèle (1)). Ils peuvent être utilisés en particulier pour un test de cette hypothèse dont le niveau sera d'un contrôle aisé. Reste à vérifier si ces avantages ne sont pas contrebalancés par une importante perte de puissance.

Une telle étude a été d'abord entreprise pour le problème de détection d'une valeur aberrante. Nous avons procédé par simulation sur le cas de la régression simple :

$Y_i = \beta_1 + \beta_2 i + U_i$, $i=1, \dots, n$. Les erreurs U_i sont des nombres pseudo-aléatoires de loi $N(0,1)$ pour $i \geq 2$ et $N(\lambda,1)$ pour $i=1$. Nous avons considéré les niveaux .10, .05 et .01, $\lambda = 2, 4$ ou 6 et $n = 10, 20, 30$. Trois tests ont été envisagés notés T_1, T_2, T_3 et fondés respectivement sur :

$$(1) \quad \sup_{i=1, \dots, n} \frac{|R_i|}{s_i \|R\|} \quad (\text{cf. } \S 2,3). \text{ Nous avons utilisé les tables de valeurs critiques}$$

approchées données en [23].

$$(2) \quad \sup_{i=1, \dots, n} |R_{mi}| \quad . \text{ Les valeurs critiques se calculent facilement à l'aide des}$$

tables de la loi normale.

$$(3) \quad \sup_{i=1, \dots, n} \frac{|R_{mi}|}{\|R_m\|} \quad . \text{ Les valeurs critiques se calculent à l'aide des tables de}$$

Student (valeurs exactes ou approchées selon n et le niveau).

Il semble inutile de reporter les résultats en détail, la conclusion s'imposant avec force : il ne faut pas utiliser les résidus modifiés dans ce problème. Le test T_1 est toujours bien meilleur que T_3 , lui-même meilleur que T_2 , les différences s'atténuant cependant quand n croît. La faible puissance de T_2 était prévisible, le vecteur R_m ayant une norme égale à $\|V\|$ (indépendante des données) qui intervient ici avec trop d'importance. Les performances de T_3 sont plus décevantes : sans doute la proximité de $\frac{R_m}{\|R_m\|}$ et $\frac{R}{\|R\|}$ précisée par la proposition 3 est-elle au sens d'une distance peu adaptée au problème envisagé ici.

Heureusement, les études sur un autre problème, celui de la corrélation des erreurs, sont beaucoup plus favorables. Elles ont été menées par M. BENHELLI et E. LAMBERT et les résultats détaillés en sont présentés ailleurs ([4], [5]).

Je remercie Mlle LAMBERT qui a réalisé les calculs cités au paragraphe 4 et M. BENHELLI qui a attiré mon attention sur plusieurs articles de la bibliographie.

BIBLIOGRAPHIE

- [1] ABRAHAMSE, A.P.J. and KOERTS J. (1971), *New Estimators of Disturbances in Regression Analysis*, J.A.S.A, 66, pp 71-74.
- [2] ANSCOMBE, F.J. and TUKEY, J.W. (1963), *The Examination and Analysis of Residuals*, *Technometrics*, 5, N° 2, pp. 141-160.
- [3] BARRA, J.R. (1971), *Notions Fondamentales de Statistique Mathématique*, Dunod, Paris.
- [4] BENHELLI, M. (1980), *Contribution à l'étude de l'autocorrélation des erreurs dans le modèle linéaire*, Thèse de 3ème cycle, Université Paul Sabatier, Toulouse.
- [5] BENHELLI, M. et LAMBERT, E. (1980), *Comparaison de quelques tests d'indépendance des erreurs dans le modèle linéaire*, en préparation.
- [6] DUBBELMAN, C. (1978), *Disturbances in the Linear Model, Estimation and hypothesis testing*, MartinusNijhoff Social Sciences Division, Leiden|Boston.
- [7] DURBIN, J. (1961), *Some Methods for Constructing exact tests*, *Biometrika*, 48, pp. 41-55.

- [8] DURBIN, J. and WATSON, G.S. (1950, 1951, 1971), *Testing for Serial Correlation in Least Squares Regression*, I, *Biometrika*, 37, pp. 409-428 ; II, *Biometrika*, 38, pp. 159-178 ; III, *Biometrika*, 58, pp. 1-19.
- [9] FIEGER, W. (1977), *Transformationen, die die Normalverteilung Charakterisieren*, *Metrika*, vol.24, pp.7-22.
- [10] HILDRETH, C. (1971), *A Note on Approximate Regression Disturbances*, Discussion Paper n° 2, Center for Economic Research, Department of Economics, University of Minnesota, Minneapolis, Minnesota, 11 pp.
- [11] LETAC, G. (1981), *Isotropy and sphericity: some characterisations of the normal distribution*, *Ann. Math. Stat.*, à paraître.
- [12] PHILOCHE, J.L. (1977), *Une condition de validité pour le test F*, *Statistique et Analyse des Données*, 1-1977, pp. 37-59.
- [13] PRESCOTT, P. (1975), *An approximate Test for Outliers in Linear Models*, *Technometrics*, 17, n° 1, pp. 129-132.
- [14] PUTTER, J. (1967), *Orthonormal Bases of Error Spaces and their Use for investigating the Normality and Variances of Residuals*, *J.A.S.A.*, 62, pp. 1022-1036.
- [15] RAMSEY, J.B. (1969), *Tests for specification errors in classical linear least-squares regression analysis*, *J.R.S.S.*, B, 31, N° 2, pp. 350-371.
- [16] SRIKANTAN, K.S. (1961), *Testing for the Single Outlier in a Regression Model*, *Sankhyā*, A, 23, pp. 251-260.
- [17] STEFANSKY, W. (1971), *Rejecting Outliers by Maximum Normed Residual*, *Ann. Math. Stat.*, 42, N° 1, pp. 35-45.
- [18] STEFANSKY, W. (1972), *Rejecting Outliers in Factorial Designs*, *Technometrics*, 14, N° 2, pp. 469-479.
- [19] STEPHENS, M.A. (1974), *EDF Statistics for Goodness of Fit and Some Comparisons*, *J.A.S.A.*, 69, N° 347, p. 730-737.

- [20] THEIL, H. (1965), *The Analysis of Disturbances in Regression Analysis*, J.A.S.A., 60, pp. 1067-79.
- [21] THEIL, H. (1968), *A Simplification of the BLUS Procedure for Analyzing Regression Disturbances*, J.A.S.A., 63, pp. 242-251.
- [22] TIAO, G.C. and GUTTMAN, I. (1967), *Analysis of Outliers with Adjusted Residuals*, Technometrics, 9, n° 4, pp. 541-559.
- [23] TIETJEN, G.L., MOORE, R.M. and BECKMAN, R.J. (1973), *Testing for a Single Outlier in Simple Linear Regression*, Technometrics, 15, N° 4, pp. 717-721.
- [24] WARD, L.L. (1973), *Is Uncorrelating the residuals worth it ?*, Thèse, Mc Gill University, Montréal.