REVUE DE STATISTIQUE APPLIQUÉE

A. MORINEAU

Régressions robustes méthodes d'ajustement et de validation

Revue de statistique appliquée, tome 26, n° 3 (1978), p. 5-28

<http://www.numdam.org/item?id=RSA_1978__26_3_5_0>

RÉGRESSIONS ROBUSTES
MÉTHODES D'AJUSTEMENT ET DE VALIDATION

A. MORINEAU

CEPREMAP, 142, rue du Chevaleret 75013, Paris

RESUME

On rappelle diverses alternatives à la technique des moindres-carrés pour effectuer des ajustements robustes dans un problème de régression. On peut détecter et éliminer les observations éloignées avant d'effectuer l'ajustement des moindres-carrés, ou utiliser des techniques robustes, en particulier dans la classe des M-estimateurs de HUBER. On pose ensuite le problème de la validation des ajustements effectués sur petits échantillons, et on envisage l'utilisation de l'estimateur de QUENOUILLE-TUKEY dont on rappelle le principe.

I – INTRODUCTION

Par analogie avec ce que l'on dit des ordinateurs, TUKEY (1977) qualifie les techniques statistiques robustes de "troisième génération" de la statistique, après la statistique classique (i.e. paramétrique, où la loi de LAPLACE-GAUSS joue un rôle important), et la statistique non-paramètrique.

La démarche de la statistique classique consiste à tirer des conclusions (par induction) en raisonnant sur des modèles qui sont supposés vrais. Elle feint le plus souvent d'ignorer que ces modèles ne sont jamais "exactement vrais" : les hypothèses sont souvent mieux justifiées par les commodités mathématiques que par les caractéristiques du phénomène étudié. En faisant de *l'estimation robuste,* le statisticien cherche à concentrer son intérêt sur le comportement de la masse des données ; il est à la recherche de méthodes qui soient suffisamment insensibles aux effets de quelques données aberrantes (permettant éventuellement de les détecter), et qui donnent des résultats stables pour un ensemble assez large de modèles qui auraient pu engendrer les observations. D'une façon générale il s'agit de limiter l'influence de toute particularité des données qui serait due à l'opération d'échantillonnage proprement dite.

Cette nouvelle tournure d'esprit, et le foisonnement des idées qui l'entourent, sont largement dus au développement des ordinateurs (comme dans le cas des *analyses de données*), et à l'emploi des méthodes de *simulation.* Simultanément à l'amélioration des outils mathématiques, la statistique semble devenir pour une part une sorte de science *expérimentale* où l'évidence empirique joue un rôle de plus en plus important. On en jugera par exemple en lisant l'ouvrage de ANDREWS *et al.* (1972) où sont comparés, essentiellement à partir de simulations, quelque 70 estimateurs robustes pour la valeur centrale d'une distribution (1).

(1) Fruit d'une "Année de la robustesse à Princeton" (1970/1971) qui a réuni ceux à qui la statistique robuste doit le plus HUBER, TUKEY, ANDREWS, HAMPEL, BICKEL etc.

Mots-clés : **Regression Robuste, Robustesse, Estimateur de QUENOUILLE-TUKEY**

Pour une présentation générale des problèmes et des techniques robustes, on consultera les articles de HUBER (1972) et HAMPEL (1973)([1]). Nous allons ici concentrer notre attention sur le problème de l'estimation d'un ensemble de coefficients à partir d'un ensemble d'observations, problème popularisé sous le titre de *régression,* abordé par les pionniers de la statistiques dès le XVIIIème siècle et demeurant la technique d'estimation et de prévision sans doute la plus utilisée.

Comme point de départ nous énoncerons le classique théorème de GAUSS-MARKOV qui a établi la suprématie du critère d'ajustement dit des *Moindres Carrés.* En rappelant ses conditions d'application, nous verrons apparaître la nécessité d'abandonner certaines d'entre elles si l'on veut s'approcher de situations réalistes.

II – LE THEOREME DE GAUSS-MARKOV

Il a régné pendant longtemps sur la destinée de la statistique un dogme puissant résumé dans la trinité :

(i) Méthode des Moindres Carrés
(ii) Moyenne et variance
(iii) Loi de LAPLACE-GAUSS

La méthode des Moindres-Carrés, suffisamment intuitive, a été utilisée semble-t-il bien avant qu'on cherche à la justifier d'un point de vue statistique. C'est seulement vers les années 1800 que GAUSS introduisit la loi qu'on appelle LAPLACE-GAUSS (dite aussi loi "normale"), comme étant la distribution pour laquelle l'estimateur des Moindres-Carrés, c'est-à-dire la moyenne arithmétique, est l'estimateur pratiquement incontesté de la valeur centrale d'une distribution.

La moyenne arithmétique, parce qu'elle donne le même poids à toutes les observations, est sensible aux extrémités des distributions. Ceci est particulièrement fâcheux car l'expérience courante montre que les distributions hypothétiques pouvant engendrer les données observées doivent avoir en général des extrémités plus épaisses que ne l'indique la loi de LAPLACE-GAUSS.

Cependant le théorème de GAUSS-MARKOV reste un solide outil statistique, et une référence dont on cherchera à s'éloigner le moins possible, tout en s'approchant le plus près possible de situations réalistes. Les conditions strictes sous lesquelles on peut dire que les Moindres-Carrés conduisent aux *meilleurs estimateurs* des coefficients α_k du modèle linéaire usuel :

$$y = \alpha_1 x_1 + \ldots + \alpha_p x_p + \epsilon$$

sont énumérées dans la liste en sept points ci-dessous :

1 – les x_k sont connus sans erreur ;

2 – la valeur centrale de y est $\Sigma \alpha_k x_k$ où les coefficients α_k sont totalement inconnus ;

3 – la variance de tous les y est σ^2 (paramètre inconnu ([2])) ;

4 – les covariances entre deux y sont nulles ([3]) ;

---

(1) Chacun comporte une bibliographie importante sur les travaux antérieurs à 1972. Pour une "préhistoire" de l'estimation robuste, voir STIGLER (1973).

(2) Ou encore $\text{var}(y_i) = w_i \sigma^2$, si w_i est un paramètre connu.

(3) Ou encore $\text{cov}(y_i, y_j) = w_{ij} \sigma^2$, si w_{ij} est un paramètre connu.

5 – on cherche des estimateurs des α_k linéaires en y ;

6 – ces estimateurs doivent être non biaisés ;

7 – la qualité des estimateurs est jugée sur leur variance.

Un complément classique du théorème de GAUSS-MARKOV dit que, si les résidus ont une distribution de LAPLACE-GAUSS, les meilleurs estimateurs doivent être *linéaires* en y. Autrement dit avec – et sans doute sans – l'hypothèse de stricte "normalité" des résidus, on ne pourra faire mieux que les Moindres-Carrés qu'avec des estimateurs biaisés et non linéaires en y.

Le papier de TUKEY (1975) évoque ce que devient l'ajustement lorsqu'on abandonne l'une ou l'autre des sept conditions. Le traitement robuste de la régression concerne également les situations où l'on ne peut invoquer l'hypothèse de "normalité", et où chacune des conditions assurant l'optimalité des Moindres-Carrés sera plus ou moins violée dans le contexte du phénomène étudié.

III – RECHERCHE ET REJET D'OBSERVATIONS ELOIGNEES

III.1 – Inspection des écarts des Moindres-Carrés

Le premier pas vers la régression robuste consiste à rechercher, dans le lot des données, celles qu'on peut considérer comme éloignées de l'ensemble des observations. Les ayant isolées, on effectuera l'ajustement définitif des *Moindres Carrés* sur les observations restant.

L'examen des écarts obtenus après ajustement des Moindres Carrés sur l'ensemble des observations est la première recommandation. Cependant, l'interprétation en est délicate comme l'illustre de façon caricaturale la *figure 1* où la seule observation assurément aberrante dans le lot, détermine l'ajustement, et apparaîtra avec un écart nul.

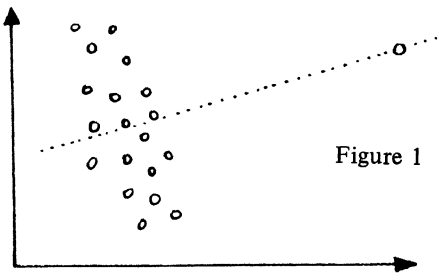


Figure 1

Les techniques non-robustes, comme les Moindres Carrés et les méthodes probabilistes du type Maximum de Vraisemblance, risquent ainsi de conduire à des écarts calculés anormalement petits, alors qu'il s'agit de s'ajuster à la "masse" des observations sans être entraîné abusivement par quelques données aberrantes ([5]).

Si l'on effectue l'ajustement sur un modèle hypothétique qui risque d'être inapproprié, la représentation graphique des écarts peut révéler le défaut sous réserve d'une grande habileté dans l'analyse (ANDREWS, 1971). On recommande

(5) Une véritable technique robuste doit résoudre le conflit entre la recherche d'un "petit" vecteur d'écarts, et le désir de ne pas rendre trop faible l'écart correspondant à une donnée aberrante non détectée.

l'examen des écarts *réduits*, c'est-à-dire l'écart divisé par son écart-type. De façon précise, définissons le modèle par :

[1] $$y = X\alpha + \epsilon$$

l'ajustement des Moindres Carrés sera

[2] $$y = Xa + e$$

où le vecteur des estimateurs des Moindres Carrés est calculé par

[3] $$a = (X'X)^{-1}\, X'y \quad \text{(X de plein rang)}$$

Le vecteur des écarts de l'ajustement est défini par :

$$e = y - Xa = y - X(X'X)^{-1}\, X'y = Qy$$

où Q est la matrice symétrique et idempotente (projection orthogonale) :

[4] $$Q = I - X(X'X)^{-1}\, X'$$

et la matrice des variances-covariances des écarts est :

[5] $$\hat{V}(e) = \sigma^2 Q$$

Sous les hypothèses classiques (ϵ suit une loi de LAPLACE-GAUSS d'espérance nulle, de matrice de variances-covariances $\sigma^2 I$), on obtient un estimateur sans biais par :

[6] $$\hat{V}(e) = s^2 Q$$

avec :

$$s^2 = (\Sigma e_i^2)/(n - p) \quad \text{(n observations, p variables exogènes)}$$

L'écart réduit est défini par :

[7] $$r_i = e_i / s\sqrt{q_{ii}}$$

où q_{ii} est le i-ème terme diagonal de Q, c'est-à-dire

[8] $$q_{ii} = 1 - h_{ii} \quad \text{avec } h_{ii} = x'(X'X)^{-1}\, x_i$$

x_i étant la i-ème ligne de X. La procédure consiste à suspecter les observations correspondant aux plus grands r_i. On trouve dans PRESCOTT (1975) une méthode de test approximatif pour le rejet d'une observation aberrante dans ce contexte classique, et LUND (1975) fournit des tables correspondant à ce test.

Après avoir détecté ces observations critiques, on s'interrogera naturellement sur les effets entraînés par leur suppression éventuelle. On trouve dans COOK (1977) la proposition d'évaluer l'effet de la i-ème observation sur l'ajustement par la statistique :

[9] $$d_i = (a_{(-i)} - a)'\, X'X(a_{(-i)} - a)/ps^2$$

où $a_{(-i)}$ est le vecteur des coefficients de l'ajustement effectué après suppression de la i-ème observation (6).

(6) On peut vérifier que d_i est aisément calculable par : $d_i = r^2 h_{ii}/p(1 - h_{ii})$.

Aux plus fortes valeurs de d_i correspondent les points ayant le plus de poids dans la détermination des coefficients. L'appréciation *simultanée* des r_i et des d_i peut guider par conséquent dans le choix des observations à extraire des données pour rendre l'ajustement des Moindres Carrés plus robuste.

Il est intéressant de regarder la matrice $H = I - Q$ qui effectue la projection orthogonale de y en $\tilde{y} = Xa$ sur le sous-espace engendré par les colonnes de X :

$$\tilde{y} = X(X'X)^{-1} X'y = Hy \text{ et } Var(\tilde{y}) = \sigma^2 H$$

Son terme h_{ij} évalue en quelque sorte l'influence de y_j sur $\tilde{y}_i$ et permet de détecter les points dans X tels que la valeur de y aura un fort impact sur l'ajustement. Plus facile à calculer, l'influence de y_i proprement dit sur l'ajustement apparaît dans le terme diagonal h_{ii} déjà introduit :

$$0 \leqslant h_{ii} \leqslant 1 \quad \text{et} \quad \Sigma h_{ii} = p$$

A la limite, si $h_{ii} = 1$ on a $\tilde{y}_i = y_i$ et si $h_{ii} = 0$ la valeur $\tilde{y}_i$ ne dépend ni de y_i ni des autres y_j. Les grandes valeurs de h_{ii} (comparées à leur moyenne p/n) permettent donc de détecter les lignes de X qui jouent un grand rôle dans l'ajustement. On notera que si X était un n-échantillon d'un vecteur de Laplace-Gauss dans R^p (modèle avec terme constant) alors la statistique

$$F = \frac{n - p}{p - 1} \frac{(h_{ii} - 1/n)}{(1 - h_{ii})}$$

suivrait une loi de FISHER-SNEDECOR à $(p - 1)$ et $(n - p)$ degrés de liberté.

A la suite de HOAGLIN *et al.* (1978) on peut s'interesser à l'écart entre y_i et sa *valeur prévue* en utilisant l'ajustement où la ligne i a été exclue :

$$e_{(-i)} = y_i - x_i' a_{(-i)}$$

On définira l'écart réduit :

$$r_{(-i)} = e_{(-i)}/s_{(-i)} \sqrt{1 - h_{ii}}$$

où $s^2_{(-i)}$ est l'estimation de σ^2 dans ce contexte ; on le calculera aisément par :

$$(n - p - 1)s^2_{(-i)} = (n - p)s^2 - e_i^2/(1 - h_{ii})$$

Sous les hypothèses classiques $r_{(-i)}$ suivrait une loi de Student à $(n - p - 1)$ degrés de liberté. Une stratégie possible consistera à étudier les couples $(h_{ii}, r_{(-i)})$: l'écart $r_{(-i)}$ peut être grand, mais une valeur faible de h_{ii} indiquera le peu d'importance de la ligne i dans l'ajustement ; inversement h_{ii} peut être grand alors que $r_{(-i)}$ reste faible parce que la valeur observée y_i est en bon accord avec la valeur prévue.

III.2 Autres méthodes pour détecter les points éloignés

Pour mémoire rappelons que des *transformations* usuelles (logarithme, racine carrée etc.) tentées sur les observations peuvent améliorer de façon notable l'aspect laplacien des distributions. Elles ne sont cependant autorisées que si le modèle n'est pas fixé définitivement (on prend la liberté de remplacer x ou y par exemple par log x et log y). Dans ce cas il est sans doute également permis d'extraire ou d'introduire de nouvelles variables dans le modèle, avec le codage le mieux approprié.

Dans tous les cas on pourra s'aider de *représentations graphiques* diverses, à la fois pour détecter les points éloignés, mettre en évidence une variance non-homogène des résidus, ou suggérer la transformation adéquate d'une variable.

Ces graphiques ne seront cependant jamais simples à interpréter compte-tenu de la complexité de la distribution des écarts, même dans le cas où les hypothèses classiques seraient satisfaites (formules [4] et [5]).Parmi les graphiques les plus utiles, citons :

a) graphique des écarts avec échelle de laplacienne (les écarts rangés, en fonction des quantiles de la distribution de LAPLACE-GAUSS) ;

b) graphique des écarts ([7]) en fonction des y ;

c) graphique portant en abscisse la variable x_k, et en ordonnée la quantité $(y - \sum_{j \neq k} a_j x_j)$; le nuage doit s'ajuster à une droite de pente a_k.

On pense naturellement aussi aux techniques *d'analyse en composantes principales* pour détecter des points éloignés dans le lot des données, mais leur utilisation ici n'est pas immédiate. Ces techniques vont en effet isoler des points "bizarres" quant à la structure intrinsèque des données (soit des observations éloignées sur les tous premiers facteurs, soit des observations déterminant à elles seules les tous derniers facteurs en dimensions supplémentaires du nuage). Mais le problème posé est d'une autre nature ; il s'agit de déterminer les points éloignés par rapport à une *structure imposée* au nuage (l'hyperplan engendré par les colonnes de X dans l'écriture $y = X\alpha + \epsilon$).

Il convient de rappeler cependant les liens étroits existant entre l'ajustement des Moindres Carrés et l'analyse en composantes principales, en particulier pour mettre en évidence le jeu des *petites valeurs propres.* Considérons les valeurs propres rangées de $X'X$:

$$\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_p \geqslant 0$$

et soit u_q le vecteur-propre normé correspondant à λ_q. Alors

$$(X'X)^{-1} = \Sigma \frac{1}{\lambda_q} u_q u'_q \quad (q = 1, 2, \ldots, p)$$

Par conséquent ([8])

$$a = \Sigma \frac{1}{\lambda_q} u_q u'_q X'y \qquad V(a) = \sigma^2 \Sigma \frac{1}{\lambda_q} u_q u'_q$$

On peut également exprimer les coefficients de l'ajustement des Moindres Carrés à l'aide des éléments propres de la matrice $(y ; X)'(y ; X)$. On trouvera les formules dans WEBSTER *et al.* (1974) ; et une comparaison de divers estimateurs de ce type dans GUNST *et al.* (1977).

---

(7) Prendre garde au fait suivant : si on ajustait une droite à ce nuage, cette droite aurait pour pente $(1 - R^2)$, où R^2 est la carré du coefficient de corrélation multiple.

(8) Cette décomposition est utilisée en particulier en cas de quasi-colinéarité dans X, détectée par : $(X u_q)' (X u_q) = \lambda_q \approx 0$ (donc $X u_q \approx 0$).

On stabilise et on précise l'estimation en restreignant la sommation aux valeurs propres assez différentes de 0. Noter que, pour résoudre le même problème, la régression bornée utilise la sommation complète sur les formules :

$$a = \Sigma \frac{1}{\lambda_q + k} u_q u'_q X'y \quad \text{et} \quad V(a) = \Sigma \frac{\lambda_q}{(\lambda_q + k)^2} u_q u'_q .$$

Il reste que le meilleur indicateur d'observation éloignée que l'on puisse espérer est l'écart entre la valeur observée et la valeur calculée après ajustement *robuste* – ce qui suppose malheureusement le problème résolu. Cependant on peut dans un premier temps mettre en jeu des procédures assez frustres, mais plus robustes que les Moindres Carrés, dans le but d'effectuer cette détection préliminaire. Le paragraphe suivant est consacré à la présentation sommaire de trois méthodes possibles. Dans l'esprit de ce paragraphe, il s'agit d'éliminer certaines observations pour effectuer ensuite l'ajustement classique des Moindres Carrés.

IV – TECHNIQUES RAPIDES D'AJUSTEMENT ROBUSTE

On sait que la médiane empirique, insensible aux extrémités des distributions, est un estimateur de valeur centrale plus robuste que la moyenne empirique. Les trois techniques suivantes peuvent être considérées comme des généralisation du calcul de la médiane.

IV.1. Minimisation de la somme des ecarts absolus

La minimisation de $\Sigma |e_i|$ avec $e_i = x_i - m$ définit la médiane empirique m des observations x_i. L'utilisation de ce critère, que l'on fait remonter à EDGEWORTH (aux alentours de 1890), s'étend au problème de la régression :

$$\min_{(a)} \Sigma | e_i | \quad \text{avec} \quad e = y - Xa$$

Comparé aux Moindres Carrés, le gain en robustesse apporté par le critère de EDGEWORTH est notable (le fait d'élever l'écart au carré fait jouer à l'observation éloignée un rôle prédominant dans la détermination des coefficients). La littérature sur le sujet est abondante. D'un point de vue numérique, le problème se ramène aisément à la résolution d'un programme linéaire :

$$\text{Minimiser } \Sigma (v_i + w_i)$$

sous les contraintes :

$$y_i = \sum_k a_k x_{ik} + v_i - w_i$$

$$v_i \geqslant 0 \; ; w_i \geqslant 0.$$

IV.2 La régression par les médianes (ANDREWS, 1974)

Supposons, pour simplifier la présentation, qu'il n'y ait qu'une variable exogène x. On range les x_i en ordre croissant. On élimine un certain nombre des plus petites et des plus grandes valeurs de x, ainsi qu'un certain nombre des valeurs les plus proches de la médiane (*figure 2*) :

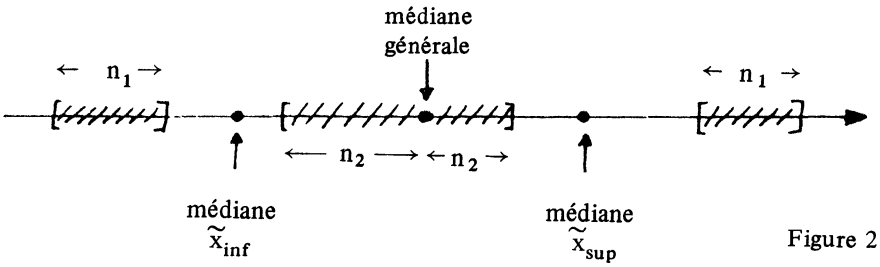


Figure 2

Soit $\tilde{x}_{inf}$ et $\tilde{x}_{sup}$ les médianes des points restants. On calcule les médianes $\tilde{y}_{inf}$ et $\tilde{y}_{sup}$ des y correspondants. On définit la pente de la droite, par :

$$a = (\tilde{y}_{sup} - \tilde{y}_{inf}) / (\tilde{x}_{sup} - \tilde{x}_{inf})$$

On peut continuer, de façon itérative, en appliquant la formule aux données modifiées par l'itération précédente ($x_i \leftarrow a\, x_i$) ; le coefficient retenu finalement est alors la somme des valeurs obtenues aux diverses itérations. Dans le cas d'une régression multiple la procédure est appliquée également aux variables exogènes afin d'éliminer la dépendance entre ces variables (voir ANDREWS, 1974).

IV.3 La méthode de HINICH-TALWAR (1975)

Dans y et X on sépare arbitrairement les observations en k parties (k de l'ordre n/p). Pour chacune d'elles on effectue l'ajustement des Moindres Carrés, et on retient comme vecteur préliminaire des coefficients la médiane des k résultats pour chaque coefficient.

On peut effectuer une seconde étape de la façon suivante. On définit une estimation robuste s pour l'écart-type σ des résidus par l'étendue à 44 % définie sur les écarts rangés $e_{(i)}$: soit e(0,28) et e(0,72) les quantiles empiriques de la distribution des écarts :

$$s = \frac{1}{1{,}654} [e\,(0{,}72) - e\,(0{,}28)]$$

On extrait alors des données, celles pour lesquelles l'écart est supérieur à 4s. On effectuera enfin l'ajustement des Moindres Carrés sur les observations restantes ([9]).

V – LES M-ESTIMATEURS DE HUBER

Un pas important vers l'estimation robuste d'une valeur centrale a été franchi avec l'introduction et l'étude par HUBER (1964) d'une catégorie d'estimateurs généralisant le Maximum de Vraisemblance, et appelés pour cette raison les *M-estimateurs.* Ces estimateurs sont construits pour fournir de bonnes estimations de la valeur centrale lorsque la fonction de répartition de la variable, pour que la distribution ait des extrémités plus épaisses que celles de la loi de LAPLACE-GAUSS, est de la forme ([10]) :

$$F(t) = (1 - \eta)\ \ \Phi(t) + \eta\, H(t) \qquad (11)$$

– Φ est la fonction de répartition de LAPLACE-GAUSS ;

– H est une fonction de répartition arbitraire (par exemple une distribution de LAPLACE-GAUSS ayant une plus grande variance) ;

– η est un nombre compris entre 0 et 1.

---

(9) On trouve dans HARVEY (1977) une comparaison des propriétés asymptotiques de ces trois techniques. Cependant il est également important d'en étudier le comportement sur les données de taille moyenne ou petite, et de comparer les difficultés et les coûts des calculs.

(10) On l'appellera la loi de LAPLACE-GAUSS *perturbée* ("contaminated normal").

On interprète cette construction en disant qu'une observation sera "convenable" avec la probabilité $(1 - \eta)$, mais pourra être "n'importe quoi" avec la probabilité (assez faible) η.

Toute estimation a pour les coefficients α du modèle $y = X\alpha + \epsilon$, définit un vecteur d'écarts :

$$\sum e = e(a) = y - Xa \qquad [12]$$

La plupart des procédures d'estimation peuvent être définies comme solutions d'une équation mettant en jeu les écarts. Ainsi l'estimation des Moindres Carrés satisfait :

$$(\Sigma e_i^2)_{MC} = \underset{(a)}{\text{Min}} \{ \sum e_i^2(a) \}$$

et il lui correspond le système d'équations du Maximum de Vraisemblance :

$$\sum_{i=1}^{n} x_{ik}\, e_i(a) = 0 \quad (k = 1, 2 \ldots, p)$$

D'une façon générale, les estimateurs du Maximum de Vraisemblance seront solutions d'équations de la forme :

$$\sum_{i=1}^{n} x_{ik}\, \psi\left(\frac{e_i(\hat{a})}{\hat{s}}\right) = 0 \quad (k = 1, 2, \ldots, p) \qquad [13]$$

où Ψ est une fonction à définir et $\hat{s}$ est un estimateur convenable de la dispersion des écarts afin que les estimateurs $\hat{a}$ soient invariant par changement d'échelle. Ce système d'équations fournit en effet la solution du problème :

$$[14] \qquad \sum_i \rho\left(\frac{e_i(\hat{a})}{\hat{s}}\right) = \underset{(a)}{\text{Max}} \sum_i \rho\left(\frac{e_i(a)}{s}\right)$$

avec :

$$[15] \qquad \psi(t) = \rho'(t) \quad \text{(fonction dérivée de } \psi \text{ continue)}$$

L'ajustement des Moindres Carrés en est clairement un exemple, avec $\rho(t) = \frac{1}{2}t^2$, et $\psi(t) = t$ (dans ce cas le facteur d'échelle s disparaît). Mais les estimateurs des Moindres Carrés ont un mauvais comportement lorsque les résidus ϵ_i du modèle sont engendrés par les distributions à extrémités épaisses comme on les définit dans [11].

Une part importante des recherches concernant la régression robuste consiste à définir des M-estimateurs (c'est-à-dire trouver des fonctions ρ, ou leurs dérivées ψ) qui soient efficaces en présence de distributions à extrémités épaisses, et restant proches de l'optimalité des Moindres Carrés en présence de résidus laplaciens. En fait le choix de ρ (ou de ψ) ne saurait être complètement arbitraire, et le bon sens permet d'imposer des conditions minimales, telles celles *d'admissibilité* énoncées par REY (1977).

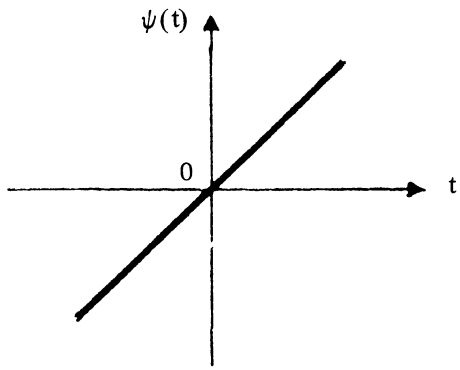
En outre les études répertoriées dans ANDREWS et al. (1972) suggèrent que le point important est le forme de la fonction $\psi(t)$ pour les grandes valeurs

de t : c'est là que se joue la pondération des grands écarts, correspondant à des observations exceptionnelles dans le corps des données. Il apparaît en particulier que la robustesse de l'estimateur est inexistante si $\psi(t)$ n'est pas bornée, elle est meilleure si $\psi(t)$ est bornée, et ses qualités augmentent encore si $\psi(t)$ revient vers 0 quand t devient grand, Ces remarques générales nous guideront dans l'exploration du catalogue des principaux M-estimateurs, le premier étant, pour mémoire, l'estimateur des Moindres-Carrés.

1) Estimateur des Moindres Carrés

[16] $$\psi(t) = t$$

La fonction $\psi(t)$ n'est pas bornée, et l'estimateur des Moindres Carrés est notoirement non-robuste.

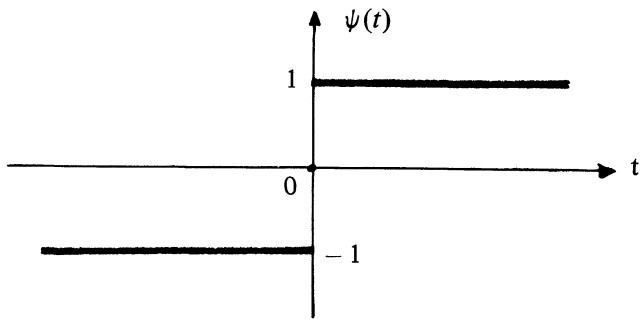


2) Minimisation de la somme des écarts absolus

On cherche les coefficients qui minimisent la somme des valeurs absolues des écarts. La procédure est réputée plus robuste que la technique des Moindres Carrés. Il lui correspond le M-estimateur défini par :

$$\rho(t) = t$$

[17] $$\psi(t) = \text{sign}(t)$$



Remarque : On peut considérer les M-estimateurs définis en 1. et 2. comme cas particuliers du M-estimateur caractérisé par :

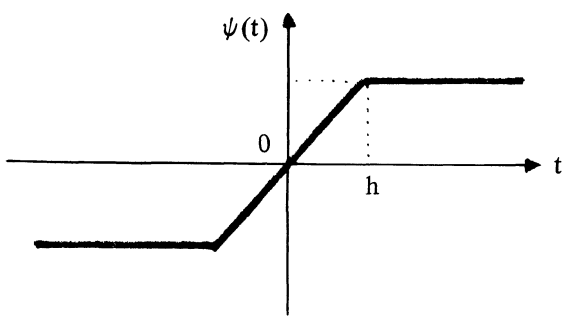
$$\rho(t) = \frac{1}{p}|t|^p$$

[18] $$\psi(t) = |t|^{p-1} \qquad \text{avec } 1 \leqslant p \leqslant 2$$

Il lui correspond le problème de minimisation du vecteur des écarts au sens de la norme L_p. La robustesse de cet estimateur a été évoquée en particulier par FORSYTHE (1972) et EKBLOM (1974). On effectuera les calculs à l'aide de l'algorithme décrit par exemple dans SPOSITO *et al.* (1977). On calculera l'ajustement en norme L_1 à l'aide, par exemple, des programmes linéaires listés dans BARRODALE *et al.* (1974) ou NARULA *et al.* (1977). L'article de SIELKEN et HARTLEY (1973) souligne les précautions à prendre pour obtenir une solution qui soit l'estimation *non biaisée* des coefficients.

3) L'estimateur de HUBER (1964, 1973)

[19] $$\psi(t) = \begin{cases} t & \text{si } |t| < h \\ h\,\text{sign}(t) & \text{si } |t| \geqslant h \end{cases}$$



La valeur h choisie pour définir ψ dépend de façon complexe des observations. D'autre part la méthode nécessite l'estimation simultanée de la valeur convenable à donner à la dispersion s des écarts e_i. La recherche d'une solution pour [19] s'effectuera de façon itérative. Nous évoquerons plus loin les problèmes de résolution numérique.

Dans son article de 1973, HUBER étudie les propriétés asymptotiques de cet estimateur, ainsi que son comportement (par simulation) sur les petits échantillons ([11]).

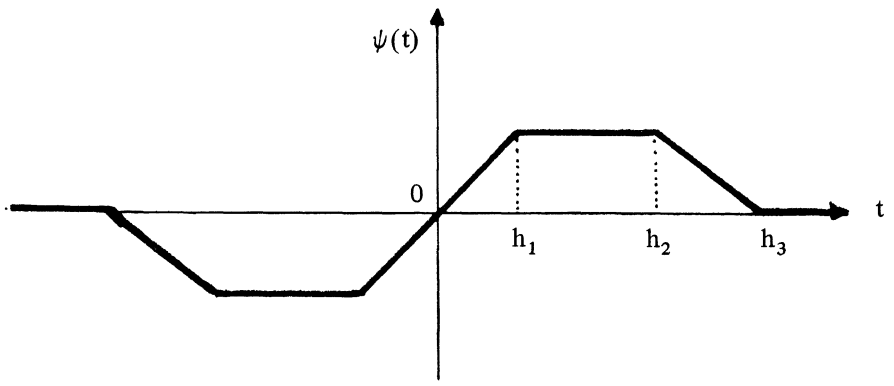
4) Estimateur de HAMPEL

Dans l'ouvrage de 1972 (ANDREWS *et al.,)* HAMPEL propose pour l'estimation d'une valeur centrale le M-estimateur défini par une fonction $\psi(t)$ symétri-

(11) Pour h fixé il s'agirait du Maximum de vraisemblance pour une distribution hypothétique de la forme $f_h(t) = \text{Cste. exp } \{-\rho_h(t)\}$

que, linéaire de 0 à h_1; puis constante jusqu'à h_2, et enfin décroissant vers 0 jusqu'á h_3.

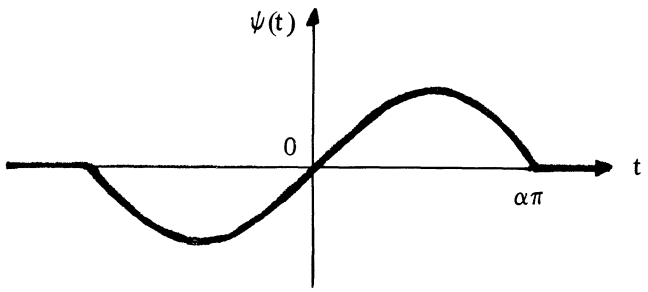
[20]
$$\psi(t) = \begin{cases} \cdot\ t \quad \text{si } |t| \leqslant h_1 \\ \cdot\ h_1 \operatorname{sign}(t) \quad \text{si } h_1 < |t| \leqslant h_2 \\ \cdot\ \dfrac{h_3 - |t|}{h_3 - h_2} h_1 \operatorname{sign}(t) \quad \text{si } h_2 < t \leqslant h_3 \\ \cdot\ 0 \quad \text{si } |t| > h_3 \end{cases}$$



Une telle fonction protège mieux l'ajustement contre des observations très éloignées. Les procédures de calcul seront analogues à celles concernant l'estimateur de HUBER.

5) Estimateur de ANDREWS

[21]
$$\psi(t) = \begin{cases} \alpha \sin\left(\dfrac{t}{\alpha}\right) \quad \text{si } |t| < \alpha\pi \\ \\ 0 \quad \text{si } |t| \geqslant \alpha\pi \end{cases}$$



Le M-estimateur proposé par ANDREWS (1974) a une forme proche de l'estimateur de HAMPEL. Les procédures de calcul seront de même nature.

Remarques sur les calculs

Il ne suffit pas de définir de nouveaux estimateurs. Il faut pour chacun d'eux étudier leurs propriétés asymptotiques, estimer leur matrice de covariances, étudier leur comportement sur les petits échantillons (en général par simulation), et enfin proposer des méthodes de calcul. C'est ce programme qui est suivi en particulier par HUBER (1973) à propos du M-estimateur caractérisé par [19].

Il faut ensuite comparer les qualités et les performances de chaque estimateur, ou de chaque famille d'estimateurs. Un travail de cette nature a été commencé par RAMSAY (1977), qui étudie les estimateurs en norme L_p, l'estimateur de ANDREWS, et un nouvel estimateur qu'il propose ; et par REY (1977) qui compare sur un exemple les estimateurs [18], [19] et [21].

Le problème numérique d'optimisation correspondant à chacun de ces esimateurs est assez complexe. Il faut tout d'abord définir, en principe en fonction des données, les meilleures valeurs à donner aux paramètres impliqués dans la fonction ψ (les auteurs proposent parfois des valeurs "passe-partout"). La résolution proprement-dite du système [13] s'effectuera en général par une *procédure itérative* où, à chaque étape, on devra estimer simultanément les coefficients $\hat{a}$ et le facteur d'échelle $\hat{s}$ (en général implicite dans les paramètres de la fonction ψ).

On a déjà proposé plus haut une estimation robuste pour s. A chaque étape on peut choisir plus simplement :

$$s = \text{médiane } |e_i(a)|$$

ou encore une moyenne tronquée :

$$s = \frac{1}{n-4} \sum_{2}^{n-2} |e_{(i)}(a)| \quad \text{où les } e_{(i)} \text{ sont rangés.}$$

On pourra par exemple estimer les nouveaux coefficients à chaque itération par ajustement des *moindres-carrés pondérés* ; ainsi dans le cas de l'estimateur [21] on utiliserait les pondérations :

$$\begin{cases} p_i = [1 - \cos\left(\frac{e_i}{\tilde{s}}\right)] / e_i^2 & \text{si } |\frac{e_i}{s}| < \alpha\,\pi \\ p_i = 1/e_i^2 & \text{si } |\frac{e_i}{s}| \geqslant \alpha\,\pi \end{cases}$$

Il est important de démarrer le processus itératif avec une solution déjà suffisamment robuste pour éviter de se déplacer vers un optimum local qui ne serait pas la solution robuste cherchée. En particulier à la solution des Moindres-Carrés on préfèrera une des trois techniques rapides présentées au paragraphe IV.

On trouvera une présentation détaillée de plusieurs algorithmes possibles pour résoudre les problèmes numériques dans DUTTER (1977). Le papier de MARTIN et al. (1975) propose un algorithme, bien adapté aux calculs sur les gros ensemble de données, opérant par *approximation stochastique* (ROBBINS-MONRO) et conduisant à un estimateur asymptotiquement équivalent à l'estimateur de HUBER.

Le papier de DENBY et MALLOWS (1977) propose des *représentations graphiques* permettant d'étudier le comportement des coefficients estimés et des écarts en fonction du paramètre h utilisé dans la fonction ψ définissant le M-

estimateur de HUBER, lorsque l'on fait varier de façon continue ce paramètre. Les graphiques permettent de voir si les estimations sont sensibles aux changements de poids affectés aux grands écarts, ce qui permet de détecter la présence et l'influence d'observations aberrantes dans le jeu des données.

VI – LES L-ESTIMATEURS et les R-ESTIMATEURS

Nous classons ici deux familles d'estimateurs robustes, les *L-estimateurs* construits par combinaison linéaire de statistiques d'ordre, et les *R-estimateurs* dérivés des tests de rang.

Evoquons sur un exemple simple le principe des *L-estimateurs.* Soit $e = y - Xa$ le vecteur des écarts d'un ajustement arbitraire ; soit $e_{(i)}$ le i-ème écart rangé. Dans l'ajustement des Moindres-Carrés on minimise :

$$\Sigma\ e^2 = \Sigma\ e_{(i)}\, e_{(i)}$$

On définit une fonction de *pondération* des écarts, le poids p(i) étant fonction du rang i de l'écart, et on minimise la somme pondérée des écarts rangés :

$$\text{Min } \{\sum p(i)\, e_{(i)}\} \qquad (\text{avec } \Sigma\, p(i) = 0)$$

Le système de pondération fonction des rangs permet de réduire l'effet des écarts aberrants, et donne de l'importance aux écarts "moyens". Une étude théorique de ces estimateurs est faite dans JAECKEL (1972). On trouvera dans HETTMANSPERGER *et al.* (1977) l'étude et la mise en œuvre d'une procédure correspondant au système de poids dit de WILCOXON :

$$p(i) = \sqrt{12}\left(\frac{i}{n+1} - \frac{1}{2}\right)$$

qui est un système monotone et antisymétrique :

$$p\,(1) \leqslant \ldots \leqslant p\,(n) \quad \text{et} \quad p\,(i) = -p\,(n - i + 1)$$

Les auteurs donnent des indications sur les algorithmes utilisables, et présentent quelques procédures de test sur les coefficients du modèle.

Présentons enfin le principe des *R-estimateurs* (dérivés des tests de rang) dans le cas de la régression simple :

$$y_i = \alpha x_i + \beta + \epsilon_i$$

Tout test de l'hypothèse $\alpha = 0$ fournit une estimation pour α de la façon suivante : chercher la première valeur a telle que le même test appliqué aux donnés transformées $(y_i - ax_i)$ conduise au rejet de l'hypothèse. Cette valeur peut être prise pour estimation de α, et la méthode permet de déterminer la variance asymptotique de l'estimation. On trouvera une étude théorique de cette classe d'estimateurs dans BICKEL et LEHMANN (1975), mais leur utilisation pour le problème d'estimation dans le modèle linéaire ne semble guère avancé. Une application de ce principe d'estimation est décrite cependant dans HOGG *et al.* (1975).

Remarque sur les procédures non-paramétriques

Il est opportun de noter ici que les problèmes de *robustesse* sont assez éloignés des problèmes résolus par les procédures *non paramétriques.* Certes les procédures non-paramétriques ont de bonnes qualités de robustesse pour résoudre le problème qui est le leur : l'estimation ou le test concernant une quantité ou une fonction bien précise (par exemple l'espérance d'une distribution, sa médiane théorique ou la fonction de répartition). Mais il est rare qu'on sache quelle quantité précisément estimer (penser au problème de l'ajustement d'un modèle linéaire). Ainsi ce que l'on entend ici par procédure robuste déborde largement le domaine classique des techniques non-paramétriques.

VII – VALIDATION D'UN AJUSTEMENT

Le problème de validation se pose dès que l'on sort du cadre classique dans lequel les hypothèses techniques *a priori* permettent d'effectuer des tests et de construire des intervalles de confiance. Il est donc crucial dans le cas d'un ajustement robuste, où toute technique conduit à un ajustement pour lequel on ne connaît à la rigueur que les propriétés asymptotiques, alors qu'en général les calculs sont effectués sur des échantillons de petite taille.

Une méthode bien connue pour étudier l'adéquation d'un ajustement consiste à retenir une partie des données pour estimer les coefficients, et utiliser le reste pour apprécier la précision des prévisions qu'ils fournissent. SNEE (1977) par exemple étudie un algorithme pour affecter astucieusement les données à l'une ou l'autre utilisation. Ce type de technique ne fournit cependant aucune indication sur la précision avec laquelle est déterminé chaque coefficient de l'ajustement.

Pour avoir une mesure de cette précision à partir des données elles-mêmes, on peut envisager de répartir les observations en sous-groupes sur chacun desquels on calcule les coefficients ; alors les variations d'un groupe à l'autre permettent d'évaluer la "variabilité" de chaque coefficient. Cependant cette procédure n'est viable qu'avec un nombre assez important d'observations. Nous allons rappeler ci-dessous le principe d'un succédané de cette méthode, puis étudier sur utilisation dans le cas des ajustements robustes.

La technique introduire par QUENOUILLE (1949 et 1956) et TUKEY (1958) joue un double rôle : réduire ou supprimer le *biais* d'un estimateur biaisé et fournir une estimation approchée mais robuste d'un *intervalle de confiance* autour de toute estimation. Elle constitue un instrument statistique de remplacement dans les cas où un calcul exact serait compliqué ou impossible. TUKEY lui a donné le nom de "JACKKNIFE" ([12]) –le couteau à usages multiples que le boy-scout emploie en toutes circonstances. On aura une idée du développement de la méthode en consultant les 45 références d'études ou d'applications parues avant 1974 et citées par MILLER (1974).

---

(12) Nous dirons estimateur Q.T pour QUENOUILLE-TUKEY.

VIII – DEFINITION ET UTILISATION DE L'ESTIMATEUR Q.T.

Soit $(X_1, X_2, \ldots, X_n)$ un échantillon aléatoire de la variable parente X. Imaginons que l'on ait à estimer un paramètre μ pour lequel on utilise l'estimateur :

$$Y_{tot} = f(X_1, X_2, \ldots, X_n)$$

L'échantillon est divisé arbitrairement en k parties disjointes, chacune contenant m = k/n individus. On appelle Y_{-j} le résultat du calcul effectué sur l'échantillon amputé de la j-ème partie (c'est-à-dire constitué des k-1 parties autres que la j-ème) :

$$Y_{-j} = f(\text{les } X_i \text{ sauf le groupe j})$$

A la suite de TUKEY, on appelle *pseudo-valeurs* les k différences pondérées :

$$\overset{*}{Y}_j = k\,Y_{tot} - (k-1)\,Y_{-j} \quad (j = 1, 2, \ldots, k)$$

L'estimateur Q.T est la moyenne des pseudo-valeurs :

$$\overset{*}{Y} = \frac{1}{k}\Sigma\,\overset{*}{Y}_j = k\,Y_{tot} - (k-1)\left\{\frac{1}{k}\Sigma\,Y_{-j}\right\} \qquad [22]$$

La variance Q.T est un estimateur de la variance de $\overset{*}{Y}$ (et de Y_{tot}) :

$$\overset{*}{S}^2 = \frac{1}{k}\left\{\frac{\Sigma(\overset{*}{Y}_j - \overset{*}{Y})^2}{k-1}\right\} = \frac{1}{k(k-1)}\left\{\Sigma\,\overset{*}{Y}_j^2 - \frac{1}{k}(\Sigma\,\overset{*}{Y}_j)^2\right\} \qquad [23]$$

On trouvera en *annexe* un programme FORTRAN illustrant le principe de calcul d'une estimation Q.T et de son écart-type approché.

NB : S'il s'agit d'estimer l'espérance mathématique $\mu = E(X)$, on a $Y_{tot} = f(X_i) \fallingdotseq \overline{X}$. Si de plus k = n (donc m = 1), il vient $\overset{*}{Y}_j = X_j$ et $\overset{*}{Y} = \overline{X}$. Alors la variance Q.T s'écrit :

$$\overset{*}{S}^2 = \frac{1}{n}\,\frac{\Sigma(X_i - \overline{X})^2}{n-1}$$

Enfin si l'échantillon est laplacien $(\overset{*}{Y} - \mu)/\overset{*}{S}$ suit exactement une loi de Student à n − 1 degrés de liberté.

1) Réduction du biais

Si Y_{tot} est un estimateur du paramètre μ ayant un biais d'ordre $\frac{1}{n}$:

$$E(Y_{tot}) = \mu + \frac{a}{n} + 0\left(\frac{1}{n^2}\right)$$

alors ce biais est éliminé pour Y. La démonstration est immédiate :

$$Y = k\,Y_{tot} - (k-1)\left(\frac{\Sigma\,Y_{-j}}{k}\right)$$

avec $$E(Y_{tot}) = \mu + \frac{a}{km} + \frac{b}{(km)^2} + \ldots \quad (km = n)$$

$$E(Y_{-j}) = \mu + \frac{a}{m(k-1)} + \frac{b}{(m(k-1))^2} + \ldots$$

donc $$E(Y) = k\left(\mu + \frac{a}{km} + \frac{b}{(km)^2} + \ldots\right) - (k-1)\left(\mu + \frac{a}{m(k-1)} + \frac{b}{(m(k-1))^2} + \ldots\right)$$

$$E(Y) = \mu - \frac{b}{m^2 k(k-1)} + \ldots = \mu + 0\left(\frac{1}{n^2}\right).$$

2) Construction d'un intervalle de confiance

En général les k pseudo-valeurs peuvent être considérées comme k variables aléatoires identiques, *approximativement* indépendantes. Dans ce sens, TUKEY conjecture que la statistique

$$T = \overset{*}{Y} / \overset{*}{S} \qquad [24]$$

suit approximativement une *loi de STUDENT* à $k - 1$ degrés de liberté. On peut alors construire un *intervalle de confiance approché* autour de l'estimation Y, quelle que soit la distribution mise en cause dans l'échantillon (intervalle "non paramétrique").

La conjecture de TUKEY s'est avérée valide dans de nombreux contextes. On peut la mettre en évidence *empiriquement* dans des problèmes précis par des procédures de *simulation* (exemple DUCAN-1973). On peut d'autre part démontrer les *théorèmes de convergence.* Ces théorèmes ne sont pas simples (exemples MILLER – 1964 ; MILLER – 1968 ; ARVESEN – 1969 ; MILLER – 1974 ; etc. . .).

Les démonstrations procèdent par développement en série de $\overset{*}{Y}$ en fonction des variables de base de l'échantillon $X_1, X_2, \ldots, X_n$. On montre que seuls les termes du premier degré sont non-négligeables, et qu'ils impliquent le comportement attendu. Plus intuitivement, on imputera la généralité du résultat à une certaine *robustesse* de la statistique de Student vis à vis de la non-indépendance des échantillons.

En fait la précision avec laquelle $\overset{*}{S^2}$ estime $\mathrm{Var}\,(\overset{*}{Y})$ dépend de la précision avec laquelle Y_{tot} peut être approché par une expression linéaire ([13]) de la forme (HINKLEY, 1977) :

$$Y_{tot} \approx \frac{1}{n} \sum_i c(X_i)$$

(13) La procédure QT s'appliquera mal par exemple à l'estimation de la médiane, ou à celle des valeurs extrêmes.

Si l'on pose $\mu = E(c(X_i))$ et $\sigma^2 = Var(c(X_i))$, il est facile de voir que, lorsqu'il y a *égalité* dans l'expression ci-dessus :

$$E(\overset{*}{Y}) = \mu \quad \text{et} \quad Var(\overset{*}{Y}) = \sigma^2/n$$

De plus :

$$E(\overset{*}{S^2}) = \sigma^2/n \quad \text{et} \quad Var(\overset{*}{S^2}) = \frac{2\sigma^4}{n^2(n-1)}\left(1 + \frac{\gamma_2}{2m}\right)$$

où γ_2 est le moment cumulant d'ordre 4 de X. S'il était nul, la distribution de $\overset{*}{S}_2$ aurait les deux premiers moments d'une distribution de CHI-2 à $(k-1)$ degrés de liberté.

L'approximation par une distribution de Student sera donc d'autant meilleure que γ_2 est petit, et que m est grand (d'où l'intérêt de supprimer m individus à la fois, sans cependant en supprimer trop pour ne pas anihiler l'efficacité de la méthode). Cette discussion, qui repose sur une égalité dans l'expression de Y_{tot}, continuera à s'appliquer si l'approximation est étroite, et la distribution de Y_{tot} proche d'une loi de LAPLACE-GAUSS.

Par ailleurs le nombre de degrés de liberté devrait être, à une unité près, le nombre de pseudo-valeurs *distinctes* que peut donner le calcul (ne pas considérer comme égales des valeurs identiques après arrondi des calculs, ou à cause de valeurs particulières des observations ; cf. MOSTELLER et TUKEY (1968)).

3) Précautions d'utilisation

On évitera les distributions d'échantillonnage ayant une extrémité abrupte, ainsi que les distributions fortement dissymétriques, ou ayant des extrémités très dispersées. On cherchera des transformations qui "polissent" le comportement de Y. On évitera également le cas où le paramètre à estimer doit appartenir à un intervalle fixé ou a une demi-droite : pour une probabilité p employer la procédure sur $\log \frac{p}{1-p}$; pour une variance σ^2, employer la procédure sur $\log \sigma^2$; pour une corrélation ρ, employer la procédure sur $\tanh^{-1} \rho$.

Remarque sur les calculs

Les pseudo-valeurs calculées par la formule [22] sont particulièrement sensibles aux *erreurs d'arrondi* (les coefficients k et $k-1$ peuvent être grands par rapport aux valeurs Y_{tot} et Y_{-j}. De plus les pseudo-valeurs sont calculées par différence entre des valeurs qui peuvent être proches). Sur ordinateur il sera prudent d'effectuer certains calculs en double précision.

IX – APPLICATION A LA REGRESSION

IX.1 Ajustement des Moindres-Carrés

On considère le modèle linéaire $y = X\alpha + \epsilon$ où les ϵ_i ont des distributions identiques, indépendantes, mais non nécessairement laplaciennes. MILLER (1974) a démontré deux théorèmes asymptotiques importants. Tout d'abord l'estimateur

Q.T pour les coefficients α, calculé par les Moindres Carrés en éliminant une observation à la fois, a une distribution asymptotique de LAPLACE-GAUSS. De plus la variance des pseudo-valeurs est une estimation convergeant vers la variance de l'estimateur Q.T sous des conditions de régularité assez naturelles.

Ce type de théorème s'étend d'ailleurs aux estimateurs usuels de la variance commune des résidus σ^2 et du coefficient de corrélation multiple pour lesquels la procédure Q.T permet par conséquent de construire des intervalles de confiance robustes.

En ce qui concerne la *robustesse* de l'ajustement, les pseudo-valeurs constituent naturellement des indicateurs privilégiés de valeurs aberrantes ([14]) ; on pourra en particulier les porter sur un graphique à échelle laplacienne. S'il existe des observations "déviantes", il est recommandé de les extraire des données et de recommencer l'estimation Q.T sur les données nettoyées, car l'estimateur Q.T s'avère assez vulnérable sur ce point ([15]) : on constaterait un grand écart entre l'estimation usuelle a et l'estimation Q.T $\overset{*}{a}$ et une variance Q.T anormalement grande.

L'étude du comportement de l'estimateur Q.T sur des données de taille moyenne ou petite est complexe, et profitera sans doute de recherches préliminaires par simulation. Notons à ce propos les principales formules utiles (notations déjà introduites) :

$$a = (X'X)^{-1} X'y$$

(estimation des Moindres Carrés)

$$a_{(-1)} = a - (X'X)^{-1} x_i e_i/(1 - h_{ii})$$

$$\overset{*}{a}_i = n\,a - (n-1)\,a_{(-i)} \qquad \text{(pseudo-valeurs)}$$

$$= a + (n-1)(X'X)^{-1} x_i e_i / (1 - h_{ii})$$

D'où l'estimation Q.T :

$$\overset{*}{a} = \frac{1}{n} \Sigma \overset{*}{a}_i = a + \frac{n-1}{n} (X'X)^{-1} \Sigma x_i e_i / (1 - h_{ii}) \qquad [25]$$

L'espérance des résidus ϵ du modèle étant nulle, celle des écarts des M.C. $e = y - X$ a l'est aussi, et par conséquent a et $\overset{*}{a}$ sont des estimations sans biais des coefficients α. La variance de $\overset{*}{a}$, supérieure à celle de a (GAUSS-MARKOV), est facile à calculer :

$$V(\overset{*}{a}) = \sigma^2 (X'X)^{-1} \{I + \left(\frac{n-1}{n}\right)^2 [D_2 - D_1 (X'X)^{-1} D_1] (X'X)^{-1}\}$$

expression dans laquelle :

$$D_k = \Sigma \{x_i x_i' / (1 - h_{ii})^k\} \quad \text{avec } k = 1, 2$$

---

(14) Voir les formules concernant $a_{(-i)}$

(15) Phénomène mis en évidence aussi à propos de l'estimation Q.T du coefficient de corrélation par HINKLEY (1978).

C'est cette variance $V(\overset{*}{a})$ qui, dans la procédure Q.T, doit être approchée par la variance $\overset{*}{S}^2$ de la moyenne des pseudo-valeurs. En général $\overset{*}{S}^2$ sera un estimateur biaisé de $V(\overset{*}{a})$ (HINKLEY, 1977) :

$$\overset{*}{S}^2 = \frac{1}{n(n-1)} (\overset{*}{a}_i - \overset{*}{a})(\overset{*}{a}_i - \overset{*}{a})' \qquad [26]$$

On remarquera que les calculs après abandon d'une ou plusieurs lignes d'observations ne nécessitent pas la ré-inversion d'une matrice, mais découlent simplement du calcul effectué sur l'ensemble des données.

IX.2. Ajustements robustes

Dans une première étape, on peut s'attacher à définir une matrice des variances des coefficients des Moindres Carrés qui soit robuste vis-à-vis d'une variance non-homogène des résidus ϵ_i du modèle. On construit par exemple (HINKLEY, 1977) des *pseudo-valeurs pondérées,* le poids étant relatif à l'importance de l'observation dans la détermination des coefficients :

$$\tilde{a}_i = a + n(1 - h_{ii})(a - a_{(-i)}) = a + n(X'X)^{-1} x_i e_i \qquad [27]$$

L'estimateur $\tilde{a}$ coïncide alors avec l'estimateur des Moindres Carrés :

$$\tilde{a} = \frac{1}{n} \Sigma \tilde{a}_i = a$$

On trouvera une estimation *robuste* des variances des coefficients dans la diagonale de la matrice :

$$\tilde{S}^2 = \frac{1}{n(n-p)} \Sigma (\tilde{a}_i - a)(\tilde{a}_i - a)' = \frac{1}{n(n-p)} (X'X)^{-1} (\Sigma e_i^2 x_i x_i') (X'X)^{-1}$$

Une autre voie de recherche consiste à appliquer la procédure Q.T usuelle à une technique d'ajustement robuste, par exemple dans la famille des M-estimateurs décrits plus haut. Des problèmes importants apparaissent alors simultanément dans le coût des calculs, et dans la précision des algorithmes. Il s'avère nécessaire (et économique) d'éliminer pour chaque pseudo-valeur plusieurs observations à la fois ; la procédure permet de stabiliser les pseudo-coefficients et conduit à des intervalles de confiance qui recouvrent la vraie valeur avec la fréquence annoncée par le seuil de confiance (on effectue les simulations avec des lois de Laplace-Gauss perturbées [11]).

ANNEXE

SOUS-PROGRAMME FORTRAN POUR LE CALCUL D'UNE ESTIMATION Q.T ET DE SON ECART-TYPE

Le sous-programme ESTQT illustre le principe du calcul d'une estimation Q.T dans le cas simple d'un paramètre construit sur un seul échantillon (cas d'une variance empirique par exemple). Ce paramètre est calculé par le sous-programme ESTIM qui doit être fourni par l'utilisateur.

Arguments en entrée

X (*) contient les valeurs de l'échantillon ;
N est la taille de l'échantillon (dimension de X (*) et de T (*)) :

NG est le nombre de groupes demandés pour le découpage de l'échantillon, donc NG ⩽ N (dimension de PSDV (*)).

Arguments en sortie

NGAL est le nombre effectif de groupes, c'est-à-dire le diviseur de N le plus proche inférieurement de NG ;
XTOT est la valeur de la statistique calculée sur l'échantillon complet ;
XQT est l'estimation Q.T (ie. la moyenne des pseudo-valeurs) ;
SQT est l'écart-type estimée de XQT (ou de XTOT) ;
PSDV(*) est le vecteur des NGCAL pseudo-valeurs ;
T(*) est un vecteur de travail de dimension N.

Remarques

Les calculs de moyenne et d'écart-type se font par accumulation afin de réduire les erreurs d'arrondi ; cependant il conviendra d'effectuer les opérations en double précision dans certaines applications. L'algorithme permet la suppression du stockage des pseudo-valeurs si celles-ci ne sont pas utilisées dans la suite.

```
      SUBROUTINE ESTQT ( X, N, NG, NGCAL, XTOT, XQT, SQT, PSDV, T )
C * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
C  ESTIMATION DE QUENOUILLE-TUKEY (XQT) ET DE SON ECART-TYPE (SQT),
C  DEMANDANT  NG  GROUPES (NG.LE.N) DANS L-ECHANTILLON X(N) .
C         NGCAL= NOMBRE REEL DE GROUPES, XTOT= ESTIMATION SUR X(N),
C         PSDV(NGCAL) = PSEUDO-VALEURS, T(N) = VECTEUR DE TRAVAIL.
C  APPEL DU SOUS-PROGRAMME  ESTIM(NZ, Z, ESTZ)  POUR CALCULER UNE
C         ESTIMATION  ESTZ  SUR UN ECHANTILLON  Z(NZ) .
C (XQT/SQT SUIT EN GENERAL UNE LOI PROCHE DE STUDENT A (NGCAL-1) DDL)
C * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
      DIMENSION  X(N), T(N), PSDV(NG)
C........ ESTIMATION  XTOT  SUR L-ECHANTILLON COMPLET  X(N)
      CALL ESTIM ( N, X, XTOT )
C........ BOUCLE  40  SUR LES CALCULS  (K = 1, NGCAL)
      NPAS  = N / NG
      NGCAL = N / NPAS
      NPSD  = N - NPAS
      SQT   = 0.0
        DO 40  K = 1, NGCAL
      II    = 1
C........ CONSTRUCTION DES DONNEES SANS LE GROUPE  K
        DO 20  I = 1, N, NPAS
        IF ( (I-NPAS-1)/NPAS) .EQ. K )  GO TO 20
      IFIN  = I - NPAS - 1
        DO 10  J = I, IFIN
      T(II) = X(I)
   10 II    = II + 1
   20   CONTINUE
C........ ESTIMATION  XK  SUR L-ECHANTILLON SANS LE GROUPE K
      CALL ESTIM ( NPSD, T, XK )
C........ CALCUL DES NGCAL PSEUDO-VALEURS
      PK    = (NGCAL*XTOT) - (NGCAL-1)*XK
      PSDV(K)= PK
C........ ACTUALISATION DE LA MOYENNE ET DE L'ECART-TYPE
        IF (K .NE. 1)                    GO TO  30
      XQT   = PK
   30 DK    = PK - XQT
      DFK   = DK / FLOAT(K)
      XQT   = XQT + DFK
      SQT   = SQT + DK*DK - DK*DFK
   40   CONTINUE
      SQT   = SQRT ( SQT / FLOAT(NGCAL-1) / FLOAT(NGCAL) )
        RETURN
        END
```

BIBLIOGRAPHIE

[1] ANDREWS D.F. (1971). – Significance Tests Based on Residuals – *Biometrika,* vol. 58, pp. 139-148.

[2] ANDREWS D.F. (1974). – A Robust Method for Multiple Linear Regression. *Technometrics,* vol. 16, pp. 523-31.

[3] ANDREWS D.F., BICKEL P.J., HAMPEL F.R., HUBER P.J., ROGERS W.H., TUKEY J.W. (1972) – Robust Estimates of Location, Survey and Advances. Princeton University Press.

[4] ARVESEN J.N. (1969). – Jackknifing U-Statistics. *Ann. Math. Statist.,* vol. 40, pp. 2076-100

[5] BARRODALE I., ROBERTS F.D.K. (1974) : Solution of an Overdetermined System of Equations in the L_1 Norm (Algorithm 478). *CACM,* vol. 17, pp. 319-20.

[6] BICKEL P.J., LEHMANN E.L. (1975). – Descriptive Statistics for Nonparametric Models (II Location). *Ann. Statist.,* vol. 3, pp. 1045-69.

[7] COOK R.D. (1977) : Detection of influential Observation in Linear Regression. *Technometrics,* vol. 19, pp. 15-8.

[8] DENBY L., MALLOWS C.L. (1977) : Two Diagnostic Displays for Robust Regression Analysis. *Technometrics,* vol. 19, pp. 1-13.

[9] DUTTER R. (1977) : Numerical solution of robust regression problems : computational aspects, a comparaison. *J. Stat. Comput. Simul.,* vol. 5, pp. 207-38.

[10] EKBLOM H. (1974) : L_p methods for robust regression. *Nordisk Tidskrift for informations behandling (BIT),* vol. 14, pp. 22-32.

[11] FORSYTHE A.B. (1972) : Robust estimation of straight line regression coefficients by minimizing pth power deviation. *Technometrics,* vol. 14, pp. 159-66.

[12] GUNST R.F., MASON R.L. – Biaised estimation in regression : an evaluation using mean squared error. *JASA*, vol. 72, pp. 616-28.

[13] HAMPEL F.R. (1973). – Robust estimation : A condensed partial survey. *Z. Wahrscheinlichkeitstheorie verw. Geb.,* vol. 27, pp. 87-104.

[14] HARVEY A.C. (1977). – A comparaison of preliminary estimators for robust regression. *JASA,* vol. 72, pp. 910-3.

[15] HETTMANSPERGER T.P., MCKEAN J.N. (1977). – A robust alternative based on ranks to least squares in analysing linear models. *Technometrics,* vol. 19, pp. 275-84.

[16] HINICH M.J., TALWAR P.P. (1975). – A simple method for robust regression. *JASA*, vol. 70, pp. 113-9.

[17] HINKLEY D.V. (1977). – Jackknifing in unbalanced situations. *Technometrics,* vol. 19, pp. 285-92.

[18] HINKLEY D.V. (1977). – Jackknifing confidence limits using Student approximations. *Biometrika,* vol. 64, pp. 21-8.

[19] HINKLEY D.V. (1978). – Improving jackknife with special reference to correlation estimation. *Biometrika,* vol. 65, pp. 13-21.

[20] HOAGLIN D.C., WELSH R.E. (1978). – The hat matrix in regression and ANOVA. *The Amer. Stat.,* vol. 32, pp. 17-22.

[21] HOGG R.V., RANDLES R.H. (1975). – Adaptive distribution free regression methods and their applications. *Technometrics,* vol. 17, pp. 399-407.

[22] HUBER P.J. (1964). – Robust estimation of a location parameter. *Ann. Math. Stat.,* vol. 37, pp. 73-101.

[23] HUBER P.J. (1972). – Robust statistics : a review. *Ann. Math. Stat.,* vol. 43, pp. 1041-67.

[24] HUBER P.J. (1973). – Robust regression. – Asymptotics, conjectures and Monte-Carlo. *Ann. Stat.,* vol. 5, pp. 799-821.

[25] JAECKEL L.A. (1972). – Estimating regression coefficients by minimizing the dispersion of the residuals. *Ann. Math. Statist.,* vol. 43, pp. 1449-58.

[26] LUND R.E. (1975). – Tables for an approximate test for outliers in linear models. *Technometrics,* vol. 17, pp. 473-6.

[27] MARTIN R.D., MASRELIEZ C.J. (1975). – Robust estimation via stochastic approximation *IEEE Trans. Inf. Theory,* vol. IT-31, pp. 263-71.

[28] MILLER R.G. (1964). – A trustworthy jackknife. *Ann. Math. Statist.,* vol. 35, pp. 1594-605

[29] MILLER R.G. (1968). – Jackknifing variances. *Ann. Math. Statist.,* vol. 39, pp. 567-82.

[30] MILLER R.G. (1974). – The jackknife – a review. *Biometrika,* vol. 61, pp. 1-15.

[31] MILLER R.G. (1974). – An unbalanced jackknife. *Ann. Stat.,* vol. 2, pp. 880-91.

[32] MOSTELLER F. (1971). – The Jackknife. *Review of the International Statistical Institute,* Vol. 39, pp. 363-8.

[33] MOSTELLER F., TUKEY J.W. (1968). – *Data Analysis Including Statistics.* In *Handbook of Social Psychology* (Eds G. LINDZEY et E. ARONSON), Addison Wesley, pp. 80-203.

[34] NARULA S.C., WELLINGTON J.F. (1977). – Multiple Linear Regression with Minimum sum of Absolute Errors. *Appl. Stat.,* vol. 26, pp. 106-11.

[35] PRESCOTT R.D. (1975). – Detection of influential Observation in Linear Regression. *Technometrics,* vol. 19, pp. 15-8.

[36] QUENOUILLE M.H. (1949). – Approximate Tests of Correlation in Time Series. *J.R. Statist. Soc. B,* vol. 11, pp. 68-84.

[37] QUENOUILLE M.H. (1956). – Notes on Bias in Estimation. *Biometrika,* vol. 43, pp. 353-60.

[38] RAMSAY J.O. (1977). – A Comparative Study of Several Robust Estimates of Slope, Intercept, and Scale in Linear Regression. *JASA*, vol. 72, pp. 608-15.

[39] REY W.J.J. (1977). – M-estimators in Robust Regression, a Case Study. in : Recent Developments in Statistics, (J.R. BARRA *and al.*, editors). North-Holland, pp. 591-4.

[40] SIELKEN R.L., HARTLEY H.O. (1973). – Two Linear Programming Algorithms for Unbiaised Estimation of Linear Models. *JASA*, vol. 68, pp. 639-41.

[41] SNEE R.D. (1977). --Validation of Regression Models, Methods and Examples. *Technometrics,* vol. 19, pp. 415-28.

[42] SPOSITO V.A., KENNEDY W.J., GENTLE J.E. (1977). – L_p Norm Fit of a Straight Line. *Appl. Stat.* vol. 26, pp. 114-8.

[43] STIGLER S.M. (1973). – Simon Newcomb, Percy Daniel, and the History of Robust Estimation 1885-1920. *JASA*, vol. 68, pp. 872-9.

[44] TUKEY J.W. (1958). – Bias and Confidence in not-quite Large Samples (Abstract). *Ann. Math. Statist.*, vol. 29, p. 614.

[45] TUKEY J.W. (1975). – Instead of Gauss-Markow Least Squares, What ? in Applied Statistics (R.P. GUPTA editor). North-Holland Publishing Company, pp. 351-72.

[46] TUKEY J.W. (1977). – Exploratory Data Analysis. Addison-Welsey.

[47] WEBSTER J.T., GUNST R.F., MASON R.L. (1974). – Latent Root Regression Analysis. *Technometrics,* vol. 16, pp. 513-22.