وزارة التعليم العالي و البحث العلمي

Université BADJI Mokhtar – Annaba
BADJI Mokhtar – Annaba University

جامعة باجي مختار – عنابة

Faculté des Sciences
Département de Chimie
Laboratoire de Sécurité Environnementale et Alimentaire (*LASEA*)

MÉMOIRE

Présenté en vue de l'obtention du diplôme de MAGISTÈRE en Chimie Analytique

THÈME

***REGRESSION (LAD); DIFFERENTES APPROCHES*:**

- ***ITERATIVELY RE-WEIGHTED LEAST SQUARES.***
- ***BASIC ITERATIVE APPROACH DISCUSSED IN [LI AND ARCE 2003].***
- ***WESELOWSKY'S DIRECT DESCENT METHOD [WESOLOWSKY,1981]***

Option : Chimie de l'Environnement

Par: *M^r^. MENACER Rafik*
DES en Chimie

Devant le jury :

Président :	M^me^. Z.HABES	MC	U.B.M. Annaba
Examinateurs:	M^r^. A .TAHAR	Pr	U.B.M. Annaba
	M^me^ . S.ALI-MOKHNACHE	Pr	U.B.M. Annaba
	M^r^. A.H.BAAZIZ	MC	U.B.M. Annaba
Rapporteur :	M^r^.D.MESSADI	Pr	U.B.M. Annaba

Année 2009

Dédicace

A mon père.

A ma mère.

A mes frères et sœurs: Marwa, Chawki, Karima, Nadia, et Nadjib.

A mon grand-père : Said.

A la mémoire de mon grand-père : Homana.

A mes grand-mères.

A mes oncles : Ferhat et Amar, et toutes mes tantes.

A mes cousins et cousines et à toute la famille sans exception.

A tous mes amis qui sont si nombreux.

Rafik MENACER

Remerciements

- **Le grand merci à Dieu pour cette gouttelette de connaissance.**

- **Un grand merci à mes parents qui ont tout fait pour que mon parcours scientifique soit un confort.**

- **J'exprime ma sincère gratitude à mon promoteur M^{r}. *Djelloul MESSADI*.**

- **Je remercie M^{r}. *BAAZIZ Abd El Halim* pour son aide précieuse.**

- **Je remercie les membres de jury qui ont bien voulu me faire l'honneur d'accepter de juger mon travail.**

- **Je remercie toute l'équipe du labo 34.**

- **Je remercie tous ceux qui ont contribué à l'élaboration de ce modeste travail sans exception.**

ملخص:

طريقة الفوارق الصغرى بالقيم المطلقة من أهم الطرق البديلة لطريقة المربعات الصغيرة عندما يتعلق الأمر بتقدير عوامل نموذج الانحدار.

شيوع طريقة المربعات الصغيرة يعود خصوصا إلى سهولة الحسابات, ولكن اليوم مع تطور الإعلام الآلي فان طريقة الفوارق الصغيرة بالقيم المطلقة يمكن استعمالها تقريبا بنفس السهولة .

المعيار المستعمل في طريقة المربعات الصغرى هو التقليل من مجموع الأخطاء المربعة : $\sum e_i^2$.

بينما المعيار المستعمل في طريقة الفوارق الصغيرة بالقيم المطلقة هو التقليل من مجموع القيم المطلقة للأخطاء: $\sum |e_i|$.

لا توجد صيغ جلية لحساب المقيم لطريقة الفوارق الصغرى بالقيم المطلقة لان السبب يكمن في أن دالة القيمة المطلقة غير قابلة للاشتقاق ولكن يمكن حسابه بتطبيق خوارزمي تكراري.

الكلمات الدالة:

طريقة الفوارق الصغرى بالقيم المطلقة، طريقة المربعات الصغرى، الخوارزمي التكراري، التقليل.

Abstract:

The least absolute deviations method (LAD) is an important alternative of the least squares method (LS) for estimating the parameters of a regression model.

The popularity of the least squares method LS is based mainly on the simplicity of the calculations. However, nowadays, with the progress of the data processing, the LAD can be used as simply as the LS.

The minimization's criteria used by the LS is the sum of the squares of the residuals: $\sum \mathbf{e_i^2}$, the minimization's criteria used by the LAD method is the sum of the absolute values of the residuals: $\sum \left| \mathbf{e_i} \right|$.

We do not dispose of an explicit formula to calculate the LAD estimator, because the absolute value function is not derivable.

However the least absolute deviations estimator can be calculated by applying an iterative algorithm. Different approaches were programmed then tested; the used descriptor was a hydrophobicity indicator.

Key words: LAD, LS, Minimization, Iterative algorithm, Hydrophobicity indicator.

Résumé :

La méthode des moindres écarts en valeurs absolues ou (LAD) pour *least absolute deviations*, est une importante alternative à la méthode des moindres carrés (LS) lorsqu'il s'agit d'estimer les paramètres d'un modèle de régression.
La popularité de la méthode des moindres carrés (LS) repose principalement sur la simplicité des calculs. Mais aujourd'hui, avec les progrès de l'informatique, la méthode LAD peut être utilisée presque aussi simplement.
Le critère de minimisation utilisé par la méthode des moindres carrés est la somme des carrés des résidus : $\sum \mathbf{e}_i^2$, alors que le critère de minimisation utilisé par la méthode LAD est la somme des valeurs absolue des résidus : $\sum \left| \mathbf{e_i} \right|$.
On ne dispose pas de formule explicite pour calculer les estimateurs LAD puisque la fonction valeur absolue n'est pas dérivable.
Les estimateurs LAD peuvent toutefois être calculés en appliquant un algorithme itératif. Différents algorithmes ont été programmés puis testés en utilisant comme descripteurs un indicateur d'hydrophobicité.

Mots clés : LAD, LS, Minimisation, Algorithme itératif, indicateur d'hydrophobicité.

LISTE DES FIGURES

LISTE DES TABLEAUX

LISTE DES SYMBOLES ET ABREVIATIONS

$\sum \lvert e_i \rvert$:	Somme d'erreurs en valeurs absolues.
ATP :	Adinozine triphosphate.
α :	Niveau de confiance.
β_0 ou ***b:***	Ordonnée à l'origine.
β_1 ou ***m :***	Pente.
β_j***:***	$j^{ème}$ coefficient de régression.
CIC50 :	Concentration d'inhibition 50 % de la croissance.
DMSO :	Diméthylsulfoxide.
E( )	Espérance mathématique.
e_i ***:***	Erreur résiduelle, $\mathbf{e_i}=Y_i - \hat{Y}_i$.
e_i***(LS) :***	Erreurs par la LS.
e_i***(LAD) :***	Erreurs par la LAD.
F :	Statistique de Fisher.
F_{calc} ***:***	Valeur de F calculée.
F_{obs} ***:***	Valeur de F observée.
H_0 ***:***	Hypothèse nulle
H_1 ***:***	Hypothèse alternative
i :	Indice de l'observation.
i.i.d :	Indépendantes et identiquement distribuées.
LAD:	Méthode des moindres écarts en valeurs absolues pour *Least Absolute Deviations*.
LAD_1***:***	LAD correspondant à l'algorithme des moindres carrés re-pondérés itérativement.
LAD_2***:***	LAD correspondant à l'algorithme itératif de base.
LAD_3***:***	LAD correspondant à l'algorithme de la descente directe.
logP ou ***(log kow)*** **:**	Coefficient de partage (Octanol /Eau).
LS**:**	Méthode des moindres carrés pour *Least Squares*.
MED :	Médiane pondérée

n:	Taille de la population (échantillon).
n-2 :	Nombre de degrés de liberté.
p :	Nombre de paramètres.
p-1 :	Nombre de descripteurs.
pCIC50 :	log 1/(CIC50).
QSAR/QSPR:	Quantitative Structure Activity/Property Relationships.
R^2 **:**	Coefficient de détermination.
S :	Erreur standard.
σ^2 **:**	Variance pour la LS.
SCE :	Somme des carrés des écarts.
SCT :	Somme des carrés totale.
t :	Statistique t de Student.
t_{calc} **:**	Valeur de t de Student calculée.
t_{obs} **:**	Valeur de t de Student observée.
$\hat{\tau}^2$ *:*	Variance pour la LAD.
W_i **:**	Poids associé à la ième observation.
X :	Variable explicative.
Y :	Variable à expliquer.
$\hat{Y}_{lad}$ **:**	Valeur de toxicité pCIC50 estimée par LAD.
$\hat{Y}_{ls}$ *:*	Valeur de toxicité pCIC50 estimée par LS.

SOMMAIRE

CHAPITRE IV : STATISTIQUE DE LA LS ET DE LA LAD

CHAPITRE V : RESULTATS ET DISCUSSION

CHAPITRE I

INTRODUCTION

Les composés organiques industriels, estimés actuellement à 120 000 avec apparition annuelle sur les marchés de 1000 produits nouveaux, ne sont pas toujours sans risques pour la santé publique et l'environnement.

Les fichiers de données expérimentales, complets, homogènes et précis les concernant, s'ils sont parfois disponibles, peuvent faire défaut même pour les composés du commerce les plus courants et les plus importants.

La détermination expérimentale systématique de toutes les données manquantes qui se traduirait par une lourde charge, économiquement insupportable pour les industriels et l'autorité de régulation, dépasse les capacités de recherche disponibles, nonobstant les larges marges d'erreurs qu'elle pourrait engendrer.

Aussi, la gestion systématique et globale des risques encourus par la présence sur le marché et dans l'environnement de la grande masse de produit chimiques, ne peut reposer uniquement sur la seule disponibilité des données expérimentales. D'où l'intérêt à développer des modèles quantitatifs qui permettent la prévision rapide et précise de la toxicité et de l'évolution dans l'environnement de polluants organiques, à partir de la seule information encodée dans leurs formules structurales.

La concentration d'inhibition 50% de la croissance (CIC 50) d'une population de *protozoaires ciliés* sert souvent d'indice de toxicité, on considère que l'action des polluants se manifeste par un dysfonctionnement des membranes cellulaires et donc la toxicité éventuelle d'une molécule dépend de sa tendance à s'y accumuler. L'octanol, milieu apolaire, constitue un modèle simple des membranes, ce qui explique que de nombreuses relations structure /activité intègrent logP comme variable explicative.

L'analyse de régression est réalisée en utilisant, souvent, la méthode des moindres carrés ordinaire.

L'utilisation de la méthode des moindres carrés dans le modèle de régression linéaire nécessite certaines hypothèses, notamment sur les erreurs.

En effet, l'estimateur des moindres carrés (LS) doit sa popularité en partie au fait qu'il possède sous certaines hypothèses, la variance minimale parmi tous les estimateurs linéaires non biaisés.

Pour construire le modèle et admettre que les coefficients de la régression sont sans biais et convergents, on montre qu'il faut poser comme hypothèses:

a) Les résidus e_i ont une espérance (E) mathématique nulle:

$$E(e_i) = 0$$

b) Le modèle choisi est correct (aucune variable explicative n'a été omise).

c) Les résidus sont indépendants entre eux:

$$E(e_j, e_i) = 0 \qquad \text{si } i \neq j$$

leurs covariances sont nulles.

d) Les résidus ont tous même variance σ^2 (propriété d'homoscédascité).

Par ailleurs, l'emploi de tests statistiques pour analyser la variation expliquée par la régression conduit à admettre que:

e) Les résidus suivent une distribution normale (de Laplace- Gauss).

Il faut de plus mentionner que, même si la majorité des erreurs dans le modèle suivent une distribution normale, il arrive souvent qu'un petit nombre d'observations suivent une distribution différente. Dans ce cas, on dit que l'échantillon est contaminé par des valeurs aberrantes. Puisque les estimateurs LAD sont peu sensibles aux données aberrantes, ils sont particulièrement adaptés à ce genre de situations.

Le but de ce travail consiste à faire une comparaison entre les méthodes LAD et LS en ce qui concerne la modélisation de la toxicité CIC50 de 21 alcools et 9 amines avec l'indicateur d'hydrophobicité logP.

I- DEFINITIONS :

I.1.1- LAD :

En anglais *least absolue déviations*, la méthode des moindres écarts en valeurs absolues, est une méthode de régression basée sur la minimisation de la somme des erreurs en valeurs absolues $\sum |e_i|$.

I.1.2- LS :

En anglais *least squares*, la méthode des moindres carrés, est une méthode de régression basée sur la minimisation de la somme des carrés des erreurs $\sum e_i^2$.

I.1.3- TOXICITE :

Propriété d'une substance (poison) capable de tuer un être vivant, pCIC50 signifie log(1/CIC50) servira d'indicateur de toxicité (CIC50 =Concentration d'inhibition 50 % de la croissance.).

I.1.4- MODELISATION:

La modélisation des données est l'art d'extraire des informations utiles d'un ensemble de données obtenues par des mesures, et de condenser cette information dans un modèle exploitable.

I.1.5- REGRESSION :

Un problème de régression consiste à étudier les changements de la valeur moyenne d'une variable (aléatoire) quand une autre variable ou plusieurs autres variables prennent différentes valeurs fixes. La première variable est appelée variable dépendante ou variable expliquée, les autres variables sont appelées variables indépendantes, variables explicatives. Comme dans notre étude il y a une seule variable explicative, on dit qu'il y a une régression simple; lorsqu'il y a au moins deux variables explicatives on dit qu'il y a une régression multiple.

I.1.6- COEFFICIENT DE DETERMINATION :

Il est ensuite possible de quantifier la plus ou moins bonne adaptation de la droite de la régression aux données grâce au coefficient de détermination R^2.

I.1.7- ESTIMATION :

L'estimation est une opération ou action de prédire une grandeur, elle a pour objectif de connaître, à partir de l'observation de l'échantillon, la véritable valeur d'une variable dans la population (sa fréquence, s'il s'agit d'une variable qualitative ; sa moyenne, s'il s'agit d'une variable quantitative).
Mais, du fait de l'incertitude liée aux fluctuations d'échantillonnage, il est impossible de connaître avec certitude la valeur exacte dans la population : on ne peut que l'*estimer* en calculant la probabilité que cette véritable valeur se trouve comprise dans un certain intervalle.

I.1.8- PROTOZOAIRES :

Les Protozoaires, étant unicellulaires, sont de petits organismes de moins d'un millimètre, pouvant s'associer en colonies.
Ils vivent exclusivement dans l'eau ou dans de la terre humide. Ils sont connus pour être responsables de nombreuses maladies telle que la malaria.

I.2- PROBLEMATIQUE :

La méthode la plus utilisée pour estimer les paramètres d'un modèle de régression linéaire simple est sans doute la méthode des moindres carrés (LS) mais cette dernière présente moins de robustesse aux valeurs aberrantes, qui sont assez fréquentes dans la recherche d'un modèle, qui prédit la toxicité pCIC50 de 21alcools et 9 amines en fonction du coefficient de partage (Octanol/Eau) logP, d'où le nécessaire recours à une méthode alternative robuste aux valeurs aberrantes.

I.3- SOLUTION PROPOSEE :

Comme il n'est pas rare de rencontrer le problème des données aberrantes, nous allons modéliser nos données via la méthode LAD, cette méthode consiste à utiliser diverses approches robustes, parmi lesquelles, les trois suivantes exploitées dans cette étude :

(1) La méthode des moindres carrés re-pondérés itérativement, désignée par LAD_1.
(2) L'approche Itérative de base, désignée par LAD_2.
(3) La méthode de la descente directe, désignée par LAD_3.

I.4- ESQUISSE DE LA SOLUTION :

La solution que nous avons proposée passe par les étapes suivantes :
(1) Collecte des données.
(2) Modélisation des données par la méthode des moindres carrés (LS).
(3) Programmation en langage Pascal des trois algorithmes de la méthode des moindres écarts en valeur absolue :
i La méthode des moindres carrés re-pondérés itérativement.
ii L'approche itérative de base.
iii La méthode de la descente directe.
(4) Modélisation des données par les trois approches LAD.
(5) Comparaison des modèles obtenus par les trois approches LAD et le modèle LS.

I.5- PLAN DU MEMOIRE :

Ce mémoire comporte six chapitres, le premier chapitre est une introduction visant à présenter le cadre de notre travail, dans lequel la problématique de cette étude a été exposée. Le deuxième chapitre intitulé état de l'art développe la méthodologie du travail.
Le troisième chapitre parle de l'histoire des deux méthodes LS et LAD, puis la statistique de ces deux méthodes sera détaillée dans le quatrième chapitre. Les résultats seront exposés et discutés dans le cinquième chapitre, et nous achèverons le mémoire par une conclusion générale dans le sixième chapitre.
A la fin de ce mémoire se trouve une bibliographie suivie par une annexe comportant le code source des programmes en langage Pascal.

CHAPITRE II

ETAT DE L'ART

II.1- TOXICITE :

Il est généralement admis que la toxicité de nombreuses substances, particulièrement les produits chimiques organiques industriels, est la conséquence de leur solubilité dans les lipides, alors que leurs caractéristiques moléculaires spécifiques ont peu ou pas d'influence. Leur mode d'action consisterait en la destruction des processus physiologiques associés aux membranes cellulaires.

Les *protozoaires* sont souvent utilisés pour l'évaluation de la toxicité, les méthodes mises en œuvre sont basées sur des critères morphologiques, ultra-structuraux, éthologiques et métaboliques

L'inhibition de la croissance d'une population est un indicateur très en vogue, parce qu'il peut être déterminé directement ou indirectement à l'aide d'un équipement électronique, ce qui permet l'acquisition rapide des observations nécessaires pour les analyses de régression. Nous considérerons la concentration d'inhibition 50% de la croissance (CIC50), dont le logarithme de l'inverse, soit $pCIC50 = \log (CIC50)^{-1}$, servira d'indicateur de toxicité.

II.2- COLLECTE DES DONNEES :

Les tests de toxicité ont été réalisés (Schultz, 1990) en examinant la croissance d'une population de *Tetrahymena pyriformis*. La température a été fixée à 27 ± 1°C et les essais ont été menés dans des erlenmeyers de 250 ml, contenant 50 ml d'un milieu dont la composition est précisée ci après :

Eau distillée	1000 mL
Protéose peptone	20 g
D-glucose	5 g
extrait de levure	1 g
FeEDTA	1 mL d'une solution à3 % (masse/v)
pH	7,35

Ce milieu est inoculé avec 0,25 ml d'une culture contenant approximativement 36000 cellules par ml. La croissance des ciliés est suivie par spectrophotométrie, en mesurant la densité optique (absorbance) à 540 nm après 48 heures d'incubation.

Plusieurs critères ont guidé au choix des composés toxiques examinés. Tous sont disponibles dans le commerce avec une pureté suffisante (95 % et plus), ce qui ne nécessite pas une re-

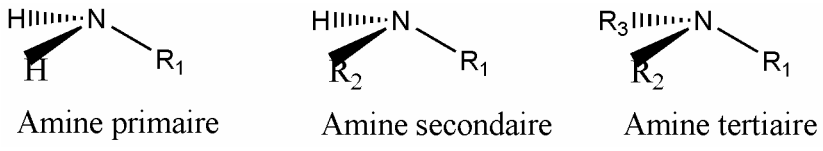
purification préalablement au test. Des précautions ont été observées afin d'assurer une diversité concernant, à la fois, les propriétés physico-chimiques et la position des substituants. Les solutions stocks des divers composés toxiques, ont été préparées dans le diméthylsulfoxide (DMSO) à des concentrations de 5, 10, 25 et 50 grammes par litre. Dans chaque cas, le volume de solution stock ajouté à chaque fiole est limité par la concentration finale de DMSO qui ne doit pas excéder 0,75 % (350 ml par fiole), quantité qui n'altère pas la reproduction de *Tétrahymena*.

II.3- COEFFICIENT DE PARTAGE (OCTANOL/EAU) LOGP:

Le coefficient de partage appelé aussi *Log Kow*, est une mesure de la solubilité différentielle de composés chimiques dans deux solvants (coefficient de partage Octanol/Eau), logP est égal au logarithme du rapport des concentrations de la substance étudiée dans l'octanol et dans l'eau, logP =Log(C_{oct} /C_{eau}), Cette valeur permet d'appréhender le caractère hydrophile ou hydrophobe (lipophile) d'une molécule. En effet, si logP est positif et très élevé,cela exprime le fait que la molécule considérée est bien plus soluble dans l'octanol que dans l'eau,ce qui reflète son caractère lipophile,et inversement. Une valeur de logP=0 signifie que la molécule se répartit de manière égale entre les deux phases et C_{oct} =C_{eau}.

II.4- AMINES :

Découvertes en 1849, par Wurtz les amines furent initialement appelées alcaloïdes artificiels. Une **amine** est un composé organique dérivé de l'ammoniac dont certains hydrogènes ont été remplacés par un groupement carboné. Si l'un des carbones liés à l'atome d'azote fait partie d'un groupement carbonyle, la molécule appartient à la
famille des amides,On parle d'amine primaire, secondaire ou tertiaire selon que l'on a un, deux ou trois hydrogènes substitués.



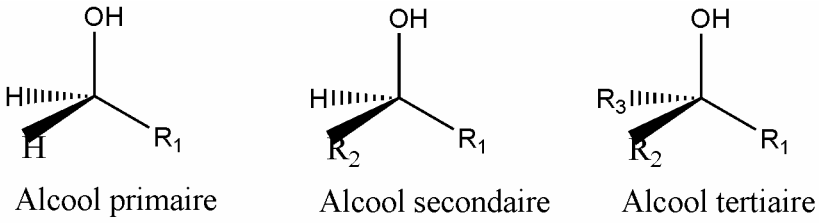
Amine primaire Amine secondaire Amine tertiaire

II.5- ALCOOLS :

En chimie organique, un **alcool** est un composé organique dont l'un des carbones (celui-ci étant tétragonal) est lié à un groupement hydroxyle (-OH). L'éthanol (ou alcool éthylique) entrant dans la composition des boissons alcoolisées est un cas particulier d'alcool, mais tous

les alcools ne sont pas propres à la consommation. En particulier, le méthanol est toxique et mortel à haute dose.

Nous distinguons trois types, l'alcool primaire, secondaire ou tertiaire selon que l'on a un, deux ou trois hydrogènes substitués.



Alcool primaire Alcool secondaire Alcool tertiaire

L'éthanol est une substance psychotrope toxique voire mortelle en grande quantité, même en quantité modérée en cas de consommation régulière, les autres alcools sont généralement beaucoup plus toxiques.

II.6- MODELISATION DES DONNEES :

Les relations LAD et LS ont été déterminées en prenant pour variable dépendante le logarithme de l'inverse de CIC50 en (mmol / litre)**,** et pour variable explicative logP.

Les données ont été modélisées en utilisant la régression par la méthode des moindres carrés de Minitab, et pour la méthode des moindres écarts en valeurs absolues (LAD), à l'aide d'applications programmées en langage Pascal et exécutables sous Windows.

Les données utilisées dans ce travail sont condensées dans le tableau1:

Tableau1 : Données.

N°	Composé	logP	pCIC50
1	Méthanol	-0.77	-2,77
2	Ethanol	-0.31	-2,41
3	Propan-1-ol	0.25	-1,84
4	Butan-1-ol	0.88	-1,52
5	Pentan-1-ol	1.56	-1,12
6	Hexan-1-ol	2.03	-0,47
7	Heptan-1-ol	2.57	0,02
8	Octan-1-ol	3.15	0,5
9	Nonan-1-ol	3.69	0,77
10	Decan-1-ol	4.23	1,1
11	Undecan-1-ol	4.77	1,87
12	Dodecan-1-ol	5.13	2,07
13	Tridecan-1-ol	5.67	2,28
14	Propan-2-ol	0.05	-1,99
15	Pentan-2-ol	1.21	-1,25
16	Pentan-3-ol	1.21	-1,33
17	2-methylbutan-1-ol	1.42	-1,13
18	3-methylbutan-1-ol	1.42	-1,13
19	3-methylbutan-2-ol	1.28	-1,08
20	(ter) pentanol	1.21	-1,27
21	(neo) pentanol	1.57	-0,96
22	1-propylamine	0.48	-0,85
23	1-butylamine	0.97	-0,7
24	1-amylamine	1.47	-0,61
25	1-hexylamine	2.06	-0,34
26	1-heptylamine	2.57	0,1
27	1-octylamine	3.04	0,51
28	1-nonylamine	3.57	1,59
29	1-decylamine	4.1	1,95
30	1-undecylamine	4.63	2,26

II.7- TROIS APPROCHES LAD EXPLOITEES :

Nous présentons dans ce paragraphe les trois approches LAD qui ont été mises en œuvre au cours de ce travail. Tous les algorithmes utilisent les paramètres du modèle trouvés par la méthode LS comme paramètres d'initialisation, pour ensuite calculer les paramètres du modèle LAD, ce qui permet une économie en temps de calcul.

II.7.1- METHODE DES MOINDRES CARRES RE-PONDERES ITERATIVEMENT :

Le principe de la méthode des moindres carrés pondérés consiste à calculer la régression et les résidus, à assigner à chaque individu un poids, d'autant plus faible que le résidu est grand, et à faire une nouvelle régression.

Avec la méthode des moindres carrés re-pondérés itérativement, on fait pareil, mais on itère jusqu'à ce que les paramètres se stabilisent.

L'algorithme effectue une minimisation pour trouver m et b:

(i) $\sum_{i=0}^{N-1} w_i (Y_i - b_{(n)} - m_{(n)} X_i)^2$, où $w_i = \frac{1}{|Y_i - b_{(n-1)} - m_{(n-1)} X_i|}$.

$b_{(n-1)}$, $m_{(n-1)}$ représentent les paramètres trouvés par l'itération précédente (n-1), et $b_{(n)}$,$m_{(n)}$ sont les paramètres à trouver lors de la présente itération (n).

Prendre les dérivées partielles de (i) par rapport à b, et m.

(ii) $\frac{d}{db} \sum_{i=0}^{N-1} w_i (Y_i - b_{(n)} - m_{(n)} X_i)^2 = -2 * \sum_{i=0}^{N-1} w_i (Y_i - b_{(n)} - m_{(n)} X_i)$

(iii) $\frac{d}{dm} \sum_{i=0}^{N-1} w_i (Y_i - b_{(n)} - m_{(n)} X_i)^2 = -2 * \sum_{i=0}^{N-1} w_i X_i (Y_i - b_{(n)} - m_{(n)} X_i)$

Poser (ii) et (iii) égal à 0, pour trouver le minimum de chaque équation.

(iv) $-2 * \sum_{i=0}^{N-1} w_i (Y_i - b_{(n)} - m_{(n)} X_i) = 0$

(v) $-2 * \sum_{i=0}^{N-1} w_i X_i (Y_i - b_{(n)} - m_{(n)} X_i) = 0$

(iv) et (v) peuvent être simplifié.

(vi) $\sum_{i=0}^{N-1} w_i Y_i = b_{(n)} \sum_{i=0}^{N-1} w_i + m_{(n)} \sum_{i=0}^{N-1} w_i X_i$

(vii) $\sum_{i=0}^{N-1} w_i X_i Y_i = b_{(n)} \sum_{i=0}^{N-1} w_i X_i + m_{(n)} \sum_{i=0}^{N-1} w_i (X_i)^2$

Dans (vi) et (vii) les seules inconnues sont b et m en utilisant la méthode préférée pour résoudre le système de deux équations de deux inconnues on trouvera (viii) et (ix) :

(viii) $$b_{(n)} = \frac{\sum_{i=0}^{N-1} w_i (X_i)^2 \sum_{i=0}^{N-1} w_i Y_i - \sum_{i=0}^{N-1} w_i X_i \sum_{i=0}^{N-1} w_i X_i Y_i}{\sum_{i=0}^{N-1} w_i \sum_{i=0}^{N-1} w_i (X_i)^2 - \left(\sum_{i=0}^{N-1} w_i X_i \right)^2}$$

(ix) $$m_{(n)} = \frac{-\sum_{i=0}^{N-1} w_i X_i \sum_{i=0}^{N-1} w_i Y_i + \sum_{i=0}^{N-1} w_i \sum_{i=0}^{N-1} w_i X_i Y_i}{\sum_{i=0}^{N-1} w_i \sum_{i=0}^{N-1} w_i (X_i)^2 - \left(\sum_{i=0}^{N-1} w_i X_i \right)^2}$$

Après un nombre suffisant d'itérations n_0; (viii) et (ix) vont définir les paramètres de la méthode des moindres carrés re-pondérés itérativement (LAD_1), La **figure1** reproduit l'algorigramme correspondant.

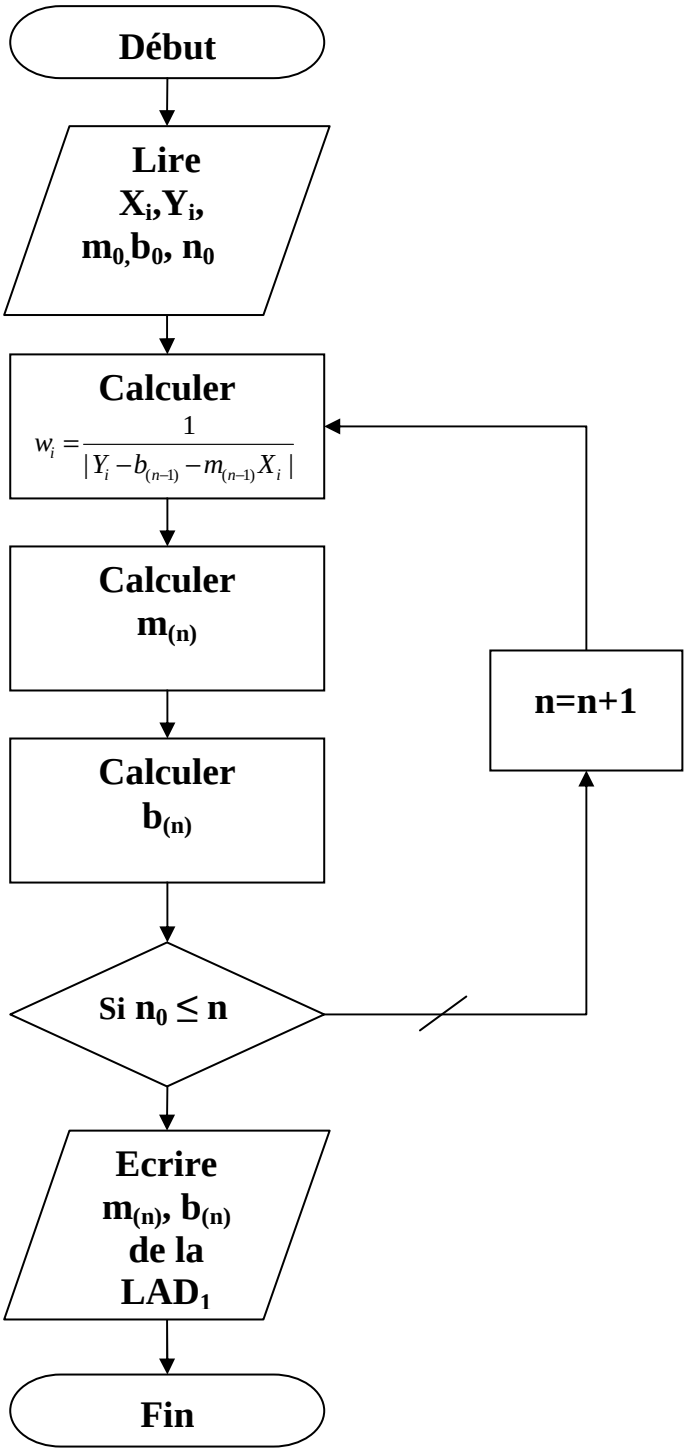


Figure1 : Algorigramme de la méthode des moindres carrés re-pondérés.

II.7.2- APPROCHE ITERATIVE DE BASE :

Cet algorithme se base sur le principe de la médiane pondérée, introduite par Laplace (voir le chapitreIII).

(i) Poser k=0, et trouver m_0 initial trouvé par la méthode des moindres carrés.

(ii) Poser k=k+1et obtenir une nouvelle estimation de b en fixant m_{k-1}.

$$b_k = MED\left(Y_i - m_{k-1} X_i \Big|_{i=1}^{N} \right)$$

(iii) Obtenir une nouvelle estimation de m en fixant b_k.

$$m_k = MED\left(|X_i| \lozenge \frac{Y_i - b_k}{X_i} \Big|_{i=1}^{N} \right)$$

(vi) Une fois que m_k et b_k ne dérivent pas au dessus de l'ordre de tolérance arrêter, m_k et b_k vont définir les paramètres de l'approche itérative de base (LAD_2).

Sinon aller à l'étape (ii).

La **figure2** reproduit l'algorigramme correspondant à l'algorithme itératif de base.

REMARQUE :

◊ : c'est l'opérateur de réplication

Exemple**: A** ◊ **B** produit B répliquer A fois.

MED: c'est la médiane pondérée des données, un poids positif est associé à chaque donnée, elle est déterminée avec la procédure suivante:

$$Y = MED\left(W_i \lozenge X_i \Big|_{i=1}^{N} \right)$$

(i) Calculer $W_0 = \frac{1}{2} \sum_{i=1}^{N} W_i$.

(ii) Associer à chaque donnée son poids correspondant, et classer les X_i par ordre croissant.

(iii) Sommer les poids ordonnés jusqu'à ce que $\sum W_i \geq W_0$

(iv) prendre le premier X_j qui satisfait $\sum W_i \geq W_0$

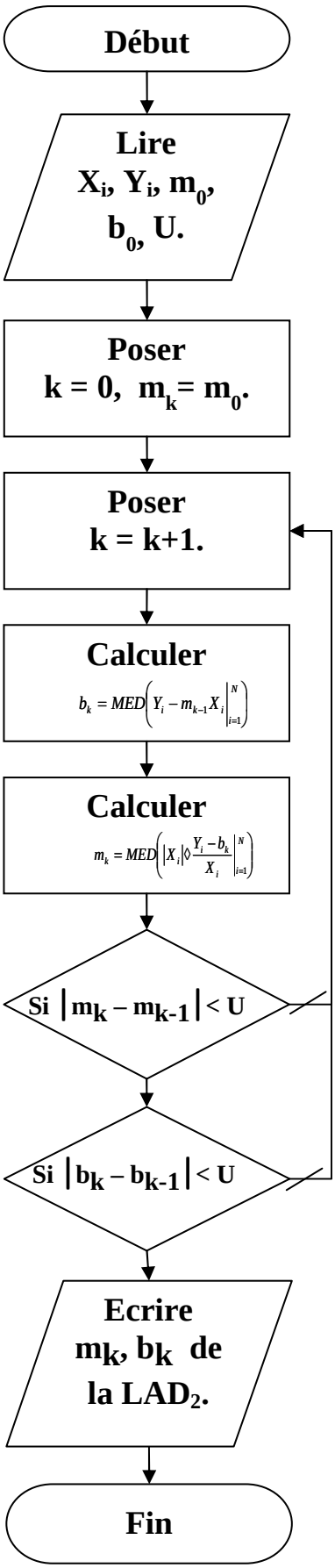


Figure2 : Algorigramme de l'approche itérative de base.

II.7.3- METHODE DE LA DESCENTE DIRECTE:

Cet algorithme se base sur le principe de la médiane pondérée, introduite par Laplace (voir le chapitreIII).

(i) poser k=0, choisir m_0, et b_0 qui sont la solution de la méthode des moindres carrés, et choisir J qui donne $\left|Y_j - m_0 X_j - b_0\right|$ minimale.

(ii) poser k=k+1 et utiliser la médiane pondérée pour trouver *b*:

$$b_k = MED\left(\left|1 - \frac{X_i}{X_j}\right| \Diamond \frac{Y_i - \frac{Y_j X_i}{X_j}}{1 - \frac{X_i}{X_j}} \Bigg|_{i=1}^{N} \right)$$

(iii) Si b_k- b_{k-1} est au dessous de l'ordre de tolérance aller à l'étape (iv), sinon poser J=i et aller à l'étape 2.

(iv) Laisser b* =b_k, m*=$\frac{y_j}{x_j} - \frac{b^*}{x_j}$ ou b*, m* vont définir les paramètres de la méthode de la descente directe (LAD_3).

La **figure3** reproduit l'algorigramme correspondant à l'algorithme de la descente directe.

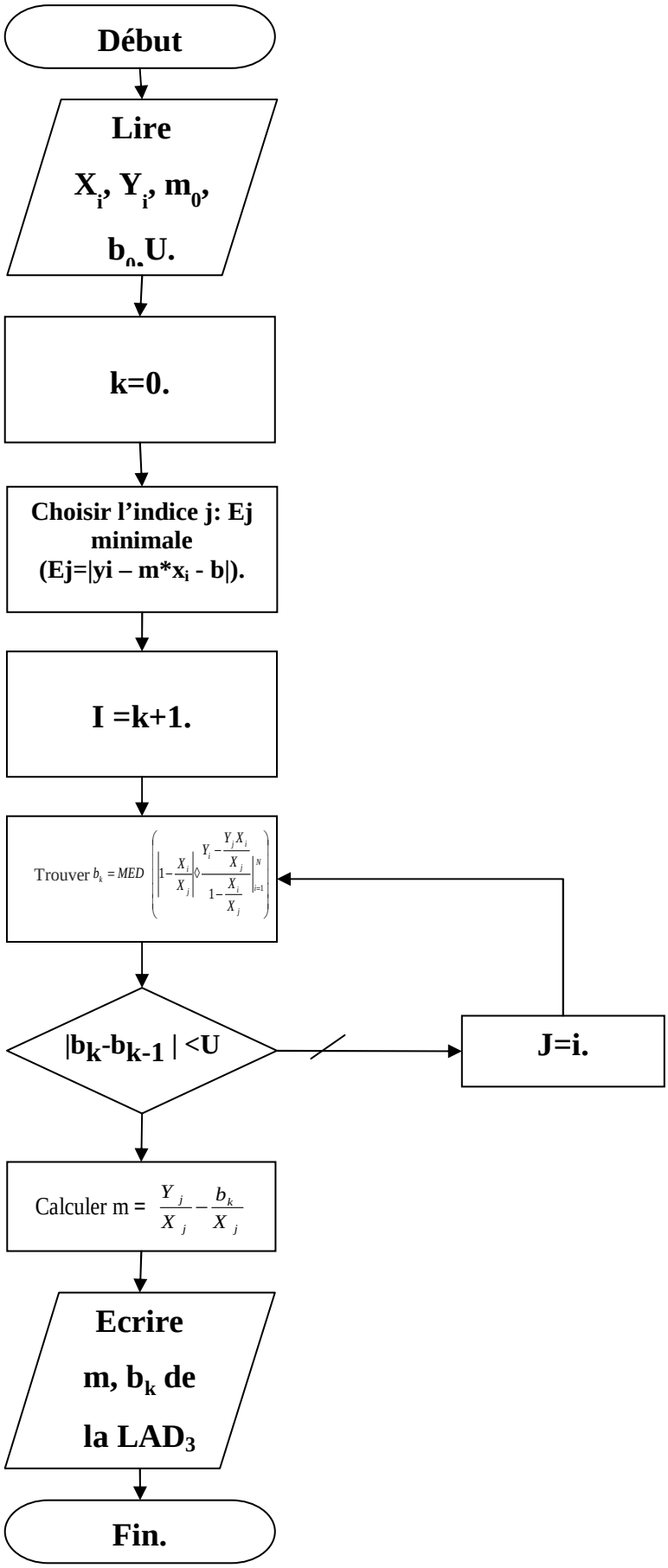


Figure3 : Algorigramme de la méthode de la descente directe.

CHAPITRE III

DECOUVERTE DES ESTIMATEURS LS ET LAD

III.1- HISTORIQUE DE L'ESTIMATION LAD (1757-1955) :

Parmi les estimateurs robustes, les estimateurs LAD ont probablement l'histoire la plus ancienne. En effet, Ronchetti (1987) mentionne qu'on en retrouve des traces dans l'oeuvre de Galilée (1632), intitulée *"Dialogo dei massimi sistemi"*, Le problème était alors de déterminer la distance de la terre à une étoile récemment découverte à cette époque. C'est cependant à Boscovich (1757) que l'on reconnaît généralement l'introduction du critère d'estimation LAD (Harter,1974 ; Ronchetti, 1987, Dielman,1992).

L'un des problèmes qui excita le plus, la curiosité des hommes de science du *XVIII*^{ème} siècle fut celui de la détermination de l'ellipticité de la terre. C'est dans ce contexte, près d'un demi-siècle avant l'annonce par (Legendre,1805) du principe des moindres carrés et vingt ans avant la naissance de Gauss en 1777, que Roger Joseph Boscovich (1757) proposa une procédure pour déterminer les paramètres du modèle de régression linéaire simple

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1, \ldots, n$$

Pour obtenir la droite $y = \hat{\beta}_0 + \hat{\beta}_1 x$ décrivant au mieux les observations, il proposa le critère de l'estimation LAD :

$$\text{Min} \sum_{i=1}^{n} \left| y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right|$$

En imposant que la droite estimée passe par le centroïde des données $\left(\overline{x}; \overline{y}\right)$, en ajoutant la condition :

$$\sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) = 0$$

Boscovich justifia cette approche de la manière suivante. Le critère de l'estimation LAD comme étant nécessaire pour que la solution soit aussi proche que possible des observations, et la condition supplémentaire pour que les erreurs positives et négatives soient de probabilité égales.

En effet cette condition signifie que la somme des erreurs positives et négatives doit être la même. De plus, elle peut se mettre sous la forme :

$$\overline{y} - \hat{\beta}_0 - \hat{\beta}_1 \overline{x} = 0$$

(1)

d'où

$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}$$

(2)

Et le problème se réduit alors à minimiser :

$$\sum_{i=1}^{n} \left| \left(y_i - \overline{y} \right) - \widehat{\beta}_1 \left(x_i - \overline{x} \right) \right|$$

Par conséquent, la détermination de la "droite de Boscovich" satisfaisant les deux critères revient à déterminer la pente $\widehat{\beta}_1$ de l'équation (2), puis à évaluer l'ordonnée à l'origine $\widehat{\beta}_0$ par l'équation (1). Ce n'est cependant que trois ans plus tard, en 1760, que Boscovich donna une procédure géométrique permettant de résoudre l'équation (2). Cette procédure est décrite en détails dans un article d'Eisenhart (1961).

Sept ans avant de s'intéresser aux estimateurs LAD , Laplace (1786) proposa une procédure permettant d'estimer les paramètres d'un modèle de régression linéaire simple en se basant sur le critère L_∞ .En d'autres termes, il proposa une solution pour trouver $\left(\widehat{\beta}_0 , \widehat{\beta}_1 \right)$ qui minimise :

$$\max_{1 \le i \le n} \left| y_i - \widehat{\beta}_0 - \widehat{\beta_1} x_i \right| = \max_{1 \le i \le n} \left| \widehat{e}_i \right|$$

Dans une publication ultérieure, Laplace (1793) proposa une procédure qu'il qualifie lui-même de plus simple. Cette procédure, basée sur les deux critères qu'avait proposé Boscovich en 1757, a l'avantage d'être analytique alors que celle proposée par Boscovich était géométrique.

L'intérêt de cette procédure analytique réside dans la facilité à obtenir les paramètres $\widehat{\beta}_0 \; et \; \widehat{\beta}_1$ lorsque le nombre d'observations augmente. Cette solution analytique de Laplace est élégante et mérite d'être rappelée ici. En adoptant les notations suivantes

$$Y_i = y_i - \overline{y} \; et \; X_i = x_i - \overline{x}$$

Le problème revient à trouver la valeur de ß qui minimise la fonction

$$f\ (\beta) = \sum_{i=1}^{n} \left| y_i - \beta X_i \right|$$

(3)

Notons que les valeurs X_i peuvent être supposées non nulles $\left(X_I \neq 0 \right)$ puisque $f\ (\beta) = \sum \left| y_i \right| + \sum \left| y_i - \beta X_i \right|$, la première somme étant prise pour

les *Xi* nuls et la seconde somme pour les X_i non nuls. Le minimum de la fonction f étant atteint pour la même valeur de β que celle rendant la seconde somme minimale. La fonction (3) peut s'écrire :

$$f(\beta) = \sum_{i=1}^{n} f_i(\beta) \text{ avec } f_i(\beta) = \left| Y_i - \beta X_i \right|$$

Chaque fonction $f_i(\beta)$ est continue, linéaire par morceaux et convexe.

Elle est formée de deux droites avec un minimum en $\left(\frac{y_i}{x_i}; 0\right)$. Sa pente est donnée par

$$f(\beta) = \begin{cases} -\left|X_i\right| & \text{si } \beta < \frac{y_i}{x_i} \\ +\left|X_i\right| & \text{si } \beta > \frac{y_i}{x_i} \end{cases}$$

Pour étudier la pente de $f(\beta)$, il s'agit d'ordonner en ordre croissant les rapports $\frac{Y_i}{X_i}$ de manière à ce que :

$$\frac{Y_1}{X_1} \leq \frac{Y_2}{X_2} \leq \ldots\ldots\ldots \leq \frac{Y_n}{X_n}$$

Ceci peut toujours être fait en renumérotant les observations. Ces rapports $\frac{Y_k}{X_k}$ seront désignés par $\beta_{(k)}, k = 1, \ldots n$ Pour $\beta < \beta_{(1)}$, chacune des fonctions $f_i(\beta)$ a une pente de $-\left|X_i\right|$ et par conséquent la pente de la fonction (I.3) est donnée par :

$$f'(\beta) = -\sum_{i=1}^{n} \left|X_i\right|$$

En chaque point $\frac{Y_k}{X_k}$ la pente de *f (ß)* augmente de $2\left|X_k\right|, k = 1, \ldots\ldots, n$.

f (ß) étant continue, linéaire par morceaux et convexe, elle atteint son minimum lorsque sa pente change de signe, c'est-à-dire pour $\beta_{(r)}$ tel que :

$$-\sum_{i=1}^{n}\left|X_i\right| + 2\sum_{i=1}^{r-1}\left|X_K\right| < 0 \quad \text{et} \quad -\sum_{i=1}^{n}\left|X_i\right| + 2\sum_{k=1}^{r}\left|X_K\right| \geq 0$$

Dans le cas où : $-\sum_{i=1}^{n}\left|X_i\right| + 2\sum_{k=1}^{r}\left|X\right| = 0$

La solution n'est *pas unique;* dans ce cas, pour toute valeur $\widehat{\beta}$ telle que :

$$\beta_{(r)} \leq \widehat{\beta} \leq \left(\beta_{(r+1)}\right), f(\beta)x \qquad \text{soit minimale.}$$

Notons encore que $\widehat{\beta} = \frac{Y_r}{X_r}$ est appelée ***médiane pondérée*** des $\frac{Y_i}{X_i}$, avec poids $\left|X_i\right|$**.** Ainsi, dans le cas où la droite LAD doit satisfaire le second critère de Boscovich, elle passe par le centroïde des données et par l'une des observations au moins.

C'est à Gauss (1809) que l'on doit une étape importante de la caractérisation des estimateurs LAD. Contrairement à Boscovich, il étudia la méthode consistant à minimiser la somme des erreurs en valeur absolue sans la restriction que leur somme soit nulle (appliquant uniquement le critère 1). A cette époque, Gauss ne semblait d'ailleurs pas savoir que cette restriction avait été introduite par Boscovich, puisqu'il l'attribue à Laplace. D'autre part, Gauss s'intéressa à l'estimation LAD dans un modèle de régression linéaire multiple, en cherchant le vecteur de paramètres $\left(\widehat{\beta}_1,....,\widehat{\beta}_p\right)$ qui minimise :

$$\sum_{i=1}^{n}\left|y_i - x_{i1}\widehat{\beta}_1 - ....x_{ip}\widehat{\beta}_p\right| = \sum_{i=1}^{n}\left|\widehat{e}_i\right|$$

Il mentionna que cette méthode fournit nécessairement *p* résidus nuls et qu'elle n'utilise les autres $(n-p)$ résidus que dans la détermination du choix des ***p*** résidus nuls. De plus, il mentionne que la solution obtenue par cette méthode n'est pas modifiée si la valeur des y_i est augmentée ou diminuée sans que les résidus changent de signe. Gauss remarqua également que la méthode consistant à minimiser

$\sum_{i=1}^{n}\left|\widehat{e}_i\right|$ avec la restriction que $\sum_{i=1}^{n}\left|\widehat{e}_i\right| = 0$ fournit nécessairement *(p* - 1) résidus nuls. Dans le cas de la régression linéaire simple *(p* = 2) traitée par Laplace avec la restriction que la somme des résidus soit nulle, on obtient effectivement une droite passant par l'une des observations, c'est-à-dire qu'un des résidus est nul.

Bloomfield et Steiger (1983) prouvent ce résultat et indiquent qu'il pourrait bien être l'un des premiers en programmation linéaire, mais pas assez profond pour que Gauss le démontre.

Avec l'annonce par Legendre (1805) de la méthode des moindres carrés et son développement par Gauss (1809, 1823, 1828) et Laplace (1812) basé sur la théorie des probabilités, la méthode d'estimation LAD joua un rôle secondaire durant la plus grande partie du *XIX*^{ème} siècle. Ce n'est qu'en 1887 que cette méthode refait surface grâce au travail d'Edgeworth.

En effet, Edgeworth supprima la restriction faite par Boscovich que la somme des résidus soit nulle. Il présenta d'un point de vue géométrique une procédure générale décrite ici dans le cas de la régression linéaire simple *(p = 2).*

Les n observations sont notées $p_1\left(x_1; y_1\right), \ldots, P_n\left(x_n; y_n\right)$.En posant $m_0 = 1$ et en traitant Pm_0 comme l'origine (en soustrayant $\boldsymbol{P_i}$ des autres observations), la procédure de Laplace peut s'appliquer. Or la droite ainsi forcée de passer par Pm_0 contiendra l'une des autres observations, disons Pm_1. Traitant cette nouvelle observation comme l'origine, on trouve une droite passant par une autre observation Pm_2 et ainsi de suite. Cette procédure ne requiert pas plus de $\boldsymbol{r = n - 1}$ étapes, chacune représentant le calcul d'une médiane pondérée, puisqu'elle se termine lorsque $Pm_r = Pm_{r-2}$

Lorsque $p > 2$, l'algorithme, bien que plus compliqué, est analogue et revient à fixer $(p - 1)$ des paramètres à estimer puis à utiliser la procédure de Laplace pour déterminer la valeur optimale du paramètre restant.

Notons finalement que l'algorithme décrit ci-dessus dans le cas de la régression linéaire simple présente certains défauts. Par exemple, il se peut que sur l'une des droites obtenues, il y ait trois observations (d'indice $\boldsymbol{i_1, i_2}$ *et* $\boldsymbol{i_3}$) conduisant la procédure à faire un cycle de la façon suivante $i_1 \rightarrow i_2 \rightarrow i_3 \rightarrow i_1 \ldots\ldots$ sans conduire pour autant à diminuer la somme des résidus en valeur absolue comme l'indique Sposito (1976). Karst (1958) mentionne que cette procédure peut s'arrêter prématurément. C'est le cas lorsque par exemple la droite forcée à passer par p_{i1} contient l'observation p_{i2} et vice-versa, mais n'est pas optimale.

Une implantation en langage Fortran de cet algorithme a été faite par Sadovski (1974) permettant d'éviter les problèmes décrits ci-dessus.

Rhodes (1930) trouva la solution graphique d'Edgeworth difficile à appliquer en pratique. Il proposa alors une procédure itérative que l'on peut résumer ainsi :

(i) Choisir arbitrairement $p - 1$ équations.

(ii) Les utiliser pour éliminer les $p - 1$ premiers paramètres du problème.

(iii) Utiliser la procédure de Laplace pour estimer le paramètre restant.

(iv) Associer l'équation correspondant au point 3 à l'ensemble des $p - 1$ équations.

(v) Si l'ensemble des p équations se répète p fois, on s'arrête. Sinon, on élimine l'équation la plus ancienne et l'on retourne au point 2.

Cette procédure reste cependant difficile à utiliser dans des problèmes pratiques compte tenu des moyens de l'époque.

C'est grâce au travail de Charnes, Cooper et Fergusson (1955) que l'intérêt porté aux estimateurs LAD a été le plus stimulé. Comme alternative à la méthode des moindres carrés, ils proposèrent l'utilisation de la programmation linéaire pour calculer les estimateurs LAD.

Charnes, Cooper et Fergusson (1955) ont montré que le problème de régression linéaire multiple basé sur la norme LAD peut se mettre sous la forme d'un problème de programmation linéaire; pour cela, ils considèrent les résidus comme la différence de deux variables non négatives.

En posant $e_i = u_i - v_i$ où $u_i ; v_i \geq 0$ représentent les déviations positives et négatives respectivement, le problème devient :

$$\text{minimiser} \sum_{i=1}^{n} \left(u_i + v_i \right)$$

$$s.c \sum_{i=1}^{p} x_{ij} + u_i - v_i = y_i$$

$$et \; u_i , v_i \geq 0, \quad i = 1, \ldots\ldots, n$$

Où $\widehat{\beta}_1, \ldots\ldots, \widehat{\beta}_p$ sont sans restriction de signe. Ainsi, le problème de régression linéaire multiple basé sur la norme LAD peut être formulé comme un problème de programmation linéaire avec $2n + p$ variables et n contraintes.

Cette formulation du problème correspond à la forme primale et a pu être résolu en utilisant la méthode du simplexe. Cependant, il a rapidement été reconnu que la structure de ce type de problème pouvait être prise en compte pour améliorer la performance des algorithmes. En effet, Wagner (1959) proposa une formulation de ce problème basée sur le dual, ce qui permit à Barrodale et Young (1966) de mettre au point un algorithme relativement rapide. Par la suite, de nombreuses publications furent consacrées à l'élaboration d'algorithmes de plus en plus performants.

III.2- DECOUVERTE DE L'ESTIMATION LS *:*

La découverte de l'estimation LS *(*méthode des moindres carrés) mérite d'être rappelée ici puisqu'elle fut à l'origine de l'une des plus grandes disputes dans l'histoire de la statistique. Adrien Marie Legendre (1805) publia le premier la méthode des moindres carrés. Il donna une explication claire de la méthode en donnant les équations normales et en fournissant un exemple numérique.

Selon Stigler (1981), Robert Adrain, un américain, publia la méthode vers la fin de l'année 1808 ou au début de l'année 1809. Selon Stigler (1977, 1978), il se pourrait que Robert Adrain

ait "découvert" cette méthode dans l'ouvrage de Legendre (1805). Cependant, quatre ans après la publication de Legendre, Gauss (1809) a le courage de réclamer la paternité de la méthode des moindres carrés, en prétendant l'avoir utilisée depuis 1795. La revendication de Gauss déclencha l'une des plus grandes disputes scientifiques dont les détails sont présentés et résumés dans un article de Plackett (1972). Bien que le doute subsiste, plusieurs faits troublants semblent indiquer que Gauss a effectivement utilisé la méthode des moindres carrés avant 1805. En particulier, Gauss prétend qu'il a parlé de cette méthode à certains astronomes (Olbers, Lindenau et von Zach) avant 1805. De plus, dans une lettre de Gauss datant de 1799, il est fait mention de "ma méthode", sans qu'un nom y soit donné. Il semble difficile de ne pas le croire, vu l'extraordinaire compétence reconnue à Gauss comme mathématicien.

Il reste cependant une question très importante : quelle importance attachait Gauss à cette découverte ? La réponse pourrait être que Gauss, bien que jugeant cette méthode utile, n'a pas réussi à communiquer son importance à ses contemporains avant 1809. En effet, dans sa publication de 1809, Gauss est allé bien plus loin que Legendre dans ses développements autant conceptuels que techniques. C'est dans cet article qu'il lie la méthode des moindres carrés à la loi normale (Gaussienne) des erreurs. Il propose également un algorithme pour le calcul des estimateurs. Son travail a d'ailleurs été discuté par plusieurs auteurs comme Seal (1967), Eisenhart (1968), Goldstine (1977), Sprott (1978) et Sheynin (1979).

Gauss a certainement été le plus grand mathématicien de cette époque, mais c'est Legendre qui a cristallisé l'idée de la méthode des moindres carrés sous une forme compréhensible par ses contemporains.

IV.1- INTRODUCTION :

La méthode des moindres écarts en valeurs absolues a plusieurs notations, comme LAD (Least absolute deviations), MAD (Minimum absolute deviations), LAR (Least absolute residuals), la norme L_1 et LAV (Least absolute values), dans ce travail nous utiliserons la notation LAD.

IV.2- METHODES D'ESTIMATIONS :

Il existe plusieurs méthodes d'estimation des paramètres; les méthodes considérées par la suite, concernent principalement l'estimation LAD, et l'estimation LS.

IV.2.1- L'ESTIMATION LAD :

La méthode des moindres écarts en valeurs absolues, dite méthode LAD**.** (En anglais : least absolute déviations), est l'une des principales alternatives à la méthode des moindres carrés lorsqu'il s'agit d'estimer les paramètre d'un modèle de régression. Elle a été introduite presque cinquante ans avant la méthode des moindres carrés, en 1757 par Roger Joseph Boscovich. Il utilisa cette procédure dans le but de concilier des mesures incohérentes dans le cadre de l'estimation de la forme de la terre. Pierre Simon Laplace adopta cette méthode trente ans plus tard, mais elle fut ensuite éclipsée par la méthode des moindres carrés développées par Legendre et Gauss, la popularité de la méthode des moindres carrés repose principalement sur la simplicité des calculs Mais aujourd'hui, avec les progrès de l'informatique, la méthode LAD, peut être utilisée presque aussi simplement.

IV.2.1.1- LE MODELE DE REGRESSION LINEAIRE SIMPLE :

Le modèle de régression simple est donné par :

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad (4)$$

A partir des n observations (x_i, y_i), il s'agit d'estimer les paramètres β_0 et β_1 du modèle par $\widehat{\beta_0}$ et $\widehat{\beta_1}$. On obtient ainsi des valeurs estimées :

$$\widehat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

qui ne doivent pas être trop éloignées de y_i, du moins si le modèle est correct. Cela signifie que les estimateurs $\widehat{\beta_0}$ et $\widehat{\beta_1}$ doivent être choisis de telle sorte que les résidus du modèle :

$$e_i = y_i - \widehat{y}_i$$

soient petits .Le critère utilisé par la méthode des moindres carrés est de minimiser la somme des carrés des résidus :

$$\Sigma e_i^2 = \Sigma \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2$$

(5)

Le critère utilisé par la méthode LAD est de minimiser la somme des valeurs absolues des résidus :

$$\Sigma \left| e_i \right| = \Sigma \left| y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i \right|$$

(6) Il s'agit donc ici de choisir $\hat{\beta}_0$ et $\hat{\beta}_1$ de telle sorte que (6) soit minimale .Dans un certain sens, il peut paraître plus naturel de vouloir minimiser (6) plutôt que (5). Pourtant, le calcul des estimateurs LAD, est plus complexe. En particulier, on ne dispose pas de formule explicite pour calculer $\hat{\beta}_0$ et $\hat{\beta}_1$ comme c'est le cas avec la méthode des moindres carrés, puisque la fonction valeur absolue n'est pas dérivable. Les estimateurs LAD, peuvent toutefois être calculés par des algorithmes itératifs.

IV.2.1.2- TEST D'HYPOTHESE SUR LA PENTE β_1:

On va voir dans cette section comment tester l'hypothèse nulle :

$$H_0 : \beta_1 = 0$$

contre l'hypothèse alternative :

$$H_1 : \beta_1 \neq 0$$

en utilisant la régression LAD.

Il s'agit tout d'abord d'estimer les paramètres du modèle β_0 et β_1 par les estimateurs LAD $\hat{\beta}_0$ et $\hat{\beta}_1$. On calcule alors les valeurs estimées LAD.

$$\widehat{y}_i = \beta_0 + \beta_1 x_i$$

Et les résidus LAD :

$$e_i = y_i - \widehat{y}_i$$

Comme la droite LAD passe par deux points, on a 2 résidus nuls (si l'on n'est pas dans un cas de dégénérescence). On classe les $m = n - 2$ résidus non nuls par ordre croissant de telle

sorte que e_1 soit le plus petit résidus non nul, e_2 le second plus petit résidu non nul, et ainsi de suite.

Soit k_1, l'entier le plus proche de $(m+1)/2 - \sqrt{m}$ et soit k_2, l'entier le plus proche de $(m+1)/2 + \sqrt{m}$, on définit :

$$\hat{\tau} = \frac{\sqrt{m}\left(e_{k_2} - e_{k_1}\right)}{4}$$

(7) L'écart type de l'estimateur LAD $\hat{\beta}_1$ est alors estimé par :

$$s(\hat{\beta}_1) = \frac{\hat{\tau}}{\sqrt{\sum (x_i - \overline{x})^2}}$$

(8)

Et la statistique pour tester H_0 est donnée par :

$$t_{LAD} = \frac{|\hat{\beta}_1|}{s(\hat{\beta}_1)}$$

(9)

On rejette H_0 à un seuil de signification α si cette statistique t_{LAD} est plus grande que la valeur critique $t_{\alpha/2,n-2}$ que l'on trouve dans une table de Student.

On remarque que cette procédure est très similaire à celle utilisée par la méthode des moindres carrés, où rappelons-le, l'écart type de l'estimateur de β_1 était estimé par :

$$s(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{\sum (x_i - \overline{x})^2}}$$

(10) Il s'agit donc de la même formule que (8) sauf que le $\hat{\tau}$ défini par (7) est remplacé dans (10) par l'estimateur habituel $\hat{\sigma}$ de l'écart type σ des erreurs ε_i du modèle le rapport entre les écarts type de l'estimateur LAD et celui des moindres carrés est donc égal à τ/σ.

IV.2.2- L'ESTIMATION LS :

IV.2.2.1- LE MODELE DE REGRESSION LINEAIRE SIMPLE :

Considérons un échantillon de n observations paires $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$. Nous avons vu que le modèle de régression linéaire, appelé ici *modèle de régression simple*, suppose, pour tout $i = 1, ..., n$, la relation:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Où les ε_i sont des quantités aléatoires inobservables, que nous appellerons dorénavant les *erreurs.* Ainsi les y_i sont des variables aléatoires (car elles dépendent des ε_i), alors que les x_i sont considérés comme des nombres fixés. En supposant l'espérance des ε_i nulle, on a ainsi:

$$\begin{aligned} E(y_i) &= \beta_0 + \beta_1 x_i + E(\varepsilon_i) \\ &= \beta_0 + \beta_1 x_i \\ Var(y_i) &= Var(\beta_0 + \beta_1 x_i + \varepsilon_i) = Var(\varepsilon_i) \end{aligned}$$

Afin de pouvoir faire de l'inférence sur la droite de régression :

$$\mu_y(x) = \beta_0 + \beta_1 x$$

à partir de la droite des moindres carrés :

$$\widehat{y}(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x$$

on fait généralement les hypothèses supplémentaires suivantes sur ces variables aléatoires ε_i .

La variance de ε_i est égale à une quantité σ^2 (inconnue) ne dépendant pas de x_i. On a donc pour tout $i = 1, ..., n$:

$$Var(\varepsilon_i) = Var(y_i) = \sigma^2$$

- Les ε_i sont indépendantes.
- Les ε_i sont normalement distribuées.

Ces trois conditions reviennent à dire que les variables aléatoires ε_i sont indépendantes et identiquement distribuées (i.i.d.) selon une loi normale d'espérance nulle et de variance σ^2.
On note parfois:

$$\varepsilon_i \; i.i.d \; .\mathrm{N}(0, \sigma^2)$$

(11)

Remarquons que : la normalité des ε_i implique la normalité des y_i (un y_i s'obtenant d'un ε_i par une simple addition de la constante $(\beta_0 + \beta_1 x_i)$ de même que l'indépendance des ε_i implique l'indépendance des y_i. On a en effet, si $i \neq j$:

$$Cov(y_i, y_j) = Cov(\beta_0 + \beta_1 x_i + \varepsilon_i, \beta_0 + \beta_1 x_j + \varepsilon_j)$$
$$= Cov(\varepsilon_i, \varepsilon_j) = 0$$

Rappelons à ce sujet que l'indépendance entre deux variables dont la distribution conjointe est normale bivariée est équivalente à la nullité de leur covariance.

Lorsque l'on utilise un modèle de régression simple, on suppose donc que l'on a tiré un échantillon de ε_i distribués selon (11). Cependant, on n'observe pas ces ε_i **,** on observe à la place les variables aléatoires y_i définies par (4) à partir de ces ε_i, pour des valeurs x_i fixées par avance et de paramètres β_0 et β_1, inconnus.

IV.2.2.2- ESTIMATION DE LA VARIANCE DES ERREURS :

Cette quantité est toutefois inconnue et doit être estimée.

Si les erreurs ε_i pouvaient être observées, un estimateur non biaisé de σ^2 serait donné par la formule habituelle :

$$\frac{\sum (\varepsilon_i - \bar{\varepsilon})^2}{n - 1}$$

où $\bar{\varepsilon}$ serait la moyenne des ε_i . Or ces quantités ne sont pas observables. Mais nous avons vu que les erreurs ε_i peuvent être estimées par les résidus du modèle :

$$e_i = y_i - \hat{y}_i$$

où les :

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$
$$= \bar{y} + \hat{\beta}_1 (x_i - \bar{x})$$

sont les valeurs estimées des y_i. Ainsi, nous allons estimer σ^2 en utilisant la somme des carrés des résidus comme estimateur de la somme des carrés des erreurs :

$$\sum (e_i - \bar{e})^2 = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2$$

où $\bar{e}$ désigne la moyenne des e_i (on a vu que la somme, donc la moyenne, des résidus est nulle).

or, comme on a vu que

$$\sum (y_i - \widehat{y}_i)^2 = \sum y_i^2 - \sum \widehat{y}_i^2$$

on a :

$$\begin{aligned} E\left(\sum (y_i - \widehat{y}_i)^2\right) &= \sum E(y_i^2) - \sum E(\widehat{y}_i^2) \\ &= \sum \left(Var(y_i) + E^2(y_i)\right) - \sum \left(Var(\widehat{y}_i) + E^2(\widehat{y}_i)\right) \end{aligned}$$

or, on a :

$$\begin{aligned} \mathrm{E}(\mathrm{y}) &= \mathrm{E}\left(\widehat{\beta}_0 + \widehat{\beta}_1 x_i\right) \\ &= E\left(\widehat{\beta}_0\right) + x_i E\left(\widehat{\beta}_1\right) \\ &= \beta_0 + x_i \beta_1 \\ &= E\left(y_i\right) \end{aligned}$$

on obtient ainsi :

$$\begin{aligned} E\left(\sum (y_i - \widehat{y}_i)^2\right) &= \sum \left(Var(y_i) - \sum Var(\widehat{y}_i)\right) \\ &= n\sigma^2 - \sum Var(\widehat{y}_i) \end{aligned}$$

d'autre part, on a :

$$\begin{aligned} Var(\widehat{y}_i) &= Var\left(\widehat{\beta}_0 + \widehat{\beta}_1 x_i\right) \\ &= Var\left(\widehat{\beta}_0\right) + x\, Var\left(\widehat{\beta}_1\right) + 2x_i Cov\left(\widehat{\beta}_0, \widehat{\beta}_1\right) \\ &= \frac{\sigma^2 \sum x_j^2}{n\sum\left(x_j - \bar{x}\right)^2} + \frac{x_i^2 \sigma^2}{\sum\left(x_j - \bar{x}\right)^2} - \frac{2x_i \bar{x}\sigma^2}{\sum\left(x_j - \bar{x}\right)^2} \\ &= \frac{\sigma^2}{\sum\left(x_j - \bar{x}\right)^2}\left(\frac{\sum x_j^2}{n} + x_i^2 - 2x_i\bar{x}\right) \\ &= \frac{\sigma^2}{\sum\left(x_j - \bar{x}\right)^2}\left(\frac{\sum x_j^2}{n} - \frac{n\bar{x}^2}{n} + x_i^2 - 2x_i\bar{x} + \bar{x}^2\right) \\ &= \frac{\sigma^2}{\sum\left(x_j - \bar{x}\right)^2}\left(\frac{\sum\left(x_j - \bar{x}\right)^2}{n} + \left(x_i - \bar{x}\right)^2\right) \\ &= \sigma^2\left(\frac{1}{n} + \frac{\left(x_i - \bar{x}\right)^2}{\sum\left(x_j - \bar{x}\right)^2}\right) \end{aligned}$$

on obtient ainsi :

$$E\left(\sum\left(x_j-\overline{x}\right)^2\right)=N\,\sigma^2-\sigma^2\left(\frac{n}{n}+\frac{\sum\left(x_i-\overline{x}\right)^2}{\sum\left(x_j-\overline{x}\right)^2}\right)$$

$$=\sigma^2(n-2)$$

on peut ainsi définir un estimateur sans biais de σ^2 en posant :

$$s^2=\frac{\sum(y_i-\widehat{y}_i)^2}{n-2}=\frac{SC_{res}}{n-2}=\frac{SC_{tot}-SC_{reg}}{n-2}=\frac{\sum(y_i-\widehat{y}_i)^2-\widehat{\beta}_1{}^2\sum\left(x_i-\overline{x}\right)^2}{n-2}$$

on estime par ailleurs l'écart type des erreurs par :

$$s=\sqrt{\frac{\sum(y_i-\widehat{y}_i)^2}{n-2}}\,.$$

IV.2.2.3- TEST SUR LA PENTE :

L'estimateur $\widehat{\beta}_1$ est normalement distribué, d'espérance β_1 et de variance que l'on note ici par :

$$\sigma^2\left(\widehat{\beta}_1\right)=\frac{\sigma}{\sum\left(x_i-\overline{x}\right)}$$

il s'ensuit que la quantité :

$$\frac{\widehat{\beta}_1-\beta_1}{\sigma\left(\widehat{\beta}_1\right)}$$

où $\sigma\left(\widehat{\beta}_1\right)$ désigne la racine carrée de $\sigma^2\left(\widehat{\beta}_1\right)$, suit une loi normale standardisée. Or cette quantité ne peut pas être utilisée pour un problème de test d'hypothèses puisque on ne connaît pas la valeur de σ. En pratique, on estime $\sigma^2\left(\widehat{\beta}_1\right)$ par :

$$s^2\left(\widehat{\beta}_1\right)=\frac{s^2}{\sum\left(x_i-\overline{x}\right)^2}$$

où s^2 est l'estimateur sans biais de σ^2 défini ci-dessus. On peut alors montrer que la quantité :

$$\frac{\widehat{\beta}_1-\beta_1}{s\left(\widehat{\beta}_1\right)}$$

où $s\left(\widehat{\beta}_1\right)$ désigne la racine carrée de $s^2\left(\widehat{\beta}_1\right)$, suit une loi de Student avec $(n-2)$ degrés de liberté.

Il s'ensuit que dans un problème de test d'hypothèses bilatéral où l'on désire tester l'hypothèses nulle :

$$H_0 : \beta_1 = b_1$$

contre l'hypothèse alternative :

$$H_t : \beta_1 \neq b_1$$

on peut utiliser la statistique :

$$t_c = \frac{\hat{\beta}_1 - b_1}{s\left(\hat{\beta}_1\right)}$$

on rejette au seuil de signification α si :

$$| t_c | \rangle t_{(\alpha/2\,n-2)}$$

où la valeur critique $t_{(\alpha/2n-2)}$ est le $(1-\alpha/2)$ quantile d'une loi de Student avec $(n-2)$ degrés de liberté que l'on trouve dans une table de Student.

Un test d'hypothèse particulièrement intéressant est le test de l'hypothèse nulle :

$$H_0 : \beta_1 = 0$$

contre l'hypothèse alternative :

$$H_1 : \beta_1 \neq 0$$

En effet, le non-rejet de l'hypothèse nulle implique un modèle avec un seul paramètre :

$$y_i = \beta_0 + \varepsilon_i$$

par contre si cette hypothèse H_0 est rejetée, c'est-à-dire si :

$$| t_c | = \frac{\hat{\beta}_1 - b_1}{s\left(\hat{\beta}_1\right)} \rangle t_{(\alpha/2n-2)}$$

on dit que la relation entre les x_i et les y_i est significative au seuil de signification α.

IV.2.2.4- INTERVALLE DE CONFIANCE :

Un intervalle de confiance au niveau $(\alpha-1)$ pour un paramètre β_j est défini par :

$$\left[\hat{\beta}_j - t_{(\alpha/2,n-p)}.s\left(\hat{\beta}_j\right); \hat{\beta}_j + t_{(\alpha/2,n-p)}.s\left(\hat{\beta}_j\right)\right]$$

c'est -à- dire par :

$$\hat{\beta}_j \pm t_{(\alpha/2,n-p)}.s\left(\hat{\beta}_j\right)$$

cet intervalle est construit de telle sorte qu'il contienne le paramètre inconnu β_j avec une probabilité de $(\alpha-1)$.

IV.2.2.5- COEFFICIENT DE CORRELATION :

Le coefficient de corrélation est défini comme suite :

$$r_{xy} = \frac{\sum\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)}{\sqrt{\left(x_i - \overline{x}\right)^2}\sqrt{\left(y_i - \overline{y}\right)^2}} = \frac{\sum\left(\widehat{y}_i - \overline{y}\right)\left(y_i - \overline{y}\right)}{\sqrt{\left(\widehat{y}_i - \overline{y}\right)^2}\sqrt{\left(y_i - \overline{y}\right)^2}} = r_{\widehat{y}y}$$

où :

$$\overline{x} = \sum x_i \Big/ n$$

$$\overline{y} = \sum y_i \Big/ n$$

IV.2.2.6- LIEN ENTRE LE COEFFICIENT DE CORRELATION ET LE COEFFICIENT DE DETERMINATION :

Il faut mettre en évidence la relation fondamentale qui existe entre le coefficient de corrélation et le coefficient de détermination R^2, qui donne le pourcentage de la variance totale des Y_i expliquée par le modèle de régression simple et le coefficient de corrélation r_{xy}. Rappelons à ce propos que l'on avait :

$$R^2 = \frac{\left(\widehat{y}_i - \overline{y}\right)^2}{\left(y_i - \overline{y}\right)^2}$$

avec :

$$\begin{aligned}\widehat{y}i &= \widehat{\beta}_0 + \widehat{\beta}_1 x_i \\ &= \overline{y} + \widehat{\beta}_1\left(x_i - \overline{x}\right)\end{aligned}$$

ce qui donne :

$$R^2 = \frac{\left(\widehat{y}_i + \widehat{\beta}_1\left(x_i - \overline{x}\right) - \overline{y}\right)^2}{\left(y_i - \overline{y}\right)^2}$$

$$= \widehat{\beta}_1^2 \frac{\sum\left(x_i - \overline{x}\right)^2}{\sum\left(y_i - \overline{y}\right)^2}$$

En reprenant les notions introduites ci-dessus, on a ainsi :

$$R^2 = \widehat{\beta}_1^2 \frac{s_x^2}{s_y^2}$$

et donc :

$$R^2 = r_{xy}^2$$

IV.3- COMPARAISON DE DEUX DROITES DE REGRESSION :

Soit $Y = a + bx = \overline{y} + b(x - \overline{x})$ l'équation d'une droite de régression, les deux droites comparées seront distinguées par les indices Y_{lad} et Y_{LS}, il faut calculer les coefficients $\overline{y}$ et b, ainsi que les variances résiduelles S. Nous nous plaçons dans le cas où l'un des nombres de couples de résultats N_I ou N_{II} est inférieur à 30.

IV.4- COMPARAISON DES ORDONNEES DE DEUX DROITES AU POINT MOYEN :

Choisir une valeur de x_0 appartenant au domaine de toutes les droites et aussi prés que possible du point moyen. Calculer la valeur de Y correspondant sur chaque droite à l'abscisse x_0 et comparer les valeurs de Y de la façon indiquée ci-*après* ; on calcule une estimation S de la variance résiduelle commune aux deux droites en faisant une moyenne pondérée de $S^2_{2(I)}$ et $S^2_{2(II)}$ suivant le nombre de degrés de liberté :

$$S_2^2 = \frac{(N_I - 2)S^2_{2(I)} + (N_{II} - 2)S^2_{2(II)}}{N_I + N_{II} + 4}$$

(12)

Cette estimation est utilisée pour calculer S^2_{YI} et S^2_{YII} par la formule :

$$S^2_{YI} = S_2^2 \left[\frac{1}{N_I} + \frac{(x_0 - \overline{x}_I)}{\sum (x_{iI} - \overline{x}_I)^2} \right]$$

(13)

La formule ci-dessous permet d'en déduire une valeur t**.**

$$t = \frac{|Y_I - Y_{II}|}{\sqrt{S^2_{Y_I} - S^2_{Y_{II}}}}$$

(14)

Cette variable suit la loi de Student à (NI +NII - 4) degrés de liberté dans le cas où les deux droites de régression vraies sont confondues.

La valeur expérimentale t est donc comparée à la limite $t_{1-\alpha/2}$ donnée par la table de Student.

Si la valeur de t est supérieure à la limite donnée par la table, on peut admettre, au niveau de confiance choisi, que les deux droites se déplacent parallèlement l'une par rapport à l'autre.

IV.5- COMPARAISON DES PENTES DES DEUX DROITES :

Utiliser l'estimation commune (12) de la variance résiduelle pour calculer

$$S^2_{b(I)} = \frac{S^2_{2(I)}}{\sum (x_{iI} - \overline{x}_I)^2}$$

(15)

En déduire t par la formule :

$$t = \frac{|b_I - b_{II}|}{\sqrt{S^2_{b_I} - S^2_{b_{II}}}}$$

(16)

Cette variable suit la loi de Student à (NI+NII - 4) degrés de liberté dans le cas où les deux droites de régression vraies sont confondues. Comparer la valeur de **t** à la limite $t_{1-\alpha/2}$ donnée par la table de Student. Si le test est significatif, on peut conclure qu'il y a rotation d'une droite par rapport à l'autre.

IV.6- COMPARAISON DES VARIANCES RESIDUELLES :

Comparer les valeurs S^2_2 par le test de Snedecor en formant le rapport :

$$F = \frac{S^2_{2I}}{S^2_{2II}}$$

Comparer ce rapport expérimental aux limites données par la table de Snedecor pour $Y_I = (N_I - 2)$ et $Y_{II} = (N_{II} - 2)$ soit, au niveau de confiance $(1-\alpha)$

$$F_{1-\alpha/2}(N_I, N_{II}) \quad \text{et} \quad F_{\alpha/2} = \frac{1}{F_{1-\alpha/2}(N_I, N_{II})}$$

L'hypothèse contrôlée : $\sigma^2_I = \sigma^2_{II}$

sera refusée si : $F > F_{1-\alpha/2}$

ou si : $F < F_{1-\alpha}$

ne sera pas refusée si : $F_\alpha < F < F_{1-\alpha/2}$.

V.1.1- Droites de régression LS et LAD_1 :

Après l'entrée des données dans l'application programmée en langage Pascal et exécutable sous Windows, nous avons calculé les paramètres du modèle LAD_1 donnés par **l'algorithme des moindres carrés re-pondérés itérativement**, le R^2 a été calculé via Excel selon la méthode décrite au chapitreIV, et les paramètres du modèle LS ont été calculés via Minitab, le tableau 2 reproduit les paramètres des modèles LS et LAD_1 :

Tableau 2 : Paramètres des modèles LS et LAD_1.

	m	**b**	**R^2**	**Variance**	$\sum\|e_i\|$
LS	0.834	- 2.070	95.20%	σ^2_{LS} =0.105	6.98682
LAD_1	0.841	- 2.149	97.54%	τ^2_{LAD} =0.089	6.55319

La figure 4 reproduit les représentations graphiques des deux droites de régression LS et LAD_1 réalisées sous Excel :

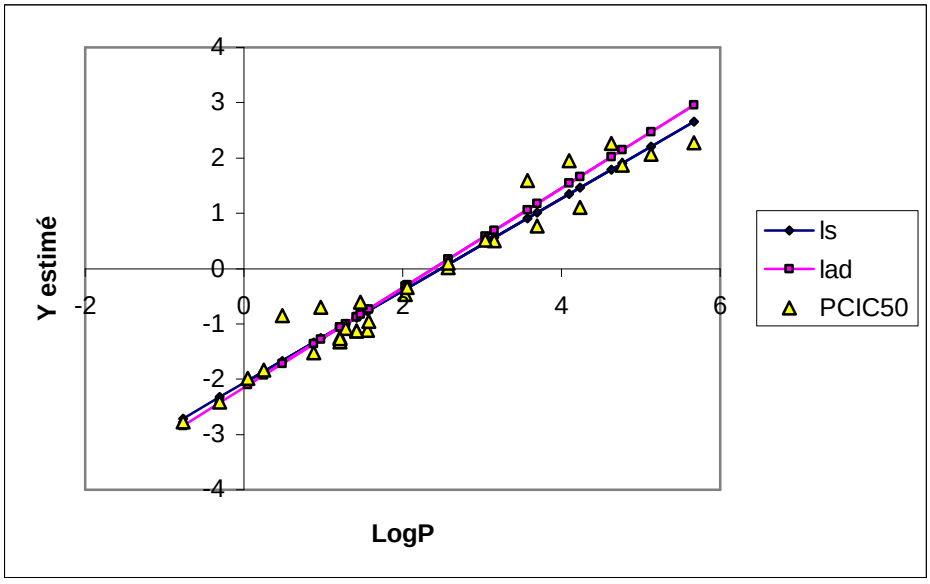


Figure 4 : Droites de régression LS, et LAD_1.

Puis nous avons calculé par les deux méthodes LS et LAD_1, les valeurs estimées de la toxicité pCIC50 et les erreurs résiduelles correspondant à chaque coefficient de partage logP des composés. Le tableau 3 reproduit la toxicité estimée et l'erreur résiduelle obtenues par régressions LS et LAD_1 pour chaque composé, la figure 5 reproduit l'histogramme des erreurs résiduelles.

Tableau 3 : Erreurs résiduelles pour la LS et la LAD_1.

N°	Composé	pCIC50	logP	$\hat{Y}_{LS}$	$\hat{Y}_{LAD1}$	$e_{i(LS)}$	$e_{i(LAD1)}$
1	Méthanol	-2.77	-0.77	-2.71018	-2.79657	-0.05982	0.02657
2	Ethanol	-2.41	-0.31	-2.32654	-2.40971	-0.08346	-0.00029
3	Propan-1-ol	-1.84	0.25	-1.8595	-1.93875	0.0195	0.09875
4	Butan-1-ol	-1.52	0.88	-1.33408	-1.40892	-0.18592	-0.11108
5	Pentan-1-ol	-1.12	1.56	-0.76696	-0.83704	-0.35304	-0.28296
6	Hexan-1-ol	-0.47	2.03	-0.37498	-0.44177	-0.09502	-0.02823
7	Heptan-1-ol	0.02	2.57	0.07538	0.01237	-0.05538	0.00763
8	Octan-1-ol	0.5	3.15	0.5591	0.50015	-0.0591	-0.00015
9	Nonan-1-ol	0.77	3.69	1.00946	0.95429	-0.23946	-0.18429
10	Decan-1-ol	1.1	4.23	1.45982	1.40843	-0.35982	-0.30843
11	Undecan-1-ol	1.87	4.77	1.91018	1.86257	-0.04018	0.00743
12	Dodecan-1-ol	2.07	5.13	2.21042	2.16533	-0.14042	-0.09533
13	Tridecan-1-ol	2.28	5.67	2.66078	2.61947	-0.38078	-0.33947
14	Propan-2-ol	-1.99	0.05	-2.0263	-2.10695	0.0363	0.11695
15	Pentan-2-ol	-1.25	1.21	-1.05886	-1.13139	-0.19114	-0.11861
16	Pentan-3-ol	-1.33	1.21	-1.05886	-1.13139	-0.27114	-0.19861
17	2-methylbutan-1-ol	-1.13	1.42	-0.88372	-0.95478	-0.24628	-0.17522
18	3-methylbutan-1-ol	-1.13	1.42	-0.88372	-0.95478	-0.24628	-0.17522
19	3-methylbutan-2-ol	-1.08	1.28	-1.00048	-1.07252	-0.07952	-0.00748
20	(ter) pentanol	-1.27	1.21	-1.05886	-1.13139	-0.21114	-0.13861
21	(neo) pentanol	-0.96	1.57	-0.75862	-0.82863	-0.20138	-0.13137
22	1-propylamine	-0.85	0.48	-1.66768	-1.74532	0.81768	0.89532
23	1-butylamine	-0.7	0.97	-1.25902	-1.33323	0.55902	0.63323
24	1-amylamine	-0.61	1.47	-0.84202	-0.91273	0.23202	0.30273
25	1-hexylamine	-0.34	2.06	-0.34996	-0.41654	0.00996	0.07654
26	1-heptylamine	0.1	2.57	0.07538	0.01237	0.02462	0.08763
27	1-octylamine	0.51	3.04	0.46736	0.40764	0.04264	0.10236
28	1-nonylamine	1.59	3.57	0.90938	0.85337	0.68062	0.73663
29	1-decylamine	1.95	4.1	1.3514	1.2991	0.5986	0.6509
30	1-undecylamine	2.26	4.63	1.79342	1.74483	0.46658	0.51517
					$\sum\|e_i\|$	**6.98682**	**6.55319**

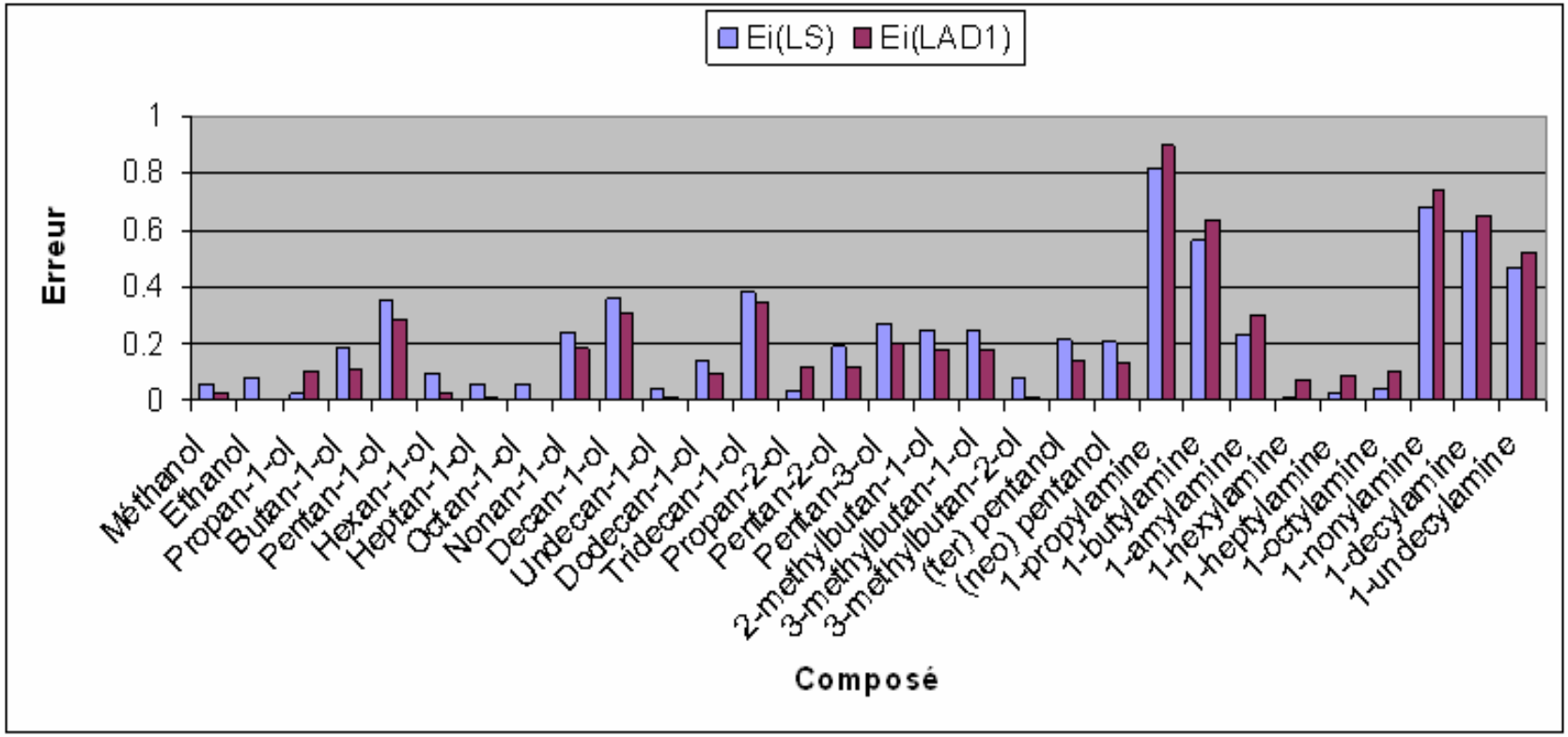


Figure 5 : Histogramme des erreurs résiduelles pour la LS et la LAD_1.

LS: $\hat{Y}_{LS}$= 0.834 logP-2.070 R^2=95.20% σ^2_{LS} =0.105

LAD_1: $\hat{Y}_{LAD1}$=0.841 logP-2.149 R^2=97.54% σ^2_{LAD} =0.089

V.1.2- TESTS STATISTIQUES :

Nous avons appliqué sur les paramètres des deux modèles linéaires LS et LAD_1, les tests statistiques que nous avons énoncés, au chapitreVI, pour α=0.05.

V.1.2.1- TEST SUR LA PENTE LAD_1 :

L'hypothèse nulle: $\mathbf{H_0 : B_{LAD} = 0.}$

L'hypothèse est acceptée (**AH**) si $\mathbf{t_{calc} < t_{obs}}$, et rejetée (**RH**) si $\mathbf{t_{calc} > t_{obs}}$.

Les calculs ont mené aux résultats ci-après :

	$\mathbf{B_{LAD}}$	$\mathbf{t_{calc}}$	$\mathbf{t_{obs}}$	$\mathbf{H_0}$
Méthode des moindres carrés re-pondérés itérativement.	0.841	25.603	2.015	R H_0

L'hypothèse est rejetée, alors la pente est significativement différente de zéro.

V.1.2.2- COMPARAISON DES DEUX PENTES OBTENUES PAR LS ET LAD_1 :

L'hypothèse nulle est: $\mathbf{H_0 : B_{LAD} = B_{LS}}$

L'hypothèse est acceptée (**AH**) si $\mathbf{t_{calc} < t_{obs}}$, et rejetée (**RH**) si $\mathbf{t_{calc} > t_{obs}}$.

Les calculs ont mené aux résultats ci-après :

	$\mathbf{B_{LAD}}$	$\mathbf{B_{LS}}$	$\mathbf{t_{calc}}$	$\mathbf{t_{obs}}$	$\mathbf{H_0}$
Méthode des moindres carrés re-pondérés itérativement.	0.841	0.834	0.145	2.015	AH_0

L'hypothèse est acceptée, alors les pentes des deux modèles LS et LAD sont significativement les mêmes.

V.1.2.3- COMPARAISON DES DEUX ORDONNEES OBTENUES PAR LS ET LAD_1 :

L'hypothèse nulle est: $\mathbf{H_0 : Y_{LAD} = Y_{LS}}$

L'hypothèse est acceptée (**AH**) si $\mathbf{t_{calc} < t_{obs}}$, et rejetée (**RH**) si $\mathbf{t_{calc} > t_{obs.}}$

Les calculs ont mené aux résultats ci-après :

	$\mathbf{Y_{LAD}}$	$\mathbf{Y_{LS}}$	$\mathbf{t_{calc}}$	$\mathbf{t_{obs}}$	$\mathbf{H_0}$
Méthode des moindres carrés re-pondérés itérativement.	-0.417	-0.350	0.129	2.0147	A H_0

L'hypothèse est acceptée, alors les ordonnées à l'origine des deux modèles LS et LAD sont significativement les mêmes.

V.1.2.4- COMPARAISON DES DEUX VARIANCES LS ET LAD_1 :

L'hypothèse nulle est : $\mathbf{H_0 :}\ \sigma^2_{LAD} = \sigma^2_{LS}$

L'hypothèse est acceptée (**AH**) si : $F_{\alpha/2} < F_{calc} < F_{1-\alpha/2}$, et rejetée (**RH**) si : $F_{calc} > F_{1-\alpha/2}$.

Les calculs ont mené aux résultats ci-après :

	σ^2_{LAD}	σ^2_{LS}	F_{calc}	$F_{\alpha/2}$	$F_{1-\alpha/2}$	H_0
Méthode des moindres carrés re-pondérés itérativement.	0.089	0.105	1.172	0.53	1.87	AH_0

L'hypothèse est acceptée, alors les variances des deux modèles LS et LAD sont significativement les mêmes.

V.1.3- Interprétation des résultats :

Le test d'hypothèse nulle $H_0 : B_{LAD}=0$ a été rejeté, ce qui révèle que la pente du modèle LAD est significativement différente de zéro, alors le modèle linéaire peut expliquer la variabilité de la toxicité pCIC50 en fonction du coefficient de partage (Octanol /Eau) logP.

Les tests statistiques de comparaison entre les modèles LS et LAD font ressortir l'absence de translation ou de rotation de l'une des droites de régression par rapport à l'autre. En outre, la comparaison des variances permet de mettre en évidence que les deux modèles ont la même dispersion.

En observant l'histogramme et le tableau des erreurs résiduelles, l'algorithme **des moindres carrés itérativement re-pondérés** donne un modèle LAD avec moins d'erreurs pour 19 molécules d'alcools, qui représentent 63% de la population, mais plus d'erreurs pour les 9 amines, le propan-1-ol et le propan-2-ol. Les amines, qui représentent 30% de la population provoquent 61% d'erreurs dans le modèle LAD et 49% d'erreurs dans le modèle LS, ce qui nous fait dire que les deux modèles linéaires simples n'arrivent pas à expliquer assez le comportement des amines.

Le modèle LS explique 95.20% de la variabilité de la toxicité pCIC50 en fonction du coefficient de partage (Octanol /Eau) logP avec $\sum|e_i| = 6.98682$.

Le modèle LAD explique 97.54% de la variabilité de la toxicité pCIC50 en fonction du coefficient de partage (Octanol /Eau) logP avec $\sum|e_i| = 6.55319$.

Le modèle LAD donné par l'algorithme **des moindres carrés itérativement re-pondérés *'Iteratively Re-weighted Least Squares'***, donne 6% de moins de $\sum|e_i|$ par rapport au modèle LS.

Ces résultats prouvent que l'application de cet algorithme a été bénéfique pour le perfectionnement du modèle linéaire simple.

V.2.1- Droites de régression LS et LAD_2 :

Après l'entrée des données dans l'application programmée en langage Pascal et exécutable sous Windows, nous avons calculé les paramètres du modèle LAD_2 donnés par **l'algorithme itératif de base**. Le R^2 et τ^2_{LAD} ont été calculés via Excel selon la méthode décrite au chapitreIV, et les paramètres du modèle LS ont été calculés via Minitab, le tableau 4 reproduit les paramètres des modèles LS et LAD_2 :

Tableau 4 : Paramètres des modèles LS et LAD_2.

	m	**b**	**R^2**	**Variance**	$\sum\|e_i\|$
LS	0.834	- 2.070	95.20%	σ^2_{LS} =0.105	6.98682
LAD_2	0.841	- 2.149	97.54%	τ^2_{LAD} =0.089	6.55319

La figure 6 reproduit les représentations graphiques des deux droites de régression LS et LAD_2 réalisées sous Excel :

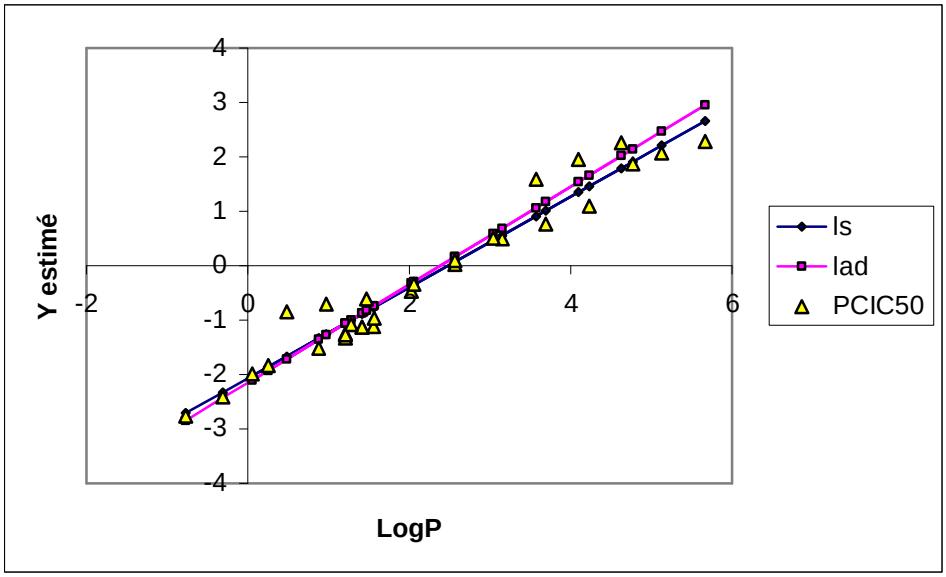


Figure 6 : Droites de régression LS, et LAD_2**.**

Puis nous avons calculé par les deux méthodes LS et LAD_2, les valeurs estimées de la toxicité pCIC50 et les erreurs résiduelles correspondant à chaque coefficient de partage logP des composés. Le tableau 5 reproduit la toxicité estimée et l'erreur résiduelle obtenues par régressions LS et LAD_2 pour chaque composé; la figure 7 reproduit l'histogramme des erreurs résiduelles.

Tableau 5 : Erreurs résiduelles pour la LS et LAD_2

N°	Composé	pCIC50	logP	$\hat{Y}_{LS}$	$\hat{Y}_{LAD2}$	$\lvert e_{i(LS)} \rvert$	$\lvert e_{i(LAD2)} \rvert$
1	Méthanol	-2.77	-0.77	-2.71018	-2.79657	-0.05982	0.02657
2	Ethanol	-2.41	-0.31	-2.32654	-2.40971	-0.08346	-0.00029
3	Propan-1-ol	-1.84	0.25	-1.8595	-1.93875	0.0195	0.09875
4	Butan-1-ol	-1.52	0.88	-1.33408	-1.40892	-0.18592	-0.11108
5	Pentan-1-ol	-1.12	1.56	-0.76696	-0.83704	-0.35304	-0.28296
6	Hexan-1-ol	-0.47	2.03	-0.37498	-0.44177	-0.09502	-0.02823
7	Heptan-1-ol	0.02	2.57	0.07538	0.01237	-0.05538	0.00763
8	Octan-1-ol	0.5	3.15	0.5591	0.50015	-0.0591	-0.00015
9	Nonan-1-ol	0.77	3.69	1.00946	0.95429	-0.23946	-0.18429
10	Decan-1-ol	1.1	4.23	1.45982	1.40843	-0.35982	-0.30843
11	Undecan-1-ol	1.87	4.77	1.91018	1.86257	-0.04018	0.00743
12	Dodecan-1-ol	2.07	5.13	2.21042	2.16533	-0.14042	-0.09533
13	Tridecan-1-ol	2.28	5.67	2.66078	2.61947	-0.38078	-0.33947
14	Propan-2-ol	-1.99	0.05	-2.0263	-2.10695	0.0363	0.11695
15	Pentan-2-ol	-1.25	1.21	-1.05886	-1.13139	-0.19114	-0.11861
16	Pentan-3-ol	-1.33	1.21	-1.05886	-1.13139	-0.27114	-0.19861
17	2-methylbutan-1-ol	-1.13	1.42	-0.88372	-0.95478	-0.24628	-0.17522
18	3-methylbutan-1-ol	-1.13	1.42	-0.88372	-0.95478	-0.24628	-0.17522
19	3-methylbutan-2-ol	-1.08	1.28	-1.00048	-1.07252	-0.07952	-0.00748
20	(ter) pentanol	-1.27	1.21	-1.05886	-1.13139	-0.21114	-0.13861
21	(neo) pentanol	-0.96	1.57	-0.75862	-0.82863	-0.20138	-0.13137
22	1-propylamine	-0.85	0.48	-1.66768	-1.74532	0.81768	0.89532
23	1-butylamine	-0.7	0.97	-1.25902	-1.33323	0.55902	0.63323
24	1-amylamine	-0.61	1.47	-0.84202	-0.91273	0.23202	0.30273
25	1-hexylamine	-0.34	2.06	-0.34996	-0.41654	0.00996	0.07654
26	1-heptylamine	0.1	2.57	0.07538	0.01237	0.02462	0.08763
27	1-octylamine	0.51	3.04	0.46736	0.40764	0.04264	0.10236
28	1-nonylamine	1.59	3.57	0.90938	0.85337	0.68062	0.73663
29	1-decylamine	1.95	4.1	1.3514	1.2991	0.5986	0.6509
30	1-undecylamine	2.26	4.63	1.79342	1.74483	0.46658	0.51517
					$\sum \lvert e_i \rvert$	6.98682	6.55319

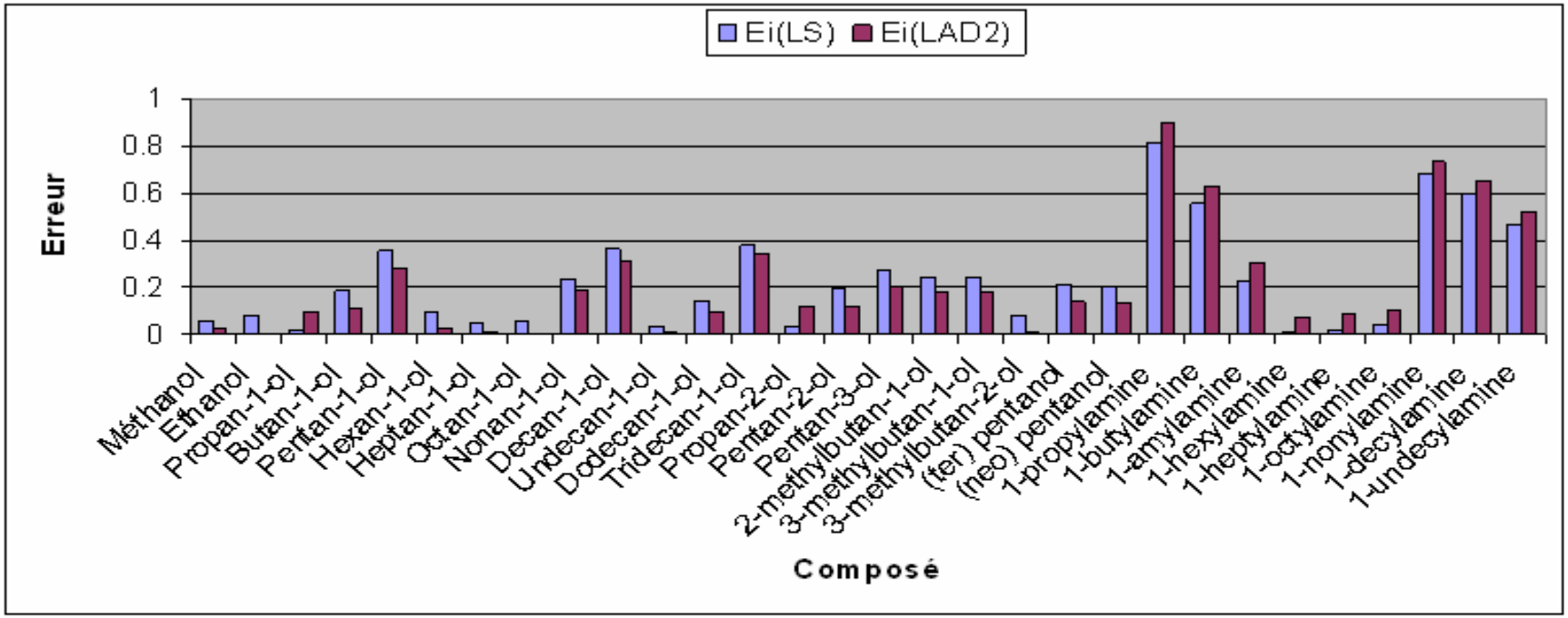


Figure 7 : Histogramme des erreurs résiduelles pour la LS et la LAD_2

LS: $\hat{Y}_{LS}$= 0.834 logP-2.070 R^2=95.20% σ^2_{LS} =0.105

LAD_2: $\hat{Y}_{LAD2}$=0.841 logP-2.149 R^2=97.54% σ^2_{LAD} =0.089

V.2.2- TESTS STATISTIQUES :

Nous avons appliqué sur les paramètres des deux modèles linéaires LS et LAD_2, les tests statistiques que nous avons énoncés, au chapitreIV, pour α=0.05.

V.2.2.1- TEST POUR LES PENTES DE LA METHODE LAD_2 :

L'hypothèse nulle: $\mathbf{H_0 : B_{LAD} = 0.}$

L'hypothèse est acceptée (**AH**) si $\mathbf{t_{calc} < t_{obs}}$, et rejetée (**RH**) si $\mathbf{t_{calc} > t_{obs}}$.

Les calculs ont mené aux résultats ci-après :

	$\mathbf{B_{LAD}}$	$\mathbf{t_{calc}}$	$\mathbf{t_{obs}}$	$\mathbf{H_0}$
Approche Itérative de base	0.841	25.603	2.015	R H_0

L'hypothèse est rejetée, alors la pente est significativement différente de zéro.

V.2.2.2- COMPARAISON DES DEUX PENTES OBTENUES PAR LS ET LAD_2 :

L'hypothèse nulle est: $\mathbf{H_0 : B_{LAD} = B_{LS}}$

L'hypothèse est acceptée (**AH**) si $\mathbf{t_{calc} < t_{obs}}$, et rejetée (**RH**) si $\mathbf{t_{calc} > t_{obs}}$.

Les calculs ont mené aux résultats ci-après :

	$\mathbf{B_{LAD}}$	$\mathbf{B_{LS}}$	$\mathbf{t_{calc}}$	$\mathbf{t_{obs}}$	$\mathbf{H_0}$
Approche Itérative de base	0.841	0.834	0.145	2.015	AH_0

L'hypothèse est acceptée, alors les pentes des deux modèles LS et LAD sont significativement les mêmes.

V.2.2.3- COMPARAISON DES DEUX ORDONNEES OBTENUES PAR LS ET LAD_2 :

L'hypothèse nulle est: $\mathbf{H_0 : Y_{LAD} = Y_{LS}}$

L'hypothèse est acceptée (**AH**) si $\mathbf{t_{calc} < t_{obs}}$, et rejetée (**RH**) si $\mathbf{t_{calc} > t_{obs.}}$

Les calculs ont mené aux résultats ci-après :

	$\mathbf{Y_{LAD}}$	$\mathbf{Y_{LS}}$	$\mathbf{t_{calc}}$	$\mathbf{t_{obs}}$	$\mathbf{H_0}$
Approche Itérative de base	-0.417	-0.350	0.129	2.0147	A H_0

L'hypothèse est acceptée, alors les ordonnées à l'origine des deux modèles LS et LAD sont significativement les mêmes.

V.2.2.4- COMPARAISON DES DEUX VARIANCES LS ET LAD_2 :

L'hypothèse nulle est : $\mathbf{H_0 :}\ \sigma^2_{LAD} = \sigma^2_{LS}$

L'hypothèse est acceptée (**AH**) si : $F_{\alpha/2} < F_{calc} < F_{1-\alpha/2}$, et rejetée (**RH**) si : $F_{calc} > F_{1-\alpha/2}$.

Les calculs ont mené aux résultats ci-après :

	σ^2_{LAD}	σ^2_{LS}	F_{calc}	$F_{\alpha/2}$	$F_{1-\alpha/2}$	H_0
Approche Itérative de base	0.089	0.105	1.172	0.53	1.87	AH_0

L'hypothèse est acceptée, alors les variances des deux modèles LS et LAD sont significativement les mêmes.

V.2.3- Interprétation des résultats :

Le test d'hypothèse nulle $H_0 : B_{LAD}=0$ a été rejeté, ce qui révèle que la pente du modèle LAD est significativement différente de zéro, alors le modèle linéaire peut expliquer la variabilité de la toxicité pCIC50 en fonction du coefficient de partage (Octanol /Eau) logP.

Les tests statistiques de comparaison entre les modèles LS et LAD font ressortir l'absence de translation ou de rotation de l'une des droites de régression par rapport à l'autre. En outre, la comparaison des variances permet de mettre en évidence que les deux modèles ont la même dispersion.

En observant l'histogramme et le tableau des erreurs résiduelles, l'algorithme **itératif de base** donne un modèle LAD avec moins d'erreurs pour 19 composés d'alcools, qui représentent 63% de la population, et plus d'erreurs pour les 9 composés d'amines, le propan-1-ol et le propan-2-ol. Les amines, qui représentent 30% de la population provoquent 61% d'erreurs dans le modèle LAD et 49% d'erreurs dans le modèle LS, ce qui nous fait dire que les deux modèles linéaires simples n'arrivent pas à expliquer assez le comportement des amines.

Le modèle LS explique 95.20% de la variabilité de la toxicité pCIC50 en fonction du coefficient de partage (Octanol /Eau) logP avec $\sum |e_i| = 6.98682$.

Le modèle LAD explique 97.54% de la variabilité de la toxicité pCIC50 en fonction du coefficient de partage (Octanol /Eau) logP avec $\sum |e_i| = 6.55319$.

Le modèle LAD donné par l'algorithme **itératif de base *'Basic iterative approach'*** donne de 6% de moins de $\sum |e_i|$ par rapport au modèle LS**.**

Ces résultats prouvent que l'application de cet algorithme a été bénéfique pour le perfectionnement du modèle linéaire simple.

V.3.1- Droites de régression LS et LAD_3 :

Après l'entrée des données dans l'application programmée en langage Pascal et exécutable sous Windows, nous avons calculé les paramètres du modèle LAD_3 donnés par **l'algorithme de la descente directe**. Le R^2 et τ^2_{LAD} ont été calculés via Excel selon la méthode décrite au chapitreIV, et les paramètres du modèle LS ont été calculés via Minitab, le tableau 6 reproduit les paramètres du modèle LS et LAD_3 :

Tableau 6: Paramètres des modèles LS et LAD_3.

	m	**b**	**R^2**	**Variance**	$\sum\|e_i\|$
LS	0.834	- 2.070	95.20%	σ^2_{LS} =0.105	6.98682
LAD_3	0.829	- 2.050	93.97%	τ^2_{LAD} =0.092	7.06315

La figure 8 reproduit les représentations graphiques des deux droites de régression LS et LAD_3 réalisées sous Excel :

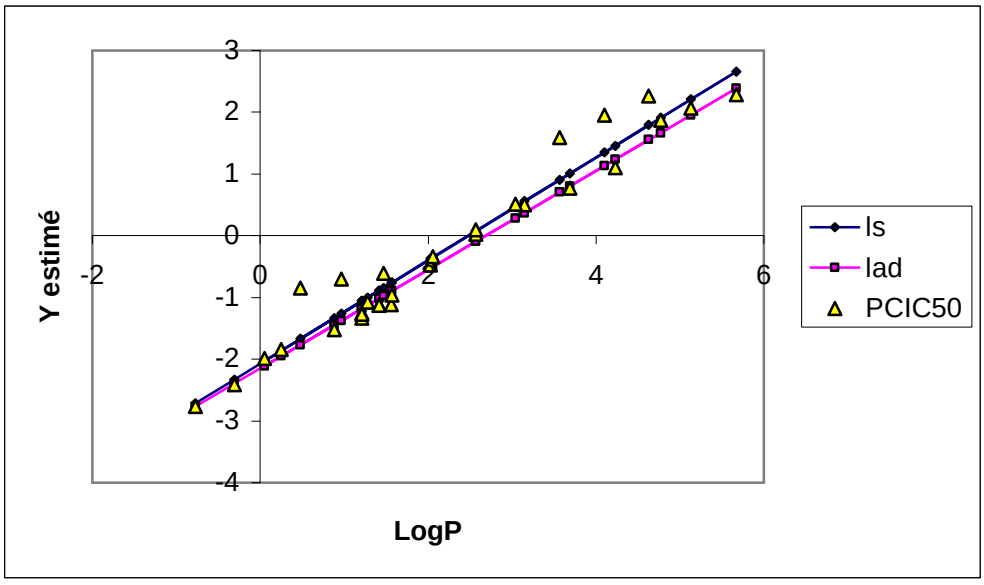


Figure 8 : Droites de régression LS, et LAD_3.

Puis nous avons calculé par les deux méthodes LS et LAD_3, les valeurs estimées de la toxicité pCIC50 et les erreurs résiduelles correspondant à chaque coefficient de partage logP des composés. Le tableau 7 reproduit la toxicité estimée et l'erreur résiduelle obtenues par régressions LS et LAD_3 pour chaque composé; la figure 9 reproduit l'histogramme des erreurs résiduelles.

Tableau7 : Erreurs résiduelles pour la LS et la LAD_3

N°	Composé	pCIC50	logP	$\hat{Y}_{LS}$	$\hat{Y}_{LAD3}$	$\lvert e_{i(LS)} \rvert$	$\lvert e_{i(LAD3)} \rvert$
1	Méthanol	-2.77	-0.77	-2.71018	-2.68533	-0.05982	-0.08467
2	Ethanol	-2.41	-0.31	-2.32654	-2.30399	-0.08346	-0.10601
3	Propan-1-ol	-1.84	0.25	-1.8595	-1.83975	0.0195	-0.00025
4	Butan-1-ol	-1.52	0.88	-1.33408	-1.31748	-0.18592	-0.20252
5	Pentan-1-ol	-1.12	1.56	-0.76696	-0.75376	-0.35304	-0.36624
6	Hexan-1-ol	-0.47	2.03	-0.37498	-0.36413	-0.09502	-0.10587
7	Heptan-1-ol	0.02	2.57	0.07538	0.08353	-0.05538	-0.06353
8	Octan-1-ol	0.5	3.15	0.5591	0.56435	-0.0591	-0.06435
9	Nonan-1-ol	0.77	3.69	1.00946	1.01201	-0.23946	-0.24201
10	Decan-1-ol	1.1	4.23	1.45982	1.45967	-0.35982	-0.35967
11	Undecan-1-ol	1.87	4.77	1.91018	1.90733	-0.04018	-0.03733
12	Dodecan-1-ol	2.07	5.13	2.21042	2.20577	-0.14042	-0.13577
13	Tridecan-1-ol	2.28	5.67	2.66078	2.65343	-0.38078	-0.37343
14	Propan-2-ol	-1.99	0.05	-2.0263	-2.00555	0.0363	0.01555
15	Pentan-2-ol	-1.25	1.21	-1.05886	-1.04391	-0.19114	-0.20609
16	Pentan-3-ol	-1.33	1.21	-1.05886	-1.04391	-0.27114	-0.28609
17	2-methylbutan-1-ol	-1.13	1.42	-0.88372	-0.86982	-0.24628	-0.26018
18	3-methylbutan-1-ol	-1.13	1.42	-0.88372	-0.86982	-0.24628	-0.26018
19	3-methylbutan-2-ol	-1.08	1.28	-1.00048	-0.98588	-0.07952	-0.09412
20	(ter) pentanol	-1.27	1.21	-1.05886	-1.04391	-0.21114	-0.22609
21	(neo) pentanol	-0.96	1.57	-0.75862	-0.74547	-0.20138	-0.21453
22	1-propylamine	-0.85	0.48	-1.66768	-1.64908	0.81768	0.79908
23	1-butylamine	-0.7	0.97	-1.25902	-1.24287	0.55902	0.54287
24	1-amylamine	-0.61	1.47	-0.84202	-0.82837	0.23202	0.21837
25	1-hexylamine	-0.34	2.06	-0.34996	-0.33926	0.00996	-0.00074
26	1-heptylamine	0.1	2.57	0.07538	0.08353	0.02462	0.01647
27	1-octylamine	0.51	3.04	0.46736	0.47316	0.04264	0.03684
28	1-nonylamine	1.59	3.57	0.90938	0.91253	0.68062	0.67747
29	1-decylamine	1.95	4.1	1.3514	1.3519	0.5986	0.5981
30	1-undecylamine	2.26	4.63	1.79342	1.79127	0.46658	0.46873
					$\sum \lvert e_i \rvert$	6.98682	7.06315

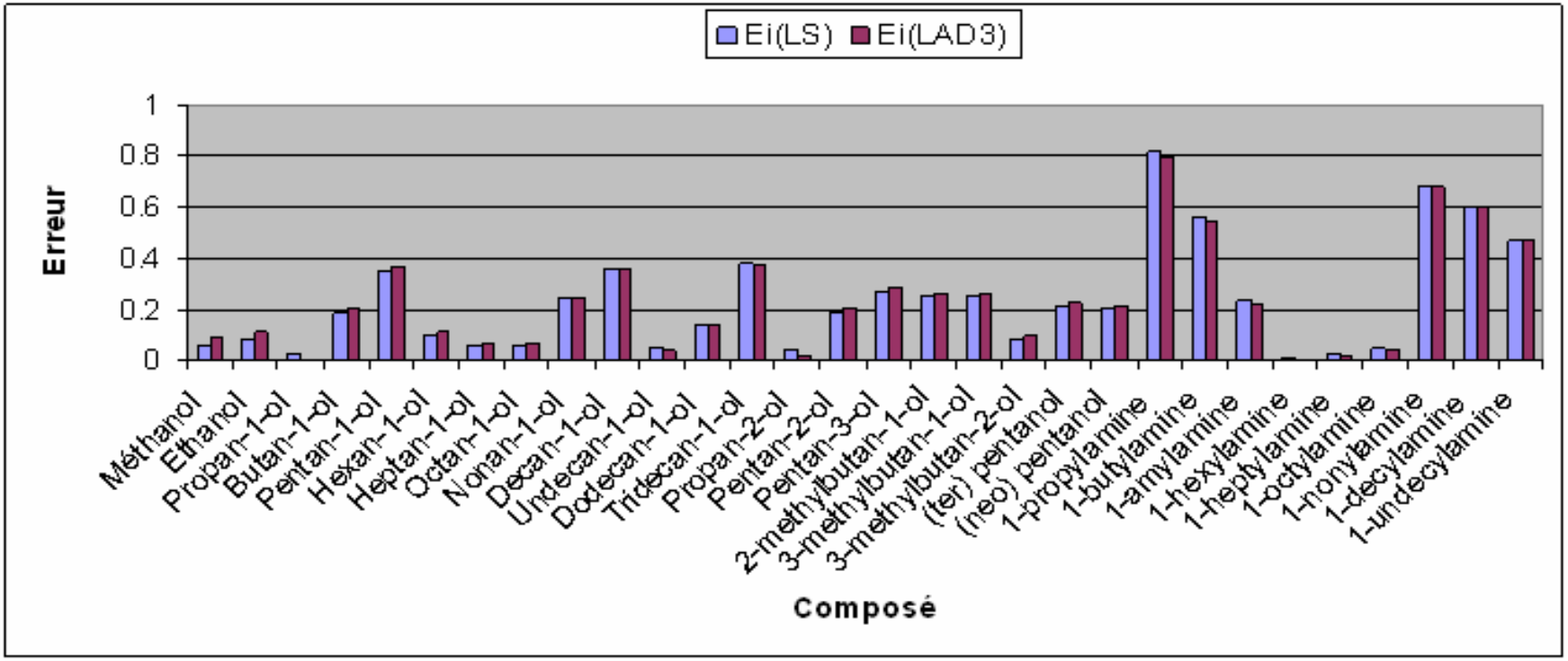


Figure 9 : Histogramme des erreurs résiduelles pour la LS et la LAD_3.

LS: $\hat{Y}_{LS}$= 0.834 logP-2.070 R^2=95.20% σ^2_{LS}=0.105

LAD_3: $\hat{Y}_{LAD3}$=0.829 logP-2.050 R^2=93.97% σ^2_{LAD}=0.092

V.3.2- TESTS STATISTIQUES :

Nous avons appliqué sur les paramètres des deux modèles linéaires LS et LAD_3, les tests statistiques que nous avons énoncés, au chapitreIV, pour α=0.05.

V.3.2.1- TEST SUR LA PENTE LAD_3 :

L'hypothèse nulle: $\mathbf{H_0 : B_{LAD} = 0.}$

L'hypothèse est acceptée (**AH**) si $\mathbf{t_{calc} < t_{obs}}$, et rejetée (**RH**) si $\mathbf{t_{calc} > t_{obs}}$.

Les calculs ont mené aux résultats ci-après :

	$\mathbf{B_{LAD}}$	$\mathbf{t_{calc}}$	$\mathbf{t_{obs}}$	$\mathbf{H_0}$
Méthode de la descente directe	0.829	24.846	2.015	R H_0

L'hypothèse est rejetée, alors la pente est significativement différente de zéro.

V.3.2.2- COMPARAISON DES DEUX PENTES OBTENUES PAR LS ET LAD_3 :

L'hypothèse nulle est: $\mathbf{H_0 : B_{LAD} = B_{LS}}$

L'hypothèse est acceptée (**AH**) si $\mathbf{t_{calc} < t_{obs}}$, et rejetée (**RH**) si $\mathbf{t_{calc} > t_{obs}}$.

Les calculs ont mené aux résultats ci-après :

	$\mathbf{B_{LAD}}$	$\mathbf{B_{LS}}$	$\mathbf{t_{calc}}$	$\mathbf{t_{obs}}$	$\mathbf{H_0}$
Méthode de la descente directe	0.828	0.834	0.108	2.015	AH_0

L'hypothèse est acceptée, alors les pentes des deux modèles LS et LAD sont significativement les mêmes.

V.3.2.3- COMPARAISON DES DEUX ORDONNEES OBTENUES PAR LS ET LAD_3 :

L'hypothèse nulle est: $\mathbf{H_0 : Y_{LAD} = Y_{LS}}$

L'hypothèse est acceptée (**AH**) si $\mathbf{t_{calc} < t_{obs}}$, et rejetée (**RH**) si $\mathbf{t_{calc} > t_{obs.}}$

Les calculs ont mené aux résultats ci-après :

	$\mathbf{Y_{LAD}}$	$\mathbf{Y_{LS}}$	$\mathbf{t_{calc}}$	$\mathbf{t_{obs}}$	$\mathbf{H_0}$
Méthode de la descente directe	-0.340	-0.350	0.129	2.0147	A H_0

L'hypothèse est acceptée, alors les ordonnées à l'origine des deux modèles LS et LAD sont significativement les mêmes.

V.3.2.4- COMPARAISON DES DEUX VARIANCES LS ET LAD_3 :

L'hypothèse nulle est : $\mathbf{H_0 :}\ \sigma^2_{LAD} = \sigma^2_{LS}$

L'hypothèse est acceptée (**AH**) si : $F_{\alpha/2} < F_{calc} < F_{1-\alpha/2}$, et rejetée (**RH**) si $F_{calc} > F_{1-\alpha/2}$.

Les calculs ont mené aux résultats ci-après :

	σ^2_{LAD}	σ^2_{LS}	F_{calc}	$F_{\alpha/2}$	$F_{1-\alpha/2}$	H_0
Méthode de la descente directe	0.092	0.105	1.124	0.53	1.87	AH_0

L'hypothèse est acceptée, alors les variances des deux modèles LS et LAD sont significativement les mêmes.

V.3.3- Interprétation des résultats :

Le test d'hypothèse nulle $H_0 : B_{LAD}=0$ a été rejeté, ce qui révèle que la pente du modèle LAD est significativement différente de zéro, alors le modèle linéaire peut expliquer la variabilité de la toxicité pCIC50 en fonction du coefficient de partage logP (Octanol /Eau).

Les tests statistiques de comparaison entre les modèles LS et LAD font ressortir l'absence de translation ou de rotation de l'une des droites de régression par rapport à l'autre. En outre la comparaison des variances permet de mettre en évidence que les deux modèles ont la même dispersion.

En observant l'histogramme et le tableau des erreurs résiduelles, les 9 amines provoquent 47% d'erreurs dans le modèle LAD et 49% d'erreurs dans le modèle LS, ce qui nous fait dire que les deux modèles linéaires simples n'arrivent pas assez à expliquer le comportement des amines.

Le modèle LS explique 95.20% de la variabilité de la toxicité pCIC50 en fonction du coefficient de partage (Octanol /Eau) logP avec $\sum|e_i| = 6.98682$.

Le modèle LAD explique 93.97% de la variabilité de la toxicité pCIC50 en fonction du coefficient de partage (Octanol /Eau) logP avec $\sum|e_i| = 7.06315$, l'algorithme **de la descente directe** donne un modèle LAD avec plus de $\sum|e_i|$.

Cet algorithme, qualifié de rapide dans la littérature, soufre de défaillances, en sachant que l'algorithme donne de très bons résultats pour d'autres données.

V.3.4- Conclusion :

Nous avons comparé un par un les modèles LAD avec le modèle LS, les deux algorithmes « **Les moindres carrés itérativement re-pondérés »** et **« L'approche itérative de base»** ont pu perfectionner le modèle LS et donner moins de 6% de $\sum|e_i|$; ces algorithmes ont donné moins d'erreurs pour 19 composés d'alcools qui présentent 63% de la population, et plus d'erreurs pour les 9 amines et le propan-1-ol et propan-2-ol, ce qui met en question l'hypothèse de la linéarité du modèle de régression (contamination par des valeurs aberrantes) de nos données où les amines sont les suspects ; mais l'estimateur LAD en donnant moins de $\sum|e_i|$, prouve sa robustesse par rapport à l'estimateur LS face aux divers effets.

VI- CONCLUSION GENERALE :

Pour étudier la régression LAD entre pCIC50 (concentration d'inhibition 50% de la croissance) et logP (coefficients de partage Octanol/Eau) d'un ensemble de 21 alcools et 9 amines, nous avons traitées ces données par trois algorithmes LAD, pour avoir des modèles de régression linéaires simples. Nous avons ensuite comparé les modèles LAD avec le modèle LS. Les deux algorithmes « **Les moindres carrés itérativement re-pondérés »** et **« L'approche itérative de base»** ont pu perfectionner le modèle LS et donner moins de $\sum|e_i|$; ces algorithmes ont donné moins d'erreurs pour 19 alcools, qui représentent 63% de la population, et plus d'erreurs pour les 9 amines, le propan-1-ol et le propan-2-ol.

L'estimateur LAD, en donnant moins de $\sum|e_i|$ et moins d'erreurs pour 63% de la population avec les deux approches LAD, prouve sa robustesse par rapport à l'estimateur LS.

Ces résultats prouvent que la LAD est une importante alternative à la LS, et nous conduisent à penser que de nouvelles voies de recherche peuvent être envisagées. Comme l'étude des modèles à plusieurs descripteurs « régression multiple ».

BIBLIOGRAPHIE

Abdelmalek, N.N. (1980). L1 Solution of Overdetermined Systems of Linear Equations. *A CM Transactions on Mathematical Software,* 6, 220-227.

Adrain, R. (1808). Research concerning the probabilities of the errors which happen in making observations. *Analyst,* 1, 93-109.

Appa, G. and Smith, C. (1973). On *L* and Chebyshev Estimation. *Mathematical Programming,* 5, 73-87.

Armstrong, R.D. andKung, M.T. (1978). Algorithm AS 132 : Least Absolute Value Estimates for a Simple Linear Regression Problem. *Applied Statistics,* 27, 363-366.

Armstrong, R.D., Frome, E.L. and Kung, D.S. (1979). A Revised Simplex Algorithm for the Absolute Deviation Curve Fitting Problem. *Commun. Statist-Simula. Computa.,* B8(2), 175-190.

Arthanari, T.S. and Dodge, Y. (1981). *Mathematical Programming in Statistics.* John Wiley, Interscience Division, New York.

Arthanari, T.S. and Dodge, Y. (1993). *Mathematical Programming in Statistics.* John Wiley Classics Library Edition, New York.

Barrodale, I. and Young, A. (1966). Algorithms for Best *L1* and $L\infty$ Linear Approximations on a Discrete Set. *Numerische Mathematik,* 8, 295-306.

Barrodale, I. and Roberts, F.D.K. (1973). An Improved Algorithm for Discrete *Ly* Linear Approximation. *SIAM J.Numer.Anal.,* 10, 839-848.

Barrodale, I. and Roberts, F.D.K. (1974). Algorithm 478 : Solution of an Overdetermined System of Equations in the *li* norm [F4]. *Communications of the ACM,* 17, 319-320.

Bassett, G.W. and Koenker, R.W. (1978). Asymptotic Theory of Least Absolute Error Regression. *Journal of the American Statistical Association,* 73, 618-622.

Belsley, D., Kuh, E. and Welsh, R.E. (1980). *Regression Diagnostics.* John Wiley, New York.

Bloomfield, P. and Steiger, W. (1980). Least Absolute Deviations Curve-Fitting. *SIAM J.Sci.Stat.Comput,* 1, 290-301.

Bloomfield, P. and Steiger, W.L. (1983). *Least absolute deviations, Theory, Applications and Algorithms.* Birkäuser, Boston.

Boscovich, R.J. (1757). De litteraria expeditione per pontificam ditionem, et synopsis amplioris operis, ac habentur plura ejus ex exemplaria etiam sensorum impressa. *Bononiensi Scientiarum et Arium Instituto Atque Academia Commentarii,* 4, 353- 396.

Boscovich, R.J. (1760). De recentissimis graduum dimensionibus, et figura, ac magnitudine terrae inde derivanda. Philosophiae Recentioris, a Benedicto Stay in Romano Archigynasis Publico Eloquentare Professore, versibus traditae, Libri X, cum

adnotianibus et Supplementis P.
Rogerii Boscovich, S.J., Tomus II, pp. 406-426, esp. 420-425. Romae.

Charnes, A., Cooper, W.W. and Fergusson, R.O. (1955). Optimal Estimation of Executive Compensation by Linear Programming. *Management Science,* 2, 138-151.

Chatterjee, S. and Hadi, A.S. (1988). *Sensitivity analysis in linear regression.* John Wiley, New York.

Dielman, T.E. and Pfaffenberger, R.C. (1984). Computational algorithms for calculating least absolute value and Chebyshev estimates for multiple regression. *American Journal of Mathematical and Management Sciences,* 4, 169-197.

Dielman, T.E. (1992). Computational algorithms for least absolute value regression. In Li-Statistical Analysis and Related Metho'ds, Dodge, Y. editor, 311-326. North-Holland, Amsterdam.

Edgeworth, F.Y. (1887). On Observations Relating to Several Quantities. *Philosophical Magazine, London,* 5th serie, 222-223.

Eisenhart, C. (1961). *Boscovich and the Combination of Observations. Roger Joseph Boscovich, S. J., F.R.S., 1711-1787 : Studies of his Life and Work on the 250th Anniversary of his Birth.* (L.L. White, Ed.). London : Allen and Unwin, Ltd., 200-212.

Eisenhart, C. (1968). Gauss, Carl Friedrich. In *International Encyclopedia of the Social Sciences,* 74-81. Reprinted 1978 in *International Encyclopedia of Statistics.* (With additions), 1, 378-386. Macmillan and Free Press, New York.

Galilei, Galileo (1632). *Dialogo sopra i due massimi sistemi del mondo : Ptolemaico e Copernicano.* Landini, Florence. (English translation, Dialogue concerning the two chief world systems, Ptolemaic and Copernican, by Stillman Drake. Univ. of Calif. Press, Berkeley, 1953.

Gauss, CF. (1809). Theoria motus corporum coelestium in sectionibus conicis solem ambientium. Frid. Perthes et I.H. Besser, Hamburgi. Reprinted 1906 in *Werke,* Band VII, Königlichen Gesellschaft der Wissenschaften, Göttingen, pp. 1-280.

Gauss, CF. (1823). Theoria combinationis observationum erroribus minimis obnoxiae. *Commentationes Societatis Regiae Scientiarum Gottingensis Recentiores,* 5. Reprinted 1880 in *Werke,* Band IV, Königlichen Gesellschaft der Wissenschaften, Göttingen, pp. 3-53, 95-104.

Gauss, CF. (1828). Supplementum theoriae combinationis observationum erroribus minimis obnoxiae. *Commentationes Societatis Regiae Scientiarum Gottingensis Recentiores,* 6. Reprinted 1880 in *Werke,* Band IV, Königlichen Gesellschaft der Wissenschaften, Göttingen, pp. 57-93, 104-108.

Gentle, J.E., Narula, S.C. and Sposito, V.A. (1987). Algorithms for Unconstrained L_1

Linear Regression. In *Statistical Data Analysis Based on the L_1- Norm,* edited by Y.Dodge, Elsevier/North-Holland, Amsterdam, 83-94.

Goldstine, H. (1977). *A history of Numerical Analysis from the 16th through the 19th Century.* Springer, New York.

Harter, H.L. (1974a). The Method of Least Squares and some Alternatives I. *International Statistical Review,* 42, 147-174.

Harter, H.L. (1974b). The Method of Least Squares and some Alternatives II. *International Statistical Review,* 42, 235-264.

Harter, H.L. (1975a). The Method of Least Squares and some Alternatives III. *International Statistical Review,* 43, 1-44.

Harter, H.L. (1975b). The Method of Least Squares and some Alternatives IV. *International Statistical Review,* 43, 125-190.

Harter, H.L. (1975c). The Method of Least Squares and some Alternatives V. *International Statistical Review,* 43, 269-278.

Harter, H.L. (1976). The Method of Least Squares and some Alternatives VI. *International Statistical Review,* 44, 113-159.

Hoerl, A.E. (1962). Application of ridge analysis to regression problems. *Chem. Eng. Progress,* 58, 54-59.

Hoerl, A.E. and Kennard, R.W. (1970a). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics,* 12, 55-67.

Hoerl, A.E. and Kennard, R.W. (1970b). Ridge Regression: Applications to Nonorthogonal Problems. *Technometrics,* 12, 69-82.

Hoerl, A.E., Kennard, R.W. and Baldwin, K.F. (1975). Ridge Regression: Some Simulations. *Communications in Statistics,* 4, 105-123.

Holland, P.W. and Welsch, R.E. (1977). Robust regression using iteratively reweighted least squares. *Communications in Statistics,* A 6, 813-827.

Huber, P.J. (1973). Robust Regression: Asymptotics, Conjectures and Monte Carlo. *Annals of Mathematical Statistics,* 1, 799-821.

Karst, O.J. (1958). Linear Curve Fitting Using Least Deviations. *American Statistical Association Journal,* 53, 118-132.

Klingman, D. and Mote J. (1982). Generalized Network Approaches for solving Least Absolute Value and Tchebycheff Regression Problems. *TIMS/Studies in Management Sciences,* 19, 53-66.

Koenker, R.W. and Bassett, G.W. (1978). Regression Quantiles. *Econometrica,* 46, 33-50.

Koenker, R.W. and d'Orey, V. (1987). Computing regression quantiles. *Appi. Statist,* 36, 383-393..

Laplace, P.S. (1786). Mémoire sur la figure de la terre. Mémoires de l'Académie royale

des Sciences de Paris, Année 1783, 17-46. Reprinted in Oeuvres complètes de Laplace, Vol. 11, pp. 3-32. Gauthier-Villars, Paris, 1895.

Laplace, P.S. (1793). Sur quelques points du système du monde. *Mémoires de l Académie royale des Sciences de Paris,* Année 1789, 1-87. Reprinted in *Oeuvres complètes de Laplace,* 11, 477-458. Gauthier-Villars, Paris, 1895.

Laplace, P.S. (1812). *Théorie analytique des Probabilités.* Third edition with new introduction and three supplements. Paris 1820 ; reprinted as Vol. VII of *Ouvres de Laplace.* Paris, 1847. National edition, Gauthier-Villars. Paris, 1886.

Legendre, A.M. (1805). Nouvelles méthodes pour la détermination des orbites et des comètes. Courcier, Paris. (Appendice sur la méthode des moindres carrés, pp. 72-80).

Li, Yinbo. Arce, Gonzalo. (2003) *a Maximum Likelihood Approach to Least Absolute Deviation Regression.* Available online

McKean, J.W. and Schrader, R.M. (1987). Least absolute errors analysis of variance. In *Statistical Data Analysis Based on the Linorm and Related Methods, Dodge. Y. editor, 297-305.* North Holland, Amsterdam.

Müller, M. (1992). *A comparative study of L_1-norm based simple and multiple regression algorithms.* Report, Postgrade in Statistics, University of Neuchâtel, 1-56.

Narula, S.C. and Wellington, J.F. (1977). An Algorithm for the Minimum Sum of Weighted Absolute Errors Regression. *Commun. Statist.-Simula. Computa.,* B6(4), 341-352.

Nyquist, H. (1985). Ridge type M-estimators. In *Linear Statistical Inference.* Edited by Calinski, T. and Klonecki, W., Springer, Berlin (1985), 246-258.

Pfaffenberger, R.C. and Dielman, T.E. (1990). A Comparison of Regression Estimators when Both Multicollinearity and Outliers are Present. In: *Robust Regression: Analysis and Applications,* edited by Lawrence, K.D. and Arthur, J.L., Marcel Dekker, Inc., New York and Basel (1990), 243-270.

Plackett, R.L. (1972). The discovery of the method of least squares. *Biornetrika,* 59, 239-251. Reprinted in *Studies in History of Statistics and Probability.* (M.G. Kendall and R.L. Plackett, eds.) Griffin, London (1977).

Rhodes, E.C. (1930). Reducing Observations by the Method of Minimum Deviations. *Philosophical Magazine, 7th serie, London,* 9, 974-992.

Ronchetti, E. (1987). Bounded Influence Inference in Regression: A Review. In *Statistical Data Analysis Based on the L\- Norm,* edited by Y.Dodge, Elsevier/North-Holland, Amsterdam, 65-80.

Sadovski, A.N. (1974). Algorithm AS74 : L1 Norm Fit of a Straight Line. *Appi. Stat,* 23, 244-248.

Seal, H.L. (1967). The historical development of the Gauss linear model. *Biornetrika,* 54,

1-24. Reprinted in *Studies in History of Statistics and Probability.* (E.S. Pearson and M.G. Kendall, eds.) Griffin, London.

Sheynin, O.B. (1979). CF. Gauss and the theory of errors. *Archive for History of Exact Science,* 20, 21-72.

Sposito, V.A. and Smith, W.C. (1976). On a Sufficient Condition and a Necessary Condition for *L*\ Estimation. *Appi. Stat.,* 25, 154- 157.

Sprott, D.A. (1978). Gauss's contributions to statistics. *Historia Mathematica,* 5, 183-203.

Stigler, S.M. (1977). An attack on Gauss, published by Legendre in 1820. *Historia Mathematica,* 4, 31-35.

Stigler, S.M. (1978). Francis Ysidro Edgeworth, Statistician. *Journal of the Royal Statistical Society,* Serie A, 141, 287-322.

Stigler, S.M. (1978). Mathematical statistics in the early states. *Annals of Statistics.,* 6, 239-265.

Stigler, S.M. (1981). Gauss and the Invention of Least Squares. *Annals of Statistics,* 9, 465-474.

Usow, K.H. (1967). On *L*\ Approximation I: Computation for Continuous Functions and Continuous Dependence. *SIAM J.Numer.Anal,* 4, 70-88.

Venables, W.N. and Ripley, B.D. (1994). *Modern Applied Statistics with S-Plus.* Statistics and Computing, Springer-Verlag, New York.

Wagner, H.M. (1959). Linear Programming Techniques for Regression Analysis. *Journal of the American Statistical Association,* 54, 206-212.

Weisberg, S. (1985). *Applied linear regression.* John Wiley, New York.

William A, Pfeil. (2006). An Interactive Qualifying Project Report submitted to the Faculty of the Worcester Polytechnic Institute in partial fulfillment of the requirements for the Degree of Bachelor of Science

Wesolowsky, G.O. (1981) A New Descent Algorithm for the Least Absolute Value Regression Problem. *Commun. Statist.-Simula. Computa.,* B10(5), 479-491.

2)Le programme de l'approche Itérative de base:

```
PROGRAM BasicIterativeApprochDiscussedin (INPUT,OUTPUT);
label
 2000;
var
X: array[1..200] of real;
Y: array[1..200] of real;
XO: array[1..200] of real;
YO: array[1..200] of real;
M: array[1..1000] of real;
B: array[1..1000] of real;
P: array[1..1000] of real;
BB: array[1..1000] of real;
BBO: array[1..1000] of real;
MM: array[1..1000] of real;
MMO: array[1..1000] of real;
A: real;
A1: real;
E: real;
B1: real;
C: real;
D: real;
Q1: real;
W: real;
W1: real;
I: integer;
K: integer;
L: integer;
N: integer;
J: integer;
J1: integer;
T: integer;
R: integer;
U: REAL;
INDICEM: integer;
INDICEB: integer;
Z: REAL;
```

```
TT: REAL;
Begin

   WRITE ('ENTREZ LA TAILLE DE VOTRE POPULATION SVP: ');
   readln (N);
   writeln (' ');

   WRITELN ('ENTREZ VOS X[i],Y[i] SVP: ');
   for I := 1 to N do
   begin
   read (X[I],Y[i]);
   end;

   WRITELN ('ENTREZ L ORDRE DE TOLERENCE MINIMAL (POSITIF) SVP: ');
   readln (U);

A:=0; (*la methode des moindres carrés*)
E:=0;
c:=0;
D:=0;
I:=0;

for I:=1 to N do
 begin
  A:=A+X[I];
end;

I:=0;

for I:=1 to N do
 begin
  E:=E+Y[I];
end;

I:=0;

for I:=1 to N do
 begin
  C:=C+X[I]*X[I];
end;

I:=0;

for I:=1 to N do
 begin
  D:=D+X[I]*Y[I];
end;

A1:=((N*D)-(A*E))/((N*C)-(A*A));
B1:=((E*c)-(A*D))/((N*C)-(A*A));

I:=0;
```

```
Q1:=0;

for I:=1 to N do
 begin
  Q1:= Q1+ abs(Y[I]-A1*X[I]-B1);
end;

for j := 1 to N do  (*ORDONANCEMENT DES X[I]Y[I] PAR ORDRE CROISSANT
DES X[I]*)
     begin
     k:=1;
     L:=1;
     for I := 1 to N do
      begin
           if X[j]>X[I] then

            k:=k+1;


           if X[j]<X[I] then

            L:=L+1;

      end;

        R:=N-(L+k-2);

     for T:=1 to R do
      begin
            XO[K+T-1] :=X[J];
            YO[K+T-1] :=Y[J];
      end;
end;

        W:=0;
        for I:= 1 to N do
        begin

         W:=W+ ABS(XO[I]);

        end;

        W1:=0;
        J1:=1;
        for I:= 1 to N do

        begin
         if W1<(W/2)then
         BEGIN
         J1:=J1+1;
         W1:=W1+ABS(XO[I])
         END;
```

```
        end;



Z:=0;
INDICEM:=1;
INDICEB:=1;

M[INDICEM]:=A1;
B[INDICEB]:=B1;

2000:

 for I:= 1 to N do
  begin
   BB[I]:=Y[I]-M[INDICEM]*X[I];
  end;

 for j := 1 to N do  (*ORDONANCEMENT DES b PAR ORDRE CROISSANT*)
     begin
     k:=1;
     L:=1;
       for I := 1 to N do
        begin
           if BB[j]>BB[I] then

            k:=k+1;


           if BB[j]<BB[I] then

             L:=L+1;

        end;

        R:=N-(L+k-2);

        for T:=1 to R do
        begin
          BBO[K+T-1] :=BB[J];
        end;
 END;

        IF (N MOD 2)=0 THEN (*CALCUL DE LA MEDIANE PARMIS LES b*)
         BEGIN
          INDICEB:=INDICEB+1;
          B[INDICEB]:= (BBO[N DIV 2]+BBO[(N DIV 2)+1])/2
         end;

         IF (N MOD 2)=1 THEN
          begin
           INDICEB:=INDICEB+1;
```

```
          B[INDICEB]:= BBO[(N DIV 2)+1]
        end;



  for I:= 1 to N do       (*CALCUL DE LA MEDIANE PONDEREE PARMIS LES m*)
 begin
  MM[I]:= (Y[I]-B[INDICEB])/X[I];
 end;

for j := 1 to N do  (*ORDONANCEMENT DES m PAR ORDRE CROISSANT*)
    begin
    k:=1;
    L:=1;
     for I := 1 to N do
      begin
         if MM[j]>MM[I] then

          k:=k+1;


         if MM[j]<MM[I] then

           L:=L+1;

       end;

       R:=N-(L+k-2);

       for T:=1 to R do
       begin
         MMO[K+T-1] :=MM[J];
         P[K+T-1] :=ABS(X[J]);
       end;
END;

       (*CALCUL DE LA MEDIANE PONDEREE PARMIS LES m*)

       Z:=0;
       J:=1;
       for I:= 1 to N do
       begin
        Z:= P[I]+Z;
        if Z <(W/2)then
        BEGIN
        J:=J+1;
        END;
       end;

 INDICEM:= INDICEM+1;
```

```
  M[INDICEM]:= MMO[J];

  TT:=0;
  FOR I:= 1 TO N DO
  BEGIN
  TT:= TT+ABS(M[INDICEM]*X[I]+B[INDICEB]-Y[I]);
  END;

  IF TT < Q1 THEN
  BEGIN
writeln ('LA DROITE DE REGRESSION DES MOINDRES ECARTS EN VALEURS
ABSOLUES (LAD) EST:');
writeln (' ');
WRITELN ('Y=',M[INDICEM],'X+',B[INDICEB]);
writeln (' ');
writeln ('LA DROITE DE REGRESSION DES MOINDRES CARREES (LS) EST:');
writeln (' ');
WRITELN ('Y=',A1,'X+',B1);
writeln (' ');
WRITELN ('E(ls)=', Q1);
writeln (' ');
WRITELN ('E(lAD)=', TT);

  END;

IF  ABS(M[INDICEM]- M[INDICEM-1]) > U  then
begin
GOTO 2000;
end;
IF ABS(B[INDICEB]-B[INDICEB-1]) > U THEN
begin
GOTO 2000;
end;
```

3)Le programme de la méthode de la descente directe.

```
PROGRAM Wesolowsky (INPUT,OUTPUT);
label
1000, 2000;
var
X: array[1..200] of real;
Y: array[1..200] of real;
B: array[1..200] of real;
P: array[1..200] of real;
BB: array[1..200] of real;
BBO: array[1..200] of real;
A: real;
A1: real;
E: real;
B1: real;
C: real;
```

```
D: real;
Q1: real;
W: real;
I: integer;
K: integer;
L: integer;
N: integer;
J: integer;
T: integer;
R: integer;
INDICEB: integer;
INDICEJ: integer;
MINIMUM: REAL;
Z: REAL;
TT: REAL;
MM:  real;
Begin

   WRITE ('ENTREZ LA TAILLE DE VOTRE POPULATION SVP: ');
   readln (N);
   writeln (' ');

   WRITELN ('ENTREZ VOS X[i],Y[i] SVP: ');
   for I := 1 to N do
   begin
   read (X[I],Y[i]);
   end;


A:=0; (*la methode des moindres carrés*)
E:=0;
c:=0;
D:=0;
I:=0;

for I:=1 to N do
 begin
  A:=A+X[I];
end;

I:=0;

for I:=1 to N do
 begin
  E:=E+Y[I];
end;

I:=0;

for I:=1 to N do
 begin
  C:=C+X[I]*X[I];
```

```
end;

I:=0;

for I:=1 to N do
 begin
  D:=D+X[I]*Y[I];
end;

A1:=((N*D)-(A*E))/((N*C)-(A*A));
B1:=((E*c)-(A*D))/((N*C)-(A*A));

I:=0;
Q1:=0;

for I:=1 to N do
 begin
  Q1:= Q1+ abs(Y[I]-A1*X[I]-B1);
end;

MINIMUM := abs(Y[1]-A1*X[1]-B1);
for I:=2 to N do
 begin
  IF MINIMUM > abs(Y[I]-A1*X[I]-B1) THEN
  BEGIN
  MINIMUM:= abs(Y[I]-A1*X[I]-B1);
  INDICEJ:= I;
  END;
end;

writeln ('LA DROITE DE REGRESSION DES MOINDRES CARREES (LS) EST:');
writeln (' ');
WRITELN ('Y=',A1,'X+',B1);
writeln (' ');
WRITELN ('E(ls)=', Q1);



INDICEB:= 1;
B[INDICEB]:=B1;

2000:

W:= 0;
 for I:=1 to N do
  begin
   W:= W+ABS(1- (X[I]/X[INDICEJ]));
 end;



  for I:= 1 to N do       (*CALCUL DE LA MEDIANE PONDEREE PARMIS LES b*)
```

```
 begin
  IF I  <> INDICEJ THEN
  BEGIN
  C:=   Y[I]-((Y[INDICEJ]*X[I])/X[INDICEJ]);
  D:=   1-(X[I]/X[INDICEJ]);
  BB[I]:= C/D;
  END;
 end;


for j := 1 to N do  (*ORDONANCEMENT DES b PAR ORDRE CROISSANT*)
    begin
    IF J  <> INDICEJ THEN
    BEGIN
    k:=1;
    L:=1;
     for I := 1 to N do
      begin
       IF I  <> INDICEJ THEN
       BEGIN
         if BB[j]>BB[I] then

          k:=k+1;


         if BB[j]<BB[I] then

           L:=L+1;
        END;
      end;

      R:=N-(L+k-2);

      for T:=1 to R do
      begin
        BBO[K+T-1] :=BB[J];
        P[K+T-1] :=ABS(1- (X[J]/X[INDICEJ]));
      end;
     END;
   END;
       (*CALCUL DE LA MEDIANE PONDEREE PARMIS LES m*)

      Z:=0;
      J:=1;
      for I:= 1 to N do
      BEGIN
      IF I  <> INDICEJ THEN
      begin
       Z:= P[I]+Z;
       if Z <(W/2)then
       BEGIN
       J:=J+1;
```

```
        END;
        END;
       end;

      for I:=1 to N do
       begin
        IF I  <> INDICEJ THEN
          BEGIN
          IF BBO[J]=BB[I]THEN
          BEGIN
          INDICEJ:=I;
          GOTO 1000;
          END;
         END;
       end;



1000:

INDICEB:=INDICEB+1;
B[INDICEB]:= BBO[J];

IF B[INDICEB]- B[INDICEB-1] <> 0 THEN
GOTO 2000;


MM:= Y[INDICEJ]/X[INDICEJ]- B[INDICEB]/X[INDICEJ];

  TT:=0;
  FOR I:= 1 TO N DO
  BEGIN
  TT:= TT+ABS(MM*X[I]+B[INDICEB]-Y[I]);
  END;

  IF TT < Q1 THEN
  BEGIN
writeln ('LA DROITE DE REGRESSION DES MOINDRES ECARTS EN VALEURS
ABSOLUES (LAD) EST:');
writeln (' ');
WRITELN ('Y=',MM,'X+',B[INDICEB]);
writeln (' ');
writeln ('LA DROITE DE REGRESSION DES MOINDRES CARREES (LS) EST:');
writeln (' ');
WRITELN ('Y=',A1,'X+',B1);
writeln (' ');
WRITELN ('E(ls)=', Q1);
writeln (' ');
WRITELN ('E(lAD)=', TT);

  END;
```

end.