T 86

UNIVERSITÉ DU QUÉBEC

MÉMOIRE

présenté

à

l'Institut national de la recherche scientifique

INRS-Eau

comme exigence partielle

de la

maîtrise es Sciences (Eau)

par

Carole Lachapelle

B.Sp. Géographie

UTILISATIONS ABUSIVES DES MÉTHODES DE RÉGRESSION ET DE CORRÉLATION: EXEMPLES D'APPLICATION EN SCIENCES DE L'EAU

REMERCIEMENT

Mes remerciements vont à:

Bernard Bobée, professeur à INRS-Eau, directeur de mémoire, qui a proposé le sujet de cette recherche et l'a dirigé;

Pierre Pinard, étudiant à INRS-Eau, pour ses conseils et l'éclairage apportés au chapitre des corrélations fictives;

Daniel Cluis et Fahim Ashkar, qui ont accepté de relire ce travail et de donner leurs commentaires;

V.T. NGuyen, qui a accepté d'agir à titre de correcteur extérieur pour cette thèse;

Claire Lortie, secrétaire pour son travail exemplaire de mise en page;

André Parent, dessinateur pour la réalisation des figures.

RÉSUMÉ

La régression et la corrélation demeurent des outils statistiques simples mais néanmoins valables dans la recherche d'une liaison quantitative entre des variables.

La consultation de plusieurs articles scientifiques nous a convaincus que les fondements théoriques de ces méthodes sont peu connus et qu'en pratique ces deux outils d'analyse sont très souvent confondus. Il en résulte fréquemment des interprétations abusives des phénomènes que l'on cherche à comprendre.

Ce mémoire comprend essentiellement deux parties. La première est une présentation synthétique de la théorie de la régression et de la corrélation et une description des principaux outils d'investigation auxquels peut se référer le chercheur. La seconde partie illustre, par des exemples concrets, les utilisations inadéquates les plus fréquentes. On y suggère aussi certains solutions méthodologiques simples permettant de vérifier si les caractéristiques des échantillons répondent aux hypothèses qui constituent les conditions nécessaires à une utilisation rigoureuse des méthodes de régression et de corrélation.

Table des matières

page

REMERCIEMENTS i

RÉSUMÉ..... ii

TABLE DES MATIÈRES..... iii

LISTE DES FIGURES viii

LISTE DES TABLEAUX..... xiv

LISTE DES SYMBOLES..... xvi

1. INTRODUCTION..... 1

2. ANALYSE DE RÉGRESSION LINÉAIRE SIMPLE (R.L.S.): ASPECTS THÉORIQUES..... 5

2.1 Introduction..... 5

2.2 Définition générale et historique de la régression..... 6

2.3 Régression linéaire simple..... 8

2.3.1 Définition..... 8

2.3.2 Modèles théoriques et empiriques..... 9

2.3.3 Hypothèses de la régression..... 12

2.3.4 Estimation et caractérisation statistique des coefficients α et β et η..... 16

2.3.5 Tests et intervalles de confiance des coefficients 21

2.3.5.1 Généralités..... 21

2.3.5.2 Tests sur α et β (tests séparés)..... 22

2.3.5.3 Test simultané pour α et β..... 25

2.3.5.4 Intervalle de confiance pour α, β et η... 25

2.3.6 Prédiction à partir de la droite de régression.... 26

2.3.6.1 Estimation de la variable dépendante et intervalle de prévision..... 26
2.3.6.2 Estimation de la variable indépendante (régression inverse)..... 30

2.3.7 Mesures de dispersion autour de la droite..... 33

2.3.8 Comparaison de deux droites de régression..... 35
2.3.8.1 Généralités..... 35
2.3.8.2 Test F..... 37

3. ANALYSE DE CORRÉLATION : ASPECTS THÉORIQUES 41

3.1 Introduction..... 42

3.2 Définition du coefficient de corrélation..... 43

3.3 Conditions d'utilisation du coefficient de corrélation... 44

3.4 Interprétation de la valeur du coefficient de corrélation 45
3.4.1 Cas d'une loi jointe normale bivariée..... 45
3.4.2 Cas d'une loi jointe quelconque..... 45
3.4.3 Commentaires..... 46

3.5 Rapport de corrélation..... 46
3.5.1 Définition..... 46
3.5.2 Interprétation du rapport de corrélation..... 48

3.6 Tests et intervalles de pour le coefficient decorrélation 49
3.6.1 Test sur le coefficient de corrélation..... 50
3.6.2 Comparaison de 2 coefficients de corrélation..... 53
3.6.3 Intervalle de confiance..... 54

3.7 Lien numérique entre le coefficient de corrélation et la droite de régression..... 56

4. ANALYSE DE RÉGRESSION : NON-RESPECT DES HYPOTHÈSES ET MAUVAISES UTILISATIONS. 58

4.1 Introduction..... 59

4.2 Analyse des hypothèses sous-jacentes à la régression..... 61
4.2.1 Hypothèse de linéarité..... 62
4.2.1.1 Vérification de la linéarité..... 62
4.2.1.2 Linéarisation:objectif et principe général 67
4.2.1.3 Méthode d'approche pour choisir une..... 70
transformation.....
4.2.1.4 Choix du meilleur modèle:..... 75
a) Caractéristiques du modèle..... 75
b) Qualités prédictives du modèle (SS_r)... 75
Comparaison entre SS_r, r^2 et $(r')^2$ ...
4.2.1.5 Tests et intervalles de confiance..... 85
(variables originales et variables
transformées).

4.2.2 Hypothèse d'homoscédasticité (variance constante).. 89
4.2.2.1 Vérification de l'homoscédasticité..... 90
4.2.2.2 Causes de l'hétéroscédasticité..... 93
4.2.2.3 Stabilisation de la variance..... 94
4.2.2.4 Commentaires et exemples..... 99

4.2.3 Hypothèse de normalité..... 121
4.2.3.1 Vérification de la normalité de la distri- 122
bution des résidus.....
4.2.3.2 Normalisation des résidus..... 128

4.2.4 Hypothèse de l'indépendance des erreurs..... 134
4.2.4.1 Causes de l'autocorrélation..... 135
4.2.4.2 Mise en évidence de l'autocorrélation..... 135
4.2.4.3 Élimination de l'autocorrélation..... 143

4.3 Problèmes liés à l'échantillonnage..... 149
4.3.1 Échantillon non-homogène..... 149
4.3.2 Échantillon non-représentatif..... 164

5. ANALYSE DE CORRÉLATION : CORRÉLATION FICTIVE ET AUTRES UTILISATIONS ABUSIVES DES COEFFICIENTS DE CORRÉLATION. 170

5.1 Introduction..... 171

5.2 Coefficient de corrélation et nature de la distribution jointe 172
5.2.1 Généralités..... 172
5.2.2 Vérification de la normalité de la distribution jointe..... 173
5.2.3 Commentaires et exemples..... 175

5.3 Corrélation fictive..... 183
5.3.1 Généralités..... 183
5.3.2 Échantillon non représentatif (points extrêmes et points aberrants)..... 185
5.3.3 Échantillon non-homogène..... 190
5.3.4 Emploi de variables communes. 201
5.3.4.1 Coefficient de corrélation pour des sommes et des différences 203
5.3.4.2 Coefficient de corrélation pour des produits et des quotients..... 208
5.3.4.3 Commentaires et exemples..... 210

5.4 Coefficient de corrélation et de détermination des variables transformées..... 219

5.5 Signification de r, R^2 et R pour une relation polynômiale 223

6. CONCLUSION 228

BIBLIOGRAPHIE..... 231

APPENDICE 1..... 240

APPENDICE 2..... 244

APPENDICE 3..... 246

APPENDICE 4..... 248

APPENDICE 5..... 254

APPENDICE 6..... 257

ANNEXE 1 260

ANNEXE 2 262

ANNEXE 3 268

ANNEXE 4 274

LISTE DES FIGURES

LISTE DES FIGURES

FIGURE 2.1 Examen des hypothèses de la régression linéaire.

FIGURE 2.2 Intervalle de confiance η et d'une valeur prédite y_0 associée à x_0.

FIGURE 2.3 Comparaison de droites de régression. A: droites à pentes égales et à faible variance résiduelle. B: droites à pentes égales et à variance résiduelle élevée.

FIGURE 4.1 Examen de l'hypothèse de linéarité.

FIGURE 4.2 Illustration de fonctions linéarisables.

FIGURE 4.3 Illustration de la linéarisation par différentes transformations.

FIGURE 4.4 Relation entre le débit de la conductivité (exemple 4.1).

FIGURE 4.5 Relation entre le débit et la conductivité; échantillon linéarisé. A: Ajustement d'une droite. B: Modèle ajusté manuellement (Hem, 1971).

FIGURE 4.6 Comparaison des différents modèles pour l'exemple 4.2.

FIGURE 4.7 Présentation des données de l'exemple 4.3.

FIGURE 4.8 Examen de l'hypothèse d'homoscédasticité.

FIGURE 4.9 Vérification de l'hypothèse d'homoscédasticité par un test graphique. Comparaison entre les résidus des variables originales (A) et transformées (B).

FIGURE 4.10 Stabilisation de la variance. Comparaison des intervalles Δa et $\Delta a'$ associés à la méthode des moindres carrés et des moindres carrés pondérés.

FIGURE 4.11 Stabilisation de la variance; comparaison entre une transformation linéaire et non-linéaire.

FIGURE 4.12 Comparaison des droites et des intervalles de prévision déduits de la méthodes des moindres carrés et des moindres carrés pondérés.

FIGURE 4.13 Comparaison des pentes suite à la stabilisation de la variance.

FIGURE 4.14 Linéarisation et intervalle de confiance associé au nouveau modèle. A: D'après Jensen (1979). B: Intervalle de confiance du modèle non-linéaire (exemple 4.5).

FIGURE 4.15 Étude de la normalité des résidus. (exemple 4.6 Phosphore).

FIGURE 4.16 Étude de la normalité des résidus. (exemple 4.6 Azote).

Figure 4.17 Examen de la normalité des résidus. A: Variables originales B: Variables transformées pour stabiliser la variance. (exemple 4.7).

FIGURE 4.18 Exemple de résidus autocorrélés en fonction du temps.

FIGURE 4.19 Illustration de l'autocorrélation positive et négative.

FIGURE 4.20 Autocorrélation due à l'hétérogénéité de l'échantillon.

FIGURE 4.21 Critères de décision pour le test d'autocorrélation.

FIGURE 4.22 Vérification de l'homogénéité de l'échantillon.

FIGURE 4.23 Étude de la nature de l'échantillon. A: Échantillon total. B: Sous-groupes homogènes (exemple 4.8).

FIGURE 4.24 Vérification de l'homogénéité de l'échantillon (exemple 4.9).

FIGURE 4.25 Comparaison de deux droites de régression (exemple 4.10).

FIGURE 4.26 Influence de points extrêmes sur les coefficients du modèle de régression linéaire (exemple 4.11).

FIGURE 5.1 Utilisation non justifiée du coefficient de corrélation. Échantillon non-linéaire, non-homogène. Distribution jointe non normale.

FIGURE 5.2 Test de normalité de la distribution jointe de X et Y. (exemple 5.1)

FIGURE 5.3 Relation non-linéaire et coefficient de corrélation. A: Données simulées (exemple 5.2). B: Tiré de Smith (1980) (exemple 5.3)

FIGURE 5.4 Influence relative des points extrêmes en fonction de leur position et du nombre d'observations considérés. Données simulées.

FIGURE 5.5 Influence d'un point extrême sur la valeur du coefficient de corrélation.

FIGURE 5.6 Illustration d'un échantillon non-homogène (exemple 5.7).

FIGURE 5.7 Étude du coefficient de corrélation, Cu: Mn. A: Échantillon total. B: Sous-groupes homogènes (exemple 5.8).

FIGURE 5.8 Étude du coefficient de corrélation, Zn : Cu. A: Échantillon total. B: Sous-groupes homogènes (exemple 5.8).

FIGURE 5.9 Subdivision de l'échantillon pour augmenter la valeur du coefficient de corrélation.

FIGURE 5.10 Corrélation fictive. A: observations originales (x, y). B: modèle déduit des variables y/x et x. C: modèle transposé après transformations. (exemple 5.9)

FIGURE 5.11 Corrélation fictive (Données simulées). A: variables originales, fonctions de variables et modèle associé. B: variables originales, fonctions de variables et log des fonctions de variables.

FIGURE 5.12 Examen du coefficient de corrélation. A: variables originales B: logarithme de fonctions de variables (exemple 5.11).

FIGURE 5.13 Coefficient de corrélation déduit des variables originales et du logarithme des variables. (exemple 5.12).

FIGURE 5.14 Comparaison entre r' et T^2_{yx} (données simulées)

FIGURE 5.15 Utilisation de $\sqrt{R^2}$ pour une relation polynômiale (exemple 5.13)

FIGURE 5.16 Comparaison entre le coefficient de corrélation r et la racine du coefficient de détermination R.

LISTE DES **TABLEAUX**

LISTE DES TABLEAUX

TABLEAU 2.1 Régression linéaire simple. Modèles théoriques et empiriques

TABLEAU 2.2 Ecart-type des estimateurs a, b et $\tilde{y}$ des modèles de la régression.

TABLEAU 2.3 Caractéristiques statistiques des modèles de régression linéaire simple.

TABLEAU 2.4 Intervalles de prévisions pour une valeur y_0 associé à x_0.

TABLEAU 4.1 Linéarisation de fonctions curvilignes et variables associées.

TABLEAU 4.3 Transformations pour stabiliser la variance.

TABLEAU 5.2 Âge des résidus miniers pour les différents sites de l'échantillon (exemple 5.7)

TABLEAU 5.3 Coefficient de corrélation entre l'âge des sites et concentration des résidus par sous-régions homogènes (exemple 5.7).

LISTE DES SYMBOLES

LISTE DES SYMBOLES

α	:	ordonnée à l'origine de la droite de régression.
α_1, α_2	:	valeurs limites du coefficient α pour un intervalle au niveau de confiance $(1-\tilde{\alpha})$.
α_1, α_2	:	ordonnée à l'origine des droites 1 et 2.
$\tilde{\alpha}$, $(1-\tilde{\alpha})$	:	niveau de signification et niveau de confiance.
β	:	pente de la droite de régression
β_1, β_2	:	valeurs limites du coefficient β pour un intervalle au niveau de confiance $(1-\tilde{\alpha})$
β_1, β_2	:	ordonnée à l'origine des droites 1 et 2
ε	:	erreur résiduelle
ξ	:	valeur théorique de la variable indépendante (régression inverse).
ξ_1, ξ_2	:	valeurs limites de ξ pour un intervalle au niveau de confiance $(1-\tilde{\alpha})$

ρ_1 : coefficent de corrélation

$\bar{\rho}_1$: coefficient d'autocorrélation d'ordre 1

σ^2_ε : variance de l'erreur résiduelle

σ^2_x, σ^2_y : variance des variables aléatoires X et Y

$\sigma^2_{y_0}$: variance de $\tilde{y}_i$ pour $X = x_0$

σ^2_Z : variance de la variable aléatoire Z

Θ^2_{YX} : rapport de corrélation

μ_X, μ_Y : moyenne des variables aléatoires X et Y

a : estimateur de α

a_1, a_2 : estimateurs des valeurs limites α_1 et α_2

a_1, a_2 : estimateurs de α_1 et α_2 pour les échantillons 1 et 2.

a_{xy} : estimteur de α (régression inverse)

b : estimateur de β

b_1 et b_2 : estimateurs de α_1 et α_2 pour les échantillons 1 et 2

b_{xy} : estimateur de β (régression inverse)

C : coefficient de variation

Cov(X,Y) : covariance des variables aléatoires X et Y

e : estimateur de ε

E(X) : espérance mathématique de la variable aléatoire X

E(Y) : espérance mathématique de la variable aléatoire Y

F : variable F de Fisher

F_0 : estimation de la quantité F dans l'échantillon

F_c : valeur critique de la variable F au niveau de signification $\tilde{\alpha}$

log : logarithme en base 10

ln : logarithme népérien (base e)

N : nombre d'observations dans l'échantillon

n_1, n_2 : nombre d'observations pour les échantillons 1 et 2

N_c : nombre d'observations des échantillons combinés 1 et 2

N_1, N_2 : nombre d'observations des sous groupes 1 et 2 de l'échantillon de taille N.

r : estimateur de ρ

$\bar{r}_1$: estimateur de $\bar{\rho}_1$

R^2 : coefficient de détermination

S_a^2 : variance estimée de α

S_b^2 : variance estimée de β

$S_{\tilde{y}}^2$: variance estimée de η

$S_{y_0}^2$: variance estimée de $\sigma_{y_0}^2$

S_e^2 : variance estimée de ε pour l'échantillon total

$S_{e_1}^2$, $S_{e_2}^2$: variance estimée de ε des échantillons 1 et 2

SS_r : somme des carrés des résidus e

SS_{r_1}, SS_{r_2} : somme des carrés des résidus des échantillons 1 et 2

$SS_{r_{12}}$: somme de SS_{r_1} et SS_{r_2}

SSR : somme des carrés des résidus des échantillons combinés

S_x^2 : variance estimée de la variable aléatoire X

S_y^2 : variance estimée de la variable aléatoire Y

S_{xy} : estimation de Cov (X,Y)

SS_{xy} : somme des produits croisés des écarts à la moyenne de la variable aléatoire X et de la variable aléatoire Y

$SS_{\tilde{y}}$: variation expliquée par la droite de régression

t : variable de Student

t_0 : valeur prise par t pour l'échantillon considéré

t_c : valeur critique de la variable t de Student pour un niveau de signification $\tilde{\alpha}$

T^2_{yx} : estimateur de Θ^2_{YX}

x_i, y_i : valeurs observées dans l'échantillon

u_i, v_i : variables standardisées

$Var(\alpha)$: variance du coefficient α

$Var(\beta)$: variance du coefficient β

$Var(\eta)$: variance du coefficient η

Z : variable transformée de Fisher

$\bar{Z}$: moyenne de la variable Z

CHAPITRE 1

INTRODUCTION

1. INTRODUCTION

Les méthodes de régression et de corrélation sont parmi les outils statistiques les plus utilisés pour modéliser des relations entre variables et pour mesurer l'importance de leur dépendance statistique (coefficient de corrélation).

La simplicité relative de ces méthodes n'est certainement pas étrangère à leur usage répandu et ce, aussi bien dans les sciences humaines que physiques.

Si pour un échantillon quelconque on se limite, comme c'est souvent le cas, à estimer les paramètres d'une droite ou encore à calculer le coefficient de corrélation, l'outil d'investigation est en effet très simple mais présente aussi un intérêt limité. En effet la fiabilité des résultats ne peut être apprêciée sans le recours à des tests objectifs qui permettent de quantifier l'incertitude associée aux différents estimateurs. La question de l'utilisation même de ces méthodes s'avère plus importante encore: elle ne sera justifiable que dans la mesure où certaines hypothèses de base auront d'abord été vérifiées.

En pratique ces hypothèses sont rarement vérifiées et c'est pourquoi de nombreux abus d'utilisation de ces méthodes peuvent être fréquemment relevés. À la limite les conclusions d'une recherche pourront être mises en doute et ce constat justifie ce mémoire qui s'adresse surtout aux utilisateurs.

L'objectif général de ce mémoire consiste en effet à mettre en relief les utilisations potentiellement inadéquates de ces méthodes d'analyse, et les conclusions erronées qui peuvent en découler.

De façon plus spécifique nous nous proposons de:

(1) présenter les différentes hypothèses théoriques de la régression en insistant sur leur importance relative et sur les conséquences qui découlent de leur non-respect;

(2) décrire les tests qui permettent de vérifier ces hypothèses, en faisant ressortir les conditions préalables à leur utilisation.

(3) analyser les différentes transformations que l'on peut opérer sur les variables originales pour qu'elles respectent les hypothèses de base et analyser les résultats obtenus à partir de ces transformations.

En plus des hypothèses qui constituent le fondement de la validité des deux méthodes statistiques considérées, d'autres aspects plus généraux, liés à la nature des échantillons, sont aussi étudiés dans la mesure où ils peuvent affecter la validité des résultats obtenus. De la même façon nous présentons pour ces sujets les tests qui permettent de vérifier si les échantillons répondent aux critères d'homogénéité et de représentativité et les solutions que l'on peut envisager pour s'y conformer.

Dans cette même optique, le problème des corrélations fictives sera également discuté. De nombreux cas ont été signalés, notamment dans les

sciences de l'eau. C'est pourquoi nous avons jugé opportun d'accorder une attention particulière à ce sujet auquel nous avons donné un sens large afin de pouvoir concilier les différentes définitions qu'en donnent divers auteurs.

Bien que sous certaines conditions la régression et la corrélation puissent être utilisées de façon complémentaire il n'en demeure pas moins que ces méthodes d'analyse sont distinctes et c'est pourquoi elles sont présentées dans des chapitres différents. Les conditions d'utilisation et les hypothèses qui sous-tendent chacune de ces approches, souvent confondues en pratique, permettront de mettre en relief leur complémentarité et leur différence.

Les chapitres deux (régression) et trois (corrélation) présentent les principaux résultats théoriques de ces méthodes. L'ensemble de ces chapitres ne prétend pas cependant être une revue exhaustive de la théorie de base. De nombreux ouvrages ont été écrits sur le sujet et on s'y référera pour des développements plus complets. On y trouvera surtout les éléments et les principales notations utiles à la compréhension des chapitres 4 et 5.

Les hypothèses de base spécifiques à ces méthodes et les conditions générales qui permettent de valider les résultats obtenus sont analysées aux chapitres 4 (régression) et 5 (corrélation). La plupart des sujets traités sont illustrés par des exemples tirés de la littérature scientifique. On pourra mettre en évidence les erreurs les plus fréquentes et on verra que les objectifs visés par les transformations de variables sont très souvent mal compris.

Dans le cadre de ce mémoire, seule la théorie de la régression linéaire simple (R.L.S.) est présentée. On discutera toutefois de relations non-linéaires dans le cas où la théorie de la R.L.S. devient applicable après certaines transformations des variables originales.

CHAPITRE 2

ANALYSE DE RÉGRESSION LINÉAIRE SIMPLE (R.L.S.) : ASPECTS THÉORIQUES

2. ANALYSE DE RÉGRESSION LINÉAIRE SIMPLE (R.L.S.) : ASPECTS THÉORIQUES

2.1 INTRODUCTION

On trouvera dans ce chapitre les principales formules utiles pour l'estimation des paramètres du modèle linéaire. D'autres modèles et qui sont en fait des cas particuliers de la R.L.S. sont également présentés. Ces modèles moins connus, qui présentent néanmoins un grand intérêt, s'avèrent utiles lorsque certaines informations "a priori" sur la population étudiée sont connues.

Outre les méthodes d'estimation des paramètres de la droite nous résumons les formules qui permettent d'éprouver de façon objective les résultats obtenus. Les tests de signification et les intervalles de confiance et de prévision sont présentés pour le modèle général. Nous tenons compte aussi des autres modèles en présentant ces mêmes résultats dans des tableaux de synthèse.

Quelques notions relatives à la régression inverse ont aussi retenu notre attention. En effet, contrairement aux cas précédents, le choix des formules se fait ici en fonction de la nature de la variable indépendante.

Une définition explicite du coefficient de détermination et de l'écart-type résiduel est également présentée. Dans les chapitres 4 (régression) et 5 (corrélation) on référera très souvent à ces coefficients (section 2.3.7), fréquemment utilisés en dehors de leur champ d'application.

Essentiellement ce chapitre vise donc deux objectifs:

(1) constituer une synthèse des principales notions théoriques associées à l'analyse de régression, en insistant sur certains résultats moins connus et néanmoins très utiles.

(2) présenter les différentes formules et notations utilisées dans les chapitres 4 et 5.

2.2 DÉFINITION GÉNÉRALE ET HISTORIQUE DE LA RÉGRESSION

La régression est une technique statistique employée pour décrire la liaison entre des variables. Les outils statistiques utilisés en analyse de régression permettent d'établir un modèle mathématique de la relation entre les variables et de quantifier l'incertitude associée à la relation qui en résulte. L'objectif le plus fréquent et le plus général est de prédire la valeur d'une certaine variable dite dépendante, connaissant la valeur d'une variable qui lui est associée (ou semble lui être associée), et que l'on nomme généralement variable indépendante ou explicative.

Le concept de régression fut introduit par Galton (1885) dont les recherches portaient sur l'hérédité. Cet auteur fut un des premiers qui décrivit les liens de dépendance entre certaines variables physiologiques. Il s'intéressa à la relation entre la taille des parents et celle de leurs enfants.

Ses résultats démontraient que les enfants dont les parents étaient très grands, s'avéraient en moyenne plus petits que leurs parents et que les

enfants dont les parents étaient très petits étaient en moyenne plus grands que leurs parents. Il en déduisit que d'une génération à l'autre la caractéristique de taille au lieu de s'accentuer (de plus en plus grande et de plus en plus petite) "régressait" c'est-à-dire se rapprochait de la moyenne.

On donne aujourd'hui une toute autre interprétation aux résultats de Galton. En fait Galton généralisa à toute la population ce qu'il avait démontré pour un individu (en l'occurence les cas extrêmes, soit les très grands et les très petits). En réalité les parents de grande taille ont des enfants de grande taille, également aussi des enfants plus petits, de sorte que globalement la taille moyenne des enfants d'une famille pourra être plus faible que celle des parents. En somme les grands ne produisent pas tous des grands ou des plus grands. Il en résulte que les tailles extrêmes (dans une famille) demeurent exceptionnelles, tout comme dans la population dont l'échantillon est issu. Ce fait implique que la distribution des tailles et la dispersion autour de la moyenne demeure la même dans la population et qu'en réalité il n'y a pas "régression" au sens ou l'entendait Galton.

Quoiqu'il en soit le terme régression a survécu mais s'applique plus généralement aux méthodes statistiques qui décrivent et étudient la relation entre des variables.

2.3 RÉGRESSION LINÉAIRE SIMPLE

2.3.1 Définition

L'équation de la liaison entre deux variables peut prendre des formes diverses. La fonction sera linéaire ou parabolique, hyperbolique, exponentielle, etc. De tous les modèles possibles, le modèle linéaire est le plus utilisé. D'une part parce qu'il est simple et qu'il constitue très souvent une approximation satisfaisante de la relation qui existe entre les variables. D'autre part parce qu'il est souvent possible de ramener une relation non-linéaire à une relation linéaire par transformation des variables.

Si la relation entre les variables semble linéaire et que d'autre part elle met en relation deux seules variables, la régression sera dite linéaire simple. L'équation générale est alors:

$$y = a + bx \qquad (2.1)$$

avec

y = variable dépendante

x = variable indépendante

a = ordonnée à l'origine de la droite

b = pente de la droite

2.3.2 Modèles théoriques et modèles empiriques

Le modèle théorique d'une régression linéaire simple est décrit par l'équation suivante:

$$\eta = \alpha + \beta X \qquad (2.2)$$

avec η = espérance de Y pour une valeur donnée de X soit $E(Y|X = x)$

x = valeur prise par la variable dépendante X

α et β = paramètres du modèle, respectivement ordonnée à l'origine et pente de la droite théorique.

Théoriquement η représente la valeur moyenne d'une distribution conditionnelle. Soit la valeur moyenne de Y étant donné $X = x$. En effet à moins que le rapport entre X et Y soit une relation mathématique parfaite on peut s'attendre à obtenir pour un x donné de la variable X plusieurs valeurs y de la variable Y. La valeur recherchée η est théoriquement au centre de la distribution et en constitue la moyenne.

Une valeur observée particulière y_i se confond rarement à la valeur ponctuelle correspondante sur la droite de régression théorique soit η_i (ou à son estimation dans l'échantillon $\tilde{y}_i$). La différence entre la valeur théorique particulière de la droite (η_i) et la valeur observée y_i constitue la partie imprévisible du modèle ou son erreur ε_i soit:

$$\varepsilon_i = (y_i - \eta_i) \qquad (2.3)$$

la valeur observée y_i est donc égale à:

$$y_i = \underbrace{\alpha + \beta x_i}_{\eta_i} + \varepsilon_i \qquad (2.4)$$

L'erreur ε_i peut être due à un ensemble de causes: soit des erreurs de mesure sur y ou en raison du fait que le modèle ne tienne pas compte de toutes les variables explicatives associées à la variable dépendante.

En pratique on ne dispose pour une valeur x_i de la variable X que d'une seule valeur y_i de la variable Y ou d'un échantillon limité de valeurs y_i et on ne peut connaître ni la vraie moyenne η ni la valeur exacte des paramètres α et β. La droite de régression théorique est alors estimée à partir d'un échantillon limité de couples x_i, y_i où i = 1 à N . Le modèle empirique est alors:

$$\tilde{y} = a + bx \qquad (2.5)$$

où $\tilde{y}$, a et b sont les estimateurs respectifs de η, α et β

Pour l'échantillon ε_i est estimé par e_i. Les résidus e_i constituent la partie imprévisible du modèle. e_i mesure donc l'écart entre une observation y_i et son estimation ponctuelle $\tilde{y}_i$ sur la droite empirique. On a:

$$e_i = y_i - \tilde{y}_i \qquad (2.6)$$

d'où l'on tire la valeur de y_i

$$y_i = \underbrace{a + bx_i}_{\tilde{y}_i} + e_i \qquad (2.7)$$

Le modèle de l'équation 2.5 est souvent présenté sous une autre forme qui est équivalente,

$$\tilde{y} = a' + b\,(x - \bar{x}) \qquad \text{(modèle 2)}$$

L'équivalence numérique entre le modèle 1 (EQ 2.5) et le modèle 2 est présenté à la section 2.3.4. D'autres modèles peuvent-être considérés si l'on a certaines connaissances a priori concernant la droite théorique. Par exemple on peut connaître avec exactitude la valeur de Y pour un X donné. Soit lorsque $X = x_s$, $Y = y_s = \eta_s$. On a donc pour ce point particulier $\eta_s = \alpha + \beta\,(x_s)$ et l'équation générale du modèle devient:

$$\eta = \eta_s + \beta\,(x - x_s)$$

estimé par:

$$\tilde{y} = \eta_s + b\,(x - x_s) \qquad \text{(modèle 3)}$$

En pratique le modèle qui présente le plus d'intérêt est un cas particulier du modèle 3, qui représente le cas d'une droite passant par l'origine. L'équation est alors $\eta = \beta X$

On peut aussi mentionner le cas où l'on connaîtrait à l'avance la

valeur de la pente (ce qui est assez rare en pratique soit, $\beta = \beta_0$) et comme cas particulier $\beta = 0$. Ces différents modèles sont présentés au tableau 2.1. Les caractéristiques des paramètres (estimation, tests et intervalles de confiance) sont spécifiques à chacun de ces modèles et ils seront présentés à la section 2.3.4.

2.3.3 Hypothèses de la régression

En analyse de régression l'on utilise la méthode des moindres carrés pour ajuster une droite à la série observée (échantillon). Cette méthode présentée à la section 2.3.4 consiste à minimiser la somme des carrés des résidus (e_i) et permet d'estimer les paramètres du modèle linéaire.

Pratiquement cette méthode peut être utilisée quelle que soit l'allure de la relation. En effet, on peut ajuster une droite à une distribution curvilinéaire (exponentielle, logarithmique, parabolique, etc.). Cette droite sera la meilleure dans le sens des moindres carrés. Cependant on conviendra qu'un tel modèle (linéaire) est peu intéressant si la relation n'est pas linéaire. Avant de considérer cette méthode d'ajustement il est donc important d'examiner si elle convient à l'allure de la relation.

Pour que le modèle de régression représente adéquatement les observations et pour utiliser les résultats théoriques associés à l'analyse de régression (tests et intervalles de confiance) on devra d'abord vérifier certaines hypothèses. La théorie de la régression linéaire simple s'appuie donc sur les hypothèses
suivantes:

- hypothèse concernant la variable indépendante.

MODELE	THEORIQUE	EMPIRIQUE	COMMENTAIRES
1	$\eta = \alpha + \beta x$	$\tilde{y} = a + bx$	Modèle général aucune connaisance à priori
2	$\eta = \alpha' + \beta(x-\bar{x})$	$\tilde{y} = a' + b\ (x-\bar{x})$	Modèle général. Équivalent au modèle 1. Présente un intérêt particulier pour certains développements théoriques.
3	$\eta = \eta_s + \beta(x-x_s)$	$\tilde{y} = \eta_s + b\ (x-x_s)$	Équation du modèle lorsque la droite théorique passe par le point (x_s, η_s)
4	$\eta = \beta x$	$\tilde{y} = bx$	Équation du modèle lorsque la droite passe par l'origine: $\alpha = 0$. Cas particulier du modèle 3(x_s=0, η_s=0)
5	$\eta = \alpha$	$\tilde{y} = a$	Équation du modèle lorsque $\beta = 0$. Cas particulier pour $\beta=\beta_o$

TABLEAU 2.1 Régression linéaire simple. Modèles théoriques et empiriques

- la variable X qui peut être fixée ou aléatoire doit théoriquement être mesurée sans erreur.

- hypothèse concernant la variable dépendante.
  - la relation statistique entre X et Y doit être linéaire.
  - la distribution conditionnelle de Y pour $X = x_i$ est une loi normale $N(\eta, \sigma^2)$ avec $\eta = E(Y|X = x_i)$
  - la variance conditionnelle de Y lorsque $X = x_i$ est constante et elle est indépendante de X. Elle est égale à σ^2 quelle que soit la valeur prise par X (caractéristique dite d'homoscédasticité).
  - les observations Y sont mutuellement indépendantes ($Cov(Y_i, Y_j) = 0$)

La figure 2.1 illustre quelques distributions conditionnelles de Y pour des valeurs x_i fixées (i = 1 à 4). Sur la figure 2.1B les variances ne sont pas constantes et la distribution conditionnelle de Y ne suit pas toujours une loi normale. Sur la figure 2.1C l'hypothèse de linéarité n'est pas respectée.

Il est aussi possible de présenter les hypothèses de la régression linéaire sous une autre forme en considérant la variable centrée ε_i. Étant donné la relation qui existe entre ε_i et y_i (EQ 2.3), on aura les hypothèses suivantes:

- le modèle est linéaire
- la distribution de ε (pour $X = x_i$) est normale de moyenne 0 et de variance σ^2_ε soit $N(0, \sigma^2_\varepsilon)$
- la variance de ε est constante et indépendante de X.
- les erreurs ε_i et ε_j sont mutuellement indépendantes, ($Cov(\varepsilon_i, \varepsilon_j) = 0$)

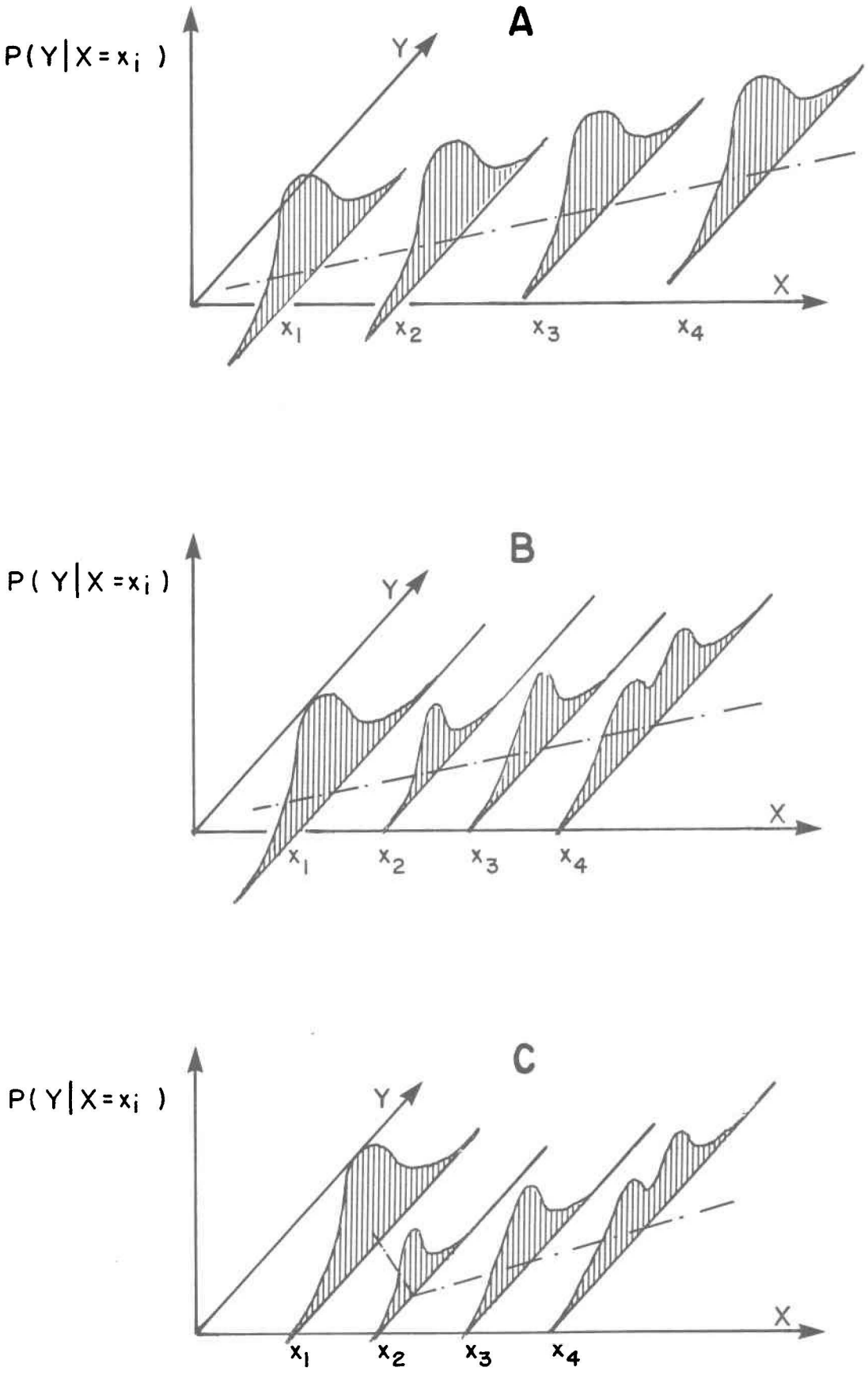


FIGURE 2.1 Examen des hypothèses de la régression linéaire. Modifié d'après Kleinbaum et Kupper (1978).

2.3.4 Estimation et caractérisation statistique des coefficients α et β et η

Lorsqu'on dispose de N couples d'observations (x_i, y_i) l'équation de la droite de régression est obtenue en estimant les paramètres α et β par la méthode des moindres carrés.

Si les différentes observations étaient parfaitement confondues sur une droite cette équation serait aisément déduite du calcul de la pente. On en déduirait la valeur de l'ordonnée à l'origine (α). Cependant les observations se présentent le plus souvent sous la forme d'un nuage de points où, à un x_i donné, correspond un ou plusieurs y_i. À travers ce nuage de points on pourrait tracer une droite approximative qui aurait le désavantage d'être différente pour chaque observateur. La meilleure droite doit donc procéder d'une méthode plus objective. Pour l'échantillon il s'agit de minimiser la somme des carrés des résidus e_i c'est-à-dire minimiser SS_2 tel que:

$$SS_r = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (y_i - \tilde{y}_i)^2 \qquad (2.8)$$

en remplacant $\tilde{y}_i$ par sa valeur (EQ 2.7) on est conduit à minimiser:

$$SS_r = \sum_{i=1}^{N} [y_i - (a + bx_i)]^2$$

En minimisant SS_r par rapport à a et b et en annulant leurs dérivés partielles on obtient:

$$a = \bar{y} - b\,\bar{x} \tag{2.9}$$

$$b = \frac{\sum_{i=1}^{N} x_i\, y_i - \bar{y} \sum_{i=1}^{N} x_i}{\sum_{i=1}^{N} x_i^{\,2} - \bar{x} \sum_{i=1}^{N} x_i} \tag{2.10}$$

il est aussi possible de démontrer que b peut encore s'écrire

$$b = \frac{1/N\text{-}1 \sum_{i=1}^{N} (x_i - \bar{x})\,(y_i - \bar{y})}{1/N\text{-}1 \sum_{i=1}^{N} (x_i - \bar{x})^2} = \frac{S_{xy}}{S_x^2} \tag{2.11}$$

où l'on reconnaît le rapport de la covariance estimmée de x et y S_{xy} sur la variance S_x^2 de la variable indépendante. En simplifiant (2.11) on obtient:

$$b = \frac{\sum_{i=1}^{N} (x_i - \bar{x})\,(y_i - \bar{y})}{\sum_{i=1}^{N} (x_i - \bar{x})^2} = \frac{SS_{xy}}{SS_x} \tag{2.12}$$

avec $$SS_{xy} = S_{xy} \ (N-1)$$

et $$SS_x = S_x^2 \ (N-1)$$

En remplacant a par sa valeur (EQ. 2.9) dans $\tilde{y} = a+bx$ on obtient pour l a droite $\tilde{y}$ une nouvelle équation.

$$\tilde{y} = \bar{y} + b \ (x - \bar{x}) \qquad (2.13)$$

En posant $\bar{y} = a'$ on obtient le modèle 2 présenté à la section 2.3.2. On voit ainsi que la droite passe nécessairement par le point $(\bar{x}, \bar{y})$ En effet, si $x = \bar{x}$, $\tilde{y} = \bar{y} + b(\bar{x} - \bar{x}) = \bar{y}$. On peut aussi vérifier que le couple $(\bar{x}, \bar{y})$ est au centre de gravité du nuage de points constitués par les N couples (x_i, y_i).

Caractéristiques statistiques des coefficients.

Si l'hypothèse de normalité est respectée c'est-à-dire si la distribution de Y pour $X = x_i$ suit une loi normale de moyenne η et de variance σ^2 (ou encore si ε est une variate normale N (0, σ_ε^2) on peut déduire les moments des estimateurs (espérance mathématique et variance), leur distribution et celle de leur variables centrés réduites. On trouvera en appendice 1 une présentation détaillée de ces caractéristiques pour le modèle 1, et aux tableaux 2.2 et 2.3, une synthèse de ces caractéristiques pour les différents modèles présentés à la section 2.3.2.

MODÈLE EMPIRIQUE	ÉCART-TYPE ASSOCIÉ À a	ÉCART-TYPE ASSOCIÉ À b	ÉCART-TYPE ASSOCIÉ À $\tilde{y}$
(1) $\tilde{y} = a + bx$	$S_a = S_e \sqrt{1/N + \bar{x}^2/SS_x}$	$S_b = S_e / \sqrt{SS_x}$	$S_{\tilde{y}} = S_e \left(1/N + (x-\bar{x})^2 / SS_x\right)^{1/2}$
(2) $\tilde{y} = a' + b\,(x - \bar{x})$	$S_{a'} = S_e / \sqrt{N}$	$S_b = S_e / \sqrt{SS_x}$	$S_{\tilde{y}} = S_e \left(1/N + (x-\bar{x})^2 / SS_x\right)^{1/2}$
(3) $\tilde{y} = n_s + b(x - \bar{x})$ La droite passe par le point (x_s, n_s)	----	$S_b = S_e / \sqrt{\sum (x_i - x_s)^2}$	$S_{\tilde{y}} = S_e \left((x_i - \bar{x}_s)^2 / \sum (x_i - x_s)^2\right)^{1/2}$
(4) $\tilde{y} = bx$		$S_b = S_e / \sqrt{\sum x_i^2}$	$S_{\tilde{y}} = S_e \left(x^2 / \sum x_i^2\right)^{1/2}$
(5) $\tilde{y} = a$	$S_a = S_e / \sqrt{N}$	----------	$S_{\tilde{y}} = S_e / \sqrt{N}$

Tableau 2.2 Ecart-type des estimateurs a, b et $\tilde{y}$ des modèles de la régression.

MODÈLES	ESTIMATEURS	MOMENTS
1. $\eta = \alpha + \beta x$	$\tilde{y} = a + bx$	$E(\tilde{y}) = \eta$ $\quad$ $\text{var}\ \tilde{y} = \sigma_\varepsilon^2 \left[1/N + \left[(x-\bar{x})^2 / SS_x\right]\right]$
	$a = \bar{y} - b\bar{x}$	$E(a) = \alpha$ $\quad$ $\text{var}\ a = \sigma_\varepsilon^2 \left[1/N + (\bar{x}^2 / SS_x)\right]$
	$b = SS_{xy} / SS_x$	$E(b) = \beta$ $\quad$ $\text{var}\ b = \sigma_\varepsilon^2 / SS_x$
2. $\eta = \alpha' + \beta(x-\bar{x})$	$\tilde{y} = a' + b(x-\bar{x})$	$E(\tilde{y}) = \eta$ $\quad$ $\text{var}\ \tilde{y} = \sigma_\varepsilon^2 \left[1/N + \left[(x-\bar{x})^2 / SS_x\right]\right]$
	$a' = \bar{y}$	$E(a') = \alpha'$ $\quad$ $\text{var}\ a' = \sigma_\varepsilon^2 / N$
	$b = SS_{xy} / SS_x$	$E(b) = \beta$ $\quad$ $\text{var}\ b = \sigma_\varepsilon^2 / SS_x$
3. $\eta = \eta_s + \beta(x-x_s)$	$\tilde{y} = \eta_s + b(x-x_s)$	$E(\tilde{y}) = \eta$ $\quad$ $\text{var}\ \tilde{y} = \sigma_\varepsilon^2 \left[(x_i-x_s)^2 / \sum (x_i-x_s)^2\right]$
	$b = \dfrac{\sum (y_i-\eta_s)(x_i-x_s)}{\sum (x_i-x_s)^2}$	$E(b) = \beta$ $\quad$ $\text{var}\ b = \sigma_\varepsilon^2 / \sum (x_i-x_s)^2$
4. $\eta = \beta x$	$\tilde{y} = bx$	$E(\tilde{y}) = \eta$ $\quad$ $\text{var}\ \tilde{y} = \sigma_\varepsilon^2 \left[x^2 / \sum x_i^2\right]$
	$b = \sum x_i y_i / \sum x_i^2$	$E(b) = \beta$ $\quad$ $\text{var}\ b = \sigma_\varepsilon^2 / \sum x_i^2$
5. $\eta = \alpha'$	$\tilde{y} = a'$	$E(\tilde{y}) = \alpha'$ $\quad$ $\text{var}\ \tilde{y} = \sigma_\varepsilon^2 / N$
	$a' = \bar{y}$	$E(a') = \alpha'$ $\quad$ $\text{var}\ a' = \sigma_\varepsilon^2 / N$

TABLEAU 2.3 Caractéristiques statistiques des modèles de régression linéaire simple.

2.3.5 Tests et intervalles de confiance des coefficients α et β.

2.3.5.1 Généralités

La droite de régression obtenue à partir de l'échantilllon constitue une estimation de la droite théorique. En effet un autre échantillon permettrait d'obtenir une droite différente, laquelle serait une nouvelle estimation de la droite théorique réelle.

À chacune de ces droites correspondraient autant d'estimateurs a et b des coefficients α et β. Dans les faits cependant on dispose pour estimer la droite d'un seul échantillon et donc d'une valeur unique pour a et b. Il est donc important d'effectuer un test sur la pertinence des valeurs particulières prises par les estimateurs.

En supposant que l'on veuille tester si l'estimateur de β est significativement différent d'une valeur β_0 quelconque on pose les hypothèses suivantes:

$$H_0 : \beta = \beta_0$$

$$H_1 : \beta \neq \beta_0 \qquad \text{test bilatéral}$$

Pour un test unilatéral l'hypothèse H_1 peut prendre la forme:

$$H_1 : \beta < \beta_0 \quad \text{ou} \quad H_1 : \beta > \beta_0$$

On compare ensuite la valeur centrée réduite $t_0 = (b - \beta_0) / S_b$ (Eq. A 1.6) à une valeur critique t_c que l'on trouve dans les tables de

Student. La valeur critique t_c est choisie en fonction du nombre de degrés de liberté ν, spécifique au modèle considéré et du niveau de signification $\tilde{\alpha}$. Le niveau de signification $\tilde{\alpha}$ est déterminé en fonction du type d'erreur que l'on veut minimiser. Plus $\tilde{\alpha}$ est petit et plus on minimise l'erreur de type I qui consiste à rejetter H_0 quand H_0 est vrai. À l'inverse plus $\tilde{\alpha}$ est grand et plus on minimise l'erreur de type II (accepter H_0 quand H_0 est faux) mais on augmente l'erreur de type I. Les valeurs de $\tilde{\alpha}$ généralement utilisés sont $\tilde{\alpha} = 0,05$ et $\tilde{\alpha} = 0,01$. En supposant $N = 12$ et $\tilde{\alpha} = 0,05$, la valeur critique de la table de Student sera:

pour un test bilatéral: $t_c = t_{N-2}(\tilde{\alpha}/2) = t_{10}\ (0,025) = 2,228$

et pour un test unilatéral: $t_c = t_{N-2}(\tilde{\alpha}) = t_{10}\ (0,05) = 1,812$

Bien qu'il soit possible de vérifier l'égalité de β ou de α à quelque valeur que ce soit β_0 ou α_0, en pratique il est surtout intéressant de vérifier si $\beta = 0$ et si $\alpha = 0$. Ces tests sont présentés dans ce qui suit.

2.3.5.2 Tests sur α et β (tests séparés)

- Test sur β

L'hypothèse H_0 que l'on désire généralement examiner pour la pente est:

$$H_0 : \beta = 0$$

$$H_1 : \beta \neq 0$$

Ce test permet de vérifier si la pente de la droite de régression est significative. En effet, la pente estimée (b) est liée à la dépendance de y sur x. Si on peut démontrer que β n'est pas significativement différent de zéro la relation 2.2.

$$\eta = \alpha + \beta x \qquad \text{(modèle 1)}$$

devient alors

$$\eta = \alpha = \text{constante} \qquad \text{(modèle 5)}$$

Dans ce cas la meilleure estimation de y est égale à sa moyenne ($\bar{y}$) . On en déduit pour une valeur particulière y_i:

$$y_i = \bar{y} + e_i$$

puisque y est distribué de façon aléatoire autour de sa moyenne.

Pour un niveau de signification $\tilde{\alpha}$ on comparera donc la valeur $t_0 = b/S_b$ à la valeur critique $t_{N-2}\ (\tilde{\alpha}/2)$. Accepter l'hypothèse nulle signifie que l'information que l'on peut tirer de l'échantillon ne permet pas de conclure que la pente est significative. On en déduit donc (avec un certain risque d'erreur) que la droite de régression est non-significative c'est-à-dire que la connaissance de la variable indépendante n'est pas utile pour estimer la variable dépendante. Ces conclusions sont cependant valables si on a d'abord vérifié l'hypothèse de linéarité. En effet si la linéarité entre x et y n'a pas été vérifiée la signification du test peut être très différente.

• Test sur α

Un test sur α présente souvent de l'intérêt lorsqu'on veut vérifier que l'ordonnée à l'origine est nulle c'est-à-dire que la droite passe par l'origine. Supposons par exemple que l'assimilation d'un certain métal (M_2) par un organisme aquatique, est nulle en l'absence d'un autre métal ($M_1 = 0$). On rendrait compte de cette connaissance en utilisant le modèle 4 pour ajuster la droite. Dans le cas cependant où un telle évidence n'est pas connue mais qu'on présume suite aux résultats expérimentaux (plusieurs points au voisinage de 0 ($M_1 = 0$, $M_2 = 0$) que la droite passe par l'origine, il peut être utile d'effectuer un test sur α. Les hypothèses que l'on doit poser sont donc pour un test bilatéral:

$$H_0 : \alpha = 0$$

$$H_1 : \alpha \neq 0$$

Pour un niveau de signification donné on compare la quantité $t_0 = a/S_a$ à la valeur critique $t_{N-2}(\tilde{\alpha}/2)$. Si on ne peut rejeter l'hypothèse nulle (H_0: $\alpha = 0$) on pourra alors utiliser le modèle 4 pour estimer la pente.

Pour les différents modèles la quantité t_0 (voir appendice 1) sera estimée en tenant compte des valeurs particulières S_a, S_b, (voir tableau 2.2) et du nombre de degrés de liberté associés à la valeur critique t, soit $t_{(N-1)}$ pour les modèles 3, 4 et 5 (voir tableau 2.3).

2.3.5.3. Test simultané sur α et β

Dans le cas du modèle 1 les estimateurs a et b ne sont pas indépendants puisque leur covariance est non nulle. En effet, on a $Cov(a,b) = -\sigma_x^2 / SS_x$. Il serait donc plus rigoureux de tester simultanément les valeurs α et β. On doit alors considérer les hypothèses suivantes:

$$H_0 : \alpha = \alpha_0 \quad \text{et} \quad \beta = \beta_0$$
$$H_1 : \alpha \neq \alpha_0 \quad \text{ou} \quad \beta \neq \beta_0$$

On peut montrer (Neter et Wasserman, 1974) que si H_0 est vrai la quantité F_0

$$F_0 = \frac{1}{2S_e^2} \left(N (a-\alpha_0)^2 + 2N\bar{x} (a-\alpha_0) (b-\beta_0) + (b-\beta_0)^2 \sum_{i=1}^{N} x_i^2 \right) \qquad (2.14)$$

suit une loi F à 2 et N-2 degrés de liberté. On compare donc F_0 avec la valeur critique $F_{2,\ N-2}(\tilde{\alpha})$. Si $F_0 < F_{2,\ N-2}(\tilde{\alpha})$, on accepte H_0.

2.3.5.4. Intervalles de confiance pour α, β et η

Connaissant la distribution de α, β et η et leurs écart-types respectifs (estimés par S_a, S_b et $S_{\tilde{y}}$), les intervalles de confiance au niveau de confiance $(1-\tilde{\alpha})$ peuvent être déduits des équations A 1.5 à A 1.7 (Appendice 1). On a

$$\begin{matrix} \alpha_1 \\ \alpha_2 \end{matrix} = a \pm t_{N-2}(\tilde{\alpha}/2) \cdot S_a \qquad (2.15)$$

$$\begin{matrix}\beta_1 \\ \beta_2\end{matrix} = b \pm t_{N-2}(\tilde{\alpha}/2) \quad . \; S_b \tag{2.16}$$

$$\begin{matrix}\eta_1 \\ \eta_2\end{matrix} = \tilde{y} \pm t_{N-2}(\tilde{\alpha}/2) \quad . \; S_{\tilde{y}} \tag{2.17}$$

On tiendra compte pour les différents modèles des valeurs particulières de S_a, S_b et $S_{\tilde{y}}$ et du nombre de degrés de liberté associés à la valeur critique t_c soit, t_{N-2} pour les modèles 1 et 2 et t_{N-1} pour les autres modèles. Si on construit l'intervalle de confiance point par point autour de $\tilde{y}$ on peut vérifier que l'écart est :

. minimum pour $X = \bar{x}$	modèles 1 et 2
. minimun au point fixé X_s	modèle 3
. nul pour X = 0 et Y = 0	modèle 4
. constant pour tout X	modèle 5

Les intervalles de confiance pour ces différents modèles sont présentés à la figure 2.2.

2.3.6 Prédiction à partir de la droite de régression.

2.3.6.1 Estimation de la variable dépendante et intervalle de prévision.

Rappellons que le principal objectif de la régression est de prédire la valeur de y pour une valeur de x donné. Pour $x = x_0$ la meilleure estimation que l'on puisse faire de la variable dépendante correspond à y_0 situé sur la

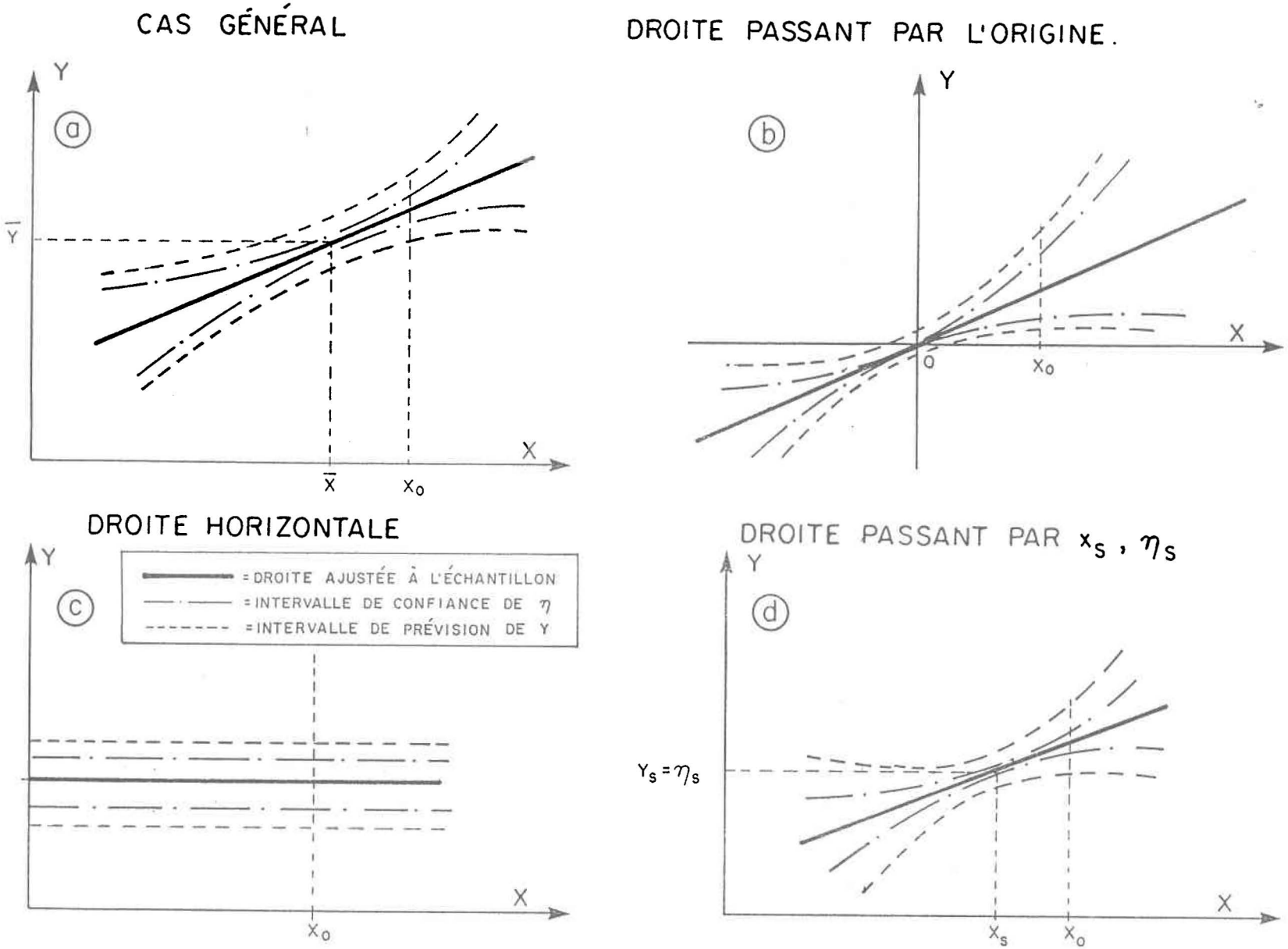


FIGURE 2.2 **Intervalle de confiance η et d'une valeur prédite y_0 associé à x_0.**

droite de régression construite à partir de l'échantillon initial. L'intervalle de confiance présenté à la section 2.3.5 représente l'intervalle de confiance autour de la droite, dont chaque point correspond théoriquement à la valeur moyenne de la distribution de Y pour $X = x_i$. De la même manière on peut construire un intervalle pour une estimation ponctuelle de Y. Cet intervalle appellé intervalle de prévision est nécessairement plus large que l'intervalle de confiance.

La valeur réelle de y_0 correspond à η_0 entaché d'une erreur ε_0. On a

$$y_0 = \eta_0 + \varepsilon_0$$

De cette relation il est possible de déterminer la variance associée à y_0 soit $\sigma^2_{y_0}$ et son estimation dans l'échantillon $S^2_{y_0}$. Soit pour le modèle 1:

$$S^2_{y_0} = S^2_e \left(1 + \frac{1}{N} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right) \qquad (2.17)$$

avec

$$S^2_e = SS_r \; / \; N - 2$$

On peut en déduire pour une valeur particulière y_0 les valeurs limites y_1 et y_2 de l'intervalle de prévision au niveau de confiance $(1 - \tilde{\alpha})$.

On trouvera au tableau 2.4 un résumé des intervalles associés à une valeur y_0 pour les modèles 1, 2, 3, 4, et 5. Bien que plus large que

VALEURS ESTIMEÉS POUR x_0	INTERVALLE DE PRÉVISION $y_2 < y_0 < y_1$
(1) $\tilde{y}_0 = a + bx_0$	$y_1, y_2 = \tilde{y}_0 \pm t_{N-2}\,(\tilde{\alpha}/2)\, S_e \left[1+1/N + \left((x_0 - \bar{x})^2 / SS_x\right)\right]^{\frac{1}{2}}$
(2) $\tilde{y}_0 = a' + b(x_0 - \bar{x})$	$y_1, y_2 = \tilde{y}_0 \pm t_{N-2}\,(\tilde{\alpha}/2)\, S_e \left[1+1/N + \left((x_0 - \bar{x})^2 / SS_x\right)\right]^{\frac{1}{2}}$
(3) $\tilde{y}_0 = a_s + b(x_0 - x_s)$ droite passant par (x_s, a_s)	$y_1, y_2 = \tilde{y}_0 \pm t_{N-2}\,(\tilde{\alpha}/2)\, S_e \left[\left(1 + (x_0 - x_s)^2 / \sum (x_i - x_s)^2\right)\right]^{\frac{1}{2}}$
(4) $\tilde{y}_0 = b\, x_0$	$y_1, y_2 = \tilde{y}_0 \pm t_{N-1}\,(\tilde{\alpha}/2)\, S_e \left[1 + (x_0^2 / \sum x_i^2)\right]^{\frac{1}{2}}$
(5) $\tilde{y}_0 = a'$	$y_1, y_2 = \tilde{y}_0 \pm t_{N-1}\,(\tilde{\alpha}/2)\, S_e\, (1 + 1/N)^{\frac{1}{2}}$
$S_e = (SS_r / (N-2)^{\frac{1}{2}}$	modèles 1 et 2
$S_e = (SS_r / (N-1)^{\frac{1}{2}}$	modèles 3,4 et 5

TABLEAU 2.4 Intervalles de prévisions pour une valeur y_0 associé à x_0.

l'intervalle de confiance (associé à la moyenne) l'intervalle de prévision conserve néanmoins la même allure et les mêmes caractéristiques. Ces intervalles sont également illustrées à la figure 2.2.

2.3.6.2 Estimation de la variable indépendante (régression inverse)

Dans ce qui précède on suppose que Y est une variable aléatoire qui suit une distribution normale. La théorie présentée jusqu'ici est valable quelque soit la nature de X (fixée ou aléatoire) et sa distribution (normale ou quelconque). Cependant si l'on désire prédire x pour une valeur de y (régression inverse) et construire l'intervalle autour de cette valeur on doit considérer la nature de la variable x et deux situations sont possibles:

(a) x est une variable aléatoire dont la distribution est normale. La droite de régression est alors construite de la façon décrite précédemment et l'on considère la régression de x sur y, soit:

$$\tilde{x} = a + by \tag{2.18}$$

Les coefficients a (a_{xy}) et b (b_{xy}) sont estimés de la façon qui a été décrite précédemment. On a donc:

$$b_{xy} = \frac{S_{xy}}{S_y^2} \tag{2.19}$$

$$a_{xy} = \bar{x} - b\bar{y} \tag{2.20}$$

Il en est de même pour les caractéristiques statistiques présentées dans les sections précédentes.

(b) x est une variable fixée (prend des valeurs particulières fixées par l'expérimentateur) ou encore x est une variable aléatoire qui ne suit pas une distribution normale. On ne peut considérer la régression de x sur y car les hypothèses de la régression ne sont pas respectées. On procède alors de la façon suivante. On considère à titre d'exemple le modèle 2, soit:

$$\eta = \alpha + \beta (\xi - \bar{x}) \tag{2.21}$$

ξ est la valeur recherchée et peut-être déduite de:

$$\xi = \bar{x} + \frac{\eta - \alpha}{\beta} \tag{2.22}$$

et l'on a pour l'échantillon:

$$x = \bar{x} + \frac{y_0 - a}{b} \tag{2.23}$$

a et b sont déduits de la régression de y sur x et l'on a, $a = \bar{y}$ et $b = S_{xy} / S_x^2$.

Dans le cas où x est une variable fixée plusieurs approches ont été suggérées pour la construction d'un intervalle autour de la valeur estimée x_0. On trouvera dans Montgomery et Peck (1982) plusieurs références et

certaines remarques concernant ces différentes approches. L'approche suggérée par Hald (1952) est présentée dans ce qui suit.

La valeur théorique ξ_0 (estimée par x_0) se situe à l'intérieur de l'intervalle de confiance dont les valeurs limites sont ξ_1 et ξ_2

$$\xi_2 < \xi_0 < \xi_1$$

L'estimation de ces valeurs limites, Hald (1969) correspond à:

$$\begin{matrix}\xi_1 \\ \xi_2\end{matrix} = \bar{x} + \frac{1}{D}\left((x_0 - \bar{x}) \pm t_{N-2}(\tilde{\alpha}/2)\,\frac{S_e}{b}\,\sqrt{D\left(\frac{1}{Q}+\frac{1}{N}\right)+\frac{(x_0-\bar{x})^2}{SS_x}}\right)$$

$$D = 1 - \frac{t_{N-2}(\tilde{\alpha}/2)\,S_e^2}{b^2\,SS_x}$$

Lorsque la pente est significativement différente de zéro il est possible de démontrer que $D \simeq 1$ (Bobée 1983) et la relation précédente est réduite à

$$\begin{matrix}\xi_1 \\ \xi_2\end{matrix} = x_0 \pm t_{N-2}(\tilde{\alpha}/2)\cdot\frac{S_e}{b}\,\sqrt{\frac{1}{Q}+\frac{1}{N}+\frac{(x_0-\bar{x})^2}{SS_x}}$$

b, $\bar{x}$, SS_x, N sont déduits de l'échantillon

Q = nombre de mesures considérées pour obtenir y_0

Il est aussi possible de déterminer l'intervalle de confiance autour ξ_0 en utilisant une méthode graphique (Bobée 1983).

2.3.7 Mesures de dispersion autour de la droite.

Lorsque la droite de régression est construite (estimation des paramètres) il est souvent utile de connaître l'adéquation du modèle et sa capacité à expliquer la variation de Y en fonction de X.

L'écart-type des résidus et le coefficient de détermination mesurent respectivement l'écart des différentes observations autour de la droite et le proportion de variation expliquée par le modèle. On peut aussi en déduire la proportion de variation non expliquée.

Pour comprendre la signification physique de ces mesures on considère d'abord l'écart entre une valeur y_i donnée et sa moyenne $\bar{y}$ dans l'échantillon, soit $(y_i - \bar{y})$. Cette différence qui constitue la variation totale autour de $\bar{y}$ est égale à la somme de deux autres différences l'une résiduelle $(y_i - \tilde{y})$, non expliquée par le modèle linéaire et l'autre expliquée $(\tilde{y}_i - \bar{y})$.

On a donc:

$$(y_i - \bar{y}) = (y_i - \tilde{y}_i) + (\tilde{y}_i - \bar{y}) \qquad (2.24)$$

En élevant au carré et en faisant la sommation pour tous les y_i on obtient.

$$\sum_{i=1}^{N} (y_i - \bar{y})^2 = \sum_{i=1}^{N} (y_i - \tilde{y}_i)^2 + \sum_{i=1}^{N} (\tilde{y}_i - \bar{y})^2 + 2 \underbrace{\sum_{i=1}^{N} (y_i - \tilde{y}_i)(\tilde{y}_i + \bar{y})}_{A}$$

$$SS_y = SS_r + SS_{\tilde{y}} + A \qquad (2.25)$$

Il est possible de démontrer (Bobée, 1983) que A = 0. L'équation 2.25 est donc réduite à

$$SS_y = SS_r + SS_{\tilde{y}} \qquad (2.26)$$

On voit donc que pour un échantillon ajusté par la méthode des moindres carrés la somme des carrés des écarts des observations y_i par rapport à la moyenne $\bar{y}$ est constituée d'une partie résiduelle non expliquée et d'une partie expliquée par la régression. De cette expression on définit deux mesures importantes en analyse de régression, soit l'écart type résiduel S_e (notamment utilisé dans les tests et dans la construction des intervalles de confiance et de prévision) et le coefficient de détermination R^2 qui mesure la proportion de variance expliquée par le modèle. On a :

$$S_e = \sqrt{\frac{SS_r}{N-2}} \qquad \text{modèles 1 et 2} \qquad (2.27)$$

et

$$R^2 = \frac{SS_{\tilde{y}}}{SS_y} = \frac{\text{variation expliquée}}{\text{variation totale}} \qquad (2.28)$$

Pour la régression linéaire la quantité R^2 peut être calculée plus simplement par r^2 qui correspond au coefficient de corrélation au carré (chapitre 3). Cependant lorsqu'il s'agit de régression non-linéaire (régression-polynôminale par exemple) cette équivalence n'est plus vraie et R^2 doit être déduit de l'équation 2.28. De plus, comme on en discutera aux chapitres 4 et 5 la quantité R^2 n'est plus définie lorsqu'on transforme la variable y pour linéariser les observations.

R^2 est un nombre sans dimension compris entre 0 et 1. La proportion de variance non expliquée correspond à $(1 - R^2)$ égale à:

$$(1 - R^2) = \frac{SS_r}{SS_y} \qquad (2.29)$$

Les quantités R^2 et $(1 - R^2)$ sont souvent multipliées par 100 et l'on parle alors de pourcentage de variation expliquée et non-expliquée.

2.3.8. Comparaison de deux droites de régression

2.3.8.1 Généralités

Il est souvent souhaitable de vérifier si deux droites de régression théoriques sont significativement différentes. Par exemple, on peut envisager deux méthodes expérimentales pour construire un modèle de régression entre x et y et être intéressé à vérifier si les résultats associés à chacune de ces méthodes sont comparables. En quelque sorte on veut vérifier si l'une ou l'autre des méthodes sont significativement différentes ou si au

contraire les échantillons obtenus pour chacune d'elles proviennent d'une même population. On peut à cette fin appliquer un test d'idendité des droites.

Avant d'appliquer un test d'idendité des droites il est préférable d'appliquer un test d'égalité des pentes. En effet si les pentes sont significativement différentes on peut conclure que les droites sont aussi différentes. Cependant la réciproque n'est pas toujours vraie. En effet on peut avoir des droites de pentes identiques très éloignées l'une de l'autre. On ne doit donc pas confondre égalité des droites et identité des pentes comme nous avons pu le relever dans certaines publications. À cet effet un exemple réel sera présenté au chapitre 4.

Deux tests différents doivent être considérés pour tester l'égalité des pentes selon que les variances $S^2_{e_1}$ et $S^2_{e_2}$ sont significativement différentes ou non. On trouvera à l'appendice 2 une présentation du test d'égalité des variances, et un test d'égalité des pentes à l'appendice 3.

Dans ce qui suit nous présentons le test F tiré de Neter et Wasserman (1974). Nous faisons ressortir par un exemple graphique que nous avons construit, la signification physique de ce test. Hald (1952) et Smillie (1966) suggèrent un autre test (voir Appendice 4) beaucoup plus long à appliquer que le test F. À partir d'un exemple que nous étudions au chapitre 4 on pourra vérifier l'intérêt que présente son utilisation. Ce test semble en effet plus sensible à la différence entre deux droites lorsque les tailles des échantillons sont faibles.

2.3.8.2 Test d'identité des droites (Test F)

Pour ce test (Neter et Wasserman, 1974) on considère les échantillons 1 et 2 et les droites associées soit:

$$\tilde{y}_1 = a_1 + b_1 x$$
$$\tilde{y}_2 = a_2 + b_2 x$$

et un échantillon combiné conduisant à la droite

$$\tilde{y}(C) = A + Bx$$

avec:

n_1 = nombre d'observations de l'échantillon 1
n_2 = nombre d'observations de l'échantillon 2
$N_C = n_1 + n_2$

Pour effectuer le test on doit calculer les quantités SS_r spécifiques aux échantillons 1 et 2 (SS_{r_1} et SS_{r_2}) et à l'échantillon combiné (SSR).

On pose:

$$SS_{r_{12}} = SS_{r_1} + SS_{r_2}$$

et on examine l'hypothèse:

$$H_0 : \alpha_1 = \alpha_2 \quad \text{et} \quad \beta_1 = \beta_2$$

$$H_1 : \alpha_1 \neq \alpha_2 \quad \text{ou} \quad \beta_1 \neq \beta_2$$

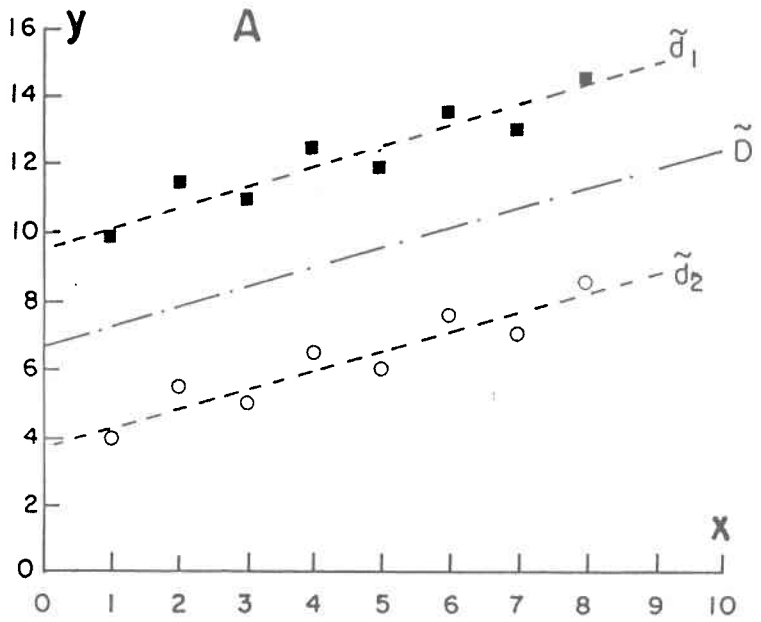
on estime la quantité F_0 soit:

$$F_0 = \frac{(SSR - SS_{r_{12}})/2}{SS_{r_{12}}/(n_1 + n_2 - 4)} \tag{2.30}$$

Pour un niveau de signification $\tilde{\alpha}$ donné on compare alors la valeur F_0 avec la valeur critique F_c pour 2 et $(n_1 + n_2 - 4)$ degrés de liberté, soit $F_{2,(n_1+n_2-4)}(\tilde{\alpha})$. Si $F_0 < F_c$ on accepte H_0. Sinon, on rejette H_0.

La signification physique de ce test peut être visualisée à la figure 2.3. Les observations qui ont servi à construire cet exemple ainsi que les principaux calculs sont présentés à l'Annexe 1. On compare ici les droites $\tilde{d}_1$ et $\tilde{d}_2$ et les droites $\tilde{m}_1$ et $\tilde{m}_2$.

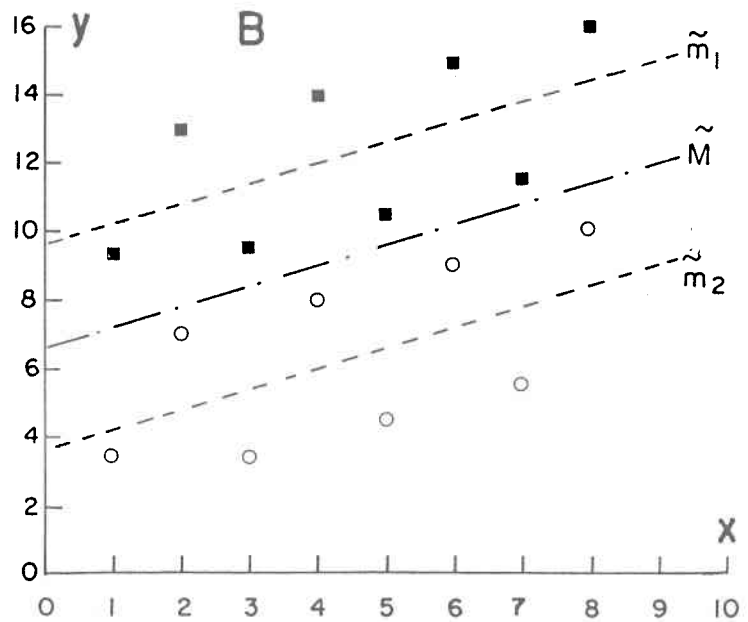
Pour $\tilde{d}_1$ et $\tilde{d}_2$ on a des pentes identiques ($b_1 = b_2$) et $\Delta d = (a_2 - b_1)$. De même pour $\tilde{m}_1$ et $\tilde{m}_2$ on a aussi des pentes égales (b') et $\Delta m = (a'_2) - a'_1)$.



$SS_{r_1} = SS_{r_2} = 1,905$

$SSR = 147,81$

$SSr_{12} = 3,810$



$SS_{r_1} = SS_{r_2} = 28,39$

$SSR = 72,78$

$SSr_{12} = 56,78$

Figure 2.3 Comparaison de droites de régression.
A: droites à pentes égales et à faible variance résiduelle.
B: droites à pentes égales et à variance résiduelle élevée.

On a $\Delta m = \Delta d$. La différence entre $\tilde{d}_1$ et $\tilde{d}_2$ est donc équivalente à la différence entre $\tilde{m}_1$ et $\tilde{m}_2$. Cependant les variances résiduelles pour les droites $\tilde{d}_1$ et $\tilde{d}_2$ sont très faibles (voir Annexe 1). On obtient d'autre part une variance combinée SSR très élevée et par suite une statistique F_0 élevée ce qui conduit au rejet de H_0.

Les variances SS_{r_1} et SS_{r_2} de $\tilde{m}_1$ et $\tilde{m}_2$ sont beaucoup plus élevées. Par suite la différence entre SSR et $SS_{r_{12}}$ est beaucoup plus faible que pour les droites $\tilde{d}_1$ et $\tilde{d}_2$. On obtient donc une valeur F_0 faible et l'on ne peut rejeter H_0.

CHAPITRE 3

ANALYSE DE CORRÉLATION: ASPECTS THÉORIQUES

3. ANALYSE DE CORRÉLATION: ASPECTS THÉORIQUES

3.1 INTRODUCTION.

Dans son sens le plus général le terme corrélation dénote l'interdépendance entre des variables quantitatives (discrètes ou continues) ou entre des variables qualitatives.

Le coefficient de corrélation mesure le degré d'interdépendance entre ces variables. Il existe plusieurs types de coefficient de corrélation dont l'utilisation est liée à la nature des variables étudiées. Ainsi parmi les coefficients non paramétriques on pourra distinguer entre corrélation d'association, de contingence ou encore corrélation des rangs.

Le coefficient d'association mesure l'interdépendance entre des variables aléatoires discrètes dont chacune est limitée à deux valeurs qualitatives. Par exemple, les variables mâle, femelle ou encore positif, négatif. Lorsque les variables peuvent prendre plus de deux valeurs mais en nombre fini on utilisera le coefficient de contingence.

Enfin si les variables peuvent être ordonnées dans un ordre quelconque, les variables sont alors considérées comme semi-quantitatives et leur interdépendance peuvent être étudiées par le coefficient de corrélation des rangs. Les coefficients de corrélation des rangs les plus connus sont le coefficient r_s de Spearman et le coefficient τ de Kendall. De tous les

coefficients non-paramétriques ce sont d'ailleurs les plus utilisés.

Le coefficient de corrélation de Pearson (ρ) dont il sera question par la suite est un coefficient paramétrique qui mesure l'interdépendance entre deux variables quantitatives. C'est le coefficient le plus usuel et c'est celui auquel on se réfère généralement lorsqu'on parle de coefficient de corrélation.

3.2 DÉFINITION

Le coefficient de corrélation est un nombre sans dimension qui peut varier entre +1 et -1. Sous certaines conditions il peut être utilisé comme outil complémentaire à l'analyse de régression. Par définition, son utilisation est conditionnelle à la nature de la variable indépendante ainsi qu'à la nature de la distribution jointe de X et Y.

Le coefficient de corrélation entre les variables X et Y représente le rapport entre la covariance de X et Y sur le produit de leur écart type respectif soit:

$$\rho = \frac{\text{Cov}\,(X,Y)}{\sigma_Y\ \sigma_X} \qquad (3.1)$$

Dans l'échantillon il est estimé par r

$$r = \frac{S_{xy}}{S_x S_y} = \frac{\sum_{i=1}^{N}\left((x_i - \bar{x})\ (y_i - \bar{y})\right) / N}{\left(\sum_{i=1}^{N} (x_i - \bar{x})^2/N \sum_{i=1}^{N} (y_i - \bar{y})^2/N\right)^{\frac{1}{2}}} \quad (3.2)$$

qui peut s'écrire sous une forme équivalente

$$r = \frac{N \sum_{i=1}^{N} x_i y_i - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i}{\left(N \sum_{i=1}^{N} x_i^2 - (\sum_{i=1}^{N} x_i)^2 \quad N \sum_{i=1}^{N} y_i^2 - (\sum_{i=1}^{N} y_i)^2\right)^{\frac{1}{2}}} \quad (3.3)$$

3.3 CONDITIONS D'UTILISATIONS DU COEFFICIENT DE CORRÉLATION

Répétons que ρ est un nombre sans dimension qui peut prendre n'importe quelle valeurs valeurs entre +1 et -1. Pour qu'il puisse être interprété sans équivoque la variable X doit être aléatoire. De plus ρ peut être estimé quelle que soit la nature de la distribution jointe entre X et Y. Cependant, l'interprétation de sa valeur numérique est largement conditionnée par la nature de la distribution jointe de X et Y. En outre, l'équivalence entre un coefficient de corrélation nul, ($\rho = 0$) et l'indépendance de X et Y n'est vérifié que lorsque la distribution jointe de X et Y est normale.

3.4 INTERPRÉTATION DE LA VALEUR DU COEFFICIENT DE CORRÉLATION

3.4.1 Cas d'une loi jointe normale bivariée

- si X et Y sont indépendants alors $\rho = 0$ et la réciproque est aussi vraie
- si $\rho = 0$ alors X et Y sont indépendants.
- si $\rho = \pm 1$ la corrélation ou dépendance (qui sont ici équivalents) entre X et Y est parfaite.
- si $-1 < \rho < 1$ il existe une corrélation ou dépendance statistique entre X et Y qui est plus ou moins forte selon la valeur particulière prise par ρ.

3.4.2 Cas d'une loi jointe quelconque

- Si X et Y sont indépendantes alors $\rho = 0$ mais la réciproque n'est pas vraie

- Si $\rho = 0$ on ne peut conclure que X et Y sont indépendants. On peut seulement dire que X et Y ne sont pas corrélés. En effet, dans le cas d'une loi jointe quelconque, indépendance et corrélation ne sont pas équivalents.

- Si $\rho \neq 0$ il y a une certaine dépendance entre les variables, mais on ne peut en préciser la nature non plus que l'intensité réelle du lien entre les variables.

3.4.3 Commentaires

La seule connaissance de ρ (estimé par r) ne peut donc suffire à nous renseigner sur la nature de la dépendance ni même sur l'existence d'une dépendance entre les variables. Pour être en mesure d'interpréter sans équivoque le coefficient de corrélation on devra donc vérifier que le loi jointe est normale. Les méthodes de vérification sont discutées au chapitre 5. Dans le cas d'une loi jointe quelconque le rapport de corrélation présenté à la section suivante peut s'avérer un outil très utile pour étudier l'intensité de la dépendance entre les variables.

3.5 RAPPORT DE CORRÉLATION

3.5.1 Définition

Lorsque la distribution jointe entre X et Y n'est pas normale, et aussi lorsque la relation n'est pas linéaire, on peut utiliser le rapport de corrélation θ^2_{YX} pour étudier l'intensité de la relation entre les variables X et Y (régression de Y sur X). En effet, le rapport de corrélation est indépendant de la distribution jointe entre les variables. On définit θ^2_{YX} (Bobée, 1983) par

$$\theta^2_{YX} = \frac{\text{Var}\{E(Y|X\}}{\text{Var } Y} \qquad (3.4)$$

avec E (Y | X) = espérance de Y étant donné X

avec $E(Y|X)$ = espérance de Y étant donné X

Pour l'échantillon le rapport de corrélation peut être estimé (Bobée 1983) par T^2_{yx}

$$T^2_{yx} = \frac{SS_{\tilde{y}} + SS_L}{SS_y} \qquad (3.6)$$

avec
$$SS_y = \sum_{i=1}^{N} (y_i - \bar{y})^2 = S_y^2 (N - 1)$$

$$SS_{\tilde{y}} = \sum_{i=1}^{N} (\tilde{y}_i - \bar{y})^2$$

$$SS_L = \sum_{i=1}^{K} \sum_{i=1}^{n_i} (\bar{y}_i - \tilde{y}_i)^2$$

où K = nombre de x_i différents

n_i = nombre de y_i différents pour $x = x_i$

Ces quantités sont plus faciles à calculer si l'on tient compte de la relation 2.28 où $SS_{\tilde{y}} = r^2 (SS_y)$. De plus si n_i est constant pour tout x_i (même nombre de réplicats), T^2_{yx} est alors donné par:

$$T_{yx}^2 = \frac{r^2(SS_y) + n_i \sum_{i=1}^{K} (\bar{y}_i - \tilde{y}_i)^2}{S_y^2 (N - 1)} \qquad (3.7)$$

3.5.2 Interprétation du rapport de corrélation.

Le rapport Θ^2_{YX} présente un grand intérêt lorsque la distribution jointe n'est pas normale. Cependant, comme l'indique sa formulation, il ne peut être utilisé que lorsque l'on dispose de réplicats.

Contrairement, au coefficient de corrélation, Θ^2_{YX} n'est pas symétrique: on a Θ^2_{YX} (régression de Y sur X) $\neq \Theta^2_{XY}$ (régression de X sur Y). Cependant si la loi jointe est une loi normale bivariée on aura par définition:

$$\Theta^2_{YX} = \Theta^2_{XY} = \rho^2$$

Théoriquement on peut montrer (Kendall et Stuart, 1979) que:

$$0 \leq \rho^2 \leq \Theta^2_{YX} \leq 1$$

et pour l'échantillon

$$0 \leq r^2 \leq T^2_{yx} \leq 1$$

Si T^2_{yx} est beaucoup plus élevé que r^2 on peut conclure qu'une relation non- linéaire serait plus appropriée. En effet l'égalité entre ρ^2 et $\Theta^2_{Y/X}$ est vérifié lorsque la relation entre les variables est linéaire. S'il ne fait aucun doute (examen graphique) que la relation est linéaire on calculera de préférence r^2 plus facile à estimer que T^2_{yx}.

Si la relation est non-linéaire (et c'est le cas lorsque la distribution jointe est non-normale) mais que l'on ne dispose pas de replicats on peut toujours estimer r(ou r^2) mais l'interprétation de sa valeur est plus limitée. (voir 3.4.2)

3.6 TESTS ET INTERVALLES DE CONFIANCE POUR LE COEFFICIENT DE CORRÉLATION

Lorsque la distribution jointe de X,Y est normale il est possible de tester ρ pour une valeur quelconque ($\rho = \rho_0$) et de déterminer l'intervalle de confiance associé à cette valeur. En pratique si l'on ne s'éloigne pas trop de l'hypothèse de normalité de la loi jointe ces tests peuvent être utilisés de manière approximative (Bobée, 1983).

3.6.1 Test sur le coefficient de corrélation

r est une variable aléatoire dont la distribution est symétrique pour ρ voisin de 0 et de plus en plus asymétrique lorsque ρ se rapproche de 1 (Neter et Wasserman, 1974). On doit donc envisager un test différent suivant que $\rho = 0$ ou $\rho = \rho_0$ (valeur quelconque différente de 0).

(A) Test pour $\rho = 0$

En pratique c'est le test le plus souvent utilisé. En effet ce test équivaut à examiner l'indépendance de X et Y à condition que la distribution jointe soit normale. On pose les hypothèses suivantes:

$$H_0 : \rho = 0$$

$$H_1 : \rho \neq 0$$

Si on ne peut rejeter l'hypothèse nulle H_0, on conviendra que les variables sont indépendantes. Si $\rho = 0$ on peut démontrer (Neter et Wasserman, 1974) que la quantité t_0

$$t_0 = \frac{r}{\sqrt{1-r^2 / (N-2)}} \qquad (3.8)$$

suit une loi de Student à (N-2) degrés de liberté.

Pour un niveau de signification $\tilde{\alpha}$ (généralement $\tilde{\alpha} = 0,01$ et $\tilde{\alpha} = 0,05$) et pour un test bilatéral on compare la quantité t_0(déduite de l'échantil-

lon) avec la valeur critique $t_{N-2}\ (\tilde{\alpha}/2)$ donnée par la table de Student.

si $t_0 < |\ t_{N-2}\ (\tilde{\alpha}/2)|$ on accepte H_0 au niveau $\tilde{\alpha}$

$t_0 > t_{N-2}\ (\tilde{\alpha}/2)$ on rejette H_0 au niveau $\tilde{\alpha}$ et $\rho > 0$

$t_0 > -t_{N-2}\ (\tilde{\alpha}/2)$ on rejette H_0 au niveau $\tilde{\alpha}$ et $\rho < 0$

(B) TEST POUR $\rho \neq 0$

Étant donnée l'asymétrie de la distribution de ρ ($\rho \neq 0$), on utilise une approximation que l'on doit à Fisher (1956). On pose les hypothèses suivantes:

$H_0 : \rho = \rho_0$ test bilatéral

$H_1 : \rho \neq \rho_0$

et on calcule la quantité Z:

$$Z = \frac{1}{2} \ln \left(\frac{1 + r}{1 - r}\right) \qquad (3.9)$$

Z étant fonction de r, on peut trouver pour une valeur r quelconque la valeur Z correspondante dans des tables construites à cette fin (voir Neter et Wasserman, 1974).

Fisher (1956) a démontré que la quantité Z suit une loi normale de moyenne $\bar{Z}$ et de variance σ_Z^2 soit $N \sim (\bar{Z}, \sigma_Z^2)$ avec:

$$\bar{Z} = \frac{1}{2} \ln \left(\frac{1 + \rho_0}{1 - \rho_0} \right) \qquad (3.10)$$

$$\sigma_Z^2 = 1 / N - 3 \qquad (3.11)$$

On détermine $\bar{Z}$ à partir de la valeur ρ_0 (fixée par le test). On calcule ensuite la variable standardisée u_0 soit:

$$u_0 = (Z - \bar{Z}) / \sigma_Z \qquad (3.12)$$

Pour un niveau de signification $\tilde{\alpha}$, on compare la valeur u_0 à la valeur critique $u(\tilde{\alpha}/2)$ que l'on trouve dans la table de la distribution normale.

$u_0 < | u(\tilde{\alpha}/2) |$ on accepte H_0 et $\rho = \rho_0$

$u_0 > u(\tilde{\alpha}/2)$ on rejette H_0 et $\rho > \rho_0$

$u_0 < -u(\tilde{\alpha}/2)$ on rejette H_0 et $\rho < \rho_0$

3.6.2 Comparaison de 2 coefficients de corrélation

On peut aussi généraliser le test précédent pour comparer 2 coefficients de corrélation. Supposons deux couples d'échantillons de taille N_1 et N_2 et les estimations r_1 et r_2 de leurs coefficients de corrélation. On suppose que les échantillons proviennent de populations indépendantes. On veut vérifier si $\rho_1 = \rho_2$. On pose:

$H_0 : \rho_1 = \rho_2 = \rho_0$ ou encore $\rho_1 - \rho_2 = 0$ $(\rho_1 = \rho_2)$

Si H_0 est vrai:

$$Z_1 = \frac{1}{2} \ln \left(\frac{1 + r_1}{1 - r_1} \right) \quad \text{suit une loi } N \left(Z_0 ; \frac{1}{N_1 - 3} \right)$$

de même

$$Z_2 = \frac{1}{2} \ln \left(\frac{1 + r_2}{1 - r_2} \right) \quad \text{suit une loi } N \left(Z_0 ; \frac{1}{N_2 - 3} \right)$$

Il n'est pas nécessaire de connaître Z_0 (valeur associée à ρ_0 par l'équation 3.10). On peut en déduire cependant la distribution de $(Z_1 - Z_2)$. On a (Hald, 1952), $(Z_1 - Z_2)$ est distribué suivant une loi $N(0, \sum_{i=1}^{2} (1/N_i - 3))$

et la quantité u :

$$u = \frac{Z_1 - Z_2}{\left((1/N_1-3) + (1/N_2-3)\right)^{\frac{1}{2}}} \qquad \text{suit une loi N (0,1)}$$

Comme pour les autres tests on comparera u avec la valeur critique $u(\tilde{\alpha}/2)$. En cas de rejet de H_0 on concluera que $\rho_1 < \rho_2$ ou que $\rho_1 > \rho_2$ selon l'hypothèse H_1 que l'on aura posée.

On peut aussi comparer K (K>2) coefficients de corrélation. On se référera à Hald (1952) pour une présentation détaillée de ce test.

3.6.3 Intervalle de confiance

Pour construire l'intervalle de confiance du coefficient de corrélation on utilise les résultats de la section précédente ($\rho = \rho_0 \neq 0$).

La variable Z est distribuée selon une loi $N(\bar{Z}, \sigma_Z^2)$. Au niveau de confiance $(1 - \tilde{\alpha})$ % l'intervalle de confiance autour de $\bar{Z}$ est tel que

$$Z - u\,(\tilde{\alpha}/2).\ \sigma_Z \leq \bar{Z} \leq Z + u\,(\tilde{\alpha}/2)\ \cdot\ \sigma_Z \qquad (3.13)$$

Z est déduit de l'équation 3.9 et σ_Z déduit de l'équation 3.11. On obtient

les limites Z_1et Z_2 associées à $\bar{Z}$ soit: $Z_1 \leqslant \bar{Z} \leqslant Z_2$ par

$$\begin{matrix} Z_1 \\ Z_2 \end{matrix} = Z \pm u\,(\tilde{\alpha}/2) \cdot \sigma_Z \qquad (3.14)$$

Par la transformation 3.9 (ou encore dans les tables construites à cette fin) on peut en déduire les limites ρ_1 et ρ_2.

Exemple numérique

On a pour un échantillon N = 10 et r = 0,90. On cherche les limites de ρ pour un intervalle de confiance à 95%. D'après 3.9 et 3.11 on peut déduire Z et σ_Z.

$$Z = 1{,}4722$$

$$\sigma_Z = 0{,}3780$$

et $u\,(\tilde{\alpha}/2) = 1{,}96$ (table de la loi normale).

Les limites Z_1 et Z_2 sont donc d'après 3.14

$$Z_1 = 1{,}4722 + 1{,}96\ (0{,}3780) = 2{,}2130$$

$$Z_2 = 1{,}4722 - 1{,}96\ (0{,}3780) = 0{,}7313$$

et l'on a $0{,}7313 \leqslant \bar{Z} \leqslant 2{,}2130$

En utilisant la table de transformation de Fisher ou en remplacant Z par sa valeur (Z_1et Z_2) dans 3.9 on déduit

$$0{,}6239 \leq \rho \leq 0{,}9764$$

L'intervalle autour de ρ permet également de vérifier si ρ est significativement différent de ρ_0 quelconque. Ici par exemple on pourra dire qu'au niveau $\tilde{\alpha}$ =0,05 ρ est significativement différent de 0,5. Il n'est pas significativement différent de 0,8 etc. Si les limites de l'intervalle comprennent la valeur $\rho = 0$, on peut alors conclure que ρ n'est pas significativement différent de zéro pour le niveau de signification $\tilde{\alpha}$ choisi.

3.7 LIEN NUMÉRIQUE ENTRE LE COEFFICIENT DE CORRÉLATION ET LA DROITE DE RÉGRESSION

Lorsque la distribution jointe de X et Y est normale, la distribution conditionnelle de Y pour $X = x_i$ suit une loi normale de moyenne η égale à:

$$\eta = E\,(Y|X = x_i) = \mu_Y + \rho \frac{\sigma_Y}{\sigma_X} (x_i - \mu_X) \qquad (3.15)$$

avec μ_X et σ_X : moyenne et écart-type de la distribution marginale de X.

μ_Y et σ_Y : moyenne et écart-type de la distribution marginale de Y.

En comparant cette relation au modèle 2 on a

$$E(Y|X = x_i) = \alpha' + \beta (X - \bar{x}) \tag{3.16}$$

Ces expressions sont équivalentes si:

$$\alpha' = \mu_Y$$
$$\bar{x} = \mu_X$$
$$\beta = \rho (\sigma_Y/\sigma_X)$$

En tenant compte de la définition de ρ on a

$$\beta = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y} \frac{\sigma_Y}{\sigma_X} = \frac{\text{Cov}(X,Y)}{\sigma_X^2} \tag{3.17}$$

et l'on retrouve la valeur de β (pente de la droite) estimée par la méthode des moindres carrés. Lorsque la distribution jointe de X et Y est normale la distribution conditionnelle de Y pour $X = x_i$ est donc équivalente au modèle 2 et $\beta = \rho\, \sigma_Y/\sigma_X$. Pour un échantillon on a donc.

$$\tilde{y} = \bar{y} + r \frac{S_y}{S_x} (x_i - \bar{x}) \tag{3.18}$$

et

$$b = r \frac{S_y}{S_x} = \frac{S_{xy}}{S_x^2} \tag{3.19}$$

CHAPITRE 4

ANALYSE DE RÉGRESSION:

NON RESPECT DES HYPOTHÈSES ET MAUVAISES UTILISATIONS.

4. ANALYSE DE RÉGRESSION: NON RESPECT DES HYPOTHÈSES ET MAUVAISES UTILISATIONS.

4.1 INTRODUCTION

Dans ce chapitre, nous présentons d'abord les hypothèses de base sous-jacentes à la régression linéaire simple et mettons en relief leur signification physique et leur importance relative.

Parallèlement à la présentation des hypothèses et des tests que l'on peut utiliser pour les vérifier, nous abordons le sujet des transformations de variables qui permettent dans certains cas de se conformer aux hypothèses initiales. Nous avons pu constater que ces transformations sont utilisées très fréquemment mais pas toujours de façon adéquate et c'est pourquoi nous mettons beaucoup d'emphase sur cet aspect. L'objectif général qui doit être poursuivi lorsqu'on transforme les variables est mis en évidence de même que l'interprétation que l'on doit donner aux variables transformées et au modèle qui en résulte.

Le problème issu de la présence des valeurs extrêmes et de leur influence sur l'estimation des paramètres est aussi discuté dans le cadre général de la représentativité des échantillons (population homogène).

Certains sujets traités dans ce chapitre concernent aussi les résutlats de l'analyse de corrélation. Par exemple:

- l'interprétation du coefficient de corrélation lorsque le modèle est non linéaire (hypothèse de linéarité non-respectée);

- interprétation du coefficient de corrélation lorsque les variables ont été transformées;

- effet des valeurs extrêmes sur la valeur du coefficient de corrélation;

- intérêt des quantités SS_r (sommation des carrés des résidus) et r (coefficient de corrélation) pour le choix du meilleur modèle;

Ces sujets (homogénéité des échantillons, valeurs extrêmes et valeurs aberrantes, etc.) sont discutés à la fois dans la section relative à la régression et à la corrélation, en raison des implications qu'ils ont sur le modèle de régression et sur la valeur du coefficient de corrélation.

Les hypothèses dont nous discutons (section 4.2) concernent la variable dépendante Y. Théoriquement on doit supposer cependant que la variable X est mesurée sans erreur ce qui constitue une hypothèse supplémentaire. En raison de la complexité et de la spécificité de cette hypothèse il était difficile d'en discuter en quelques pages et de façon satisfaisante dans le cadre de ce mémoire. Lorsque X est mesuré avec erreur il n'est pas possible dans ce cas de transformer les variables originales pour se ramener aux hypothèses de la régression. On doit alors envisager d'autres méthodes d'estimation des paramètres du modèle linéaire et utiliser des tests différents de ceux présentés aux chapitres précédents.

Dans la littérature scientifique cette hypothèse est discutée dans le cadre des relations structurelles et fonctionnelles et nous référons donc le lecteur à des ouvrages spécialisés (Kendall et Stuart, 1979; Johnston, 1977; Mandel, 1984).

Les exemples présentés aux chapitres 4 et 5 ne sont pas toujours des cas rares. D'une part certaines erreurs méthodologiques sont très fréquentes et nous n'en avons retenu que quelques exemples pour fins d'illustrations. D'autre part beaucoup d'articles scientifiques ne contiennent pas suffisamment d'informations sur les données originales pour que l'on puisse conclure rigoureusement à des applications erronées; mais il n'en demeure pas moins que l'examen graphique de leurs données laisse souvent suspecter de telles erreurs.

4.2 ANALYSE DES HYPOTHÈSES SOUS-JACENTES À LA RÉGRESSION

La vérification des hypothèses qui sous-tendent la théorie de la régression permet de vérifier l'adéquation du modèle et son aptitude à représenter le phénomène étudié.

La vérification des hypothèses peut être faite dans plusieurs cas par des méthodes graphiques. Ces méthodes sont simples d'application et c'est ce qui en fait l'avantage principal. Cependant, ces méthodes ne reposent pas sur des critères quantitatifs et c'est pourquoi, dans certains cas litigieux, leur interprétation est difficile. Lorsque la méthode graphique ne démontre pas clairement le respect ou le non respect d'une hypothèse

quelconque, il est possible, sous certaines conditions, d'appliquer des tests plus rigoureux.

Dans ce qui suit, les principaux tests et méthodes de détection sont présentés. Lorsque les hypothèses de la régression ne sont pas respectées, il est possible, dans certains cas, de transformer les données originales et d'utiliser par la suite la régression pour ajuster un modèle sur les nouvelles variables.

4.2.1 Hypothèse de linéarité

4.2.1.1 Vérification de la linéarité

La pertinence du recours à un modèle linéaire peut ressortir de façon évidente de l'examen graphique des différents couples (x_i, y_i). Si tel n'est pas le cas, une meilleure visualisation peut être obtenue en reportant sur un graphique les différents x_i et les résidus correspondant $e_i = (y_i - \tilde{y}_i)$ où $\tilde{y}_i$ représente l'estimation ponctuelle de y_i déduite de la droite $\tilde{y}$. Les résidus sont reportés de part et d'autre d'une droite horizontale de valeur $e_i = 0$. On conclura à la représentativité d'un modèle linéaire si les résidus sont compris à l'intérieur d'une bande horizontale de part et d'autre de cette droite (voir figure 4.1 A).

Il est souvent plus révélateur d'étudier le graphique des variables (x_i, e_i) plutôt que (x_i, y_i). D'une part, si la relation entre x et y n'est

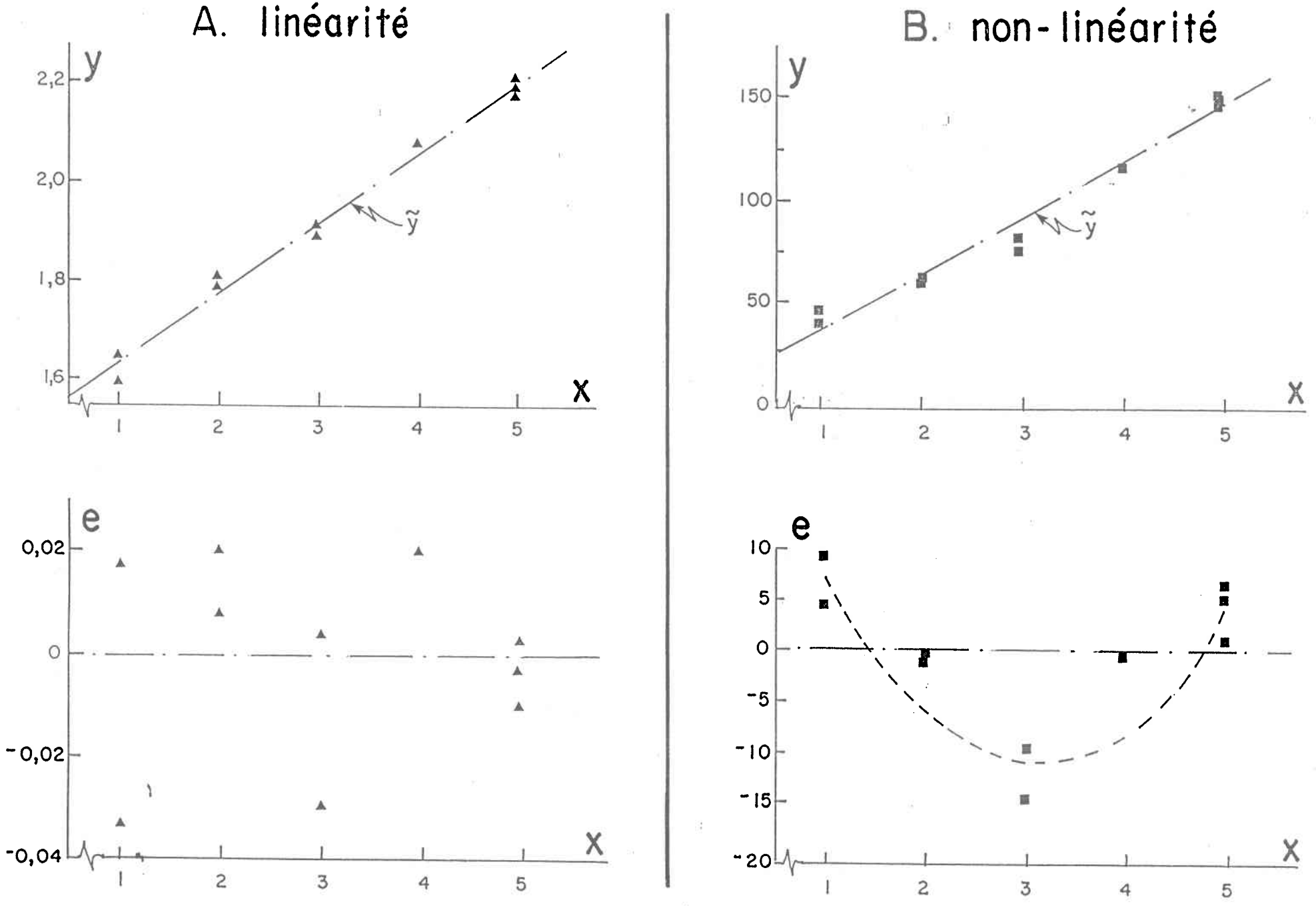


FIGURE 4.1 Examen de l'hypothèse de linéarité

pas linéaire, le graphique des résidus fait souvent ressortir la structure résiduelle et permet dans certains cas d'identifier la fonction appropriée (Neter et Wasserman, 1974). D'autre part, l'écart à la linéarité peut être masquée dans le graphique original (x_i, y_i) du seul fait de l'échelle utilisée. En effet, une faible échelle sur l'axe des y aura pour résultat de comprimer les différents y_i sur la droite $\tilde{y}$, et la présence d'une composante non linéaire peut alors être masquée comme on peut le voir à la figure 4.1 B.

La distribution des résidus ne présente pas toujours un patron très clair pour lequel il est facile de conclure ou non à la linéarité. On peut alors effectuer un test de linéarité (Test F). Les conditions d'applications du test sont les suivantes:

- toutes les autres hypothèses de la régression sont respectées ou supposées respectées;

- on doit avoir au moins trois valeurs différentes de x;

- on doit pour les différents x_i (i = 1 à K) disposer de réplicats. Cependant, il n'est pas indispensable de disposer de réplicats pour tous les niveaux de x (Baillargeon, 1977; Neter et Wasserman, 1974).

Ce test permet d'examiner les hypothèses suivantes:

H_0: $\eta = \alpha + \beta x$ (on accepte l'hypothèse de linéarité)

H_1: $\eta \neq \alpha + \beta x$ (on rejette l'hypothèse de linéarité)

Si le modèle linéaire est adéquat, il est possible de démontrer (Bobée, 1983) que la quantité F_0 définie par:

$$F_0 = \frac{SS_L/(K-2)}{SS_e/(N-K)} \tag{4.1}$$

suit une loi F avec (K-2) et (N-K) degrés de liberté. Le numérateur et le dénominateur du rapport F_0 correspondent respectivement aux quantités suivantes:

$$SS_L = \sum_{i=1}^{K} \sum_{i=1}^{n_i} (\bar{y}_i - y_i)^2 \tag{4.2}$$

$$SS_E = \sum_{i=1}^{K} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 \tag{4.3}$$

où K = nombre de x_i différents

n = nombre de y_i différents pour chacun des x_i;

$N = \sum_{i=1}^{K} n_i$

On trouvera dans Bobée (1983) et Baillargeon et Rainville (1977) des exemples détaillés de ce test. En pratique, il est suggéré Bobée (1983) de calculer SS_L et SS_r (équation 2.27), plus faciles à déterminer que SS_E. La relation qui existe entre ces quantités permet alors de déterminer SS_E. En effet, on a

$$SS_r = SS_E + SS_L \qquad (4.4)$$

On calcule d'abord SS_r. D'après 2.29 on peut déduire que $SS_r = (1 - r^2)\, SS_y$ avec $SS_y = S_y^2 (N-1)$. On détermine ensuite SS_L et on déduit SS_E. Pour un niveau de signification $\tilde{\alpha}$ donné, on compare la quantité F_0 avec la valeur critique $F_{K-2,N-K}(\tilde{\alpha})$ que l'on trouve dans les tables de Fisher.

Si $F_0 \leq F_{K-2,N-K}(\tilde{\alpha})$ on accepte H_0

$F_0 > F_{K-2,N-K}(\tilde{\alpha})$ on rejette H_0

Il est important de souligner que les réplicats pour un niveau x_i donné le sont vraiment si chaque observation est un essai indépendant. Lorsqu'on effectue plus d'une mesure sur un même échantillon, on ne peut parler d'observations multiples véritables. Dans ce cas, pour un x_i donné, chaque mesure y_i n'est pas indépendante de l'observation suivante. En mesurant le même échantillon plusieurs fois, on contrôle la valeur y_i en éliminant la variabilité inhérente à la relation étudiée. En effet, y_i peut être différent de $\tilde{y}_i$ en raison d'erreurs de mesures mais aussi d'erreurs de structure. Or ici en mesurant n_i fois un même échantillon toute variabilité de cette nature (erreur de structure) est éliminée. On ne peut donc dans ce cas appliquer un test de linéarité, puisqu'on ne possède pas de véritables réplicats indépendants.

4.2.1.2 Linéarisation: objectif et principe général

Lorsque l'hypothèse de linéarite n'est pas respectée deux solutions peuvent être envisagées. Ajuster les données par une méthode autre que la régression linéaire, (par exemple: régression non linéaire) ou encore transformer les variables initiales de telle sorte qu'elles respectent l'hypothèse de linéarité. Il est en effet possible, dans certains cas, d'opérer des transformations qui permettent de se ramener à une relation linéaire. Ce faisant, il est alors permis d'utiliser la méthode des moindres carrés pour estimer les paramètres de l'échantillon linéarisé. L'objectif général des transformations est de se ramener au modèle le plus simple possible.

Le principe des transformations qui permettent de linéariser les données est en soi assez simple. Les transformations les plus utiles sont les transformations logarithmiques et les transformations réciproques. Par exemple un modèle multiplicatif de type $y = ax^b$ peut être linéarisé en utilisant le logarithme de la fonction soit:

Fonction théorique : $y = ax^b$

Logarithme de la fonction: $\log y = \log a + b \log x$

En posant: $y' = \log y$

$x' = \log x$

$a' = \log a$

On obtient:

Fonction linéarisée: $y' = a' + bx'$

Plusieurs de ces fonctions linéarisables sont indiquées au tableau 4.1. Nous présentons un exemple numérique qui correspondrait au modèle théorique 1 de ce tableau.

Exemple numérique

On suppose que l'équation théorique du modèle est $y = ax^b$. On utilise les variables $y' = \log y$ et $x' = \log x$. Le modèle linéaire de y', en fonction de x' est:

$$\tilde{y}' = 0{,}3010 + 5x'$$

Le modèle théorique exprimé en fonction de x et y, est déduit en utilisant l'antilog du modèle linéaire. On a:

$$\log \tilde{y} = 0{,}3010 + 5 \log x$$
$$10^{\log \tilde{y}} = 10^{(0{,}3010 + 5 \log x)}$$
$$\tilde{y} = 2{,}0\ x^5$$

Connaissant l'équation de la fonction théorique il serait facile de la linéariser en choisissant la transformation appropriée. Cependant, ces fonctions ne sont pratiquement jamais connues à priori. C'est en effet ce qu'on cherche justement à identifier. Il faut donc dans la plupart des cas

TABLEAU 4.1 Linéarisation de fonctions curvilignes et variables associées.

Fonctions théoriques	Linéarisation (modèle résultant)	Variables à utiliser pour linéariser
1. $y = ax^b$	$\log y = \log a + b \log x$ $y' = a' + bx'$	y' et x' $y' = \log y$ $x' = \log x$
2. $y = ae^{bx}$	$\ln y = \ln a + bx$ $y' = a' + bx$	y' et x $y' = \ln y$
3. $y = ab^x$	$\log y = \log a + \log b^x$ $y' = a' + b'x$	y' et x $y' = \log y$
4. $y = a + b\ (\ln x)$ ou $y = a + b\ (\log x)$	$y = a + bx'$	y et x' $x' = \ln x$ ou $x' = \log x$
5. $y = \frac{x}{ax + b}$	$\frac{1}{y} = \frac{ax + b}{x}$ $\frac{1}{y} = a + \frac{b}{x}$ $y' = a + bx'$	y'et x' $y' = \frac{1}{y}$ $x' = \frac{1}{x}$
6. $y = a + \frac{b}{x}$	$y = a + bx'$	y et x' $x' = \frac{1}{x}$
7. $y = \frac{1}{a + bx}$	$\frac{1}{y} = a + bx$ $y' = a + bx$	y' et x $y' = \frac{1}{y}$
8. $y = ae^{b/x}$	$\ln y = \ln a + \frac{b}{x}$ $y' = a' + bx'$	y' et x' $y' = \ln y$ $x' = \frac{1}{x}$
9. $y = \frac{1}{a + be^{-x}}$	$\frac{1}{y} = a + be^{-x}$ $y' = a + bx$	y' et x' $y' = \frac{1}{y}$ $x' = e^{-x}$
10. $y = a + b \log (c-x)$	$y = a + bx'$	y et x' $x' = \log(c-x)$
11. $y = c + ae^{bx}$	$(y-c) = ae^{bx}$ $\ln (y-c) = \ln a + bx$ $y' = a' + bx$	y' et x $y' = \ln (y-c)$

pour les modèles 10 et 11 on a $c = c_o$

essayer différentes transformations jusqu'à ce que l'on obteinne une relation linéaire.

La méthode par essai et erreur est plutôt fastidieuse et inutilement longue. Deux types de méthodes peuvent être envisagées pour choisir la transformation appropriée: (1) Une méthode graphique (2) une méthode analytique.

4.2.1.3 Méthodes d'approche pour choisir une transformation

Méthode graphique

Du seul examen graphique de la série, on peut d'emblée éliminer certaines transformations. Plusieurs des fonctions possiblement utiles et linéarisables (fonctions intrinsèquement linéaires) sont illustrées à la figure 4.2. Cette figure (voir aussi le tableau 4.1) présente l'équation de la fonction théorique (souvent inconnue), et sa représentation graphique qui tient compte notamment des différentes valeurs du paramètre b. Finalement les variables que l'on devrait utiliser pour linéariser l'une ou l'autre de ces fonctions peuvent être déduites du tableau 4.1. À la lumière de cette figure, le choix d'une transformation peut être "guidé" par un des aspects suivants, qui peut être connu à priori ou que l'on peut observer à partir des données.

- existence d'asymptote verticale et/ou horizontale
- valeurs centrées à l'origine

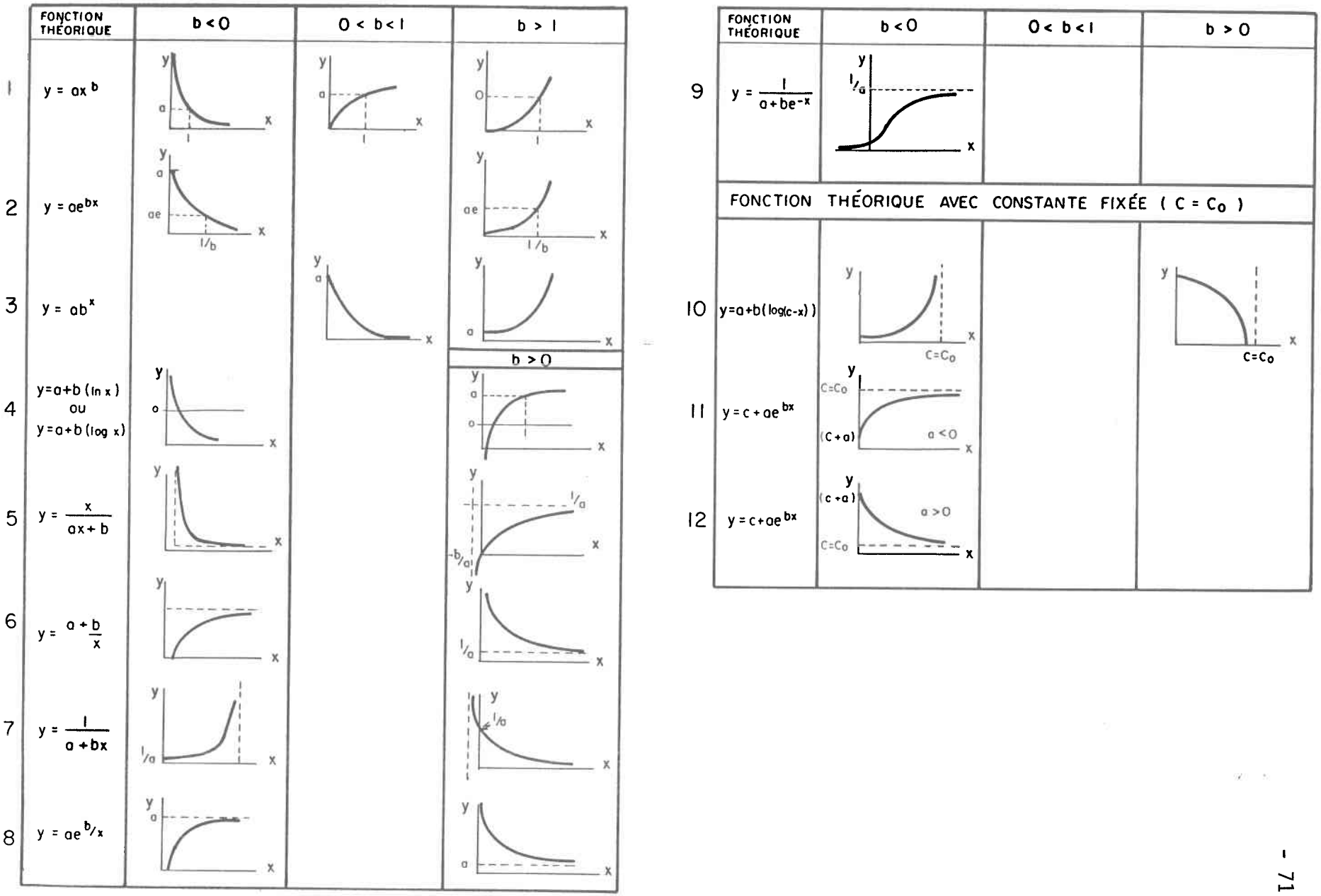


FIGURE 4.2 Illustration de fonctions linéarisables

- concavité ou convexité de la série

La figure 4.2 et le tableau 4.1 présentent une synthèse des fonctions que l'on trouve le plus souvent dans plusieurs ouvrages de référence. Nous avons aussi ajouté d'autres fonctions rarement suggérées et qui n'en demeurent pas moins très utiles. Par exemple la fonction $y = a + b\,(\log(c-x))$ suggérée par Smilie (1966) qui peut être utile si l'on connaît à priori une valeur asymptotique quelconque ou si l'échantillon permet de déduire une telle asymptote.

Méthode analytique

On peut également déterminer par une méthode analytique la transformation la plus appropriée pour les variables x ou y. Ces méthodes ne peuvent être utilisées conjointement cependant. Étant donné que la linéarisation requièrt parfois une transformation simultanée des deux variables, ces méthodes ont donc un intérêt limité à moins que l'on sache a priori laquelle des variables doit être transformée. Box et Tidwell (1962) décrivent une méthode qui permet de trouver la meilleure transformation pour la variable x. Cette méthode requiert l'utilisation de la régression multiple et c'est pourquoi nous n'en discutons pas ici. On trouvera dans Montgomery et Peck (1982) une présentation simplifiée de cette procédure. D'autres références et exemples pourront également être trouvés dans Gunst et Mason (1980) et dans Dolby (1963). On retiendra une méthode suggérée par Box et Cox (1964) pour une transformation de la variable y. Étant donné que, pour respecter l'hypothèse d'homoscédasticité, on doit souvent transformer la variable y, nous reportons à la section 4.2.2 la présentation de cette méthode.

Exemple graphique

Les variables du graphique 1 de la figure 4.3 représentent un échantillon non linéaire. On peut vérifier qu'une droite ajuste mal les données et ce, même si le coefficient de corrélation (linéaire) est élevé. Six transformations furent utilisées pour linéariser les variables. Si on se réfère aux fonctions de la figure 4.2, les transformations suivantes semblent appropriées.

- log y et log x
- x et log y
- x et 1/y
- log(c-x) et y

Deux transformations qui auraient pu être éliminées à priori furent aussi utilisées soit 1/x et y et log x et y (voir modèles 6 et 4, figure 4.2). On peut vérifier pour ces modèles la non linéarité des variables transformées (graphiques 2A et 3A, figure 4.3) et les modèles qui résultent de transformations inappropriées (graphiques 2B et 3B, figure 4.3). On peut aussi noter que même si une transformation peut sembler convenable à priori, on n'obtient pas nécessairement un résultat satisfaisant, ce que démontre d'ailleurs l'utilisation des variables log x et log y (modèle présumé: $y = ax^b$)

Les modèles $y = ab^x$, $y = 1/(a + bx)$, $y = a + b(\log(c-x))$ ajustent bien les observations (graphiques 5B, 6B et 7B, figure 4.3). Le choix de l'un ou l'autre devra être fait en fonction de critères quantitatifs ou encore à la

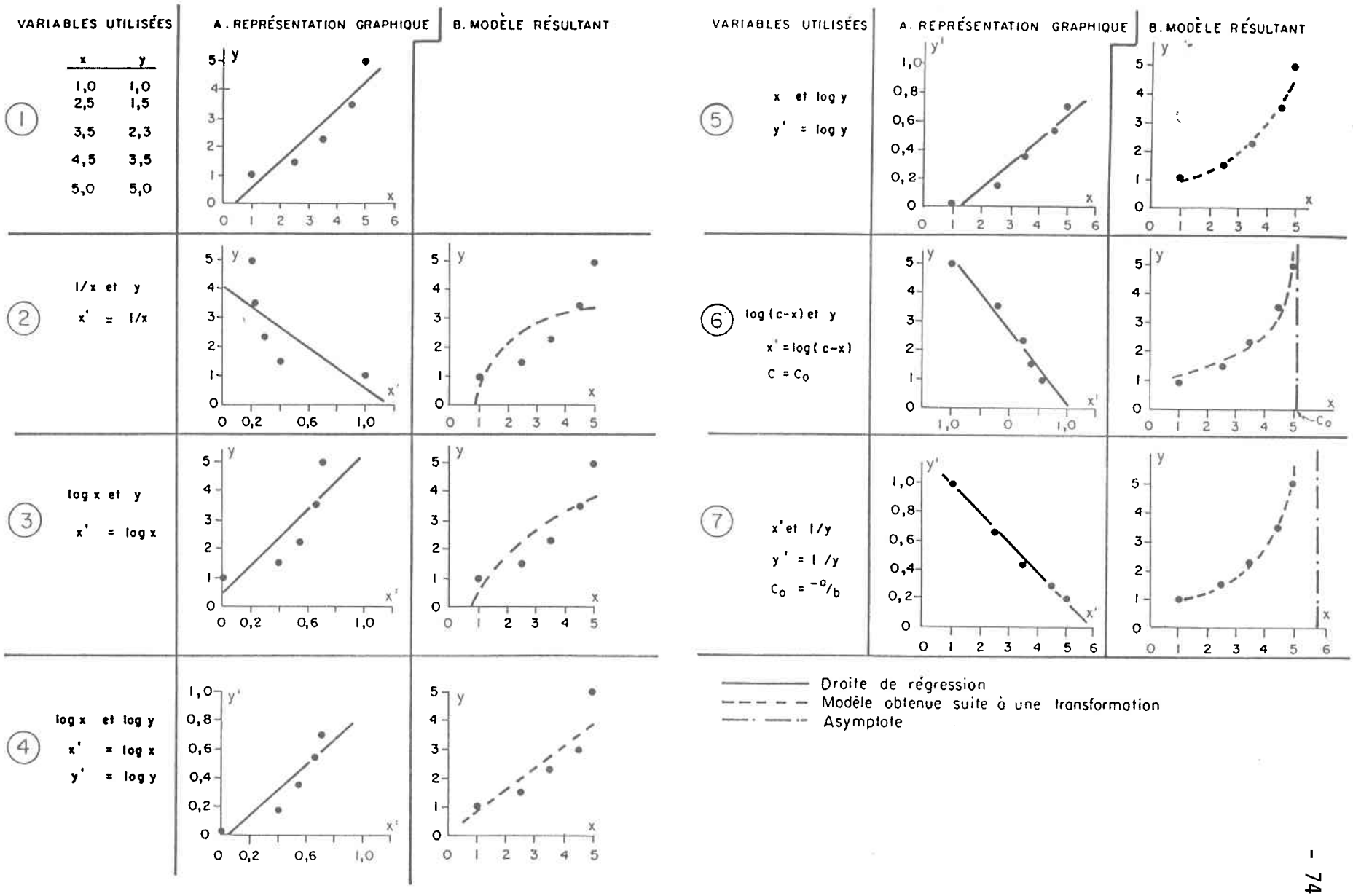


FIGURE 4.3 Illustration de la linéarisation par différentes transformations

lumière de certaines connaissances à priori que l'on peut avoir du phénomène étudié.

4.2.1.4 Choix du meilleur modèle

Si plusieurs transformations permettent d'obtenir un modèle satisfaisant, on devra tenir compte des caractéristiques du modèle ou encore de la qualité prédictive du modèle représentée par la quantité SS_r (équation 2.8) pour le choix de l'une ou de l'autre des transformations.

(a) caractéristiques du modèle

Le choix final de l'un des modèles peut être décidé en fonction de certaines caractéristiques. Valeurs centrées à l'origine, modèle asymptotique, etc. Par exemple on pourra retenir le modèle 7B (figure 4.3) si une asymptote de valeur -a/b semble appropriée. Cependant, on peut supposer que le modèle est asymptotique mais que des asymptotes de valeur -a/b ou de valeur $c = c_0$ ne sont pas du tout plausibles pour le phénomène considéré. On pourra retenir alors le modèle dont la valeur prédictive est la meilleure pour les données observées.

b) qualités prédictives du modèle (quantité SS_r)

Lorsque plusieurs transformations permettent de linéariser les différents couples x',y' et qu'aucune connaissance à priori ne permet de privilégier un modèle plutôt qu'un autre, le modèle pour lequel la quantité SS_r est

minimum devra être retenu. Ce modèle aura en effet la meilleure valeur prédictive. Il n'est pas inutile de noter que les quantités SS_r sont calculées sur le graphique des différents modèles résultant des transformations utilisées. En effet, la quantité SS_r n'est d'aucune utilité sur les grahiques des variables transformées (SS_r") puisque ces dernières sont reportées selon des échelles différentes.

Commentaires et exemples

Comparaison entre SS_r, r^2 et r'^2

Plusieurs exemples tirés de la littérature démontrent que l'on utilise très souvent le coefficient de corrélation ou le coefficient de détermination comme critère d'appréciation du meilleur modèle suite à différentes transformations. Comme on l'a vu aux sections 2.3.7 et 3.4 le coefficient de détermination et le coefficient de corrélation sont dans certains cas des outils complémentaires à l'analyse de régression. L'équivalence numérique entre SS_r et r soit $SS_r = (1 - r^2)\ SS_y$ (équation 2.29) permet de constater que plus la variance expliquée par le modèle linéaire est élevée et plus la variabilité résiduelle SS_r est faible.

Lorsqu'on transforme la variable y cette équivalence n'est plus vraie. Dans ce qui suit, on pourra comparer pour l'exemple précédent (figure 4.3) les quantités SS_r, $SS_{r'}$, et $SS_{r''}$ et en tirer certaines conclusions quant à r^2 (coefficient de détermination des variables originales) et $(r')^2$ (coefficient de détermination des variables transformées x' et y'). Ces quantités sont ainsi définies:

$$SS_r = \sum_{i=1}^{N} (y_i - \tilde{y}_i)^2 \text{ ou encore } SS_r = (1 - r^2)\, SS_y$$

$(y_i - \tilde{y}_i)$ correspond à la différence entre la variable observée y_i et son estimation ponctuelle déduite de l'un des modèles illustrés aux graphiques 2B à 7B (figure 4.3).

$$SS_{r'} = 1 - (r')^2\, SS_y$$

$(r')^2$ correspond au coefficient de détermination des variables transformées et

$$SS_y = \sum_{i=1}^{N} (y_i - \bar{y})^2 = S_y^2\,(N-1)$$

$$SS_{r''} = \sum_{i=1}^{N} (y'_i - \tilde{y}'_i)^2 = 1 - (r')^2\, SS_{y'}$$

où $$SS_{y'} = \sum_{i=1}^{N} (y' - \bar{y}')^2 = S_{y'}^2\,(N-1)$$

Au tableau 4.2 le calcul de ces différentes quantités a été reproduit. La quantité $SS_{r''}$ n'est pas indiquée lorsque la transformation affecte la variable indépendante uniquement. En effet, dans ce cas, y' n'existe pas et l'on a donc $SS_{r'} = SS_{r''}$. Des résultats de ce tableau, on peut tirer les conclusions suivantes:

- La quantité $SS_{r''}$ est peu utile puisque la sommation des résidus au

TABLEAU 4.2

Comparaison des quantités r^2 , $SS_{r'}$ et $SS_{r''}$. (variables originales et variables transformées).

	Variables originales	Modèles ($\tilde{y}$)	r	r^2	$\sum_{i=1}^{N}(y_i-\tilde{y})^2 = (1-r^2)SS_y$		
1	x, y	$\tilde{y}$ =-0,4350 +0,9379 x	0,9328	0,8701	1.3522	1.3522	
	Variables transformées	Modèle ($\tilde{y}$)	r'	$(r')^2$	$SS_r = \sum_{i=1}^{N}(y_i-\tilde{y}_i)^2$	$SS_{r'} = 1-(r')^2 SS_y$	$SS_{r''} = 1-(r')^2 SS_{y'}$
2	x' = 1/x	$\tilde{y}$ = 4,16 - 3,55 x'	-0,732	0,537	4,82	4,82	
3	x' = log x	$\tilde{y}$ = 0,450 + 4,82 x'	0,841	0,706	3,06	3,06	
4	x' = log x et y' = log y	$\tilde{y}'$=-0,069 + 0,927 x' ou $\tilde{y}$ = 0,853 ($x^{0,927}$)	0,934	0,872	1,91	1,33	0,040
5	y' = log y	$\tilde{y}'$=-0,212 + 0,172 x ou $\tilde{y}$ = 0,614 ($1,49^x$)	0,989	0,977	0,378	0,236	0,007
6	x' = log (c - x) c = 5,1	$\tilde{y}$ =2,67 - 2,50 x'	-0,992	0,985	0,161	0,161	
7	y' = 1/y	$\tilde{y}$ =1,18 - 0,200 x ou $\tilde{y}$ = 1/(1,18 -0,200 x)	-0,996	0,993	0,484	0,076	0,003

carré des modèles 1A et 7A n'est pas comparable étant donné que les variables sont à des échelles différentes.

- L'examen de la quantité $SS_{r'}$ permet de constater que lorsque la variable y est transformée, les quantités SS_r et $SS_{r'}$ ne sont pas identiques, c'est-à-dire $\sum_{i=1}^{N} (y_i - \tilde{y}_i)^2 \neq 1 - (r')^2 \, SS_y$. Par exemple, à la quantité SS_r minimum (modèle 6) ne correspond pas le coefficient de détermination $(r')^2$ maximum (modèle 7).

La quantité SS_r ne peut donc pas être remplacée par $(r')^2$ comme critère de comparaison de différents modèles. Au chapitre 5, nous discuterons de façon plus explicite de la signification et de l'utilisation des coefficients de détermination et de corrélation après une transformation de variables, soit:

(1) comparaison de $(r')^2$ et SS_r pour plusieurs modèles ajustés à un même échantillon,

(2) comparaison de $(r')^2$ et SS_r pour différents échantillons auxquels on a ajusté un modèle identique.

EXEMPLE 4.1 (linéarisation)

Les données de l'exemple qui suit sont tirés d'un manuel de référence en chimie de l'eau (Hem, 1971). Les données de la figure 4.4 représentent

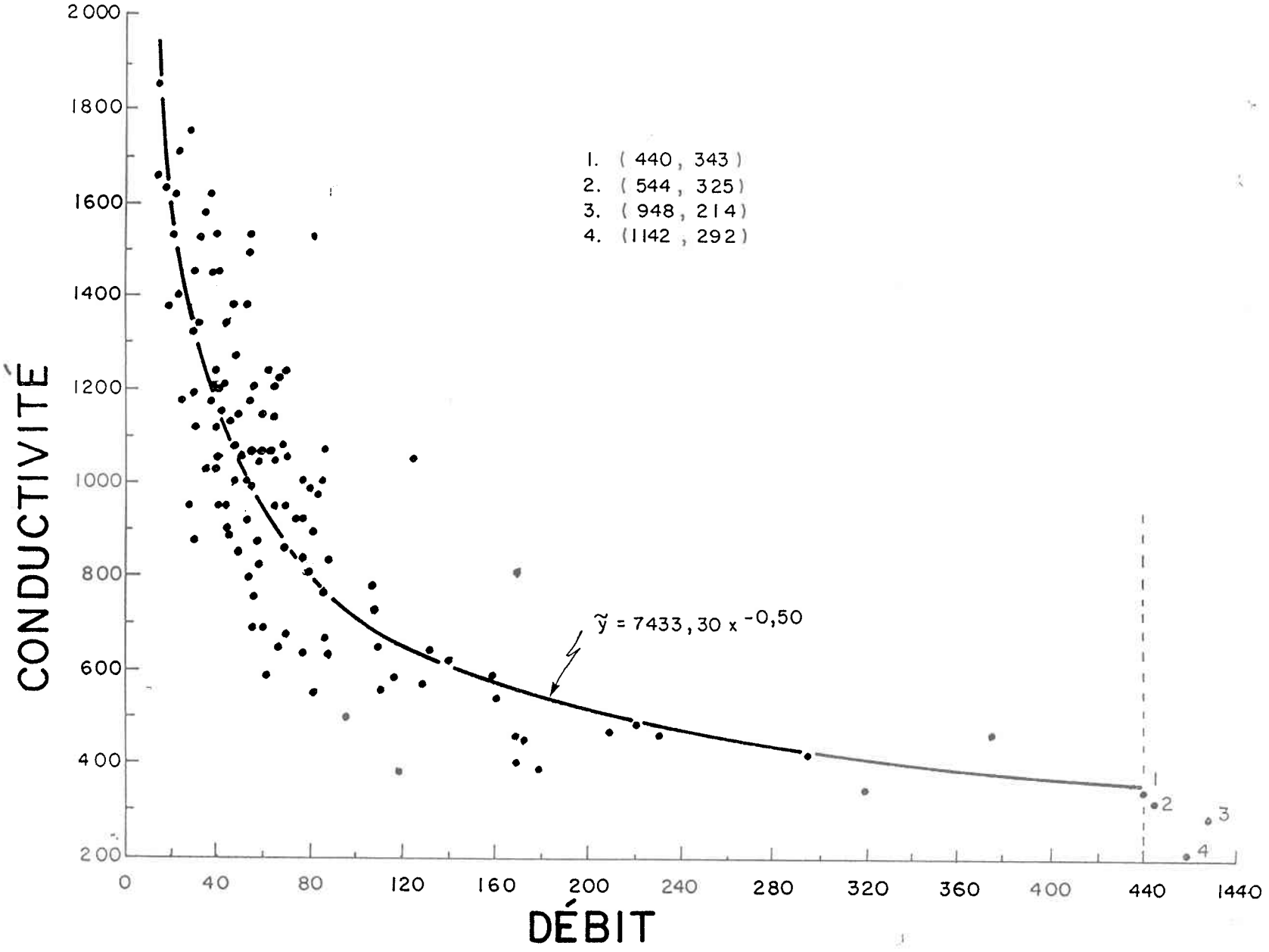


FIGURE 4.4 Relation entre le débit et la conductivité (exemple 4.1)
Données reconstituées d'après Hem (1971).

le débit (Q) et la conductivité (C) mesurés dans une rivière pendant une année. Les variables furent reconstituées à partir d'un graphique à échelles logarithmiques. On peut constater qu'un modèle linéaire serait peu adéquat pour représenter l'échantillon de Q,C. En utilisant les variables $x' = \log C$ et $y' = \log Q$ (figure 4.5), nous avons ajusté un modèle linéaire (4.5A). On obtient l'équation $\tilde{y}' = 3,8712 - 0,5003x'$

L'examen de la figure 4.5A, de même que l'étude des résidus de l'équation obtenue, permet de confirmer la linéarité des observations transformées. On peut donc conclure que la transformation utilisée est adéquate, du moins pour l'échantillon des valeurs dont on dispose.

L'auteur ajuste une courbe (figure 4.5B) (aucune équation présentée) qui selon lui convient mieux à l'échantillon qu'un modèle linéaire. "The scatter of point is substantial and it would be difficult to fit them to a straight line." Cette courbe tient compte de 2 caractéristiques du modèle théorique présumée. On suppose que pour $C \approx 0$, $Q = 16000$ et que pour $C > 1000$, $Q = 200$ = constante.

Il n'est pas de notre ressort de vérifier si ces valeurs ont un certain fondement théorique. Comme il fut discuté au chapitre 2, on peut construire un modèle en tenant compte d'une valeur fixée à l'avance ou encore utiliser un modèle asymptotique. Encore faut-il vérifier que ce modèle est compatible pour les observations mesurées. Or, dans ce cas, pour des valeurs de concentrations très faibles, les valeurs mesurées sont très éloignées de la valeur limite suggérée. On a en effet pour $C \approx 15$ (qui correspond à la plus faible valeur mesurée) $Q = 1600$. De plus, 98% des valeurs de débits obser-

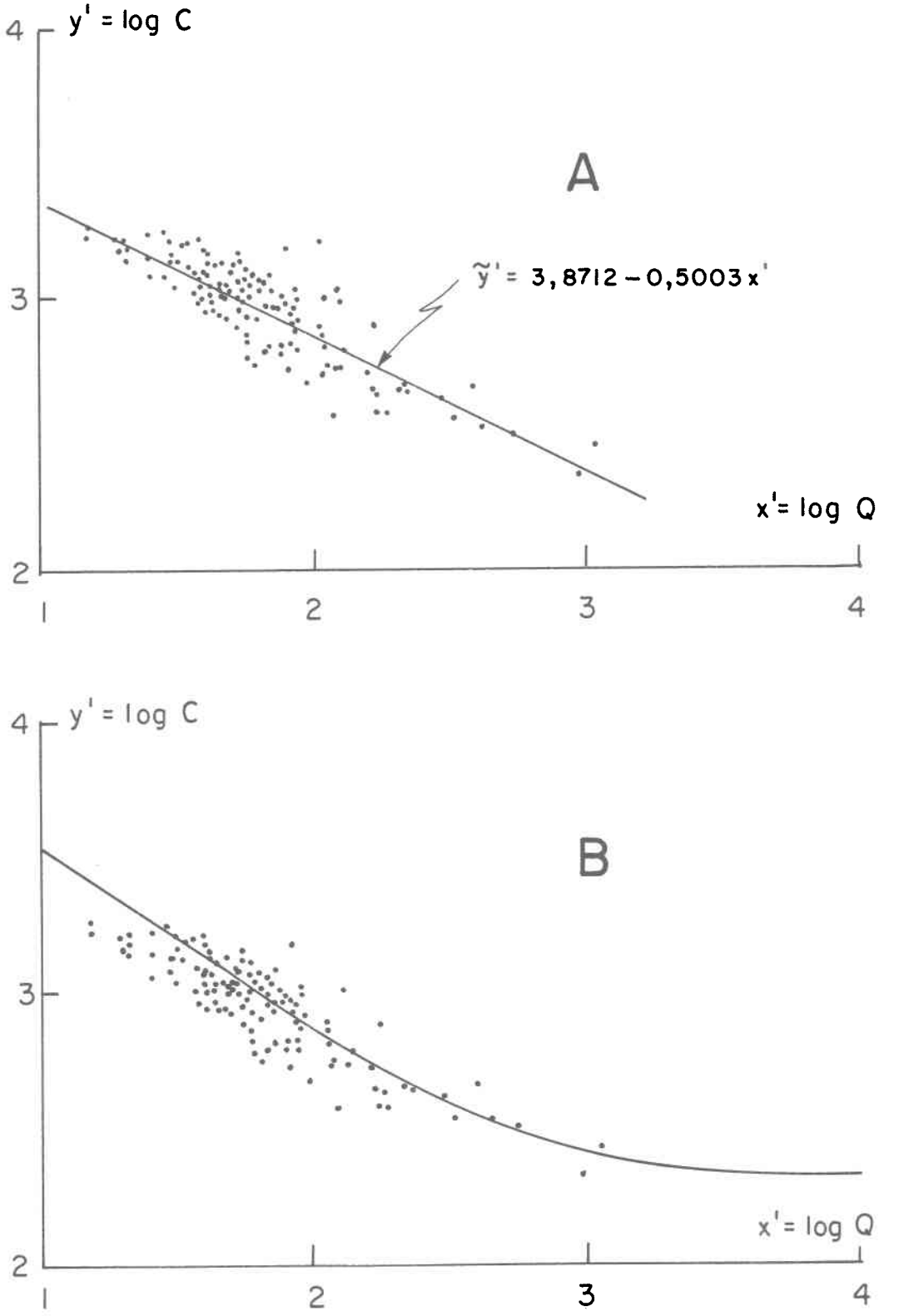


FIGURE 4.5 Relation entre le débit et la conductivité; échantillon linéarisé . A: Ajustement d'une droite. B: Modèle ajusté manuellement (Hem, 1971).

vées sont inférieurs à 440. Or le modèle fixe les valeurs de concentration à 200 pour Q > 1000, valeur de débits pour laquelle une seule observation fut mesurée (Q ≈ 1240, C = 300). Il s'agit donc là d'extrapolation plutôt risquée. Selon que l'on considère que l'échantillon est représentatif ou non, on pourra selon nous conclure de la façon suivante.

(1) L'échantillon est représentatif.

L'utilisation des variables log Q et log C linéarise les variables et par suite le modèle $\tilde{y}' = a + bx'$ ($\tilde{y} = ax^b$) est valable pour ajuster l'échantillon dont on dispose. Le modèle exprimé en fonction de x et y est reproduit à la figure 4.5A.

(2) L'échantillon est non-représentatif.

Puisque le modèle théorique proposé par l'auteur ajuste mal l'échantillon, on devrait donc conclure que l'échantillon est peu représentatif pour l'étude de la relation entre le débit et la conductivité. On peut d'ailleurs remarquer que pour la courbe suggérée par l'auteur, environ 70% (figure 4.5 B) des observations sont sous la courbe.

EXEMPLE 4.2 (linéarisation)

L'échantillon de la figure 4.6A met en relation le nombre de prises (y) d'une espèce piscicole en fonction de l'âge (x). Les variables y' = ln y et x utilisées par l'auteur (Winters, 1983) sont présentées à la figure 4.6B. Comme on peut le voir, la relation n'est pas linéaire et la transformation

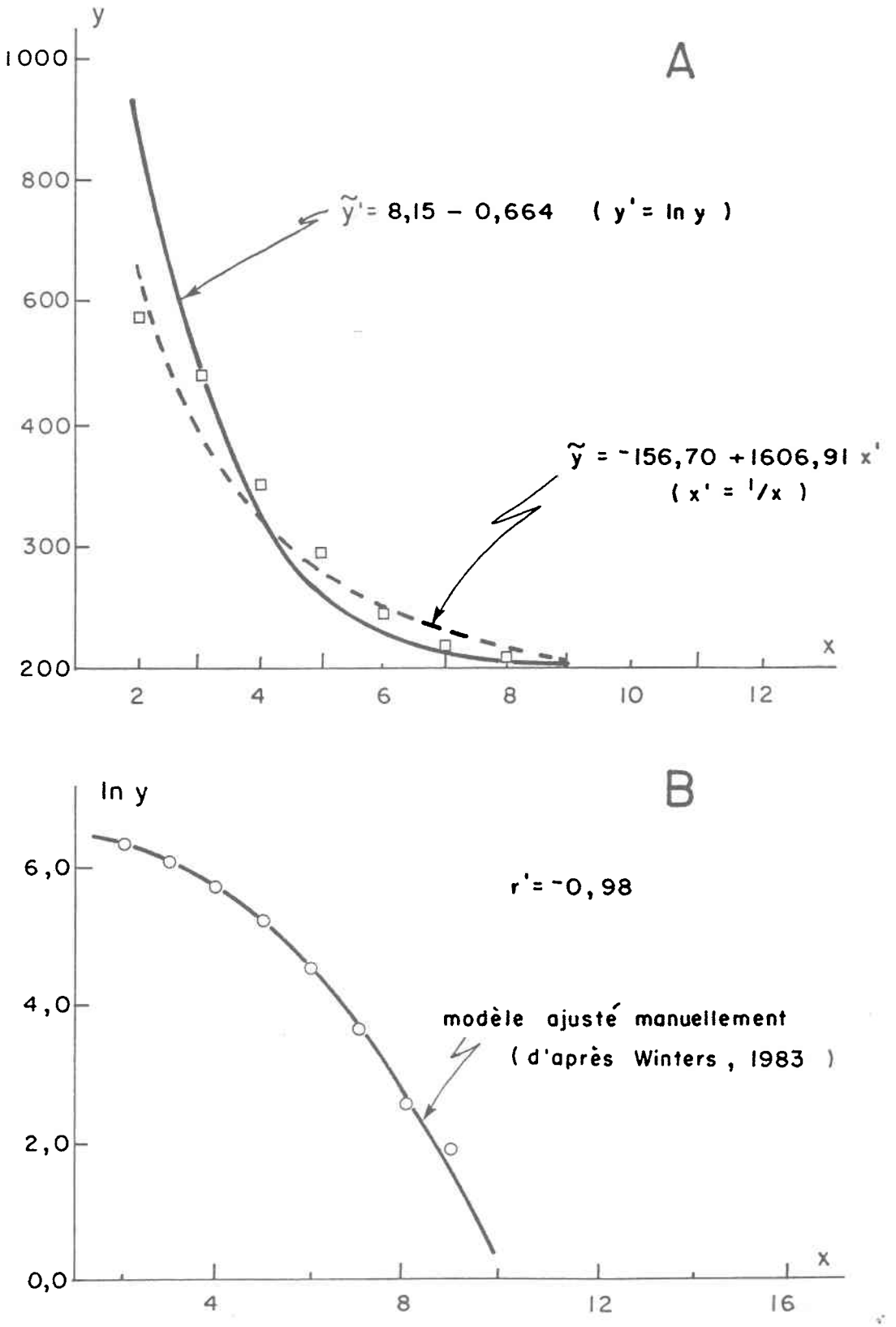


FIGURE 4.6 Comparaison des différents modèles pour l'exemple 4.2
Données tirées de Winters (1983).

utilisée est donc peu adéquate. Nous avons vérifié qu'aucune des transformations simples décrites à la figure 4.2 ne permettait d'obtenir une linéarisation parfaite. Cependant, si l'on s'en tient à l'un de ces modèles, le modèle ln y = a + bx n'est certainement pas le plus adéquat. En utilisant les variables y et x' = 1/x, on obtient en effet un modèle pour lequel la quantité SS_r est beaucoup plus faible. On a:

$\tilde{y} = 8,15 - 0,664\ x$ $\quad (\tilde{y}' = \ln y)$ $\quad SS_r = 157867,83$

$\tilde{y} = -156,70 + 1606,91x'$ $\quad (x' = 1/x)$ $\quad SS_r = 19720,30$

Comme on l'a dit, le coefficient de corrélation des variables transformées, soit r', ne peut être utilisé comme critère pour vérifier l'adéquation du modèle ou encore pour comparer plusieurs modèles entre eux. On peut d'ailleurs vérifier ici que le coefficient de corrélation le plus élevé correspond au modèle logarithmique (r' = -0,98) pour lequel la quantité SS_r est aussi la plus élevée. En utilisant les variables 1/x et y on obtient r' = 0,97.

4.2.1.5 Tests et intervalles de confiance (Variables originales et variables transformées)

Si la linéarisation semble acceptable et que les autres hypothèses sont vérifiées, il est possible d'utiliser les tests et de construire les intervalles de confiance associés à la droite.

L'hypothèse concernant l'indépendance des erreurs ne dépend pas du modèle utilisé, mais du mode d'échantillonnage. Il n'est donc pas intéressant d'en discuter dans le contexte de linéarisation. En théorie, l'hypothèse de la normalité des résidus doit être vérifiée. Cependant, si l'on ne s'en éloigne pas trop et que l'échantillon est grand, on peut appliquer les résultats concernant les tests et les intervalles de confiance (Bobée, 1983). En pratique, cette hypothèse n'est donc pas très restrictive. De plus, l'écart à la normalité est souvent due au fait que la variance résiduelle n'est pas constante. Il est donc suggéré (Neter et Wasserman, 1974) de vérifier d'abord l'hypothèse concernant la variance des résidus, hypothèse beaucoup plus restrictive en pratique.

Il ne faut pas perdre de vue que les hypothèses doivent être vérifiées pour le modèle transformé. Une variance résiduelle constante pour les variables transformées n'équivaut pas en général à variance constante pour le modèle original.

EXEMPLE 4.3

Supposons l'utilisation des variables x et y' = log y et les résidus $e'_i = y'_i - \tilde{y}'_i$. Si les résidus e'_i présentent une variance constante quelque soit x, l'intervalle de confiance autour de $\tilde{y}'$ peut être construit et les résultats transposés. Comme les valeurs des résidus sont logarithmiques, une variance constante. (variables x et log y) reflète en fait une variance résiduelle proportionnelle à x pour l'échantillon original (variables x et

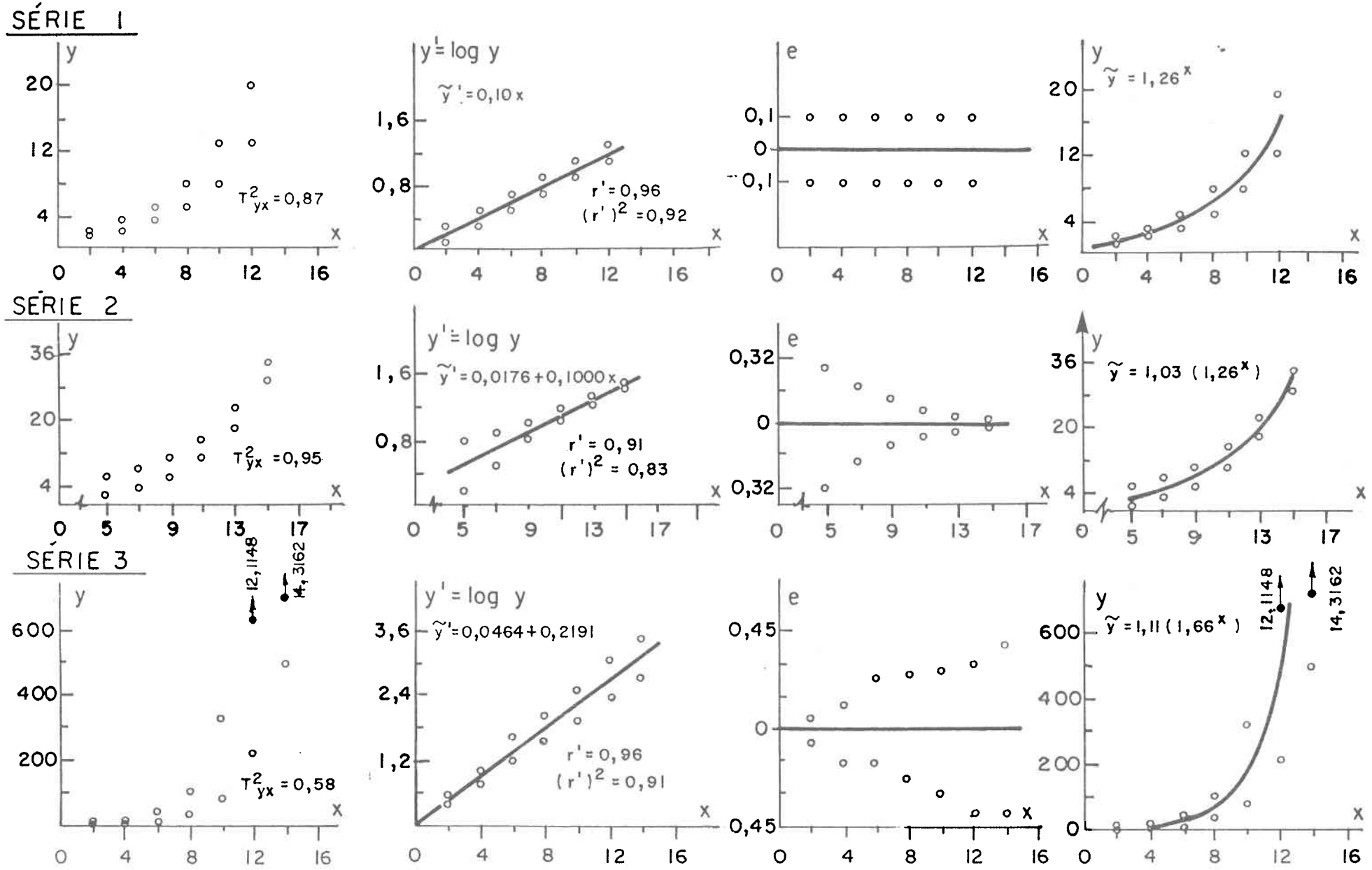


Figure 4.7 Présentation des données de l'exemple 4.3

y), variance dès lors minimum à l'origine et croissante avec x. Un tel échantillon est présenté à la figure 4.7, série 1. De même, si le patron résiduel de e'_i présente une variance décroissante en fonction de x (série 2, figure 4.7), on ne doit pas en conclure que, dans le modèle original, la dispersion de la variable dépendante diminue quand x augmente. Au contraire, une décroissance de la variance en fonction de x peut refléter une variance constante dans le modèle original (série 2, figure 4.7) et même une variance croissante. On doit aussi envisager le cas où l'on obtient pour les résidus une variance croissante avec x. Étant donnée la transformation utilisée ici, une variance croissante avec x (série 3) reflète cette fois une variance croissante avec x dans le modèle original, variance souvent beaucoup plus importante que ne le laissent supposer les variables transformées.

Dans ce contexte (variables transformées) une variance non constante n'affecte pas nécessairement la validité du modèle. Dans certains cas, cela peut refléter un modèle dont la valeur prédictive est très bonne (voir série 2, figure 4.7). Lorsque les variables transformées ne présentent pas une variance constante, on doit en conclure que le patron résiduel est tel qu'il n'est pas possible d'utiliser les tests et de construire les intervalles de confiance à partir des résultats de la régression, ce qui n'invalide pas le modèle obtenu après transformation.

Comme il n'est pas toujours facile d'imaginer l'échantillon original à partir des variables transformées, on se méfiera lorsque seules ces dernières sont présentées dans l'étude d'une relation entre les variables. Cette procédure, très souvent utilisée, masque souvent une dispersion telle que le

modèle résultant n'a pratiquement aucun intérêt pour estimer y. Plusieurs exemples réels seront présentés.

La présentation unique des variables transformées peut donc être trompeuse. Nous avons aussi remarqué l'utilisation fréquente du coefficient de corrélation r' et de détermination $(r')^2$ comme mesure de la forte dépendance entre les variables x et y, alors que r' mesure seulement l'interdépendance entre les variables transformées.

À partir de l'exemple de la figure 4.7, on pourra aussi vérifier le peu d'intérêt du coefficient de corrélation r' ou de $(r')^2$ pour la comparaison de plusieurs échantillons dont les variables sont transformées. Ces discussions et les exemples sont reportées au chapitre 5. On pourra y vérifier d'ailleurs l'intérêt d'utiliser le rapport de corrélation T^2_{yx}.

4.2.2 Hypothèse d'homoscédasticité (variance constante)

Une autre hypothèse qu'il est important de vérifier concerne la constance de la variance de la distribution conditionnelle de y pour $x = x_i$ soit $\sigma^2_{y|x_i}$. Lorsque cette variance est non constante, elle est alors dite hétéroscédastique. On peut toujours estimer les paramètres de la distribution par la méthode des moindres carrés. Les estimateurs de α et β sont alors non biaisés et efficaces. Cependant, ils ne possèdent plus une variance minimum, c'est-à-dire qu'ils ne sont pas exhaustifs (Neter et Wasserman, 1974), ce qui de ce fait invalide l'utilisation des tests et la construction des intervalles de confiance et de prévision.

4.2.2.1 Vérification de l'hypothèse d'homoscédasticité

Plusieurs tests permettent de vérifier l'hypothèse d'homoscédasticité. Très souvent cependant les observations présentent un patron caractéristique qui suggère de façon évidente une variance non constante. On aura par exemple une variance nettement croissante ou décroissante en fonction de x. Pour des cas moins évidents, on pourra vérifier cette hypothèse par une méthode graphique. Certains tests permettent aussi de vérifier si la variance est constante.

La méthode graphique est très simple. Elle consiste à reporter en abcisse les résidus $e_i = (y_i - \tilde{y}_i)$ en fonction de x. L'étude des résidus e_i permet donc de vérifier simultanément l'hypothèse de linéarité et d'homoscédasticité. Quelques patrons caractéristiques sont présentés à la figure 4.8. On conclut à l'homoscédasticité si les résidus e_i sont compris à l'intérieur d'une bande parallèle à l'axe des x (figure 4.8A), c'est-à-dire que la valeur des résidus positifs et négatifs sont approximativement constants et ne varient pas de façon systématique en fonction de x. S'il est difficile de conclure à partir du graphique, on peut utiliser certains tests pour vérifier l'égalité de la variance. Selon que l'on dispose ou non de réplicats, on appliquera le test de Bartlett ou le test F.

Test de Bartlett

Ce test peut être utilisé si l'on dispose de réplicats. On calcule pour chacun des x_i; i=1 à K la variance des y_{ij} correspondants; j = 1 à n_i.

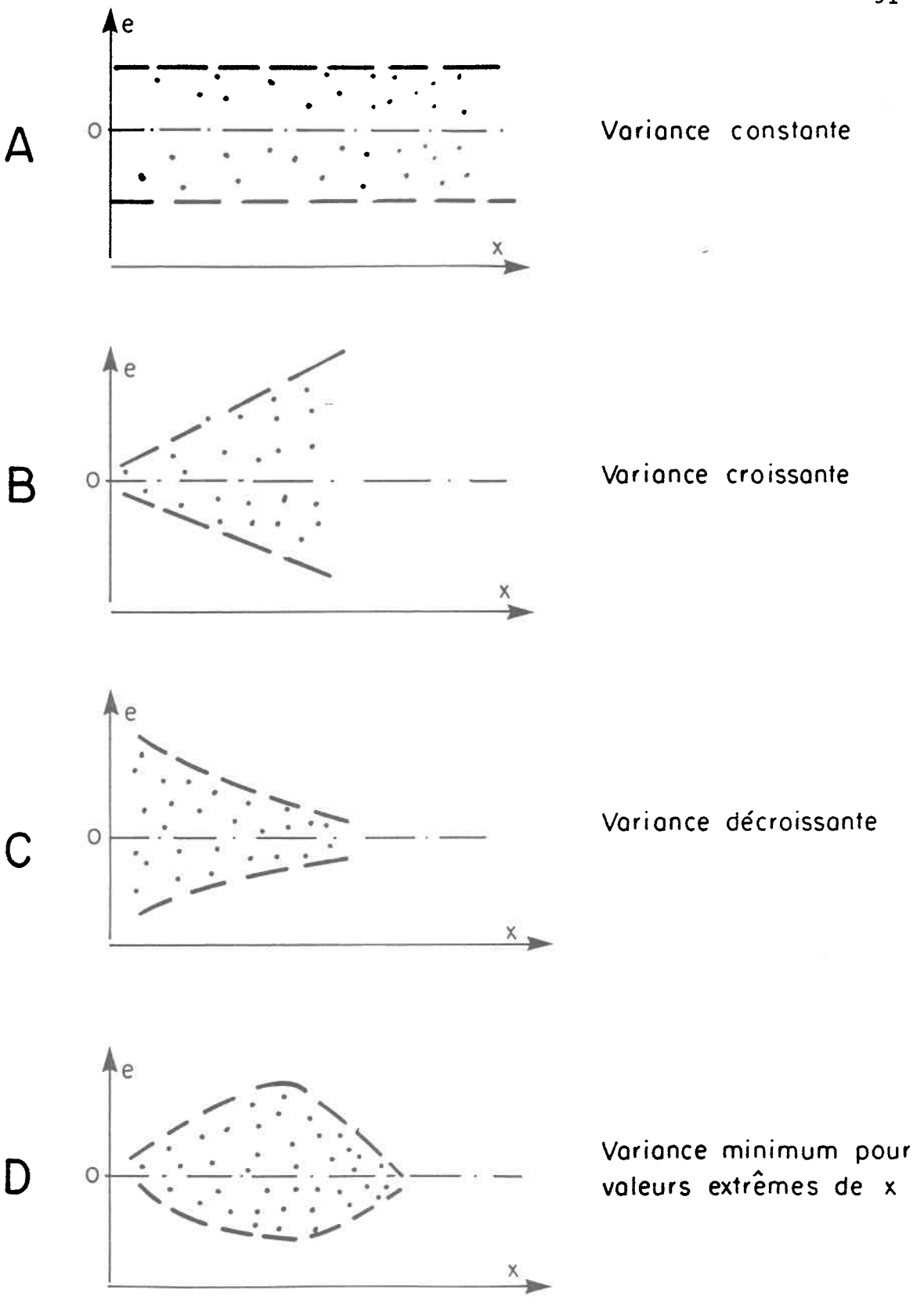


Figure 4.8 Examen de l'hypothèse d'homoscédasticité.

Il s'agit alors de vérifier l'égalité des K variances ainsi obtenues.

La qualité de ce test dépend de la taille des échantillons; en pratique, on considère que $n_i > 5$ est acceptable (Scherrer, 1984). De plus, ce test est également sensible au non-respect de l'hypothèse de normalité.

L'utilisation du test est donc assez limitée. D'une part on doit disposer d'au moins 5 replicats pour chacun des x_i. D'autre part parce qu'une variance non constante (ou que l'on soupçonne non-constante) est souvent observée lorsque la distribution de y suit une distribution autre que la distribution normale. Ce test est néanmoins utile dans certains cas et est présenté en détail à l'appendice 5.

Test F

Lorsqu'on ne dispose pas de réplicats on peut diviser l'échantillon en deux ou plusieurs sous-groupes. En supposant deux sous-groupes N_1 et N_2, le premier serait constitué par les N/2 plus faibles valeurs de x et le second par les N/2 plus fortes valeurs de x. En supposant un nombre pair d'observations, on aura $N_1 = N_2$. Si le nombre est impair la valeur médiane pourra être associée à l'un des groupes, les variances S_e^2 (EQ. 2.27) correspondant à chacun des groupes.

On calcule ensuite le rapport des variances F_0 que l'on compare à une valeur critique F_c (table de Fischer). Par exemple pour un niveau de signification $\tilde{\alpha}$ et un test unilatéral et en supposant $S_{e_2}^2 > S_{e_1}^2$ on concluera à une varian-

ce significativement différente si $F_0 > F_c$

avec $F_0 = S_{e_2}^2 / S_{e_1}^2$

et $F_c = F_{\nu_2,\nu_1}(\tilde{\alpha})$ ν_2 = d.d.l.du numérateur

ν_1 = d.d.l. du dénominateur

Il est important de noter que le test dépend du nombre de degré de liberté des groupes N_1 et N_2 et donc de la taille de l'échantillon (N). Pour de petits échantillons, des différences importantes entre les variances ne conduisent pas nécessairement au rejet de l'hypothèse nulle (H_0: $\sigma_1^2 = \sigma_2^2$). Par exemple pour un échantillon de 8 observations ($N_1 = N_2 = 4$) et un niveau $\tilde{\alpha} = 0,025$ (test unilatéral) le rapport des variances F_0 doit être supérieur à 39,2 (et à 19,2 si le niveau choisi est $\tilde{\alpha} = 0,05$) pour que l'on puisse conclure à une différence significative. Pour un échantillon de 16 observations ($N_1 = N_2 = 8$) la valeur critique sera égale à 6,98 au niveau $\tilde{\alpha} = 0,025$ et à 4,95 au niveau $\tilde{\alpha} = 0,05$.

4.2.2.2 Causes de l'hétéroscédasticité

La distribution de y sera nécessairement hétéroscédastique si la variance de y est fonctionnellement reliée à la moyenne (Montgomery et Peck, 1982). C'est le cas par exemple si y suit une distribution de Poisson ou une distribution binôminale. On observera aussi une distribution hétéroscédastique si l'erreur de mesure sur y croît en fonction de x.

4.2.2.3 Stabilisation de la variance

De même que pour une relation non-linéaire on peut transformer les variables pour les ramener à une droite, certaines transformations permettent de stabiliser la variance, ce qui valide l'utilisation des tests pour les paramètres α et β et permet aussi de construire l'intervalle de confiance associé aux paramètres et à la droite. La variance est stabilisée si les résidus e_i des variables transformées sont constants quelque soit x. On devra envisager deux types de transformations fort différente selon que:

(1) La distribution est hétéroscédastique parce que Y ne suit pas une distribution normale.

(2) La distribution de Y est normale mais l'erreur de mesure croît (ou décroît) avec x.

- Y ne suit pas une distribution normale

On sait par exemple que pour une distribution de Poisson la moyenne est égale à la variance. On aura donc, si Y croît en fonction de X une moyenne ($\bar{y}_i$, pour $x = x_i$) croissante et une variance (S_i^2 pour $x = x_i$) elle aussi croissante. Le tableau 4.3 met en relief le type de transformation approprié en fonction de la relation qui existe entre la moyenne et la variance, et spécifie la distribution correspondante.

TABLEAU 4.3 TRANSFORMATIONS POUR STABILISER LA VARIANCE [1]

Relation entre la variance de Y (σ_Y^2) et la moyenne E(Y)		Transformations appropriées
σ_Y^2 = constante	distribution normale	aucune transformation
$\sigma_Y^2 = E(Y)$	distribution de Poisson	$Y' = \sqrt{Y}$
$\sigma_Y^2 = E(Y)\left[1-E(Y)\right]$	distribution Binominale pour $0 < Y_i < 1$	$Y' = \text{Sin}^{-1}\sqrt{Y}$
$\sigma_Y^2 = \left[E(Y)\right]^2$		$Y' = \ln Y$
$\sigma_Y^2 = \left[E(Y)\right]^3$		$Y' = Y^{-\frac{1}{2}}$
$\sigma_Y^2 = \left[E(Y)\right]^4$		$Y' = Y^{-1}$

(1) tiré de Montgomery et Peck (1982).

- L'erreur de mesure est fonction de x

Même si la distribution conditionnelle de y par rapport à x est normale, on peut aussi observer une variance non-constante dans le cas où la précision des mesures de y n'est pas toujours la même. Par exemple, si l'on mesure la concentration d'un élément chimique (y) en fonction du pH(x), il est possible que les mesures soient d'autant moins précises que les concentrations sont fortes. L'écart-type de la distribution conditionnelle de y_i (ou de e_i) est alors proportionnelle à x, soit:

$$\sigma_{y|x_i} = Cx_i \qquad C = \text{constante}$$

ou encore

$$\sigma^2_{y|x_i} = C' x_i^2 \qquad \text{avec } C' = C^2$$

De manière plus générale, on aura:

$$\sigma^2_{y|x_i} = C\, h(x) \qquad h(x) = \text{fonction de } x$$

Comme on le sait, la méthode des moindres carrés consiste à minimiser la quantité $SS_r = \sum_{i=1}^{N} (y_i - \tilde{y}_i)^2$. Lorsque la variance conditionnelle de y est une fonction de x, on suggère d'utiliser la méthode des moindres carrés pondérés (Montgomery et Peck, 1982). Cette méthode (weigthed least-squares) consiste à minimiser la somme des carrés des résidus en introduisant un

poids w_i pour chaque x_i. On est alors conduit à minimiser:

$$Q = \sum_{i=1}^{N} w_i (y_i - \tilde{y}_i)^2 \qquad (4.5)$$

Le poids est inversement proportionnel à la variance conditionnelle de y_i. En supposant $\sigma^2_{y|x_i} = C'x_i^2$, on minimiserait la quantité:

$$Q = \sum_{i=1}^{N} \frac{1}{C'x_i^2} (y_i - \tilde{y}_i)^2 \qquad (4.6)$$

On obtient finalement les équations 4.7 et 4.8 qui permettent d'estimer les paramètres a et b.

$$\sum_{i=1}^{N} \frac{y_i}{x_i^2} = a \sum_{i=1}^{N} \frac{1}{x_i^2} + b \sum_{i=1}^{N} \frac{1}{x_i} \qquad (4.7)$$

$$\sum_{i=1}^{N} \frac{y_i}{x_i} = a \sum_{i=1}^{N} \frac{1}{x_i} + Nb \qquad (4.8)$$

Pour le cas particulier où la variance croît avec x, le procédé consiste en somme à accorder une moindre importance aux fortes valeurs de x et à mettre plus d'emphase sur les faibles valeurs de x pour lesquelles la variance est plus faible. Il est possible de démontrer (Neter et Wasserman, 1974) qu'on obtient les mêmes résultats en divisant la droite y = a + bx par x, soit:

$$\frac{y}{x} = a\frac{1}{x} + b \qquad (4.9)$$

$$y' = ax' + b \qquad (4.10)$$

En d'autres termes a et b peuvent être estimés en effectuant une régression avec les variables y' et x' (y' = y/x et x' = 1/x). Si l'on veut obtenir par la suite l'équation de la droite par rapport à y (et non y'), il suffit de multiplier l'équation par x.

$$y'x = (ax' + b)x \qquad (4.11)$$

$$y'x = ax'(x) + bx \qquad (4.12)$$

$$\tilde{y} = a + bx \qquad (4.13)$$

Cette transformation est particulière car elle met en relation des fonctions de variables possédant une variable commune, technique discutée dans le cadre des corrélations fictives. On doit donc souligner ici qu'il ne s'agit pas d'interpréter les variables transformées (ou encore d'utiliser le coefficient de corrélation des variables transformées) mais bien de stabiliser par ce moyen la variance pour être en mesure d'utiliser certains résultats de la régression. Nous proposons à cet effet quelques exemples qui font ressortir d'une part l'utilité du procédé voir figures 4.10 et 4.11 et l'utilisation abusive de ce même procédé (exemple 4.4).

Exemple numérique

On a un échantillon dont les variables x et y présentent une variance non constante. On suppose que les variables $1/x = x'$ et $y/x = y'$ stabilisent la variance. La régression de y' et x' aboutit au modèle suivant:

$\tilde{y}' = 3,80 + 0,5x'$ (modèle pour les variables transformées)

On aura simplement pour y et x:

$\tilde{y} = 0,5 + 3,80x$ (nouveau modèle pour les variables x et y)

4.2.2.4 Commentaires et exemples

- Comparaison entre le test F et le test graphique

Nous avons appliqué le test F à un exemple présenté dans Chatterjee et Price (1977). Par le test graphique (y_i, e_i) les auteurs concluent que la variance est croissante avec x (figure 4.9). Comme le nombre d'observations était impair (N = 9), le test fut appliqué en considérant des groupes non-égaux. Le rapport des variances $F_0 = S_{e_2}^2 / S_{e_1}^2$ est estimé et est comparé au rapport critique F_c pour un niveau de signification $\tilde{\alpha} = 0,025$ et $\tilde{\alpha} = 0,05$. On obtient:

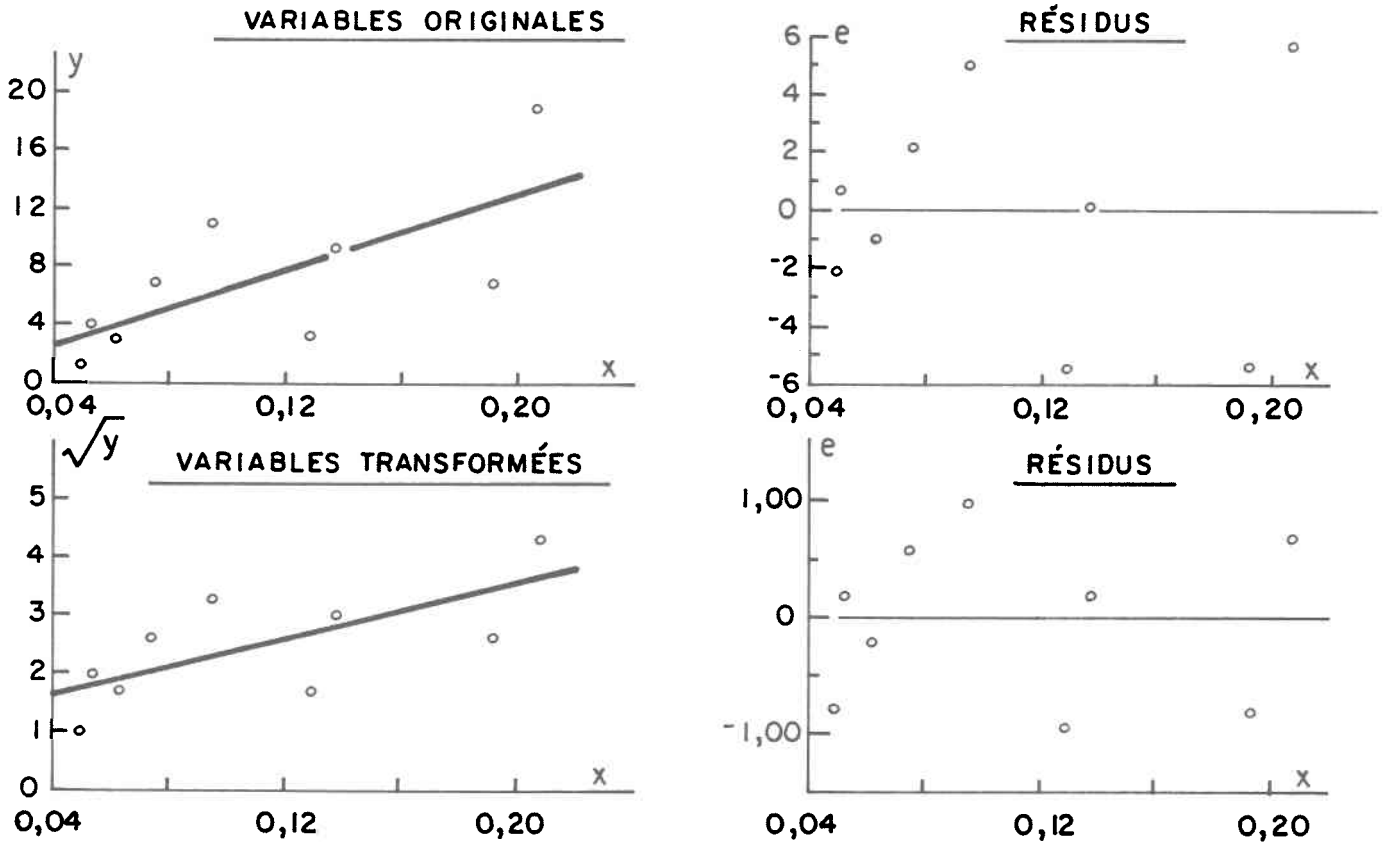


FIGURE 4.9 Vérification de l'hypothèse d'homoscédasticité par un test graphique. Comparaison entre les résidus des variables originales (A) et transformées (B). Données tirées de Chatterjee et Price (1977).

$F_1 = S_{e_2}^2 / S_{e_1}^2 = 3,52 < F_c = 16,0$ $\quad N_2 = 4 \quad N_1 = 5$ $\quad F_c = F_{2,3}(0,025)$

$F_2 = S_{e_2}^2 / S_{e_1}^2 = 7,72 < F_c = 39,2$ $\quad N_2 = 5 \quad N_1 = 4$ $\quad F_c = F_{3,2}(0,025)^{0,0}$

On obtient donc toujours $F_0 < F_c$ et l'on ne peut démontrer que les différences sont significatives, même en choisissant un niveau de signification plus élevé. En effet pour $\tilde{\alpha} = 0,05$ on aurait: $F_1 = 3,52 < F_c = 9,55$ et $F_2 = 7,72 < F_c = 19,2$

On n'aboutit donc pas forcément aux mêmes conclusions en utilisant le test F et le test graphique. Le test F a été construit pour vérifier si les variances provenant de deux populations étaient significativement différentes. Il est donc utilisé ici pour répondre à un autre objectif. Indépendemment du test F il est souhaitable de considérer d'abord les résultats du test graphique si ce dernier permet de conclure sans ambiguité à un problème d'hétéroscédasticité. Le test F pourra être considéré pour des cas plus litigeux et lorsque la taille de l'échantillon est assez élevée.

- Choix d'une transformation

Théoriquement, le choix d'une transformation devrait être fonction de ce qui cause l'hétéroscédasticité. En pratique on dispose rarement de cette connaissance. Même en sachant que Y ne suit pas une distribution normale

encore faudrait-il identifier à quelle distribution Y appartient. De plus il n'est pas facile de faire un choix à partir de la seule observation des couples
(x_i, y_i). Le graphique des résidus en fonction de x permet de mettre en relief un problème d'hétéroscédasticité, mais ne donne pas d'indication sur le moyen d'y remédier.

En pratique donc, on devra procéder par essai et erreur, c'est-à-dire essayer plusieurs transformations, construire le patron résiduel correspondant, et retenir celle qui permet de stabiliser la variance. De plus, dans certains cas, plusieurs transformations semblent stabiliser la variance et il est difficile de choisir objectivement la meilleure transformation. L'examen des résidus étant souvent beaucoup plus difficile à interpréter que lorsqu'il s'agit d'une linéarisation, on peut aussi envisager une méthode analytique pour comparer plusieurs transformations.

Box et Cox (1964) proposent une méthode générale qui permet de choisir, pour la classe des transformations puissances, la transformation la plus appropriée. Cette méthode consiste à déterminer pour la variable y^λ la valeur du paramètre λ qui minimise les résidus $SS_r(\lambda)$. Pour différentes valeurs de λ et pour que les quantités $SS_r(\lambda)$ soient comparables (échelles comparables), on considérera les variables $y^{(\lambda)}$ soit:

$$y^\lambda = \begin{cases} \dfrac{y^\lambda - 1}{\lambda y_G^{\lambda-1}} & \lambda \neq 0 \\ y_G \ln y & \lambda = 0 \end{cases} \qquad (4.14)$$

Y_G correspond à la moyenne géométrique des différents y_i (i=1 à N) soit:

$$y_G = ((y_1)\ (y_2)\ \ldots\ (y_N))^{1/N}$$

Cette méthode procède par itération et l'on choisi arbitrairement la première valeur de λ. En supposant $\lambda = 0,5$ on effectuera une régression (méthode des moindres carrés) en utilisant les variables x et $y^{(0,5)}$, soit, d'après (4.14), et pour tous les y_i:

$$y_i^{0,5} = \frac{y_i^{(0,5-1)}}{0,5[y_G\ ^{(0,5-1)}]} = \frac{y_i^{-0,5}}{0,5y_G^{-0,5}}$$

On calcule ensuite pour les variables $y_i^{(0,5)}$ et x la quantité $SS_r(\lambda)$. On compare ainsi cette quantité pour d'autres valeurs de λ

Exemple numérique

On suppose les résultats suivants pour différentes valeurs de λ

λ	y^{λ}	$SS_r(\lambda)$	Variables à utiliser
-1	y^{-1}	400	$1/y$
-0,5	$y^{-0,5}$	200	$1/\sqrt{y}$
0	y^{0}	75	ln y
0,5	$y^{0,5}$	2	$\sqrt{y}$
1,0	y^{1}	18	y
2,0	y^{2}	79	y^2

On obtient ici une quantité $SS_r(\lambda)$ minimum pour $\lambda = 0,5$. On en déduit que, pour l'échantillon dont on dispose, les variables qui stabilisent la variance sont x et $\sqrt{y}$. On trouvera dans Montgomery et Peck (1982) des exemples détaillés de cette méthode.

- Effet des transformations

Certaines transformations ont peu d'influence sur l'équation théorique du modèle. C'est le cas lorsqu'on utilise les variables y/x et 1/x. Cette transformation est intéressante si on suppose qu'en dépit d'une variance non-constante, un modèle linéaire est plausible. On aura donc un modèle linéaire comme pour une régression en utilisant y et x. Cependant, cette droite sera plus centrée aux valeurs de x possédant une variance minimum. L'effet de cette transformation est mis en évidence à la figure 4.10. Les données de cet exemple sont tirées de Chatterjee et Price (1977).

En supposant une variance minimum pour les faibles valeurs de x, la stabilisation réduira l'écart-type associé aux paramètres a et b, soit S_a et S_b. Pour les modèles 1 et 2 (figure 4.10), les valeurs minimum et maximum des estimateurs a et b sont présentées au tableau 4.4. L'intervalle de confiance à 90% autour de a est illustré à la figure 4.10.

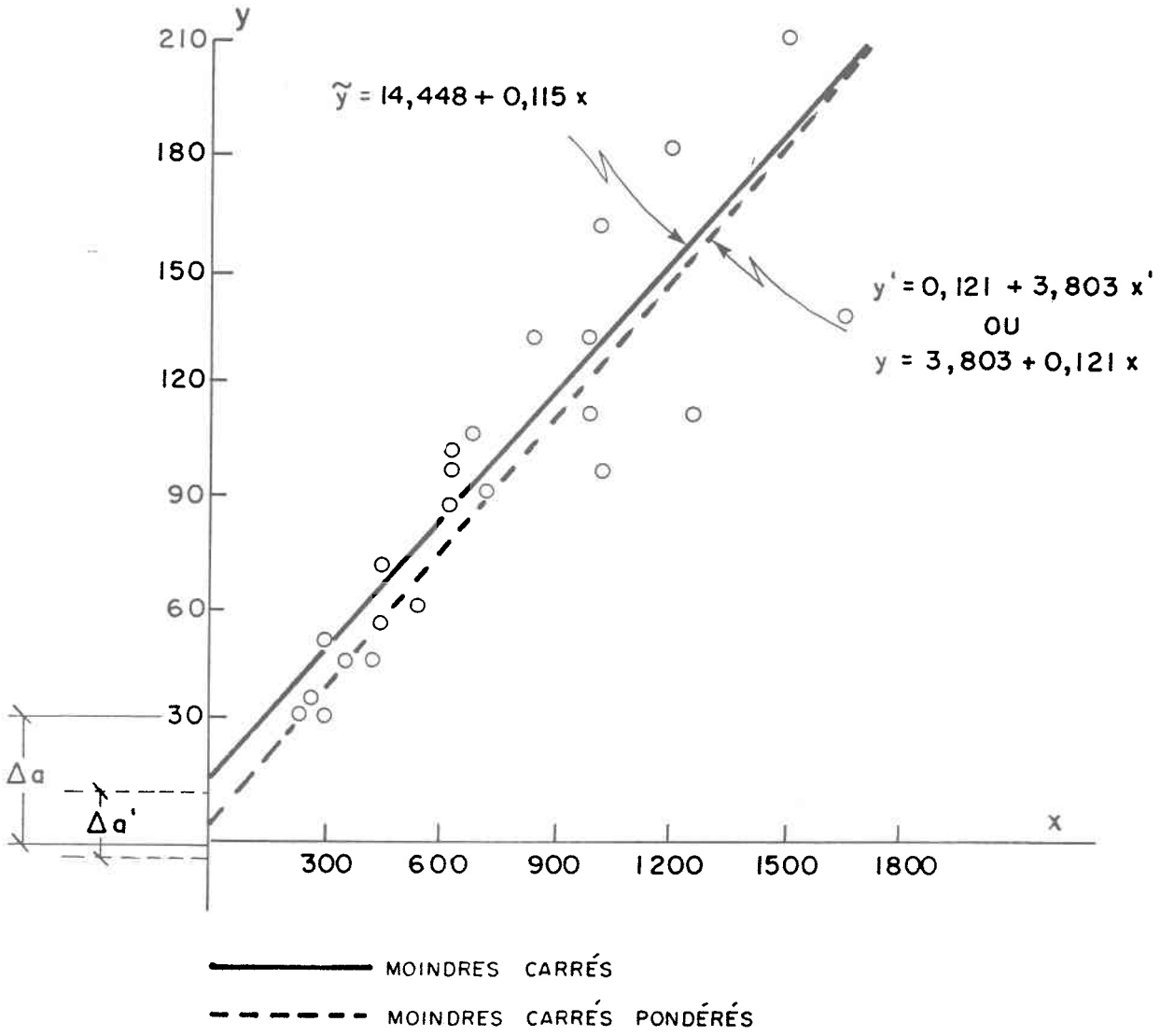


Figure 4.10 Stabilisation de la variance. Comparaison des intervalles Δa et $\Delta a'$ associés à la méthode des moindres carrés et des moindres carrés pondérés. Modifié d'après Chatterjee et Price (1977).

TABLEAU 4.4 - STABILISATION DE LA VARIANCE: ÉCART-TYPE DES ESTIMATEURS a ET b. (Variables originales et variables transformées)

Variables utilisées	a	(S_a)	a_1	a_2	b	(S_b)	b_1	b_2
(x,y)	14,4	(9,56)	-1,88	30,8	0,115	(0,011)	0,096	0,134
1/x, y/x)	3,80	(4,57)	-4,45	12,1	0,121	(0,009)	0,105	0,137

a_1, a_2, b_1, b_2 sont les valeurs limites des paramètres pour un intervalle à 90%.

La plupart des transformations qui stabilisent la variance ne sont pas des modèles linéaires (tableau 4.3). On vérifiera donc s'il est pertinent d'utiliser un modèle non-linéaire. Par exemple, en utilisant les variables x et $\sqrt{y}$, le modèle est de la forme suivante:

$$\sqrt{y} = a + bx \qquad (4.15)$$

ce qui se traduit en fonction de y par

$$y = a^2 + 2ab\,x + b^2\,x$$

ou encore

$$y = A + Bx + Cx^2 \qquad (4.16)$$

Un tel modèle est présenté à la figure 4.11. L'étude des résidus a permis de vérifier que l'utilisation des variables $\sqrt{y}$ et x stabilisent la variance. Les observations sont tirées de Montgomery et Peck (1982), de même que la transformation suggérée $\sqrt{y}$ et x. Nous avons construit les modèles (courbe et droite) correspondant aux transformations ($\sqrt{y}$ et x) et (y/x et 1/x). On peut vérifier que les variables y/x et 1/x stabilisent aussi la variance. On devra dès lors choisir entre un modèle linéaire et un modèle non-linéaire qui aboutissent à des valeurs prédictives très différentes pour $x \geq 2400$. Lorsque les transformations à comparer affectent et la variable dépendante et la variable indépendante, il n'est plus possible d'utiliser un modèle analytique pour comparer plusieurs transformations. On peut alors considérer le modèle qui permet d'obtenir une variance minimum pour les paramètres. L'écart-type S_a et les valeurs minimum et maximum (intervalle à 90%) permettent de vérifier que le modèle non linéaire ($\sqrt{y}$ et x) est plus efficace. Les résultats pour les variables x et y ainsi que pour les deux transformations sont présentés au tableau 4.5.

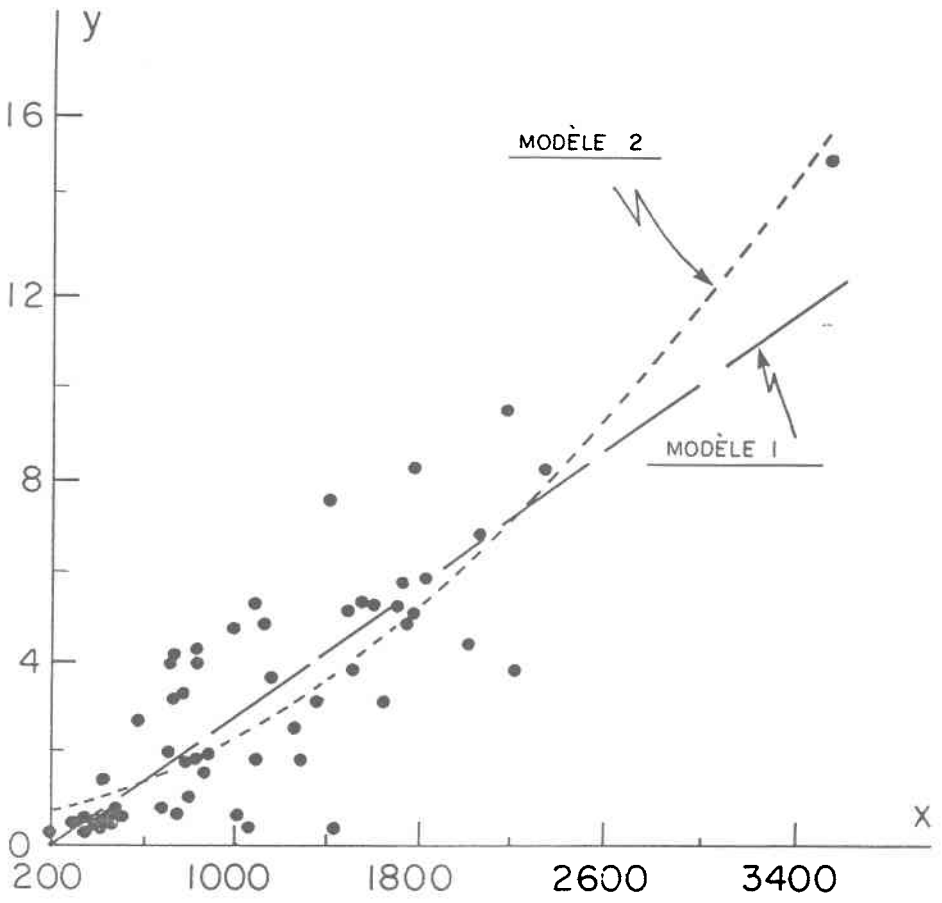


Figure 4.11 Stabilisation de la variance; comparaison entre une transformation linéaire et non-linéaire. Données tirées de Montgomery et Peck (1982).

TABLEAU 4.5- ÉCART-TYPE ET VALEURS LIMITES DE a POUR DIFFÉRENTS MODÈLES LINÉAIRES ET NON LINÉAIRES

Variables utilisées	a	(S_a)	a_1	a_2
x et y	-0,8313	(0,7405)	-1,157	-0,0908
y/x et 1/x	-0,7324	(0,2677)	-1,181	-0,2835
$\sqrt{y}$ et x	0,3390		0,1328	0,6400

L'intervalle autour de a doit d'abord être calculé pour les variables transformées. On estime ensuite les résultats pour le graphique des variables x et y. Par exemple, pour les variables $\sqrt{y}$ et x, on obtient les résultats suivants:

$$y' = 0,5822 + 0,00095x \qquad \text{avec } y' = \sqrt{y}$$

$$= a \pm t_{N-2}(\tilde{\alpha}/2)S_a$$

$$= 0,5822 \pm 1,677(0,1299)$$

Pour l'ordonnée à l'origine a = 0,5822 on obtient les valeurs limites a_1 = 0,3644 et a_2 = 0,8000. D'après 4.16, on peut en déduire l'ordonnée à l'origine (A) et les valeurs limites en terme des variables x et y. On a: $A = a^2 = 0,3390$, $A_1 = a_1^2 = 0,1328$ et $A_2 = a_2^2 = 0,6400$.

La différence la plus "visible" suite à une transformation apparaît surtout lorsque l'on construit l'intervalle de confiance associé au nouveau modèle. À la figure 4.12, on peut observer un échantillon dont les résidus sont croissants avec x. La figure 4.12 permet de comparer la droite ajustée par la méthode des moindres carrés et la droite des moindres carrés pondérés Ces droites diffèrent assez peu. On remarque cependant que l'intervalle associé au modèle transformé (figure 4.12c) reflète la dispersion des points (croissante avec x), ce qui n'est pas le cas de l'intervalle déduit du modèle $y = a + bx$ (figure 4.12b).

TABLEAU 4.6 - ÉCART-TYPE DES ESTIMATEURS a ET b (Modèles linéaires)

Variables utilisées	a	(S_a)	a_1	a_2	b	(S_b)	b_1	b_2
x,y	4,2	(3,3)	1,2	7,3	4,5	(0,53)	3,6	5,5
1/x,y/x	5,7	(0,40)	4,9	6,4	4,2	(0,97)	2,4	5,9

a_1, a_2 b_1, b_2 sont les valeurs limites estimées pour un intervalle à 90% ($\tilde{\alpha} = 0,05$)

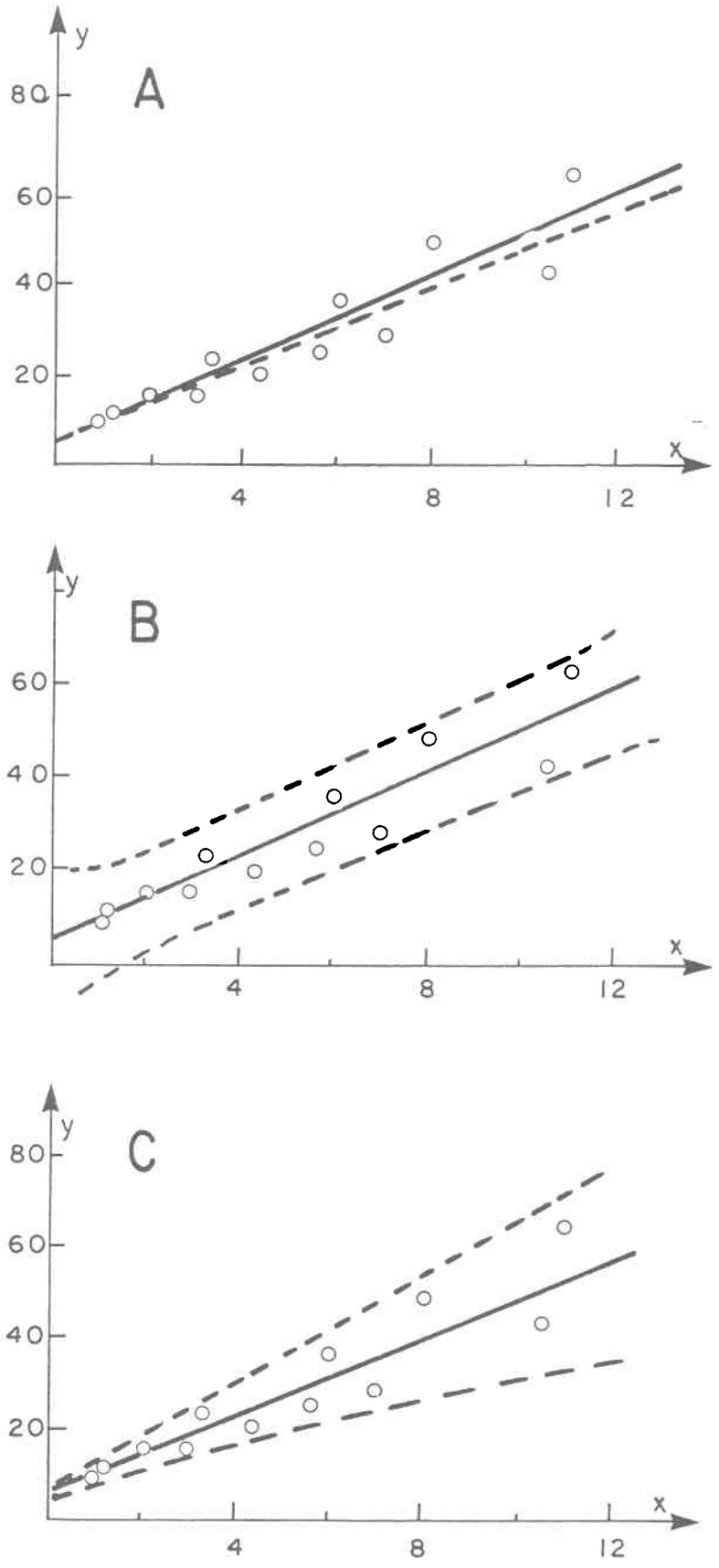


Figure 4.12 Comparaison des droites et des intervalles de prévision déduits de la méthodes des moindres carrés et des moindres carrés pondérés. Données tirées de Neter et Wasserman (1974).

EXEMPLE 4.4

Les variables de la figure 4.13A sont tirés de Younger (1979). L'équation $\tilde{y} = 2,3873 + 0,3066x$ représente la droite ajustée à l'échantillon par la méthode des moindres carrés. Le test de la pente ou encore du coefficient de corrélation ($r = 0,3201$) permet de conclure à une relation non-significative entre les variables x et y. Cependant, le test graphique (x_i, e_i) et le test F mettent en évidence un problème d'hétéroscédasticité. On obtient en effet pour les sous-groupes N_1 et N_2 un rapport $F_0 > F_c$ (valeur critique). On a:

$$F_0 = 9,42 \quad \text{et} \quad F_c = F_{5,5}(0,05) = 5,5$$

Étant donné la variance non constante, les tests de la pente (ou encore du coefficient de corrélation) ne sont pas rigoureusement applicables. Il est donc souhaitable de transformer les variables pour stabiliser la variance. Les variables $y' = y/x$ et $x' = 1/x$ suggérées pour Younger (1979) stabilisent la variance, comme on peut le vérifier à la figure 4.13B. L'équation $\tilde{y}' = 0,3000 + 2,4026x'$ est déduite des variables y' et x'. On obtient donc en fonction de x et y (voir EQ 4.10 à 4.13) $\tilde{y} = 2,4026 + 0,3000x$.

L'auteur qui étudie l'échantillon en conclut que les variables sont stabilisées et que la nouvelle équation donne des valeurs de y légèrement différentes de l'équation originale. Cependant, même si pour la droite $\tilde{y}'$,

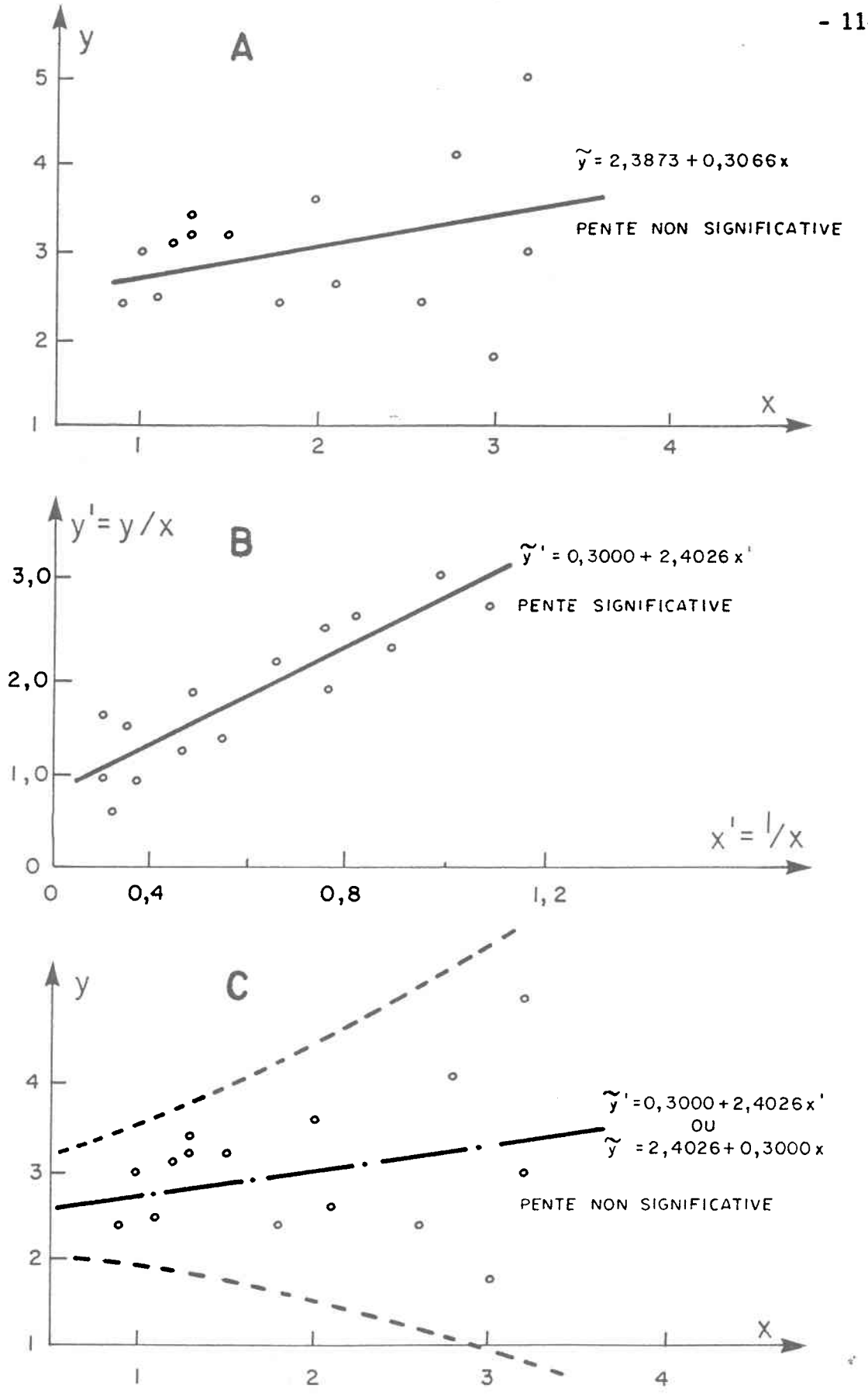


Figure 4.13 Comparaison des pentes suite à la stabilisation de la variance. Données tirées de Younger (1979) Exemple 4.4.

la pente (b' = 2,4026) et le coefficient de corrélation (r' = 0,89) sont significatifs, on ne peut rien en conclure quant aux variables originales.

En effet, on ne peut conclure, comme le suggère l'auteur, que la relation entre x et y est significative, du seul fait que b' (ou r') sont significatifs. Ce qui est vrai pour les variables y' et x' ne l'est pas nécessairement pour les variables originales (x et y). Comme on l'a mentionné précédemment, la stabilisation de la variance permet d'obtenir pour les paramètres une variance minimum. Puisque a' correspond à la pente sur le graphique de y et x, on doit comparer S_b (écart-type de b) et S'_a (écart-type de a'). Les résultats sont présentés au tableau 4.7

TABLEAU 4.7 - STABILISATION DE LA VARIANCE, INTERVALLE AUTOUR DE LA PENTE MODÈLE ORIGINAL ET MODÈLE TRANSFORMÉ

Variables utilisées	Pente	Écart-type	Valeurs limites de l'intervalle de la pente	Intervalle $(1-\tilde{\alpha})x100$
x et y	b = 0,3066	0,2516	-0,1392 0,7522	(90%)
			-0,2369 0,8500	(95%)
x'=1/x y'=y/x	a'= 0,3000	0,2267	-0,1015 0,7015	(90%)
			-0,1896 0,7896	(95%)

On vérifie bien que $S'_a < S_b$. Cependant, que l'on considère les variables x, y ou x' et y', les limites de l'intervalle à 90% et à 95% contiennent la valeur zéro, ce qui permet de vérifier que la pente n'est pas significativement différente de zéro. On arriverait aux mêmes conclusions en effectuant un test de signification de la pente, c'est-à-dire un test sur a'. On calcule t_0 que l'on compare à la valeur critique t_c.

$$t_0 = \frac{a'}{S'_e \quad (1/N + (\bar{x}'^2 / SS_{x'})^{\frac{1}{2}}} = 1,32$$

Pour $\tilde{\alpha} = 0,025$ la valeur critique $t_c = t_{13}\ (0,025) = 2,160$. Puisque $t_0 < t_c$, H_0 est accepté (voir section 2.3.5.2).

Nous en concluons donc qu'étant donné la stabilisation de la variance le test de la pente est valide, et la relation est non-significative puisque la pente n'est pas significativement différente de 0.

EXEMPLE 4.5

Les données de l'exemple suivant sont tirées d'une publication de Jensen (1979). Nous avons étudié un des graphiques présenté dans l'article de cet auteur. Les variables, l'équation de la courbe et l'intervalle de confiance sont présentés à la figure 4.14A. L'équation de la courbe a été déduite en utilisant les variables x et $y' = \ln(y-c)$ avec $c = 21,7$. Par la méthode des moindres carrés on obtient l'équation

$$\tilde{y}' = -7,684 + 37,85\ x.$$

ce qui se traduit en fonction de y et x par:

$$\tilde{y} = 21,7 + (0,00046\ e^{37,85x})$$

L'étude de ces résidus e'_i en fonction de x_i nous a permis de

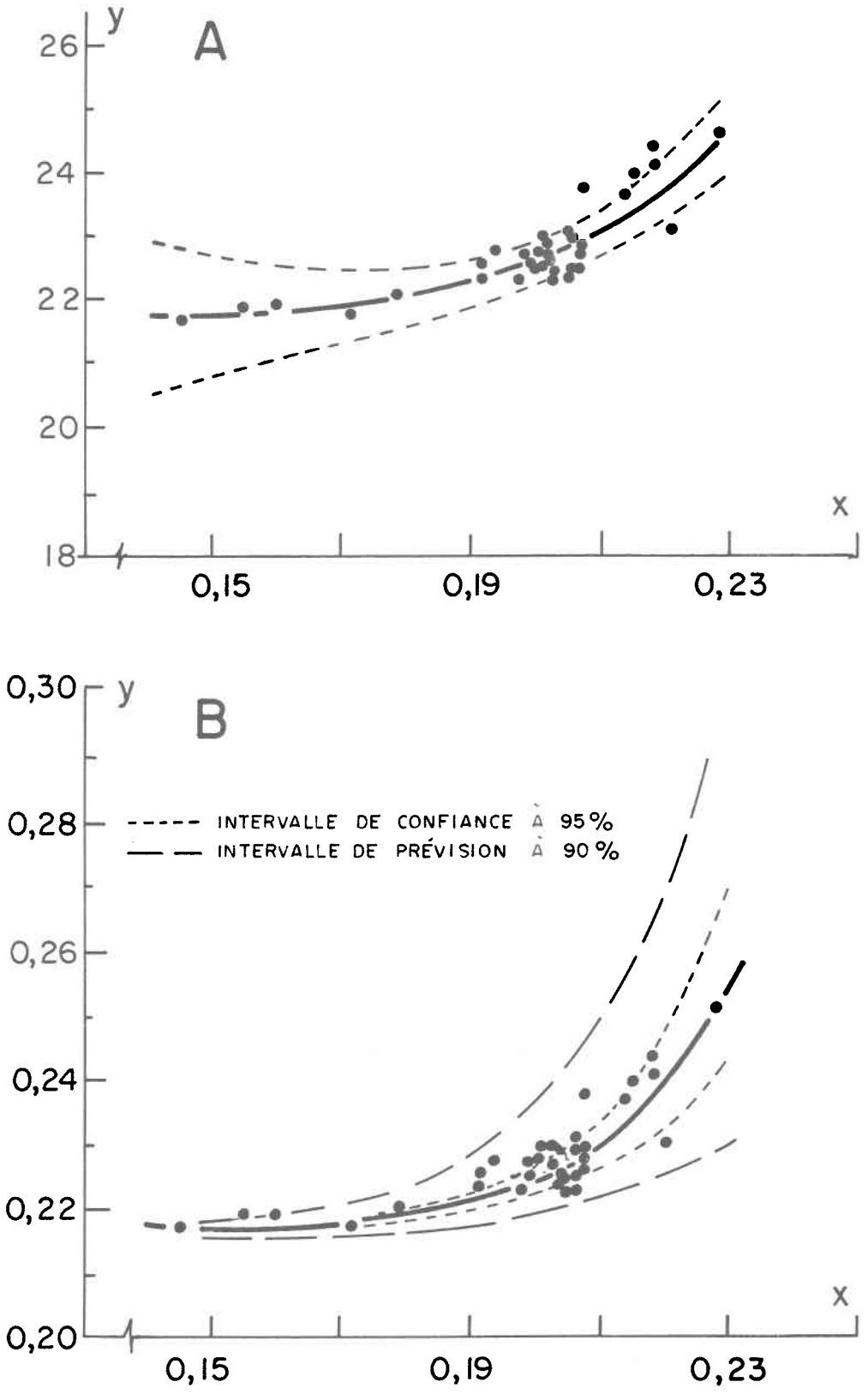


Figure 4.14 Linéarisation et intervalle de confiance associé au nouveau modèle. A: D'après Jensen (1979). B: Intervalle de confiance du modèle non-linéaire (Exemple 4.5).

vérifier la linéarité des variables y' et x. Le même graphique (x_i, e'_i) peut être utilisé pour vérifier l'hypothèse d'homocédasticité. Comme la variance ne semblait pas parfaitement constante, nous avons appliqué le test F d'égalité des variances, test moins subjectif que le test graphique. Les résultats suivants furent obtenus pour les niveaux de signification $\tilde{\alpha} = 0,025$ et $\tilde{\alpha} = 0,05$ et N = 34

$$F_0 = S_2^2 / S_1^2 = 2,69 \qquad \text{avec } N_1 = N_2 = 17$$

$$F_c = F_{15,15}(0,025) = 2,86$$

$$F_c = F_{15,15}(0,05) = 2,40$$

Au niveau de signification $\tilde{\alpha} = 0,025$ $F_0 < F_c$, on pourrait donc accepter l'hypothèse d'égalité des variances. Ce faisant, on peut utiliser les résultats de la régression linéaire pour construire l'intervalle autour de y' (intervalle de confiance) ou encore l'intervalle de prévision. Cet intervalle doit évidemment être construit à partir des variables y' et x. En effet, les hypothèses de linéarité et d'homoscédasticité sont respectées en fonction de ces variables. Les résultats peuvent ensuite être transposés sur le graphique de y et x. Les limites de l'intervalle de prévision $(1-\tilde{\alpha}/2) \times 100$ au point $x = x_0$ sont η'_1 et η'_2 estimés par:

$$\begin{matrix} y'_1 \\ y'_2 \end{matrix} = \tilde{y}' \pm t_{(N-2)}^{(\tilde{\alpha}/2)} \; S'_e \left(1 + \frac{1}{N} + \frac{(x_0 - \bar{x})^2}{SS_x}\right)^{\frac{1}{2}}$$

Par exemple, on obtient pour x = 0,19 les résultats suivants:

x	$\tilde{y}'$	$\tilde{y}'_1$	$\tilde{y}'_2$
0,19	-0,6779	- 1,6793	0,3234

Les limites y_1 et y_2 sur le graphique de y et x sont données par

$$y_1 = e^{-1,6793} + 21,7 = 21,89$$
$$y_2 = e^{0,3234} + 21,7 = 32,08$$

L'intervalle de prévision à 95% a été construit point par point et les résultats transposés sur le graphique y et x (figure 4.14 B). Comme on peut le voir, cet intervalle (figure 4.14 A) est très différent de l'intervalle présenté par l'auteur. Il est aussi intéressant de noter que cet intervalle (figure 4.14 A) est très large là où les observations sont peu dispersées ($x < 0,18$) et qu'il est minimum là où les observations présentent une dispersion importante $x \approx 20$. On notera aussi que le coefficient de régression r = 0,90 est déduit des variables y' et x et que cela n'implique pas nécessairement une corrélation importante pour les variables y et x comme cela est suggéré. On se rappellera l'exemple précédent où l'on obtenait un fort coefficient pour les variables transformées et néanmoins une relation non significative en terme des variables y et x.

4.2.3 Hypothèse de normalité

Si l'on désire connaître la distribution de a, b et de $\tilde{y}$, on doit aussi s'assurer que les résidus sont distribués normalement. Cette condition est en effet nécessaire pour une utilisation rigoureuse des tests et pour la construction des intervalles de confiance.

Comme on l'a déjà signalé cependant, cette hypothèse n'est pas très restrictive en pratique. En effet, les résultats concernant les tests et les intervalles de confiance peuvent être utilisés si l'écart à la normalité n'est pas trop grand et si la taille de l'échantillon n'est pas trop petit (Bobée, 1983; Neter et Wasserman, 1974). L'écart à la normalité peut avoir plusieurs causes:

- modèle non adéquat pour représenter l'échantillon;
- variance non constante;
- présence de valeurs extrêmes.

Il est donc souhaitable de vérifier d'abord ces conditions avant d'effectuer un test de normalité. Très souvent d'ailleurs, une stabilisation de la variance aura aussi pour effet de normaliser la distribution des résidus. À l'inverse, cependant, on pourra aussi détruire l'hypothèse de normalité en effectuant une transformation quelconque et c'est pourquoi il est utile de vérifier cette hypothèse pour les variables transformées. À cet effet un exemple est présenté.

4.2.3.1 Vérification de la normalité des résidus

Plusieurs méthodes et tests graphiques peuvent être utilisés pour vérifier l'hypothèse de normalité de la distribution conditionnelle de y/x. Nous en résumons ici quelques unes.

- Méthodes graphiques

Les méthodes graphiques sont assez simples à appliquer, mais de façon générale, elles ne présentent un intérêt que si N est élevé.

(a) On peut construire l'histogramme de la distribution des $e_i = (y_i - \tilde{y}_i)$ et vérifier la forme et la symétrie de cette histogramme. Cette méthode d'appréciation est très subjective et ne présente un intérêt réel que si N est très élevé. Elle peut cependant être utilisée conjointement à une autre méthode.

(b) On peut aussi calculer la proportion des résidus e_i comprise dans le domaine $(-KS_e + KS_e)$ ou S_e est estimé par l'équation 2.27. Pour la loi normale, la proportion des valeurs qui devraient être comprises à l'intérieur de ce domaine est connue et peut être déduite à partir de la table de la loi normale. Quelques valeurs limites généralement utilisées sont présentées au tableau 4.8. Par exemple, on vérifiera qu'environ 38% des

résidus sont compris à l'intérieur ± 0,5 S_e et 68% à l'intérieur de ± 1S_e.

TABLEAU 4.8 - PROBABILITÉS ASSOCIÉES À LA LOI NORMALE

K	$Pr(-KS_e < e < KS_e)$
0.5	38.29%
1.0	68.27%
1.5	86.64%
2.0	95.45%
2.5	98.76%
3.0	99.73%

On peut aussi utiliser la variables standardisées u_i, soit:

$$u_i = \frac{e_i - \bar{e}_i}{S_e} = \frac{e_i}{S_e} \qquad (4.17)$$

et vérifier la proportion comprise à l'intérieur du domaine (-K; +K).

(c) La normalité peut aussi être vérifiée sur un papier de probabilité normale. On présente généralement ce test (Baillargeon Rainville, 1977; Neter et Wasserman, 1983; Scherrer, 1984) pour des données groupées en

en classe ce qui suggère un échantillon très élevé. Nous considérons ici le test pour des observations non groupées. Les observations standardisées sont d'abord classées par ordre croissant:

$$u_1 \leq u_{i+1} \leq u_k \ldots\ldots \leq u_N \qquad \text{avec } i = 1 \text{ à } N$$

On associe à chaque résidu standardisé une probabilité empirique F_K. Il existe plusieurs formules pour calculer F_K. On trouvera dans Bobée (1983) quelques-unes des formules fréquemment utilisées. Nous considérons les formules de Weibull et de Hazen, suggérée par Daniel et Wood (1971) et par Gunst et Mason (1980).

$$F_K = \frac{K}{N+1} \quad \text{(formule de Weibull)} \qquad (4.18)$$

$$F_K = \frac{K - 0,5}{N} \quad \text{(formule de Hazen)} \qquad (4.19)$$

avec K = rang des observations classées

N = nombre d'observations

Les variables standardisées u_i et leur probabilité associée F_K (fréquences cumulées) sont ensuite reportées sur le papier de probabilité normale. Sur ce papier les points correspondant à la loi normale sont situés sur une droite. Donc si la distribution des u_i est normale les observations devraient être approximativement sur une droite.

La fiabilité de cette méthode repose d'abord sur la taille de

l'échantillon. On trouvera dans Daniel & Wood (1971) plusieurs exemples pour des échantillons variant de 8 à 384 observations. On ne peut s'attendre au tracé d'une droite parfaite. Il faut donc interpréter le graphique; en ce sens on peut utiliser les graphiques illustrés dans Daniel & Wood (1971) comme guide d'interprétaton. On peut voir que même si u_i provient d'une distribution normale, on obtient rarement une droite lorsque le nombre d'observations est faible. Cette méthode graphique devrait être utilisée (Montgomery et Peck, 1982) lorsqu'on dispose d'un échantillon d'au moins 20 observations. Neter et Wasserman (1983) proposent aussi de calculer le coefficient de corrélation entre u_i et F_k. Un coefficient supérieur à 0,90 serait l'indice d'une normalité acceptable.

- Tests

Lorsqu'il est difficile de conclure à partir d'une méthode graphique, certains tests permettent de vérifier la normalité des résidus.

Les tests du CHI-2 et de Kolmogorov-Smirnov suggérés par Neter et Wasserman (1974) ne sont pas très puissants pour étudier la validité de la loi normale ajustée à un échantillon (Bobée, 1983). Dans le contexte de la régression, on devrait considérer surtout le test basé sur l'asymétrie et l'aplatissement (Scherrer, 1984) ou le test W (Shapiro et Wilk, 1965).

Le test basé sur l'asymétrie et l'aplatissement est fiable si $N > 15$. Le test W peut être appliqué même si la taille de l'échantillon est faible. Il présente donc un avantage sur le test précédent et sur les méthodes graphiques déjà présentées. Il sera sommairement présenté dans ce qui suit.

- Description du test W

Les résidus e_i sont classés par ordre croissant.

$$e_i < e_{i+1} < \dots e_N$$

On calcule ensuite la quantité SS_r et on détermine la quantité B soit:

$$B = \sum_{j=1}^{K} a_j \, (e_{N-j+1} - e_j) \qquad (4.20)$$

avec $K = \frac{N}{2}$ si N est pair

$K = \frac{N-1}{2}$ si N est impair

a_j est un coefficient donné par des tables (test W de normalité) et il est fonction du nombre d'observations N. On calcule par la suite W_0

$$W_0 = \frac{B^2}{SS_r} \qquad (4.21)$$

On compare alors W_0 à la valeur critique W_c également donnée par des tables. W_c est fonction de N et de $\tilde{\alpha}$. Si $W_0 < W(\tilde{\alpha})$ on rejette l'hypothèse de normalité.

Exemple numérique

Soit un échantillon de taille N = 6 et les résidus e_i

	e_i
1	-2,7
2	-2,8
3	-3
4	1
5	3,5
6	4,0

on a $K = \frac{N}{2} = 3$ et d'après 4.20 et 4.21 calcule B et W_0

$$B = \sum_{i=1}^{3} a_1(4 + 2,7) + a_2(3,5 + 2,8) + a_3(1 + 3,0) = 6,37$$

avec $a_1 = 0,6341, \quad a_2 = 0,2806, \quad a_3 = 0,0875$

$$W_0 = \frac{B^2}{SS_r} = \frac{(6,37)^2}{53,38} = 0,7593$$

Pour $\tilde{\alpha} = 0,05$ on a $W(0,05) = 0,788$. Comme $W_0 < W(0,05)$ on ne peut accepter l'hypothèse de normalité.

4.2.3.2 Normalisation des résidus

Les transformations les plus utiles pour normaliser les résidus sont les transformations logarithmiques et la transformation racine carré. Comme on l'a déjà mentionné, les transformations qui permettent de stabiliser les variances peuvent aussi dans certains cas normaliser les résidus. Comme ce n'est pas toujours le cas, on doit donc s'assurer que ces transformations ne détruisent pas l'hypothèse de normalité.

Deux exemples sont examinés dans ce qui suit. Pour le premier, on examine la normalité des résidus pour deux échantillons présentés dans une publication (Sharpley et al., 1985). Pour le second, on examine les résidus e_i et e'_i (résidus des variables transformées). Par ces exemples, on pourra aussi comparer les résultats d'un test graphique et du test W.

EXEMPLE 4.6

Les variables de la figure 4.15 (encadré) et 4.16 (encadré) représentent les valeurs de phosphore et d'azote mesurées (x) et prédites (y) par un modèle. Comme l'échantillon est grand N = 31, la vérification de la normalité peut être faite à l'aide d'un test graphique (papier de probabilité

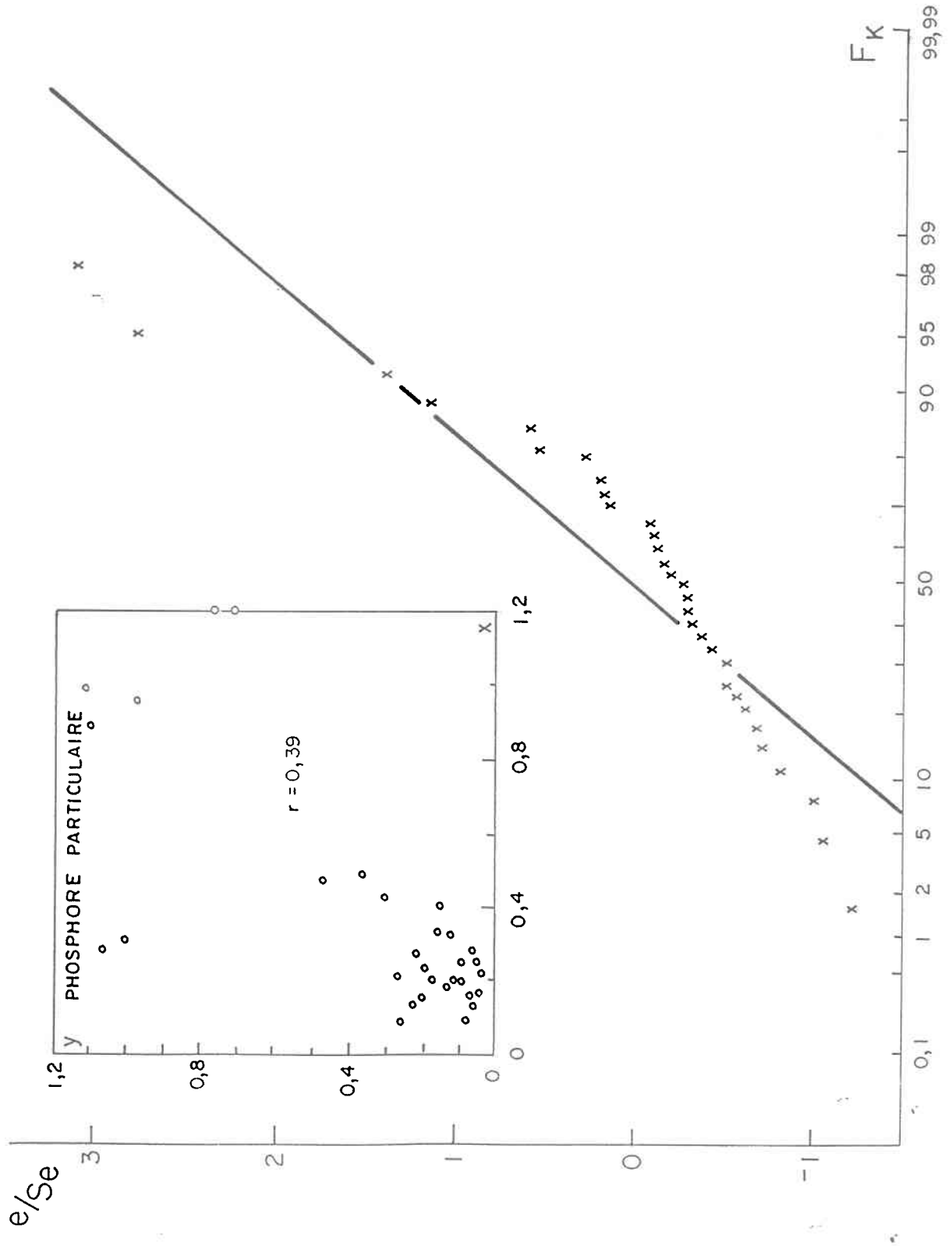


Figure 4.15 Etude de la normalité des résidus. (Exemple 4.6 Phosphore).

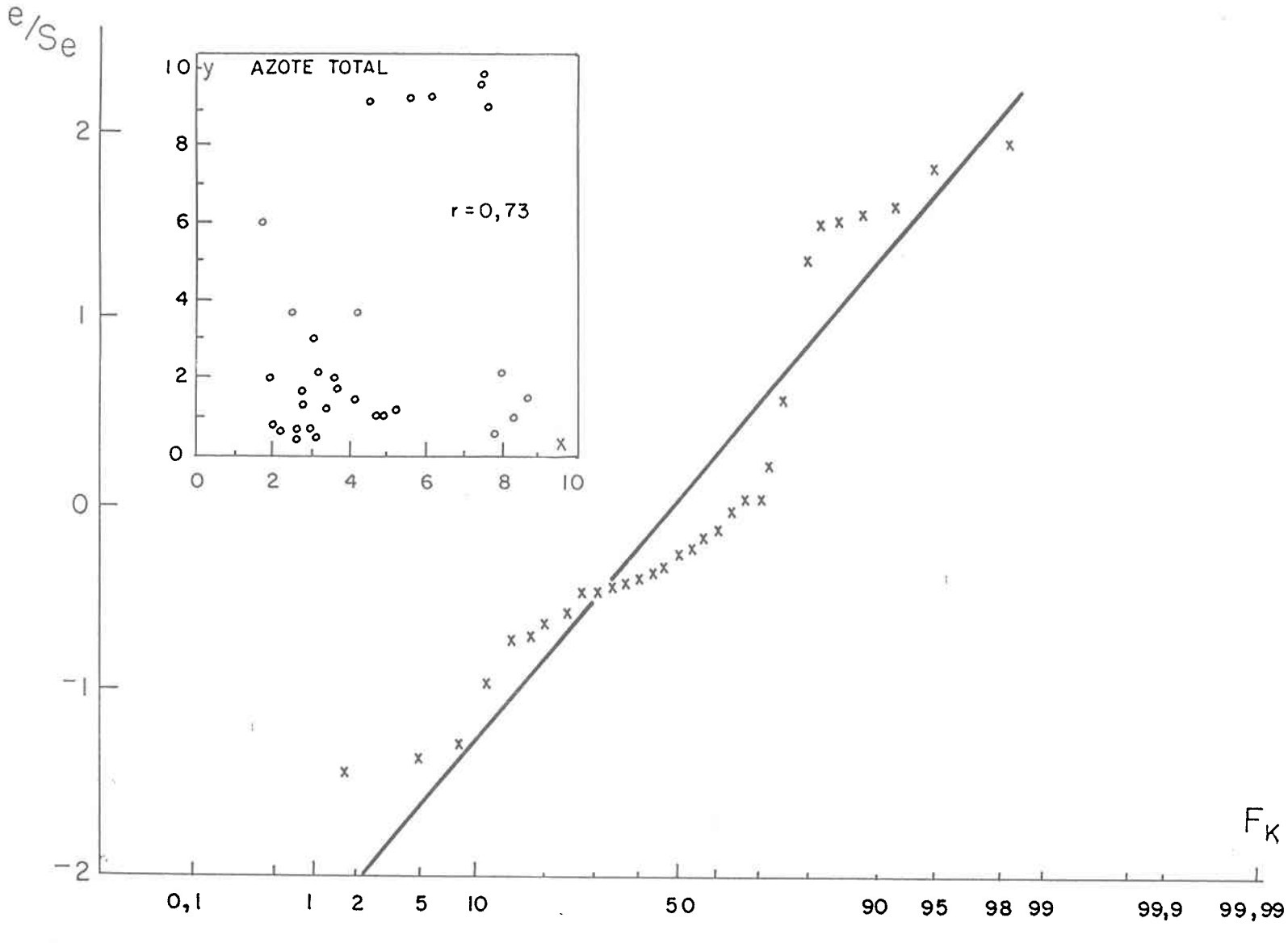


Figure 4.16 Etude de la normalité des résidus (Exemple 4.6 Azote).

normale). Pour ce test la formule de Hazen fut utilisée (EQ. 4.19)

L'examen des figures 4.15 et 4.16 permettent de conclure au non respect de l'hypothèse de normalité. Bien que les résultats du test graphique soient fiables (étant donné la taille de l'échantillon) le test W fut aussi appliqué. On obtient pour le phosphore: $W_0 = 0,8165$ et $W(0,05) = 0,929$. On a donc $W_0 < W(0,05)$, et l'hypothèse de normalité doit être rejetée. On arrive aux mêmes conclusions pour un niveau $\tilde{\alpha} = 0,02$ et $\tilde{\alpha} = 0,01$. En effet $W(0,02) = 0,914$ et $W(0,01) = 0,902$

Le non-respect de l'hypothèse de normalité ne fait donc aucun doute. Pour l'azote, on doit aussi rejeter l'hypothèse de normalité puisque $W_0 = 0,8864$. On ne peut donc souscrire aux conclusions suggérées dans l'article d'une relation significative entre les variables. En effet, les tests de signification ne sont guère applicables, étant donné le non-respect évident de l'hypothèse de normalité.

Cet exemple sera aussi étudié au chapitre 5. On pourra vérifier également par un test que les conditions d'utilisation et d'interprétation du coefficient de corrélation ne sont pas respectées.

EXEMPLE 4.7

Comme on l'a dit, certaines transformations qui stabilisent la variance peuvent cependant conduire au non-respect de l'hypothèse de normalité pour les variables transformées. L'exemple présenté à la figure 4.12 est repris ici.

Les résidus standardisés e_i/S_e des variables originales ont d'abord été étudiés par un test graphique (figure 4.17A). Comme on peut le voir, les résidus sont linéaires et l'hypothèse de normalité est respectée, et ce, même si pour ce modèle, la variance est non constante. Les résultats du test W confirment le test graphique. On obtient, $W_0 = 0,9688$. Pour $N = 12$ et $\tilde{\alpha} = 0,05$ et $0,01$, les valeurs de W de la table sont $W(0,05) = 0,803$ et $W(0,01) = 0,859$. Par suite $W_0 > W(\alpha)$ et l'hypothèse de normalité est confirmée.

Dans un pareil cas, si les résidus des variables originales suivent une distribution normale, une transformation pour stabiliser la variance peut alors détruire la normalité. Le même test graphique a été utilisé pour les variables $x' = 1/x$ et $y' = y/x$ choisies pour stabiliser la variance. Les résidus standardisés e'/S'_e (figure 4.17B) ne présentent plus une linéarité aussi bonne que les variables originales. Étant donné la taille de l'échantillon ($N = 12$), on peut difficilement conclure à partir de ce test. Le test W a donc été appliqué. On obtient, $W_0 = 0,859$. D'autre part comme précédemment $W(0,05) = 0,803$ et $W(0,01) = 0,859$.

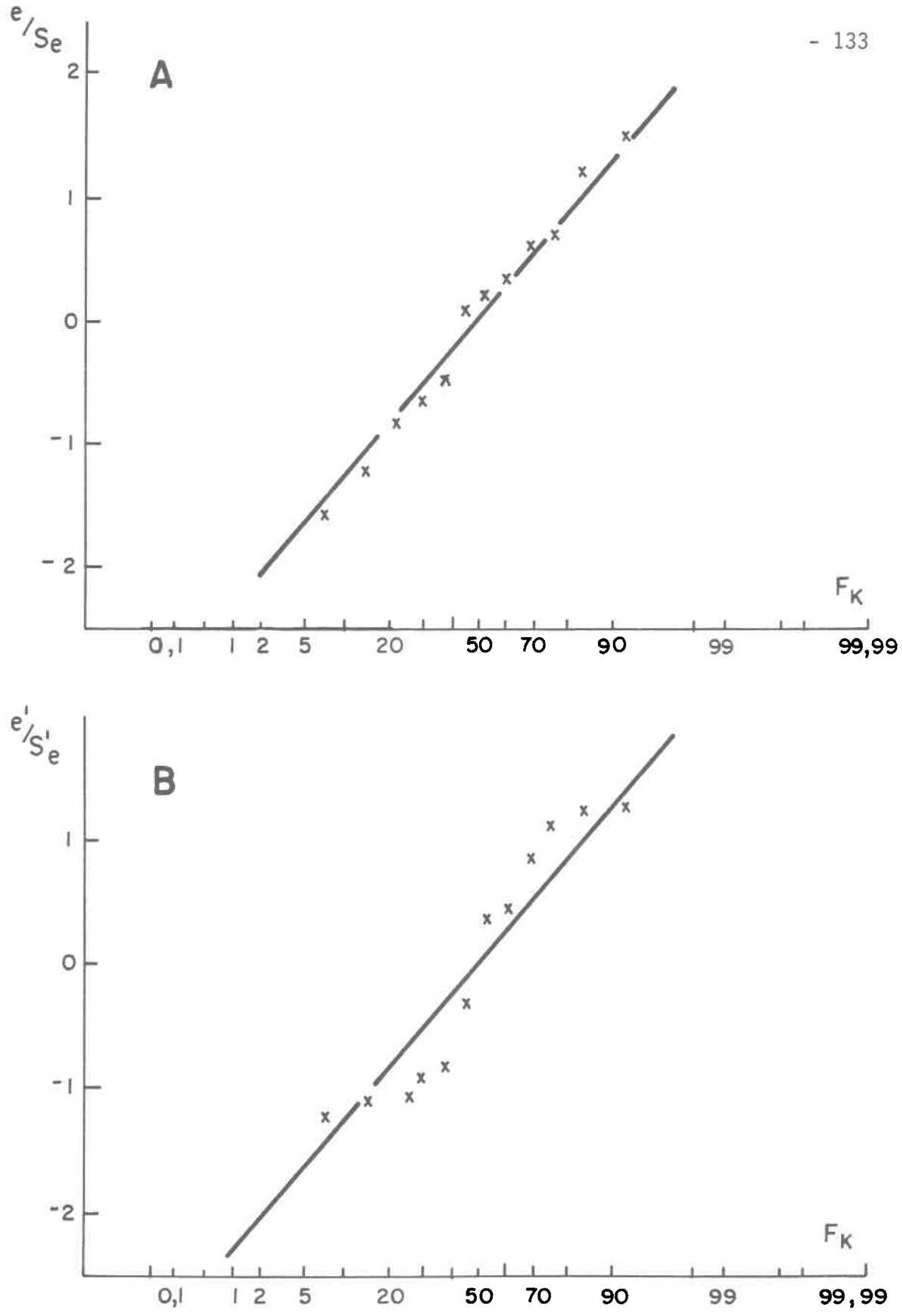


Figure 4.17 Examen de la normalité des résidus.
A: Variables originales. B: Variables transformées pour stabiliser la variance (exemple 4.7).

On peut donc conclure que les variables transformées stabilisent la variance et que ces mêmes variables permettent aussi de respecter l'hypothèse de normalité, bien que l'on soit à la limite pour $\tilde{\alpha} = 0,01$.

4.2.4 Hypothèse de l'indépendance des erreurs

Si dans une certaine mesure, on peut s'éloigner un peu de l'hypothèse de normalité des résidus, et de celle d'homocédasticité, il n'en est pas de même en ce qui concerne l'indépendance des erreurs. Cette hypothèse est fondamentale et ne pas la respecter peut conduire dans certains cas à un modèle tout à fait erroné. On doit, pour cette hypothèse, vérifier que l'erreur e_i pour $x = x_i$ est indépendante de e_{i-1} pour $x = x_{i-1}$.

Il convient de distinguer entre deux types d'autocorrélation très différents (Chatterjee et Price, 1977). On différencie en effet l'autocorrélation fictive (apparent autocorrélation) qui est due à l'omission d'une ou de variables importantes dans le modèle et l'autocorrélation véritable (pure autocorrélation) qui est due à un problème de persistance. Il n'est pas superflu de faire cette distinction. En effet, les causes et les solutions qu'on doit envisager sont très différentes, selon que l'on se réfère à un type ou l'autre d'autocorrélation.

4.2.4.1 Causes de l'autocorrélation

On observe souvent une dépendance entre les erreurs lorsque les observations peuvent être ordonnées selon une séquence particulière. De nombreux cas de non indépendance des résidus sont relevés lorsque les observations sont obtenues selon une séquence temporelle. De façon générale, la non indépendance des résidus peut être liée à d'autres facteurs. En sciences de l'eau, par exemple, on pourra mettre en évidence l'influence d'une variable spatiale (altitude, localisation voisine des points de mesure dans un bassin, embouchure versus tête d'un bassin versant, etc.).

4.2.4.2 Mise en évidence de l'autocorrélation

Pour s'assurer de l'indépendance des résidus, on peut utiliser un test graphique. Un exemple fictif est présenté à la figure 4.18. Cette figure représente la relation entre deux variables physico-chimiques. Les mesures ont été prises à des dates différentes (figure 4.18A). Les résidus ont été classés en fonction de la date de mesure et les variables e_i, t_i sont illustrées à la figure 4.18C. Ce graphique fait ressortir la dépendance systématique des résidus e_i en fonction de t_i, ce qui ne peut être mis en évidence à la figure 4.18B (x_i, e_i). Pour conclure à l'indépendance des résidus, on devra vérifier que ces derniers sont répartis de façon aléatoire, en fonction de t (figure 4.18D).

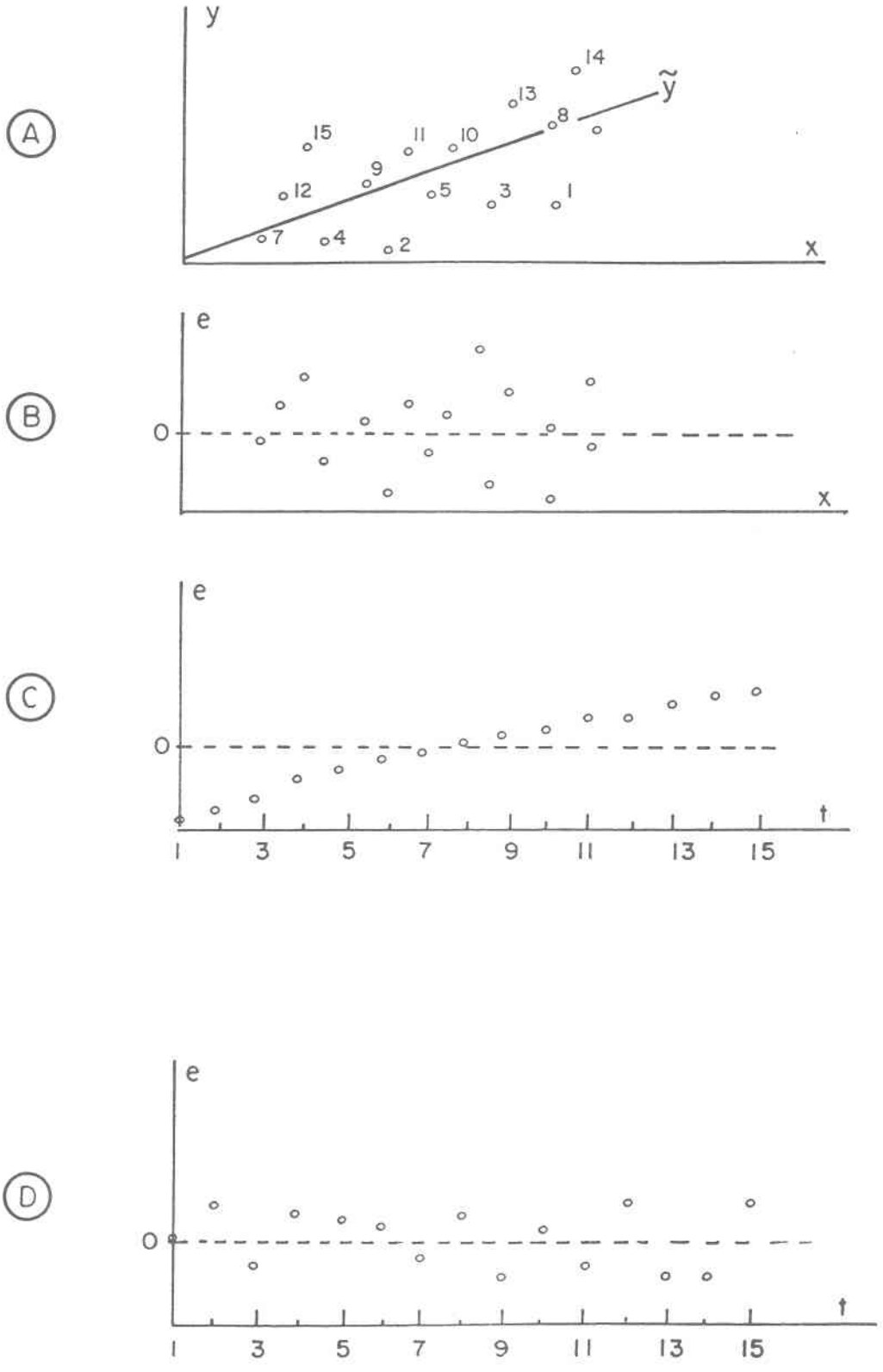


FIGURE 4.18 Exemple de résidus autocorrélés en fonction du temps. Modifié d'après Bobée (1983).

L'allure du graphique permet aussi de distinguer s'il s'agit d'autocorrélation positive ou négative (figure 4.19). Dans le premier cas, les résidus adjacents sont du même ordre de grandeur et de même signe (4.19A). Si l'autocorrélation est négative, les résidus alternent systématiquement entre les valeurs positives et négatives (4.19 C). D'après plusieurs auteurs (Chatterjee et Price, 1977; Montgomery et Peck, 1982; Bobée, 1983; Younger, 1979), les cas d'autocorrélation négative sont peu fréquents et de plus beaucoup plus difficiles à identifier.

On peut aussi retrouver le même type de problème pour des périodes homogènes (avant et après une pluie, saisons, etc.). Par exemple, chaque observation ne dépend pas d'une date de mesure, mais est liée à une saison. La figure 4.20 représente des variables physico-chimiques. Sauf exception, on peut observer que les résidus sont positifs durant l'automne et négatifs durant l'été, ce qui indiquerait qu'il faut introduire une variable explicative supplémentaire et utiliser la régression multiple. Cet aspect est repris à la section 4.3 où l'on discute de la représentativité de l'échantillon.

L'autocorrélation peut donc être mise en évidence par un test graphique. Il est parfois difficile de conclure à partir de ce test dont l'allure ne présente pas une variation aussi systématique que les graphiques présentés ici. On peut donc utiliser certains tests qui permettent de vérifier cette hypothèse de manière moins subjective.

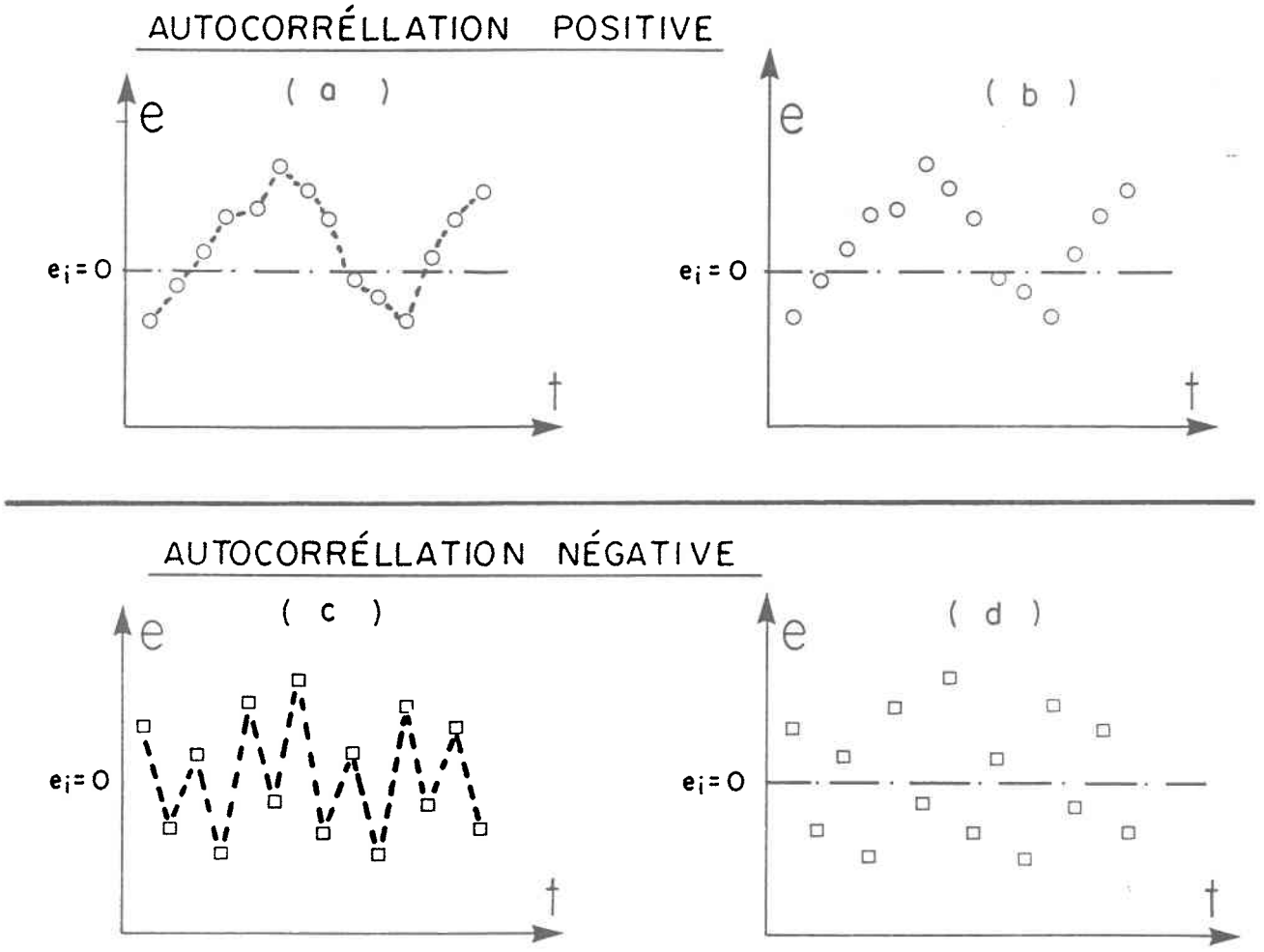


FIGURE 4.19 Illustration de l'autocorrélation positive et négative.

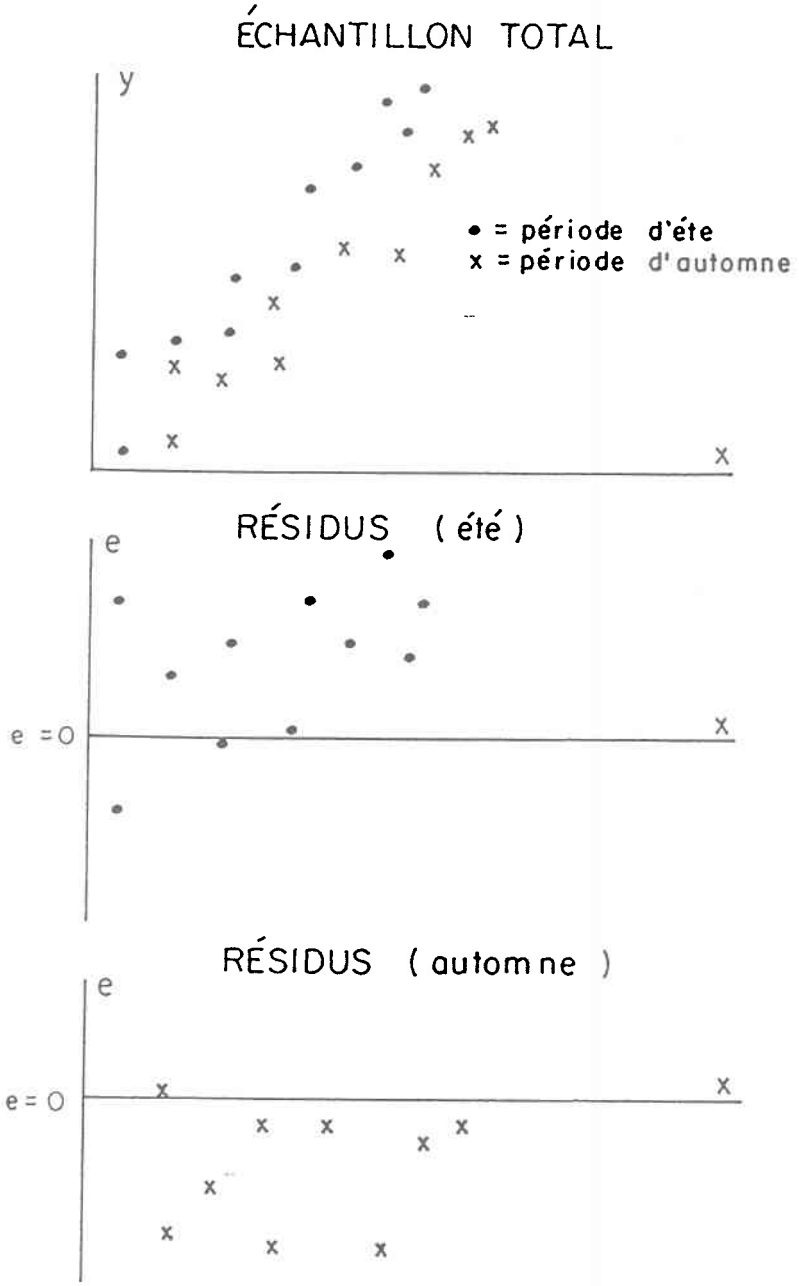


FIGURE 4.20 Autocorrélation due à l'hétérogénéité de l'échantillon

Le test de Durbin-Watson (1951) spécialement adapté à l'ajustement d'un modèle par la méthode des moindres carrés (Bobée, 1983) est décrit dans la suite. On trouvera d'autres tests également utiles dans Theil et Nagar (1961) et de nombreuses références pour différentes approches dans Montgomery et Peck (1982).

Nous considérons ici le cas où chaque mesure est prise à une date différente (altitude différente, etc.). Pour le cas où les variables semblent groupées en nuages distincts, figure 4.20, on vérifiera si les populations sont homogènes et si les droites de régression qu'on obtiendrait pour chacun des groupes sont significativement différentes. Ce problème particulier est abordé à la section 4.3.

- Test de Durbin-Watson

Le test de Durbin-Watson (1951), permet de vérifier s'il existe une certaine corrélation entres les diverses observations ordonnées en fonction du temps (t). On étudie le coefficient d'autocorrélation d'ordre 1 ($\bar{\rho}_1$) soit le coefficient d'autocorrélation entre les séries e_t et $e_{(t-1)}$, décalés d'une unité de temps. On considère la statistique D telle que

$$D = \frac{\sum_{t=2}^{N} (e_t - e_{t-1})^2}{\sum_{t=1}^{N} e_t^2} \qquad (4.22)$$

Pour tester si un coefficient d'autocorrélation est significativement positif, on considère les hypothèses suivantes:

$$H_0: \quad \rho = 0$$
$$H_1: \quad \rho > 0$$

On compare la quantité D aux valeurs critiques d_u et d_ℓ que l'on trouve dans les tables et que l'on choisi en fonction des critères suivants:

(1) Niveau de signification $\tilde{\alpha}$ ($\tilde{\alpha} = 0,01$ ou $\tilde{\alpha} = 0,05$)

(2) Nombre d'observations N. Pour appliquer ce test, on doit disposer d'un échantillon d'au moins 15 observations

(3) Nombre de variables indépendantes (k). Pour la régression simple on a donc: k = p-1 où p est le nombre de paramètres.

En fonction de la quantité D et des variables d_u et d_ℓ, on conclut de la façon suivante:

on rejette H_0 si $D < d_\ell$

on accepte H_0	si $D > d_u$
on ne peut conclure	si $d_\ell < D < d_u$

Pour tester si une autocorrélation est significativement négative, on étudie la quantité (4-D) et on considère l'hypothèse H_1, H_1: $\tilde{\rho} < 0$

on rejette H_0	si $4-D < d_u$
on accepte H_0	si $4-D > d_\ell$
on ne peut conclure	si $d_u < 4-D < d_\ell$

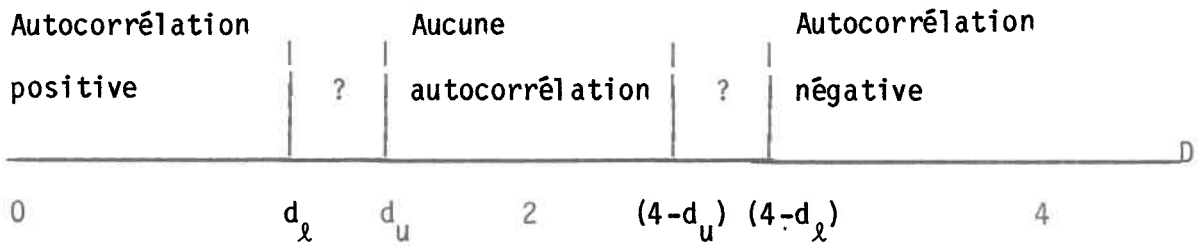
Lorsqu'on ne peut conclure en principe on devrait recueillir plus d'informations (échantillon plus grand). Si cela est impossible et c'est souvent le cas pour une série temporelle, on peut utiliser une approximation du test de Durbin-Watson, applicable si l'échantillon a une taille élevée (N > 40) (Durbin et Watson, 1951).

On peut présenter le test d'une autre façon qui est équivalente, mais qui fait mieux ressortir la rationalité du test. On compare la statistique D aux valeurs critiques d_ℓ et d_u (test d'autocorrélation positive) ou aux valeurs $(4-d_u)$ et $(4-d_\ell)$. La relation approximative (voir Chatterjee et Price, 1977) qui existe entre D et $\tilde{r}_1$ (estimation de $\tilde{\rho}_1$) $D = 2(1-\tilde{r}_1)$

avec
$$\bar{r}_1 = \frac{\sum_{t=2}^{N} e_t e_{t-1}}{\sum_{t=2}^{N} e_{t-1}} \qquad (4.23)$$

permet de voir que D peut être compris entre 0 et 4 et qu'une valeur proche de 2 permet de conclure à une autocorrélation nulle ou non significative. En fonction des différentes valeurs prises par D, on conclura de la façon illustrée à la figure 4.21

Figure 4.21 Critères de décision pour le test d'autocorrélation



4.2.4.3 Élimination de l'autocorrélation

Lorsqu'on peut mettre en évidence un problème d'autocorrélation, plusieurs approches peuvent être envisagées pour construire un modèle dont les résidus soient indépendants. Si on peut démontrer par exemple qu'il existe une dépendance des résidus en fonction du temps, on doit vérifier si cette dépendance est liée à une variable omise par le modèle (et qui varie dans le temps) ou s'il s'agit d'un problème de persistance (dépendance réelle des observations dans le temps).

Pour le premier cas, on doit identifier la variable clé omise par le modèle et qui est influencée par le temps de mesure. En sciences de l'eau, ce pourrait être par exemple le débit ou la température qui varie de façon systématique pendant la période d'échantillonnage. Si une telle variable peut être identifiée, elle devra être incluse dans le modèle. Le modèle résultant serait donc une régression multiple. On pourra par la suite vérifier l'indépendance des résidus par le test de Durbin-Watson (1951). Si les observations semblent regroupées en fonction d'un aspect particulier (saison, altitude, etc.), on peut envisager une régression pour chaque sous-groupe. Ce problème est discuté à la section 4.3.

Si l'on ne peut mettre en relief l'influence déterminante d'une seconde variable, on pourra alors utiliser un modèle autorégressif. Plusieurs des résultats théoriques de la régression ne sont plus valides lorsque les résidus sont autocorrélés. Bien que les paramètres soient non biaisés, la variance résiduelle S_e^2 sous estime σ_e^2. On obtient donc pour α et β des estimateurs S_a et S_b beaucoup plus petits qu'ils ne devraient l'être en réalité. Conséquemment, les résultats concernant les tests et la construction des intervalles de confiance et de prévision ne sont plus rigoureusement applicables. La prise en compte du coefficient d'autocorrélation permet d'obtenir des estimateurs beaucoup plus fiables.

Description générale du modèle autorégressif

On suppose dans ce qui suit une série où les observations peuvent être ordonnées dans le temps (t). Un modèle autorégressif tient compte du coefficient d'autocorrélation $\bar{\rho}_1$ entre les variables e_t et e_{t-1} qui permet de définir des nouvelles variables x'_t et y'_t. Les variables sont des fonctions linéaires de x_{t-1} et y_{t-1} de pente $\bar{\rho}_1$, soit:

$$x'_t = x_t - \bar{\rho}_1 x_{t-1}$$
$$y'_t = y_t - \bar{\rho}_1 y_{t-1} \qquad (4.24)$$

x_t, y_t sont les variables classées en fonction du temps. En utilisant ces nouvelles variables, on obtient la droite $\tilde{y}'$, soit:

$$\tilde{y}'_t = \alpha' + \beta' x'_t \qquad (4.25)$$

Deux approches sont généralement utilisées pour estimer les coefficients ce sont les méthodes de Cochrane-Orcutt (méthode itérative) et la méthode des "premières différences". Comme $\bar{\rho}_1$ n'est pas connu, il est estimé dans l'échantillon par $\bar{r}_1$ soit $\bar{r}_1 = \bar{r}_0$ (méthode de Cochrane-Orcutt). Pour la méthode des premières différences, on assume a priori que $\bar{\rho}_1 = 1$.

- Méthode de Cochrane-Orcutt

Comme le modèle 4.25 est linéaire, les coefficients α' et β' peuvent être estimés par la méthode des moindres carrés en effectuant une régression en fonction des variables x'_t et y'_t

$$x'_t = x_t - \bar{r}_1 x_{t-1}$$
$$y'_t = y_t - \bar{r}_1 y_{t-1}$$

avec $\bar{r}_1$ calculé dans l'échantillon d'après l'équation 4.23. On obtient donc un nouveau modèle

$$\tilde{y}_t = a' + b'x'_t \qquad (4.26)$$

On peut démontrer (Montgomery et Peck, 1982) qu'il existe entre les coefficients α et β ($\tilde{y} = \alpha + \beta x$) et α' et β' ($\tilde{y}'_t = \alpha' + \beta' x'_t$) la relation suivante:

$$\alpha = \frac{\alpha'}{(1-\bar{\rho}_1)} \qquad (4.27)$$

$$\beta = \beta' \qquad (4.28)$$

α' sera d'autant plus différent de α que $\bar{\rho}_1$ (estimé par $\bar{r}_1$) sera plus élevé. On doit par suite vérifier que les résidus du nouveau modèle ne sont pas autocorrélés en utilisant à nouveau le test de Durbin-Watson. Si les résidus sont encore interdépendants, on définit de nouvelles variables x" et y", en réestimant $\bar{\rho}$ à partir des résidus e'_i du modèle transformé et ainsi de

suite jusqu'à ce que l'autocorrélation soit éliminée.

Cette méthode n'est pas toujours très performante. Cela s'explique entre autres par le fait que lorsque les résidus sont corrélés positivement, $\bar{r}_1$ sous estime $\bar{\rho}_1$ (Montgomery et Peck, 1982; Neter et Waserman, 1974). Certains auteurs suggèrent de s'en tenir à une (ou deux) itération(s) après quoi on devrait utiliser une autre technique (Montgomery et Peck 1982; Chatterjee et Price, 1977).

En supposant une autocorrélation nulle suite à une première itération, on devra vérifier que la pente du modèle résultant est approximativement égale à celle du modèle original, ($b' \approx b$) sans quoi la méthode n'est pas appropriée.

- Méthode des premières différences

Cette méthode constitue un cas particulier de l'approche de Cochrane-Orcutt. On suppose cependant que $\bar{\rho}_1 = 1$. Étant donné (4.27), $\alpha' = \alpha(1 - \bar{\rho}_1) = 0$. La variable y'_t est donc égale à:

$$y'_t = \beta' x'_t + \varepsilon$$

et le modèle

$$y'_t = \beta' x'_t$$

Comme le modèle est centré à l'origine, β' est estimé par b' soit:

$$b' = \frac{\sum_{i=1}^{N} y'_t \, x'_t}{\sum_{i=1}^{N} x'^2_t}$$

et l'on doit là aussi vérifier que $b' \approx b$. Pour ce modèle, l'indépendance des erreurs ne peut être vérifiée par le test de Durbin-Watson qui n'est pas valide lorsque la droite est centrée à l'origine.

4.3 PROBLÈMES LIÉS À L'ÉCHANTILLONNAGE

4.3.1 Échantillon non homogène

On entend par échantillon non homogène, un échantillon dont les différents couples (x, y) proviennent de plusieurs populations différentes. Ce problème peut être illustré par un exemple. Prenons l'étude de la relation entre deux variables physico-chimiques x et y. Les différents couples sont présentés à la figure 4.22 A. Les échantillons proviennent de deux lacs dont les observations apparaissent comme deux groupes distincts. Les pentes respectives b_1 et b_2 ne sont pas significativement différentes de zéro. Par contre, la pente obtenue pour toutes les observations est très significativement différente de zéro et laisse croire à une relation importante entre les variables x et y, relation uniquement due à la mise en commun des deux groupes.

D'un point de vue graphique, ce problème peut souvent être mis en évidence si les observations sont distribuées en groupes distincts correspondant à différentes populations. Cependant, si tous les couples proviennent de populations différentes, par exemple 15 mesures et 15 lacs, le problème sera plus difficile à déceler à moins d'avoir des informations précises concernant le mode d'échantillonnage.

On ne doit pas conclure cependant que deux échantillons obtenus de façon indépendante (deux lacs, deux méthodes de mesure, etc.) sont

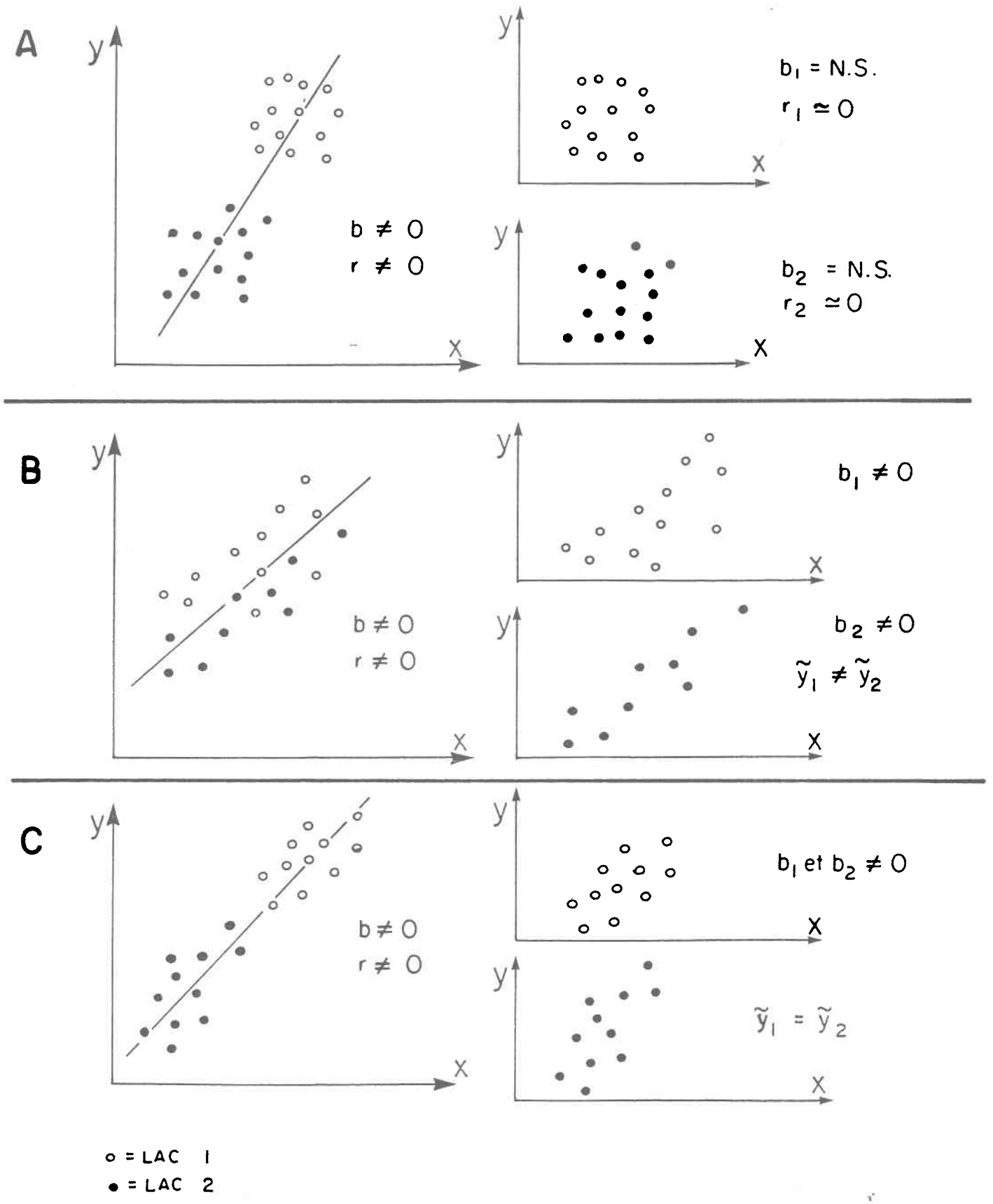


FIGURE 4.22 Vérification de l'homogénéité de l'échantillon

nécessairement issus de populations différentes. Si les variables obtenues semblent regroupées en fonction d'un aspect quelconque (lac_1, lac_2, lac_3 ou lac de tête, lac à l'embouchure), on doit d'abord vérifier si les pentes des droites de régression sont significativement différentes de zéro pour chaque groupe. En supposant les pentes b_1 et b_2 significatives, on peut ensuite vérifier par la méthode décrite au chapitre 2 (section 2.3.8) si les droites $\tilde{y}_1$ et $\tilde{y}_2$ sont significativement différentes. Les graphiques B et C de la figure 4.22 présentent ces différents cas. En B les pentes b_1 et b_2 sont significatives mais les droites $\tilde{y}_1$ et $\tilde{y}_2$ diffèrent. En C les droites $\tilde{y}_1$ et $\tilde{y}_2$ ne sont pas significativement différentes. Des exemples réels correspondant à ces différents cas sont présentés dans ce qui suit.

EXEMPLE 4.8

L'exemple suivant représente une expérience où l'on tente de mesurer la capacité d'une plante à régulariser la quantité d'eau absorbée (y) en fonction de la capacité de rétention de l'eau d'un sol (x). Les observations mesurées sont présentées à la figure 4.23A.

Les auteurs Ungar et al. (1979) concluent à une relation très significative entre les variables x et y, ce qui est vrai si on applique un test de signification de la pente (section 2.3.5.2) pour les 56 observations. On a $t_0 = 9,12$ et $t_{54}\ (0,05) = 1,67$.

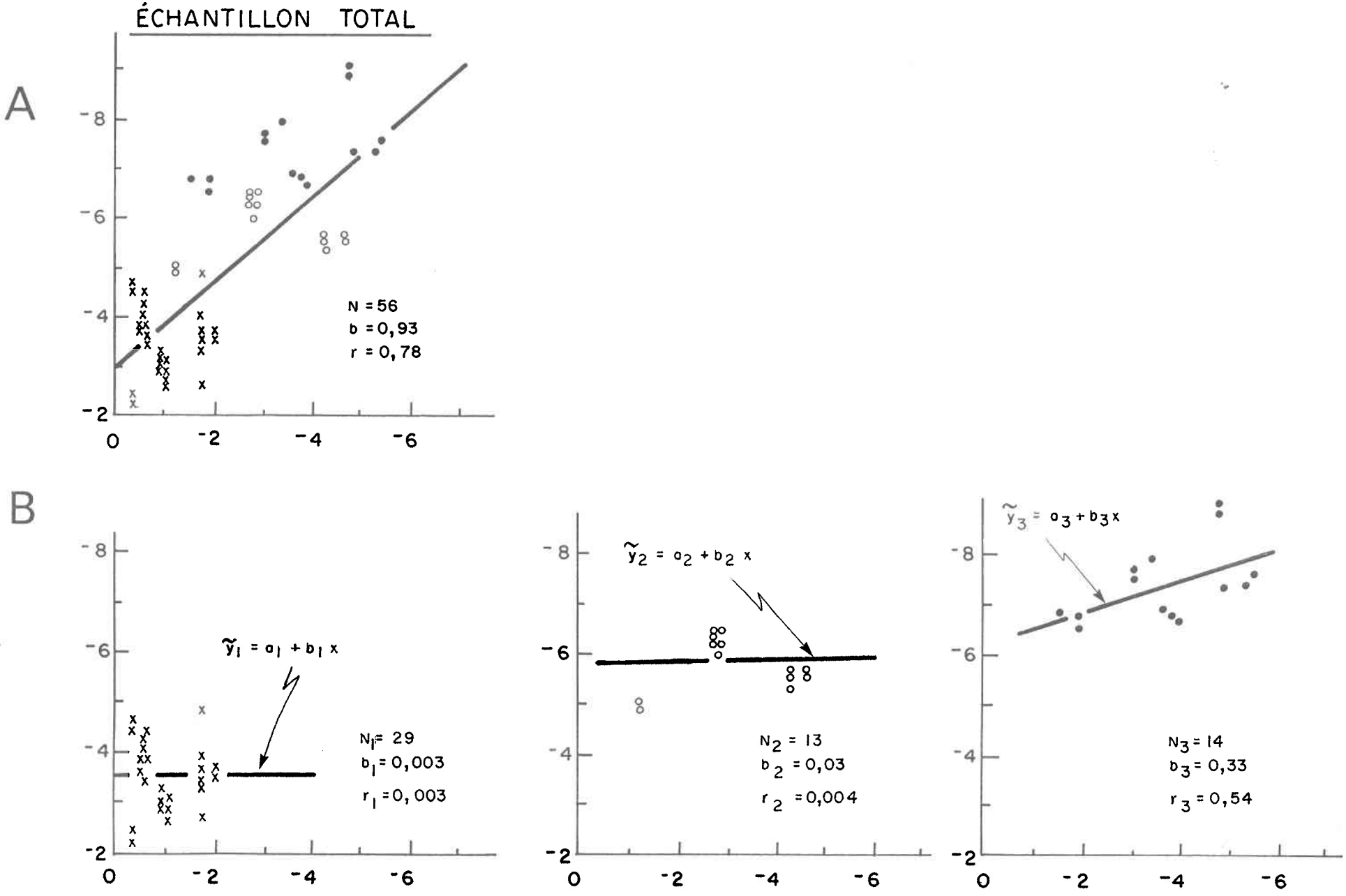


FIGURE 4.23 Étude de la nature de l'échantillon. A : Échantillon total. B : Sous-groupes homogènes (exemple 4.8). Données tirées de Ungar et al. (1979)

Étant donné que $t_0 > t_c$, et ce, même pour un niveau de signification égal à 0,0005 (t_{54} (0,0005) = 3.49) on peut rejeter l'hypothèse nulle et conclure à une relation très significative.

Comme on l'a dit, les résultats des tests sont valides si les hypothèses de la régression sont respectées (linéarité, homoscédasticité, normalité des résidus) ce qui a été vérifié pour cet exemple. Cependant, on peut mettre en doute la représentativité de l'échantillon. En effet, les variables ne forment pas un groupe homogène. L'échantillon est constitué de valeurs obtenues en laboratoire: groupe 1 figure 4.23B et de valeurs obtenues sur le terrain à deux périodes différentes: groupes 2 et 3 figure 4.23B. Comme on peut l'observer, ces observations forment des groupes distincts. Or si l'on teste les pentes de chacun des groupes, on peut vérifier qu'elles ne sont pas significatives (pentes b_1 et b_2) ou très peu significatives (b_3). De même, on obtient pour chaque groupe un coefficient de corrélation pratiquement nul (r_1 et r_2) ou faiblement significatif (r_3).

Il en résulte donc une relation significative due en grande partie à la mise en commun des sous groupes 1, 2 et 3. On peut surtout critiquer ici la prise en compte du groupe des valeurs obtenues en laboratoire (visiblement utilisées ici pour combler le manque d'observations entre x = 0 et x = -2). En effet, sur le terrain, on peut supposer l'influence conjugée de plusieurs autres variables (vent, humidité, température, etc.) lesquelles demeurent

constantes dans une expérience contrôlée en laboratoire. En supposant que la période de mesure n'influence pas ou influence peu la relation de y et x on pourrait étudier les groupes 2 et 3. On peut vérifier que pour ce groupe, la pente est significative du moins pour $\tilde{\alpha} = 0,01$. De même, on obtient un coefficient de corrélation significatif r = 0,38 pour N = 27, qui est cependant beaucoup moins élevé que le coefficient obtenu pour l'échantillon total r = 0,78 pour N = 56. On obtient aussi des estimateurs a et b très différents lorsque l'on considère l'échantillon total (tableau 4.10).

TABLEAU 4.10 - COMPARAISON DES ESTIMATEURS a ET b DE L'ÉCHANTILLON TOTAL ET DE SOUS-GROUPES HOMOGÈNES (exemple 4.8)

Groupe considéré	Paramètres N	a	b	t_0	Test de la pente $t_{(N-2)}(\tilde{\alpha})$	
					$\tilde{\alpha} = 0,01$	$\tilde{\alpha} = 0,05$
1	29	-3,48	0,003	0,0138	2,473	1,073
2	13	-5,73	0,029	0,2004	2,718	1,796
3	14	-6,20	0,332	2,211	2,681	1,782
2 et 3	27	-5,51	0,331	2,068	2,485	1.708
TOTAL	56	-2,97	0,928	9,117	2,400	1,675

EXEMPLE 4.9

L'exemple qui suit est tiré d'un article présenté par Hedges et al., (1984). Le graphique de la figure 4.24 a été construit à partir des données

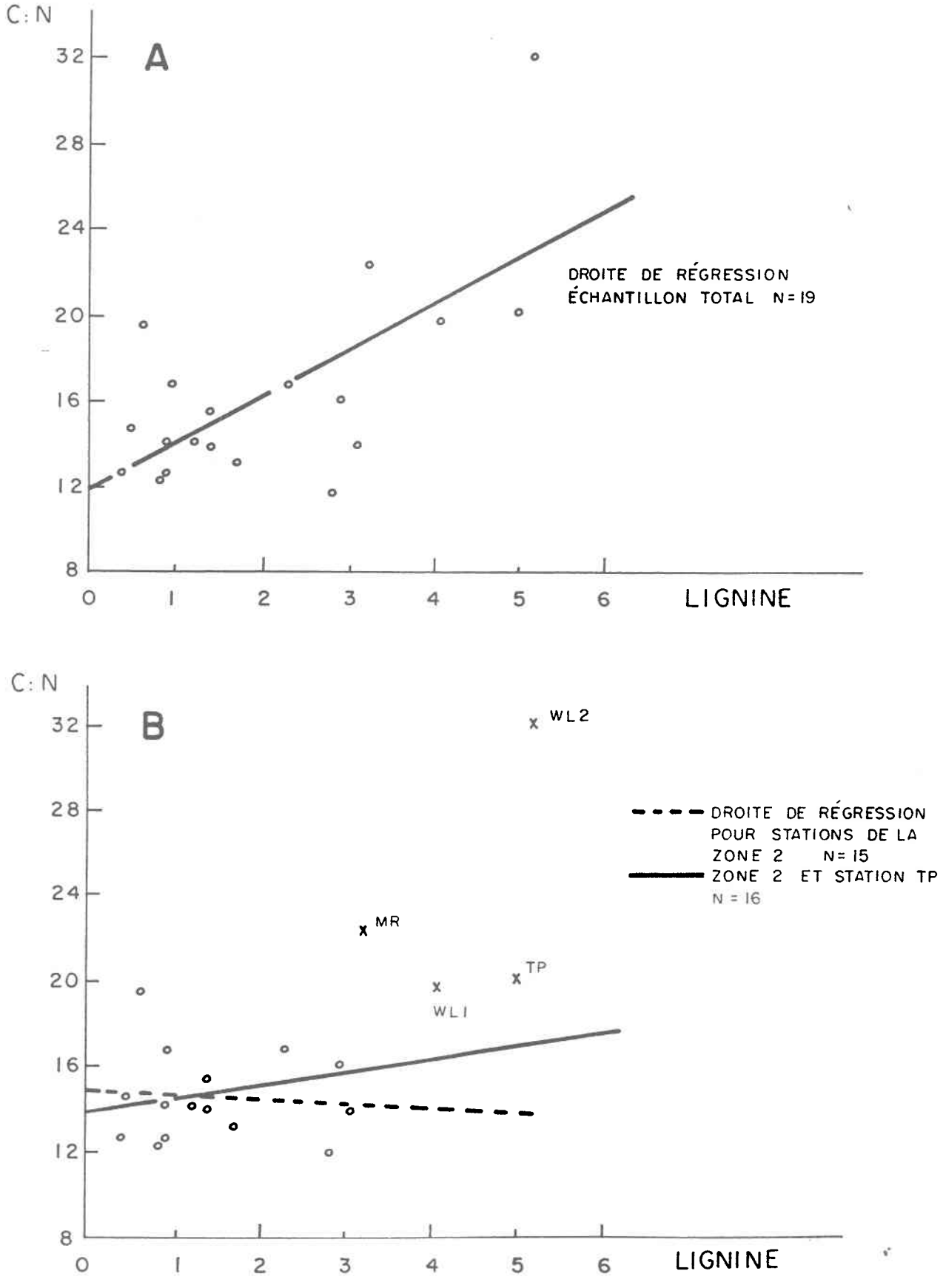


FIGURE 4.24 Vérification de l'homogénéité de l'échantillon (exemple 4.9)
Données tirées de Hedges et al.(1984)

originales. Ce graphique met en relation la quantité totale de lignine et le rapport carbone-azote dans les sédiments de la rivière Columbia et de ses principaux tributaires. Pour cette étude, 19 sites furent échantillonnés pour l'ensemble du bassin de la rivière Columbia. Le test de signification de la pente (ou encore du coefficient de corrélation) permet de conclure à une relation très significative ($r = 0,68$).

Dans le cadre général de l'étude qui porte sur plusieurs autres aspects, les auteurs ont identifié deux régions très distinctes du point de vue de la végétation que l'on nommera ici zone 1 et zone 2. Les différentes stations peuvent être incluses dans l'une ou l'autre de ces zones. La rivière "WL" (stations WL 1 et WL 2) coulent entièrement dans la zone 1. La rivière "MR" (station MR) coule presqu'entièrement dans cette même zone sauf pour une faible partie de son cours. Il en est de même pour la station identifiée par TP (situé à l'embouchure du bassin versant). Ces différentes stations sont identifiées à la figure 4.24B. Les 15 autres points de mesure sont situés dans la zone 2. Les 4 stations de la zone 1 (WL 1, WL 2, MR et TP) se différencient aussi des stations de la zone 2 par la nature des dépôts meubles constitués principalement de sables grossiers (MR et TP) ou d'un mélange de matériel grossier et fin (WL 1 et WL 2). Les dépôts meubles de la majorité des 15 autres stations sont constitués de matériel très fin (sauf une exception).

L'étude de la relation entre la lignine et le rapport C/N (carbone/azote) par région homogène ne permet plus de conclure à une relation significative. En effet, pour les 15 points de mesure de la zone 2, on obtient une pente pratiquement nulle et donc non-significative (voir figure 4.24B). On arrive d'ailleurs aux mêmes conclusions en incluant la station TP qui, étant située à l'embouchure du bassin principal, est dans une certaine mesure influencée par la zone 2. L'inclusion des trois autres stations (WL 1, WL 2, et MR), qui sont strictement comprises dans la zone 1 permet cette fois d'obtenir une relation très significative. On est donc en droit de se demander si la relation de y et x n'est pas significative uniquement en raison de ce regroupement.

En supposant que ces caractéristiques physiques n'aient aucune influence sur les variables étudiées, il ne semble pas que le modèle linéaire soit le plus adéquat pour décrire la relation de y et x . De plus, le faible pouvoir explicatif du modèle linéaire est en grande partie masqué, si l'on peut dire, par l'influence très importante de la station WL 2. En effet, on obtient pour les 18 premières stations un coefficient de détermination très faible soit $R^2 = 0,2886$. Or, l'inclusion du point extrême WL 1 a pour effet d'augmenter ce coefficient de près de 20%: $R^2 = 0,4659$ (N = 19).

EXEMPLE 4.10

Le test de la pente permet de vérifier si la relation obtenue par groupe homogène est significative. Comme on l'a vu pour les deux exemples précédents, dans certains cas, seule la mise en commun des différents groupes permet d'obtenir une relation significative.

En supposant d'autre part des pentes significatives pour les groupes considérés, on ne peut cependant conclure que les variables proviennent d'une même population, et ce, même si les pentes obtenues étaient identiques. En effet, deux échantillons dont les pentes sont identiques peuvent néanmoins représenter des populations fort différentes. On doit donc dans l'ordre effectuer:

- un test de signification des pentes;
- un test d'égalités des pentes des groupes considérés;
- un test d'identité des droites.

Nous avons appliqué cette procédure pour l'exemple suivant lequel met en relation l'énergie totale (lumineuse et calorifique) appliquée à une eau résiduaire et le taux d'épuration de cette eau. Cette expérience fut menée dans deux bassins de profondeurs différentes. Les droites obtenues pour le bassin 1 (35 cm) et le bassin 2 (25 cm) sont illustrées à la figure 4.25. L'auteur (Pouliot, 1985) concluait qu'étant donné l'égalité des pentes ont pouvait conclure à l'identité des procédés utilisant des bassins de profondeurs différentes.

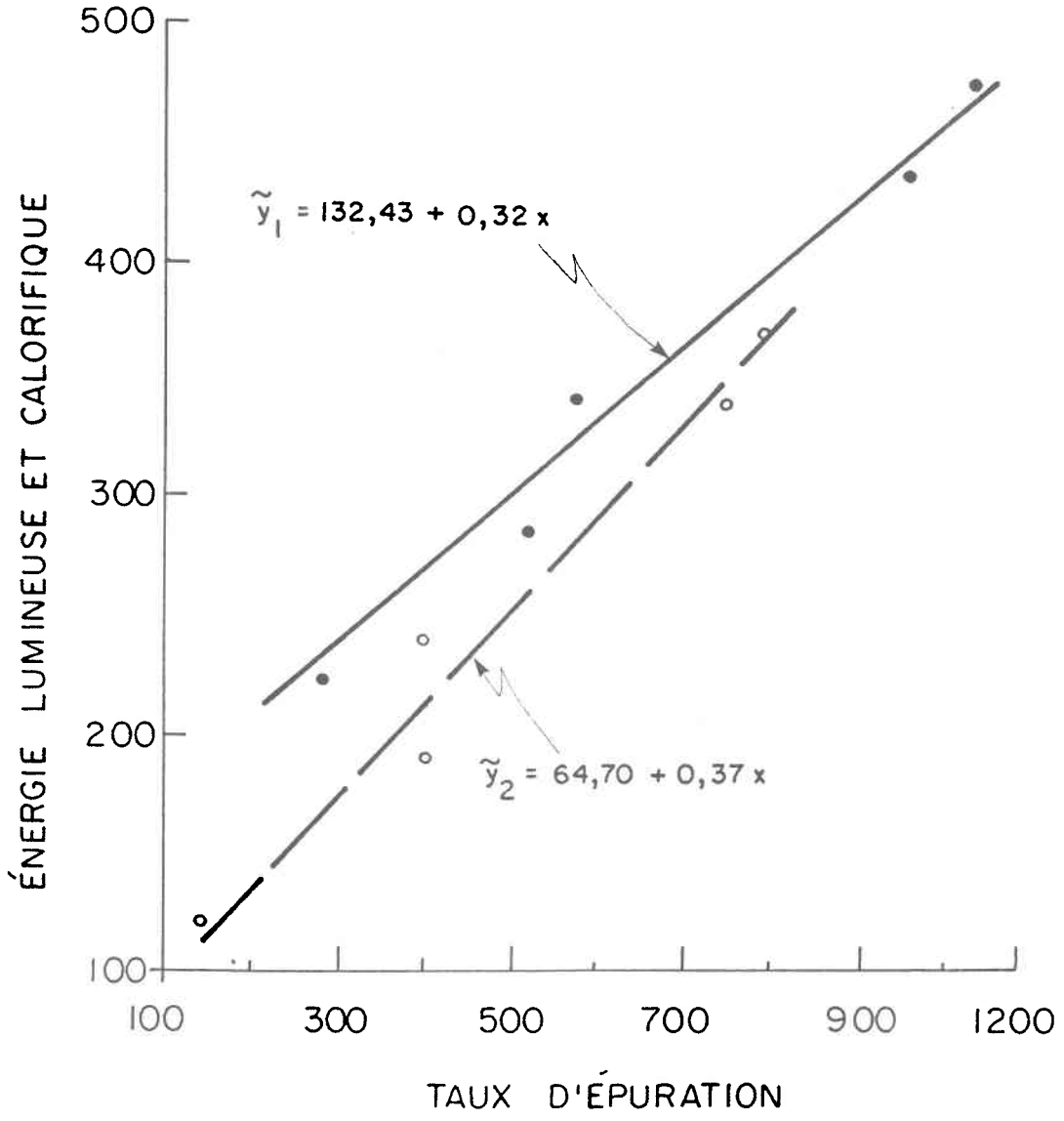


FIGURE 4.25 Comparaison de deux droites de régression (exemple 4.10)
Données tirées de Pouliot (1985)

Nous avons repris le test de comparaison des pentes. En effet, on ne pouvait se fier aux résultats du test utilisé dans l'étude, ce test étant inadéquat. La quantité t (équation A 1.6) utilisée pour effectuer le test soit: $t = b-\beta/S_b$ ne permet pas de comparer deux pentes (supposons b_1 et b_2), mais est utile pour vérifier si une pente b est significativement différente d'une pente théorique $\beta = \beta_0$.

Dans ce qui suit, nous avons d'abord comparé les variances résiduelles $S^2_{e_1}$ (bassin 1) et $S^2_{e_2}$ (bassin 2) pour être en mesure de choisir le test adéquat de comparaison des pentes (voir discussion à cet effet à la section 2.3.8.2). Les droites furent ensuite comparées. Les tests présentés à la section 2.3.8 et à l'appendice 4 furent appliqués.

(a) Vérification de l'égalité des variances (Test F).

On calcule d'abord F_0 que l'on compare à F_c avec $\tilde{\alpha} = 0,05$

$$F_0 = 456,93/292,05 = 1,56$$

$$F_c = F_{3,3}\ (0,05) = 9,28$$

Puisque $F_0 < F_c$ on accepte l'hypothèse H_0 d'égalité des variances.

(b) Vérification de l'égalité des pentes

Étant donné que les variances ne sont pas significativement différentes, on peut utiliser le test t pour vérifier l'égalité des pentes. Puisque N_1 et N_2 sont petits ($N_1 = N_2 = 5$) on utilise $\tilde{S}^2_{et}$ défini à l'appendice 4.

$$t_0 = \frac{b_1 - b_2}{\tilde{S}^2_{et}\left((1/SS_{x_1} + 1/SS_{x_2})\right)^{\frac{1}{2}}} = -1,05$$

Pour un test bilatéral avec $\tilde{\alpha} = 0,05$ et $(N - 2)$ degrés de liberté on a, $t_c = t_8\ (0,025) = 2,306$. Puisque $t_0 = < |t_c|$ on accepte H_0.

(c) Vérification de l'identité des droites (test t)

Étant donné que les pentes ne sont pas significativement différentes, on peut donc vérifier l'identité des droites.

$$\bar{b} = \frac{0,3215(394\ 693,30) + 0,3717\ (292\ 496,80)}{394\ 693,30 + 292\ 496,80} = 0,3429$$

Étant donné que N est petit, on utilise $\tilde{S}^2_{et}$ plutôt que S^2_{et} pour estimer la variance de $\bar{b}$ et de $\tilde{b}$ (voir Appendice 4).

$$S^2_{\tilde{b}} = \frac{\tilde{S}^2_{et}}{SS_{x_1} + SS_{x_2}} = \frac{381,34}{687\ 190,10} = 0,000555$$

$$\tilde{b} = \frac{\bar{y}_1 - \bar{y}_2}{\bar{x}_1 - \bar{x}_2} = \frac{350,96 - 251,2}{679,80 - 501,8} = 0,5604$$

$$S^2_{\tilde{b}} = \frac{\tilde{S}^2_{et}}{(\bar{x}_1 - \bar{x}_2)} \left(\frac{1}{N_1} + \frac{1}{N_2}\right) = \frac{381,34\ (0,4)}{(31\ 684,00)} = 0,0048143$$

$$Var\ (\tilde{b} - b) = S^2_{\tilde{b}} + S^2_{b} = 0,0053693$$

$$t_0 = \frac{\tilde{b} - b}{\sqrt{Var\ (\tilde{b} - b)}} = \frac{0,5604 - 0,3429}{0,0732799} = 2,968$$

$$t_c = t_6(0,025) = 2,447$$

Pout un test bilatéral et un niveau $\tilde{\alpha} = 0,05$, on obtient $t_0 > t_c$ et on conclue donc que les droites sont significativement différentes.

(d) Vérification de l'identité des droites (Test F)

$$F_0 = \frac{(SSR - SS_{r_{12}})\ /2}{SS_{r_{12}}\ /\ (N_1 + N_2 - 4)} = \frac{(6032,9381 - 224,06)\ /2}{2246,06\ /\ 6} = 5,06$$

Pour que les résultats du test soient comparables au test précédent, on choisit un test bilatéral et un niveau $\tilde{\alpha} = 0,05$. La valeur critique $F_c = F_{2,6}\ (0,025)$ est égale à 7.26. On a donc; $F_0 < F_c$ et on ne peut conclure que les droites sont différentes.

Commentaires

Comme on l'a souligné précédemment, l'égalité des pentes n'est pas équivalente à l'identité des droites. Par le test t (test de Hald), on peut en effet vérifier que, pour un test bilatéral et un niveau $\tilde{\alpha} = 0,05$ les droites sont significativement différentes bien que les pentes n'étaient pas significativement différentes. Il est important de souligner que le test F de comparaison des droites ne permet pas cependant de rejeter H_0. Par le test t et le test F et un niveau $\tilde{\alpha} = 0,05$, on arrive donc à des conclusions différentes. A notre connaissance, la puissance respective de ces tests n'a pas été mise en évidence mais le test t semble plus rigoureux puisqu'il propose certaines corrections lorsque l'échantillon est de faible taille. On doit aussi considérer qu'en réduisant la valeur de $\tilde{\alpha}$ ($\tilde{\alpha} = 0,01$), les résultats des tests sont identiques et $H_0 : \eta_1 = \eta_2$ est acceptée. Il est donc assez difficile de conclure quant à l'identité des droites. Seul un échantillon plus élevé permettrait de minimiser le risque d'erreur qui est d'autant plus grand que l'échantillon est faible, comme c'est ici le cas.

4.3.2 Échantillon non représentatif (valeurs extrêmes et valeurs aberrantes.)

L'examen graphique des observations permet très souvent de déceler dans l'échantillon la présence d'un ou de plusieurs valeurs extrêmes (outliers). Dans certains cas, on peut démontrer que ces valeurs sont aberrantes. En effet, pour des raisons théoriques, il est parfois possible de mettre en évidence la non-validité de certaines mesures qui ont pu être obtenues à cause d'une erreur de lecture de l'expérimentateur, d'une erreur de transcription ou encore en raison d'une défectuosité de l'appareil de mesure utilisé. Ces valeurs peuvent alors être éliminées.

Si tel n'est pas le cas, on peut toujours effectuer un test de normalité et vérifier si ces résidus ou encore les variables x et y sont comprises à l'intérieur de ± K écarts types. Ce test est plutôt grossier et c'est pourquoi il est suggéré d'utiliser un test plus objectif (Gunst et Mason, 1980; Bobée, 1983). Nous présentons ici le test de Grubbs (1969) suggéré par Bobée (1983).

Test de Grubbs

L'hypothèse H_0 est que toutes les valeurs appartiennent à la même population. Si H_0 est rejetée, on est amené à conclure que la valeur extrême n'appartient pas à la distribution des (N - 1) autres observations. On suppose ici une valeur extrême par rapport à l'axe des x. Les valeurs de x sont classées par ordre croissant

$$x_1 \leq x_2 \leq \ldots x_N$$

où x_N constitue la valeur extrême. On détermine ensuite la quantité SS_x:

$$SS_x = \sum_{i=1}^{N} (x_i - \bar{x})^2 = S_x^2 \quad (N - 1)$$

On estime ensuite la même quantité en éliminant la valeur qui semble douteuse soit SS'_x . On a donc SS_x pour les N observations et SS'_x pour les (N - 1) observations. On calcule par suite le rapport B:

$$B = \frac{SS'_x}{SS_x}$$

Plus x_N est extrême et plus le rapport B devient petit. Cette valeur est comparée à une valeur critique B_c calculée par Grubbs (1969). Pour un échantillon de N observations et un niveau de signification $\tilde{\alpha}$, les quantités B et B_c seront comparées de la façon suivante.

Si $B < B_c (\tilde{\alpha}, N)$ la valeur x_N doit être rejetée

Si $B > B_c (\tilde{\alpha}, N)$ la valeur x_N est acceptée.

Ce test est valide si l'hypothèse de normalité est respectée. Les résultats du test peuvent donc être faussés si l'échantilon considéré ne respecte pas cette hypothèse. Ce peut être le cas par exemple si la relation entre les variables est non-linéaire.

Le test peut aussi s'appliquer si deux observations semblent extrêmes. On calcule alors SS_X et SS''_X, ou SS''_X est constitué des (N-2) observations restantes.

Il faut bien comprendre que ce test ne permet pas de décider si nous sommes en présence de valeurs aberrantes ou simplement de valeurs réelles mais exceptionnelles. Seules certaines connaissances théoriques du phénomène permettront d'en décider.

Commentaires et exemples

Une valeur extrême peut se situer dans l'axe principal du nuage de points des autres observations et être approximativement au centre de celui-ci. Dans ce cas, cette valeur aura peu d'effet sur l'équation théorique du modèle (estimation des coefficients a et b). D'autre part, une valeur peut être tout à fait marginale par rapport aux autres observations et la considérer peut donner lieu à un modèle peu représentatif pour les (N - 1) autres points. Il est surtout important de souligner ici l'influence très grande que peut avoir un point extrême sur les tests de signification et particulièrement pour le test de pente, mais aussi du coefficient de corrélation. À cet effet, nous ferons ressortir par quelques exemples une tendance assez répandue chez plusieurs chercheurs. Lorsque certaines valeurs affaiblissent le modèle (ou la valeur du coefficient de corrélation), on a plutôt tendance à rejeter ces valeurs, les considérant d'emblée comme valeurs aberrantes.

À l'inverse, lorsqu'une ou quelques valeurs permettent d'obtenir une relation significative, ces valeurs sont cette fois-ci retenues et considérées comme significatives. A la limite, ce procédé aboutit parfois à l'établissement de corrélations fictives.

Pour la majorité des articles consultés, les tests utilisés par les auteurs portent sur le coefficient de corrélation plutôt que sur la pente. C'est pourquoi nous reportons au chapitre 5 la majorité des exemples qui traitent de ce problème. Un seul exemple sera présenté ici. Cet exemple illustre bien l'influence des points extrêmes sur l'équation du modèle linéaire et sur le test de la pente.

Exemple 4.11

L'échantillon (N = 15) analysé ici est présenté par Bonny (1978). Tous les graphiques de cette étude faisaient nettement ressortir la présence de 2 points extrêmes. Pour un des graphiques (figure 4.26A) nous avons vérifié l'influence de ces points sur l'estimation des coefficients du modèle linéaire et sur les tests de signification. Le modèle présenté dans cet article possède les caractéristiquees suivantes (voir figure 4.26B): l'ordonnée à l'origine est négative ce qui physiquement n'a aucun sens pour ce phénomène représentant la quantité de pollen déposé en fonction de la pluie. La pente est très significative. On obtient une quantité $t_0 = 4,86 > t_{13}\ (0,05) = 1,77$. En éliminant les deux points extrêmes (N = 13) on obtient pour le modèle linéaire (figure 4.26B) un ordonnée à

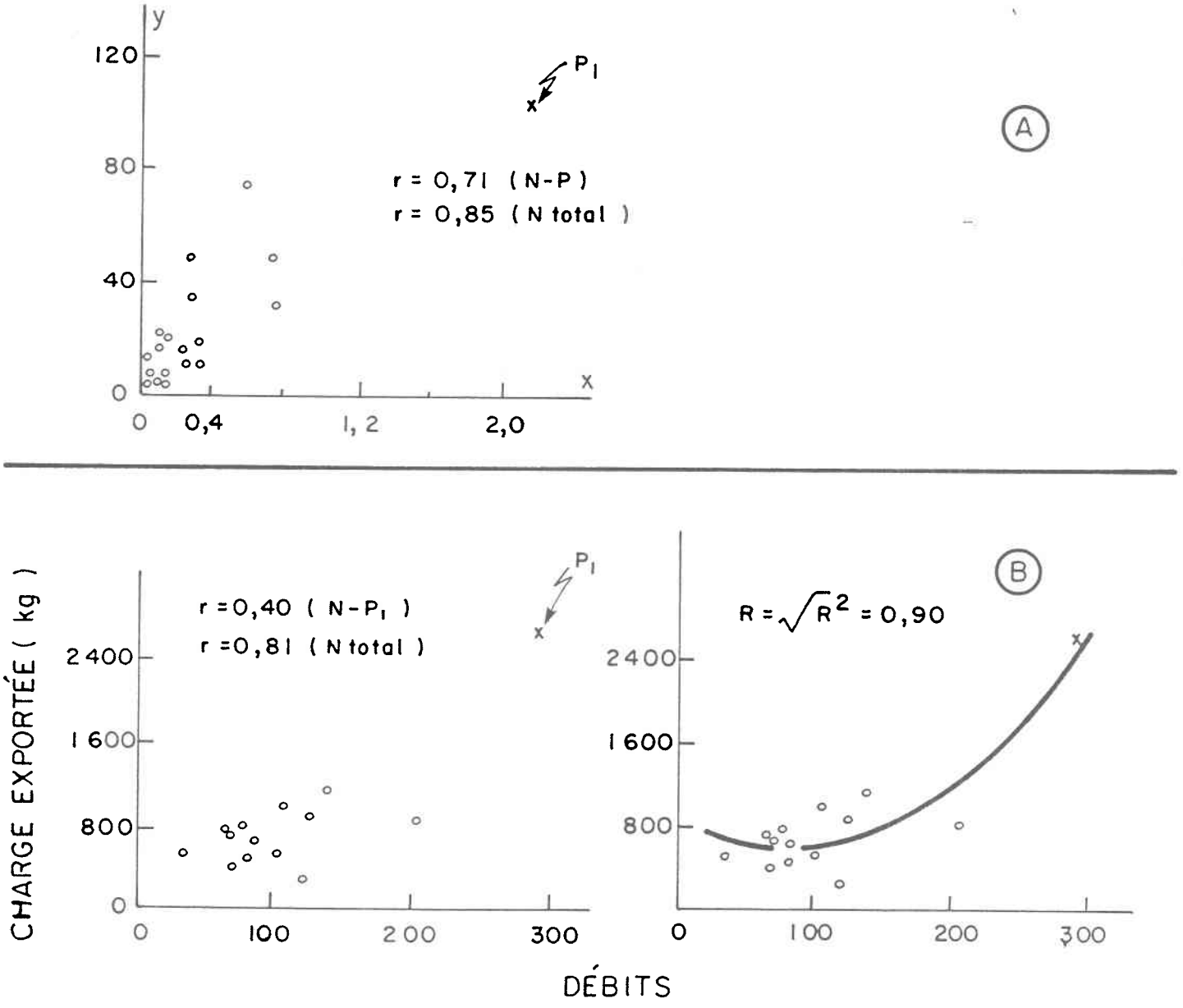


FIGURE 4.26 Influence de points extrêmes sur les coefficients du modèle de régression linéaire (exemple 4.11)
Données tirées de Barny (1978)

l'origine positive mais une pente non significative (b = 0,0015) avec $t_0 = 1,63 < t_{11}\ (0,05) = 1,80$.

En considérant l'échantillon total on obtient donc une relation significative mais qui représente mal les 13 premières observations (valeur de y négative pour $x < 50$ et pente très forte alors que la croissance est pratiquement nulle). En éliminant les deux points extrêmes (qui par ailleurs sont rejetés par le test de Grubbs s'applique) la relation est non significative. On peut aussi ajouter que la prise en compte de ces deux points fait pratiquement doubler la valeur du coefficient de corrélation égal à 0,44 pour N = 13 et à 0,80 pour N = 15.

Si on accepte les résultats du test de Grubbs on en vient à conclure que la relation est non-significative pour les 13 premières observations. Cependant, on peut aussi supposer que la population d'origine est non-normale, et que par exemple la relation est non-linéaire. À ce titre nous avons ajusté deux autres modèles. On pourra d'ailleurs vérifier que pour ces modèles (figure 4.26C) on obtient une quantité SS_r moindre que pour le modèle linéaire.

CHAPITRE 5

ANALYSE DE CORRÉLATION : CORRÉLATION FICTIVE ET AUTRES UTILISATIONS ABUSIVES DU COEFFICIENT DE CORRÉLATION.

CHAPITRE 5 : ANALYSE DE CORRÉLATION : CORRÉLATION FICTIVE ET AUTRES UTILISATIONS ABUSIVES DU COEFFICIENT DE CORRÉLATION.

5.1 INTRODUCTION

Les aspects théoriques utiles pour l'estimation et l'interprétation du coefficient de corrélation ont été présentés au chapitre 3.

Dans le présent chapitre nous étudions les principaux abus concernant l'utilisation de ce coefficient. Tout comme en analyse de régression on doit pour utiliser et interpréter le coefficient de corrélation vérifier que certaines conditions et hypothèses sont respectées. Ces hypothèses concernent particulièrement le mode d'échantillonnage (X aléatoire) et la nature de la distribution jointe entre X et Y.

D'autres aspects doivent être considérés lorsqu'on analyse la valeur de r. D'une part si l'on peut démontrer que la distribution jointe est normale cela n'assure pas que le coefficient soit significatif, ce qu'il faudra vérifier par un autre test (test de signification).

Nous aborderons également le problème des corrélations fictives. Il suffit de consulter quelques auteurs pour se rendre compte des sens très différents donnés à ce terme. À cet effet nous avons jugé utile de faire certaines distinctions entre, par exemple, la corrélation due au hasard (non-sense corrélation) et qui n'a aucun sens physique et la "corrélation

fictive" (spurious correlation) liée à l'introduction de variables communes.

Enfin un sujet peu discuté dans la littérature concerne l'utilisation du coefficient de corrélation suite à la transformation des variables originales X et Y. Cet aspect sera également abordé dans ce chapitre.

5.2 COEFFICIENT DE CORRÉLATION ET NATURE DE LA DISTRIBUTION JOINTE

5.2.1 Généralités

Comme on l'a déjà signalé (section 3.4), l'interprétation de la valeur du coefficient de corrélation est sans équivoque si la distribution jointe entre X et Y est une loi normale. Il est important de noter que si la distribution jointe bivariée est normale les distributions marginales sont aussi normales. Cependant, la réciproque n'est pas toujours vraie. En effet, si les distributions marginales sont normales on ne peut rien conclure quant à la nature de la distribution jointe (Neter et Wasserman, 1974, Hald, 1952).

On peut envisager deux procédures pour s'assurer que les conditions d'utilisation du coefficient de corrélation sont respectées: vérifier si les distributions marginales sont normales. Si au moins une d'entre elles ne l'est pas on peut immédiatement conclure à la non-normalité de la distribu-

tion jointe. Si les distributions marginales sont normales, on devra vérifier aussi la nature de la distribution jointe. Si l'on soupçonne qu'au moins une des deux distributions marginales est non normale on devrait envisager cette procédure car elle est beaucoup moins longue. Sinon il vaut mieux vérifier uniquement la nature de la distribution jointe. Les principaux tests et méthodes graphiques présentés à la section 4.2.3 pour vérifier la normalité des résidus peuvent également être appliqués pour vérifier la normalité des distributions marginales des variables X et Y. Le test qui permet de vérifier la normalité de la distribution jointe est beaucoup moins connu et c'est pourquoi nous avons jugé opportun de la présenter de façon détaillée. En fait, de toutes les publications consultées, où l'on employait le coefficient de corrélation nous n'avons jamais vu l'utilisation de ce test.

5.2.2 Vérification de la normalité de la distribution jointe de X, Y. (Hald, 1952).

Nous résumons ici les principales étapes à suivre pour vérifier la normalité de la distribution jointe de X et Y par un test graphique. À partir de l'échantillon des observations x_i, y_i on estime le coefficient de corrélation r. On considère aussi les variables u et v qui représentent respectivement les variables standardisées de x et y. Pour les différents couples u_i, v_i on calcule la quantité B_i soit:

$$B_i = \frac{1}{(1-r^2)} \left(u_i^2 - 2r\, u_i v_i + v_i^2\right) \qquad (5.1)$$

Si la distribution jointe de u_i, v_i est normale, B_i suit une loi CHI-2 à 2 degrés de liberté (Hald, 1952). On doit alors comparer les points expérimentaux (B_i, G_i) aux points théoriques (χ_i^2, G_i), lesquels représentent une droite de pente -0,217 sur un papier de probabilité semi-logarithmique.

- Tracé de points théoriques (χ_i^2 , G_i)

La fonction de distribution cumulée (FDC) de la loi CHI-2 étant:

$$F(\chi_i^2) = 1 - e^{-\chi_i^2/2} \qquad (5.2)$$

par transformation et pour le log (base 10) on obtient:

$$\chi_i^2 = - 4{,}605 \log [1 - F(\chi_i^2)] \qquad (5.3)$$

En posant $G(\chi_i^2) = 1 - F(\chi_i^2)$ qui correspond à la probabilité au dépassement on a:

$$\chi_i^2 = - 4{,}605 \log [G(\chi_i^2)]$$

$$-0{,}217\ \chi_i^2 = \log [G(\chi_i^2)] \qquad (5.4)$$

Si on reporte en abcisse les quantités G_i (échelle log) et en ordonnée χ_i^2 sur un graphique on obtient une droite de pente négative.

Exemple de calcul pour χ_i^2 (avec i = 0 à 4)et G_i

χ_i^2	Log G_i =0-217 χ_i^2	G_i=10 $^{\log G_i}$ (en %)	Coordonnées sur le graphique
0	0	100	(0, 100,00)
1	-0,217	60,67	(1, 60,67)
2	-0,434	36,81	(2, 36,81)
3	-0,651	22,34	(3, 22,34)

- Tracé de points expérimentaux (B_i, G_i)

Les quantités B_i sont d'abord classées par ordre croissant soit:

$$B_i \leq B_{i+1} \leq \ldots B_{N-1} \leq B_N \qquad \text{avec } i = 1 \text{ à } N$$

On associe aux valeurs classées une probabilité empirique F_i (Eq. 4.18 et 4.19). On calcule G_i (en %) soit, $G_i = (1 - F_i)\ 100$. On reporte ensuite sur le graphique les différents couples (B_i, G_i(%)). Si B_i suit une loi CHI-2 les différents points seront situés au voisinage de la droite théorique.

5.2.3 Commentaires et exemples

Un des critères les plus visuellement vérifiables qui légitime l'utilisation du coefficient de corrélation est sans doute la linéarité. Si les

observations x_i, y_i ne sont pas linéaires il n'est pas utile d'effectuer des tests de normalité. On peut en effet conclure immédiatement que r n'est pas le meilleur outil pour mesurer l'intensité de la relation entre x et y. Ainsi dans un étude de Karam et Cescas (1984) on conclue à une relation très significative entre les variables pH et A (constante de Freundlich) obtenus à deux sites différents (figure 5.1). Or l'examen des graphiques permet de jeter un doute sérieux sur la pertinence d'utiliser le coefficient de corrélation et ce pour plusieurs raisons qui peuvent être déduites graphiquement. Outre la non-linéarité on peut aussi observer que les observations sont regroupés en nuages distincts. Ainsi on peut voir que 20 des 23 observations ont des valeurs de pH peu différentes situés entre 6.75 et 8.05 tandis que seulement 3 observations ont de faibles valeurs, d'abcisse $x \simeq 4$.

D'autres aspects liés à la représentativité de l'échantillon peuvent aussi limiter l'interprétation du coefficient de corrélation: l'hétérogénéité de l'échantillon, la présence de points extrêmes. Ces sujets seront développés à la section qui traite de la corrélation fictive. Nous voulons simplement souligner que les conditions d'utilisation de r peuvent aussi être évaluées graphiquement, par l'examen des résidus et qu'il n'est pas toujours nécessaire d'effectuer un test de normalité.

Cependant l'écart à la normalité n'est pas toujours très évident et c'est pourquoi il sera alors utile de la vérifier par un test. À cet effet nous présentons l'exemple suivant.

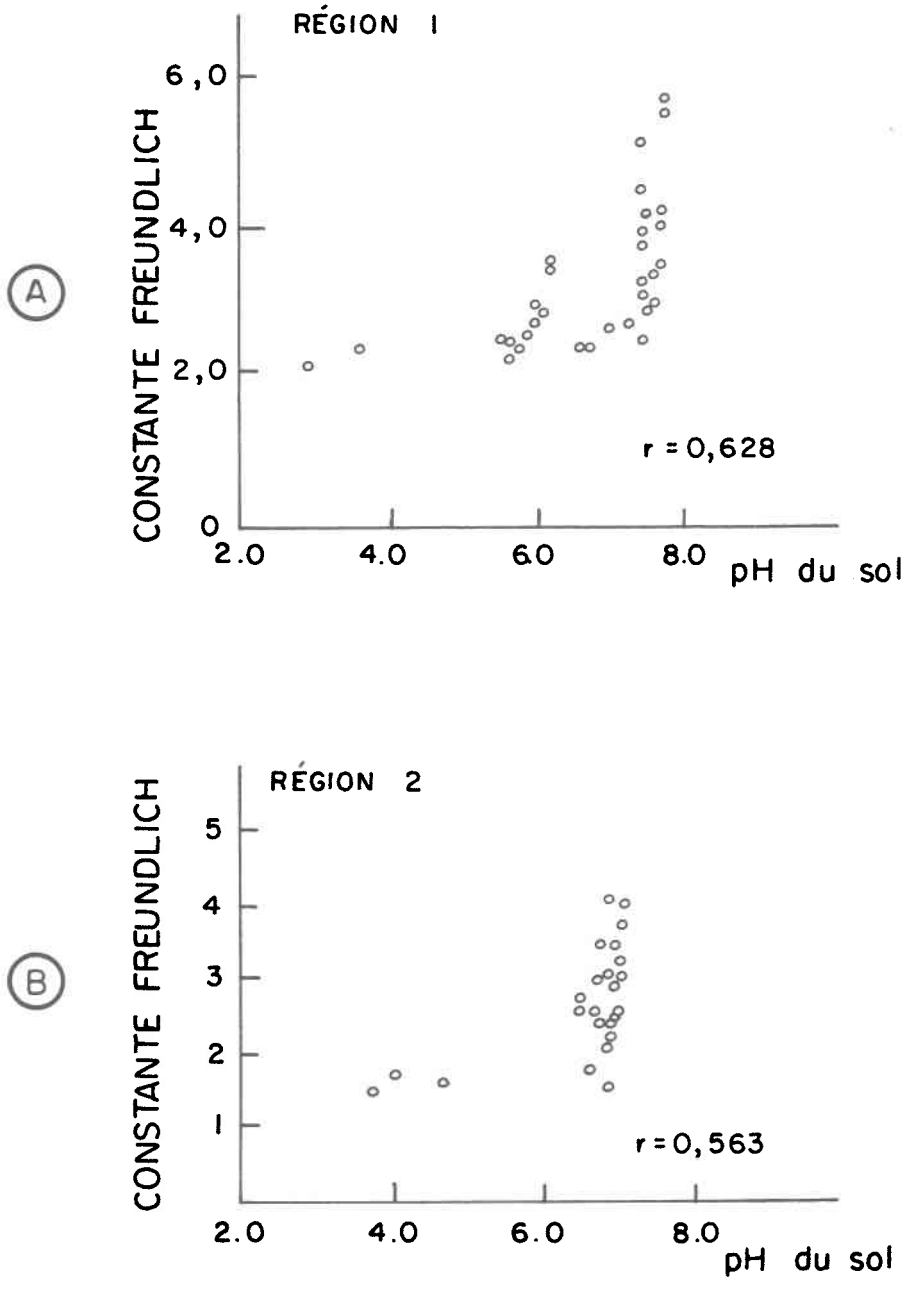


FIGURE 5.1 Utilisation non justifiée du coefficient de corrélation. Echantillon non-linéaire, non-homogène. Distribution jointe non normale. Tiré de Karam et Cescas (1984).

EXEMPLE 5.1

Nous reprenons ici l'exemple 4.6 . Les auteurs Sharpley et al., (1985) concluaient qu'en dépit du faible pouvoir prédictif du modèle les coefficients de corrélation (voir figures 4.15 et 4.16) étaient significatifs (du moins au niveau $\tilde{\alpha} = 0,05$ pour l'Azote). Comme on l'a mentionné les tests sur ρ sont valables si la distribution jointe de X et Y est normale. Cependant ces résultats peuvent être utilisés de manière approximative si la distribution jointe n'est pas trop éloignée de la normalité. Le test de normalité fut appliqué aux échantillons des figures 4.15 et 4.16. Les résultats du test graphique sont présentés aux figures 5.2 A et B. On peut conclure sans équivoque que la distribution jointe est très éloignée de la normalité, ce qui invalide le test de signification des coefficients de corrélation présenté dans cet article.

EXEMPLE 5.2

Bien que le coefficient de corrélation ρ mesure l'intensité d'une relation linéaire entre les variables X et Y il est très souvent employé lorsque la relation est non-linéaire. À la limite on pourra mettre en évidence une relation de forte dépendance entre les variables X et Y et, néanmoins un coefficient de corrélation nul. Un exemple fictif est présenté à la figure 5.3 A. La variable X représente la température de l'eau et la variable Y le taux de survie des oeufs d'une espèce piscicole quelconque. Comme on peut le voir l'échantillon est parfaitement ajusté à un polynôme

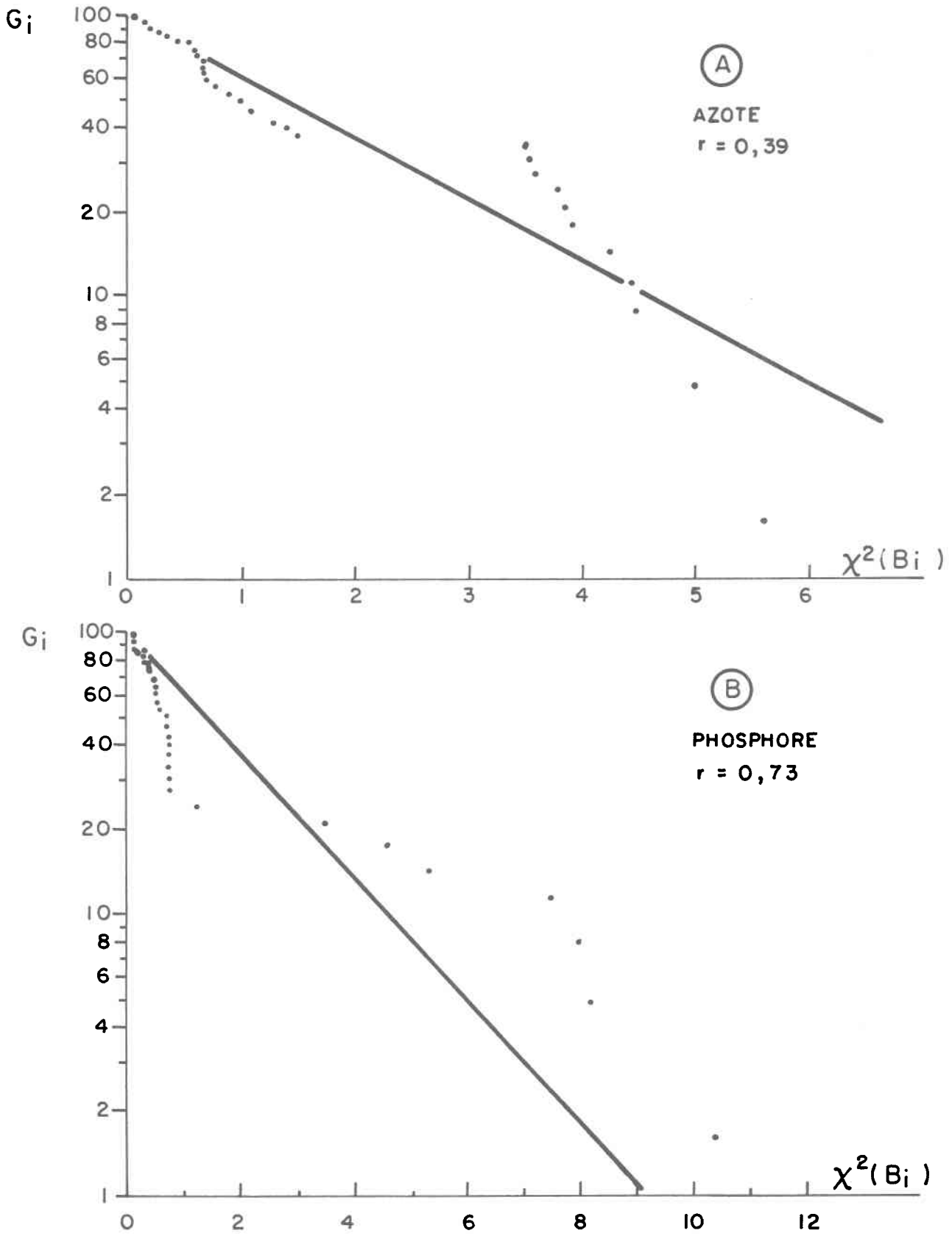


FIGURE 5.2 Test de normalité de la distribution jointe de X et Y. (Exemple 5.1)

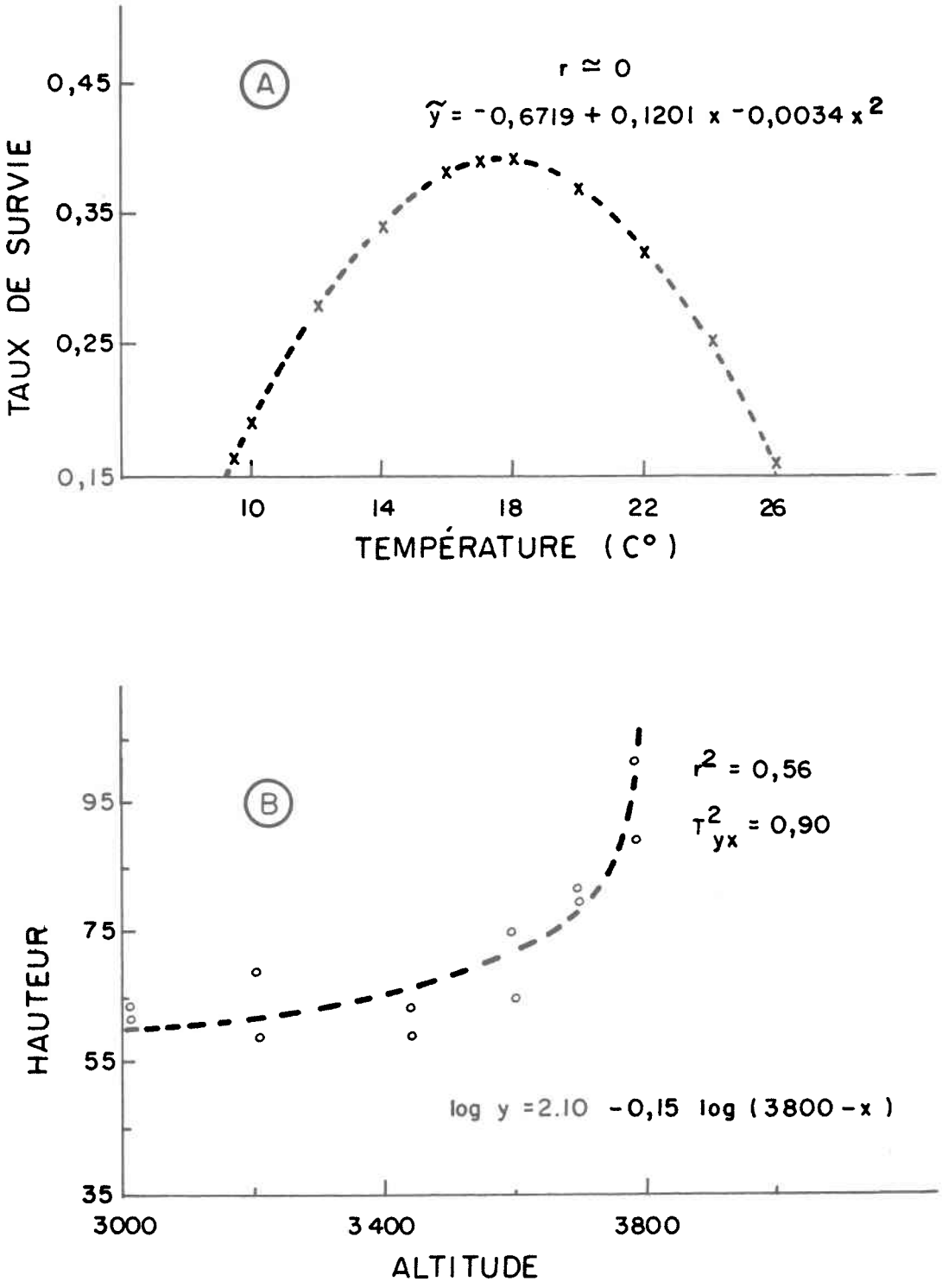


FIGURE 5.3 Relation non-linéaire et coefficient de corrélation. A: Données simulées - exemple 5.2. B: Tiré de Smith (1980) - exemple 5.3

deuxième degré. On a : $y = -0,6719 + 0,1201\ x\ -0,0034x^2$. Néanmoins le coefficient de corrélation est nul.

EXEMPLE 5.3

L'exemple de la figure 5.3 B est tiré d'une étude de Smith (1980). Toutes les relations présentées dans cet article sont non-linéaires. Bien que l'auteur ne suggère aucun modèle on y présente néanmoins les coefficients de corrélation linéaire. Comme on l'a déjà dit, lorsque la distribution jointe est non-normale (et c'est le cas notamment lorsque la relation est curviligne) on devrait utiliser le rapport de corrélation Θ^2_{YX}, défini à la section 3.5. Ce paramètre estimé dans l'échantillon par T^2_{yx} (équation 3.6) est indépendant de la nature de la distribution jointe. Pour cet exemple on pourra comparer le rapport de corrélation au coefficient de corrélation r et au coefficient de détermination $(r)^2$. On a: $r = 0,76$ $r^2 = 0,56$ et $T^2_{yx} = 0,90$. Puisque T^2_{yx} est beaucoup plus élevé que r^2 on peut conclure qu'un modèle non-linéaire serait beaucoup plus adéquat qu'un modèle linéaire et que r est un mauvais estimateur pour apprécier l'intensité de la dépendance entre les variables X et Y.

Un modèle non linéaire fut ajusté à l'échantillon en utilisant les variables log y et log (C-x) avec C = 3800. La comparaison des quantités SS_{r_1} (modèle linéaire) et SS_{r_2} (modèle non-linéaire) permet de vérifier l'intérêt d'un modèle non-linéaire.

(1) $\tilde{y} = -61,94 + 0,04\ x$ $\qquad SS_{r_1} = 870,36$

(2) $\log \tilde{y} = 2,10 - 0,15 \log (3800 - x)$ $\qquad SS_{r_2} = 278,59$

$\tilde{y} = 127,09\ (3800 - x)^{-0,15}$

Le modèle logarithmique est illustré à la figure 5.3 B

EXEMPLE 5.4

Comme on l'a mentionné on peut interpréter sans équivoque la valeur du coefficient de corrélation si la distribution jointe est normale bivariée. Cette condition vérifiée n'est cependant pas synonyme de "coefficient de corrélation significatif". De plus un coefficient de corrélation élevé pour un petit échantillon n'est pas nécessairement significatif et à l'inverse un faible coefficient de corrélation pour un grand échantillon peut être très significatif, car le test (équation 3.8) tient compte du nombre d'observations de l'échantillon.

Une étude présentée dans Marier et al.,(1980) porte sur la relation entre la quantité de magnésium et la dureté de l'eau. On obtient pour cet échantillon (N = 4 observations) un coefficient de corrélation de 0,90. L'auteur conclut à une relation "très fortement significative". Or si l'on compare la statistique t_0 à une valeur critique $t_c = t_{N-2}(\tilde{\alpha})$ on obtient les résultats suivants pour différents niveaux de significations.

$$t_0 = \frac{r}{(1 - r^2 / N - 2)^{\frac{1}{2}}} = 2.92$$

avec $t_2(0,05) = 2.92$, $t_2(0,025) = 4,30$ et $t_2(0,01) = 6,93$

On peut donc conclure pour des niveaux de signification de 0,025 et 0,01 que le coefficient est non-significativement différent de zéro, puisque $t_0 < t_c$. Cependant en considérant un niveau $\tilde{\alpha} = 0,05$ on a, $t_0 = t_c$ et à la limite le coefficient est significatif.

Il n'est donc pas inutile d'indiquer le niveau $\tilde{\alpha}$ pour lequel le coefficient de corrélation est significatif. Car à la limite, quelque soit la valeur de r, on peut toujours choisir un niveau $\tilde{\alpha}$ suffisamment élevé pour être en mesure de conclure à une relation significative mais on augmente aussi le risque d'erreur de type 1. Pour ce cas particulier, il serait plus juste de dire que le coefficient de corrélation est faiblement significatif.

5.3 CORRÉLATION FICTIVE

5.3.1 Généralités

On peut mettre en relation deux variables qui en soi n'ont aucun lien et obtenir néanmoins un coefficient de corrélation élevé ou du moins statistiquement significatif. Nombreux sont les exemples signalés. Citons un exemple de Kendall et Yule (1950), rapporté dans Montgomery et Peck (1982). Des données relatives au nombre de déficients mentaux et au nombre de postes de radio détenus par la population au Royaume-Uni furent recueillies pour 14

années consécutives (1924-1937). Le coefficient de corrélation entre ces variables était pratiquement égal à l'unité r = 0,9921. Un tel coefficient pouvait laisser croire qu'il existait une relation d'interdépendance très étroite entre ces variables, alors qu'en réalité cette association était purement circonstantielle (méthode de plus en plus raffinée pour le dépistage des troubles mentaux et, parallèlement, des progrès technologiques qui contribuaient à rendre les appareils de moins en moins coûteux et dès lors plus accessibles.)

Quelques unes de ces corrélations (non-sense corrélation) sont maintenant des classiques. Par exemple la corrélation élevée entre le nombre de nids de cigognes et le nombre d'enfants dans certaines villes d'Europe (Wallis et Roberts (1956) cité par Tate, 1978).

Il est important cependant de différencier ce qu'il conviendrait d'appeler une corrélation dénuée de sens et ce que l'on désignera sous le terme de corrélation fictive. Dans le premier cas, même si le rapprochement de certaines variables peut être contestable et même fantaisiste il n'en demeure pas moins que ces relations sont statistiquement réelles. Le terme corrélation fictive est ici réservé à une association apparente entre des variables qui en réalité ne sont pas interdépendantes d'un point de vue statistique. Une corrélation sera dite fictive si elle peut être remise en cause d'un point de vue statistique et ce pour diverses raisons. Nous identifions les cas suivants de corrélation fictive:

- corrélation fictive due à l'hétérogénéité de l'échantillon
- corrélation fictive due à un échantillon non représentatif
- corrélation fictive due à l'introduction de variables communes que Kenney (1982) particularise par le terme "spurious self-corrélation".

Dans tous les cas il serait plus rigoureux de distinguer entre corrélation totalement fictive et partiellement fictive. D'une part on peut obtenir un coefficient de corrélation significatif, alors qu'en réalité il n'existe aucune association entre les variables et le coefficient obtenu est alors totalement fictif. D'autre part on pourra mesurer un coefficient beaucoup plus élevé qu'il ne l'est en réalité (notamment si l'échantillon comporte un point extrême), et l'on pourra dire alors que le coefficient est partiellement fictif.

5.3.2 Échantillon non-représentatif.

Comme on l'a mentionné à la section 4.3.2 un point extrême, s'il n'est pas marginal, aura peu d'influence sur l'équation de régression du modèle empirique. Cependant quelque soit sa position par rapport à l'axe principal la valeur du coefficient de corrélation est toujours influencée par ce point et ce, d'autant plus fortement qu'il est éloigné des autres observations, et que la taille de l'échantillon est faible.

Pour apprécier l'influence relative d'un point en fonction de sa position et de la taille de l'échantillon, nous avons simulé l'exemple de la

figure 5.4. On peut observer l'influence des points extrêmes P_1 et P_2 sur la valeur des coefficients de corrélation des échantillon n et N. Soient les échantillons n (9 observations) et N (49 observations). On définit les coefficients r_1, r'_1, r_2 et r'_2 comme suit:

r_1 = coefficient pour n + P_1

r'_1 = coefficient pour N + P_1

r_2 = coefficient pour n + P_2

r'_2 = coefficient pour N + P_2

Les coefficients de corrélation r_0 et r'_0 des échantillons initiaux de taille n et N sont nuls. Si on ajoute P_1 on obtient pour le petit échantillon un coefficient très élevé $r_1 = 0,948$. L'influence de P_1, est beaucoup moins importante pour l'échantillon de taille N : $r'_1 = 0,376$. On peut aussi remarquer la valeur encore plus élevée des coefficients lorsque l'on considère l'observation P_2. On a $r_2 = 0,99$ et $r'_2 = 0,69$. Bien que r_0 et r'_0 soient nuls, les coefficients r_1 et r'_1 sont néanmoins significativement différents de zéro, même au niveau de signification $\tilde{\alpha} = 0,005$

On a pour r_1 : $t_1 = 8,42 > t_c = t_8\ (0,005) = 3,35$

et pour r'_1 : $t'_1 = 2,81 > t_c = t_{48}\ (0,005) = 2,69$

EXEMPLE 5.5

Les exemples suivants permettent d'apprécier l'influence plus ou moins importante d'un ou de quelques points extrêmes sur la valeur de r. Le premier exemple est tiré de Jenkyn (1984). Les variables mises en relation

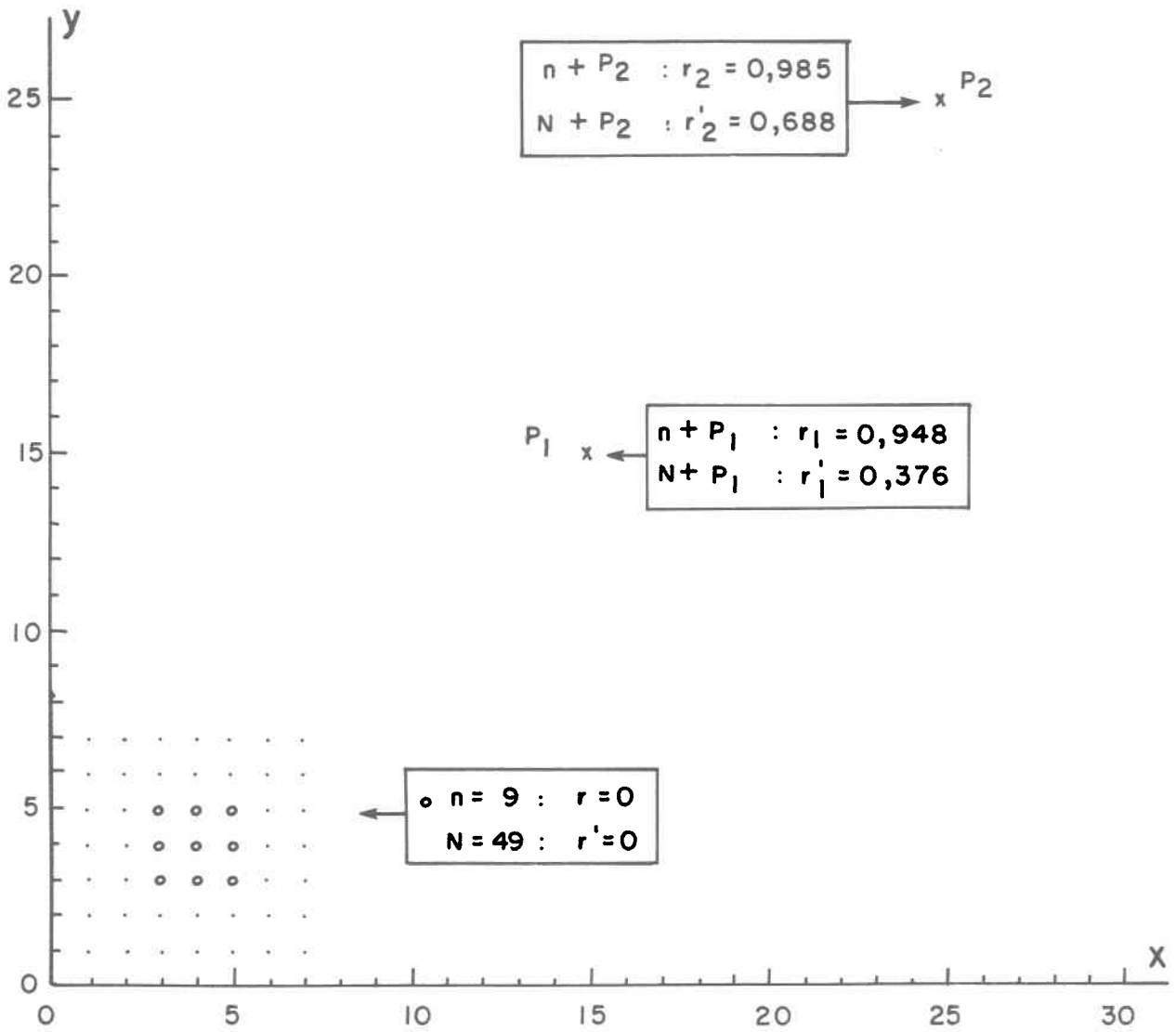


FIGURE 5.4 Influence relative des points extrêmes en fonction de leur position et du nombre d'observations considérés. Données simulées.

sont le nombre de jours requis pour obtenir une température cumulative de 600°C (x) et le taux d'accroissement de la production de grains céréaliers (y) suite à l'application d'un fongicide. Chaque observation représente une année (1969-1979). Cet exemple illustre bien l'influence d'un point extrême sur la valeur du coefficient de corrélation et sur le test de signification. En effet, la prise en considération du point 1 (figure 5.5A) permet d'obtenir un coefficient de corrélation très élevé. En considérant l'échantillon total on a $r = -0,62$. L'élimination de l'observation P_1 donne un coefficient très faible ($r = -0,33$) et non-significatif au seuil de 20% ($\tilde{\alpha} = 0,20$), alors que le coefficient de l'échantillon total est significatif pour $\tilde{\alpha} = 0,05$.

Aucune information n'étant fournie sur ce point extrême nous ne pouvons affirmer que ce point est aberrant ou simplement exceptionnel. Cependant il eut été nécessaire d'obtenir d'autres observations pour $x < 38$ pour être en mesure de vérifier si la relation entre les variables est significative ou non. A partir de l'échantillon dont les auteurs disposent il semble donc très imprudent de conclure à une relation significative. Par ailleurs il est très étonnant que les auteurs considèrent le point P_2 comme possiblement aberrant, en raison de son influence sur la valeur de r. En effet, ce point ne nous semble aucunement extrême et son influence sur la valeur de r est peu importante. En éliminant ce point on obtient en effet un coefficient peu différent $r = -0,69$.

Nous avons pu relever de nombreux exemples de ce type. Nous nous contenterons d'en citer un second. Une étude tirée de Howe et Vande Kerckhove (1981) présentait un coefficient déduit d'un échantillon de 17

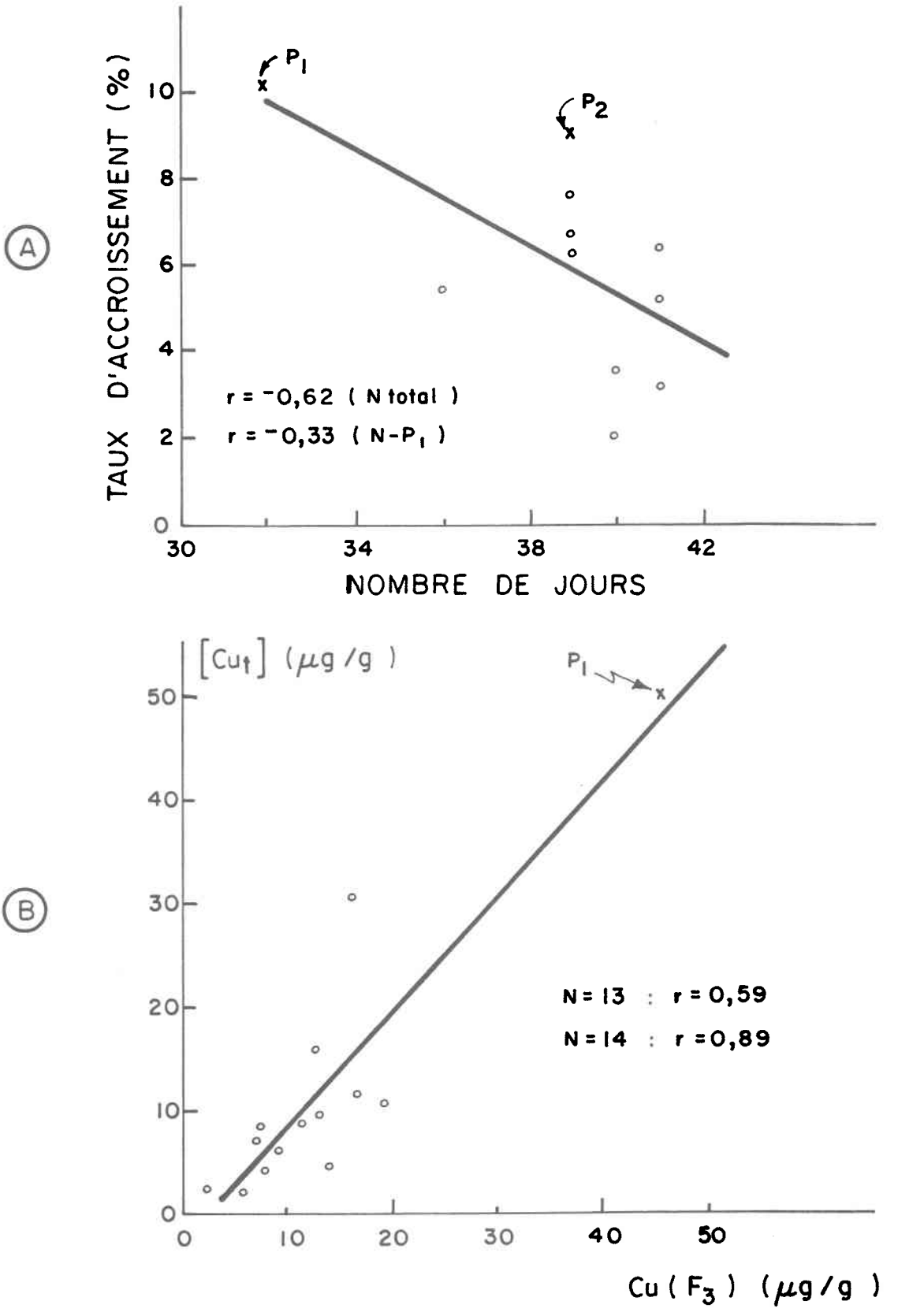


FIGURE 5.5 Influence d'un point extrême sur la valeur du coefficient de corrélation. 5.5 A: d'après Jenkyn (1984). 5.5 B: d'après Tessier et al.(1982).

considération des 15 premières observations ($0 < x_i < 4600$) conduit à un coefficient égal à 0,42 et l'échantillon total où l'on considère deux observations éloignées sur l'axe des x (x = 9000 et x = 10 400) permet de doubler la valeur du coefficient de corrélation (r = 0,80)

EXEMPLE 5.6

La présence d'un point extrême ne conduit pas toujours à l'établissement d'une corrélation fictive. Cependant, dans certains cas, ce point aura une influence beaucoup trop importante sur la valeur de r. Un tel exemple est présenté à la figure 5.5 B. Ce graphique tiré de Tessier et al., (1982) met en relation les teneurs en cuivre (Cu_t) des tiges de Nuphar Variegatum et les concentrations de cuivre liées aux oxydes de fer ($Cu(F_3)$) et de manganèse dans les sédiments superficiels. En considérant l'échantillon total le coefficient de corrélation est très élevé soit r = 0,89. Si on élimine le point extrême, le coefficient demeure significatif mais s'avère néanmoins beaucoup plus faible et r = 0,59.

5.3.3 Échantillon non homogène

Ce qui a été discuté à la section 4.3.1 s'applique également au coefficient de corrélation. En effet, si une relation est significative du seul fait de la mise en commun de deux groupes distincts, de la même façon on obtiendra un coefficient de corrélation significatif, qui en réalité sera totalement fictif.

On pourra très souvent détecter un problème d'hétérogénéité à l'examen graphique de l'échantillon. Le regroupement des observations en nuages de

points distincts, ou encore la présence de points extrêmes, sont souvent l'indice d'une population hétérogène. Les exemples suivants présentent de telles caractéristiques. De plus les informations que nous avons pu obtenir quant à la nature de l'échantillon confirment ce que nous avions observé graphiquement.

EXEMPLE 5.7

Les variables de cet exemple représentent les concentrations de matière organique, azote (N), phosphore (P), potassium (K), calcium (Ca) et magnésium (Mg) contenus dans des déchets miniers (china clay) d'âges différents (16 à 116 ans).

Les valeurs observées sont représentées à la figure 5.6. Selon les auteurs (Roberts et al, 1981) on observe une accumulation significative de ces éléments (sauf dans le cas du potassium) en fonction du temps. Se référant à la figure 5.6 ils affirment (page 157): ".... and it is apparent that organic mater and all nutrients (except potassium) have accumulated as time as passed".

À partir de l'examen graphique il y a lieu de se demander si on peut ainsi conclure à l'accumulation des éléments mentionnées en fonction de l'âge des résidus. En effet, pour tous les éléments significativement corrélés, on peut remarquer la présence de 1 ou 2 points extrêmes qui semblent avoir une influence déterminante sur la pente, et par extension sur la valeur de r. Il est intéressant de vérifier l'allure du nuage de points

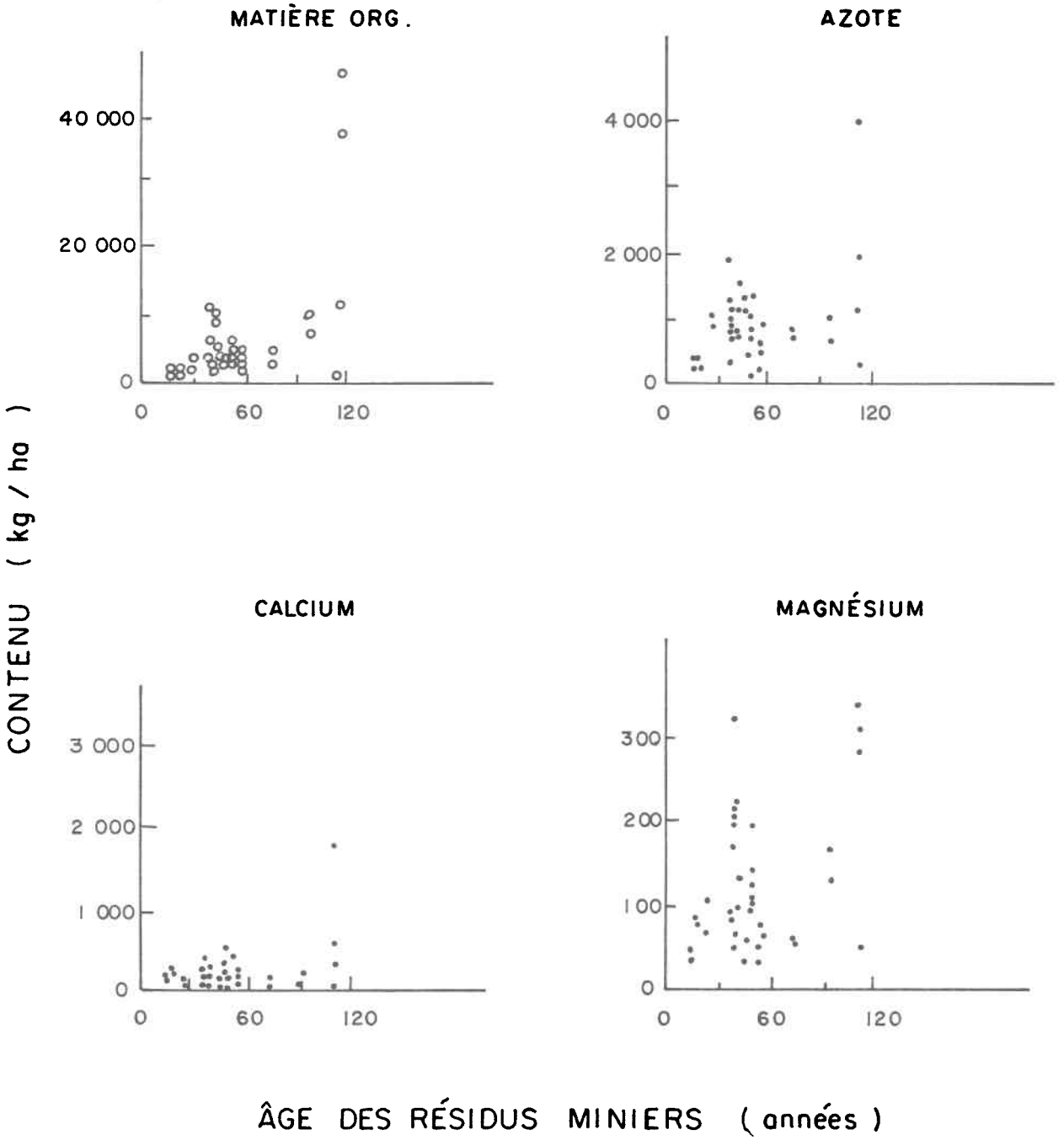


FIGURE 5.6 Illustration d'un échantillon non-homogène (exemple 5.7) D'après Roberts et al., (1981).

lorsqu'on exclue ces valeurs ou cette valeur extrême que l'on retrouve toujours pour x = 116.

Pour être en mesure d'apprécier objectivement l'influence des valeurs extrêmes, les coefficients de corrélation furent calculés en excluant 1, 2, 3 et 4 valeurs qui correspondent aux réplicats d'abcisse x = 116. Les résultats sont présentés au tableau 5.1. En excluant le point le plus extrême sur un total de 38 observations, les coefficients de corrélation des éléments ne sont plus significatifs pour $\tilde{\alpha}$ = 0,05, sauf celui de la matière organique.

En excluant 2 points les coefficients sont pratiquement nuls pour l'azote et le calcium, ou non-significativement différent de zéro ($\tilde{\alpha}$ = 0,01) pour le magnésium et la matière organique.

Tableau 5.1 Valeur du coefficient de corrélation en fonction du nombre d'observations considérées. (Exemple 5.7)

NOMBRE D'OBSERVATIONS	VALEURS DU COEFFICIENT DE CORRÉLATION r			
	Mat. organique	Azote	Calcium	Magnésium
N - 4	0,37	0,13	-0,08	0,06
N - 3	0,17	0,00	-0,14	-0,03
N - 2	0,34	0,07	-0,02	0,16
N - 1	0,50	0,23	0,19	0,30
N TOTAL	0,58	0,42	0,41	0,41

Même si toutes les observations sont retenues, on peut se demander s'il y a vraiment une accumulation, compte tenu de la variance extrêmement importante de l'échantillon au point x = 116. En effet à partir des graphiques on pourra vérifier que pour tous les éléments, il existe toujours au moins une observation où la concentration des éléments est aussi faible que pour x = 16 (plus faible valeur). Or, si on élimine ces 4 réplicats, les coefficients sont pratiquement nuls (calcium, magnésium) ou non-significatifs (azote). Seul le coefficient de la matière organique demeure significatif. Cependant, pour la matière organique, on notera l'influence importante des 4 observations sur la valeur de r: r= 0,37 pour N =34 et r = 0,58 pour N = 38.

- Examen de la nature de l'échantillon

Les remarques précédentes peuvent être faites à partir du seul examen graphique des observations. Le calcul de r ne fait que préciser ce qui peut être déduit de l'observation des données. Cependant il n'est pas inutile de se reférer à la nature de l'échantillon, dont l'hétérogénéité explique souvent une telle distribution (présence d'un point extrême ou d'un groupe de points extrêmes).

La lecture de l'article de Roberts et al.,(1981) fait ressortir que les observations ne représentent pas un échantillon homogène à des époques différentes (qui correspondrait au "vieillissement", ou encore à l'évolution des résidus miniers avec le temps). En fait les données sont mesurés à 8 sites différents, dont les âges varient de 16 à 116 ans. L'âge d'un site

correspond comme on l'a dit au nombre d'années écoulées depuis la fermeture d'une mine. D'après les coordonnées (voir article), les sites représenteraient donc des régions disparates d'un point de vue géographique. On peut donc supposer que l'aspect géographique recouvre de nombreux facteurs (précipitations, température, topographie, etc.) qui influencent de multiples façons la relation entre la concentration des éléments et l'âge des résidus (si cette relation existe vraiment).

On peut supposer qu'il y a une plus grande homogénéité à l'intérieur d'un site, et l'étude âge-concentration pour un même site serait à notre avis plus révélatrice. Quatre des sites étudiés présentent des échantillons d'âges différents (voir tableau 5.2). Les sites 5 et 7 furent les seuls retenus en raison, d'une part, du faible nombre d'observations au site 4 (6 observations) et, d'autre part, du nombre réduit (2) de classes d'âge des sites 4 et 6.

Les résultats pour les sites retenus sont présentés au tableau 5.3. Il est intéressant de noter que pour tous les éléments on obtient maintenant une relation inverse entre l'âge et la concentration. Nous ne pouvons en conclure que la relation est inverse. En effet, plusieurs de ces coefficients sont non-significatifs. Avec peu d'observations (site 7: N = 8 et site 5: N = 10) le coefficient doit être très élevé pour être significati-

	Âge des résidus (années)	Nombre de réplicats
site 1	16	2
site 2	18	2
site 3	26	2
site 4	39	2
	51	4
site 5	40	6
	51	2
	76	2
site 6	42	2
	98	2
site 7	43	2
	48	2
	57	4
site 8	116	4

TABLEAU 5.2 Âge des résidus miniers pour les différents sites de l'échantillon (exemple 5.7)

	Mat. Organique	Azote	Calcium	Magnésium
site 5 (N=10)	- 0,29	- 0,40	- 0,45	- 0,84
site 7 (N=8)	- 0,61	- 0,88	- 0,45	- 0,58

TABLEAU 5.3 Coefficient de corrélation entre l'âge des sites de la concentration des résidus par sous-régions homogènes (exemple 5.7).

vement différent de zéro. Cependant ces résultats font encore plus sérieusement douter des conclusions tirés par les auteurs. Il faut également ajouter que les concentrations initiales à l'âge 0 soit l'année où les mines cessèrent d'être exploitées ne sont pas connues, ce qui jette un doute encore plus sérieux sur les conclusions proposées.

EXEMPLE 5.8

L'exemple suivant est tiré d'une étude (Tessier, 1982) portant notamment sur la relation entre les concentrations de différentes variables physico-chimiques contenues dans les tiges de plantes aquatiques. Quatre lacs comportant plusieurs stations furent échantillonés. Plusieurs des variables étudiées présentaient des coefficients de corrélation sinon élevés du moins significatifs. Cependant on se rend compte à l'examen des graphiques que plusieurs observations sont regroupées en nuages de points distincts correspondants à des sous-bassins hydrographiques. Soit le sous-bassin 1 (lac Routhier) et le sous-bassin 2 (lacs Beauchastel, Bruère et Pelletier).

Le graphique de la figure 5.7A met en relation la teneur en zinc et en manganèse dans les tiges d'une plante. Pour l'échantillon total (N = 21) on obtient un coefficient de corrélation peu élevé r = 0,56 mais cependant très significatif: t_c = -3,02 et $t_{20}(0,05)$ = -1,73. Cependant si l'on étudie chaque groupe homogène, on obtient des résultats très différents. On a en effet pour le sous-bassin 2 (figure 5.7B) un coefficient de corrélation

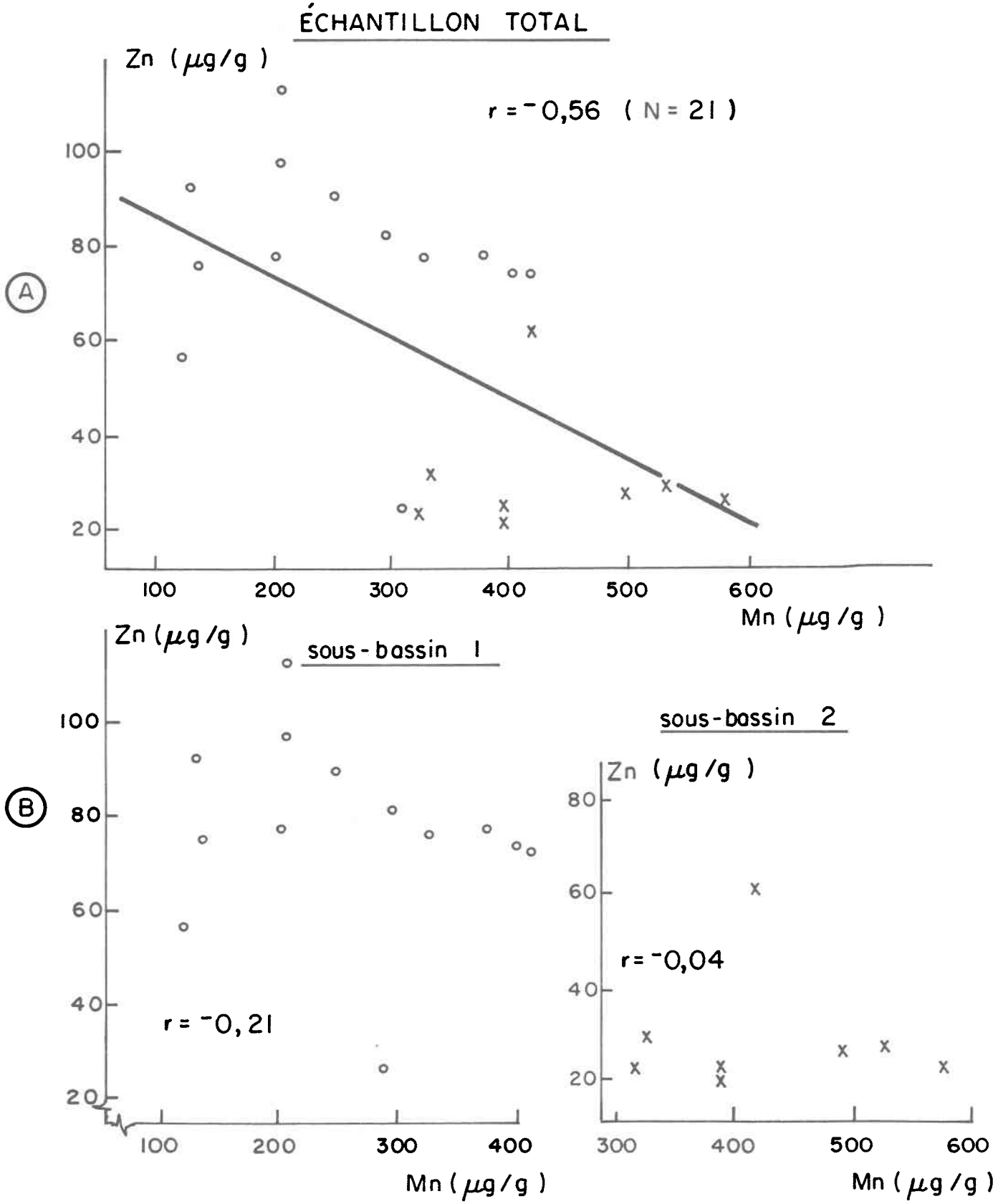


FIGURE 5.7 Étude du coefficient de corrélation, Cu: Mn.
A : Echantillon total. B : Sous-groupes homogènes (exemple 5.8). D'après Tessier et al., (1982).

pratiquement nul (r = 0,04), et pour le sous-bassin 1, comportant 13 observations, un coefficient très faible et non-significatif (r = -0,21). Il semble donc que la forte dépendance statistique entre les variables soit due en grande partie à la mise en commun des sous-bassins 1 et 2.

On peut de la même façon comparer les coefficients de corrélation obtenus entre les concentrations du cuivre (Cu) et de zinc (Zn) pour l'échantillon total (5.8A) et par sous-bassin (5.8B). Il faut noter cependant que le coefficient présenté dans l'article (r = 0,81) exclut les deux observations les plus extrêmes (voir 5.8A). En effet, pour l'échantillon total (N = 22) on obtiendrait un coefficient significatif mais beaucoup plus faible r = 0,41. Or ces deux points semblent avoir été éliminés sans critère objectif mais uniquement parce qu'ils contribuaient à diminuer de façon importante la valeur de r.

En appliquant par ailleurs le test de Grubbs, seule l'observation 2 devrait être rejetée. On remarquera que cette observation (Cu = 50 mg /litre) correspond à l'observation extrême présentée à la figure 5.5 B celle-là même qui contribuait à augmenter de façon très significative la valeur de r. Si une valeur est considérée aberrante dans un cas, il n'y a pas de raison pour qu'elle soit retenue dans un autre cas.

De façon générale il est donc prudent de s'assurer que l'échantillon est homogène, et que le coefficient n'est pas significatif uniquement en raison de la mise en commun de différents groupes. À l'autre extrême, nous avons relevé plusieurs cas où l'échantillon était sans raison valabe divisé en deux groupes.

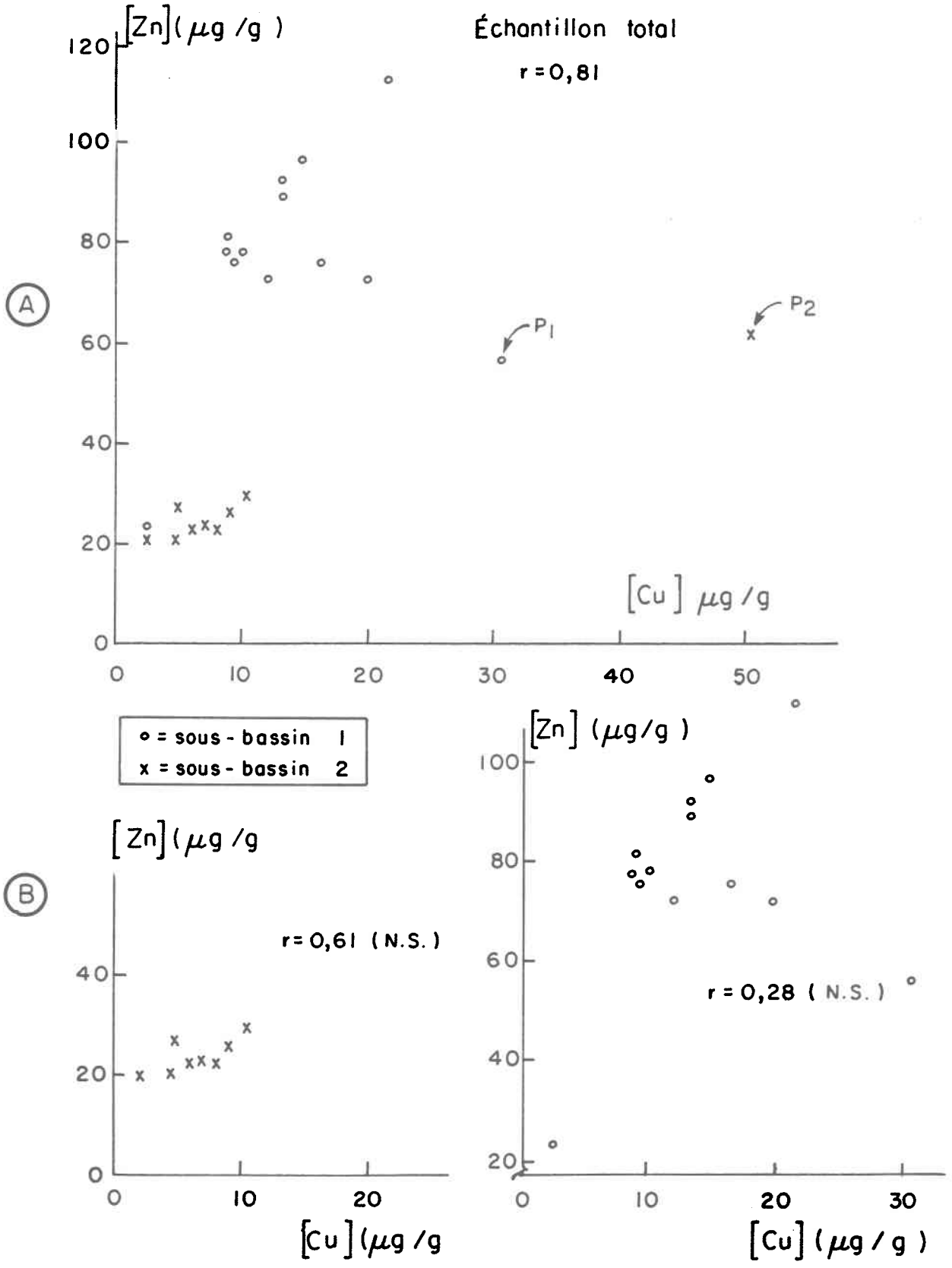


FIGURE 5.8 Étude du coefficient de corrélation, Zn : Cu.
A : Échantillon total. B : Sous-groupes (exemple 5.8)
D'après Tessier et al., (1982)

Une telle subdivision n'est pas justifiable si elle a simplement pour objectif d'augmenter la valeur du coefficient de corrélation comme cela semblait être le cas des exemples examinés.

Certains auteurs présentent parfois deux modèles pour lesquels à une même valeur de x correspondent deux estimateurs $\tilde{y}$ différents, ce qui apparaît assez hasardeux comme procédure. Pour les deux exemples présentés aux figures 5.9 A et B, certaines observations sont communes aux deux droites de régressions estimées, et aux deux coefficients de corrélation.

5.3.4 Emploi de variables communes

De façon générale si deux ou plusieurs variables présentent entre elles un coefficient de corrélation non significativement différents de 0 (ou faible), des fonctions (quotient, produit, somme ou différence) de ces mêmes variables ayant au moins une variable en commun auront un coefficient de corrélation non-nul ou plus élevé. Prenons un cas très simple où l'on considère d'une part deux variables x_1 et x_2 et d'autre part une variable x_3 qui représente une fonction additive des deux premières soit $x_3 = x_1 + x_2$. En supposant un coefficient nul entre x_1 et x_2 ($r_{x_1x_2} = 0$), on obtiendra sous certaines conditions un coefficient non nul entre les variables x_1 et x_3, ou entre x_2 et x_3 soit, $r_{x_2x_3} \neq 0$ et $r_{x_1x_3} \neq 0$

Pearson (1897) fut le premier à mettre en évidence des corrélations fictives entre des produits de variables. D'autres auteurs ont repris les

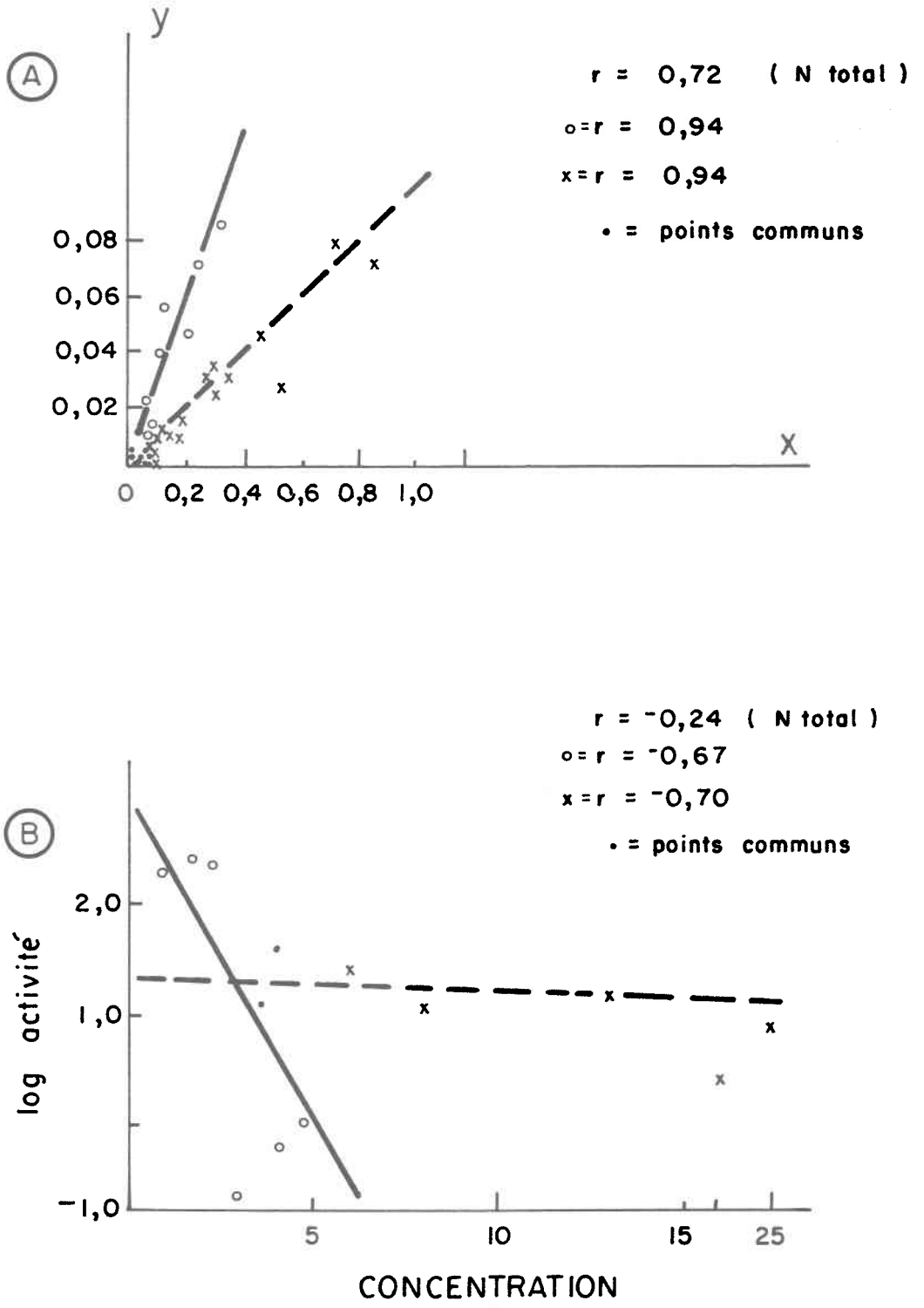


FIGURE 5.9 Subdivision de l'échantillon pour augmenter la valeur du coefficient de corrélation. A: Tiré de Watson (1979) B: Tiré de Francko (1984)

travaux de Pearson et présentent pour différents domaines les abus auxquels l'emploi des variables communes peuvent conduire. Citons notamment les travaux de Chayes (1949), Benson (1965), et plus récemment de Kenney (1982), à qui l'on doit le terme de "spurious self-correlation".

Lorsque des fonctions de variable sont utilisées, on peut vérifier que la valeur du coefficient de corrélation qui en résulte dépend du coefficient de variation des variables originales, lorsque les fonctions se définissent à partir de produits ou de quotients de variables. Lorsque les fonctions sont des sommes et des différences le coefficient dépend uniquement des écart-types des variables originales. Nous présentons quelques éléments théoriques relatifs à ces deux cas, qui sont considérés séparément dans les sections suivantes.

5.3.4.1 Coefficient de corrélation pour des sommes et des différences

Nous considérons ici les variables originales x et y et le coefficient qui résulterait de la mise en commun de la variable x pour une fonction additive soit :

$$u = x$$

$$v = x + y$$

Pour une fonction additive (ou une fonction différence) on peut vérifier que le coefficient de corrélation entre u et v (r_{uv}) dépend de la valeur des écarts-types des variables x et y soit S_x et S_y. En utilisant la

définition du coefficient de corrélation (Eq. 3.2) et en exprimant u et v en fonction de x et y (voir Appendice 6) on arrive à l'expression simplifiée suivante.

$$r_{uv} = \frac{S_x + r_{xy}\, S_y}{(S_x^2 + S_y^2 + 2r_{xy}\, S_x S_y)^{\frac{1}{2}}} \qquad (5.5)$$

Nous présentons dans ce qui suit trois cas extrêmes qui permettent de faire ressortir que la valeur de r_{uv} dépend de S_x et S_y.

- cas particuliers

On considère ici 3 cas extrêmes

(1) $S_y \gg S_x$ soit $S_x/S_y \to 0$

(2) $S_x \gg S_y$ soit $S_y/S_x \to 0$

(3) $S_y = S_x$

(1) En reprenant l'équation 5.5 et en divisant le numérateur et le dénominateur par S_y on obtient:

$$r_{uv} = \frac{S_x/S_y + r_{xy}}{\left[(S_x/S_y)^2 + 1 + 2\, r_{xy}\, (S_x/S_y)\right]^{\frac{1}{2}}}$$

si $S_x/S_y \to 0$ alors $r_{uv} \to r_{xy}$

(2) En divisant le numérateur et le dénominateur par S_x on a d'après 5.5:

$$r_{uv} = \frac{1 + r_{xy} \quad (S_y/S_x)}{\left[1 + (S_y/S_x)^2 + 2\, r_{xy} \quad (S_y/S_x)\right]^{\frac{1}{2}}}$$

$$\text{si } S_y/S_x \to 0 \qquad \text{alors } r_{uv} \to 1$$

(3) On peut aussi envisager le cas où $S_x = S_y$. D'après 5.5 on aura:

$$r_{uv} = \frac{1 + r_{xy}}{\left[2\ (1 + r_{xy})\right]^{\frac{1}{2}}}$$

Pour le cas particulier où $r_{xy} = 0$ alors $r_{uv} = 0{,}71$.

En résumé, si l'écart-type de y (S_y) est beaucoup plus élevé que l'écart-type de la variable commune x (S_x), le coefficient qui dérive d'une fonction de variables (r_{uv}) demeure à peu près égal au coefficient obtenu pour les variables originales. À l'autre extrême, si l'écart-type de la variable commune est beaucoup plus élevé que l'écart-type de y, on aurait pour les variables u et v un coefficient proche de l'unité et ce quelque soit la valeur du coefficient entre les variables originales. De plus, si r_{xy} est nul et que les écarts-types des variables originales sont égaux, le coefficient entre u et v est une constante.

- formules générales

En utilisant la procédure déjà présentée on peut dériver une formulation générale du coefficient qu'on obtiendrait pour deux fonctions quelconque (sommes et différences). Ces formules sont tirés de Kenney (1982). Soit les variables u et v définies par:

$$u = \pm x_1 \pm x_2$$
$$v = \pm x_3 \pm x_4$$

et S_1, S_2, S_3, S_4 qui sont les écarts-type des variables x_1, x_2, x_3 et x_4
Le coefficient de corrélation général entre u et v est alors:

$$r_{uv} = \frac{\pm r_{13} \; S_1S_3 \pm r_{14} \; S_1S_4 \pm r_{23} \; S_2S_3 \pm r_{24} \; S_2S_4}{\sqrt{(S_1^2 + S_2^2 \pm 2 \, r_{12} \; S_1S_2) \quad (S_3^2 + S_4^2 \pm 2 \, r_{34} \; S_3 \, S_4)}} \qquad (5.6)$$

Pour cette expression le signe des différents r dépend du produit des signes devant les variables x_1, x_2, x_3 et x_4. Par exemple si $u = x_1 + x_2$ et $v = x_3 - x_4$, on aurait alors: $+r_{13}$, $-r_{14}$, $+r_{23}$ et $-r_{24}$

exemple numérique:

En supposant la variable commune x et une fonction différence.

$$u = x$$

$$v = x - y$$

on peut dériver de l'équation 5.6 le coefficient r_{uv} suivant:

$$r_{uv} = \frac{+r_{xx} \; S_x \; S_x - r_{xy} \; S_x \; S_y}{\sqrt{(S_x^2) \quad (S_x^2 + S_y^2 - 2 \; r_{xy} \; S_x \; S_y)}}$$

comme $r_{xx} = 1$

$$r_{uv} = \frac{S_x^2 - r_{xy} \; S_x \; S_y}{S_x \sqrt{(S_x^2 + S_y^2 - 2 \; r_{xy} \; S_x \; S_y)}}$$

En donnant des valeurs relatives à S_x et S_y on peut vérifier que cette formule conduit aux mêmes résultats que les cas particuliers présentés précédemment. Si on suppose que $r_{xy} = 0$, l'équation se simplifie et on a:

$$r_{uv} = \frac{S_x}{\sqrt{(S_x^2 + S_y^2)}}$$

Si $S_y = 10\ S_x$ $\qquad r_{uv} = \dfrac{S_x}{(S_x^2 + 100\ S_x^2)^{\frac{1}{2}}} = 0{,}0995 \simeq 0$

Si $S_x = 10\ S_y$ $\qquad r_{uv} = \dfrac{10\ S_y}{(100\ S_y^2 + S_y^2)^{\frac{1}{2}}} = 0{,}995 \simeq 1$

Si $S_x = S_y$ $\qquad r_{uv} = \dfrac{S_x}{(2S_x^2)^{\frac{1}{2}}} = 0{,}71 = \text{constante}$

5.3.4.2 Coefficient de corrélation pour des produits et des quotients

Lorsque les fonctions sont des produits ou des quotients on peut vérifier (Pearson 1897, Reed 1921) que le coefficient résultant dépend du coefficient de variation C, qui représente le rapport de l'écart-type à la moyenne. Soit pour la variable x, $C_x = S_x / \bar{x}$. La formulation générale du coefficient impliquant un produit de variables a été défini par Pearson (1897). En posant:

$$u = x_1\ x_2^E$$

$$v = x_3\ x_4^E$$

E = + 1 si on considère des produits ($u = x_1x_2$ et $v = x_3x_4$)

E = - 1 si on considère des quotients ($u = x_1/x_2$ et x_3 / x_4)

on a :

$$r_{uv} = \frac{r_{13}\ C_1C_3 + E\ (r_{14}\ C_1C_4 + r_{23}\ C_2C_3) + r_{24}\ C_2C_4}{\left(C_1^2 + C_2^2 + 2Er_{12}\ C_1C_2\right)^{\frac{1}{2}} \left(C_3^2 + C_4^2 + 2Er_{34}\ C_3C_4\right)^{\frac{1}{2}}} \qquad (5.7)$$

On trouvera dans Benson (1965) la présentation du coefficient r_{uv} pour les différents cas possibles (variables communes en abscisse, ou en ordonnée, variables communes sur les deux axes etc...), ainsi que les formules qui en dérivent. Ces formules sont asymptotiques donc sont d'autant plus justes que la taille de l'échantillon est grande (Bobée, 1983). En remplacant l'écart-type par le coefficient de variation on obtient la même estimation de r_{uv} pour les cas extrêmes présentés pour les sommes et les différences. Supposons par exemple :

$$u = x$$

$$v = y/x$$

on aurait :

$$r_{uv} = \frac{r_{xy}\, C_y - C_x}{(C_x^2 + C_y^2 - 2\, r_{xy}\, C_x\, C_y)^{\frac{1}{2}}} \qquad (5.8)$$

Pour $r_{xy} = 0$ cette équation est réduite à

$$r_{uv} = \frac{-C_x}{(C_x^2 + C_y^2)^{\frac{1}{2}}} = \frac{-1}{(1 + (C_y / C_x)^2)^{\frac{1}{2}}}$$

En posant:

(1) $C_y = 10\, C_x$ $(C_y > C_x)$ $\qquad r_{uv} = 0{,}0995 \approx 0$

(2) $C_x = 10\, C_y$ $(C_x > C_y)$ $\qquad r_{uv} = 0{,}9950 \approx 1$

(3) $C_x = C_y$ $\qquad\qquad$ $r_{uv} = 0,71$ = constante

Comme on l'a dit on peut obtenir un coefficient très élevé entre des fonctions de variables ayant une variable commune, même si le coefficient entre les variables originales est faible ou nul. Sous certaines conditions il est aussi possible d'obtenir pour des fonctions un coefficient plus faible que pour les variables originales. Cela se présente lorsque le coefficient des variables originales est très élevé. Comme on s'en doute, ce type de manipulation est beaucoup moins fréquent. C'est pourquoi nous nous contenterons de présenter un seul cas limite à titre indicatif. On utilise les mêmes variables que précédemment, soit x et y, et $u = x$ et $v = y/x$. Si $r_{xy} = 1$ et que $C_x = C_y$ la formule 5.8 devient:

$$r_{uv} = \frac{C_y + C_x}{(C_x^2 + C_y^2 - 2C_x C_y)^{\frac{1}{2}}}$$

et puisque $C_x = C_y$ on en déduit $\quad r_{uv} = 0$

5.3.4.3. Commentaires et exemples

En supposant les variables $u = x_1/x_2$ et $v = x_3/x_2$ on peut toujours estimer le coefficient de corrélation entre u et v (soit r_{uv}) et prédire la quantité v à partir de u (on suppose $v = f(u)$). Ces procédés sont tout à fait légitimes, bien que peu utiles. Cependant, même si r_{uv} est très élevé, on ne peut rien conclure quant à la relation entre x_1 et x_2 ou entre x_2 et x_3. De même aucune de ces variables ne peuvent être prédites individuel-

lement à partir de la relation globale déduite entre u et v. En effet, même dans le cas où l'estimation de x_3/x_2 est connue, cela n'implique pas la connaissance individuelle des termes du rapport. Et c'est justement ce procédé (déduire x_3 à partir de x_1 et x_2 à partir de x_1) qui est dangereux, puisque rien ne permet d'affirmer que les coefficients entre ces variables sont significatifs.

Très souvent les fonctions de variables impliquent non pas les variables originales, mais des variables transformées (logarithmiques par exemple). Ces transformations conduisent dans certains cas à des coefficients encore plus élevés. Cette double manipulation obscurcit davantage l'allure de la fonction originale. Pour certaines transformations les formules présentées aux sections précédentes pourront être utilisées directement. Par exemple, si $u = \log x_1$ et $v = \log (x_2/x_1)$, la formule utilisée pour une fonction différence est applicable, puisque $\log (x_2/x_1) = \log x_2 - \log x_1$.

De nombreux cas de corrélations fictives impliquant une variable commune ont été relevés dans la littérature. À cet effet on consultera les articles de Benson (1965), et particulièrement un article de Kenney (1982) dont plusieurs exemples concernent le domaine des sciences de l'eau. Entre autres on retiendra les exemples de corrélations fictives de certains modèles d'eutrophisation proposés par Vollenweider (1976). De même Bobée (1983) note des abus fréquents dans l'étude de la relation entre la concentration

d'un élément chimique (C) et le débit massique qui est en fait la concentration divisée par le débit (C/Q).

Quelques exemples sont présentés dans ce qui suit. Ils constituent tous des quotients de variables qui, d'après les publications consultées, semblent être les cas les plus fréquents.

Exemple 5.9

Cet exemple cité par Kenney (1982) est tiré d'un publication de Morgan (1980). L'auteur étudie le taux d'assimilation d'épiphyte (A) consommé par des crevettes en fonction de leur poids (P). Le coefficient de corrélation présenté est significatif r = 0,832. Cependant ce coefficient n'est pas calculé en fonction des variables A et P, mais pour les variables A/P et P. On met donc en relation une variable et une fonction dont l'élément commun est le poids. Comme les observations originales sont également présentées sur un graphique, nous avons pu déduire le coefficient entre le poids et le taux d'assimilation. Ce dernier est non significatif au seuil de 20% r_{PA} = 0,268. Les observations qui ont servi ou calcul sont présentées à la figure 5.10A. Du modèle obtenu (voir figure 5.10 B) pour les variables A/P et P l'auteur en déduit un modèle pour les variables A et P.

on pose $A\tilde{/}P = a + bP$

et on déduit $A\tilde{/}P \cdot P_0 = A_0$

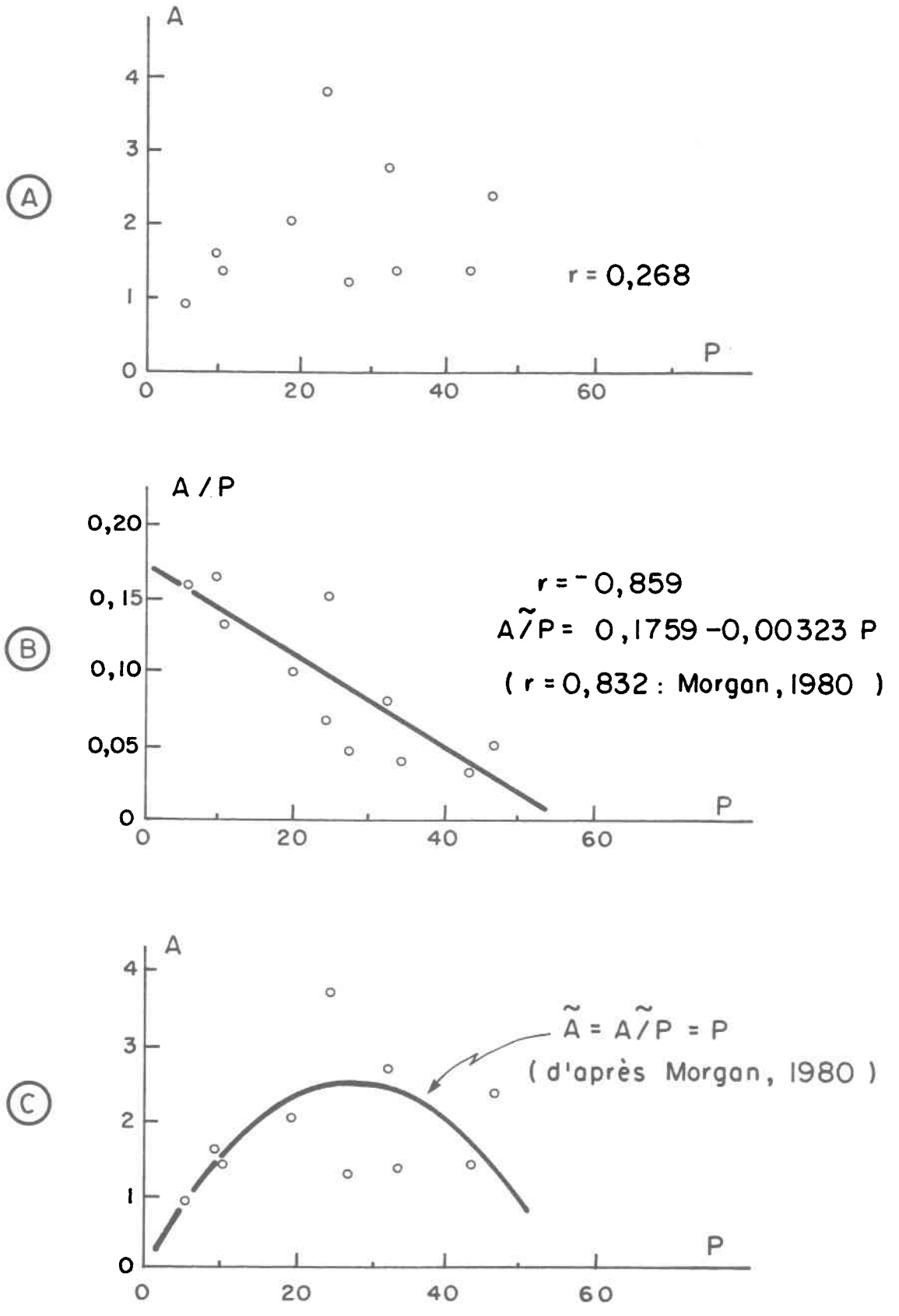


FIGURE 5.10 Corrélation fictive
A : observations originales (x, y). B : modèle déduit des variables y/x et x. C : modèle transposé après transformations. (exemple 5.9)

où A_0 est l'estimation du taux d'assimilation pour $P = P_0$. Tel que suggéré dans l'article, ce modèle fut construit point par point, et il est représenté à la figure 5.10 C.

On a simulé un cas encore plus extrême pour mieux faire ressortir ce genre de manipulation. Au lieu d'un coefficient non-significatif supposons un coefficient de corrélation nul entre les variables x et y (figure 5.11A). En utilisant les variables $v = x/y$ et $u = x$ on aurait cette fois-ci $r_{uv} = -0,92$ et on pourrait conclure, comme plusieurs auteurs le font, qu'il existe une relation significative entre nos variables de départ. Le modèle résultant celui illustré à la figure 5.11 A. De la même façon que pour l'article précédent, la parabole fut obtenue à partir de la relation de u/v. On a:

$\tilde{v} = a + bu$ comme $v = y/x$ et $u = x$

on en déduit $y\tilde{/}x = a + bx$

et pour y_i $y_i = y\tilde{/}x \cdot x_i$

Exemple 5.10

Le second exemple est présenté dans Limnology et Oceanography. On y étudie les quantités de ARN mesurées dans diverses espèces de zooplancton dont le poids diffère (P). Toutes les relations présentées sont significatives. Les auteurs Bamstedt et Skjoldal (1980) diront: "There was commonly a negative relationship between RNA concentration and individual dry weight (W). The correlation was usually highly significant and the regression most often explained 60% of the total intraspecific variation in RNA".

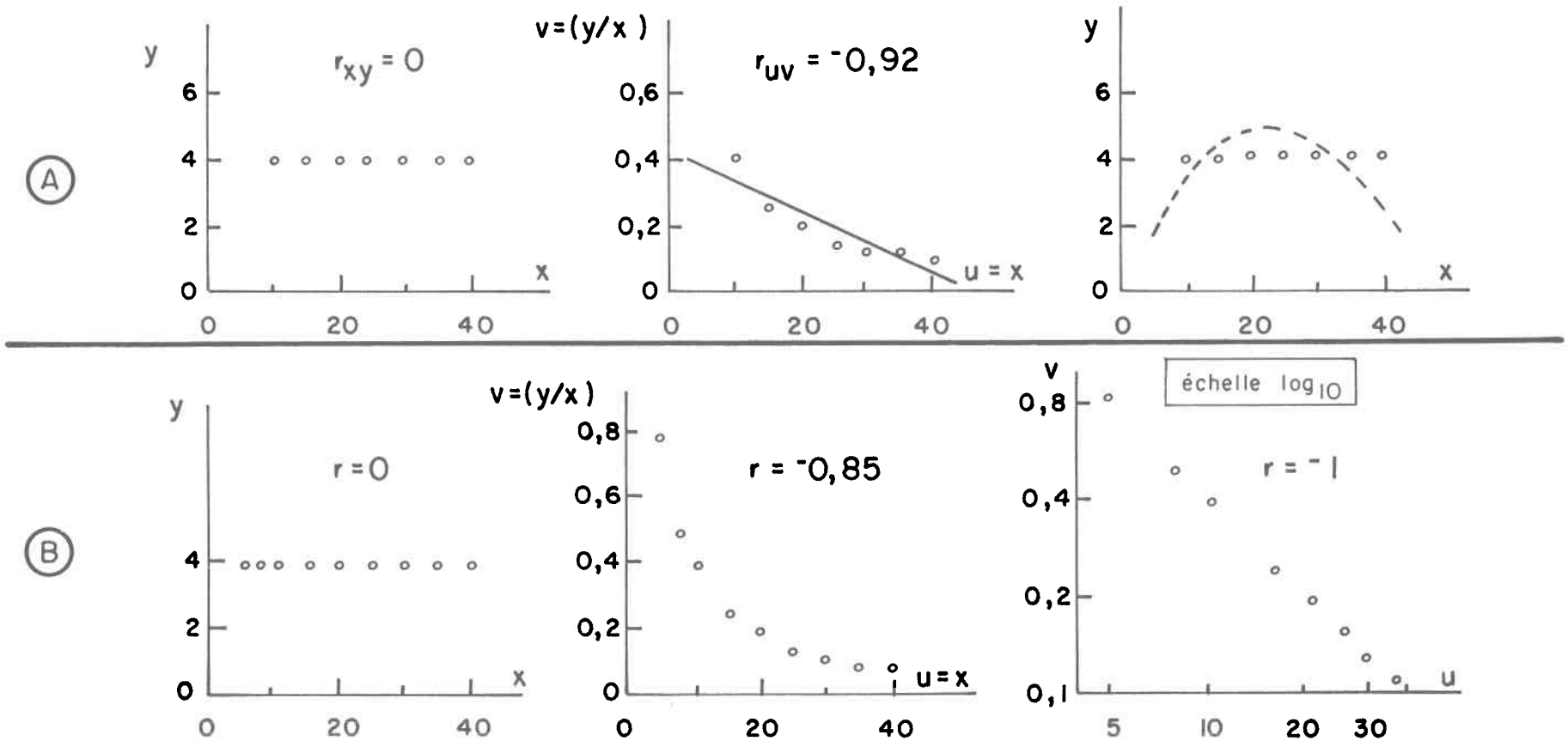


FIGURE 5.11 Corrélation fictive (Données simulées)
A : variables originales, fonctions de variables et modèle associé.
B : variables originales, fonctions de variables et log des fonctions de variables.

L'examen des graphiques présentés dans l'article permet de voir que la relation implique RNA/P avec P et non pas la quantité de RNA et P. D'autre part on a aussi utilisé le log de ces variables soit log (RNA/P) et log P. À la suite de ces deux manipulations, on conceptualise très mal les variables originales. Puisque les données des variables RNA et P n'étaient pas présentées, on n'a pu vérifier si le coefficient était totalement fictif.

EXEMPLE 5.11

L'article de Small et al., (1974) présente ce même type de manipulation. On cherche ici à vérifier la relation entre le taux d'excrétion du zinc (zinc/heure) en fonction du poids des crustacés. Si l'on retient le coefficient de l'article, on conviendra qu'il s'avère très significatif ($r = -0,87$). Cependant l'observation du graphique 5.12 A permet de vérifier que les variables mises en relation sont en fait: zinc/heure/poids en fonction du poids. De plus, non seulement a-t-on utilisé des variables comportant un élément commun mais encore le coefficient présenté est calculé à partir du logarithme (base 10) de ces mêmes variables. Nous avons estimé les coefficients de corrélation qu'on obtient en fonction des variables utilisées. De plus, nous avons aussi tenu compte du fait que dans les résultats de l'article une observation fut éliminée. Cette dernière fut jugée aberrante de manière arbitraire. Comme on peut l'observer l'élimination de cette donnée contribue aussi à augmenter la valeur du coefficient de corrélation (tableau 5.4).

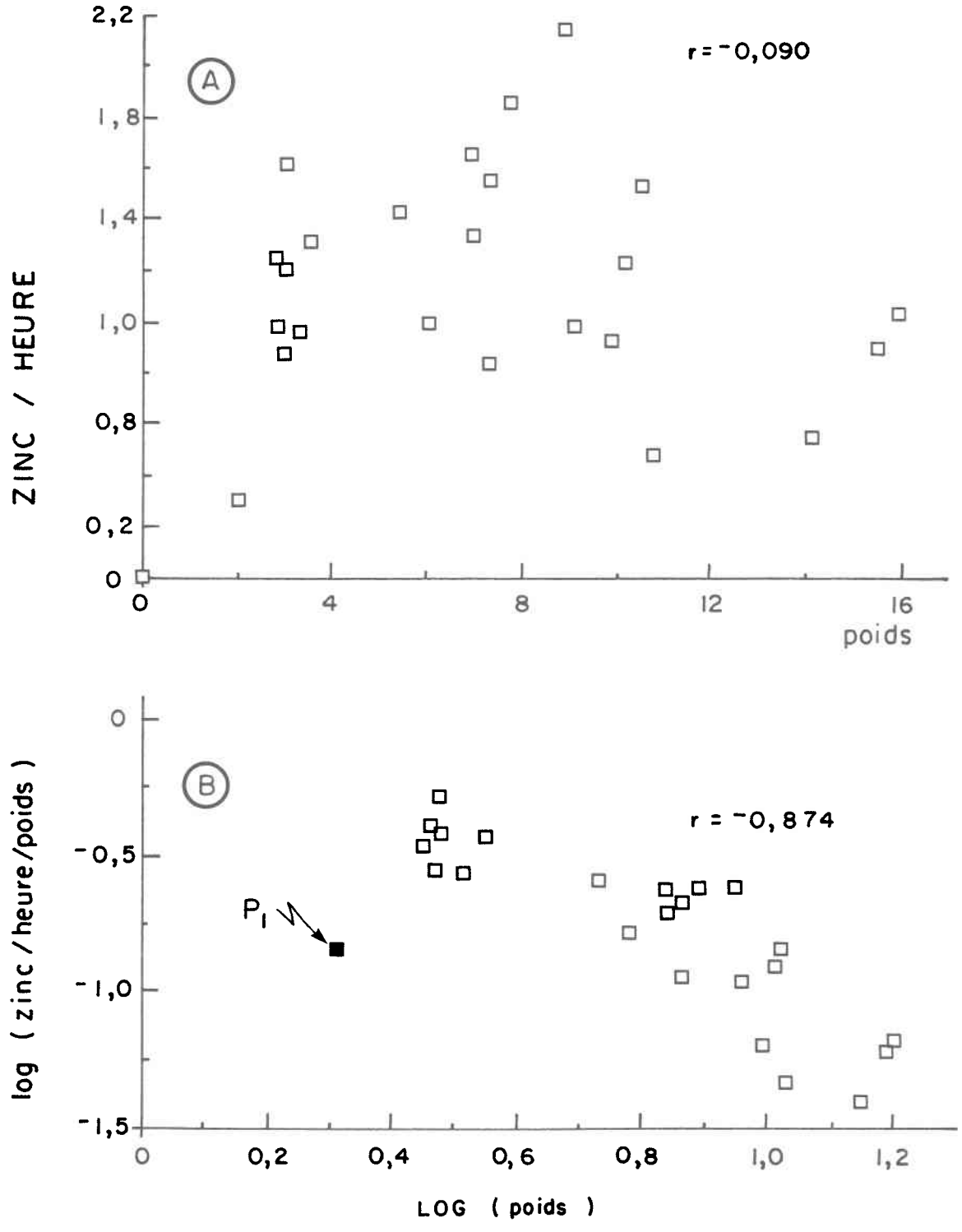


FIGURE 5.12 Examen du coefficient de corrélation
A : variables originales. B : logarithme de fonctions de variables (exemple 5.11). Données tirées de Small et al., (1974).

TABLEAU 5.4 Coefficient de corrélation en fonction des variables utilisées. (Données de l'exemple 5.11)

Variables utilisées			Coefficient de corrélation	
			N = 24	N = 23
Zn/hre	et	P	-0,0900	-0,2358
Zn/hre/P	et	P	-0,6989	-0,7529
log(Zn/hre/P)	et	log P	-0,7885	-0,8740

Les variables non-transformées et les variables de l'article sont présentées à la figure 5.12. Encore une fois on a simulé un exemple théorique encore plus extrême que les deux exemples précédents. On a d'une part un échantillon de variables x et y dont le coefficient de corrélation est nul $r_{xy} = 0$ (figure 5.11 B). En utilisant les variables u = x et v = y/x, on obtient un coefficient très significatif (r = -0,85), et le coefficient entre u' = log x et v est égal à l'unité $r_{uv} = -1$. On comprend alors comment il est erroné d'en conclure que les variables x et y sont parfaitement dépendantes.

Pour mettre en évidence les conclusions qui peuvent résulter de la manipulation des données originales, nous proposons cet exemple farfelu. Supposons un échantillon de 8 familles dont les salaires (S) s'échelonnent entre 10,000$ et 50,000$ dollars. On dénombre un réfrigérateur (R) par ménage, ce qui conduit à un coefficient nul entre les variables S et R. En utilisant cette fois les variables S/R en fonction de R on obtiendrait un

coefficent de corrélation négatif. Pourrait-on en conclure que plus on est riche et moins on possède de réfrigérateurs ?

5.4 COEFFICIENT DE CORRÉLATION ET DE DÉTERMINATION DES VARIABLES TRANSFORMÉES

Lorsque les hypothèses de la régression ne sont pas respectées pour les variables originales, la transformation des variables peut s'avérer une technique très utile pour utiliser certains des résultats de la régression. Cependant le coefficient de corrélation qui est calculé à partir des variables transformées ne reflète pas l'interdépendance entre les variables originales. Nombreux sont les exemples de telles utilisations du coefficient de corrélation et dans certains cas on pourrait là aussi parler de corrélation fictive, ou du moins d'utilisation abusive du coefficient de corrélation.

EXEMPLE 5.12

L'exemple qui suit illustre assez bien ce genre d'extrapolation. Les observations de la figure 5.13 (encadré) tirées d'un article de Pleasants (1980) représentent les observations en valeurs logarithmiques dont le coefficient est $r' = 0,70$. L'auteur en conclue qu'il y a une relation très significative entre le nombre de visites effectués par divers insectes pollinisation) et l'abondance d'une espèce florale. La renconstitution des variables originales (figure 5.13) permet de constater que cette dépendance est très faible comme en fait foi la valeur du coefficient de corrélation $r = 0,29$.

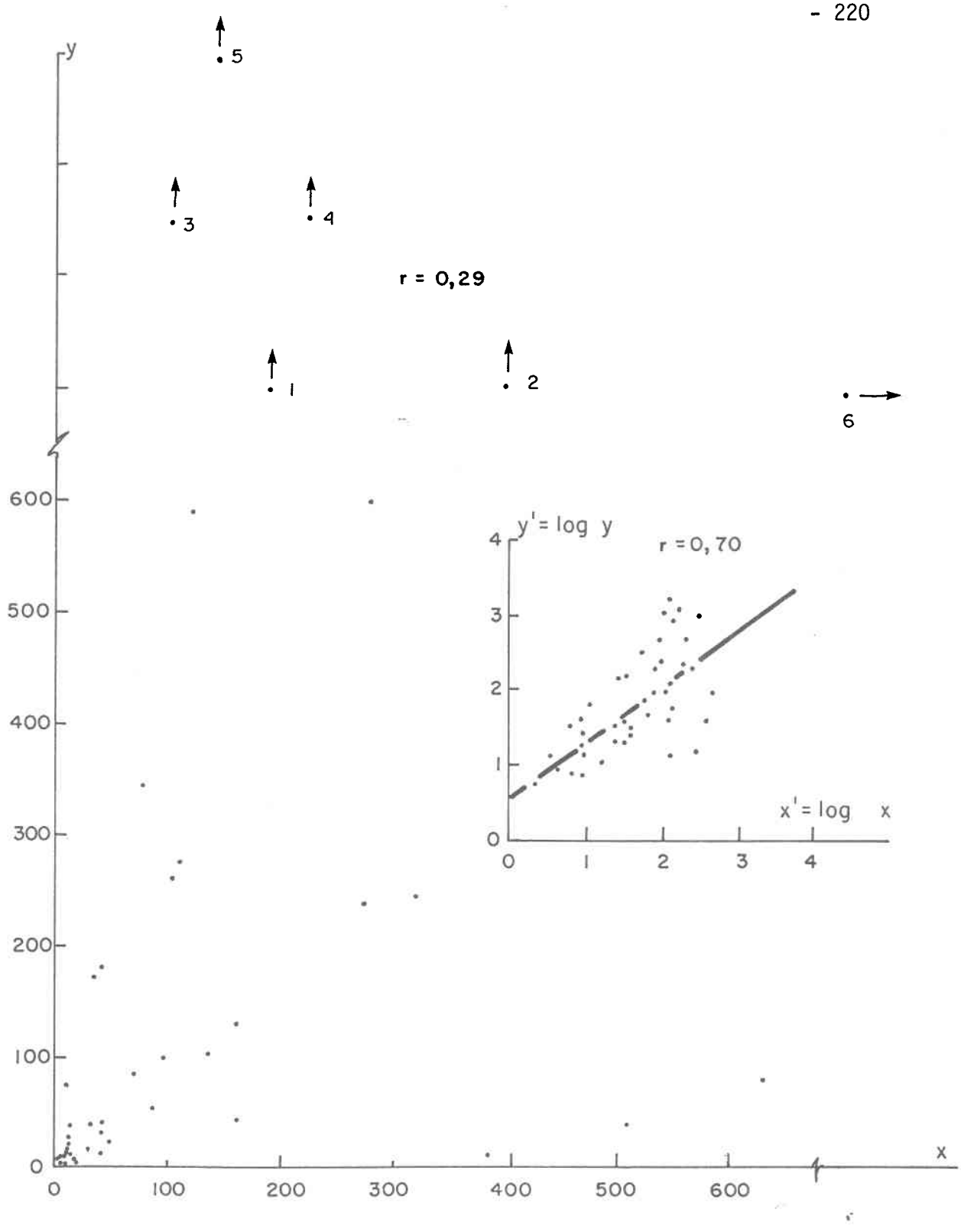


FIGURE 5.13 Coefficient de corrélation déduit des variables originales et du logarithme des variables. Exemple 5.12. Données tirées de Pleasants (1980).

À titre de comparaison on peut reprendre ici les graphiques 2 et 3 de la figure 4.7. À partir des estimateurs r' on pourrait croire à une dépendance plus importante entre les variables de la figure 5.14 B (r' = 0,96) qu'entre celles de la figure 5.14 A (r' = 0,91). Or si l'on estime T^2_{yx} on arrive à des conclusions très différentes (voir aussi tableau 5.5). Malheureusement T^2_{yx} s'applique uniquement lorsque l'on dispose de réplicats ce qui est assez rare en pratique. Dans un pareil cas on pourrait construire l'intervalle de confiance autour de la courbe ce qui permettrait au moins d'apprécier la qualité de la prédiction.

TABLEAU 5.5 Comparaison entre les estimateurs r, r^2 et T^2_{yx} (exemple de la figure 5.13)

Variables utilisées		r et (r')	r^2 et $(r')^2$	T^2_{yx}
A.	x, y (x, log y)	0,92 (0,91)	0,86 (0,83)	0,95
B.	x, y (x, log y)	0,61 (0,96)	0,37 (0,91)	0,58

Comme on l'a déjà mentionné au chapitre 3, r et r^2 présentent un intérêt lorsque la relation est linéaire. Dans ce qui précède on a aussi pu vérifier la non utilité de r' ou encore de $(r')^2$ pour rendre compte de l'interdépendance entre x et y. D'autres chercheurs utilisent le coefficient de détermination R^2 qui est équivalent à r^2 uniquement dans le cas d'une régression linéaire (voir section 2.37). Cette utilisation très fréquente de R^2 est dans plusieurs cas inadéquate. On peut vérifier que lorsque la variable dépendante est transformée la quantité R^2 n'est pas définie. En effet, dans ce cas la quantité A de l'équation 2.25 n'est pas nulle et dès lors on ne peut plus estimer R^2 (équation 2.28).

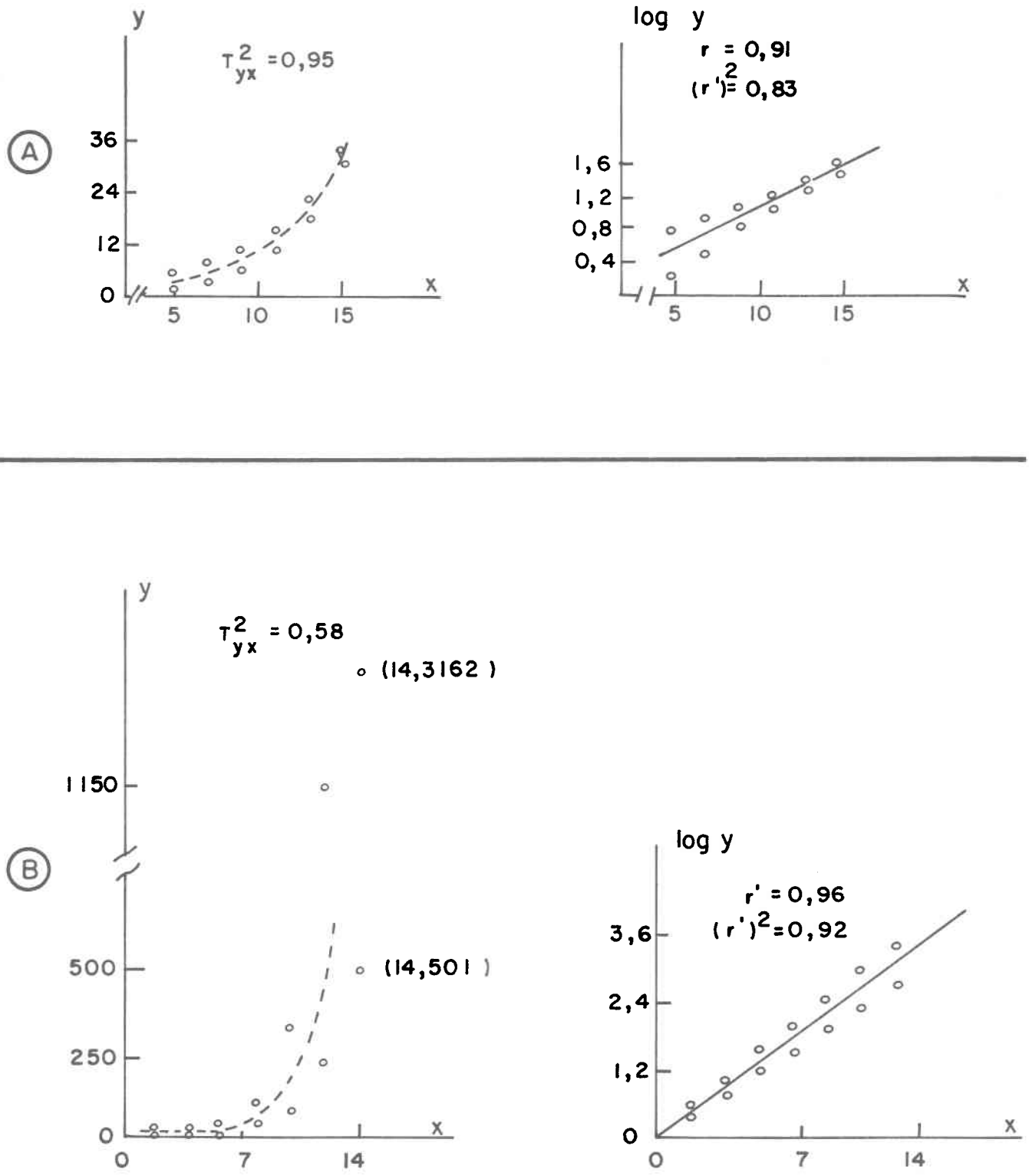


FIGURE 5.14 Comparaison entre r' et T^2_{yx} (données simulées)

Le coefficient de détermination peut cependant être utilisé dans le cas d'une régression polynômiale. Cependant là aussi il semble y avoir beaucoup de confusion et la racine de R^2 soit R ne peut être équivalente au coefficient r. Cet aspect est discuté à la section suivante.

5.5 SIGNIFICATION DE r', R^2 et R POUR UNE RELATION POLYNÔMIALE.

Les exemples sont nombreux dans la littérature scientifique où les statistiques r^2 et R^2 sont confondues. Cela vient en partie du fait croyons-nous que pour une régression linéaire simple ou multiple le coefficient de corrélation au carré (r^2) et le coefficient de détermination R^2 sont équivalents. Ce n'est cependant plus vrai lorsqu'il s'agit d'une régression polynômiale. Par exemple à la figure 5.3 A le polynôme présenté ajuste parfaitement l'échantillon. On aurait donc $R^2 = 1$ mais néanmoins $r = 0$ et $r^2 = 0$. De ce fait la racine de R^2 ne devrait pas être interprétée comme c'est souvent le cas comme équivalent à r.

EXEMPLE 5.13

Il est intéressant à cet égard de comparer les figures 5.15A et 5.15B tirés d'un article de Bilby et Likens (1980). Il nous semblait étonnant que le coefficient de corrélation de B soit plus élevé que celui de la figure 5.15 A. Nous avons d'abord calculé le coefficient de corrélation des deux échantillons avec et sans les deux points extrêmes qui selon nous faussaient

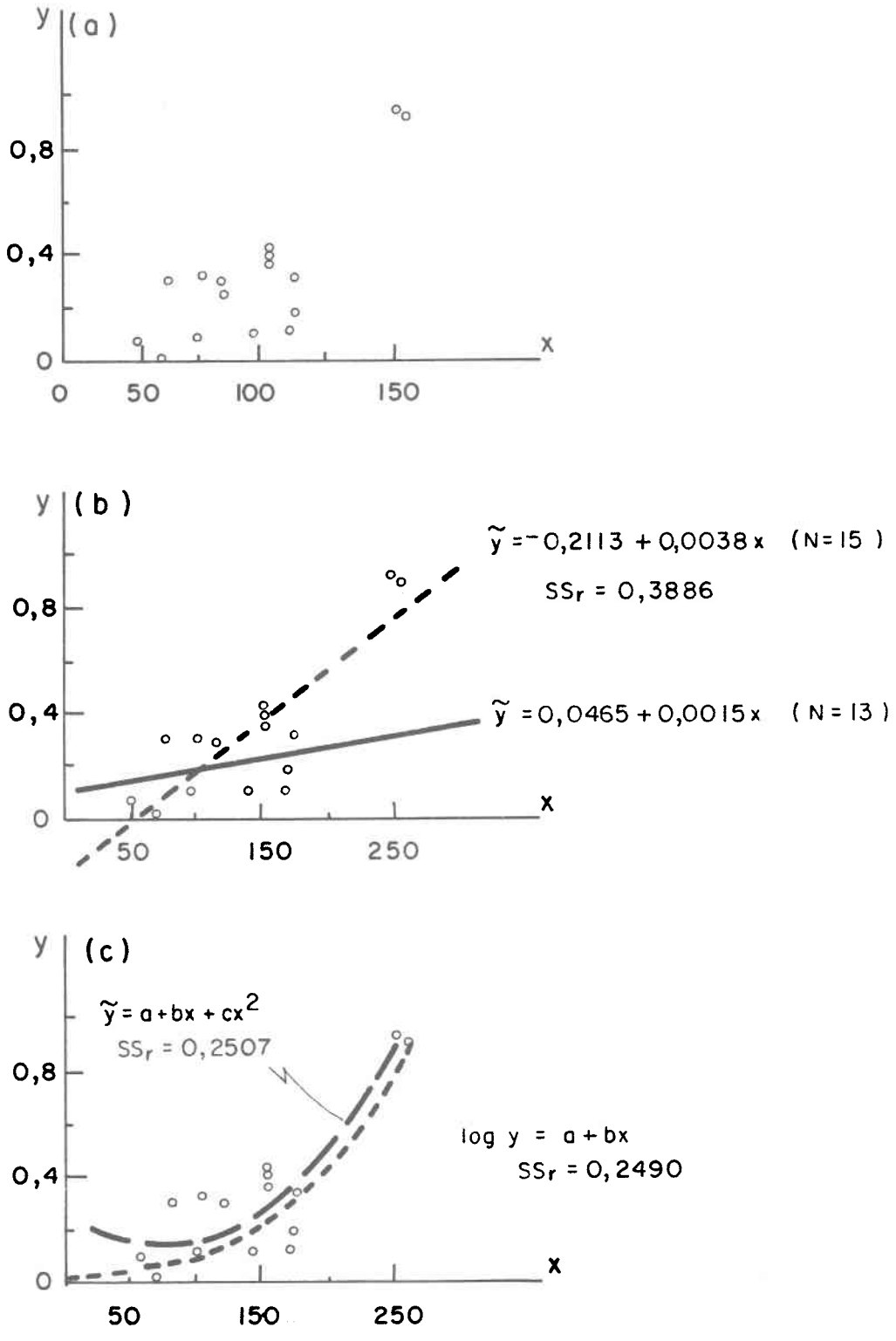


FIGURE 5.15 Utilisation de $\sqrt{R^2}$ pour une relation polynômiale (exemple 5.13). Données tirées de Bilby et Likens (1980).

considérablement la valeur de r surtout pour l'échantillon B. On obtient dans le premier cas r = 0,70 (18 observations) et r = 0,85 (échantillon total figure 5.15A) et pour le second r = 0,40 (13 observations) et r = 0,81 (échantillon total soit 14 observations). Dans un premier temps nous avions sélectionné cet article pour mettre en évidence l'influence plus ou moins importante d'un point extrême sur la valeur de r. Or il nous est apparu que la valeur du coefficient présenté dans l'article soit r = 0,90 était encore plus élevé que notre propre estimation faite en reconstituant les observations des graphiques. Une valeur aussi élevée s'approche d'une relation presque parfaite ce qui est surprenant lorsqu'on observe le graphique où la pente est quasiment nulle pour les 13 premières observations. Ce coefficient présenté comme r est en fait la racine du coefficient de détermination soit R = $\surd$ 0,81 = 0,90. C'est d'ailleurs un polynôme du second degré qui ajuste l'échantillon de la figure B. Tous les exemples de cet article où un polynôme du second degré est utilisé présentent également ce r qui est en réalité $\surd R^2$.

On peut donc se demander quel peut être la signification physique d'un coefficient positif pour une relation du second degré où la pente est comme on le sait négative nulle et positive, comme la démontre d'ailleurs l'équation du polynôme (figure 5.15 C). De plus, le seul fait que R^2 soit élevé ne permet pas de conclure à une relation significative entre les variables. En effet il s'agirait alors d'utiliser un polynôme de degré élevé pour que R^2 soit aussi très élevé et ce d'autant plus que l'échantillon est petit. On dispose ici de 14 observations dont une est extrême et devrait être considérée avec beaucoup de prudence.

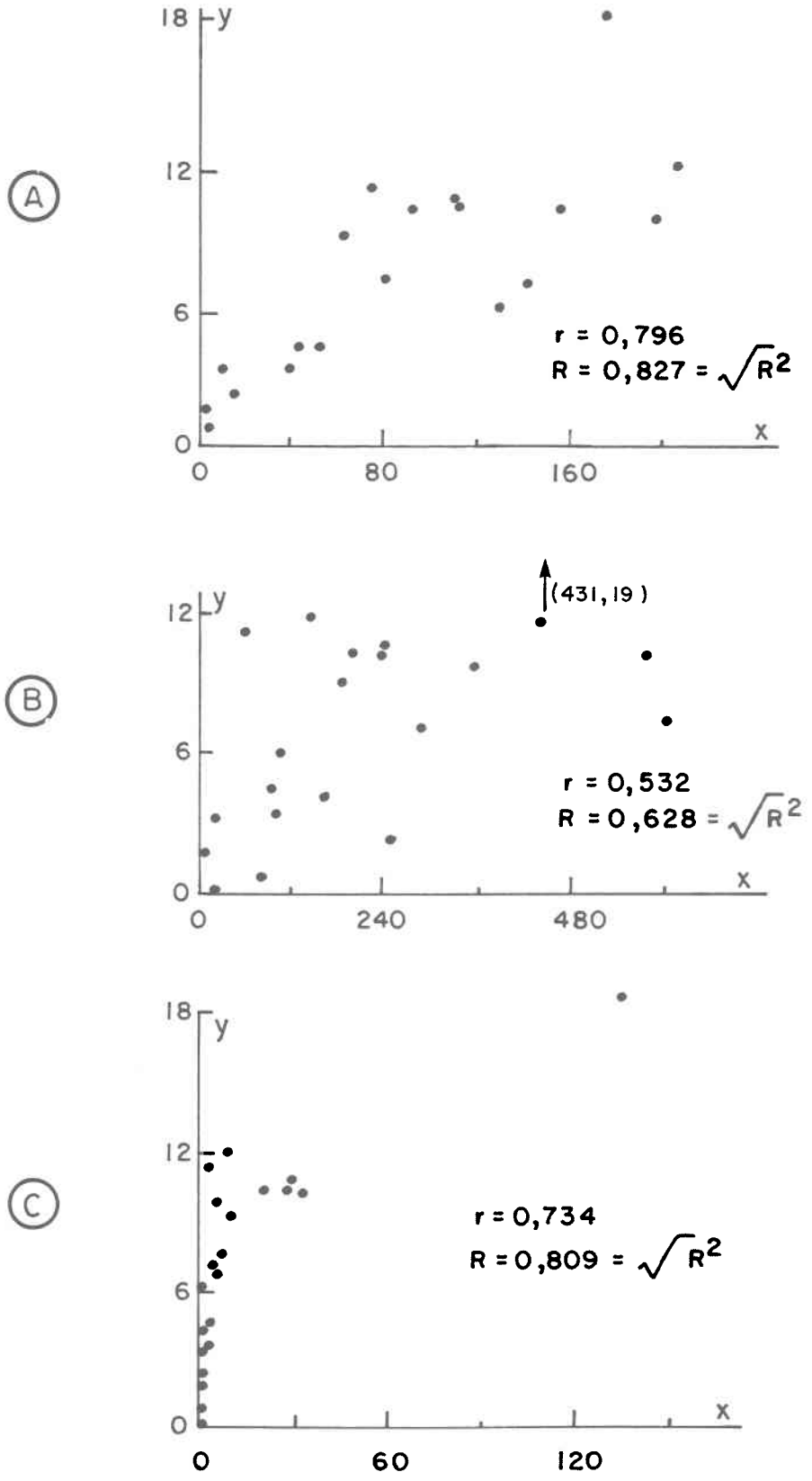


FIGURE 5.16 Comparaison entre le coefficient de corrélation r et la racine du coefficient de détermination R. Données tirées de Hernando et al. (1970).

Mentionnons aussi un article de Giroux et Bordeleau (1984) où l'on utilise un polynôme du second degré pour ajuster un échantillon de 9 observations. Même s'il apparaît peu évident que la relation soit de forme parabolique on obtient tout de même un R^2 élevé ($R^2 = 0,70$) ce qui n'est pas étonnant étant donné la taille de l'échantillon.

On trouvera aussi dans Hernando et al., (1970) de nombreux exemples où l'on présente cette fois r et R (= $\sqrt{R^2}$). L'observation des différents graphiques (figure 5.16) nous laisse très sceptique quant à la pertinence d'utiliser un polynôme du second degré, qui n'est pas justifiable du seul fait que R soit plus élevé que le coefficient de corrélation r. En fait il n'y aurait alors qu'à utiliser un polynôme de degré encore plus élevé pour s'assurer d'un R pratiquement égal à l'unité.

CHAPITRE 6

CONCLUSION

6. CONCLUSION

Dans ce qui précède on aura pu vérifier que des techniques simples du point de vue de la théorie de base peuvent néanmoins susciter des problèmes au niveau de l'interprétation et des applicatons.

Tous les aspects relatifs à ces méthodes d'analyse n'ont pu cependant être traités dans le cadre de ce mémoire. L'examen de la littérature fait ressortir, par exemple qu'une forte dépendance statistique est souvent interprétée comme une relation causale. Or les méthodes présentées ne permettent en aucun cas de mettre en évidence une relation de cause à effet entre les variables. Seules des considérations théoriques liées à la particularité du phénomène étudié autorisent à faire de telles déductions. Le caractère très spécialisé des articles consultés ne nous permettaient pas cependant de juger de la pertinence des recherches concluant à la causalité. Nous tenons cependant à rappeler qu'il n'y a pas nécessairement équivalence entre corrélation élevée et relation de causalité.

On retiendra aussi que le nombre des exemples présentés pour les différents aspects abordés n'est peut-être pas proportionnel à la fréquence des utilisations inappropriées. Si plusieurs problèmes peuvent être mis en évidence par un examen graphique d'autres s'avèrent beaucoup plus difficles à déceler sans une connaissance approfondie du phénomène étudié. Il est notamment très difficile de déceler un problème d'autocorrélation de la seule observation des données d'un graphique. Ainsi dans plusieurs articles faisant état de variables physico-chimiques recueillies en séquence temporelle, il eût été difficile de mettre en évidence ou encore de vérifier s'il

y avait autocorrélation car de façon générale on ne donne jamais d'indications précises sur les dates de mesure des différents observations. Dans cet ordre d'idées soulignons que plusieurs corrélations apparemment significatives peuvent s'avérer être des corrélations totalement fictives lorsque l'on prend connaissance de la signification même des variables mises en relation. En ce sens certains cas de corrélation fictive seront moins évidents que ceux présentés au chapitre 5. Par exemple, l'étude de la relation entre l'efficacité photosynthétique et la chlorophylle semble à priori peu contestable. Or l'efficacité photosynthétique est en fait une fonction représentant le rapport de la production primaire sur la chlorophylle.

Dans la recherche d'une relation quantitative entre des variables un nombre imposant de méthodes sont à la disposition du chercheur. La régression linéaire et la corrélation sont sans doute les plus accessibles à celui qui tout en ayant une bonne connaissance de la statistique ne peut prétendre être un spécialiste. D'autres méthodes regroupées sous le terme d'analyses multivariées peuvent sembler au profane plus performantes en raison peut-être de leur complexité. Cependant, elles ne sauraient se substituer aux méthodes qui ont été décrites ici puisqu'elles répondent à des besoins d'analyse faisant appel à un grand nombre de variables et à des échantillons de tailles importantes.

BIBLIOGRAPHIQUE

BIBLIOGRAPHIQUE

BAILLARGEON, G. et J. RAINVILLE (1977). Statistique appliquée, 5ième édition, Les éditions SMG, 594 pages.

BAMSTEDT, U. et H. R. SKJOLDAL (1980). RNA concentration of zooplankton : Relationship with size and growth. Limnology et Oceanography, 25(2): 304-316.

BARTLETT, M.S. (1937). Properties of sufficiency and statistical tests. Proc. Roy. Soc., A, 160: 268-282.

BILBY, R.E. et G.E. LIKENS (1980). Importance of organic debris dams in the structure and function of stream ecosystems. Ecology, 61(5): 1107-1113.

BENSON, M.A. (1965). Spurious Correlation in hydraulics and hydrology. Journal of the Hydraulics Division, ASCE, 91(HY4), proc. paper 3630.

BOBÉE, B. (1983). Notes de cours, 2ième année de maîtrise, INRS-Eau.

BONNY, A.P. (1978). The effect of pollen recruitment processes on pollen distribution over the sediment surface of a small lake in Columbia. Journal of Ecology, 66(2): 385-416.

BOX, G.E.P. et D.R. COX (1964). An analysis of transformations, Journal of the Royal Statistical Society, Serie B, 26: 211-243.

BOX, G.E.P. et P.W. TIDWELL (1962). Transformations of the independent variables, Technometrics, 4: 531-550.

CHATTERJEE, S. et B. PRICE (1977). Regression Analysis by example, John Wiley and Sons, 228 pages.

CHAYES, F. (1949). On ratio correlation in petrography, J. Geol., 57: 237-254.

DANIEL, C. et F.S. WOOD (1971). Fitting equations to data, Wiley-Interscience, New-York, 342 pages.

DOLBY, J.L. (1963). A quick method of choosing a transformation, Technometrics. 5: 317-325.

DURBIN, J. et G.S. WATSON (1951). Testing for serial correlation in least squares regression II. Biometrika, 38: 159-178.

FISHER, R.A. (1956). Statistical Methods and Scientific Inferences. Olivier and Boyd, Edinburgh, 407 pages.

FRANCKO, D.A. (1984). Relationships between phosporus functional classes and alkaline phosphatase activity in reservoir lakes. Journal of freshwater ecology, 2(6): 541-547.

GALTON, F. (1885). J. Anthropol. Inst., 15: 246-263.

GIROUX, M. et L.M. BORDELEAU (1984). Effet du potassium et du magnésium sur la croissance et la nodulation des plantules de luzerne. Le naturaliste canadien, Université Laval. 111(2): 167-173.

GRUBBS, F.E. (1950). Sample criteria for testing outlying observations. Ann. Math. Stat., 21(1): 28-58.

GUNST, R.F. et R.L. MASON (1980). Regression Analysis and its application, Marcel Dekker Inc., 402 pages.

HALD, A. (1952). Statistical theory of Engineering Applications, John Wiley and Sons., New-York, 783 pages.

HEDGES, J.I., TURIN, H.J. et J.R. ERTEL (1984). Sources and distributions of sedimentary organic matter in the Columbia River drainage basin, Washington and Oregon. Limnology and Oceanography, 29(1): 35-45.

HEM, J.D. (1971). Study and interpretation of the chemichal caracteristics of natural water, 2nd edition. Geological Survey Water Supply, paper 1473, 363 pages.

HERNANDO, V., LOMBARDIA, V. et M.T. PARDO (1970). Corrélations entre les fractions actives du phosphore dans le sol et leur assimilabilité pour le blé. Phosphore et agriculture. no. 56: 41-48.

HOWE, H.F. et G.A. VANDE KERCKHOVE (1981). Removal of wild nutmeg crops by birds. Ecology, 64(4): 1093-1106.

JENKYN, J.F. (1984). Effects of powdery mildew on grain filling in spring barley in contrasting environments. Annals of applied biology, 105(2): 195-212.

JENSEN, A.J. (1979). Energy content analysis from weight and liver index measurements of immature pollock. Journal of the Fisheries Research Board of Canada, 36(10): 1207-1213.

JOHNSTON, J. (1972). Econometrics Methods, Mc Graw-Hill, Inc. New-York, 437 pages.

KARAM, A. et M.P. CESCAS (1984). Adsorption du bore en relation avec le pH et les bases échangeables du sol. Le Naturaliste Canadien, 111(4): 429-434.

KENDALL, M. et A. STUART (1979). The advanced theory of statistics. Charles Griffin and Company Limited, Vol. 2, 4e édition, 748 pages.

KENDALL, M.G. et G.V. YULE (1950). An introduction to the Theory of Statistics, Charles Griffin, London.

KENNEY, B.C. (1982). Beware of spurious self-correlations. Water Resources Research, 18(4): 1041-1048.

KLEINBAUM, D.G. et L. KUPPER (1978). Applied regression analysis and other multivariable methods. Durbury Press, Massachussets, 556 pages.

MARIER, J.R., NERI, L.C. et T.W. ANDERSON (1980). Dureté de l'eau, santé et importance du magnésium. Conseil national de recherches Canada. 123 pages.

MANDEL, J. (1984). Fitting straight lines when both variables are subject to error. Journal of quality technology, 60(1): 1-14.

MONTGOMERY, D.C. et E.A. PECK (1982). Introduction to linear regression analysis, John Wiley & Sons. 504 pages.

MORGAN, M.D. (1980). Grazing and predation of the grass shrimp. Limnology and Oceanography, 25(5): 896-902.

NETER, J. et W. WASSERMAN (1974). Applied linear statistical models. Richard D. Irwin Inc., 842 pages.

NETER, J., W. WASSERMAN et M.H. KUTNER (1983). Applied linear regression models. Richard D. Irwin Inc., 547 pages.

PEARSON, K. (1897). On a form of spurious correlation which may arise when indices are used in the measurement of organs. Proc. R. Soc. London, 60: 489-502.

PLEASANTS, J.M. (1980). Competition for bumblebee pollinators in Rocky Mountain plant communities, Ecology 6(16): 1446-1459.

POULIOT, Y. (1985). Epuration des eaux usées et production de micro-algues: Étude de faisabilité. Thèse de maîtrise, No. 63-21, Université Laval, 78 pages.

REED, J.L. (1921). On the correlations between any two functions and its application to the general case of spurious correlation. Journal of the Washington Academy of Sciences. 11(19): 450-455.

ROBERTS, R.D., MARRS, R.H., SKEFFINGTON, R.A. et A.D. BRADSHAW (1981). Ecosystem development on naturally- colonized china clay wastes. Journal of Ecology, 69: 153-161.

SCHERRER, B. (1984). Biostatistique. Gaétan Morin, éditeur, 850 pages.

SHAPIRO et WILK (1965). An analysis of variance test for normality. Biometrika 52(3): 591-611.

SHARPELY, A.N., SMITH, S.J., BERG, W.A. et J.R. WILLIAMS (1985). Nutrient runoff losses as predicted by annual and monthly soil sampling. Journal of Environmental Quality. 14(3): 354-358.

SMALL, L.F., KECKES, S. et S.W. FOWLER. (1974). Excretion of different forms of zinc by the prawn. Limnology and Oceanography, 19(5): 789-793

SMILLIE, K.W. (1966). An introduction to regression and correlation. Academic Press Inc. 168 pages.

SMITH, A.P. (1980). The paradox of plant height in an Andean giant rosette species, Journal of Ecology. 68(1): 63-73.

TATE, R.F. (1978). Correlations methods. International Encyclopedia of Statistics. The Free Press, New-York, 615-627.

TESSIER, A., CAMPBELL, P.G.C., AUCLAIR, J.C., BISSON, M. et H. BOUCHER. (1982). Évaluation de l'impact de rejets miniers sur des organismes biologiques, Université du Québec, INRS-Eau, rapport no. 146: 257 pages.

THEIL, H. et A.L. NAGAR (1961). Testing the independance of regression disturbances. Journal of the American Statistical Association, 56: 793-806.

UNGAR, I.A., BENNER, D.K. et D.C. McGRAW (1979). The distribution and growth of Salicornia Europaea on an inland salt pan. Ecology, 69(2): 329-336.

VOLLENWEIDER, R.A. (1976).

Advances in defining critical loading levels for phosphores in lake entrophication. Mem. Ist. Ital. Idrobiol. Dott, Marco de Marchi, 33: 53-83.

WALLIS, W.A. et H.V. ROBERTS (1956). Statistics: A New Approach. The Free Press of Glencoe Inc., Glencoe Illinois, 646 pages.

WATSON, M.A. (1979). Age structure and mortality within a group of closely related mosses. Ecology, 60(5): 988-997.

WINTERS, G.H. (1983). Analysis of the Biological and Demographic Parameters of Northern Sand Lance, Ammodytes dubius, from the Newfoundland Grand Bank, Journal Canadien des sciences halieutiques et aquatiques, 40(4): 409-419.

YOUNGER, M.S. (1979). A handbook for linear regression, Duxbury Press, 569 pages.

APPENDICE 1

CARACTÉRISTIQUES STATISTIQUES DES ESTIMATEURS DE LA RÉGRESSION.

CARACTÉRISTIQUES STATISTIQUES DES ESTIMATEURS DE LA RÉGRESSION.

a suit une distribution normale de

moyenne $E(a) = \alpha$

variance $$var(a) = \sigma_a^2 = \sigma_\varepsilon^2 \left[\frac{1}{N} + \frac{\bar{x}^2}{\sum_{i=1}^{N} (x_i - \bar{x})^2} \right]$$

b est distribué normalement avec

une moyenne $E(b) = \beta$

une variance $$var(b) = \sigma_b^2 = \frac{\sigma_\varepsilon^2}{\sum_{i=1}^{N} (x_i - \bar{x})^2}$$

$\tilde{y}$ est distribué normalement avec

une moyenne $E(\tilde{y}) = \eta$

une variance $\quad \text{var}(\tilde{y}) = \sigma_{\tilde{y}} = \sigma_\varepsilon \left[\frac{1}{N} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{N} (x_i - \bar{x})^2} \right]$

σ_ε^2 n'étant généralement pas connu on l'estime par S_e^2

$$S_e^2 = \frac{SS_r}{N - 2} = \sum_{i=1}^{N} \frac{(y_i - \tilde{y}_i)^2}{N - 2} \qquad \text{A 1.1}$$

et on peut en déduire les variances associées de a, b et $\tilde{y}$ soit:

$$S_a^2 = S_e^2 \left[\frac{1}{N} + \frac{\bar{x}^2}{\sum_{i=1}^{N} (x_i - \bar{x})^2} \right] \qquad \text{A 1.2}$$

$$S_b^2 = \frac{S_e^2}{\sum_{i=1}^{N} (x_i - \bar{x})^2} \qquad \text{A 1.3}$$

$$S_{\tilde{y}}^2 = S_e^2 \left[\frac{1}{N} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{N} (x_i - \bar{x})^2} \right] \qquad \text{A 1.4}$$

De ces quantités ont peut déduire l'écart-type des estimatateurs soit S_a, S_b et $S_{\tilde{y}}$. Sachant que la quantité $S_e^2 / \sigma_e^2 (N-2)$ suit une loi KHI-2 à

ν =N-2 degrés de liberté et en utilisant la définition de la loi de Student (Bobée, 1983) on peut en déduire que les variables centrées réduites suivent une loi de Student avec (N-2) degrés de liberté et l'on a :

$$t_{N-2} = \frac{a - \alpha}{S_a} \qquad \text{A 1.5}$$

$$t_{N-2} = \frac{b - \beta}{S_b} \qquad \text{A 1.6}$$

$$t_{N-2} = \frac{\tilde{y} - \eta}{S_{\tilde{y}}} \qquad \text{A 1.7}$$

Le nombre de degrés de liberté (ν) pour les modèles 1 et 2 est ν = N-2 puisque l'on doit estimer deux paramètres. Pour les autres modèles où l'on connaît à priori soit α soit β on aura ν = N-1. Pour les différents modèles les quantités S_a, S_b et $S_{\tilde{y}}$ sont définis au tableau 2.2.

APPENDICE 2

TEST D'ÉGALITÉ DES VARIANCES

TEST D'ÉGALITÉ DES VARIANCES

On suppose les droites 1 et 2 dont on veut comparer les variances. On pose les hypothèses suivantes (test unilatéral):

$$H_0 : \sigma^2_{\varepsilon_1} = \sigma^2_{\varepsilon_2}$$
$$H_1 : \sigma^2_{\varepsilon_1} > \sigma^2_{\varepsilon_2}$$

Les quantités $S^2_{e_1}$ et $S^2_{e_2}$ sont les estimateurs des variances théoriques. Pour vérifier l'égalité des variances et en supposant $S^2_{e_1} > S^2_{e_1}$ on doit comparer la quantité F_0 soit:

$$F_0 = \frac{S^2_{e_1}}{S^2_{e_2}} \qquad \text{A 2.1}$$

à la quantité F_c de la table de Fisher

$$F_c = F_{\nu_1 \nu_2}\ (\tilde{\alpha})$$

avec ν_1 : degré de liberté du numérateur = n_1-2

ν_2 : degré de liberté du dénominateur = n_2-2

n_1 et n_2 : nombres d'observations des droites 1 et 2

Si $F_0 < F_c$ on accepte H_0

Si $F_0 > F_c$ on rejette H_0

APPENDICE 3

TEST D'ÉGALITÉ DES PENTES

TEST D'ÉGALITÉ DES PENTES

On suppose les droites 1 et 2 dont on veut comparer les pentes. Ce test, Hald (1952) est valide lorsque les variances ne sont pas significativement différentes. Si tel n'est pas le cas un autre test devra être utilisé. Il est également présenté dans Hald (1952). Pour un test unilatéral on pose les hypothèses:

$$H_0 : \beta_1 = \beta_2$$
$$H_1 : \beta_1 < \beta_2$$

et on compare la quantité t_0

$$t_0 = \frac{b_1 - b_2}{S_{et} \left(1/SS_{x_1} + 1/SS_{x_2} \right)^{\frac{1}{2}}} \qquad \text{A 3.1}$$

$$\text{avec} \quad S_{et}^2 = \frac{(n_1-2)\, S_{e_1}^2 + (n_1 -2)\, S_{e_2}^2}{(n_1 + n_2 -4)} \qquad \text{A 3.2}$$

à la valeur critique $t_c = t_\nu, (\tilde{\alpha})$ avec $\nu = (n_1 + n_2 -4)$

si $t_0 < t_c$ H_0 est accepté

$t_0 > t_c$ H_0 est rejeté

APPENDICE 4

TEST D'IDENDITÉ DES DROITES (Test t)

TEST D'IDENDITÉ DES DROITES (Test t)

Comme pour le test F décrit en 2.3.8.2 on doit préalablement appliquer un test d'égalité des variances. Les résultats de ce test permettront de choisir le test adéquat pour vérifier l'égalité des pentes. En supposant que les pentes ne sont pas significativement différentes on considérera:

(a) test d'identité des droites avec variances non significativement différentes.

(b) test d'identité des droites avec variances significativement différentes.

Pour simplifier la présentation du test on se référera au tableau synthèse reporté à la fin de cette appendice qui résume les principales notations. Dans ce qui suit on comparera les droites en utilisant le modèle 2 présenté à la section 2.3.4.2. Les droites théoriques que l'on veut comparer sont:

$$\eta_1 = \alpha_1' + \beta_1 (x - \bar{x}_1)$$

et

$$\eta_2 = \alpha_2' + \beta_2 (x - \bar{x}_2)$$

estimés dans l'échantillon par

$$\tilde{y}_1 = a_1' + b_1 (x - \bar{x}_1)$$

et

$$\tilde{y}_2 = a_2' + b_2 (x - \bar{x}_2)$$

(a) test d'identité des droites (les variances $\sigma^2_{\varepsilon_1}$ et $\sigma^2_{\varepsilon_2}$ ne sont pas significativement différentes).

Par ce test (Hald, 1952) étudie la quantité t_o telle que:

$$t_o = \frac{\tilde{b} - \bar{b}}{\sqrt{\text{Var}\,(\tilde{b} - \bar{b})}} \qquad \text{A 4.1}$$

$\bar{b}$ correspond à la moyenne pondérée des pentes b_1 et b_2

$$\bar{b} = \frac{SS_{x_1} b_1 + SS_{x_2} b_2}{SS_{x_1} + SS_{x_2}} \qquad \text{A 4.2}$$

Sa variance est estimée par $S^2_{\bar{b}}$

$$S^2_{\bar{b}} \simeq \frac{S^2_{et}}{SS_{x_1} + SS_{x_2}} \qquad \text{A 4.3}$$

Lorsque N est petit Hald (1952), suggère d'utiliser $\tilde{S}^2_{et}$ plutôt que S^2_{et}

- $\tilde{b}$ correspond à l'estimation de la pente théorique $\tilde{\beta}$ en supposant que les deux droites théoriques ont une pente identique. On a donc:

$$\eta_1 = \alpha_1' + \beta\,(x - \bar{x}_1) = \alpha_1' - \beta\bar{x}_1 + \beta x$$
$$\eta_2 = \alpha_2' + \beta\,(x - \bar{x}_2) = \alpha_2' - \beta\bar{x}_2 + \beta x$$

En supposant les termes constants égaux on a

$$\alpha_1' - \beta\bar{x}_1 = \alpha_2' - \beta\bar{x}_2$$

et $\tilde{\beta}$ est estimé dans l'échantillon par:

$$\tilde{b} = \frac{\bar{y}_1 - \bar{y}_2}{\bar{x}_1 - \bar{x}_2} \qquad \text{A 4.4}$$

la variance de $\tilde{b}$, est estimé par $S^2_{\tilde{b}}$

$$S^2_{\tilde{b}} = \frac{S^2_{et}}{(\bar{x}_1 - \bar{x}_2)^2}\left(\frac{1}{N_1} + \frac{1}{N_2}\right) \qquad \text{A 4.5}$$

et on a d'autre part : Var $(\bar{b} - \tilde{b})$ = Var $(\bar{b})$ + Var $(\tilde{b})$ estimé par:

$$S^2(\bar{b} - \tilde{b}) = S^2_{\bar{b}} + S^2_{\tilde{b}} \qquad \text{A 4.6}$$

Si les pentes théoriques sont égales la quantité $u = \tilde{b} - \bar{b} \,/\, \sqrt{\text{Var}\,(\tilde{b} - \bar{b})}$ suit une normale N(0,1). Puisque la variance théorique est estimé par S^2

cette quantité suivra une distribution de student et l'on devra comparer t_o

$$t_o = \frac{\tilde{b} - \bar{b}}{\sqrt{s^2(\tilde{b} - \bar{b})}} \qquad \text{A 4.7}$$

à la valeur critique t_c . On accepte l'hypothèse d'identité des droites si $t_o < t_\nu(\tilde{\alpha})$ avec $\nu = N_1+N_2-4$

(b) test d'identité des droites (variances significativement différentes).

On considère la quantité u telle que:

$$u = \frac{\tilde{b} - \bar{b}}{\sqrt{Var\ (\tilde{b} - \bar{b})}} \qquad \text{A 4.8}$$

qui suit une loi normale, N(0,1). En remplacant Var $(\bar{b} - \tilde{b})$ par S^2 $(\bar{b} - \tilde{b})$ cette quantité sera aussi distribuée normalement si N_1 et N_2 sont grands. Les quantités $\bar{b}$, $\tilde{b}$, Var$(\bar{b})$ et Var $(\tilde{b})$ correspondent à :

$$\bar{b} = \left((b_1 SS_{x_1} / S^2_{e_1}) + (b_2 SS_{x_2} / S^2_{e_2})\right) \Big/ \left((SS_{x_1} / S^2_{e_1}) + (SS_{x_2} / S^2_{e_2})\right)$$

$$Var(\bar{b}) = \left((SS_{x_1} / S^2_{e_1}) + (SS_{x_2} / S^2_{e_2})\right)^{-1}$$

$$\tilde{b} = (\bar{y}_1 - \bar{y}_2) / (\bar{x}_1 - \bar{x}_2)$$

$$\text{Var}(\tilde{b}) = 1/\ (\bar{x}_1 - \bar{x}_2)^2 \left((S_{e_1}^2 /N_1 + S_{e_2}^2 /N_2) \right)$$

PRINCIPALES NOTATIONS Échantillon 1

$\bar{x}_1$ = moyenne, variable indépendante

$\bar{y}_1$ = a_1' = moyenne, variable dépendante

$SS_{x_1} = S_{x_1}^2 (N_1 - 1)$

$SS_{y_1} = S_{y_1}^2 (N_1 - 1)$

σ_e^2 = variance résiduelle théorique estimé par $S_{e_1}^2 = SS_{r_1} / (N_1 - 2)$

σ_{et}^2 variance théorique si $\beta_1 = \beta_2$ estimé par S_{et}^2

$$S_{et}^2 = \frac{(N_1 - 2)\ S_{e_1}^2 + (N_2 - 2)\ S_{e_2}^2}{(N_1 + N_2 - 4)}$$

$$\tilde{S}_{et}^2 = \frac{S_{et}^2 + C}{(N_1 + N_2 - 3)}$$

avec

$$C = \frac{(b_1 - b_2)^2}{\frac{1}{SS_{x_1}} + \frac{1}{SS_{x_2}}}$$

APPENDICE 5

TEST DE BARTLETT

TEST DE BARTLETT

Le test de Bartlett (1937) a été développé pour vérifier si les variances de K échantillons différents étaient égales. On compare alors les variances de ces échantillons constitués chacun d'un certain nombre d'observations n.

Ce test peut être utilisé pour vérifier l'hypothèse d'homoscédasticité en considérant K différents x_i (i = 1 à K) constitués des y_{ij} (j = 1 à n_i) où n_i correspond au nombre de réplicats pour chacun des différents x_i. On comparera donc les variances conditionnelles $\sigma^2_{y|x_i} = \sigma^2_i$ (i = 1 à K), estimées dans l'échantillon par S^2_i. Les hypothèses que l'on doit poser sont:

H_0 : $\sigma^2_1 = \sigma^2_2 = \ldots \sigma^2_K$

H_1 : les variances ne sont pas toutes égales

Pour appliquer ce test on considère les quantités suivantes:

$$\sum_{i=1}^{N} n_i = N \qquad \text{N = nombre total d'observations}$$

$$M = \frac{1}{N-K} \sum_{i=1}^{K} (n_i - 1)\, S_i^2$$

$$G = \left[(S_1^2)^{n_1-1} (S_2^2)^{n_2-1} \ldots (S_K^2)^{n_K-1} \right]^{1/N-K}$$

M correspond à la moyenne arithmétique des K variances et G à leur moyenne géométrique.

Bartlett a démontré que pour de larges échantillons, la quantité (log M - log G) suit approximativement une distribution de χ^2 avec (K - 1) degrés de liberté. Pour vérifier l'égalité des variances on devra donc calculer la quantité B soit.

$$B = \frac{2.302585}{C} (N - K) (\log M - \log G)$$

qui peut être réduit à

$$B = \frac{2.302585}{C} \left((N - K) \log M - \sum_{i=1}^{K} (n_i - 1) \log S_i^2 \right)$$

2.302585 est un facteur de conversion pour passer de log ($\log_{10}$) à ln ($\log_e$)

$$C = 1 + \frac{1}{3 (K-1)} \left(\left(\sum_{i=1}^{K} \frac{1}{n_i - 1} \right) - \frac{1}{N-K} \right)$$

si $B < \chi^2 (1 - \tilde{\alpha} ; K - 1)$ H_0 est accepté.

$B > \chi^2 (1 - \tilde{\alpha} ; K - 1)$ H_0 est refusé.

APPENDICE 6

EXPRESSION DU COEFFICIENT DE CORRÉLATION RÉSULTANT D'UNE FONCTION DE VARIABLES

(cas d'une fonction additive)

EXPRESSION DU COEFFICIENT DE CORRÉLATION RÉSULTANT D'UNE FONCTION DE VARIABLES (cas d'une fonction additive).

On suppose les variables x et y et les fonctions u et v

$$u = x$$

$$v = x + y$$

D'après l'équation 3.3 r_{xy} peut s'exprimer par:

$$r_{xy} = \frac{S_{xy}}{S_x \, S_y} = \frac{N \sum_{i=1}^{N} xy - \sum_{i=1}^{N} x \sum_{i=1}^{N} y}{\left[N \sum_{i=1}^{N} x^2 - \left(\sum_{i=1}^{N} x \right)^2 \quad N \sum_{i=1}^{N} y^2 - \left(\sum_{i=1}^{N} y \right)^2 \right]^{1/2}}$$

et l'on a pour les variables u et v exprimées en fonction de x et y

$$r_{uv} = \frac{N \sum_{i=1}^{N} x\,(x+y) - \sum_{i=1}^{N} x \sum_{i=1}^{N} (x+y)}{\left[N \sum_{i=1}^{N} x^2 - \left(\sum_{i=1}^{N} x \right)^2 \quad N \sum_{i=1}^{N} (x+y)^2 - \left(\sum_{i=1}^{N} (x+y) \right)^2 \right]^{1/2}}$$

En simplifiant et en regroupant certains termes on obtient une expression équivalente

$$r_{uv} = \frac{S_x^2 + S_{xy}}{S_x \left[S_x^2 + S_y^2 + 2\, S_{xy} \right]^{1/2}}$$

comme $S_{xy} = r_{xy}\, S_x\, S_y$ (équation 3.3) on a:

$$r_{uv} = \frac{S_x + r_{xy} S_y}{\left[S_x^2 + S_y^2 + 2r_{xy}\, S_x\, S_y\right]^{\frac{1}{2}}}$$

ANNEXE 1

ILLUSTRATION DES PRINCIPAUX CALCULS POUR LA COMPARAISON

DES DROITES DE RÉGRESSION PAR LE TEST F

ILLUSTRATION DES PRINCIPAUX CALCULS POUR LA COMPARAISON DES DROITES DE REGRESSION PAR LE TEST F

Échantillon d_1		Échantillon d_2	
x	y	x	y
1	4,0	1	10,0
2	5,5	2	11,5
3	5,0	3	11,0
4	6,5	4	12,5
5	6,0	5	12,0
6	7,5	6	13,5
7	7,0	7	13,0
8	8,5	8	14,5

$\tilde{d}_1 = 3,79 + 0,55\ x$

$\tilde{d}_2 = 9,79 + 0,55\ x$

$\tilde{D} = 6,79 + 0,55\ x$

$SS_{r_1} = SS_{r_2} = 1,905$

$SS_{r_{12}} = 3,81$

$SSR = 147,81$

$F_0 = \dfrac{(147,81 - 3,81)\ /2}{3,81/4} = 75,59$

$F_0 > F_c$: H_0 rejeté

Échantillon m_1		Échantillon m_2	
x	y	x	y
1	3,5	1	9,5
2	7,0	2	13,0
3	3,5	3	9,5
4	8,0	4	14,0
5	4,5	5	10,5
6	9,0	6	15,0
7	5,5	7	11,5
8	10,0	8	16,0

$\tilde{m}_1 = 3,64 + 0,61\ x$

$\tilde{m}_2 = 9,64 + 0,61\ x$

$\tilde{M}_2 = 6,64 + 0,61\ x$

$SS_{r_1} = SS_{r_2} = 28,39$

$SS_{r_{12}} = 56,78$

$SSR = 72,78$

$F_0 = \dfrac{(72,78 - 56,78)\ /\ 2}{56,78/\ 14} = 1,97$

$F_0 < F_c$: H_0 accepté

$F_c = F_{2,14}\ (0,05) = 3,74$

$F_c = F_{2,14}\ (0,01) = 6,51$

ANNEXE 2

DONNÉES DES EXEMPLES 4.2 à 4.10

Annexe 2.1

Exemple 4.2 (figure 4.6)	
x	ln x
2	6,36
3	6,14
4	5,72
5	5,23
6	4,53
7	3,64
8	2,59
9	1,98

Exemple 4.3 (figures 4.7 et 5.14)					
Série 1		Série 2		Série 3	
x	y	x	y	x	y
2	1,26	5	6,31	2	2,63
2	2,00	5	1,58	2	3,47
4	2,00	7	8,16	4	5,75
4	3,16	7	3,43	4	10,96
6	3,16	9	11,09	6	15,84
6	5,01	9	6,37	6	39,81
8	5,01	11	15,74	8	36,30
8	7,94	11	11,01	8	112,20
10	7,94	13	23,10	10	83,18
10	12,59	13	18,38	10	331,13
12	12,59	15	34,77	12	218,77
12	19,95	15	30,05	12	1148,15
				14	501,18
				14	3162,27

Annexe 2.2

Exemple 4.5

x	y	x	y
0,1450	21,72	0,2033	22,41
0,1548	21,90	0,2038	22,30
0,1596	21,96	0,2042	22,50
0,1716	21,75	0,2058	23,08
0,1787	22,08	0,2063	23,00
0,1916	22,36	0,2075	22,66
0,1916	22,58	0,2075	22,83
0,1936	22,82	0,2058	22,33
0,1975	22,33	0,2063	22,50
0,1992	22,73	0,2073	23,50
0,1998	22,56	0,2075	23,80
0,2008	22,42	0,2133	23,75
0,2008	22,78	0,2150	24,00
0,2022	22,91	0,2183	24,41
0,2016	23,01	0,2183	24,16
0,2025	22,71	0,2208	23,08
0,2033	22,63	0,2283	24,58

Exemple 4.4

x	y
2,40	0,9
3,00	1,0
2,50	1,1
3,10	1,2
3,20	1,3
2,40	1,3
3,20	1,5
2,40	1,8
3,60	2,0
2,60	2,1
2,40	2,6
4,10	2,8
1,80	3,0
3,00	3,2
5,00	3,2

Annexe 2.3

Exemples 4.6 et 5.1 (figures 4.15, 4.16, 5.12)			
Azote		Phosphore	
x	y	x	y
2,17	0,63	0,093	0,087
2,63	0,40	0,133	0,077
3,10	0,45	0,167	0,053
2,67	0,67	0,167	0,067
3,00	0,77	0,196	0,107
2,83	1,30	0,177	0,140
3,00	1,07	0,200	0,117
3,40	1,17	0,217	0,050
2,83	1,67	0,247	0,063
3,17	2,17	0,367	0,073
3,63	2,00	0,250	0,107
3,70	2,67	0,327	0,133
4,17	3,60	0,200	0,183
2,53	3,67	0,230	0,200
4,20	1,40	0,326	0,170
4,67	0,97	0,400	0,160
4,90	1,00	0,213	0,267
5,17	1,23	0,280	0,233
7,83	0,60	0,083	0,263
8,00	2,17	0,133	0,233
8,27	1,00	0,150	0,200
8,73	1,50	0,433	0,307
4,63	9,17	0,493	0,373
5,57	9,30	0,283	1,070
6,17	9,30	0,306	1,000
7,73	9,03	0,473	0,476
7,50	9,57	0,963	0,976
7,57	9,73	0,906	1,140
2,00	0,83	1,030	1,180
1,93	1,93	1,197	0,726
1,77	6,07	1,197	0,783

Annexe 2.4

Exemple 4.8 (figure 4.23)					
Groupe 1(1)		Groupe 2(2)		Groupe 3(3)	
x	y	x	y	x	y
- 0,40	- 2,26	- 1,28	- 4,89	- 1,53	- 6,79
- 0,40	- 2,44	- 1,28	- 5,00	- 1,89	- 6,47
- 0,40	- 4,42	- 2,82	- 6,00	- 1,95	- 6,77
- 0,40	- 4,63	- 2,74	- 6,29	- 3,05	- 7,58
- 0,55	- 3,67	- 2,74	- 6,42	- 3,05	- 7,68
- 0,55	- 3,84	- 2,74	- 6,53	- 3,37	- 7,95
- 0,55	- 4,04	- 2,84	- 6,32	- 3,58	- 6,89
- 0,55	- 4,25	- 2,86	- 6,50	- 3,82	- 6,79
- 0,65	- 3,42	- 4,26	- 5,32	- 3,89	- 6,68
- 0,65	- 3,52	- 4,26	- 5,52	- 4,79	- 9,10
- 0,65	- 3,84	- 4,26	- 5,63	- 4,79	- 8,84
- 0,65	- 4,42	- 4,63	- 5,58	- 4,94	- 7,32
- 0,90	- 2,89	- 4,63	- 5,68	- 5,42	- 7,58
- 0,90	- 3,05			- 5,32	- 7,42
- 0,90	- 3,16				
- 0,90	- 3,26				
- 1,01	- 2,63				
- 1,01	- 2,74				
- 1,01	- 2,89				
- 1,01	- 3,11				
- 1,74	- 2,68				
- 1,74	- 3,29				
- 1,74	- 3,47				
- 1,74	- 3,68				
- 1,74	- 4,00				
- 1,74	- 4,84				
- 2,00	- 3,47				
- 2,00	- 3,58				
- 2,00	- 3,68				

(1) Observations mesurées en laboratoire

(2) Observations mesurées sur le terrain (période 1)

(3) Observations mesurées sur le terrain (période 2)

Annexe 2.5

Exemple 4.9 (figure 4.24)	
x	y
0,47	14,7
0,85	12,4
0,65	19,5
3,10	14,0
2,90	16,0
2,30	16,8
2,80	11,8
1,70	13,1
0,41	12,8
1,40	15,5
0,90	14,2
0,97	16,8
1,40	14,0
0,87	12,5
1,20	14,1
3,20	22,4 (MR)
4,10	19,8 (WL1)
5,20	32,1 (WL2)
5,00	20,2 (TP)

Exemple 4.11 (figure 4.26)	
x	y
52,50	0,0750
70,00	0,0125
84,00	0,3050
96,00	0,1000
102,50	0,3200
117,50	0,2950
141,00	0,1050
151,50	0,3500
152,50	0,3875
154,00	0,4375
169,00	0,1150
172,50	0,1800
175,00	0,3250
247,50	0,9500
256,00	0,9250

Exemple 4.10 (figure 4.25)			
Bassin 1 (35cm)		Bassin 2 (25cm)	
x	y	x	y
290,00	224,3	148,0	121,3
526,5	284,9	408,0	190,2
580,0	341,2	405,0	239,3
1037,5	470,4	796,0	368,0
965,0	434,0	752,0	337,2

ANNEXE 3

DONNÉES DES EXEMPLES 5.2 à 5.14

Annexe 3.1

Exemple 5.2 Figure 5.3A	
x	y
9,5	0,1622
10,5	0,1891
12,0	0,2797
14,0	0,3431
16,0	0,3793
17,0	0,3872
18,0	0,3883
20,0	0,3701
22,0	0,3247
24,0	0,2521
26,0	0,1523

Exemple 5.3 Figure 5.3B	
x	y
3 000	47
3 000	49
3 210	45
3 210	55
3 440	45
3 440	49
3 600	50
3 600	61
3 700	65
3 700	67
3 780	75
3 780	87

Exemple 5.5 Figure 5.5A	
x	y
32,00	10,19
36,00	5,44
38,95	9,10
38,95	7,60
38,95	6,65
38,84	6,24
40,00	3,55
40,00	2,00
40,88	6,38
40,88	5,13
40,88	3,09

Exemple 5.6 Figure 5.5B	
x	y
11,4	4,7
7,1	7,0
13,7	5,1
7,8	4,7
18,8	10,6
9,0	6,1
7,4	8,1
45,9	50,0
12,4	15,7
12,6	10,0
16,6	11,90
2,2	2,4
15,7	30,7
5,6	2,4

Annexe 3.2

Exemple 5.7 (figure 5.6)				
Âge	[Mat. org.]	[Azote]	[Mg]	[Ca]
16	1636	250	34,1	123
16	2000	320	48,5	192
18	1636	290	74,2	192
18	2000	400	81,8	246
26	1818	900	69,7	62
26	3364	1000	103,7	115
39	3182	300	81,8	69
39	10456	700	88,9	246
40	1636	800	166,7	38
40	2273	900	196,3	115
40	3636	950	207,4	154
40	5455	1150	213,9	192
40	8182	1250	221,5	254
40	9545	1900	324,4	377
42	2545	700	50,0	61
42	3182	800	67,7	115
43	4091	1120	100,0	154
43	5000	1550	128,6	192
48	2727	1150	32,3	38
48	3636	1300	57,7	77
51	2273	75	96,2	31
51	3182	450	101,9	77
51	4091	700	116,3	154
51	4545	850	122,2	192
51	5455	1050	140,7	323
51	6182	1320	190,4	538
57	1818	275	34,6	54
57	2727	500	50,0	115
57	3182	600	66,9	192
57	4091	850	76,9	400
76	2273	775	50,0	38
76	4364	875	57,7	115
98	6545	680	130,8	77
98	9545	1000	176,9	192
116	455	320	50,0	62
116	10910	1150	280,7	323
116	36964	1950	309,2	577
116	46364	4000	342,3	1770

Annexe 3.3

Exemple 5.8 (figure 5.7)		
Sous-bassin 1		
Mn	Zn	Cu
205	77,8	10,0
310	23,6	2,4
124	55,8**	30,7**
208	96,7	15,2
137	91,2	13,5
140	75,4	16,4
212	112,0	21,9
253	89,7	13,5
300	81,3	9,0
330	76,5	9,4
380	77,0	9,1
414	72,6	19,9
404	73,1	12,0
Sous-bassin 2		
292*	20,1*	2,4
494	26,3	4,7
589	23,4	7,0
529	27,2	5,1
390	20,5	4,7
330	29,8	10,6
390	22,5	6,1
319	22,8	8,1
420	61,0**	50,0**

* (292-20,1) non considéré dans les calculs. (Fig. 5.7).

** (55,8-30,7)= P_1
(61,0-50,0)= P_2

Exemple 5.9 (figure 5.10)	
P	A
5,08	0,85
9,38	1,51
10,78	1,36
19,53	1,94
25,10	3,65
27,73	1,17
33,98	1,28
32,81	2,64
42,97	1,24
47,27	2,25

Annexe 3.4

Exemple 5.11 (figure 5.12)	
x	y
2,06	0,2989
2,84	0,9852
3,03	0,8735
3,30	0,9493
2,95	1,2342
3,03	1,1925
3,03	1,6263
3,58	1,3236
5,43	1,4336
6,07	0,9981
6,97	1,6659
6,97	1,3480
7,37	1,5546
7,79	1,8613
7,37	0,8348
8,95	2,1382
9,19	0,9787
10,28	1,2384
10,57	1,5353
9,94	0,9280
10,86	0,4834
14,18	0,5444
15,58	0,8897
16,02	1,0364

Exemple 5.12 (figure 5.13)			
x	y	x	y
161,2	130,2	84,3	53,1
630,0	81,0	75,3	343,2
16,9	11,5	28,4	21,9
136,6	104,8	32,0	39,7
8,4	7,6	125,0	589,0
44,0	27,4	273,4	240,0
10,2	28,2	6,9	37,9
3,4	13,9	320,0	247,7
12,8	77,9	8,8	43,1
100,4	100,2	387,5	15,0
181,9	14,5	35,2	176,8
69,4	85,6	510,6	43,1
178,0	59,7	7,8	7,8
39,2	182,0	42,2	32,9
112,1	274,0		
10,1	13,4	190,3	1100,2 (Pt1)
38,0	41,4	400,1	1210,0 (Pt2)
9,9	19,1	105,0	1433,2 (Pt3)
2,4	6,4	223,6	1446,0 (Pt4)
39,4	21,9	141,7	1868,4 (Pt5)
161,7	45,5	1992,0	862,2 (Pt6)
104,7	258,2		
279,0	602,7		
8,1	9,8		

Annexe 3.5

Exemple 5.13 (figure 5.15A)		(figure 5.15B)	
x	y	x	y
0,04	4	38	500
0,05	8	63	750
0,05	12	66	705
0,10	5	70	400
0,10	16	75	790
0,10	23	81	500
0,13	5	85	630
0,14	20	100	515
0,15	8	107	995
0,25	15	121	300
0,26	12	125	900
0,30	37	138	1125
0,30	50	206	830
0,35	10	290	2620
0,35	18		
0,62	74		
0,75	49		
0,78	32		
2,18	100		

ANNEXE 4

DONNÉES DES FIGURES 5.1 - 5.4 - 5.9

Annexe 4.1

Données de la figure 5.1A			
ph	A	ph	A
7,35	4,48	6,00	2,79
7,40	5,14	5,90	2,53
7,45	5,15	6,20	2,48
7,45	5,12	6,25	3,36
7,45	3,85	6,20	3,40
7,35	3,21	7,60	3,95
7,35	3,72	7,65	3,94
7,40	4,11	7,20	2,65
7,40	2,99	6,90	2,55
7,45	3,15	6,60	2,33
7,35	2,47	3,65	2,30
7,60	3,95	2,90	2,15
7,70	4,41	5,50	2,45
7,70	5,46	5,60	2,38
7,70	5,72	5,60	2,24
7,45	2,80	5,70	2,35
7,50	2,93	5,80	2,47
7,60	3,46	5,85	2,89
7,50	3,32	5,80	2,61

Données de la figure 5.1B	
ph	A
6,85	3,01
6,85	1,52
6,95	2,47
6,85	4,02
6,60	1,72
6,70	2,88
6,70	2,45
6,85	2,07
6,70	2,36
6,50	2,69
6,75	3,70
7,05	3,16
7,05	3,64
7,05	2,95
6,95	3,36
6,90	2,37
6,70	2,38
6,50	2,50
6,85	2,04
6,85	2,29
6,95	2,86
4,60	1,53
4,05	1,69
3,70	1,50

Données de la figure 5.4 N=49							
x_i	y_1	y_2	y_3	y_4	y_5	y_6	y_7
1	1	2	3	4	5	6	7
2	1	2	3	4	5	6	7
3	1	2	3	4	5	6	7
4	1	2	3	4	5	6	7
5	1	2	3	4	5	6	7
6	1	2	3	4	5	6	7
7	1	2	3	4	5	6	7

Annexe 4.2

Données de la figure 5.9A					
Groupe 1		Données communes aux gr. 1 et 2		Groupe 2	
x	y	x	y	x	y
0,0753	0,0118	0,0047	0,0047	0,0353	0,2118
0,0659	0,0353	0,0047	0,0235	0,0541	0,2282
0,0588	0,0753	0,0047	0,0471	0,0941	0,3600
0,0894	0,0894	0,0353	0,0047	0,1059	0,5412
0,0941	0,1294	0,0306	0,0282	0,2000	0,4588
0,1247	0,1059	0,0471	0,0282	0,1882	0,7882
0,1529	0,0988	0,0471	0,0353	0,2282	0,7129
0,1694	0,1600			0,3059	0,8471
0,2518	0,3059				
0,2753	0,2471				
0,2753	0,3459				
0,3294	0,3059				
0,4424	0,4706				
0,5129	0,2753				
0,7224	0,7882				
0,8541	0,7224				

Données de la figure 5.9B			
Groupe 1		Données communes	
x	logy	4,11	1,65
		3,48	1,09
0,89	2,28		
1,61	2,40	Groupe 2	
2,23	2,32		
4,73	0,03		
4,14	- 0,18	5,92	1,41
2,94	- 0,65	7,94	1,09
		12,89	1,19
		20,00	0,44
		25,00	0,93