Centre de Montpellier

Unité Formation

Formation statistique

au modèle linéaire général

****

support de cours

****

Cellule de Biométrie et statistique

24 - 28 avril 1995

Modèle Linéaire Statistique

F. Labé

Introduction

Le modèle linéaire est l'un des outils essentiels de la statistique.

"Tout le monde" connaît l'existence la régression, l'analyse de la variance, l'analyse de la covariance. Le fait que ces "techniques" soient associées à des modèles qui sont tous des cas particuliers du modèle linéaire statistique en général est par contre beaucoup moins connu. L'objet de cette semaine de formation est de présenter ce qu'est un modèle linéaire statistique, en vue de mettre à disposition un outil dont le pouvoir de représentation est extrèmement large.

Le modèle linéaire statistique est un outil de représentation. On ajuste des modèles, on estime des paramètres, on fait des tests, des intervalles et des régions de confiance. Mais que représente-t-on et comment le fait-on ? Que signifie le terme linéaire ? Peut-on rendre compte de relations qui ne sont pas des droites ? Tenir compte d'effets de seuil ? Les résultats classiques associés aux modèles linéaires restent ils valables –et sous quelles conditions– dans d'autres contextes, non linéaires, non gaussiens etc. ?

L'objet de cette semaine de formation est aussi d'apporter des réponses à de telles questions.

La présentation se fera en neuf parties.

1. La forme générale d'un modèle linéaire.
2. L'estimation des paramètres
3. Intervalles et régions de confiance
4. Tests d'hypothèses linéaires
5. Quelques résultats associés à l'analyse de variance
6. Robustesse
7. Un exemple de modèle moins "classique"
8. Modèles à effets fixes et/ou aléatoires
9. Quelques prolongements...

Un exemple de traitement général d'un jeu de données est donné en annexe. Il s'agit de la description de la distribution de tailles de crevettes en Casamance au Sénégal, conditionnellement à des informations sur la vitesse du courant et la salinité.

Cet exemple est traité avec le logiciel SAS, il est présenté avec le programme.

Trois exemples sont par ailleurs donnés à l'issue des parties 5 (analyse de variance), 7 (exemple de modèle "moins classique") et 8 (modèles à effets fixes et/ou aléatoires).

Ces trois exemples sont traités avec des données simulées et avec un logiciel (Genstat) relativement peu connu. Leur intérêt ne réside donc pas en la présentation d'un logiciel particulier, mais plutôt dans la présentation de "sorties" de programmes d'intérêt général, dans lesquelles on peut retrouver les informations très classiques asssociées à un modèle linéaire en général...

1 La forme générale d'un modèle linéaire.

Un modèle statistique s'intéresse à la description, la caractérisation de distributions de variables aléatoires. Pour présenter une partie des notations qui seront utilisées tout long de cette semaine, un premier exemple, le plus simple possible peut être donné.

1.1 Un premier exemple simple

On sélectionne n individus dans une population. La sélection est faite selon une procédure d'échantillonnage aléatoire simple avec remise. On mesure une quantité quelconque, la taille de chacun de ces individus par exemple.

Chaque mesure est la réalisation d'une variable aléatoire $Y_i, i = 1 \cdots n$ Toutes les variables Y_i sont indépendantes, elles ont la même distribution que l'on peut apprécier à partir d'un histogramme des observations.

Si on admet (suppose) que cette distribution est normale on peut écrire :

$$Y_i \sim \mathcal{N}(m, \sigma^2) \quad i = 1 \cdots n$$

Dans cette équation m est l'espérance de la distribution et σ^2 sa variance. Cette équation ne fait pas apparaître le fait que les distributions des Y_i sont indépendantes. Pour celà, il est nécessaire d'écrire que les covariances entre les Y_i sont nulles. La matrice de variance covariance des distributions des Y_i doit donc être diagonale, les valeurs de la diagonale étant toutes égales à σ^2 (rappelons que la matrice de variance covariance de n variables est une matrice carrée à n lignes et n colonnes dont le terme de la ligne i et de la colonne j est la covariance entre les variables Y_i et Y_j).

On peut donc écrire de façon préfèrable :

$$Y \sim \mathcal{N}_n(m, I_n\sigma^2)$$

Où I_n est la matrice identité d'ordre n (des 1 sur la diagonale et des 0 en dehors). Cette notation est matricielle, Y est un vecteur colonne des n variables $Y_1, Y_2 \cdots Y_n$ m est également un vecteur colonne de n valeurs toutes égales et $\sigma^2 I_n$ est la matrice de variance covariance des variables Y

Dans ce modèle très simple, on remarque qu'il n'y a que deux paramètres inconnus, m et σ^2. On peut présenter séparément la description de l'espérance des Y_i et celle de leur composante aléatoire à l'aide de σ^2. L'idée est alors d'écrire que chaque Y_i est la somme de son espérance et d'une variable aléatoire qui est alors d'espérance nulle et qui ne contient donc plus que de l'information sur la variabilité des Y :

$$Y = m + \varepsilon \; ; \quad \varepsilon \sim \mathcal{N}_n(0, I_n\sigma^2)$$

ε est ici un vecteur colonne de n variables indépendantes de moyennes nulles et de variances égales à σ^2

Ce modèle peut être écrit de façon explicite sous forme matricielle :

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} m \\ m \\ \vdots \\ m \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

A partir de deux exemples un peu moins caricaturaux, la forme générale du modèle linéaire peut être précisée

1.2 Une régression linéaire simple

On suppose ici que l'espérance de chaque Y_i est fonction de la valeur prise par une quantité connue x_i, que cette fonction s'exprime selon l'équation d'une droite

$$E(Y|_x) = a \times x + b$$

Cette formule ne constitue en aucun cas un modèle statistique satisfaisant, puisqu'elle ne fait état d'aucune information sur la distribution des Y, si ce n'est leur espérance.

Il est nécessaire en fait d'écrire par exemple :

$$Y_i \sim \mathcal{N}(a \times x_i + b,\ \sigma^2)$$

ou de façon équivalente

$$Y_i = a \times x_i + b + \varepsilon_i\ ;\quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

Dans les deux équations ci dessus il n'apparaît pas d'hypothèse sur l'indépendance des Y_i, ceci peut être fait par la formulation suivante faisant de nouveau intervenir une notation matricielle :

$$Y = a \times x + b + \varepsilon\ ;\quad \varepsilon \sim \mathcal{N}_n(0, I_n\sigma^2)$$

Dans cette formule x est un vecteur colonne contenant les n valeurs prises par x.

Dans l'équation ci-dessus, on peut observer que toutes les espérances des distributions des Y_i se présentent sous la forme d'une combinaison linéaire des deux paramètres a et b. chacune des ces combinaisons peut s'écrire sous la forme du produit matriciel d'un vecteur ligne contenant les coefficients de la combinaison par le vecteur colonne constitué des deux valeurs a et b :

$$E(Y_i) = \begin{pmatrix} x_i & 1 \end{pmatrix} \times \begin{pmatrix} a \\ b \end{pmatrix} = a \times x_i + b$$

En "empilant" ces produits on obtient l'équation matricielle explicite suivante :

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{pmatrix} \times \begin{pmatrix} a \\ b \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

1.3 Analyse de variance à un facteur

Les observations sont maintenant des réalisations de variables aléatoires Y_{ik}. L'indice i indique l'appartenance à l'une des p catégories d'une variable qualitative. n_i observations sont faites dans la catégorie i, l'indice k indiquant un numéro d'ordre. Y_{ik} est donc la variable aléatoire dont la

$k^{\text{ième}}$ observation dans la catégorie i est une réalisation. Toutes les variables Y_{ik} sont supposées normales, indépendantes de variance σ^2 et d'espérance m_i. On peut donc écrire :

$$Y_{ik} = m_i + \varepsilon_{ik} \; ; \quad i = 1 \cdots p \; ; \quad k = 1 \cdots n_i \; ; \quad n = \sum_{i=1}^{p} n_i$$

avec $\varepsilon \sim \mathcal{N}_n(0, I_n\sigma^2)$

Les espérances des Y_{ik} sont comme dans l'exemple précédent des combinaisons linéaires de paramètres (ici $m_1, m_2 \cdots m_p$) on a par exemple pour chacune des observations de la deuxième catégorie :

$$E(Y_{2k}) = 0 \times m_1 + 1 \times m_2 + 0 \times m_3 + \cdots + 0 \times m_p$$

En empilant comme dans l'exemple précédent ces combinaisons on obtient l'équation matricielle explicite présentée ci dessous dans le cas où il y aurait trois catégories ($p = 3$) avec deux observations au sein des deux premières catégories et trois dans la troisième ($n_1 = n_2 = 2$, $n_3 = 3$)

$$\begin{pmatrix} Y_{11} \\ Y_{12} \\ Y_{21} \\ Y_{22} \\ Y_{31} \\ Y_{32} \\ Y_{33} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} \times \begin{pmatrix} m_1 \\ m_2 \\ m_3 \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{31} \\ \varepsilon_{32} \\ \varepsilon_{33} \end{pmatrix}$$

Nous pouvons observer maintenant que le premier modèle très simple présenté peut aussi s'écrire :

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \times \begin{pmatrix} m \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

1.4 Analyse de la covariance

On peut maintenant combiner analyse de variance et régression au sein de modèles souvent appelés "analyse de la covariance". On peut ainsi supposer que la relation sous forme de droite entre l'espérance de Y et une quantité x peut ne pas être la même selon les catégories définies par une variable qualitative. On peut ainsi, avec l'exemple précédent comportant trois catégories et au total 7 observations, rechercher trois relations en écrivant :

$$Y_{ik} = a_i \times x_{ik} + b_i + \varepsilon_{ik}$$

toujours avec les mêmes hypothèses sur la distribution des variables aléatoires ε_{ik}.

Nous avons maintenant six paramètres, trois pentes et trois ordonnées à l'origine, avec l'écriture matricielle complète suivante :

$$\begin{pmatrix} Y_{11} \\ Y_{12} \\ Y_{21} \\ Y_{22} \\ Y_{31} \\ Y_{32} \\ Y_{33} \end{pmatrix} = \begin{pmatrix} x_{11} & 1 & 0 & 0 & 0 & 0 \\ x_{12} & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & x_{21} & 1 & 0 & 0 \\ 0 & 0 & x_{22} & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & x_{31} & 1 \\ 0 & 0 & 0 & 0 & x_{32} & 1 \\ 0 & 0 & 0 & 0 & x_{33} & 1 \end{pmatrix} \times \begin{pmatrix} a_1 \\ b_1 \\ a_2 \\ b_2 \\ a_3 \\ b_3 \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{31} \\ \varepsilon_{32} \\ \varepsilon_{33} \end{pmatrix}$$

De tous les exemples présentés ressort une expression commune

$$Y = X \times \theta + \varepsilon$$

où

- Y est un vecteur de n variables aléatoires indépendantes, gaussiennes, de même variance σ^2
- X est une matrice de n lignes et p colonnes de quantités connues (non aléatoires)
- θ est un vecteur colonne de p valeurs, des paramètres, permettant à travers le produit $X \times \theta$ de donner les espérances de chacune des variables Y_{ik}.
- ε est un vecteur colonne de n variables aléatoires indépendantes gaussiennes d'espérance nulle et de variance σ^2. σ^2 est donc également un paramètre du modèle.

Il s'agit là de la forme générale d'un modèle linéaire gaussien

(en fait on peut aussi réunir sous cette appellation quelques extensions, en supposant l'existence de corrélations entre les variables ou des variances non toutes égales...)

Les quelques exemples ci-dessus laissent entrevoir que ce modèle permet de rendre compte d'un très grand nombre de situations. Nous en rencontrerons d'autres plus loin. *Ces exemples, en particulier celui de la régression linéaire simple montrent aussi que le terme linéaire n'est pas associé à l'idée de droite rectiligne, mais à celle de représentation des espérances des variables aléatoires à l'aide de combinaisons* linéaires *de paramètres.* en représentant par exemple l'espérance d'une variable aléatoire par l'équation

$$E(Y_i) = a_O \times 1 + a_1 \times x_i + a_2 \times x_i^2$$

on se réfère à un modèle linéaire (les espérances sont combinaisons linéaires des trois paramètres a_0, a_1, a_2) **même si la relation entre l'espérance de Y et x est celle d'une parabole.**

2 L'estimation des paramètres

2.1 Estimateurs, estimations.

Revenons au tout premier exemple très simple où toutes les variables ont la même distribution, et supposons que 100 personnes fassent chacune, indépendamment les unes des autres une sélection de n individus.

On obtient ainsi 100 échantillons indépendants, et on peut calculer la moyenne de chacun d'entre eux. On a donc 100 moyennes $\overline{Y_1}, \overline{Y_2} \cdots \overline{Y_{100}}$

On peut bien évidemment faire un histogramme de ces 100 moyennes et ainsi pouvoir apprécier la distribution de ces moyennes. Cela signifie que la moyenne calculée sur n observations est elle même une variable aléatoire avec une espérance, une variance etc.

Si l'une des 100 personnes ayant collecté un échantillon désire faire l'estimation de l'espérance de la variable étudiée, elle peut faire la moyenne des n observations qu'elle a réalisée. le résultat qu'elle obtient ainsi est une réalisation d'une variable aléatoire.

Cette variable aléatoire est un *estimateur*

Le résultat obtenu (la réalisation de cette variable estimateur) est une *estimation*

En étant une variable aléatoire, un estimateur possède une distribution pouvant être décrite à partir de caractéristiques parmi lesquelles son espérance et sa variance.

Si l'espérance de l'estimateur est égale à la valeur que l'on cherche à estimer, l'estimateur est dit sans biais.

Parmi tous les estimateurs sans biais possibles, celui ayant la plus petite variance est qualifié "estimateur sans biais de variance minimum"

Il n'est pas acquis qu'un estimateur sans biais soit celui dont l'erreur quadratique soit en moyenne la plus petite comme le montre le petit exemple suivant :

Admettons que l'estimateur sans biais de variance minimum d'un paramètre m ait une variance supérieure à m^2 (*ie* le coefficient de variation est supérieur à 100%). L'erreur quadratique moyenne associée à cet estimateur est égale à sa variance et est donc supérieure à m^2.

Prenons maintenant l'estimateur constant "0" (j'affirme, quelles que soient les circonstances que m est égal à 0). Cet estimateur est de variance nulle, mais il est biaisé, la valeur du biais étant $0 - m = -m$. L'erreur quadratique est donc toujours égale à m^2.

L'estimateur "0" a donc erreur quadratique moyenne inférieure à l'estimateur sans biais.

L'intérêt de cet exemple est ici de montrer qu'il n'y a pas de de façon unique et définitivement meilleure que toutes les autres pour parler d'un estimateur.

2.2 Qu'estime-ton ?

Les exemples donnés pour illustrer la forme générale d'un modèle linéaire font apparaître deux choses bien distinctes.

- Des suppositions sont faites sur la nature des distributions des Y. Selon ces suppositions les variables Y ont été considérées gaussiennes, indépendantes, de même variance et leurs espérances sont supposées être des combinaisons linéaires de p paramètres réunis dans un vecteur (une liste) θ. Les poids de ces combinaisons sont connus, des 0, des 1 ou des x dans les exemples donnés. Il convient évidemment de vérifier autant que faire se peut la véracité de ces suppositions à partir des données dont on dispose.
- Les valeurs exactes des $p+1$ paramètres (θ et variance σ^2) ne sont pas connues. Nous chercherons donc à les estimer, *en admettant vraies les suppositions faites sur la nature des distributions des* Y.

Bien entendu, on ne se restreindra pas aux seules estimations de θ et σ^2, mais on il est clair que tout ce qui pourra être estimé sera en définitive fonction de ces estimations.

2.3 Estimation par maximum de vraisemblance

La solution de bon sens généralement adoptée est de choisir un estimateur dont la réalisation (l'estimation) donne une vraisemblance –une probabilité d'apparaître– maximum aux informations disponibles (à l'échantillon effectivement observé).

Imaginons que nous ne disposions que d'une observation y, réalisation d'une variable gaussienne Y d'espérance m et variance σ^2. On désire estimer m. Si l'espérance de Y était égale à une valeur donnée m_0 la vraisemblance de y, appréciée par la fonction de densité de la loi de Y serait, en fonction de y et m_0 :

$$L(y, m_0) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{\frac{-(y-m_0)^2}{2\sigma^2}}$$

Pour une valeur donnée quelconque de σ^2, cette vraisemblance est maximum pour $m_0 = y$. On choisit donc $m_0 = y$ comme estimation du maximum de vraisemblance pour m. Cette estimation est donc égale à la réalisation connue de Y, l'estimateur de m est donc la variable Y. Cet estimateur est sans biais, par définition. Sa variance est égale à σ^2.

L'intérêt de cet exemple est de faire apparaître que l'estimateur de m est bien une variable aléatoire (puisqu'il s'agit de la variable Y elle même) et que la réalisation disponible de Y, la valeur *connue* y, est l'estimation de m

En général on notera les estimateurs par des majuscules surmontées d'un chapeau, et les estimations par des minuscules également surmontées d'un chapeau :

- $\widehat{M}$ est l'estimateur de m, vraie valeur inconnue (dans l'exemple ci-dessus, $\widehat{M} = Y$)
- $\widehat{m}$ est une estimation de m, réalisation connue de la variable aléatoire $\widehat{M}$; ($\widehat{m} = y$ dans l'exemple ci-dessus).

On peut considérer l'exemple un peu moins caricatural où l'on dispose de n observations $y_1, y_2 \cdots y_n$ de variables gaussiennes indépendantes $Y_1, Y_2 \cdots Y_n$ de même espérance m et de même variance σ^2 La vraisemblance de l'échantillon est le produit des vraisemblances de chaque observation séparée (l'hypothèse d'indépendance joue ici son rôle essentiel) :

$$L(y_1, y_2 \ldots y_n, m_0) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(y_i - m_0)^2}{2\sigma^2}}$$

Le logarithme de cette vraisemblance est à une valeur indépendante de m_0 près :

$$-\sum_{i=1}^{n}(y_i - m_0)^2$$

La valeur de m_0 qui maximise cette somme de carrés multipliée par -1 est donc celle qui minimise la somme de carrés :

$$\sum_{i=1}^{n}(y_i - m_0)^2$$

En dérivant cette somme par rapport à m_0 on montre qu'elle est maximum si m_0 est égal à la moyenne simple des y_i. On a donc :

$$\widehat{M} = \sum_{i=1}^{n} Y_i/n$$

et

$$\widehat{m} = \sum_{i=1}^{n} y_i/n$$

L'estimateur du maximum de vraisemblance est donc constitué de la moyenne simple des observations, et il correspond à l'estimateur des moindres carrés. Ce résultat est obtenu en faisant intervenir toutes les suppositions du modèle, égalité des variances, indépendance des variables et loi de Gauss. Nous verrons que lorsque certaines de ces suppositions sont fausses, l'usage de l'estimateur des moindres carrés peut être plus ou moins justifié selon les cas.

On peut montrer que l'espérance de $\widehat{M}$ est égale à $n.m/n = m$ ($\widehat{M}$ est un estimateur sans biais) et sa variance est égale à σ^2/n

2.3.1 Application au modèle linéaire général

On pourrait bien entendu, dans chaque cas présenté en illustration de la forme général d'un modèle linéaire expliciter les estimateurs de θ... Et le refaire dans chaque nouveau cas rencontré.

Il est préférable de tirer profit de la forme générale du modèle. L'estimateur du maximum de vraisemblance de θ est dans le cas gaussien équivalent à celui des moindres carrés. L'espérance de Y étant égale à $X \times \theta$, l'estimateur de l'espérance de Y peut s'écrire $\widehat{Y} = X \times \widehat{\theta}$[1] dès lors qu'on dispose d'un estimateur $\widehat{\theta}$ de θ. Le principe est donc de rechercher la valeur concrète des paramètres qui rend minimum la somme de carrés des écarts entre valeurs observées et valeurs ajustées.

[1] L'estimateur de θ sera noté $\widehat{\theta}$, sans recourrir à l'usage d'une lettre majuscule $\widehat{\Theta}$ qui alourdirait trop ici la notation.

En notant X' la transposée de X, on montre que ces valeurs concrètes sont égales au produit matriciel :

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

où y est le vecteur des réalisations des Y. On a donc

$$\widehat{\theta} = (X'X)^{-1}X'Y$$

et

$$\widehat{Y} = X\widehat{\theta} = X(X'X)^{-1}X'Y$$

comme estimateurs des paramètres et des espérances des Y

On peut illustrer ces résultats sur quelques uns des exemples donnés en introduction.

Dans le cas de l'analyse de la variance à 1 facteur, la matrice $X'X$ est une matrice diagonale, de dimension (p,p), le $i^{\text{ième}}$ terme de la diagonale étant égal à n_i. La matrice $(X'X)^{-1}$ est également diagonale, les termes diagonaux étant les $1/n_i$.

Le produit $(X'X)^{-1}X'y$ est quant à lui égal au vecteur des moyennes des observations par catégories.

$\widehat{\theta}$ est donc le vecteur des p variables $1/n_i \sum_{k=1}^{n_i} Y_{ik}$

Il est tout à fait important de "voir" que $\widehat{\theta}$ et $X\widehat{\theta}$ sont des combinaisons linéaires des variables Y. Connaissant la matrice de covariance de Y $(I_n\sigma^2)$ et les poids de ces combinaisons, on a

$\text{var}(\widehat{\theta}) = (X'X)^{-1}X'\text{var}(Y)X(X'X)^{-1}$

soit $\text{var}(\widehat{\theta}) = \sigma^2 I_n(X'X)^{-1}(X'X)(X'X)^{-1}$

soit enfin $\text{var}(\widehat{\theta}) = \sigma^2(X'X)^{-1}$

Ce résultat matriciel est très intéressant. Il contient l'information sur la variance des estimateurs $\widehat{\theta}$ et sur les covariances entre les divers paramètres de θ. Celà nous permettra d'estimer la variance de n'importe quelle combinaison linéaire des éléments de θ

La distribution des Y étant normale, toute combinaison linéaire des éléments de Y est elle-même une variable normale. On peut montrer enfin que l'espérance de $\widehat{\theta}$ est égale à θ (estimateur sans biais).

Toutes ces propriétés se résument selon les caractéristiques de la distribution de la variable aléatoire $\widehat{\theta}$:

$$\widehat{\theta} \sim \mathcal{N}_p(\theta, (X'X)^{-1}\sigma^2)$$

2.4 Estimation de σ^2

σ^2 est la variance des ε, il s'agit donc de l'espérance des carrés des écarts entre les observations et l'espérance des variables dont ces observations sont des réalisations.

Une estimation des ε est donc simplement obtenue par la différence $Y - \widehat{Y} = Y - X\widehat{\theta}$. La moyenne des $\widehat{\varepsilon}$ étant nulle, un estimateur naturel de σ^2 est la moyenne des carrés des écarts entre valeurs observées Y et estimées $\widehat{Y}$.

On peut montrer que cet estimateur est en fait biaisé et qu'en divisant la somme des carrés des écarts par $n - p$ au lieu de n on obtient un estimateur sans biais de σ^2 :

$$\widehat{\sigma^2} = \frac{\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2}{n-p}$$

Par ailleurs, la loi de cet estimateur est proportionnelle à celle d'un χ^2 :

$$\widehat{\sigma^2} \sim \sigma^2 \frac{\chi^2_{n-p}}{n-p}$$

(Une loi de χ^2_n est la loi d'une somme de n variables $\mathcal{N}(0,1)$ indépendantes élévées au carré).

Tous les résultats présentés jusqu'ici sont valables pour le modèle linéaire gaussien. On montre pour ce modèle que les estimateurs possèdent de très bonnes qualités. $\widehat{\theta}$ est ainsi l'estimateur sans biais de variance minimum.

Si le modèle n'est pas gaussien si toutes les autres suppositions sont satisfaites (indépendance, espérances combinaisons linéaires des paramètres et égalité des variances), $(X'X)^{-1}X'Y$ est encore un estimateur sans biais, dont la variance est minimum parmi tous les estimateurs se présentant sous la forme de combinaisons linéaires des Y.

3 Intervalles de confiance sur un paramètre ou une combinaison linéaire de paramètres

Si σ^2 était connu, il en serait de même pour $(X'X)^{-1}\sigma^2$; On connaitrait donc la variance de l'estimateur de chaque élément de θ et de toute combinaison linéaire de ces éléments, permettant la construction immédiate d'intervalles de confiance. En recourrant à l'estimateur $\widehat{\sigma^2}$ de σ^2, pour n'importe quel élément de θ, la pente a ou l'ordonnée à l'origne b dans la régression simple ou l'espérance m_i relative à un niveau i dans l'analyse de variance d'un modèle à un facteur de variation, la recherche d'un intervalle de confiance reste relativement simple.

Ainsi dans le cas de la régression simple, $\widehat{\theta}$ est la variable aléatoire de dimension $p = 2$ réunissant les deux estimateurs $\widehat{A}$ et $\widehat{B}$ de a et de b. La loi de $\widehat{A}$ est normale, d'espérance a et de variance $(X'X)^{-1}_{11}\sigma^2 = \text{var}(\widehat{A})$. En centrant et réduisant $\widehat{A}$, on obtient donc une variable aléatoire $\mathcal{N}(0,1)$:

$$\frac{\widehat{A} - a}{\sqrt{\sigma^2_{\widehat{A}}}} \sim \mathcal{N}(0,1)$$

Ce résultat est valable pour n'importe quel élément θ_i de θ.

Mais σ^2 est inconnu et est remplacé par son estimateur. On a donc :

$$\frac{\widehat{\theta_i} - \theta_i}{\sqrt{\widehat{\sigma^2_{\widehat{\theta_i}}}}} \sim \frac{\mathcal{N}(0,1)}{\sqrt{\frac{\chi^2_{n-p}}{n-p}}}$$

c'est-à-dire

$$\frac{\widehat{\theta_i} - \theta_i}{\sqrt{\widehat{\sigma^2_{\widehat{\theta_i}}}}} \sim T_{n-p}$$

T_{n-p} est une loi de Student à $n-p$ degrés de libertés. Par définition, une distribution T_{n-p} est le rapport d'une loi $\mathcal{N}(0,1)$ et de la racine carrée d'une loi de $\chi^2_{n-p}/(n-p)$, indépendantes.

L'intervalle de confiance sur θ_i est l'ensemble des valeurs $c\theta_i$ telles que

$$\frac{|\widehat{\theta_i} - c\theta_i|}{\sqrt{\widehat{\sigma^2_{\widehat{\theta_i}}}}} < t_{\alpha,(n-p)}$$

où $\widehat{\theta_i}$ est la réalisation (estimation) de l'estimateur du paramètre θ_i et où $t_{\alpha,(n-p)}$ est le seuil de rejet au niveau α qu'on peut lire dans une table de distribution de variables de Student.

On peut également s'intéresser à nimporte quel paramètre se présentant sous la forme d'une combinaison linéaire des éléments de θ. Ainsi, si on s'intéresse à l'espérance de la variable Y pour une valeur donnée x_o de la variable explicative, on s'intéresse à une combinaison liéaire $ax_o + b$ de a, b, dont les poids sont x_o et 1

$$\frac{(a\widehat{x_o+b})-(ax_o+b)}{\sqrt{\widehat{\sigma^2_{(ax_o+b)}}}} \sim T_{n-2}$$

La variance de $(a\widehat{x_o+b})$ est égale à

$$x_o^2 \mathrm{var}\widehat{A} \; + \; 2x_o \mathrm{cov}(\widehat{A}\widehat{B}) \; + \; \mathrm{var}\widehat{B}$$

Sachant que

$$\sigma^2(X'X)^{-1} = \begin{pmatrix} \mathrm{var}(\widehat{A}) & \mathrm{cov}(\widehat{A},\widehat{B}) \\ \mathrm{cov}(\widehat{A},\widehat{B}) & \mathrm{var}(\widehat{B}) \end{pmatrix} = \frac{\sigma^2}{n\sum_{i=1}^n (x_i-\overline{x})^2} \begin{pmatrix} n & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix}$$

On a donc

$$\widehat{\sigma^2_{(ax_o+b)}} = \widehat{\sigma^2}\frac{\sum_{i=1}^n x_i^2 + nx_o^2 - 2n\overline{x}x_o}{n\sum_{i=1}^n (x_i-\overline{x})^2} = \widehat{\sigma^2}h(x_o)$$

L'intervalle de confiance au niveau α sur ax_o+b est donc l'ensemble des valeurs v telles que

$$|v-(a\widehat{x_o+b})| < t_{\alpha,n-2}\sqrt{\widehat{\sigma^2}h(x_o)}$$

Il faut bien remarquer que l'intervalle de confiance présenté ici est relatif à *l'espérance* de Y pour $x=x_o$. Si l'on s'intéresse à un intervalle sur la valeur que peut prendre une *réalisation* de Y pour $x=x_o$, il convient de tenir compte de la variabilité de ε qui s'ajoute à celle de l'estimateur de $a\times\widehat{x_o}+b$...

4 Tests d'hypothèses linéaires

Si l'on fait un intervalle sur la pente de la régression simple par exemple, on peut s'intéresser au fait de savoir si une valeur particulière est ou non dans l'intervalle. Il s'agira en général de la valeur 0 pour la pente, indiquant alors que la quantité x n'apporte pas d'information sur la distribution de Y dans le cadre du modèle postulé. Dans certains cas on peut s'intéresser à un ensemble de paramètres. Ainsi, dans l'analyse de variance à un facteur avec p niveaux, on s'intéressera d'abord au fait de savoir si la source de variation asssociée au facteur étudié apporte de l'information, ce qui revient à dire que les espérances des Y dépendent du niveau du facteur. On s'intéresse donc alors à l'hypothèse H_o : $m_i = m \quad \forall i$.

L'hypothèse ainsi faite ne porte pas sur une combinaison linéaire de paramètres mais sur plusieurs combinaisons linéaires. Si il y a p niveaux, on peut montrer qu'il faut annuler au moins $p-1$ combinaisons de paramètres pour se trouver dans le domaine de l'hypothèse H_o :

Il faut par exemple que m_1 soit égal à tous les $m_i, i = 2 \cdots p$. Celà entraîne la nullité de toutes les différences $(m_i - m_j)$ puisque $(m_i - m_j) = (m_1 - m_j) - (m_1 - m_i) = 0$

L'hypothèse H_o revient donc à supposer la nullité d'au moins $p-1$ combinaisons linéaires de θ ; on dit alors qu'elle est une hypothèse linéaire de dimension $p-1$.

Dans l'exemple donné en introduction de l'analyse de la covariance avec un facteur à 3 niveaux, on peut s'intéresser à l'hypothèse selon laquelle les trois pentes a_1, a_2, a_3 sont égales. Cette hypothèse nécessite d'annuler au moins deux combinaisons linéaires de θ : $a_1 - a_2 = a_1 - a_3 = 0$, il s'agit donc d'une hypothèse linéaire de dimension 2.

D'une façon générale, à partir d'un modèle, supposé correct, comportant p paramètres, l'hypothèse linéaire revient à envisager un modèle plus simple, se déduisant du premier (c'est-à-dire inclus dans le premier) par l'annulation de certains paramètres ou combinaisons de paramètres. Le modèle simplifié est alors un modèle linéaire à $p-h$ paramètres, h étant la dimension de l'hypothèse.

L'hypothèse H_o testée sera rejetée lorsque la qualité de l'ajustement obtenu par le modèle simplifié est "significativement" moins bonne que celle obtenue avec le modèle initial supposé correct.

La qualité de l'ajustement est appréciée par la somme des carrés des écarts entre les valeurs observées et les valeurs ajustées par le modèle. La comparaison des divers modèles se fait donc à partir de leurs sommes de carrés résiduelles.

En notant SR_1 et SR_2 ces sommes de carrés pour le modèle initial (1) et le modèle simplifié (2), on montre que, sous l'hypothèse H_o considérée :

$$\frac{(SR_2 - SR_1)/h}{SR_1/(n-p)} \sim F_{h,(n-p)}$$

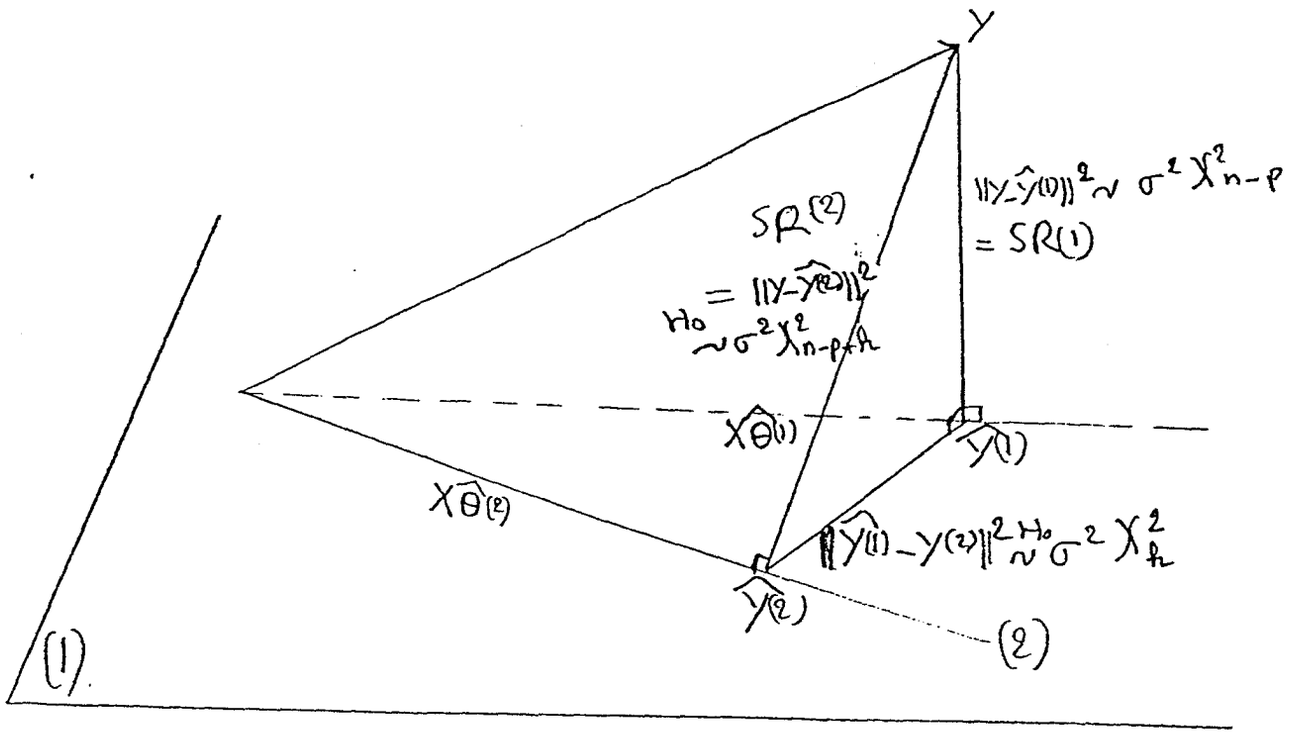
où $F_{h,(n-p)}$ est une loi de Fisher à h et $n-p$ degrés de libertés.

(Une loi de Fisher à n_1 et n_2 degrés de liberté est un rapport

$$\frac{\chi^2_{n_1}/n_1}{\chi^2_{n_2}/n_2},$$

les deux χ^2 étant indépendants.

Une représentation géométrique permet d'illustrer "simplement" ce résultat.



Cette représentation fait apparaître les points importants suivants :

- Le vecteur Y est dans un espace de dimension n. On projette Y sur le "plan" (1) de dimension p associé au modèle supposé correct. La projection orthogonale sur le plan correspond à l'estimation $\widehat{Y^{(1)}}$ de Y par les moindres carrés. La somme ST des carrés des Y est donc décomposée en une somme de carrés expliqués $SC_{(1)}$ et une somme de carrés résiduels $SR_{(1)}$ (théorème de pythagore). On réalise la même opération pour le modèle (2) qui se déduit de (1) par des contraintes linéaires et qui est donc un sous espace de dimension $p-h$ représenté par la "droite" (2) au sein du "plan" (1). On obtient ainsi la décomposition de ST selon la somme de $SC_{(2)}$ et $SR_{(2)}$,

- les projections de Y sur (1) et (2) sont les points $\widehat{Y^{(1)}}$ et $\widehat{Y^{(2)}}$. Le triangle rectangle (théorème des trois perpendiculaires) $Y\widehat{Y^{(1)}}\widehat{Y^{(2)}}$ indique (théorème de Pythagore) que la somme de carrés résiduelle $SR_{(2)}$ est égale à la somme de $SR_{(1)}$ et du carré de la distance entre $\widehat{Y^{(1)}}$ et $\widehat{Y^{(2)}}$ qui est donc égal à $SR_{(2)} - SR_{(1)}$

- $SR_{(1)}$ suit une loi de $\sigma^2\chi^2_{n-p}$.
  Sous l'hypothèse H_o testée, $SR_{(2)}$ suit une loi de $\sigma^2\chi^2_{n-p+h}$.
  Ce $\sigma^2\chi^2_{n-p+h}$ se décompose selon une somme de deux $\sigma^2\chi^2$ indépendants de degrés de libertés $n-p$ et h correspondant à $SR_{(1)}$ et à la différence $SR_{(2)} - SR_{(1)}$. Cette différence suit donc une loi de $\sigma^2\chi^2_h$

- Sous H_o, le rapport de ces deux $\sigma^2\chi^2$ divisés par leurs degrés de libertés respectifs

$$\frac{(SR_2 - SR_1)/h}{SR_1/(n-p)}$$

  suit, par définition, une loi de Fisher de degrés de liberté h et $n-p$

- Si H_o est fausse, la différence entre les deux résiduelles contient de l'information relative à la structure du modèle. Le numérateur du rapport

$$\frac{(SR_2 - SR_1)/h}{SR_1/(n-p)}$$

  ne suit plus une loi de χ^2 divisée par son nombre de degrés de libertés, mais quelques chose de plus grand, prenant en compte les carrés des erreurs commises en supposant vraie l'hypothèse H_o alors qu'elle est fausse.

4.1 Des intervalles aux régions de confiance

Nous n'avons envisagé dans la partie précédente que des intervalles de confiance relatifs à un seul paramètre ou à une seule combinaison de paramètres. Ainsi, l'intervalle relatif à l'espérance de Y pour x_o donné dans la régression linéaire simple :

$$(v) \quad \text{tels que} \quad |v - (a\widehat{x_o} + b)| < t_{\alpha,n-2}\sqrt{\widehat{\sigma^2}h(x_o)}$$

ne vaut que pour le x_o donné. Si l'on cherche un intervalle valable pour n'importe qu'elle valeur de x, la question porte en fait sur une région de dimension 2, contenant des points ordonnées et

abscisses. La question est donc analogue à celle posée dans un test d'une hypothèse de dimension $h = 2$.

On a alors une région de confiance au niveau $(1 - \alpha)$ réunissant les points de coordonnées :

$$(v, x) \quad \text{tels que} |v - (\widehat{ax + b}| < \sqrt{2 f_{\alpha,2,n-2} \sigma^2 h(x)}$$

Dans le cas d'un x_o particulier, on avait :

$$(v) \quad \text{tels que} |v - (\widehat{ax_o + b})| < \sqrt{1 f_{\alpha,1,n-2} \sigma^2 h(x)}$$

$t_{\alpha,n-2}$ est en effet égal à $\sqrt{1 f_{\alpha,1,n-2}}$

Cette difficulté est associée à un problème assez connu dans le cas de l'analyse de variance d'un modèle à un facteur de variation lorsqu'on s'intéresse à la question de savoir quelles sont les niveaux du facteur dont les moyennes sont effectivement différentes après avoir rejeté l'hypothèse selon laquelle les moyennes ne sont pas toutes égales.

En effet, imaginons que toutes les moyennes soient égales. S'il y a p niveaux, donc p espérances m_i, on peut réaliser $p(p-1)/2$ tests de type $m_i - m_j$ avec un risque d'erreur α. La probabilité de rejeter au moins une des hypothèses d'égalité croît avec p (si on réalise au niveau 5% 100 tests d'hypothèses H_o toutes vraies, on en rejette en moyenne 5 !).

Il convient donc de réaliser des comparaisons multiples de moyennes. Plusieurs méthodes existent. La méthode la plus "proche" de la logique du test F est la méthode de Scheffé :

On rejette l'hypothèse $(m_i - m_j) = 0$ si

$$|\widehat{m_i} - \widehat{m_j}| > \sqrt{(p-1) f_{\alpha,p-1,n-p} \, \widehat{\sigma^2} \, (\frac{1}{n_i} + \frac{1}{n_j})}$$

On remarque que si p=2, on retrouve le test T de comparaison de deux moyennes.

5 Analyse de variance

L'analyse de la variance concerne ici les modèles à une ou plusieurs sources de variation associées à des variables qualitatives.

Dans les logiciels de type "SAS", il est possible de réaliser des analyses de variance soit à partir de procédures spécifiques (proc ANOVA pour les modèles orthogonaux), soit à partir de procédure de type "proc GLM" permettant la mise en œuvre du modèle linéaire général tel que nous le présentons.

Cette situation est justifiée par une difficulté issue du fait que les paramétrisations habituelles associées à ces modèles conduisent à des matrices $(X'X)$ non inversibles. Il convient de donner des contraintes liant les paramètres et il n'est pas toujours possible de choisir de "bonnes" contraintes.

5.1 Les modèles orthogonaux

Prenons par exemple un modèle à deux facteurs de variations avec interactions.

$$Y_{ijk} = m + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$$

où :

- $\varepsilon_{ijk} \sim \mathcal{N}(0, \sigma^2)$ indépendants
- α_i, β_j, γ_{ij} sont des constantes fixes, inconnues
- $Y_{ijk} \sim \mathcal{N}(\alpha_i + \beta_j + \gamma_{ik}, \sigma^2)$ est la variable aléatoire dont une réalisation est la $k^{ième}$ observation faite dans le niveau i du premier facteur et le niveau j du second.
- $i = 1 \cdots p, \quad j = 1 \cdots q, \quad k = 1 \cdots n_{ij}$

  n_{ij} est le nombre d'observations faites dans la combinaison de niveaux i et j

  $n_{oo} = \sum_{i,j} n_{ij}$ est le nombre total d'observations.

  On note $n_{io} = \sum_{j=1}^{q} n_{ij}$ le nombre total d'observations faites dans le niveau i du premier facteur. On note $n_{oj} = \sum_{i=1}^{p} n_{ij}$ le nombre total d'observations faites dans le niveau j du second facteur.

On remarque que les espérances des Y_{ijk} sont toutes égales pour une combinaison ij donnée à une quantité

$$m_{ij} = m + \alpha_i + \beta_j + \gamma_{ik}$$

Il n'y a que $p \times q$ combinaisons ij, $p \times q$ est donc la dimension du modèle. Il y a par contre $1 + p + q + p \times q$ paramètres α ou β ou γ. Il y en a donc $p + q + 1$ "en trop".

On va résoudre ce problème d'entropologie en imposant des contraintes de nullité de combinaisons linéaires des paramètres, définissant un espace de dimension $p + q + 1$

Ces combinaisons peuvent être recherchées de façon à ce que les paramètres soient le plus naturellement interprétables possible. Un solution pour s'approcher d'un tel objectif est de faire en sorte que les estimateurs soient les "plus indépendants" possible, auquel cas les hypothèses que l'on pourra faire sur certains d'entre eux ne modifieront pas l'estimation des autres.

On montre que la condition liant les effectifs n_{ij} :

$$n_{ij} = \frac{n_{io}n_{oj}}{n_{oo}} \quad \forall i,j$$

est une condition nécessaire et suffisante pour qu'il existe un système de contraintes rendant mutuellement indépendants les estimateurs de m, des α, des β et des γ. Le modèle est alors dit "orthogonal".

Ces contraintes sont obligatoirement alors les suivantes :

- $\sum_{i=1}^{p} n_{io}\alpha_i = 0$ (Une contrainte)
- $\sum_{j=1}^{q} n_{oj}\beta_j = 0$ (Une contrainte)
- $\sum_{j=1}^{q} n_{ij}\gamma_{ij} = 0 \quad \forall i$ (p contraintes)
- $\sum_{i=1}^{p} n_{ij}\gamma_{ij} = 0 \quad \forall j$ (q contraintes)

Il y a $p+q+2$ contraintes dans le systèmes ci-dessus, en fait, les $p+q+1$ premères entraînent la dernière.

Le cas particulier des modèles équilibrés (les n_{ij} tous égaux) correspond à un modèle orthogonal très sympathique puisque le système de contraintes est alors :

- $\sum_{i=1}^{p} \alpha_i = 0$
- $\sum_{j=1}^{q} \beta_j = 0$
- $\sum_{j=1}^{q} \gamma_{ij} = 0 \quad \forall i$
- $\sum_{i=1}^{p} \gamma_{ij} = 0 \quad \forall j$

Les estimateurs des divers paramètres m, α, β et γ sont alors assez simples :
$\widehat{m} = Y_{...}$
$\widehat{\alpha_i} = Y_{i..} - Y_{...}$
$\widehat{\beta_j} = Y_{.j.} - Y_{...}$
$\widehat{\gamma_{ij}} = Y_{ij.} - Y_{i..} - Y_{.j.} + Y_{...}$

et la variabilité des écarts à la moyenne générale des Y se décompose en une somme "simple" de quatre sommes de carrés :

$$\underbrace{\sum_{ijk}(Y_{ijk} - Y_{...})^2}_{S} = \underbrace{\sum_{ijk}(Y_{i..} - Y_{...})^2}_{S_1} + \underbrace{\sum_{ijk}(Y_{.j.} - Y_{...})^2}_{S_2} + \underbrace{\sum_{ijk}(Y_{ij.} - Y_{i..} - Y_{.j.} + Y_{...})^2}_{S_3} + \underbrace{\sum_{ijk}(Y_{ijk} - Y_{ij.})^2}_{S_4}$$

La quatrième somme (S_4) est la somme de carrés résiduelle qui suit une loi de $\sigma^2\chi^2_{n-pq}$
La première (S_1) suit une loi de $\sigma^2\chi^2_{p-1}$ sous l'hypothèse de nullité des α_i. La seconde (S_2) suit une loi de $\sigma^2\chi^2_{q-1}$ sous l'hypothèse de nullité des β_j. La troisième (S_3) suit une loi de $\sigma^2\chi^2_{(p-1)(q-1)}$ sous l'hypothèse de nullité des γ_{ij}.

Ces lois de χ^2 sont par ailleurs indépendantes et permettent de construire immédiatement les tests des diverses hypothèses.

Cette décomposition en sommes de carrés est donnée dans les tableaux d'analyse de variance donnés sur les listings en sortie par les logiciels statistiques. Un exemple en est donné dans le tableau ci-dessous reprenant l'analyse de variance d'un modèle équilibré à deux facteurs A et B de variation ayant respectivement p et q niveaux.

source	ddl	SS	MS	F	P
A	$p-1$	S_1	$M_1 = S_1/(p-1)$	M_1/M_4	
B	$q-1$	S_2	$M_2 = S_2/(q-1)$	M_2/M_4	
A.B	$(p-1)(q-1)$	S_3	$M_3 = S_3/(p-1)(q-1)$	M_3/M_4	
résiduelle	$n-pq$	S_4	$M_4 = S_4/(n-pq)$		

On lit dans la colonne "P" la probabilité d'obtenir une réalisation d'une variable de Fisher supérieure à la valeur donnée dans la colonne "F". Concrètement, ceci conduit à des tableaux tels que celui présenté ci-dessous, obtenu avec un logiciel auquel on a demandé une analyse de la variance du modèle équilibré

$$Y_{ijk} = m + \alpha_i + \beta_j + \gamma_{ik} + \varepsilon_{ijk}$$

avec deux facteurs de variation A et B ayant respectivement $p = 3$ et $q = 4$ niveaux avec 5 répétitions pour chaque combinaison.

Les données ont été obtenues par simulation, toutes les variables Y_{ijk} étant gaussiennes, indépendantes de moyenne nulle et de variance égale à 1. Dans ce cas les hypothèses de nullité des coefficients α, β et γ sont toutes vraies et l'on remarque qu'en choisissant un risque de première espèce de 5% on rejette celle de nullité des γ (P=0.024).

Le risque de première espèce est donc bien un risque...

On notera que le logiciel utilisé ici nomme "v.r." (variance ratio) la colonne généralement intitulée "F". De même la colonne "F prob." correspond à l'habituelle colonne "P". Enfin la ligne "residual" est fréquemment appelée "error" dans les autres logiciels...

```
  1  job 'anova2'
  2  unit [60]

DEFINITION DES FACTEURS A ET B

LE FACTEUR K EST LE FACTEUR DE REPETITION

  3  fact [lev=3;nval=60] A
  4  fact [lev=4;nval=60] B
  5  fact [lev=5;nval=60] K
  6  gene A,B,K

SIMULATION DE LA VARIABLE ETUDIEE, SELON DES LOIS NORMALES CENTREES
REDUITES INDEPENDANTES...

  7  calc x=urand (1)
  8  calc x=ned(x)

DEFINITION DU MODELE (A*B), ET ANALYSE DE VARIANCE

  9  treat A*B
 10  anova [fprob=yes]x

10.....................................................................

***** Analysis of variance *****

Variate: x

Source of variation     d.f.       s.s.       m.s.     v.r.  F pr.
A                          2     0.2137     0.1068     0.16  0.852
B                          3     5.5486     1.8495     2.78  0.051
A.B                        6    10.7795     1.7966     2.70  0.024
Residual                  48    31.9561     0.6658
Total                     59    48.4978

***** Tables of means *****


        A         1         2         3
              0.27      0.29      0.16

        B         1         2         3         4
              0.60      0.33      0.27     -0.24

        A         B         1         2         3         4
        1               1.26     -0.13      0.76     -0.78
        2               0.21      0.85     -0.20      0.31
        3               0.34      0.28      0.26     -0.24
```

5.2 Les modèles non orthogonaux

Lorsque la condition d'orthogonalité n'est pas satisfaite, la situation est assez grave. Il n'existe plus de possibilité de recourir aux formules relativement simples données ci-dessus.

En fait, il faut écrire de façon explicite la matrice X du modèle en s'assurant de l'inversibilité de $(X'X)$. On choisit pour celà un système de contraintes arbitraires. Les "grands" logiciels vont ainsi en général considérer que tous les effets associés au premier ou au dernier niveau d'un facteur sont nuls (voir le résultat de programme donné page 32 avec l'absence de l'effet "mois 1" arbitrairement exclu du modèle. Dans ce cas, le paramètre "constant" est la moyenne du premier mois, les effets "mois i", $i = 1 \cdots 12$ sont en fait les différences de moyenne entre le mois 1 et le mois i).

Les logiciels calculent $(X'X)^{-1}$, les différentes choses utiles etc. Il faut bien voir que ces opérations sont très couteuses en calculs et ne peuvent être réalisées sans ordinateurs.

Il ne faut pas croire cependant, que disposer d'un ordinateur résoud la difficulté. S'il est possible de faire les calculs, toutes les difficultés associées à la dépendance des estimateurs restent présentes.

En particulier, la décomposition en sommes de carrés n'est plus valable et les sorties d'ordinateurs sont beaucoup moins sympathiques (voir par exemple les sommes de carrés de type 1 ou de type 3 offertes par proc GLM de SAS).

6 Robustesse

Nous avons défini le modèle linéaire par une équation générale :

$$Y = X \times \theta + \varepsilon$$

avec

$$\varepsilon \sim \mathcal{N}_n(0, \sigma^2 I_n)$$

Cette équation et les suppositions faites sur la distribution conjointe des résidus comportent plusieurs aspects.

- L'espérance des Y est bien représentée par $X \times \theta$, ce qui entraîne la nullité des espérances des ε
- La distribution conjointe des Y et celle des ε sont normales.
- Les ε sont indépendants (ainsi que les Y)
- les variances des ε sont toutes égales.

Si ces suppositions sont toutes faites à bon escient, les résultats présentés jusqu'ici sont valables et les méthodes employées pour y parvenir sont optimales, selon les considérations de bons sens liées aux idées de vraisemblance, d'erreur quadratique mimimale, de risque d'erreur minimal etc.

Si l'une ou plusieurs de ces suppositions ne sont pas valables les choses peuvent plus ou moins bien se passer... selon les cas et selon la nature des questions auxquelles on cherche à répondre.

Il se peut qu'une supposition erronnée n'ait pratiquement aucune importance et que l'usage de la méthode, dont on sait qu'elle est optimale lorsque la supposition faite reflète la réalité, reste tout à fait acceptable. On parlera alors de robustesse.

Mais il se peut qu'une supposition erronnée rende totalement faux ou illusoires les résultats obtenus. On parlera alors de sensibilité.

Il n'est pas vraiment possible de donner une liste des choses qu'on peut ou non accepter. Nous pouvons présenter ici deux exemples permettant d'illustrer cette difficulté, tout en indiquant néanmoins quelques idées de bases sur ces aspects.

Sensibilité du test T de comparaison de deux moyennes à l'inégalité des deux variances intra populations.

Lorsqu'on réalise un test T de comparaisons des moyennes de deux populations, on calcule à partir des observations faites dans chacune des deux populations :

$$\frac{(\widehat{m_1} - \widehat{m_2})}{\widehat{\sigma}(\sqrt{1/n_1 + 1/n_2})}$$

où

- $\widehat{m_1}$ et $\widehat{m_2}$ sont les moyennes observées pour chaque population.

- $\widehat{\sigma}$ et la racine carré de la somme des carrés des écarts entre chaque observation et la moyenne des observations dans la population dont est issue cette observation; cette somme de carrés étant divisée par son nombre de degrés de libertés $(n_1 + n_2 - 2)$.

On peut vérifier sans peine à partir des précédentes parties que tout ceci n'est que la procédure décrite relative à l'analyse de la variance à un facteur, dans le cas où ce facteur a deux niveaux $(p = 2)$

Lorsqu'on demande à un logiciel de réaliser l'analyse de variance de ce modèle, le test d'égalité des deux moyennes proposé en sortie indique le risque d'erreur commis en rejetant l'hypothèse d'égalité si cette hypothèse H_o est vraie (ie $m_1 = m_2$). Ce risque est donné sous l'hypothèse que toutes les suppositions associées au modèle linéaire gaussien sont vraies.

Qu'en est-il de la valeur de ce risque si les deux variances intra populations σ_1^2 et σ_2^2 ne sont pas égales, les autres suppositions restant vraies ?

Le tableau ci dessous (Hsu, 1938 *in* Coursol 1981) donne des indications très intéressantes sur le risque de première espèce d'un test d'égalité de deux moyennes *ie* sur la probabilité de rejeter l'égalité des deux moyennes lorsque celles ci sont égales et que l'on décide de rejeter H_o lorsque la statistique T est supérieur au seuil $t_{\alpha=0.05,n_1+n_2-2}$ lu dans une table.

Lorsque les deux variances sont égales, cette probabilité est effectivement de 0.05 ; lorsque le rapport σ_1^2/σ_2^2 est différent de 1, la probabilité de rejeter H_o n'est plus nécessairement égale à 0.05. Cette probabilité dépend du rapport entre les deux variances, mais aussi des effectifs n_1 et n_2

σ_1^2/σ_2^2 n_1, n_2	0	0.1	0.5	1	2	10	∞
15 5	0.32	0.23	0.10	0.05	0.025	0.005	0.002
5 3	0.22	0.14	0.07	0.05	0.04	0.03	0.03
7 7	0.07	0.07	0.06	0.05	0.05	0.06	0.07

Ainsi on observe que la robustesse du test est relativement acceptable lorsque les effectifs sont égaux, et qu'elle devient illusoire lorsque les effectifs sont très différents...

La robustesse est très largement liée à la question posée. Ainsi l'on sait (théorème central limite) que l'estimateur moyenne a une distribution tendant vers une loi normale lorsque le nombre d'observations tend vers l'infini. Généralement, on considère que si n_1 et n_2 sont plus grands que 30, le test T est acceptable en cas de non normalité des Y (les autres suppositions étant vraies). D'une façon générale la normalité de $\widehat{\theta}$ est asymptotiquement acquise. L'intervalle de confiance donné sur $a \times x_o + b$ est donc peu sensible à la normalité de Y dans l'exemple de la régression linéaire simple.

Par contre l'intervalle de confiance que l'on peut donner sur une réalisation de Y pour $x = x_o$ est lui très sensible puisqu'on s'intéresse alors à une réalisation unique et que la forme de la distribution de Y intervient complètement...

D'une façon générale par contre, la non satisfaction de l'indépendance des observations intervient de façon dramatique. Le problème est alors que l'on ne sait plus quelle est la question à laquelle on répond.

Nous reviendrons sur ce point dans la discussion sur les effets fixes ou aléatoires dans l'analyse de la variance ainsi que dans les modèles à mesures répétées...

Quelques principes pour l'analyse des résidus estimés

Dans le "travaux pratique" donné en annexe sur les données de crevettes en Casamance, plusieurs modèles sont proposés. Le modèle I de départ est un modèle supposant une relation quadratique entre la taille des crevette et la salinité, relation dépendant du niveau de courant de surface.

$$Y_{ik} = a_{o_i} + a_{1_i} \times x_{ik} + a_{2_i} \times x_{ik}^2 \quad + \varepsilon_{ik}$$

où les ε sont supposés être indépendants et suivre des lois normales centrées de variances toutes égales à σ^2.

Ces suppositions associées à ce modèle ne sont pas remises en cause par la suite, les hypothèses testées n'étant relatives qu'à des simplifications conduisant à conserver en définitive le modèle 3 :

$$Yik = a_{o_i} + a_{1_i} \times x_{ik} + \varepsilon_{ik}$$

Avec un regroupement des courants en deux niveaux seulement pour les pentes. ($a_{i_1} = a_{i_2} = a_{i_3}$ et $a_{i_4} = a_{i_5}$). On a donc réduit le nombre de paramètres de 15 à 7).

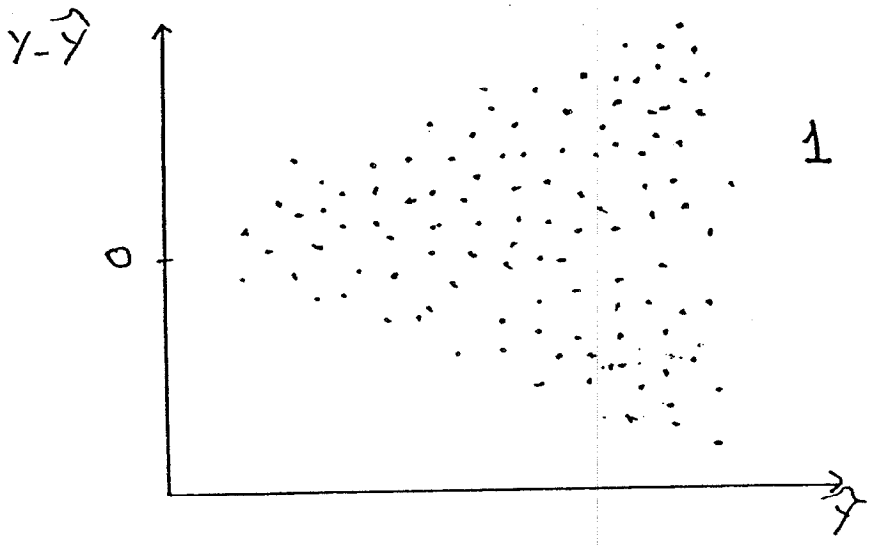
On peut toutefois observer la répartition des résidus estimés pour évaluer leur distribution. On constate que les histogrammes obtenus sont pour tous les ajustements relativement proches de lois normales. On peut observer aussi les graphes des résidus estimés en fonction des valeurs prédites (qui sont orthogonaux comme l'indique le schéma de la page 14). En théorie, les résidus ne doivent pas présenter de tendance ni pour leurs moyennes, ni pour leur variabilité (cf dessins présentés page suivante).

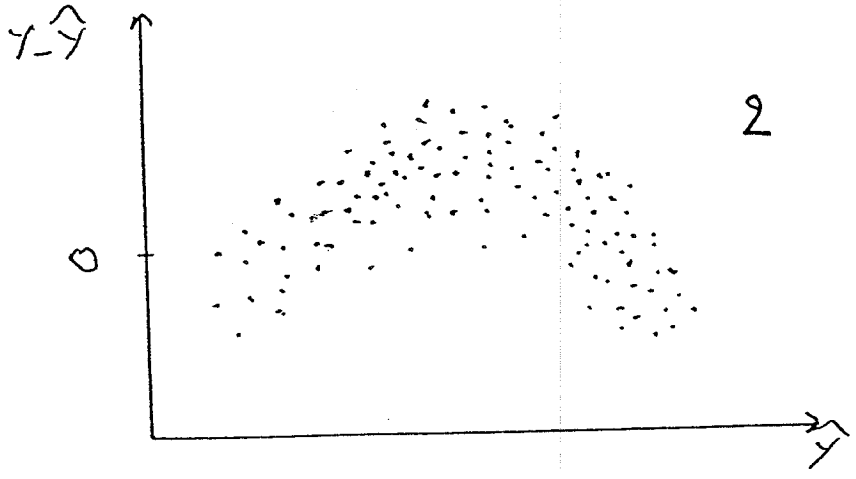
Mais ceci n'est pas toujours suffisant. Ainsi dans l'exemple des crevettes, le graphe des résidus en fonction des valeurs prédites pour le modèle rejeté numéro 4 où l'on suppose l'égalité de toutes les pentes, n'est pas très différent à première vue de celui du modèle qui est retenu. Cependant si on s'intéresse à chaque niveau de courant (avec les symboles donnés sur les graphes) on observe pour chaque niveau des tendances très marquées dans le cas du modèle rejeté. Ces tendances n'apparaissent pas dans le modèle conservé.

La seule conclusion pouvant être donnée ici est la nécessité de bien examiner les écarts aux ajustements, éventuellement de réaliser des tests de normalité... en ayant bien conscience qu'il n'y a pas de procédure définitivement établie permettant de valider un modèle (et qu'il n'y en aura peut-être jamais selon MacCullagh et Nelder).

Par exemple, dans le graphe 1 les résidus sont globalement de moyenne nulle pour tout Y, mais leur variance augmente avec $\widehat{Y}$. Cette situation conduit à envisager une transformation logarithmique pour Y.

Dans le graphe 2, la moyenne des résidus dépend de $\widehat{Y}$, le modèle ne parvient pas à rendre compte des espérances de Y, il convient d'envisager une autre formulation avec par exemple des régressions sur polynômes.





7 Un exemple de modèle moins classique

Les modèles présentés jusqu'ici sont relativement classiques. Il existe une très grande variété de modèles construits pour répondre à des questions très spécifiques. Nous illustrons cet aspect par un modèle adapté à l'analyse des phénomènes périodiques qui nous permettra de mieux illustrer la diversité des formes possibles d'un même modèle.

Régression périodique

Admettons que l'on dispose de données d'une série chronologique périodique dont la période T est connue. Les n observations disponibles sont faites à des pas de temps réguliers, et l'on dispose de données relatives à l périodes.

A titre d'exemple, on peut disposeer sur cinquante années de données mensuelles pour une série de températures. On sait que la variable étudiée est caractérisée par une période annuelle et on admet qu'il n'y a pas de tendance interanuelle.

On peut alors s'intéresser au modèle :

$$Y_{ik} = m_i + \varepsilon_{ik} \quad (i = 1 \cdots 12) \quad \varepsilon_{ik} \sim \mathcal{N}(0, \sigma^2)$$

correspondant à un modèle à un facteur de variation (effet mois).

On peut aussi de référer au résultat selon lequel une fonction périodique de période T peut s'écrire sous la forme d'un somme de sinus et de cosinus de périodes T, $T/2$, $T/3 \cdots$.

On peut donc faire une régression sur des variables sinus et cosinus.

Le modèle considéré est de dimension 12; on va donc faire la regression multiple de Y sur les fonctions suivantes [2] :

$$Y_t = a_o + \sum_{k=1}^{6} a_k \cos(2\pi\ k\ t/T) + \sum_{k=1}^{5} b_k \sin(2\pi\ k\ t/T)$$

Un résultat intéressant est asssocié à la matrice $(X'X)^{-1}$ du modèle qui est diagonale (à condition que le pas de temps soit régulier, diviseur de T et qu'on dispose d'un nombre entier de périodes)

$$\begin{pmatrix} 1 & \cos(2\pi 1/12) & \cos(2\pi 2/12) & \cdots & \sin(2\pi 5/12) \\ 1 & \cos(2\pi 2/12) & \cos(2\pi 4/12) & \cdots & \sin(2\pi 10/12) \\ 1 & \cos(2\pi 3/12) & \cos(2\pi 6/12) & \cdots & \sin(2\pi 15/12) \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & \cos(2\pi t/12) & \cos(2\pi 2t/12) & \cdots & \sin(2\pi 5t/12) \end{pmatrix}$$

[2] D'une manière générale, si T est paire on prend les $T/2$ premiers cosinus et les $T/2-1$ premiers sinus ; si T est impaire, on prend les $(T-1)/2$ premiers cosinus et sinus

L'ajustement donne un résultat identique à celui de l'analyse de variance (même $\widehat{Y}$) (cf. sorties de programme ci-après).

Ceci est tout à fait normal, l'analyse de variance conduit à estimer chaque $E(Y_{ik})$ par la moyenne des données du mois i. La "question" posée à la régression multiple périodique est de rechercher pour chaque mois la valeur qui va minimiser la somme des carrés des écarts entre les observations le la valeur ajustée relative au mois dont elles relèvent. La question est donc la même dans les deux cas et l'équation de la régression périodique, faisant appel à des combinaisons linéaires de 12 variables indépendantes, permet de "passer" par n'importe quel ensemble de 12 points, en particulier celui constitué des 12 moyennes. L'ajustement obtenu par l'analyse de variance est donc accessible et est "trouvé" par la régression multiple.

On aurait pu avoir le même résultat avec un polynôme de degré 11 selon le numéro du mois.

Ces modèles sont donc équivalents, il correspondent cependant à des représentations différentes des données disponibles. On peut s'en rendre compte en recherchant quelles hypothèses sont testables plus ou moins naturellement selon ces formulations.

Ainsi, dans le modèle avec les effets mois, on pourra s'intéresser à l'égalité des moyennes mensuelles au sein de saisons, dans le cas de la regression périodique, on s'interessera de façon plus naturelle à la présence des divers harmoniques.

La sortie d'ordinateur donnée ci-après donne une illustration de ces résultats avec des données simulées avec les deux écritures, analyse de variance d'un modèle avec effet mois ou régression périodique. A titre d'illustration, on montre la matrice $(X'X)^{-1}$, qui est bien diagonale.

```
   1  job"regper"
```

TROIS FACONS DE TRAITER UNE SERIE CHRONOLOGIQUE ...
AVEC UNE PERIODE CONNUE T=12

```
   2  scal n,nv,T : calc n=6  &nv=600  &T=12
   3  fact [lev=12;val=(1...12)50]mois
   4  unit [n=nv]
   5  vari t;value=!(1...nv)
   6  calc t2=t*t/90000
   7  point [nv=n]co &si
   8  calc pi=2*arcsin(1)
```

CALCUL DES COSINUS ET SINUS.

CO[i]=cos(2*pi*t*i/T)
SI[i]=sin(2*pi*t*i/T)

```
   9  for c=co[1...n];s=si[1...n];i=1...n
  10  calc c=cos(2*pi*i*t/T)
  11  calc s=sin(2*pi*i*t/T)
  12  endf
```

SIMULATION DE tt SELON DES LOIS NORMALES CENTREES REDUITES INDEPENDANTES

```
  13  calc tt=urand(1) &tt=ned(tt)
```

AJOUT D'UNE COMPOSANTE DETERMINISTE PERIODIQUE

```
  14  calc tt=4*tt+co[1]+si[1]+co[4]+si[5]
```

$\varepsilon \sim N(0, I_n \sigma^2)$ La composante periodique
$\sigma^2 = 16$

PREMIER AJUSTEMENT AVEC LES SINUS ET COSINUS

```
  15  model tt;resi=r
  16  fit [fprob=yes]co[],si[1...5]

16.....................................................................
```

***** Regression Analysis *****

Response variate: tt
Fitted terms: Constant, co[1], co[2], co[3], co[4], co[5],
co[6], si[1], si[2], si[3], si[4], si[5]

*** Summary of analysis ***

	d.f.	s.s.	m.s.	v.r.	F pr.
Regression	11	1841.	167.38	10.82	<.001
Residual	588	9098.	15.47		
Total	599	10940.	18.26		

Percentage variance accounted for 15.3
Standard error of observations is estimated to be 3.93
* MESSAGE: The following units have large standardized residuals:

573	-3.28
579	3.34

*** Estimates of regression coefficients ***

	estimate	s.e.	t(588)
Constant	0.187	0.161	1.17
co[1]	1.382	0.227	6.08
co[2]	0.202	0.227	0.89
co[3]	-0.079	0.227	-0.35
co[4]	1.096	0.227	4.83
co[5]	-0.070	0.227	-0.31
co[6]	0.043	0.161	0.27
si[1]	0.810	0.227	3.56
si[2]	0.020	0.227	0.09
si[3]	0.293	0.227	1.29
si[4]	0.135	0.227	0.59
si[5]	1.487	0.227	6.55

RECUPERATION DE LA MATRICE (X'X)-1

```
  17  rkeep inverse=xx
  18  print xx
```

LA MATRICE (X'X)-1

	xx				
Constant	0.0016667				
co[1]	0.0000000	0.0033333			
co[2]	0.0000000	0.0000000	0.0033333		
co[3]	0.0000000	0.0000000	0.0000000	0.0033333	
co[4]	0.0000000	0.0000000	0.0000000	0.0000000	0.0033333
co[5]	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
co[6]	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
si[1]	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
si[2]	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
si[3]	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
si[4]	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
si[5]	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000

	Constant	co[1]	co[2]	co[3]	co[4]
co[5]	0.0033333				
co[6]	0.0000000	0.0016667			
si[1]	0.0000000	0.0000000	0.0033333		
si[2]	0.0000000	0.0000000	0.0000000	0.0033333	
si[3]	0.0000000	0.0000000	0.0000000	0.0000000	0.0033333
si[4]	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
si[5]	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000

	co[5]	co[6]	si[1]	si[2]	si[3]
si[4]	0.0033333				
si[5]	0.0000000	0.0033333			

	si[4]	si[5]

DEUXIEME AJUSTEMENT REGRESSION SUR LA VARIABLE QUALITATIVE MOIS

```
  19  model tt;resi=r
  20  fit [fprob=yes]mois

20.....................................................................
```

VERIFIER QUE LE TABLEAU D'ANALYSE DE LA VARIANCE EST SEMBLABLE AU PRECEDENT...

***** Regression Analysis *****

Response variate: tt
Fitted terms: Constant, mois

*** Summary of analysis ***

	d.f.	s.s.	m.s.	v.r.	F pr.
Regression	11	1841.	167.38	10.82	<.001
Residual	588	9098.	15.47		
Total	599	10940.	18.26		

Percentage variance accounted for 15.3
Standard error of observations is estimated to be 3.93
* MESSAGE: The following units have large standardized residuals:

573	-3.28
579	3.34

*** Estimates of regression coefficients ***

	estimate	s.e.	t(588)
Constant	2.529	0.556	4.55
mois 2	-2.898	0.787	-3.68
mois 3	0.512	0.787	0.65
mois 4	-4.170	0.787	-5.30
mois 5	-2.781	0.787	-3.53
mois 6	-2.234	0.787	-2.84
mois 7	-5.397	0.787	-6.86
mois 8	-3.196	0.787	-4.06
mois 9	-3.494	0.787	-4.44
mois 10	-1.526	0.787	-1.94
mois 11	-3.150	0.787	-4.00
mois 12	0.233	0.787	0.30

pas de "mois 1" (cf p. 22)

TROISIEME AJUSTEMENT AVE ANALYSE DE VARIANCE.

ON OBTIENT LE MEME RESULTAT MAIS AVEC DES CONTRAINTES
D'IDENTIFICATIONS DIFFERENTES (SOMME DES EFFETS NULLE)

AVEC LE SECOND AJUSTEMENT (REGRESSION), L'EFFET DU DERNIER MOIS ETAIT NUL !!

```
  21  treat mois
  22  anova [fprob=yes]tt;fitted=f
```

22.....

***** Analysis of variance *****

Variate: tt

Source of variation	d.f.	s.s.	m.s.	v.r.	F pr.
mois	11	1841.13	167.38	10.82	<.001
Residual	588	9098.41	15.47		
Total	599	10939.54			

* MESSAGE: the following units have large residuals.

units 573	-12.76	s.e. 3.89
units 579	12.99	s.e. 3.89

***** Tables of means *****

Variate: tt

Grand mean 0.19

mois	1	2	3	4	5	6	7
	2.53	-0.37	3.04	-1.64	-0.25	0.30	-2.87

mois	8	9	10	11	12
	-0.67	-0.96	1.00	-0.62	2.76

*** Standard errors of differences of means ***

Table	mois
rep.	50
s.e.d.	0.787

8 Modèles à effets fixes et/ou aléatoires

Supposons que nous disposions de données relatives à une variable quantitative Y. Des observations ont été réalisées dans une ville au sein de laquelle il y a $p = 14$ arrondissements. Dans chacun d'eux on prend au hasard $q = 2$ quartiers et on mesure la quantité étudiée auprès de $r = 4$ individus dans chaque quartier. On a donc $n = pqr = 112$ observations.

On peut écrire un modèle d'analyse hiérarchisée (que nous critiquerons par la suite) :

$$Y_{ijk} = m + \alpha_i + \beta_{ij} + \varepsilon_{ijk}$$

avec $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$

α_i est l'effet du $i^{ième}$ arrondissement, β_{ij} est l'effet du quartier j de l'arrondissement i

Le système de contraintes :

$\sum_{i=1}^{p} n_{io}\alpha_i = 0$ et $\sum_{j=1}^{q} n_{ij}\beta_{ij} = 0 \ \ \forall i$ rend les estimateurs de m, des α et des β mutuellement indépendants. On a de plus la décomposition en sommes de carrés suivante :

$$\underbrace{\sum_{ijk}(Y_{ijk} - Y_{...})^2}_{S} = \underbrace{\sum_{ijk}(Y_{i..} - Y_{...})^2}_{S_1} + \underbrace{\sum_{ijk}(Y_{ij.} - Y_{i..})^2}_{S_2} + \underbrace{\sum_{ijk}(Y_{ijk} - Y_{ij.})^2}_{S_3}$$

(Il s'agit d'une décomposition analogue à celle de l'analyse de variance d'un modèle à deux facteurs avec répétition, à ceci près que le "second" facteur, le quartier dans notre exemple est inféodé au premier (arrondissement) et qu'il n'y a donc pas d'effet principal quartier).

La troisème somme (S_3) est la somme de carrés résiduelle qui suit une loi de $\sigma^2\chi^2_{n-pq}$
La première (S_1) suit une loi de $\sigma^2\chi^2_{p-1}$ sous l'hypothèse de nullité des α. La seconde (S_2) suit une loi de $\sigma^2\chi^2_{p(q-1)}$ sous l'hypothèse de nullité des β.

Ces lois de $\sigma^2\chi^2$ sont par ailleurs indépendantes et permettent de construire immédiatement les tests des diverses hypothèses.

La décomposition donnée ci-dessus est donnée en sortie par les bons logiciels (cf exemple page 37).

On remarque que le test relatif à l'absence d'effet arrondissement est réalisé en comparant la somme des carrés des écarts "inter arrondissements" à la somme de carrés résiduelle c'est-à-dire ici à la somme de carrés des écarts intra quartiers.

Dans l'extrait de listing, le test de la nullité des effets arrondissements conduit à rejeter cette hypothèse avec un risque d'erreur inférieur à 0.1%. On observe également que le test de nullité des effets quartiers intra arrondissement conduit également au rejet de cette hypothèse avec un risque inférieur à 0.1%.

On peut se poser la question de la signification de l'effet arrondissement dès lors que les quartiers d'un même arrondissement peuvent être différents. Il est en effet naturel que la moyenne de deux quartiers quelconques d'un d'arrondissement soit alors différente de deux autres quartiers quelconques d'un autre arrondissement... ou de ce même arrondissement. Le rejet de l'hypothèse de nullité de l'effet arrondissement ne signifie pas nécessairement l'existence d'une différence réelle

entre arrondissement mais seulement celle d'une différence entre les quartiers choisis selon les arrondissements. Ces différences sont le fruit des choix des quartiers qui pourraient être différents si l'on refaisait la sélection des quartiers pour l'enquête ce qui pourrait se traduire par des classsements différents des arrondissements.

Il conviendrait donc ici de pouvoir répondre à la question de savoir si les différences entre arrondissements sont effectivement liées à une source de variation à leur niveau et tenir compte pour répondre à cette question du fait que les quartiers ont été tirés au hasard.

Pour ce faire on considère que la moyenne du quartier j de l'arrondissement i n'est pas une quantité $m+\alpha_i+\beta_{ij}$ donnée mais une réalisation d'une variable aléatoire de moyenne $m+\alpha_i$. L'effet quartier intra arrondissement est *aléatoire* et représenté par une variable aléatoire B_{ij} d'espérance nulle. Ce modèle est ici plus réaliste dans la mesure où si l'on faisait une deuxième fois l'enquête, on ne reprendrait pas les mêmes quartiers. Si tel était le cas, l'écriture du modèle à effet fixe serait justifiée.

Le modèle considérant aléatoire l'effet quartier intra arrondissement s'écrit :

$$Y_{ijk} = m + \alpha_i + B_{ij} + \varepsilon_{ijk}$$

On supposera ici que les B_{ij} sont des variables normales indépendantes entre elles, d'espérance nulle et de variance σ_B^2. Les B_{ij} sont de plus indépendants des ε.

Dans ce modèle les espérances et variance des Y_{ijk} ne sont plus les mêmes que dans le modèle à effet fixe.

L'espérance de Y_{ijk} est maintenant $m + \alpha_i$ au lieu de $m + \alpha_i + \beta_{ij}$ et la variance de Y_{ijk} est maintenant $\sigma_B^2 + \sigma^2$ au lieu de σ^2.

De plus les Y_{ijk} ne sont plus tous indépendants puisque la covariance entre deux variables d'un même quartier est de σ_B^2

Avec ce modèle, la décomposition de la somme des carrés des écarts à la moyenne

$$\underbrace{\sum_{ijk}(Y_{ijk} - Y_{...})^2}_{S} = \underbrace{\sum_{ijk}(Y_{i..} - Y_{...})^2}_{S_1} + \underbrace{\sum_{ijk}(Y_{ij.} - Y_{i..})^2}_{S_2} + \underbrace{\sum_{ijk}(Y_{ijk} - Y_{ij.})^2}_{S_3}$$

reste valable mais ses différents termes ne suivent plus exactement les mêmes lois :

La somme de carrés résiduelle (S_3) suit encore une loi de $\sigma^2\chi^2_{n-pq}$
La première (S_1) suit une loi de $(r\sigma_B^2 + \sigma^2)\chi^2_{p-1}$ sous l'hypothèse de nullité des α. La seconde (S_2) suit une loi de $(r\sigma_B^2 + \sigma^2)\chi^2_{p(q-1)}$.

Le rapport des carrés moyens relatifs à la variabilité interquartier intra arrondissement et à la variabilité résiduelle suit maintenant une loi de

$$\frac{r\sigma_B^2 + \sigma^2}{\sigma^2} F_{p(q-1),pq(r-1)}$$

soit un $F_{p(q-1),pq(r-1)}$ sous l'hypothèse de nullité de σ_B^2

Le rapport de carrés moyens relatifs à la variabilité interarrondissements et à la variabilité résiduelle suit, sous l'hypothèse de nullité des α une loi de

$$\frac{r\sigma_B^2 + \sigma^2}{\sigma^2} F_{p-1,pq(r-1)}$$

Il s'agit d'un F multiplié par quelque chose de plus grand que 1 dès lors que σ_B^2 est positif. Il est donc naturel de rejeter l'hypothèse de nullité des α dès qu'il existe une variabilité des quartiers au sein des arrondissements.

Si la question est de savoir s'il existe un effet arrondissement général, c'est-à-dire de savoir si la moyenne de la distribution des moyennes de quartiers est différente d'un arrondissement à l'autre, il faut bien entendu s'intéresser au rapport des carrés moyens relatifs à la variabilité interarrondissements et à la variabilité interquartiers intra arrondissement qui suit sous l'hypothèse de nullité des α une loi de $F_{p-1,p(q-1)}$.

Les deux résultats peuvent être contradictoires comme c'est le cas dans l'exemple donné dans le listing qui suit. Dans cet exemple les données on été obtenues par simulation à partir d'un modèle où il n'y a pas d'effet arrondissement mais où la variabilité inter quartiers intra arrondissement est présente :

$$Y_{ijk} = m + B_{ij} + \varepsilon_{ijk}$$

avec $\sigma_B^2 = 25$ et $\sigma^2 = 1$

Le test de l'hypothèse de nullité des α conduit à ne pas rejeter cette hypothèse lorsque l'effet quartier est considéré comme aléatoire (page 38).

Cette hypothèse est par contre rejetée si tous les effets sont fixes. Ceci est normal puisque dans ce cas on constate simplement que les moyennes de couples de quartiers sont différentes (page 37).

Genstat 5 Release 3.1 (Sun 4 SPARC / SunOS 4) Sun Apr 23 15:25:11 1995

JEU DE DONNEES
MODELE HIERARCHIQUE EQUILIBRE

Yijk =ai+bij+eijk

i=1...14 ,p=14
j=1,2 ,q=2
k=1...4 ,r=4

DEFINITION DU MODELE A EFFET FIXES ET AJUSTEMENT

(TREAT EQUIVAUT A "TRAITEMEMT")

```
  31  treat zone/quartier
  32  anova[fprob=yes] y
```

***** Analysis of variance *****

Variate: y

Source of variation	d.f.	s.s.	m.s.	v.r.	F pr.
zone	13	1327.8716	102.1440	114.54	<.001
zone.quartier	28	2178.9114	77.8183	87.26	<.001
Residual	126	112.3616	0.8918		
Total	167	3619.1445			

***** Tables of means *****

Variate: y

Grand mean -0.601

zone	1	2	3	4	5	6	7
	-7.007	-4.443	3.463	1.760	2.302	-0.123	1.000

zone	8	9	10	11	12	13	14
	1.745	-2.425	-1.189	-2.562	-2.656	0.746	0.979

zone	quartier	1	2	3
	1	-9.131	-10.326	-1.563
	2	-5.930	-1.692	-5.706
	3	1.610	4.895	3.886
	4	3.973	1.131	0.175
	5	1.243	-0.846	6.508
	6	-6.511	0.109	6.034
	7	-3.200	7.870	-1.671
	8	-1.887	3.235	3.886
	9	-7.883	1.108	-0.500
	10	-4.826	-0.904	2.164
	11	-6.436	4.781	-6.032
	12	-5.250	-7.652	4.934
	13	2.848	0.951	-1.562
	14	-3.072	2.793	3.217

MODELE HIERARCHIQUE EQUILIBRE AVEC EFFET QUARTIER ALEATOIRE
(ANNONCE PAR LA DIRECTIVE ("bloc zone/quartier)

Yijk =ai+Bij+eijk

```
  34  bloc  zone/quartier
  35  treat zone
  36  anova [fprob=yes]y;fitted=fy
```

***** Analysis of variance *****

Variate: y

Source of variation	d.f.	s.s.	m.s.	v.r.	F pr.
zone.quartier stratum					
zone	13	1327.8716	102.1440	1.31	0.263
Residual	28	2178.9114	77.8183	87.26	
zone.quartier.*Units* stratum	126	112.3616	0.8918		
Total	167	3619.1445			

***** Tables of means *****

Variate: y

Grand mean -0.601

zone	1	2	3	4	5	6	7
	-7.007	-4.443	3.463	1.760	2.302	-0.123	1.000
zone	8	9	10	11	12	13	14
	1.745	-2.425	-1.189	-2.562	-2.656	0.746	0.979

ON OBSERVE L'ABSENCE DES MOYENNES RELATIVES AU QUARTIERS, CE QUI EST NORMAL PUISQU'IL S'AGIT D'UN EFFET ALEATOIRE... AVEC LE MODELE A EFFETS FIXES CES MOYENNES AVAIENT UN SENS.

Pourquoi dire qu'un effet est fixe ou aléatoire ?

Dans l'exemple donné ici, l'effet quartier doit être considéré comme aléatoire. En effet, si on renouvelle l'expérience, on sélectionne de nouveaux quartiers et les moyennes des couples de quartiers par arrondissement seront de nouveau différentes. mais le classement de ces moyennes et donc celui des arrondissements ne sera plus le même.

Si tous les quartiers étaient sélectionnés, les moyennes par arrondissement seraient différentes et en renouvelant l'expérience, les mêmes quartiers seraient de nouveau sélectionnés et le classement des arrondissements serait stable. L'effet quartier peut alors être considéré comme fixe.

En fait, même dans ce dernier cas, ce choix n'est pas évident. Si on s'intéresse par exemple à une variable indiquant l'état nutritionnel d'une population et que l'on désire indiquer à une administration les arrondissements où il conviendrait d'agir en priorité, le modèle à effet fixe s'impose. Si on désire comprendre la nature de la variabilité de la variable étudiée, on peut considérer que les quartiers composant un arrondissement sont le produit d'un tirage au sein d'une population infinie de quartiers. Dans ce cas le modèle avec un effet quartier aléatoire s'impose.

La nature du modèle est donc complètement inféodée à la question posée. Des réponses différentes ne signifient pas une incohérence de l'outil statistique mais indiquent simplement que les questions n'étaient pas les mêmes.

La question de savoir si la question à laquelle on répond est bien la question que l'on se pose est donc une question cruciale...

Plans d'expériences, mesures répétées...

L'usage des modèles mixtes, c'est-à-dire combinant des effets fixes et des effets aléatoires est très répandu dans la planification des expériences, dès lors que des observations sont réalisées dans des dispositifs comprenant des sources de variations incontournables et ne faisant pas l'objet de la question conduisant à la mise en place de l'expérience.

C'est le cas par exemple lorsque la recherche de différences éventuelles entre diverses variétés d'une plante selon une variable quelconque implique la mise en culture de plusieurs "champs". Dans chaque champ on peut par exemple délimiter un nombre de parcelles égal au nombre de variétés étudiées. Il y a alors deux sources de variation, variété et champ. L'effet variété est naturellement fixe et l'effet champ est aléatoire, il s'agit ici d'un plan en blocs complets sans répétition. Le modèle permet de tenir compte du fait que les observations faites dans un même champ ne sont pas indépendantes.

Cette référence aux plans d'expériences est très classique, avec des variétés de plantes et des champs, mais cette situation peut se présenter dans une multitude de circonstances, par exemple lorsqu'on mesure une quantité quelconque en divers points d'un organe, chaque organe étant pris sur un individu. Dans ce cas le lieu de prélèvement est représenté par un effet fixe et l'individu (organe) par un effet aléatoire. Le modèle s'écrit alors

$$Y_{ik} = m + \alpha_i + O_k + \varepsilon_{ik}$$

où $O_k \sim \mathcal{N}(0, \sigma_O^2)$ et $\varepsilon_{ik} \sim \mathcal{N}(0, \sigma^2)$ ces variables étant bien sûr indépendantes. Avec ce modèle, les variables relatives à un même organe sont bien toutes corrélées (covar(Y_{i_1k}, Y_{i_2k})=σ_O^2)

On se trouve assez fréquemment confronté à un problème de confusion entre des sources de variations correspondant à des effets fixes ou aléatoires. C'est le cas par exemple lorsqu'on installe chaque variété dans un seul champ et que chaque champ ne reçoit qu'une seule variété. Si cette situation peut sembler ridicule, elle est souvent rencontrée parce que les dispositifs expérimentaux sont lourds à mettre en place et l'expérimentateur est très souvent persuadé qu'en suivant de façon méticuleuse les protocoles techniques, il supprime toutes les sources indésirables de variation. Cette confiance est souvent excessive et il est parfois possible, sans grande difficulté de modifier le protocole expérimental de façon à éviter des situations où l'on ne peut pas conclure sur l'origine d'une source de variation, sauf en refaisant l'expérience.

La prise en compte de sources de variation selon un effet aléatoire peut être également faite dans le cas de plan "à mesures répétées", lorsque la quantité étudiée est observée à plusieurs reprises auprès des mêmes individus, chacun d'eux pouvant alors être considéré comme un niveau d'un facteur aléatoire. Mais il existe d'autres façons de rendre compte des corrélations entre observations faites auprès d'un même individu, en particulier lorsqu'on considère que les corrélations entre les mesures s'estompent au cours du temps.

9 Prolongements...

Cette présentation est bien évidemment très loin d'être complète. Un certain nombre d'ouvrages dont des références sont données ci après peuvent être consultés pour la prolonger.

La théorie du modèle linéaire peut être prolongée par des généralisations selon plusieurs directions.

Les modèles non linéaires

Les résultats présentés jusqu'ici étaient relatifs au modèle

$$Y = X \times \theta + \varepsilon$$

avec des suppositions sur la distribution des ε

Nombre des résultats présentés restent valables asymptotiquement (lorsque le nombre d'observations devient grand) dans la cas où l'espérance de Y n'est pas linéaire *ie* dans le cas où le modèle s'écrit :

$$Y = f(x, \theta) + \varepsilon$$

en gardant les mêmes suppositions sur la distribution des ε

Ces modèles posent le problème qu'on ne sait pas trouver directement les valeurs $\widehat{\theta}$ minimisant la somme des carrés des écarts entre valeurs observées et estimées. Il faut alors recourrir à des algorithmes itératifs présents dans les bons logiciels (Proc NLIN de SAS par exemple). Les difficultés peuvent provenir de l'existence minimums locaux de la fonction dont on cherche le minimum, et les solutions obtenues peuvent dépendre des valeurs initiales données aux paramètres...

Les ouvrages donnés en référence comportent généralement des chapitres consacrés aux modèles non linéaires.

Les modèles linéaires généralisés.

Nous n'avons évoqué jusqu'ici que le modèle linéaire gaussien. La normalité des variables n'est pas d'une manière générale l'hypothèse la plus contraignante. Par ailleurs, un ajustement par les moindres carrés reste un ajustement intéressant en général. La non normalité des variables est souvent évoquée pour faire autre chose que des moindres carrés. Ceci peut être justifié si l'intérêt de l'alternative est clairement présenté, ce qui n'est pas toujours le cas...

Il existe des situations dans lesquelles il est justifié de choisir un ajustement selon un critère autre que celui des moindres carrés. C'est le cas lorsque la distribution Y est l'une des distributions de la famille exponentielle, famille de distributions dont relèvent, outre les lois normales, les lois de Poisson, binomiales, gamma et quelques autres.

Dans ce cas la théorie des modèles linéaires généralisés est très utile. On peut en effet écrire le modèle sous la forme

$$Y_{ik} = \mu_i + \varepsilon_{ik}$$

où μ_i est une fonction éventuellement non linéaire d'un prédicteur linéaire η_i la relation entre μ_i et η_i est donnée par une fonction de lien ("link function") G :

$$\eta_i = G(\mu_i)$$

Si par exemple la distribution de Y_i est une loi de Poisson, on montre que la fonction de lien la mieux adaptée est la fonction Logarithme. La régression de Y sur p variables $x^{(j)}$ $j = 1 \cdots p$ consiste à représenter l'espérance de Y_i par :

$$\log(E(Y_i)) = \alpha + \sum_j \beta_j x_i^{(j)}$$

on a donc également $E(Y_i) = e^{\eta_i}$

$$E(Y_i) = e^{\left(\alpha + \sum_j \beta_j x_i^{(j)}\right)}$$

Plus généralement, en revenant à la notation $E(Y) = X \times \theta$, on a

$$G(E(Y)) = X \times \theta \quad \text{et} \quad E(Y) = G^{-1}(X \times \theta)$$

Le modèle gaussien est un cas particulier de ce modèle généralisé avec la fonction G identité ($\eta_i = \mu_i$).

Les bons logiciels tirent pleinement partie de cette présentation, en permettant de présenter la combinaison linéaire $X \times \theta$ d'une façon semblable à celle utilisée dans le modèle gaussien (par exemple la pocédure GENMOD de SAS).

Les procédures de tests sont également assez semblables à celles présentées pour le modèle gaussien, avec toutefois certaines difficultés supplémentaires.

L'utilisation des modèles généralisés est en plein essor, en particulier pour le traitement des variables binomiales avec les regressions logistiques avec

$$G(E(Y)) = \log(\mu/(n - \mu))$$

et

$$G^{-1}(\eta) = n\ e^{\eta}/(1 + e^{\eta})$$

n étant le nombre total d'observations parmi lesquelles Y est le le nombre présentant la qualité dont on étudie la proportion.

Conclusion

Cette présentation du modèle linéaire est bien sûr très insuffisante. Elle peut être complétée par de nombreux ouvrages existants dont quelques uns sont donnés en référence ci-après.

Nous espérons avoir fait sentir que le modèle linéaire n'est pas qu'une technique d'ajustement mais un outil de représentation de distributions aléatoires. La puissance des ordinateurs et la convivialité des logiciels actuels permettent de définir des modèles de plus en plus divers.

Les généralisations actuellement en plein développement, telles que les méthodes d'estimation fonctionnelles (régressions non paramétriques, modèles additifs généraux...), sont de plus en plus utilisées et semblent pouvoir améliorer de façon notable la capacité de représentation et d'exploration. Ces méthodes ne peuvent guère cependant être utilisées sans références au modèle linéaire dont elles constituent des prolongements. Ainsi les quelques principes exposés ici sur la validation des modèles, sur la recherche de modèles simplifiés sont très largement communs à toutes ces méthodes.

La présentation faite ici du modèle linéaire a donc une portée bien plus générale que celle de l'utilisation de :

$$Y = X \times \theta + \varepsilon$$

Quelques références

Les Références données ci-dessous regroupent quelques ouvrages de base en anglais et en français.

Ces ouvrages sont tous disponibles au centre de documentation du centre Orstom de Montpellier.

- P. Dagnélie. 1969. Théorie et méthodes statistiques. 2 volumes, les presses agronomiques de Gembloux.
- P. Dagnélie. 1981. Principes d'expérimentation. Les presses agronomiques de Gembloux.
- N. Draper and H. Smith 1981. Applied Regression Analysis 2ème édition. Wiley
- Huet S., E. Jolivet, A. Messéan, 1992. La régression non linéaire, méthode et applications en biologie, INRA.
- P. Mac Cullagh and J.A. Nelder 1983. Generalized Linear models. Monographs on statistics and applied probability, 37, Chapman and Hall.
- Tomassone R. S. Audrain, E. Lesquoy-de Turkheim, C. Millier, 1992 2ème édition. La régression, nouveau regard sur une ancienne méthode statistique statistique. Inra
- Tomassone R. , C. Dervin et J.P. Masson, 1993. Biométrie : modélisation de phénomènes biologiques. Masson.

Annexe

Les Crevettes de Casamance

M. SIRIER -

```
/*---------------------------------------- */
/*                                         */
/*  Programme SAS pour cours               */
/*     Modeles lineaires  24-28 avril 95   */
/*                                         */
/* ---------------------------------------- */

/*------- Definition options generales  -------------------------*/

libname  trav '';
title  'TP MODELE LINEAIRE 24-28 AVRIL 95';
options pagesize=65 nodate nonumber;

/*--- Configuration pour sorties en graphique haute resolution --*/

filename gsasfile pipe 'lpr -Plp';

goptions device=xbw
         target=applelw
         colors=(black)
         cback=white
         gaccess=gsasfile
         ;

/*------- Lecture des donnees -----------------------------------*/

data trav.ml;
  infile 'ml.dat';
  input taille crtm crts crtf sal;
  sal2=sal*sal;
  keep taille crts sal sal2;
run;

proc sort data=trav.ml out=trav.ml;
     by crts sal;
run;

proc print data=trav.ml;
     title2 'FICHIER DE DONNEES';
run;

symbol1 i=none v=plus;
symbol2 i=none v=X;
symbol3 i=none v=star;
symbol4 i=none v=square;
symbol5 i=none v=diamond;

proc gplot data=trav.ml;
     title2 'TAILLE OBSERVEE EN FONCTION DE LA SALINITE';
     plot taille*sal=crts /
           haxis=0 to 80 by 10
           vaxis=10 to 35 by 5;
run;

/* ------------------------------------------------ */
/* Modele 1 taille=f(crts, sal*crts, sal2*crts)     */
/* ------------------------------------------------ */

proc glm data=trav.ml;
     title1 'MODELE 1 : TAILLE = F( CRTS, SAL*CRTS, SAL2*CRTS )';
     class crts;
     model taille= crts sal*crts sal2*crts
          /solution;
     output out=model1 p=yhat1 r=resid1;
run;
```

```
symbol1 i=none v=plus h=0.3cm;
symbol2 i=none v=X h=0.3cm;
symbol3 i=none v=star h=0.3cm;
symbol4 i=none v=square h=0.3cm;
symbol5 i=none v=diamond h=0.3cm;

proc gchart data=model1;
     title2 'HISTOGRAMME DES RESIDUS';
     vbar resid1;
run;

proc gplot data=model1;
     title2 'RESIDUS EN FONCTION DE LA SALINITE';
     plot resid1*sal=crts;
run;

symbol1 i=join v=plus h=0.2cm;
symbol2 i=join v=X h=0.2cm;
symbol3 i=join v=star h=0.2cm;
symbol4 i=join v=square h=0.2cm;
symbol5 i=join v=diamond h=0.2cm;

proc gplot data=model1;
      title2 'VALEURS DE TAILLE PREDITES EN FONCTION DE LA SALINITE';
      plot yhat1*sal=crts  /
           haxis=0 to 80 by 10
           vaxis=10 to 35 by 5;
      plot2 taille*sal=crts  /
           haxis=0 to 80 by 10
           vaxis=10 to 35 by 5;
run;

/* ---------------------------------- */
/* Modele 2 taille=f(crts, sal*crts) */
/* ---------------------------------- */

proc glm data=trav.ml;
     title1 'MODELE 2 : TAILLE = F( CRTS, SAL*CRTS )';
     class crts;
     model taille= crts sal*crts
          /solution;
     output out=model2 p=yhat2 r=resid2;
run;

proc gchart data=model2;
     title2 'HISTOGRAMME DES RESIDUS';
     vbar resid2;
run;

symbol1 i=none v=plus;
symbol2 i=none v=X;
symbol3 i=none v=star;
symbol4 i=none v=square;
symbol5 i=none v=diamond;

proc gplot data=model2;
     title2 'RESIDUS EN FONCTION DE LA SALINITE';
     plot resid2*sal=crts;
run;

symbol1 i=join v=plus;
symbol2 i=join v=X;
symbol3 i=join v=star;
symbol4 i=join v=square;
```

```
symbol5 i=join v=diamond;

proc gplot data=model2;
      title2 'VALEURS DE TAILLE PREDITES EN FONCTION DE LA SALINITE';
      plot yhat2*sal=crts  /
           haxis=0 to 80 by 10
           vaxis=10 to 35 by 5;
      plot2 taille*sal=crts  /
           haxis=0 to 80 by 10
           vaxis=10 to 35 by 5;
run;

/* ---------------------------------- */
/* Modele 3  taille=f(crts, sal*nc) */
/* ---------------------------------- */
data m1;
     set trav.m1;
     if crts<=65 then nc=1;
     else nc=2;
run;

proc glm data=m1;
     title1 'MODELE 3 : TAILLE = F( CRTS, SAL*NC )';
     class crts nc ;
     model taille= crts sal*nc
          /solution ;
     output out=model3 p=yhat3 r=resid3;
run;

proc gchart data=model3;
     title2 'HISTOGRAMME DES RESIDUS';
     vbar resid3;
run;

symbol1 i=none v=plus;
symbol2 i=none v=X;
symbol3 i=none v=star;
symbol4 i=none v=square;
symbol5 i=none v=diamond;

proc gplot data=model3;
     title2 'RESIDUS EN FONCTION DE LA SALINITE';
     plot resid3*sal=crts;
run;

symbol1 i=join v=plus;
symbol2 i=join v=X;
symbol3 i=join v=star;
symbol4 i=join v=square;
symbol5 i=join v=diamond;

proc gplot data=model3;
      title2 'VALEURS DE TAILLE PREDITES EN FONCTION DE LA SALINITE';
      plot yhat3*sal=crts  /
           haxis=0 to 80 by 10
           vaxis=10 to 35 by 5;
      plot2 taille*sal=crts  /
           haxis=0 to 80 by 10
           vaxis=10 to 35 by 5;
run;


/* ------------------------------------ */
/* Modele 4 taille=f(crts, sal)         */
/* ------------------------------------ */
```

```
proc glm data=trav.ml;
     title1 'MODELE 4 : TAILLE = F( CRTS, SAL )';
     class crts;
     model taille= crts sal
          /solution ;
     output out=model4 p=yhat4 r=resid4;
run;

proc gchart data=model4;
     title2 'HISTOGRAMME DES RESIDUS';
     vbar resid4;
run;

symbol1 i=none v=plus;
symbol2 i=none v=X;
symbol3 i=none v=star;
symbol4 i=none v=square;
symbol5 i=none v=diamond;

proc gplot data=model4;
     title2 'RESIDUS EN FONCTION DE LA SALINITE';
     plot resid4*sal=crts;
run;

symbol1 i=join v=plus;
symbol2 i=join v=X;
symbol3 i=join v=star;
symbol4 i=join v=square;
symbol5 i=join v=diamond;

proc gplot data=model4;
      title2 'VALEURS DE TAILLE PREDITES EN FONCTION DE LA SALINITE';
      plot yhat4*sal=crts  /
           haxis=0 to 80 by 10
           vaxis=10 to 35 by 5;
      plot2 taille*sal=crts  /
           haxis=0 to 80 by 10
           vaxis=10 to 35 by 5;
run;


/* ---------------------------------- */
/* Modele 5  taille=f(nc, sal*nc) */
/* ---------------------------------- */

proc glm data=ml;
     title1 'MODELE 5 : TAILLE = F( NC, SAL*NC )';
     class crts nc ;
     model taille= nc sal*nc
          /solution;
     output out=model5 p=yhat5 r=resid5;
run;

proc gchart data=model5;
     title2 'HISTOGRAMME DES RESIDUS';
     vbar resid5;
run;

symbol1 i=none v=plus;
symbol2 i=none v=X;
symbol3 i=none v=star;
symbol4 i=none v=square;
symbol5 i=none v=diamond;
```

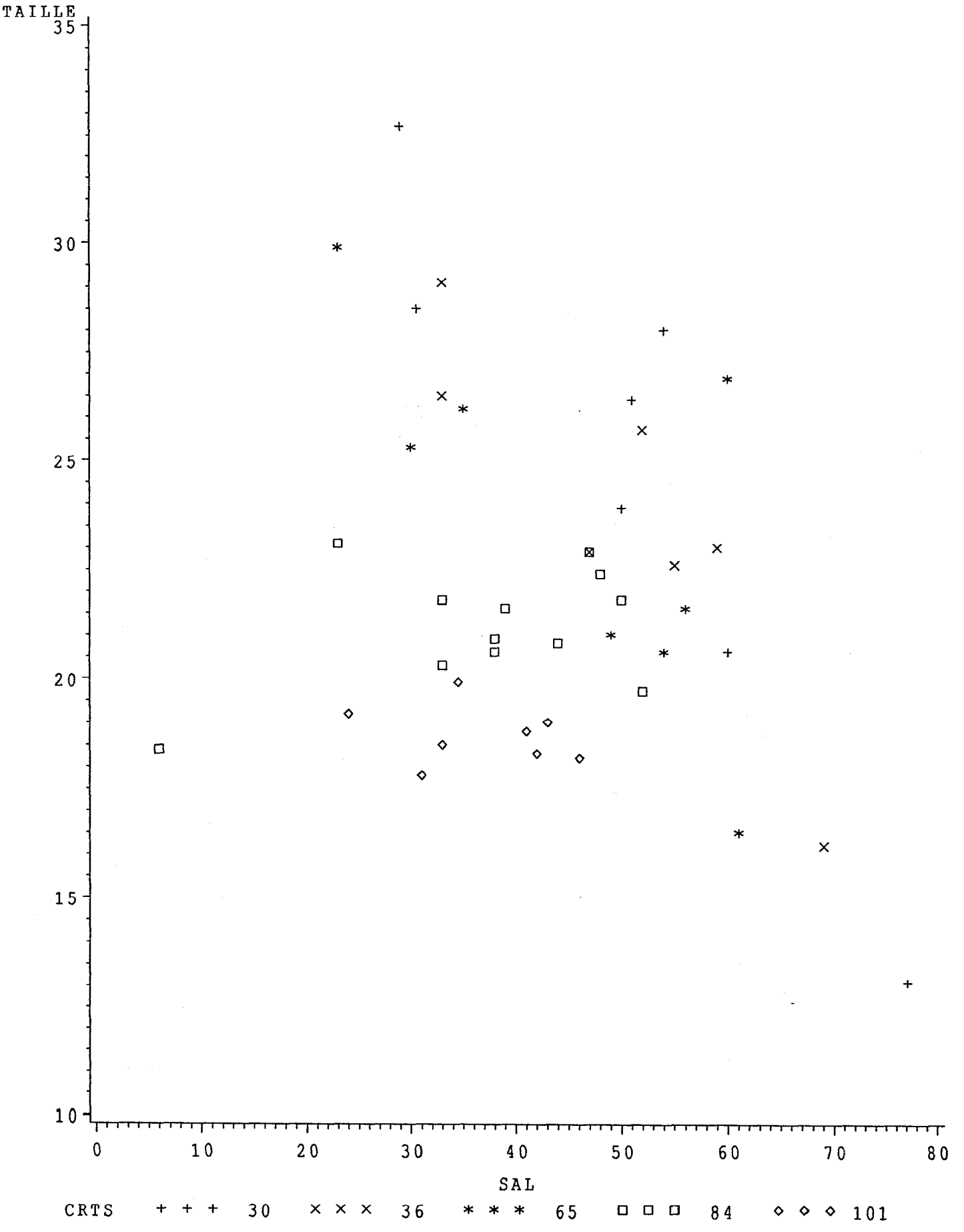
```
proc gplot data=model5;
     title2 'RESIDUS EN FONCTION DE LA SALINITE';
     plot resid5*sal=crts;
run;

symbol1 i=join v=plus;
symbol2 i=join v=X;
symbol3 i=join v=star;
symbol4 i=join v=square;
symbol5 i=join v=diamond;

proc gplot data=model5;
      title2 'VALEURS DE TAILLE PREDITES EN FONCTION DE LA SALINITE';
      plot yhat5*sal=crts  /
           haxis=0 to 80 by 10
           vaxis=10 to 35 by 5;
      plot2 taille*sal=crts  /
           haxis=0 to 80 by 10
           vaxis=10 to 35 by 5;
run;
```

TP MODELE LINEAIRE 24-28 AVRIL 95
FICHIER DE DONNEES

OBS	TAILLE	CRTS	SAL	SAL2
1	32.7	30	29.0	841.00
2	28.5	30	30.5	930.25
3	23.9	30	50.0	2500.00
4	26.4	30	51.0	2601.00
5	28.0	30	54.0	2916.00
6	20.6	30	60.0	3600.00
7	13.1	30	77.0	5929.00
8	19.6	36	.	.
9	26.5	36	33.0	1089.00
10	29.1	36	33.0	1089.00
11	22.9	36	47.0	2209.00
12	25.7	36	52.0	2704.00
13	22.6	36	55.0	3025.00
14	23.0	36	59.0	3481.00
15	16.2	36	69.0	4761.00
16	20.9	65	.	.
17	21.2	65	.	.
18	21.7	65	.	.
19	20.4	65	.	.
20	29.9	65	23.0	529.00
21	25.3	65	30.0	900.00
22	26.2	65	35.0	1225.00
23	.	65	43.0	1849.00
24	21.0	65	49.0	2401.00
25	20.6	65	54.0	2916.00
26	21.6	65	56.0	3136.00
27	26.9	65	60.0	3600.00
28	16.5	65	61.0	3721.00
29	23.1	84	.	.
30	18.4	84	6.0	36.00
31	23.1	84	23.0	529.00
32	21.8	84	33.0	1089.00
33	20.3	84	33.0	1089.00
34	20.6	84	38.0	1444.00
35	20.9	84	38.0	1444.00
36	21.6	84	39.0	1521.00
37	20.8	84	44.0	1936.00
38	22.9	84	47.0	2209.00
39	22.4	84	48.0	2304.00
40	21.8	84	50.0	2500.00
41	19.7	84	52.0	2704.00
42	17.7	101	.	.
43	19.2	101	24.0	576.00
44	17.8	101	31.0	961.00
45	18.5	101	33.0	1089.00
46	19.9	101	34.5	1190.25
47	18.8	101	41.0	1681.00
48	18.3	101	42.0	1764.00
49	19.0	101	43.0	1849.00
50	18.2	101	46.0	2116.00

TP MODELE LINEAIRE 24-28 AVRIL 95

TAILLE OBSERVEE EN FONCTION DE LA SALINITE



MODELE 1 : TAILLE = F(CRTS, SAL*CRTS, SAL2*CRTS)

General Linear Models Procedure
Class Level Information

Class	Levels	Values
CRTS	5	30 36 65 84 101

Number of observations in data set = 50

NOTE: Due to missing values, only 42 observations can be used in this analysis.

MODELE 1 : TAILLE = F(CRTS, SAL*CRTS, SAL2*CRTS)

General Linear Models Procedure

Dependent Variable: TAILLE

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	14	562.996485	40.214035	9.14	0.0001
Error	27	118.809229	4.400342		
Corrected Total	41	681.805714			

R-Square	C.V.	Root MSE	TAILLE Mean
0.825743	9.461272	2.09770	22.1714

Source	DF	Type I SS	Mean Square	F Value	Pr > F
CRTS	4	184.302083	46.075521	10.47	0.0001
SAL*CRTS	5	349.206869	69.841374	15.87	0.0001
SAL2*CRTS	5	29.487533	5.897507	1.34	0.2776

Source	DF	Type III SS	Mean Square	F Value	Pr > F
CRTS	4	32.7996798	8.1999200	1.86	0.1459
SAL*CRTS	5	18.9794835	3.7958967	0.86	0.5187
SAL2*CRTS	5	29.4875331	5.8975066	1.34	0.2776

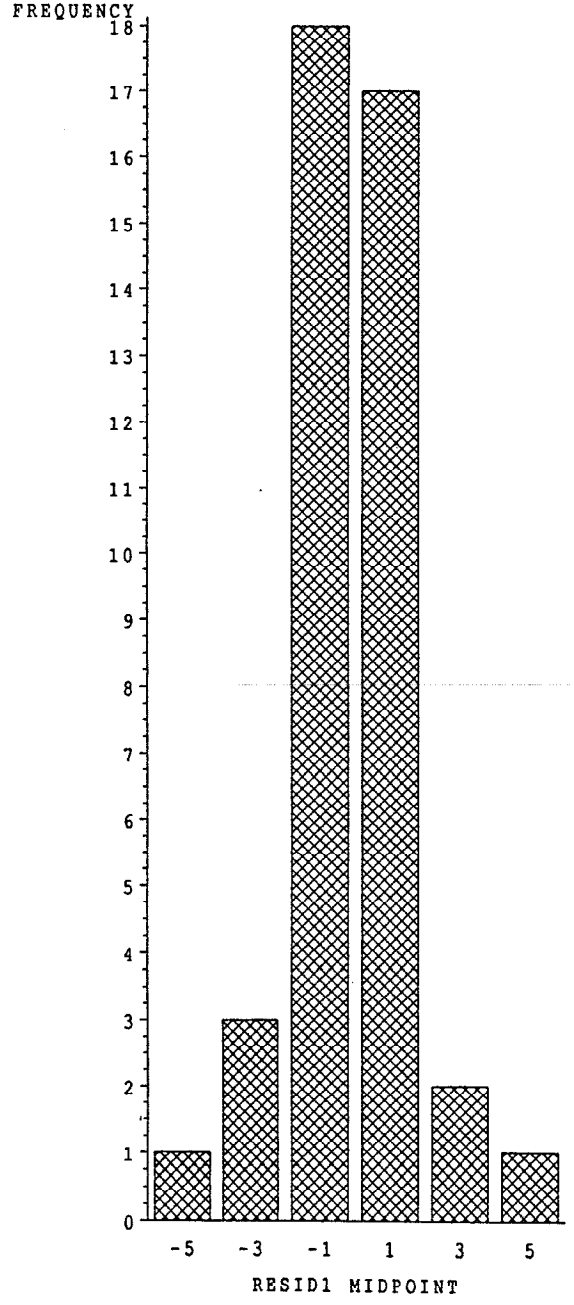
Parameter		Estimate	T for H0: Parameter=0	Pr > |T|	Std Error of Estimate
INTERCEPT		18.55258058 B	0.92	0.3675	20.24040839
CRTS	30	10.65099036 B	0.49	0.6269	21.65905085
	36	3.75005186 B	0.16	0.8774	24.07782856
	65	25.64406773 B	1.13	0.2690	22.72237477
	84	-0.85468517 B	-0.04	0.9670	20.45402319
	101	0.00000000 B	.	.	.
SAL*CRTS	30	0.21342139	0.68	0.5042	0.31523766
	36	0.37616372	0.68	0.5011	0.55165249
	65	-0.79236586	-1.50	0.1452	0.52823499
	84	0.21210033	1.07	0.2953	0.19873535
	101	0.03250428	0.03	0.9781	1.17521891
SAL2*CRTS	30	-0.00549988	-1.81	0.0818	0.00304242
	36	-0.00659241	-1.18	0.2473	0.00557515
	65	0.00683837	1.11	0.2763	0.00615455
	84	-0.00285732	-0.89	0.3794	0.00319753
	101	-0.00073873	-0.04	0.9648	0.01656342

NOTE: The X'X matrix has been found to be singular and a generalized inverse was used to solve the normal equations. Estimates followed by the letter 'B' are biased, and are not unique estimators of the parameters.

54

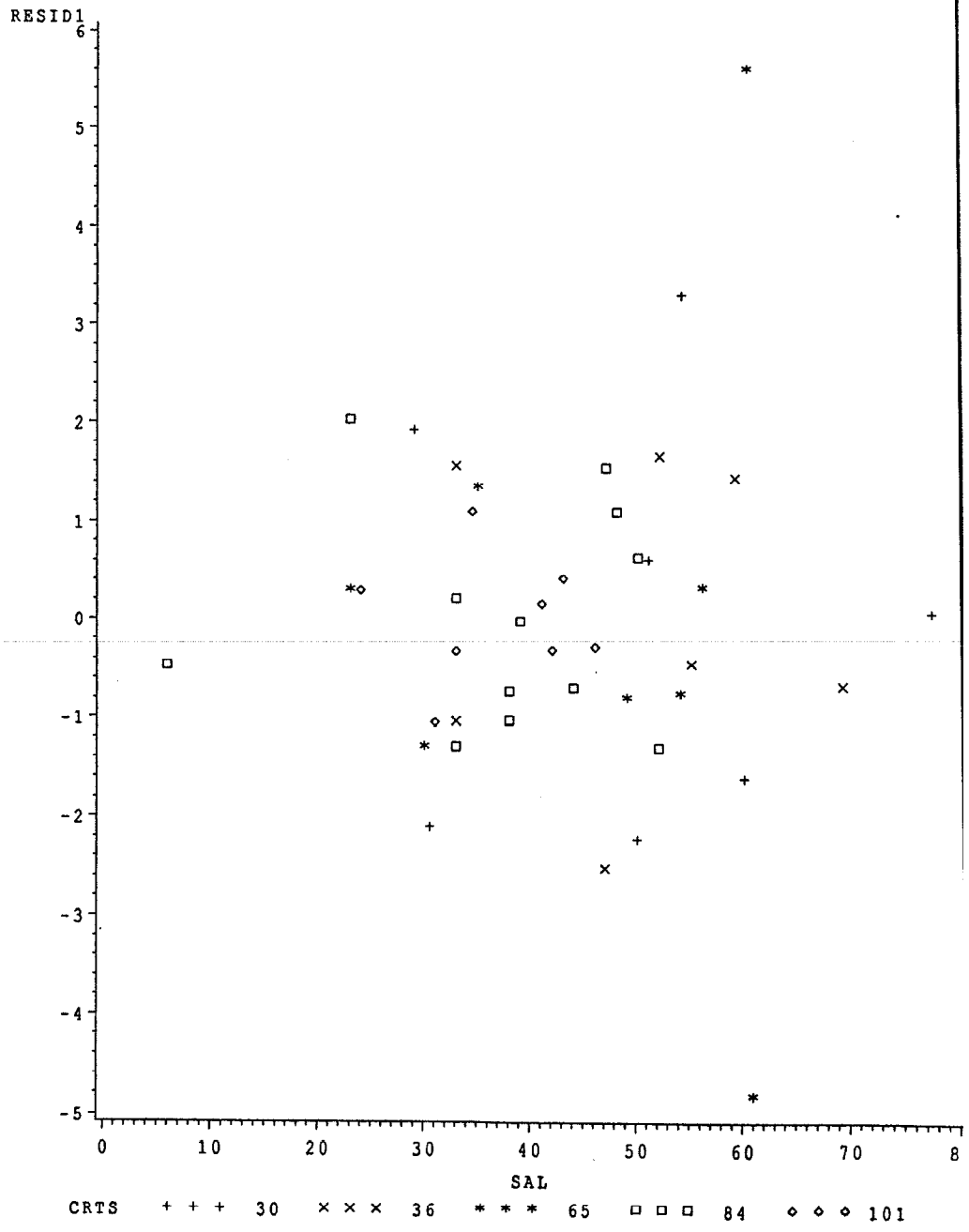
MODELE 1 : TAILLE = F(CRTS, SAL*CRTS, SAL2*CRTS)

HISTOGRAMME DES RESIDUS



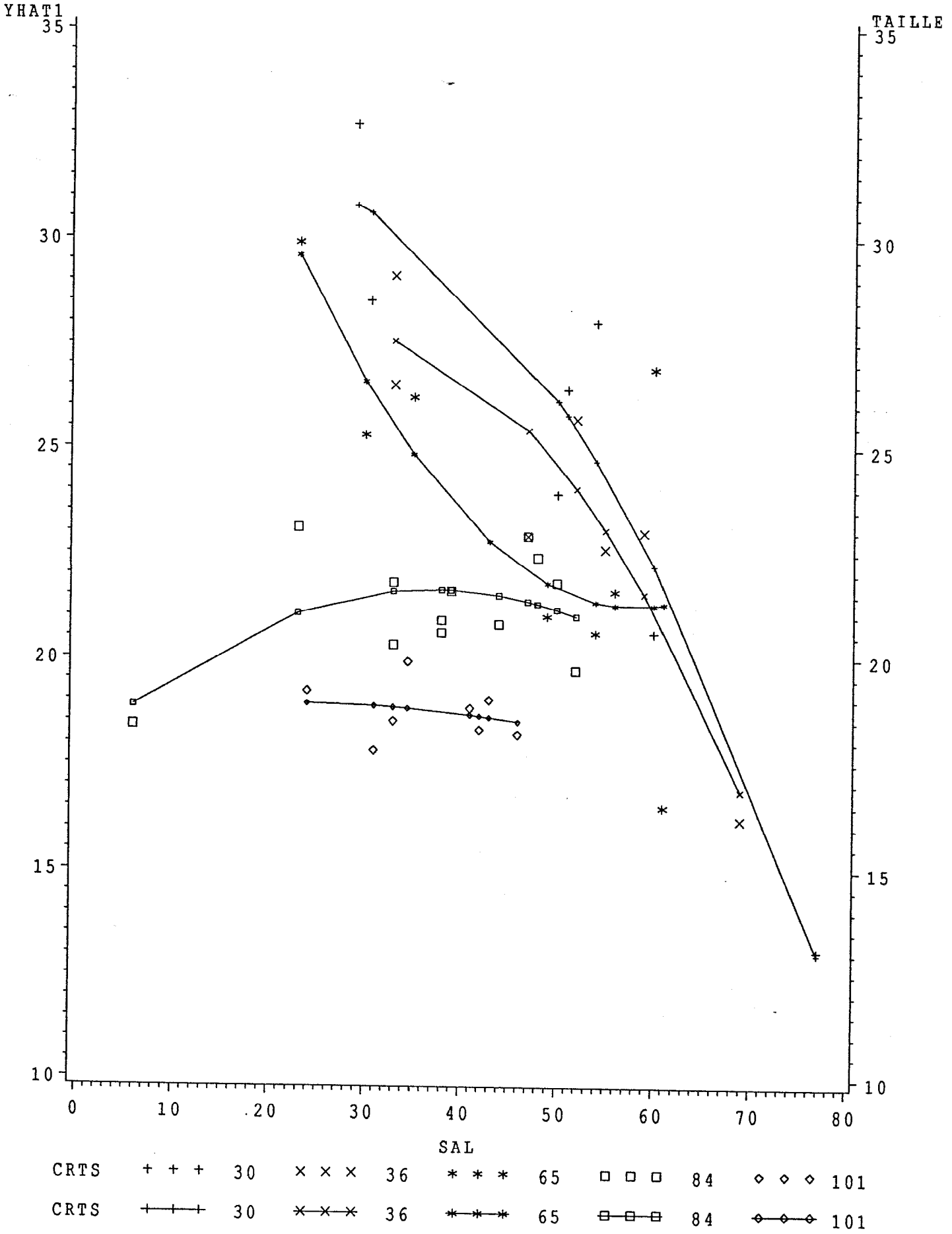
MODELE 1 : TAILLE = F(CRTS, SAL*CRTS, SAL2*CRTS)

RESIDUS EN FONCTION DE LA SALINITE



MODELE 1 : TAILLE = F(CRTS, SAL*CRTS, SAL2*CRTS)

VALEURS DE TAILLE PREDITES EN FONCTION DE LA SALINITE



MODELE 2 : TAILLE = F(CRTS, SAL*CRTS)

General Linear Models Procedure
Class Level Information

Class	Levels	Values
CRTS	5	30 36 65 84 101

Number of observations in data set = 50

NOTE: Due to missing values, only 42 observations can be used in this analysis.

MODELE 2 : TAILLE = F(CRTS, SAL*CRTS)

General Linear Models Procedure

Dependent Variable: TAILLE

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	533.508952	59.278772	12.79	0.0001
Error	32	148.296762	4.634274		
Corrected Total	41	681.805714			

R-Square	C.V.	Root MSE	TAILLE Mean
0.782494	9.709507	2.15274	22.1714

Source	DF	Type I SS	Mean Square	F Value	Pr > F
CRTS	4	184.302083	46.075521	9.94	0.0001
SAL*CRTS	5	349.206869	69.841374	15.07	0.0001

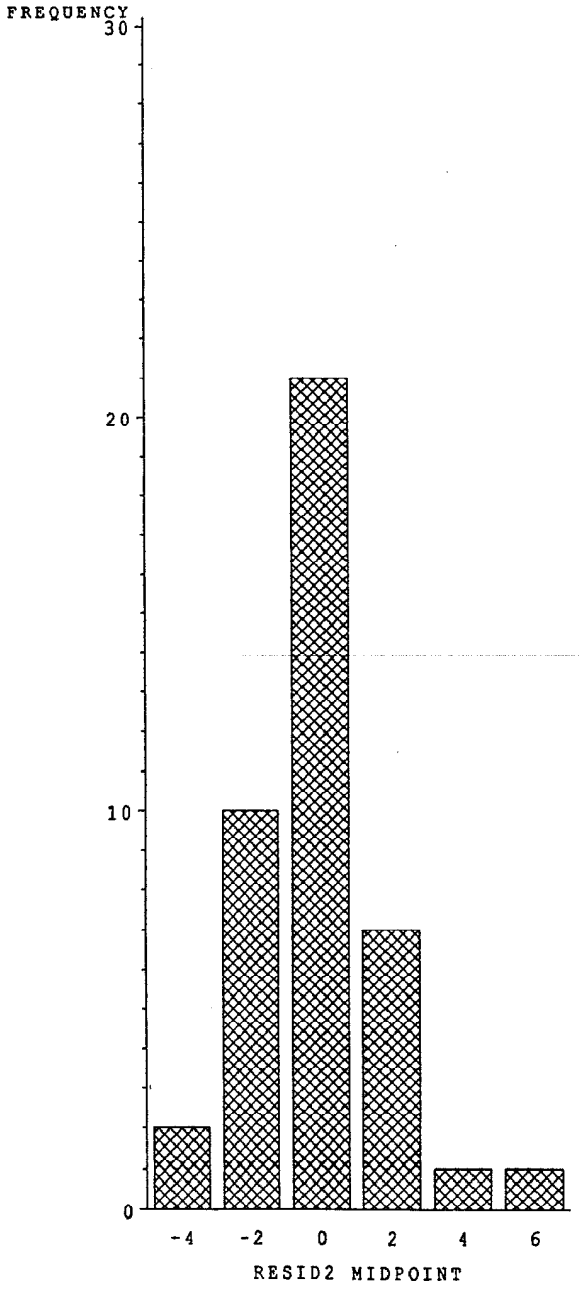
Source	DF	Type III SS	Mean Square	F Value	Pr > F
CRTS	4	271.958490	67.989622	14.67	0.0001
SAL*CRTS	5	349.206869	69.841374	15.07	0.0001

Parameter		Estimate	T for H0: Parameter=0	Pr > |T|	Std Error of Estimate
INTERCEPT		19.43745434 B	4.73	0.0001	4.11008182
CRTS	30	22.82133857 B	4.60	0.0001	4.95629217
	36	17.78270238 B	3.34	0.0022	5.32789102
	65	13.65507899 B	2.79	0.0089	4.89994616
	84	0.25368183 B	0.06	0.9560	4.56157115
	101	0.00000000 B	.	.	.
SAL*CRTS	30	-0.34882376	-6.62	0.0001	0.05272691
	36	-0.27166982	-4.10	0.0003	0.06620112
	65	-0.20853333	-3.75	0.0007	0.05558341
	84	0.03992542	0.80	0.4303	0.04998373
	101	-0.01969316	-0.18	0.8587	0.10971806

NOTE: The X'X matrix has been found to be singular and a generalized inverse was used to solve the normal equations. Estimates followed by the letter 'B' are biased, and are not unique estimators of the parameters.

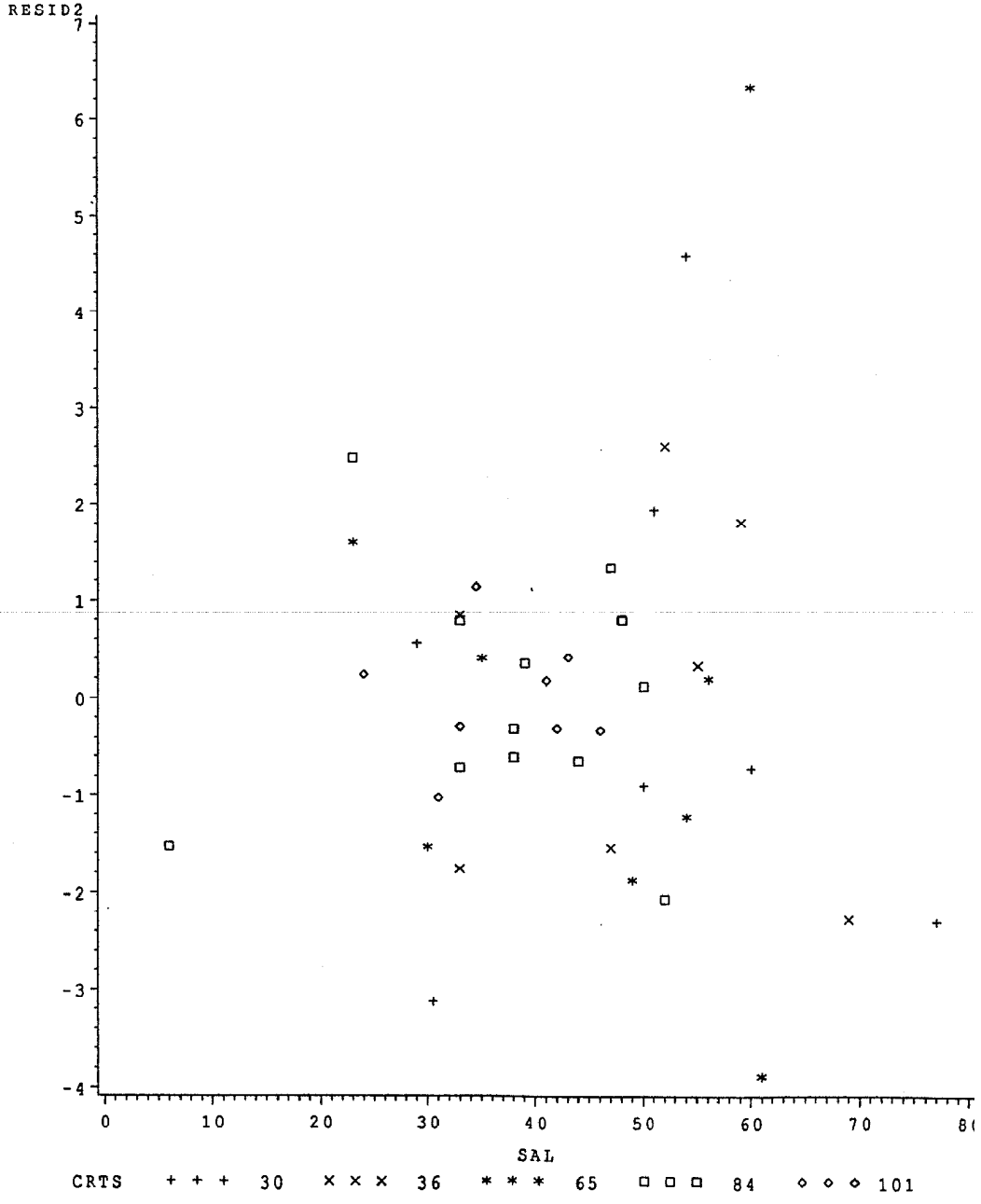
MODELE 2 : TAILLE = F(CRTS, SAL*CRTS)

HISTOGRAMME DES RESIDUS



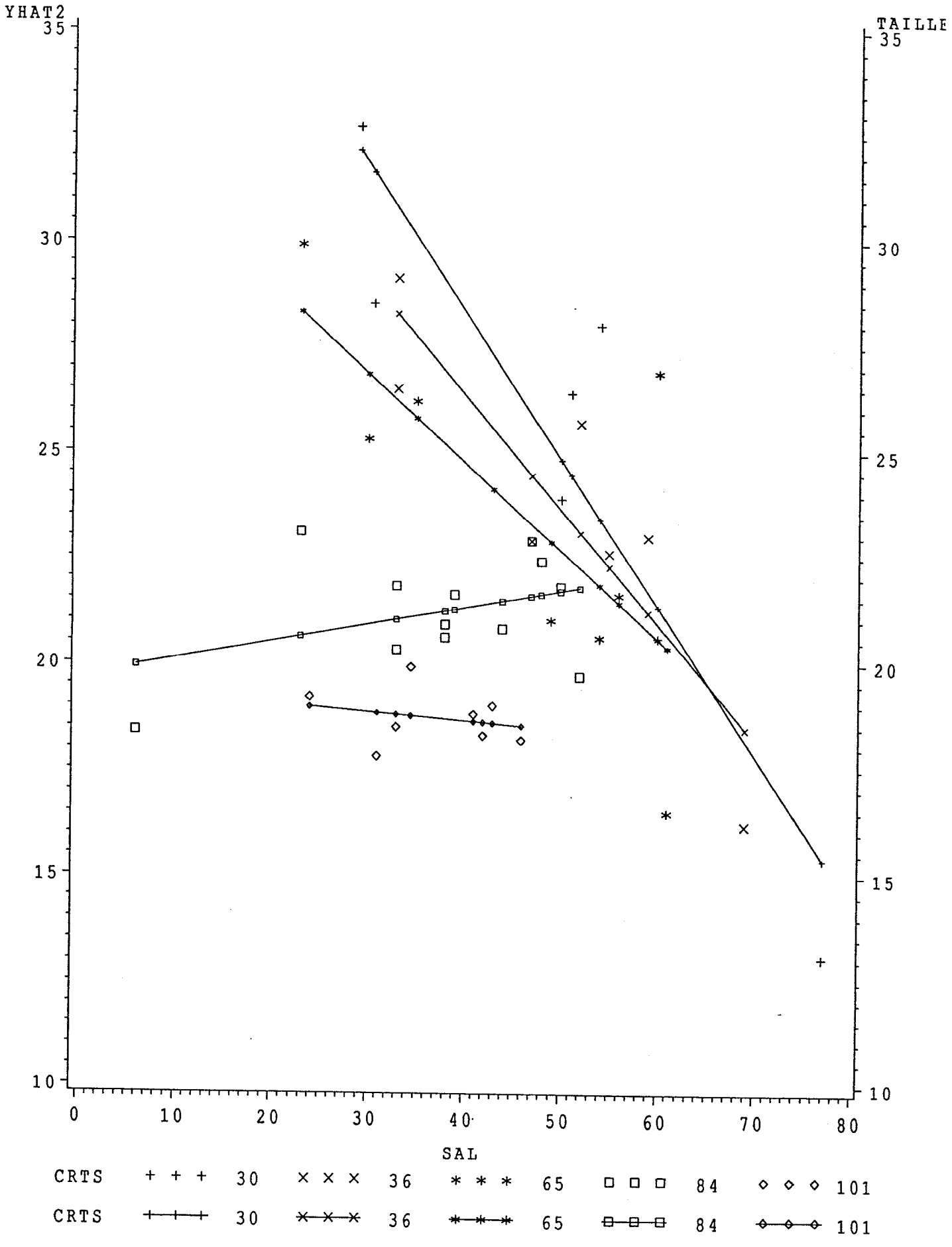
MODELE 2 : TAILLE = F(CRTS, SAL*CRTS)

RESIDUS EN FONCTION DE LA SALINITE



MODELE 2 : TAILLE = F(CRTS, SAL*CRTS)

VALEURS DE TAILLE PREDITES EN FONCTION DE LA SALINITE



MODELE 3 : TAILLE = F(CRTS, SAL*NC)

General Linear Models Procedure
Class Level Information

Class	Levels	Values
CRTS	5	30 36 65 84 101
NC	2	1 2

Number of observations in data set = 50

NOTE: Due to missing values, only 42 observations can be used in this analysis.

MODELE 3 : TAILLE = F(CRTS, SAL*NC)

General Linear Models Procedure

Dependent Variable: TAILLE

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	516.745833	86.124306	18.26	0.0001
Error	35	165.059881	4.715997		
Corrected Total	41	681.805714			

R-Square	C.V.	Root MSE	TAILLE Mean
0.757908	9.794743	2.17163	22.1714

Source	DF	Type I SS	Mean Square	F Value	Pr > F
CRTS	4	184.302083	46.075521	9.77	0.0001
SAL*NC	2	332.443750	166.221875	35.25	0.0001

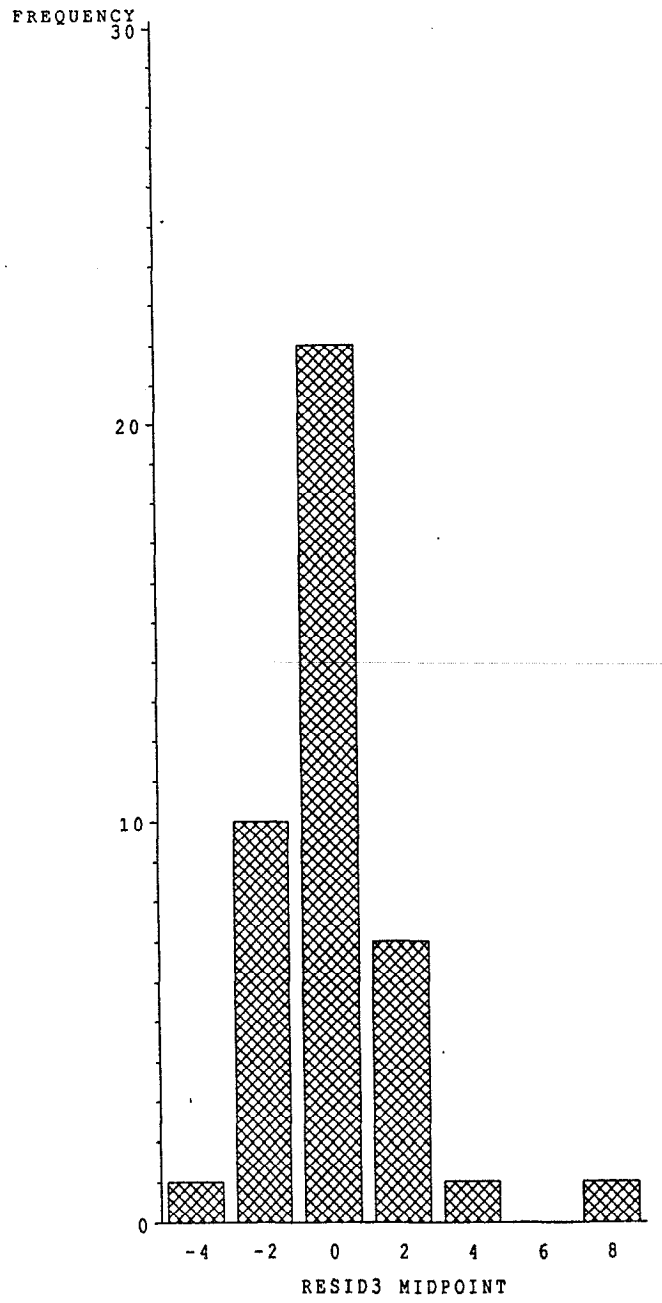
Source	DF	Type III SS	Mean Square	F Value	Pr > F
CRTS	4	311.639829	77.909957	16.52	0.0001
SAL*NC	2	332.443750	166.221875	35.25	0.0001

Parameter		Estimate	T for H0: Parameter=0	Pr > |T|	Std Error of Estimate
INTERCEPT		17.61994963 B	9.50	0.0001	1.85546210
CRTS	30	21.16764140 B	8.04	0.0001	2.63275185
	36	19.99922198 B	7.63	0.0001	2.62213712
	65	18.74606548 B	7.41	0.0001	2.52873558
	84	2.45628927 B	2.48	0.0182	0.99184189
	101	0.00000000 B	.	.	.
SAL*NC	1	-0.27969598	-8.37	0.0001	0.03341230
	2	0.02967879	0.65	0.5220	0.04588533

NOTE: The X'X matrix has been found to be singular and a generalized inverse was used to solve the normal equations. Estimates followed by the letter 'B' are biased, and are not unique estimators of the parameters.

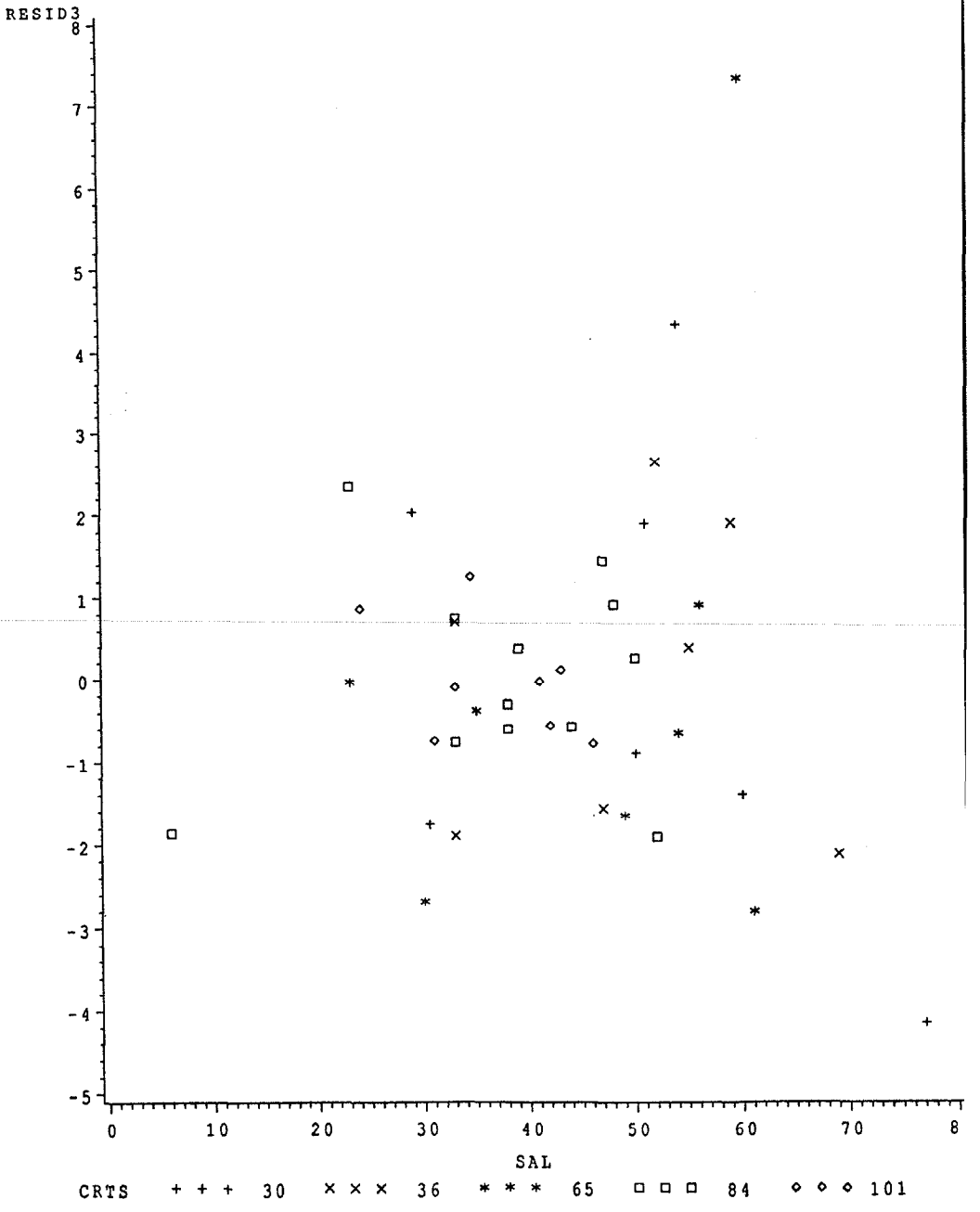
MODELE 3 : TAILLE = F(CRTS, SAL*NC)

HISTOGRAMME DES RESIDUS



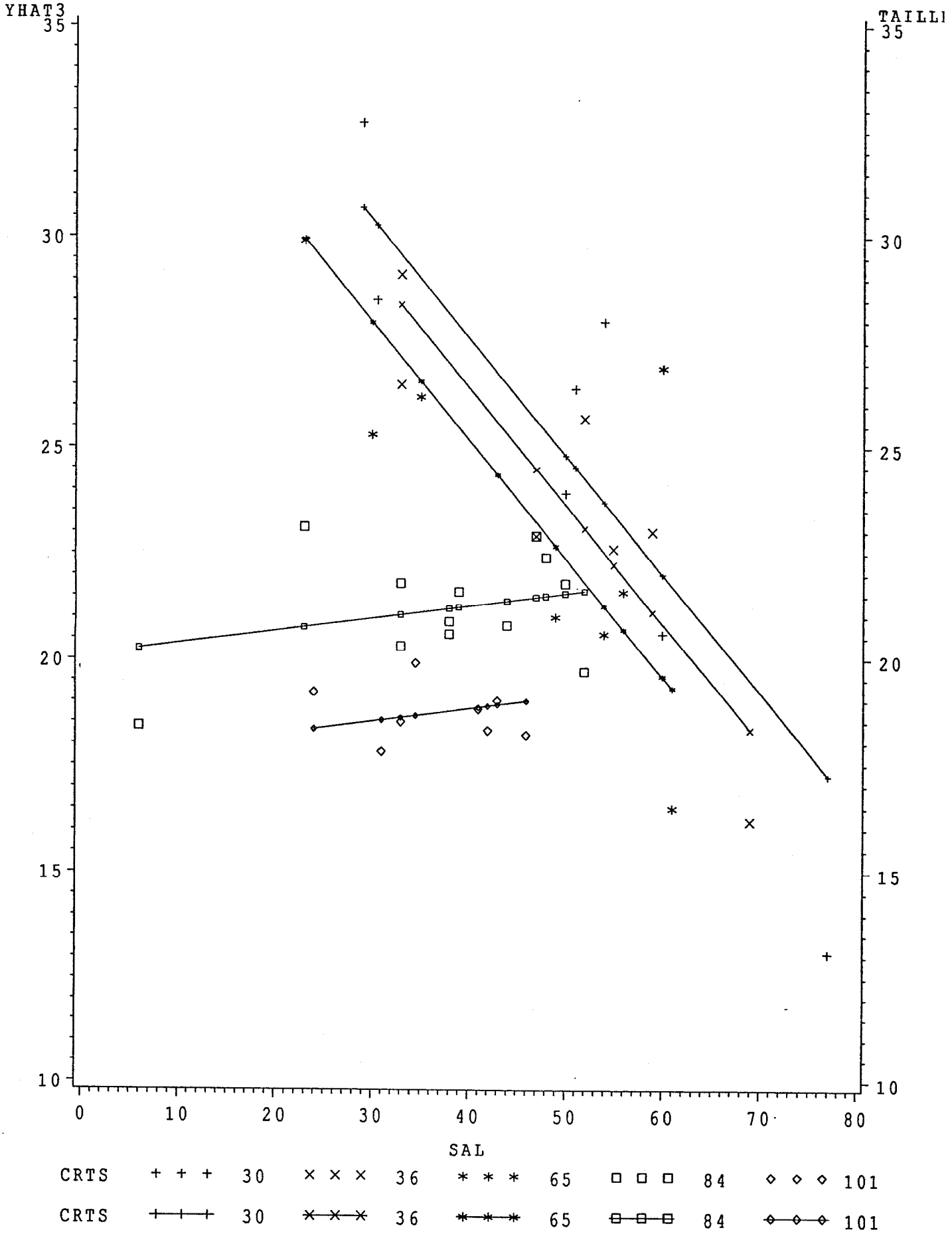
MODELE 3 : TAILLE = F(CRTS, SAL*NC)

RESIDUS EN FONCTION DE LA SALINITE



MODELE 3 : TAILLE = F(CRTS, SAL*NC)

VALEURS DE TAILLE PREDITES EN FONCTION DE LA SALINITE



MODELE 4 : TAILLE = F(CRTS, SAL)

General Linear Models Procedure
Class Level Information

Class	Levels	Values
CRTS	5	30 36 65 84 101

Number of observations in data set = 50

NOTE: Due to missing values, only 42 observations can be used in this analysis.

MODELE 4 : TAILLE = F(CRTS, SAL)

General Linear Models Procedure

Dependent Variable: TAILLE

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	376.645693	75.329139	8.89	0.0001
Error	36	305.160021	8.476667		
Corrected Total	41	681.805714			

R-Square	C.V.	Root MSE	TAILLE Mean
0.552424	13.13164	2.91147	22.1714

Source	DF	Type I SS	Mean Square	F Value	Pr > F
CRTS	4	184.302083	46.075521	5.44	0.0016
SAL	1	192.343610	192.343610	22.69	0.0001

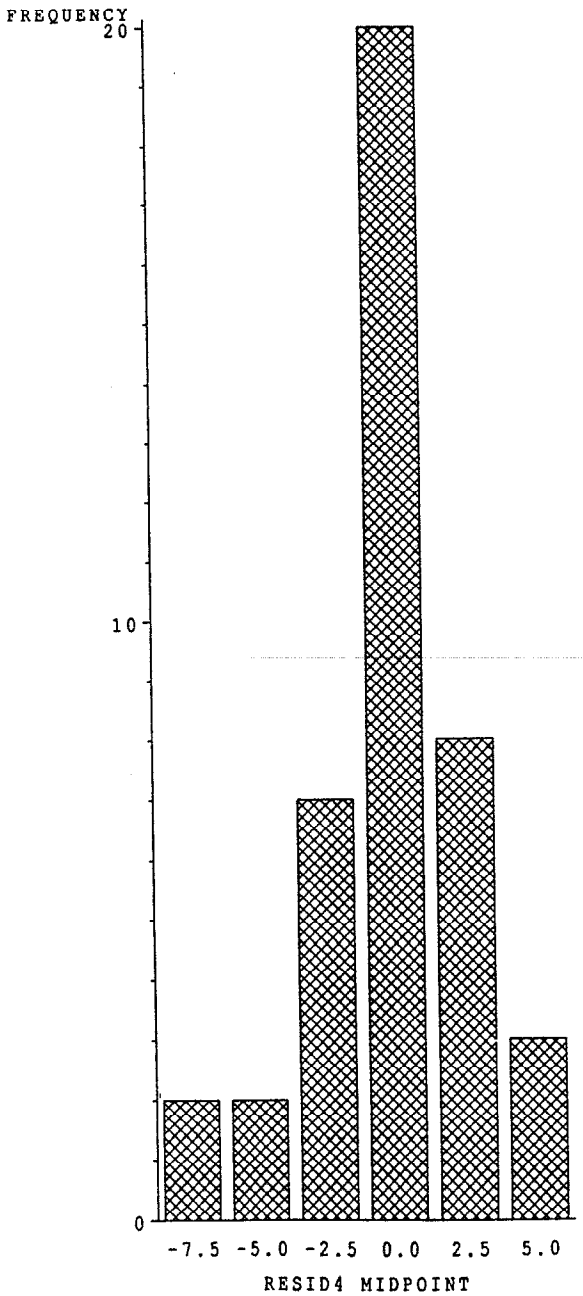
Source	DF	Type III SS	Mean Square	F Value	Pr > F
CRTS	4	323.658376	80.914594	9.55	0.0001
SAL	1	192.343610	192.343610	22.69	0.0001

Parameter		Estimate	T for H0: Parameter=0	Pr > |T|	Std Error of Estimate
INTERCEPT		25.06252407 B	14.88	0.0001	1.68422894
CRTS	30	8.34211706 B	5.27	0.0001	1.58305357
	36	7.22729742 B	4.58	0.0001	1.57759705
	65	6.37231076 B	4.27	0.0001	1.49326998
	84	2.61213265 B	1.97	0.0571	1.32919205
	101	0.00000000 B	.	.	.
SAL		-0.17249641	-4.76	0.0001	0.03621210

NOTE: The X'X matrix has been found to be singular and a generalized inverse was used to solve the normal equations. Estimates followed by the letter 'B' are biased, and are not unique estimators of the parameters.

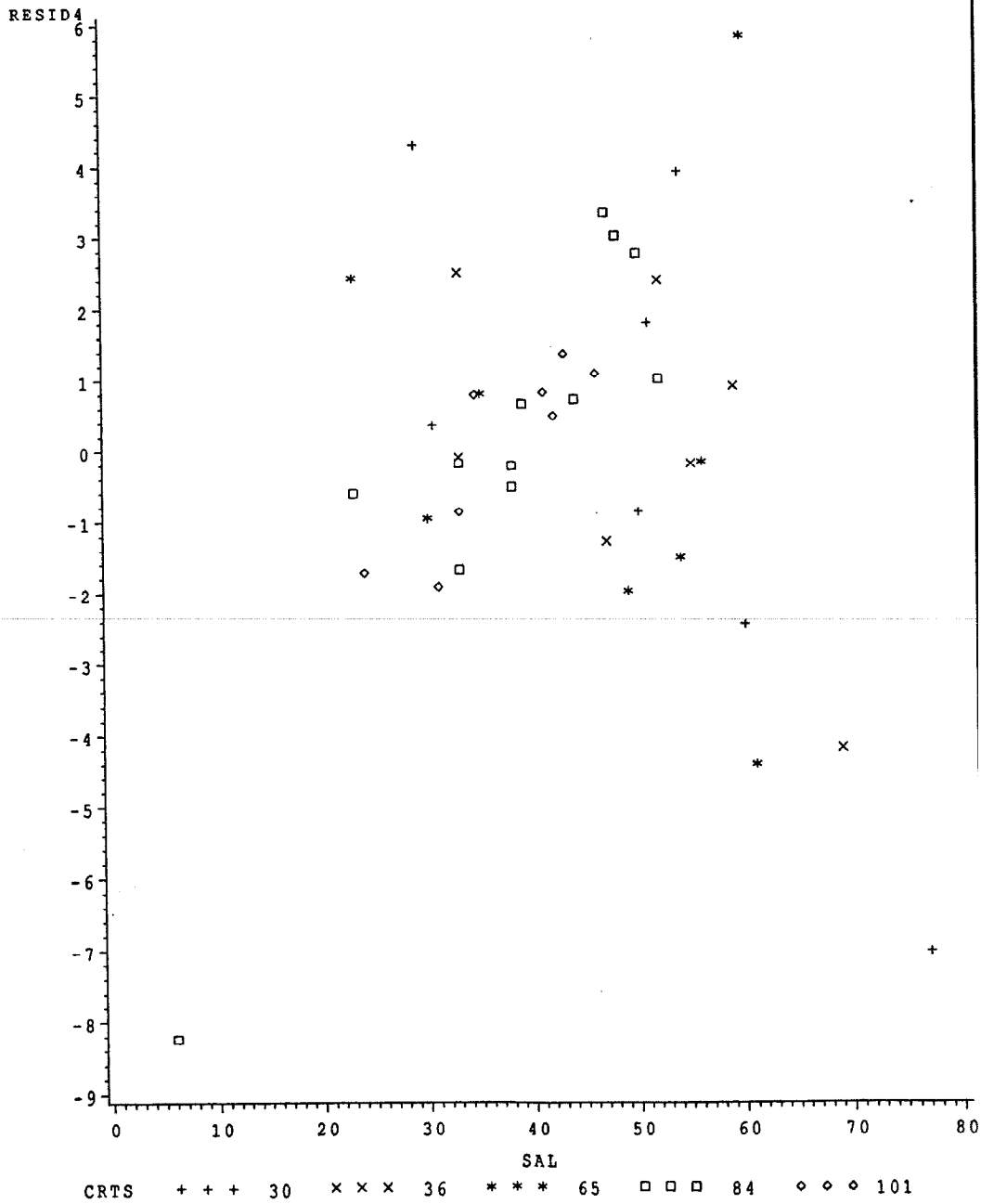
MODELE 4 : TAILLE = F(CRTS, SAL)

HISTOGRAMME DES RESIDUS



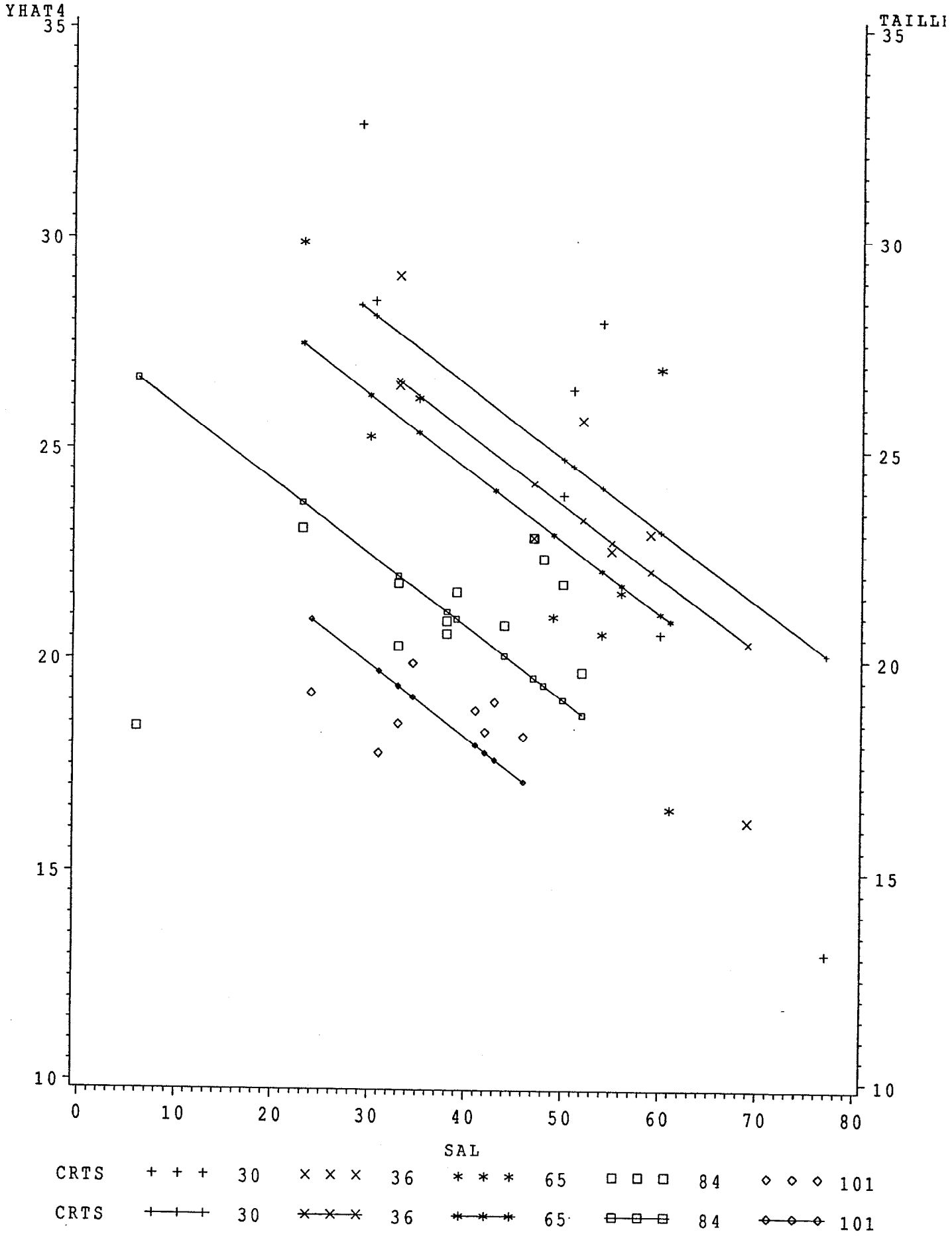
MODELE 4 : TAILLE = F(CRTS, SAL)

RESIDUS EN FONCTION DE LA SALINITE



MODELE 4 : TAILLE = F(CRTS, SAL)

VALEURS DE TAILLE PREDITES EN FONCTION DE LA SALINITE



MODELE 5 : TAILLE = F(NC, SAL*NC)

General Linear Models Procedure
Class Level Information

Class	Levels	Values
CRTS	5	30 36 65 84 101
NC	2	1 2

Number of observations in data set = 50

NOTE: Due to missing values, only 42 observations can be used in this analysis.

MODELE 5 : TAILLE = F(NC, SAL*NC)

General Linear Models Procedure

Dependent Variable: TAILLE

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	466.213439	155.404480	27.39	0.0001
Error	38	215.592275	5.673481		
Corrected Total	41	681.805714			

R-Square	C.V.	Root MSE	TAILLE Mean
0.683792	10.74314	2.38191	22.1714

Source	DF	Type I SS	Mean Square	F Value	Pr > F
NC	1	148.394805	148.394805	26.16	0.0001
SAL*NC	2	317.818634	158.909317	28.01	0.0001

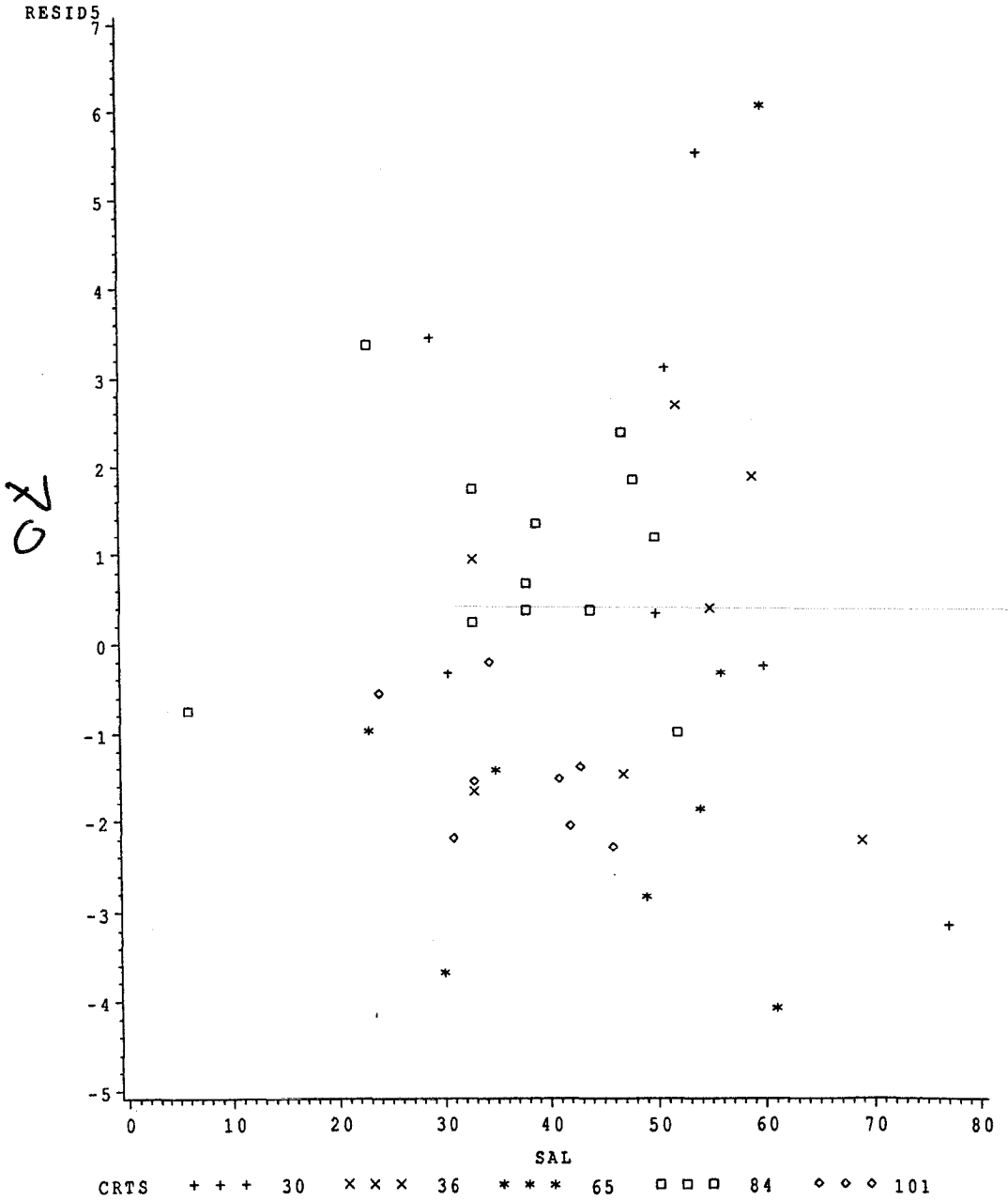
Source	DF	Type III SS	Mean Square	F Value	Pr > F
NC	1	261.107435	261.107435	46.02	0.0001
SAL*NC	2	317.818634	158.909317	28.01	0.0001

Parameter		Estimate	T for H0: Parameter=0	Pr > |T|	Std Error of Estimate
INTERCEPT		18.94267330 B	9.72	0.0001	1.94897956
NC	1	18.15159901 B	6.78	0.0001	2.67565472
	2	0.00000000 B	.	.	.
SAL*NC	1	-0.27060795	-7.45	0.0001	0.03630163
	2	0.03373110	0.67	0.5065	0.05029625

NOTE: The X'X matrix has been found to be singular and a generalized inverse was used to solve the normal equations. Estimates followed by the letter 'B' are biased, and are not unique estimators of the parameters.

MODELE 5 : TAILLE = F(NC, SAL*NC)

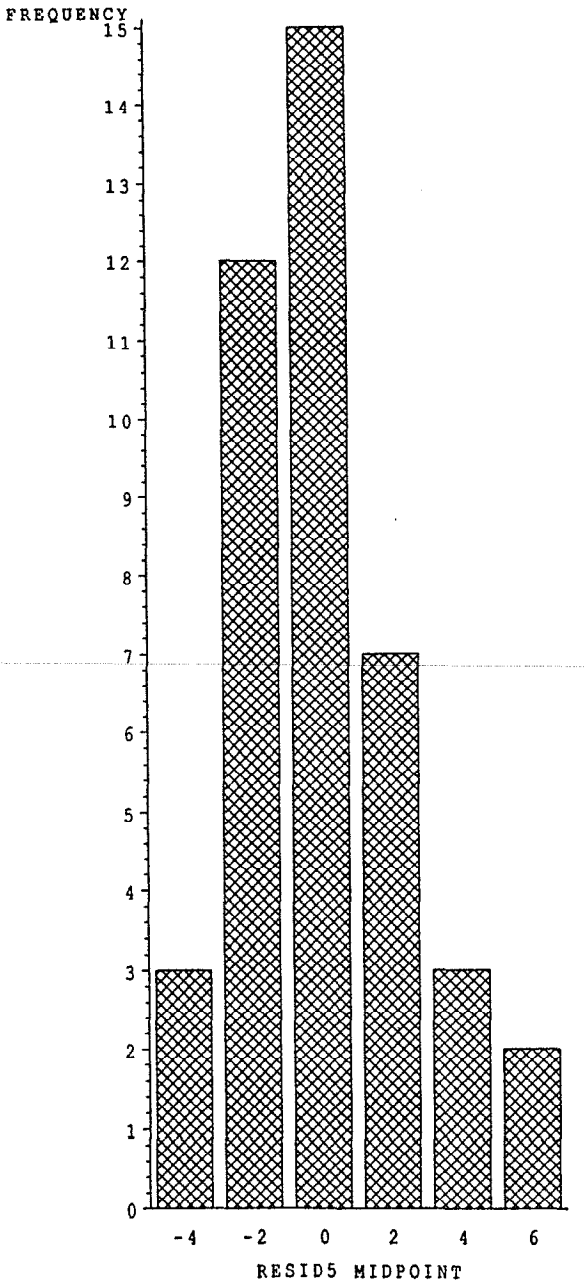
RESIDUS EN FONCTION DE LA SALINITE



02

MODELE 5 : TAILLE = F(NC, SAL*NC)

HISTOGRAMME DES RESIDUS



MODELE 5 : TAILLE = F(NC, SAL*NC)

VALEURS DE TAILLE PREDITES EN FONCTION DE LA SALINITE