Regressions et correlations multiples en hydrologie

P. TOUCHEBEUF DE LUSSIGNY
Ingénieur en Chef à E.D.F.

Résumé

Le calcul automatique permet maintenant une plus large utilisation en hydrologie des régressions et corrélations multiples. Leur théorie est ici exposée dans ses grandes lignes et un programme FORTRAN est proposé pour la sélection des variables explicatives dans l'application pratique des régressions linéaires multiples.

Abstract

The automatic computers make it possible now to use on a larger scale in hydrology the multiple regressions and correlations. The main outlines of their theory are explained here and a FORTRAN programme is proposed for the selection of the independent variables in the practical application of the multiple linear regressions.

Sommaire

Introduction

Les régressions et corrélations multiples sont d'un grand intérêt en hydrologie pour étudier les relations entre les variables hydrologiques, climatologiques, morphologiques, etc.

Leur utilisation peut se classer sous trois rubriques:

a) extension dans le temps de séries d'observations hydrologiques qui sont de trop courte durée ou comportent des lacunes;

b) prévisions de données hydrologiques (apports mensuels, crues, étiages, etc.) en fonction des conditions hydro-météorologiques observées au moment de la prévision;

c) extension géographique à des bassins non observés des caractéristiques hydrologiques déterminées sur divers bassins versants de régime analogue.

La théorie complète des régressions et corrélations multiples est d'un abord assez difficile et il existe déjà à son sujet une abondante littérature. Nous nous bornerons ici à en donner les grandes lignes et insisterons plus particulièrement sur les points qui sont essentiels pour une application correcte de la théorie.

2. Ajustement analytique d'une relation quelconque

(Méthode des moindres carrés)

Considérons une population finie qui comprenne n observations, ces observations portant elles-mêmes sur k variables X_1, X_2, ..., X_k.

L'approximation de X_1 en fonction des autres variables aléatoires s'appelle la régression de X_1 en X_2, ..., X_k. Elle peut s'envisager sous la forme de la relation stochastique suivante:

$$X_1 = f(X_2, X_3, \ldots, X_k) + \varepsilon \qquad (2 - 1)$$

On appelle:

— X_1 la variable dépendante ou variable à expliquer;

— X_2, X_3, ..., X_k les variables indépendantes, ou explicatives;

— ε le résidu.

A chacune des n observations correspond une valeur particulière du résidu qui constitue une variable aléatoire de moyenne nulle. Ce résidu provient de trois causes:

a) erreurs aléatoires des mesures;

b) non prise en compte dans les variables explicatives de tous les facteurs conditionnels;

c) imperfection de la forme analytique de la fonction choisie.

Si l'on connaît la forme analytique de cette fonction f, le problème revient à estimer un certain nombre de paramètres d'ajustement, dits « coefficients de régression », à partir de l'échantillon.

On peut écrire la relation (2 — 1) sous une forme plus explicite:

$$X_{1i} = f(X_{2i}, \ldots, X_{ki}, b_0, b_2, \ldots, b_m) + \varepsilon_i$$

dans laquelle:

— b_0, b_2, ..., b_m sont les paramètres d'ajustement;

— et i un indice variant de 1 à n qui caractérise chacune des observations.

Pour déterminer les paramètres, on utilise la méthode des moindres carrés qui consiste à minimiser la somme des carrés des résidus, en annulant les dérivées partielles de cette somme par rapport aux coefficients b_0, b_2, ..., b_m (voir « Hydrologie de Surface » de M. ROCHE, p. 49).

3. Hypothèses fondamentales de l'analyse de régression et de corrélation multiple

Il y a d'abord lieu d'établir une distinction entre « régression multiple » et « corrélation multiple », car il existe entre ces deux types d'analyse une différence importante dans leur interprétation statistique, comme on le verra plus loin.

Dans le cas de la régression multiple, la variable dépendante est, bien entendu, une variable aléatoire, mais il n'est pas nécessaire que toutes les variables explicatives soient aléatoires. Dans le cas de la corrélation multiple toutes les variables doivent obligatoirement être aléatoires. Si de plus la corrélation multiple est linéaire, il est nécessaire que toutes les variables aléatoires aient chacune une distribution propre (appelée « distribution marginale ») qui soit normale. Autrement dit, l'échantillon total des valeurs observées doit être tiré d'une distribution normale à k variables.

Les hypothèses fondamentales de l'analyse de régression et de corrélation multiple sont au nombre de trois:

a) les variables explicatives doivent être connues avec une erreur de mesure négligeable par rapport à leur variabilité. Seule la variable dépendante peut être entachée de certaines erreurs aléatoires de mesure;

b) la variable dépendante doit pouvoir être considérée comme une variable aléatoire « intérieurement indépendante », c'est-à-dire que l'auto-corrélation des valeurs observées successives doit être négligeable. Dans le cas de la corrélation multiple, cette condition d'indépendance intérieure doit être également remplie par toutes les variables explicatives;

c) les écarts de la variable dépendante autour de l'hyper-surface de régression (*), c'est-à-dire les résidus, doivent être distribués normalement et avec le même écart-type en tout point de l'hyper-surface. Cette condition, dite d'« homoscédasticité », peut encore s'énoncer de la façon suivante: la distribution de X_1 liée par X_2, ..., X_k doit être normale et de même variance quels que soient les éléments ΔX_2, ..., ΔX_k considérés (voir fig. 1 pour $k = 2$).

Dans le cas de la corrélation, la condition d'homoscédasticité doit être remplie non seulement par rapport à la variable X_1 mais par rapport à toutes les autres variables.

La première hypothèse — précision des variables explicatives — n'est pas toujours facile à respecter rigoureusement en hydrologie par suite des erreurs ou des lacunes

(*) Cette hyper-surface dans un espace à k dimensions est définie par la relation:

$$X1 = f(X_2, ..., X_k, b_0, b_2, ..., b_m)$$

Elle devient un hyperplan dans le cas d'une régression linéaire. Elle se réduit à un plan si $k = 3$ et à une droite si $k = 2$.

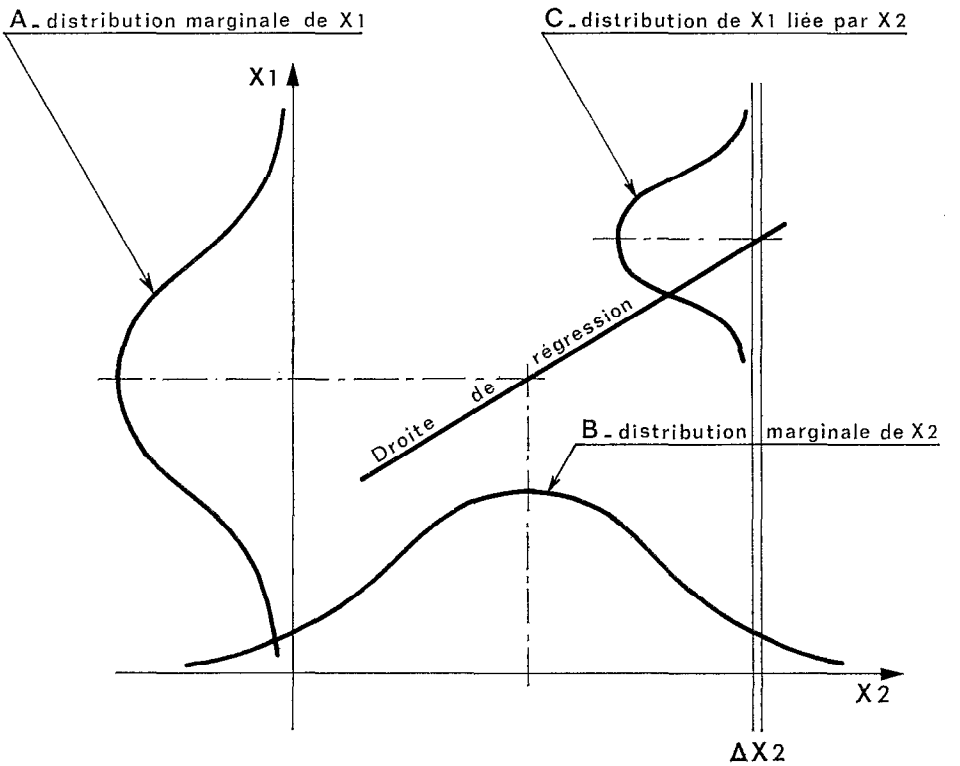


Fig. 1

DISTRIBUTION NORMALE A DEUX VARIABLES
Homoscédasticité : la distribution C reste constante quel que soit l'élément $\Delta \times 2$ considéré

d'observation, dans les données pluviométriques notamment. Certains caractères physiques du bassin versant, comme la densité de la couverture végétale, sont également difficiles à définir par un indice quantitatif précis.

La seconde hypothèse — absence d'auto-corrélation — ne peut être satisfaite que si les observations successives de la variable dépendante (et de toutes les variables explicatives pour une étude de corrélation) sont suffisamment éloignées, dans le temps ou dans l'espace suivant le problème étudié.

La troisième hypothèse — normalité et homoscédasticité des résidus — n'est pas toujours remplie par les données hydrologiques. Il est fréquent que les résidus aient une distribution dissymétrique et croissent avec la variable dépendante. On parvient cependant à satisfaire cette hypothèse de façon plus ou moins approximative en ayant recours à des transformations de variables, comme on le verra plus loin. On ne peut guère se montrer très exigeant sur ce point, car le nombre d'observations est le plus souvent insuffisant pour vérifier correctement la normalité des résidus et l'homogénéité de leur variance.

4. Choix d'un modèle d'ajustement

Avant de procéder à l'étude analytique d'une relation entre variables la question se pose de savoir quelle est la forme analytique de la fonction:

$$X_1 = f(X_2, X_3, \ldots, X_k)$$

La forme de cette fonction doit être aussi simple que possible et permettre d'expliquer le maximum de la variance de la variable dépendante avec le minimum de variables explicatives ($2k < n$).

Des considérations théoriques et des essais graphiques préalables guident le choix de la fonction qui paraît à priori la mieux appropriée.

4.1 *Ajustement linéaire*

C'est le modèle le plus simple auquel on doit avoir recours dans toute la mesure du possible, quitte à effectuer des transformations de variables si nécessaire.

Une condition suffisante pour que le modèle linéaire soit applicable est que toutes les variables prises en considération soient des variables aléatoires tirées d'une distribution normale à k variables. Si cette condition est remplie, il est possible d'effectuer une étude complète de régression et de corrélation multiple et d'appliquer tous les tests de signification.

4.2 *Ajustement curviligne*

S'il n'est pas possible de normaliser la distribution des variables par transformation, on pourra tenter un ajustement curviligne soit par une méthode graphique, soit par une méthode analytique.

4.2.1. Ajustement graphique

En dehors de la méthode coaxiale développée par Linsley (pour quatre variables ou plus), la méthode la plus employée est celle bien connue à l'O.R.S.T.O.M. des déviations résiduelles (voir « Hydrologie de Surface » de M. Roche, p. 51).

Cette méthode est très souple puisqu'on ne s'impose pas de donner aux liaisons entre variables une forme analytique quelconque, mais elle présente de sérieux inconvénients. Elle est assez laborieuse et en partie subjective. Le choix de l'ordre d'importance des différentes variables explicatives n'est pas toujours évident et peut avoir une influence non négligeable sur le résultat final. Chose plus grave, on n'a aucune possibilité de tester objectivement la validité des liaisons obtenues. La souplesse même de la méthode revient à augmenter aveuglément le nombre p de paramètres d'ajustement et à réduire ainsi de façon inconsidérée le nombre ν de degré de liberté ($\nu = n - p$) de la régression multiple. Or, comme on le verra plus loin, lorsque ν se rapproche de zéro les liaisons obtenues perdent toute signification.

La méthode peut fournir des ajustements, plus ou moins sinueux, qui sont excellents en apparence. En réalité leur qualité est assez illusoire, car ces ajustements épousent trop étroitement des écarts accidentels qu'ils devraient négliger parce que liés à des facteurs secondaires qui n'ont pas été englobés dans les variables explicatives.

On a cherché à remédier aux inconvénients de la méthode des déviations résiduelles qui conserve un intérêt certain pour nos hydrologues d'outre-mer qui ne disposent

pas encore de moyens de calcul automatique. On y parvient dans une bonne mesure en sacrifiant la souplesse de la méthode. On s'impose de tracer des courbes d'ajustement de forme très simple qui ne comportent jamais plus d'un point d'inflexion, tout en s'assurant, notamment dans leur extrapolation, qu'elles sont compatibles avec les réalités physiques. On peut même s'imposer un ajustement de forme analytique simple (exponentielle par exemple, en utilisant un graphique semi-logarithmique). On limite également de façon sévère le nombre des variables explicatives: $k = 2$ pour $n < 15$, $k = 3$ pour $15 \leqslant n < 30$, $k = 4$ pour $30 \leqslant n < 60$ et $k = 5$ pour $n \geqslant 60$.

4.2.2. Ajustement analytique

De nombreuses possibilités s'offrent pour des mathématiciens avertis. Nous n'insisterons pas sur ce point et indiquerons seulement qu'on peut en particulier utiliser des polynomes.

5. Normalisation des variables

Lorsqu'une variable X présente une distribution nettement dissymétrique, ce qui est assez fréquent en hydrologie, on peut chercher à définir une nouvelle variable $Z = f(X)$ qui, elle, ait une distribution sensiblement normale ainsi qu'une variance stabilisée et qui puisse donc être introduite dans un modèle d'ajustement linéaire.

5.1 *Transformation logarithmique*

La normalisation des variables peut s'effectuer par une transformation logarithmique de la forme:

$$Z = \log (X - X_0) \qquad \text{avec } X > X_0$$

Cette transformation qui a l'avantage de la simplicité, est la plus fréquemment employée.

Le paramètre X_0 constitue une borne inférieure de la variable X, dont la détermination peut apparaître assez arbitraire lorsque le nombre des observations est réduit. On lui donne parfois une valeur égale à zéro par souci de simplicité, mais cette valeur nulle ne convient pas lorsque les observations de la variable X comportent elles-mêmes des valeurs nulles, comme ce peut être le cas pour des études en région aride. On préfère alors une transformation par puissance fractionnaire.

5.2 *Transformation par puissance fractionnaire*

Cette transformation est de la forme:

$$Z = X^q \qquad \text{avec } X \geqslant 0$$

q étant un nombre compris entre 0 et 1.

La valeur de q peut être déterminée en fonction des coefficients β_1 et β_2 de Pearson de la variable X, en utilisant l'abaque que M. Milu Rosenberg a établi par des considérations théoriques (voir fig. 2).

Rappelons que β_1 et β_2 mesurent respectivement les degrés d'asymétrie et d'aplatissement d'une distribution. Ils ont pour expression:

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} \qquad \text{et} \qquad \beta_2 = \frac{\mu_4}{\mu_2^2}$$

avec:

$$\frac{\mu}{k} = \frac{1}{n} \sum_{1}^{n} (x_i - \bar{x})^k$$

(Pour une distribution normale $\beta_1 = 0$ et $\beta_2 = 3$.)

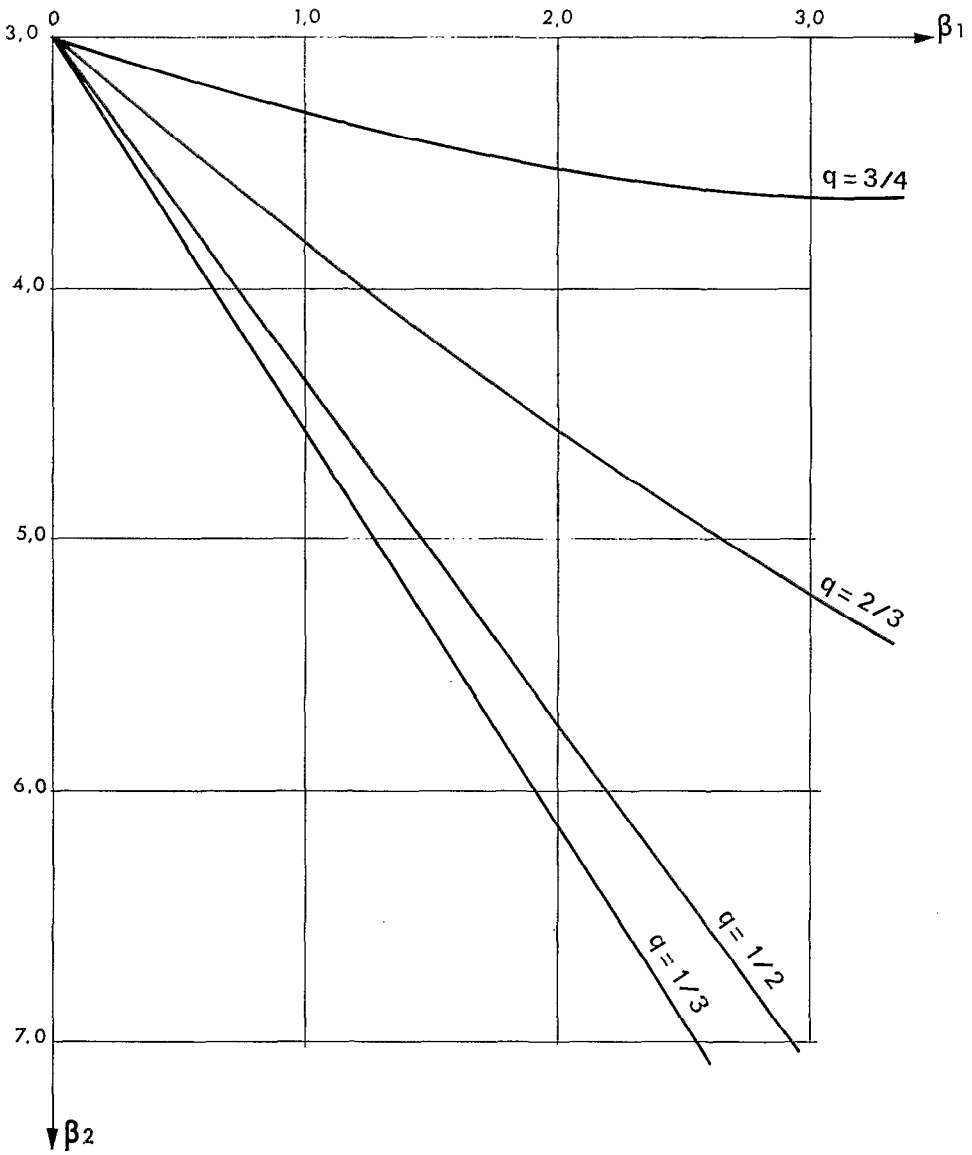


Fig. 2

TRANSFORMATION PAR PUISSANCE FRACTIONNAIRE q
Valeur de q en fonction de β1 et de β2 (d'après M. Milu Rosenberg)

On doit noter que β_1 et β_2 sont entachés de fortes erreurs d'échantillonnage lorsque le nombre d'observations est réduit. L'abaque ne donne donc qu'une indication préliminaire sur la valeur de q dans le cas où $n < 50$.

5.3 *Anamorphose*

D'une façon plus générale, on peut toujours passer d'une loi de distribution quelconque F(X) à une autre loi de distribution G(Z) fixée à l'avance. La loi F(X) est, par exemple, une distribution de PEARSON III et la loi G(Z) une distribution normale de GAUSS.

A toute valeur de X on peut faire correspondre une valeur de Z telle que les fréquences cumulées F(X) et G(X) soient égales. Cette opération appelée « anamorphose » peut facilement être réalisée par ordinateur pour toutes les lois couramment utilisées en hydrologie, en utilisant les sous-programmes mis au point par Y. BRUNET-MORET. Dans l'exemple cité, la FUNCTION FGAMA permet de calculer la fréquence F correspondant à une valeur donnée de X, puis la FUNCTION VNORM calcule la valeur de Z correspondant à la fréquence F (voir *Cah. ORSTOM sér. Hydrol.*, vol. VII, n° 3, 1969).

Les transformations de variables indiquées aux paragraphes 5.1 et 5.2 ne sont en fait que des cas particuliers d'anamorphose où l'on peut expliciter la relation liant Z à X.

6. Distribution normale à k variables

Soit une distribution continue à k variables aléatoires $X_1, X_2, \ldots, X_k$ distribuées normalement avec des écarts-types respectifs $\sigma_1, \sigma_2, \ldots, \sigma_k$ et mesurées par rapport à leurs moyennes respectives.

Si les k variables sont indépendantes, la densité de probabilité de la distribution est donnée par:

$$f(X_1, X_2, \ldots, X_n) = \frac{e^{-\frac{1}{2}\varnothing}}{2\pi^{k/2}\,\sigma_1.\sigma_2 \ldots \sigma_k}$$

avec:

$$\varnothing = \frac{X_1^2}{\sigma_1^2} + \frac{X_2^2}{\sigma_2^2} + \ldots + \frac{X_k^2}{\sigma_k^2}$$

Dans le cas plus général où les k variables ne sont pas indépendantes, c'est-à-dire présentent entre elles prises deux à deux des coefficients de corrélation significatifs, la densité de probabilité peut s'exprimer sous la forme suivante:

$$f = \frac{e^{-\frac{1}{2}\varnothing}}{2\pi^{k/2}\,\sigma_1.\sigma_2 \ldots \sigma_k \sqrt{\Delta}}$$

avec:

$$\varnothing = \frac{1}{\Delta}\left\{\sum_{i=1}^{k} \Delta_{ii}\frac{X_i^2}{\sigma_i^2} + 2\sum_{i,j=1}^{k}\frac{X_i\ X_j}{\sigma_i \cdot \sigma_j}\cdot\Delta_{ij}\right\}$$

pour $i \neq j$.

On désigne par Δ le déterminant d'ordre k:

$$\Delta = \begin{vmatrix} 1 & r_{12} & \ldots & r_{1k} \\ r_{21} & 1 & \ldots & r_{2k} \\ & & & \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \\ r_{k1} & r_{k2} & \ldots & 1 \end{vmatrix}$$

et par Δ_{ij} son mineur obtenu en supprimant la i^e ligne et la j^e colonne.

r_{ij} est le coefficient de corrélation simple (ou « totale ») entre les variables X_i et X_j. Si n est le nombre d'observations, on a:

$$r_{ij} = \frac{\sum_{l=1}^{n} X_{il} \cdot X_{jl}}{\sqrt{\sum_{l=1}^{n} X^2_{il}} \cdot \sqrt{\sum_{l=1}^{n} X^2_{jl}}}$$

On notera que l'expression de la densité de probabilité est symétrique par rapport à tous les indices. On peut montrer que la régression de n'importe quelle variable avec une partie ou toutes les autres est strictement linéaire. Les résidus de n'importe quel ordre sont distribués normalement et de façon homoscédastique.

7. Définitions et notations

La régression linéaire multiple de X_1 en X_2, ..., X_k peut s'exprimer sous la forme:

$$X_1 = b_0 + b_2 X_2 + \ldots + b_k X_k + \varepsilon \qquad (7.1)$$

Si l'on rapporte chacune des variables aléatoires à sa moyenne arithmétique, la relation (7.1) s'écrit:

$$X_1 = b_{12.34\ldots K} X_2 + b_{13.24\ldots K} X_3 + \ldots + b_{1K.23..(K-1)} X_K + \varepsilon$$

On désigne par $b_{12.34\ldots K}$ le coefficient de régression partielle de X_1 par rapport à X_2 compte tenu des autres variables. Les indices 1 et 2 sont dits primaires et les indices 3, 4, ..., K secondaires.

Les résidus ε, écarts entre les valeurs observées X_1 et les valeurs correspondantes $X'_{1.23\ldots K}$ estimées par la régression, sont notés $X_{1.23\ldots K}$:

$$X_{1.23\ldots K} = \varepsilon = X_1 - X'_{1.23\ldots K}$$

La « variance résiduelle » est définie par la relation:

$$S^2_{1.23\ldots K} = \frac{\sum_{i=1}^{n} \varepsilon_i^2}{n}$$

Le « coefficient de corrélation partielle » $r_{12.3\ldots K}$ est tel que:

$$r_{12.34\ldots k} = \sqrt{b_{12.34\ldots k} \cdot b_{21.34\ldots k}}$$

On verra plus loin ce que représente ce coefficient.

Le nombre p d'indices secondaires du coefficient de régression, de la variance résiduelle et du coefficient de corrélation partielle définit leur ordre; ainsi l'ordre de $S_1{}^2{}_{.23\ldots k}$ est de k — 1, tandis que celui de $r_{12.34\ldots k}$ est de k — 2.

Les « écarts-types marginaux » de chacune des k variables sont désignées par $S_1, S_2, \ldots, S_k$:

$$S_1 = \sqrt{\frac{\sum_{i=1}^{n} X_{1i}^2}{n}} \qquad S_k = \sqrt{\frac{\sum X_{ki}^2}{n}}$$

On écrit parfois l'équation de régression (7.1) sous une forme non dimensionnelle:

$$\frac{X_1}{S_1} = \beta_0 + \beta_2 \cdot \frac{X_2}{S_2} + \ldots + \beta_K \cdot \frac{X_K}{S_K} + \varepsilon$$

Les *coefficients* β sont des paramètres sans dimension qui mesurent l'effet des variables explicatives sur la variable indépendante. Ils sont liés aux coefficients b par les relations:

$$\beta_0 = \frac{b_0}{S_1}, \qquad \beta_i = b_1 \cdot \frac{S_i}{S_1}$$

8. Régression linéaire et corrélation avec deux variables

Le cas de la régression linéaire simple est bien connu. Nous rappellerons brièvement les résultats classiques avec les notations qui viennent d'être définies:

En supposant les variables aléatoires X_1 et X_2 mesurées par rapport à leur moyenne, les régressions de X_1 en X_2 et de X_2 en X_1 s'écrivent:

$$X_1 = b_{12} X_2 \text{ et } X_2 = b_{21} X_1$$

L'application de la méthode des moindres carrés conduit aux résultats suivants:

a) coefficients de régression:

$$b_{12} = \frac{\sum_{i=1}^{n} (X_1 X_2)_i}{\sum_{i=1}^{n} (X_2^2)_i} = r_{12} \cdot \frac{S_1}{S_2}$$

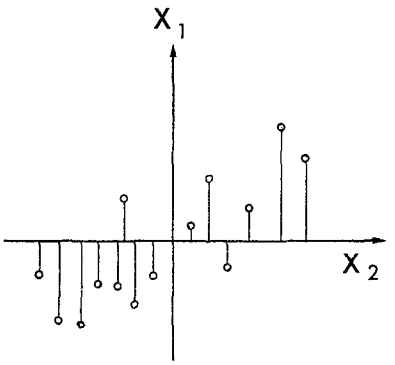
$$b_{21} = \frac{\Sigma (X_1 X_2)_i}{\Sigma (X_1^2)_i} = r_{21} \cdot \frac{S_2}{S_1}$$

$$S_1^2 = S'^2_{1-2} + S^2_{1-2}$$

Variance marginale :

$$S_1^2 = \frac{\Sigma(X_1^2)}{n}$$

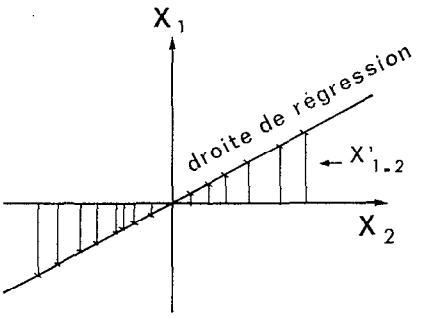
valeurs observées



Variance dûe à la régression :

$$S'^2_{1-2} = \frac{\Sigma(X'^2_{1-2})}{n}$$

valeurs calculées



Variance résiduelle :

$$S^2_{1-2} = \frac{\Sigma(X_1 - X'_{1-2})^2}{n}$$

valeurs observées

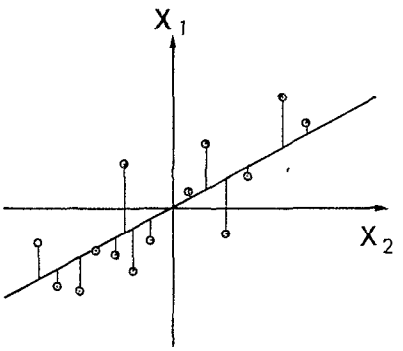


Fig. 3

RÉGRESSION LINÉAIRE A DEUX VARIABLES
Analyse de la variance

b) coefficients de corrélation:

$$r_{12} = r_{21} = \frac{\Sigma(X_1 X_2)_i}{n S_1 S_2} = \sqrt{b_{12} \cdot b_{21}}$$

c) variance résiduelle:

$$S_{1.2}^2 = \frac{\Sigma\,(X_{1.2}^2)_i}{n} = S_1^2\,(1 - r_{12}^2)$$

$$S_{2.1}^2 = \frac{\Sigma\,(X_{2.1}^2)_i}{n} = S_2^2\,(1 - r_{21}^2)$$

(toutes les sommes Σ s'entendent pour i variant de 1 à n).

9. Régression linéaire multiple (k > 2)

Si les k variables sont mesurées à partir de leurs moyennes respectives, l'application de la méthode des moindres carrés pour déterminer les coefficients de régression revient à résoudre un système d'équations linéaires constitué par l'équation:

$$\sum_{i=1}^{n} (X_1 - b_{12.3\ldots k}\,X_2 - \ldots - b_{1k.23\ldots(k-1)}\,X_k)_i = 0$$

et les (k — 1) autres équations dont la forme générale est la suivante:

$$\sum_{i=1}^{n} X_j\,[X_1 - b_{12.3\ldots k}\,X_2 - \ldots - b_{1k.23\ldots(k-1)}\,X_k]_i = 0$$

j variant de 2 à k.

La résolution de ce système d'équations aboutit à l'expression suivante du coefficient de régression partielle de $X_{i(p)}$ par rapport à $X_{j(p)}$:

$$b_{ij(p)} = \frac{\sum_{l=1}^{n} [X_{i(p)} \cdot X_{j(p)}]\ l}{\sum_{l=1}^{n} [X^2_{j(p)}]\ l}$$

où (p) désigne un groupe d'indices secondaires ne comprenant ni i, ni j.

Cette expression est rarement utilisée. On lui préfère pour les calculs la forme suivante:

$$b_{ij(p)} = -\frac{S_i}{S_j} \cdot \frac{\Delta_{ij}}{\Delta_{ii}}$$

10 Analyse de corrélation multiple

Cette analyse, qui suppose remplies les hypothèses les plus restrictives énoncées au paragraphe 3, vise à mesurer le degré de liaison des variables intervenant dans une régression linéaire multiple. Le degré de liaison est défini par quatre paramètres:

— écart-type résiduel;

— coefficient de corrélation partielle;
— coefficient de corrélation multiple;
— coefficient de détermination.

10.1 ***Ecart-type résiduel***

Ce paramètre caractérise les fluctuations de la variable dépendante par rapport à l'hyper-plan de régression. On peut montrer que sa valeur est donnée par la relation suivante:

$$S_{1.23\ldots k} = \sqrt{\frac{\Delta}{\Delta_{11}} \cdot S_1^2}$$

ou d'une façon plus générale par:

$$S_{i.p} = \sqrt{\frac{\Delta}{\Delta_{ii}} \cdot S_i^2}$$

L'écart-type résiduel permet de calculer pour une probabilité donnée, l'intervalle de confiance des valeurs estimées $X'_{1.23\ldots k}$ de la variable dépendante:

— intervalle à 80% (intervalle dans lequel il y a 80 chances sur 100 que se situe la valeur vraie de la variable estimée):

$$X'_{1.23\ldots k} \pm 1{,}28\ S_{1.23\ldots k}$$

— intervalle à 90%:

$$X'_{1.23\ldots k} \pm 1{,}64\ S_{1.23\ldots k}$$

— intervalle à 95%:

$$X'_{1.23\ldots k} \pm 1{,}96\ S_{1.23\ldots k}$$

10.2 ***Coefficient de corrélation partielle***

Ce coefficient mesure la corrélation totale entre les variables X_i et X_j lorsque toutes les autres variables désignées par p sont maintenues constantes. La linéarité de la régression multiple implique que cette corrélation totale ne dépends pas des valeurs constantes attribuées aux variables désignées par p.

Le coefficient de corrélation partielle caractérise donc la liaison entre deux variables X_i et X_j, lorsque l'effet de toutes les autres variables est éliminé.

Il peut être estimé par:

$$r_{ij(p)} = \frac{\sum_{l=1}^{n} [X_{i(p)} \cdot X_{j(p)}]_l}{\sqrt{\sum_{l=1}^{n} [X_{i(p)}]_l \cdot \sum_{l=1}^{n} [X_{j(p)}]_l}}$$

ou plus commodément par:

$$r_{ij(p)} = \frac{-\Delta_{ij}}{\sqrt{\Delta_{ii} \cdot \Delta_{jj}}}$$

On notera que le coefficient de régression partielle et le coefficient de corrélation partielle sont liés par la relation:

$$b_{ij(p)} = r_{ij(p)} \frac{S_{i(p)}}{S_{j(p)}}$$

Le coefficient de corrélation partielle peut aussi être exprimé en fonction des variances résiduelles:

$$r_{ij(p)} = \sqrt{1 - \frac{S^2_{i(jp)}}{S^2_{i(p)}}}$$

Cette dernière expression montre que le coefficient de corrélation partielle mesure la diminution relative de la variance résiduelle lorsqu'on ajoute la variable X_j au groupe p des variables explicatives.

Le coefficient de corrélation partielle reste toujours compris entre — 1 et + 1. Son signe est le même que celui du coefficient $b_{ij(p)}$ correspondant.

10.3 *Coefficient de corrélation multiple*

Il correspond au coefficient de corrélation totale entre la variable dépendante X_1 et son estimation $X'_{1.23...k}$ donnée par la régression linéaire multiple. D'où:

$$R_{1.23...k} = \frac{\sum (X_1 \cdot X'_{1.23...k})_i}{\sqrt{\sum (X_1^2)_i \cdot \sum (X'^2_{1.23...k})_i}}$$

Le coefficient de corrélation multiple peut se ramener à une expression plus simple:

$$R_{1.23...k} = \sqrt{1 - \frac{S^2_{1.23...k}}{S^2_1}}$$

Le coefficient de corrélation multiple mesure la liaison de la variable dépendante avec les k variables explicatives. S'il est nul, la liaison est inexistante. S'il est égal à 1, la relation linéaire n'est plus aléatoire mais fonctionnelle.

10.4 *Coefficient de détermination multiple*

Ce coefficient n'est autre que le carré du coefficient de corrélation multiple:

$$D_{1.23...k} = R^2_{1.23...k} = 1 - \frac{S^2_{1.23...k}}{S^2_1}$$

Il indique la proportion de la variance marginale de la variable dépendante X_1 qui est expliquée par la régression multiple. La part de cette variance qui reste inexpliquée est donnée par $(1 - D_{1.23...k})$.

11. Tests de signification

Du fait que les n observations des k variables aléatoires ne constituent qu'un échantillon limité d'une population beaucoup plus vaste, le problème se pose de savoir quelle est la signification des paramètres de liaison et des coefficients de régression que l'on détermine à partir de l'échantillon.

En d'autres termes, les paramètres de liaison de la population totale ont-ils une probabilité raisonnable d'être réellement différents de zéro? C'est pour répondre objectivement à cette question que l'on fait appel aux tests de signification.

L'application de ces tests suppose que les hypothèses fondamentales de l'analyse de corrélation multiple sont satisfaites. En effet, s'il n'en est pas ainsi, les coefficients de corrélation ne mesurent pas correctement les degrés de liaison des variables et sont affectés plus ou moins gravement par la mauvaise adéquation de l'ajustement ou la non-normalité de la distribution marginale des variables.

Avant d'examiner les tests de signification il est nécessaire de procéder à une analyse de la variance pour donner une estimation sans biais d'échantillonnage des paramètres de liaison définis au paragraphe 10.

11.1 *Analyse de la variance*

La détermination des coefficients de régression par la méthode des moindres carrés implique la relation:

$$\Sigma X_1^2 = \Sigma X'^2_{1.23\ldots k} + \Sigma X^2_{1.23\ldots k}$$

ou la relation équivalente:

$$S_1^2 = S'^2_{1.23\ldots k} + S^2_{1.23\ldots k}$$

en désignant par $S'^2_{1.23\ldots k}$ la variance due à la liaison linéaire. On voit que la somme de cette variance et de la variance résiduelle est égale à la variance marginale (voir fig. 3 pour $k = 2$).

Mais ces variances sont entachées d'erreur d'échantillonnage et pour en obtenir une estimation plus correcte, il faut faire intervenir dans leurs expressions non pas le nombre n observations, mais le nombre de degrés de liberté qu'elles comportent.

L'analyse de la variance peut alors se résumer sous la forme des tableaux suivants:

a) *Analyse de la variance de X_1 en X_2, X_3, ..., X_k*

Source de variation	Somme des carrés	Degrés de liberté	Variance sans biais
Liaison linéaire	$\Sigma X'^2_{1.23\ldots k} = R^2_{1.23\ldots k} \Sigma X_1^2$	$k - 1$	$S'^2_{1.23\ldots k} = \frac{R^2_{1.23\ldots k}}{k-1} \cdot \Sigma X_1^2$
Résiduelle	$\Sigma X^2_{1.23\ldots k} = (1 - R^2_{1.23\ldots k}) \Sigma X_1^2$	$n - k$	$S^2_{1.23\ldots k} = \frac{1 - R^2_{1.23\ldots k}}{n-k} \cdot \Sigma X_1^2$
Marginale	ΣX_1^2	$n - 1$	$S_1^2 = \frac{1}{n-1} \Sigma X_1^2$

b) *Analyse de la variance de X_1 en X_2 liés par X_3, ..., X_k*

Source de variation	Somme des carrés	Degrés de liberté	Variance sans biais
Liaison linéaire	$\Sigma (X'_{1.23\ldots k} - X'_{1.3\ldots k})^2 = r^2_{12.3\ldots k} \Sigma (X_1 - X'_{1.3\ldots k})^2$	1	$r^2_{12.3\ldots k} \Sigma (X_1 - X'_{1.3\ldots k})^2$
Résiduelle	$\Sigma (X_1 - X'_{1.3\ldots k})^2 = (1 - r^2_{12.3\ldots k}) \Sigma (X_1 - X'_{1.3\ldots k})^2$	$n - k$	$\frac{1 - r^2_{12.3\ldots k}}{n - k} \Sigma (X_1 - X'_{1.3\ldots k})^2$
Totale	$\Sigma (X_1 - X'_{1.3\ldots k})^2$	$n - k + 1$	$\frac{1}{n - k + 1} \Sigma (X_1 - X'_{1.3\ldots k})^2$

11.2 *Estimation sans biais des paramètres de liaison*

Compte-tenu de l'analyse de la variance qui précède on peut montrer que les paramètres de liaison sont estimés de façon correcte (sans biais d'échantillonnage) en adoptant les expressions qui suivent:

— écart-type résiduel:

$$\hat{S}_{i.p} = \sqrt{\frac{n}{n - k}} \cdot S_{i.p} = \sqrt{\frac{n}{n - k} \cdot \frac{\Delta}{\Delta_{ii}} S^2_i}$$

— coefficient de corrélation partielle:

$$\hat{r}_{ij.p} = \sqrt{1 - \frac{n - k + 1}{n - 1} (1 - r^2_{ij.p})}$$

— coefficient de corrélation multiple:

$$\hat{R}_{1.p} = \sqrt{\frac{(n - 1) R^2_{1.p} + 1 - k}{n - k}}$$

11.3 *Signification du coefficient de corrélation multiple*

11.3.1. Test F global

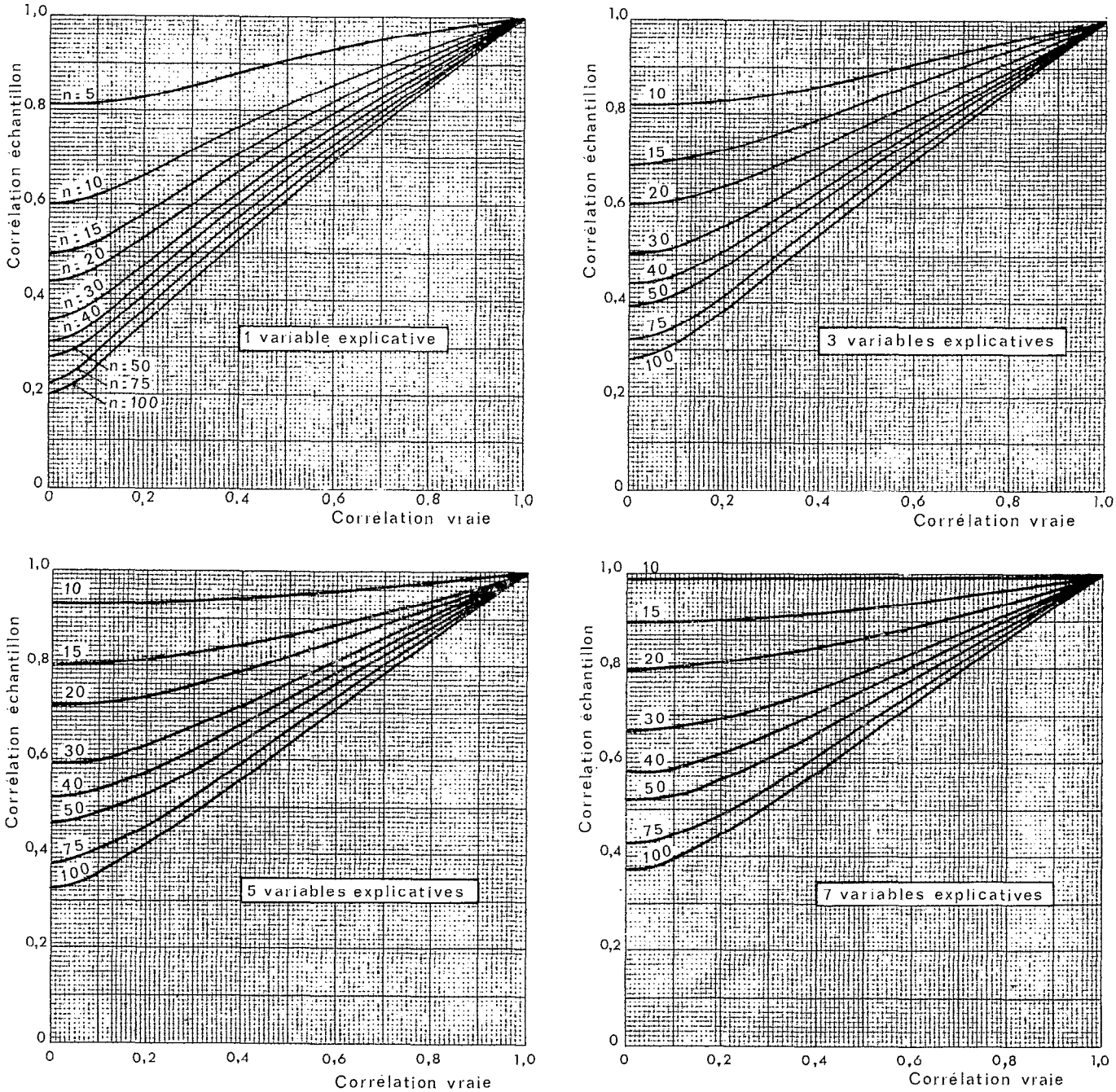
On utilise le test de Fisher-Snedecor pour déterminer si le coefficient de corrélation multiple de la population entière est significativement différent de zéro.

$R^2_{1.23\ldots k}$ est considéré comme significatif si la variance due à la régression linéaire $S'^2_{1.23\ldots k}$ est significativement supérieure à la variance résiduelle $S^2_{1.23\ldots k}$, c'est-à-dire si:

$$F = \frac{n - k}{k - 1} \cdot \frac{R^2_{1.23\ldots k}}{1 - R^2_{1.23\ldots k}}$$

est significativement supérieur à l'unité.

Pratiquement, on calcule la valeur de F et l'on consulte une table de Snedecor pour voir si cette valeur est supérieure à celle donnée par la table pour $v_1 = k - 1$,



Abscisses : Coefficient de corrélation "vraie" déterminé sur la population entière.

Ordonnées : Coefficient de corrélation déterminé sur échantillon et ayant 5 chances sur 100 d'être dépassé. (l'échantillon est tiré au hasard et comporte n observations)

(d'après M. EZEKIEL et K.A. FOX)

Fig. 4
Signification du cœfficient de corrélation multiple

$\nu_2 = n - k$ (degrés de liberté) et pour une probabilité fixée (5% ou 1% en général). C'est seulement s'il en est bien ainsi que l'on conclut que la corrélation multiple est significative.

Les graphiques de la figure 4 établis d'après M. EZEKIEL et K. A. FOX sur la base des travaux de FISHER, donnent une bonne idée des erreurs d'échantillonnage que comporte la détermination des coefficients de corrélation multiple.

On notera, par exemple, que si le coefficient de corrélation multiple de la population est nul, il y a 5 chances sur 100 pour le coefficient de corrélation multiple déterminé avec un échantillon de 20 observations, atteigne une valeur de:

0,43 s'il y a une variable explicative
0,60 s'il y a trois variables explicatives
0,71 s'il y a cinq variables explicatives
0,79 s'il y a sept variables explicatives

Cet exemple montre bien que le coefficient de corrélation multiple déduit d'un échantillon limité peut avoir une valeur assez élevée, mais parfaitement illusoire en ce qui concerne la population entière, et ceci d'autant plus que le nombre des variables explicatives est grand. L'utilisation des tests de signification, loin d'être un raffinement superflu, apparaît donc comme absolument indispensable.

11.3.2. TEST F PARTIEL

Le test F partiel sert à juger de la signification, non plus de l'ensemble des variables explicatives, mais d'une seule ou de quelques-unes des variables explicatives.

Soit $S'^2_{1.p}$ la variance due à la régression linéaire dans le cas de p variables explicatives et $S'^2_{1.q}$ la même variance dans le cas de q variables explicatives, q étant plus grand que p.

L'apport des variables X_{p+1}, X_{p+2}, ... X_q est considéré comme significatif s'il conduit à une augmentation de la variance liée à la régression significativement supérieure à la variance résiduelle, c'est-à-dire si:

$$F = \frac{S'^2_{1.q} - \dfrac{S'^2_{1.p}}{q - p}}{S^2_{1.p}}$$

est significativement supérieure à l'unité.

La valeur de F est, comme dans le cas du test global, comparée à la valeur donnée par la table de SNEDECOR pour $\nu_1 = q - p$, $\nu_2 = n - q - 1$ et une probabilité de 0,05 ou 0,01.

11.4 *Signification du coefficient de corrélation partielle*

Le test de signification consiste à voir si la variance due à la liaison est supérieure à la variance résiduelle, c'est-à-dire si:

$$F = \frac{r^2_{ij.p}}{1 - r^2_{ij.p}} \cdot (n - k)$$

est significativement supérieur à l'unité. Comme dans le cas précédent on utilisera

une table de SNEDECOR avec $\nu_1 = 1$ et $\nu_2 = n - k$, pour déterminer si le coefficient de corrélation partielle est significativement différent de zéro.

On peut également utiliser le test de STUDENT-FISHER, en considérant la variable statistique t définie par:

$$t = \sqrt{F}$$

On utilise alors une table de STUDENT en prenant un degré de liberté égal à $(n - k)$.

11.5 *Signification du coefficient de régression multiple*

On peut juger de la signification du coefficient de régression en considérant la variable t de STUDENT définie par:

$$t = \frac{b_{ij.p}}{S_{bij.p}} \cdot \sqrt{n - k}$$

$S_{bij.p}$ étant l'écart-type de $b_{ij.p}$.

On compare la valeur de t à celle donnée par la table de STUDENT pour un degré de liberté égal à $(n - k)$ et une probabilité fixée. Si la première valeur est supérieure à la seconde, le coefficient de régression est significativement différent de zéro.

On peut tout aussi bien utiliser une table de SNEDECOR avec $F = t^2$, $\nu_1 = 1$ et $\nu_2 = n - k$.

Il est par ailleurs facile de voir que le test sur $b_{ij.p}$ est équivalent à celui sur $r_{ij.p}$.

12. Sélection des variables explicatives

(Méthode « Stepwise»)

La sélection des variables explicatives ne peut guère s'opérer de façon systématique qu'avec l'aide d'un ordinateur, sinon l'on se heurte à des calculs extrêmement laborieux dès que le nombre des variables à sélectionner dépasse deux ou trois.

En toute rigueur la sélection devrait s'effectuer en testant toutes les combinaisons possibles de toutes les variables jugées à priori susceptibles d'être explicatives.

En pratique, pour économiser les temps de calcul à l'ordinateur, on se contente d'une méthode moins rigoureuse telle que la méthode « Stepwise », ainsi appelée parce que les variables explicatives sont introduites progressivement dans la régression multiple.

La méthode Stepwise comporte en premier lieu, les calculs suivants:

— moyennes et écarts-types (sans biais) des valeurs observées de la variable dépendante et de toutes les variables soumises à la sélection;

— coefficients de corrélation totale de toutes les variables prises deux à deux.

Ensuite la méthode procède par paliers successifs:

Premier palier:

Parmi les variables soumises à la sélection on considère celle qui a le coefficient de corrélation totale le plus élevé avec la variable dépendante. On lui applique le test F en se basant sur un seuil de signification fixé à l'avance et désigné ici par FIN.

Si le résultat du test est négatif la sélection est terminée. S'il est positif, la variable considérée est introduite dans une régression linéaire provisoire de la forme:

$$Y = b_0 + b_1 X_1$$

où y désigne la variable dépendante et X_1 la première variable explicative sélectionnée. On calcule l'écart-type résiduel, le terme constant b_0 ainsi que le coefficient b_1 avec son écart-type, sa variable t de STUDENT et le coefficient β_1 correspondant. (L'écart-type et la variable t permettent de juger de la signification de b_1 en consultant une table de STUDENT et de calculer éventuellement son intervalle de confiance.)

Deuxième palier:

Parmi les variables non introduites dans la régression on considère celle qui permet d'expliquer la plus grande part de la variance résiduelle. Elle est retenue ou rejetée comme deuxième variable explicative suivant qu'elle satisfait ou non au test F partiel basé sur le seuil de signification FIN. Si elle est rejetée, la sélection s'arrête et la régression provisoire du premier palier est adoptée de façon définitive. Si elle est retenue, la nouvelle équation de régression adoptée provisoirement est de la forme:

$$Y = b'_0 + b'_1 X_1 + b'_2 X_2$$

Sont alors calculées les nouvelles valeurs de l'écart-type résiduel, du coefficient de corrélation multiple, de la variable F pour le test global d'ajustement, du terme constant b'_0 et enfin des coefficients de régression avec leurs écarts-types, t de STUDENT et coefficients β correspondants.

Troisième palier:

La première variable introduite dans la régression est à nouveau testée pour voir si elle reste significative quand la seconde est ajoutée dans la régression. Le test F partiel lui est appliqué avec un seuil de signification fixé à l'avance qui est égal ou inférieur à FIN et que nous désignons par FOUT. Si le test n'est pas passé avec succès la nouvelle équation de régression adoptée provisoirement est de la forme:

$$Y = b''_0 + b''_2 X_2$$

Toutes les caractéristiques de la régression sont à nouveau calculées comme aux paliers précédents et l'on passe au palier suivant.

Si le test F partiel donne un résultat favorable, la sélection d'une troisième variable explicative s'effectue d'une façon analogue à celle de la seconde variable au 2e palier, etc.

D'une façon générale, chaque fois qu'une nouvelle variable a subi favorablement le test F partiel (avec seuil FIN) et a été introduite dans la régression, toutes les autres variables explicatives introduites précédemment sont à nouveau testées (avec seuil FOUT) comme si elles avaient été ajoutées en dernière position. Suivant le résultat de ce test, elles sont conservées ou éliminées de la régression.

Remarques:

Par souci de simplification les tests F partiels sont effectués avec des seuils de significations FIN et FOUT fixés une fois pour toutes à l'avance. En toute rigueur, ils devraient varier d'un palier à l'autre. En effet, quand le nombre de variables explicatives introduites dans la régression passe de (p — 1) à p, le test F partiel doit faire

intervenir des degrés de liberté $\nu_1 = 1$ et $\nu_2 = n - p - 1$. En fait, la table de SNEDECOR montre que la valeur de F varie assez lentement avec ν_2, si celui-ci dépasse 10, comme on peut le voir d'après les valeurs repères ci-dessous:

	$\nu_2 = 1$ et P = 0,05						
ν_2	5	10	15	20	30	40	60
F	6,61	4,96	4,54	4,35	4,17	4,08	4,00

Dans la plupart des cas les valeurs de FIN et FOUT peuvent être fixées entre 4 et 5. La valeur de FOUT peut être inférieure ou égale à celle de FIN, mais non supérieure pour éviter qu'une variable déjà éliminée soit ré-introduite au palier suivant.

En ce qui concerne le test F global, il n'est pas effectué de façon automatique mais doit être opéré manuellement d'après les éléments fournis par l'ordinateur à chaque palier. Il en est de même du test t des coefficients de régression.

Il est à remarquer que lorsque les variables explicatives ne sont pas extérieurement indépendantes, c'est-à-dire présentent entre elles des corrélations totales non négligeables, l'utilisation des tests de signification perd de sa rigueur. Ceci explique, au moins partiellement, certaines anomalies. M. MASSON a fait remarquer que dans de nombreuses applications avec n de l'ordre de 30, lorsque 5 ou 6 variables explicatives ont déjà été introduites dans la régression et que le coefficient de corrélation multiple atteint une valeur voisine de 0,95, il arrive que la valeur de F calculée pour le test partiel devienne instable et prenne de fortes valeurs. Certaines variables tenues jusque-là à l'écart de la régression ont tout à coup un effet significatif sur la variable dépendante et donnent au coefficient de corrélation multiple une valeur très voisine de l'unité.

On en retiendra qu'il faut éviter de soumettre à la sélection des variables trop étroitement corrélées entre elles et que la régression devient suspecte lorsque les variables sélectionnées sont en trop grand nombre.

On devra également beaucoup se méfier de l'extrapolation de l'équation de régression en-dehors du champ d'observations sur lequel elle a été établie.

13. Programme Fortran de la méthode Stepwise

Pour l'application pratique de la méthode Stepwise aux problèmes hydrologiques nous proposons une version quelque peu modifiée d'un programme FORTRAN utilisé par la Direction générale des Eaux du Québec.

Nous n'exposerons pas ce programme dans le détail. Le lecteur suffisamment familiarisé avec le calcul matriciel pourra trouver une description détaillée de la méthode Stepwise dans « Mathematical Method for digital Computers » de A. RALSTON et H. S. WILF. Nous nous bornerons à donner les indications strictement nécessaires à l'utilisateur.

13.1 *Programme principal*

Le programme principal, dont on trouvera le listing en annexe, est conçu pour traiter successivement plusieurs études complètes de régression du même type. Exemple: étude des douze régressions liant les débits mensuels de chacun des mois de l'année aux indices pluviométriques du mois correspondant et des mois antérieurs. (L'application correcte de la méthode suppose que les indices pluviométriques de mois successifs ne sont pas étroitement corrélés entre eux.)

13.2 *Subroutines*

Le programme principal est de portée très générale, mais il fait appel à plusieurs subroutines qui, elles, doivent être adaptées à chaque problème particulier. Ces subroutines sont au nombre maximal de quatre:

a) *Subroutine LECTUR:*

Elle sert à lire les données d'observation brutes et les met dans une matrice Y (J, K). Dans l'exemple déjà cité, la matrice Y contient les débits mensuels (K = 1, 12), le module (K = 13), les indices pluviométriques mensuels (K = 14, 25) et l'indice pluviométrique annuel (K = 26) pour un certain nombre d'années d'observation (J = 1, NJA).

b) *Subroutine ANAMOR:*

Elle sert éventuellement à effectuer des anamorphoses diverses sur les valeurs observées de certaines variables (transformations logarithmiques pour J = 1, NJA et K = 1,13 par exemple). La matrice Y conserve son nom après ces transformations.

c) *Subroutine MATRIX:*

Contrairement aux deux subroutines précédentes qui ne sont appelées qu'une fois, celle-ci est utilisée pour chaque régression de la série à étudier. Elle sert à constituer à partir de la matrice Y une nouvelle matrice X (J, K) contenant les données d'observation transformées qui sont utiles pour l'étude de régression en cours. Les variables explicatives à sélectionner correspondent aux indices K compris entre 1 et NINDV. La variable dépendante correspond à l'indice K = NVAR = NINDV + 1. Exemple: si l'on étudie la régression liant le débit mensuel de janvier aux indices pluviométriques de janvier et des onze mois précédents, on a K = 1, 12 pour les indices de janvier, décembre, novembre, ..., février et K = 13 pour le débit mensuel de janvier. L'indice J correspond aux années d'observation et varie de 1 à NØBS = NJA — 1.

d) *Subroutine SORTIE:*

Elle est utilisée comme la précédente pour chaque régression de la série à étudier. Elle sert à imprimer ou à perforer les résultats de l'étude de régression en cours sous une forme adaptée à toute la série. Elle imprime et/ou perfore, par exemple, les coefficients de régression et les numéros des mois dont l'indice pluviométrique a été sélectionné pour la régression du débit de janvier. Elle peut éventuellement effectuer une anamorphose inverse des valeurs estimées de la variable dépendante.

Remarques:

Si l'on a une seule étude de régression à effectuer (NBREG = 1), la subroutine

MATRIX devient superflue. La matrice X peut être directement établie par la subroutine LECTUR.

La subroutine ANAMOR devient également sans objet s'il n'y a pas à effectuer de transformations de variables ou si celles-ci sont effectuées préalablement par un autre programme.

La subroutine SORTIE n'est pas non plus toujours indispensable, notamment dans le cas où il n'y a pas de transformations de variables.

Entre le programme principal et les subroutines, les données sont transmises au moyen de six COMMON nommés:

— COM 1: entre Principal et MATRIX;

— COM 2: entre Principal et LECTUR;

— COM 3: entre Principal et LECTUR, ANAMOR et MATRIX;

— COM 4: entre Principal et SORTIE;

— COM 5: entre Principal et MATRIX et SORTIE;

— COM 6: entre Principal et LECTUR, ANAMOR et SORTIE.

On trouvera en annexe, les listings de subroutines LECTUR, MATRIX et SORTIE convenant pour l'exemple cité dans le cas où il n'y a pas lieu d'effectuer d'anamorphose. La subroutine LECTUR fait elle-même appel à deux sous-programmes (subroutine BMØDUL pour le calcul des modules en fonction des débits mensuels et fonction ARONDI pour arrondir les débits à trois chiffres significatifs avec précision maximale de 1 l/s).

13.3 *Données d'entrée*

a) 1 carte FORMAT (2 I 2, 2 F 6.3, I 1) fixant les valeurs de:

— *NBREG:* Nombre d'études successives de régression multiple. Exemple: NBREG = 12, dans le cas où l'on cherche à relier les débits mensuels de chacun des douze mois de l'année aux indices pluviométriques concomitants et antérieurs;

— *NVAR:* Nombre total de variables dépendantes et explicatives à sélectionner introduites dans la matrice X. Dans l'exemple cité, NVAR = 13;

— *FIN et FOUT:* Niveaux de signification F respectivement pour introduire et retirer une variable de la régression;

— *IRES:* Indicateur pour l'impression d'un tableau donnant les valeurs observées, les valeurs calculées et leurs différences, puis l'écart quadratique moyen — ainsi que pour l'appel de la subroutine SORTIE.

Si IRES $\neq$ 0 impression et appel de SORTIE;

Si IRES = 0 pas d'impression, ni appel de SORTIE.

b) un jeu de cartes dont la lecture est commandée par la subroutine LECTUR;

c) éventuellement un jeu de cartes dont la lecture est commandée par la subroutine ANAMOR;

d) éventuellement un jeu de cartes dont la lecture est commandée par la subroutine MATRIX.

13.4 *Sorties*

13.4.1. Sur imprimante

1) Sorties commandées éventuellement par les subroutines LECTUR et ANAMOR.

2) Pour chacune des NBREG études de régression:

a) sorties éventuelles commandées par la subroutine MATRIX;

b) moyenne et écart-type des valeurs observées pour la variable dépendante et toutes les variables soumises à la sélection;

c) coefficients de corrélation totale des mêmes variables que ci-dessus prises deux à deux;

d) à chaque palier:

- — variable introduite ou variable éliminée,
- — écart-type résiduel,
- — coefficients de corrélation multiple,
- — test global d'ajustement: niveau F,
- — coefficient de régression b_0,
- — coefficients de régression bi avec écarts-types, variables t student et coefficients β;

e) éventuellement (si IRES $\neq$ 0) tableau des valeurs observées, calculées et de leurs différences. Ecart quadratique moyen;

f) éventuellement (si IRES $\neq$ 0), sorties commandées par la subroutine SORTIE.

On trouvera en annexe un exemple de sorties sur imprimante.

13.4.2. Sur cartes perforees

Sorties commandées éventuellement par la subroutine SORTIE.

13.5 *Précisions complémentaires*

Il reste à préciser la signification de certaines variables qui ont été introduites dans les « COMMON » des listings donnés en annexe et qui n'ont pas encore été définies:

NEREUR	Nombre d'erreurs détectées dans les données d'entrée par LECTUR.
BSUBO	Terme constant b_0 de l'équation de régression définitive.
B(I)	Coefficients b_1, b_2 etc. de l'équation de régression définitive.
NIN	Nombre de variables explicatives sélectionnées.
ID (I)	Valeur de l'indice K de la matrice X correspondant à la variable sélectionnée de rang I.
NØREG	Numéro de la régression étudiée (varie de 1 à NBREG).
L(K)	Variable spécifique de l'exemple cité, non utilisée par le programme principal (sert à définir les correspondances entre indices des matrices X et Y).
MØDEB	Variable spécifique de l'exemple cité, non utilisée par le programme principal (désigne le numéro du premier mois de l'année hydrologique).

14. Conclusion

Le présent article n'a pas eu d'autre objectif que de rassembler à l'intention des hydrologues — qui ne sont pas nécessairement des statisticiens avertis — les éléments essentiels qui leur permettent d'utiliser avec profit l'analyse des régressions linéaires et corrélations multiples.

Notre exposé ne prétend pas épuiser le sujet ni y apporter une contribution réellement originale. Il s'est inspiré des ouvrages cités en bibliographie et plus particulièrement de la thèse récente de M. MILU ROSENBERG dont les deux premiers chapitres apportent beaucoup de clarté en la matière.

Bibliographie

MORLAT (G.) – 1952 – Les méthodes statistiques. Electricité de France, Direction des Etudes et Recherches. Paris.

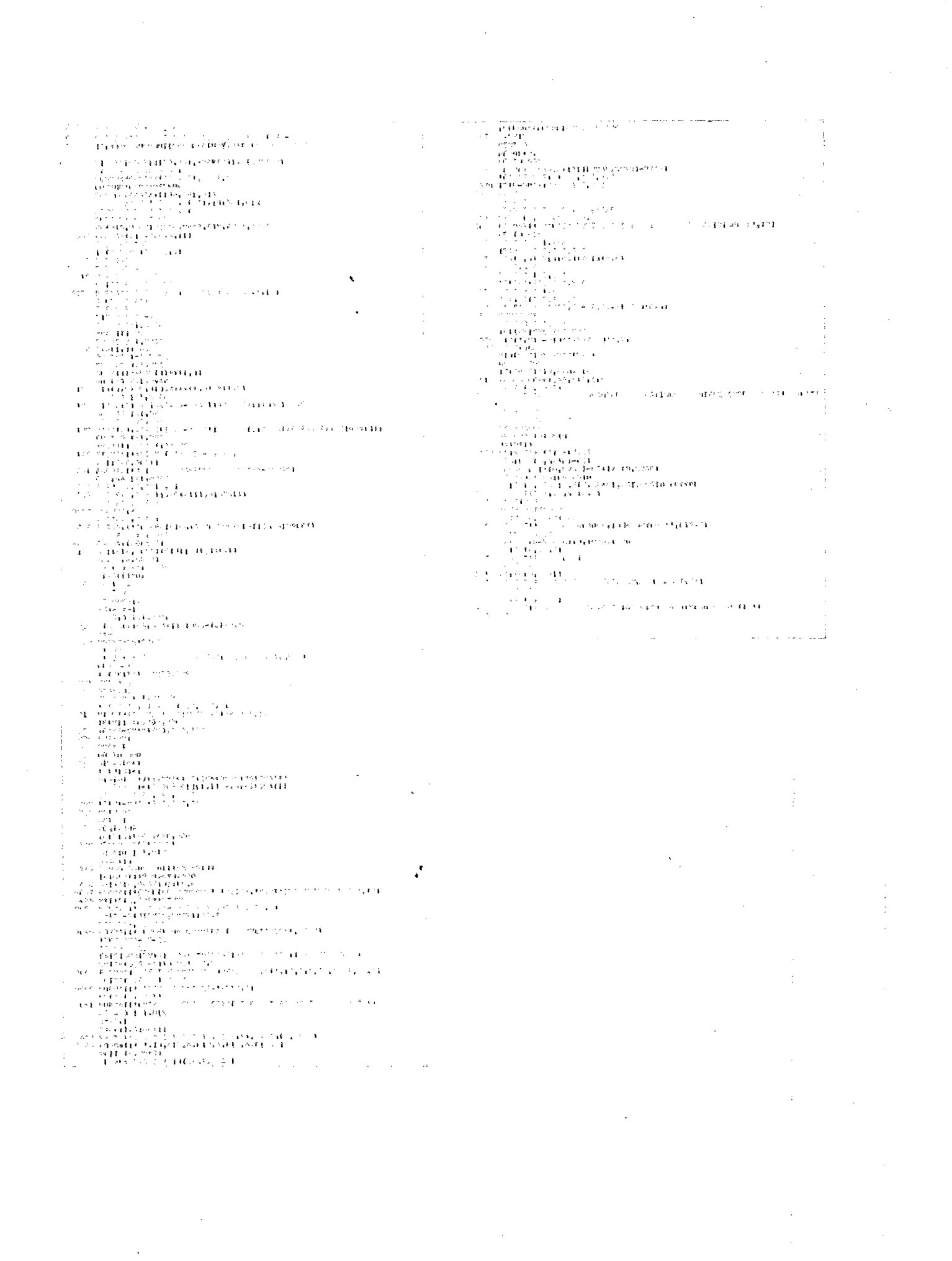
MORICE (E.), CHARTIER (F.) – 1954 – Méthode statistique. INSEE, Paris.

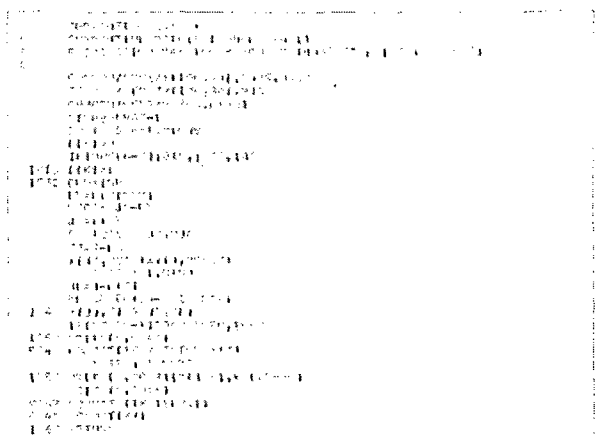
EZECKIEL (M.), FOX (K. A.) – 1959 – Methods of correlation and regression analysis. 3e édition, John Wiley and sons, New York.

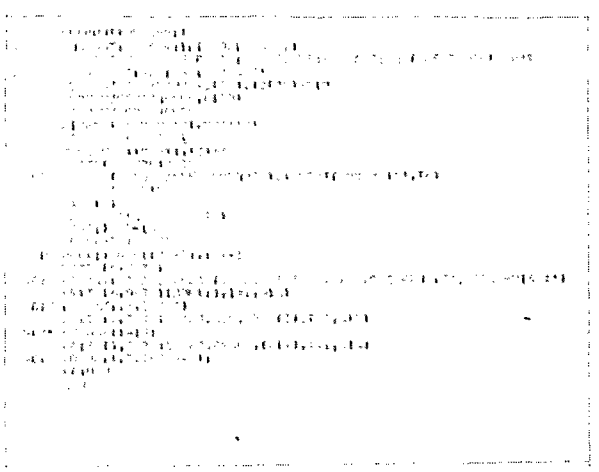
RALSTON (A.), WILF (H. S.) – 1962 – Mathematical methods for digital computers. John Wiley and Sons, New York.

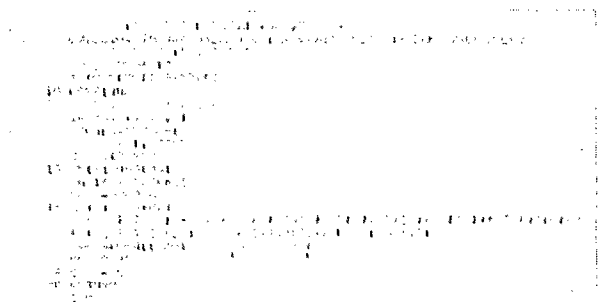
ROSENBERG (M.) – 1971 – Thèse de Doctorat ès Sciences Physiques. Université scientifique et médicale de Grenoble.

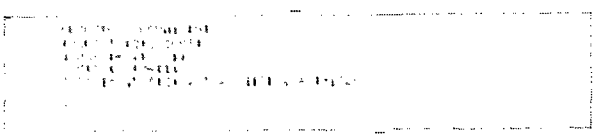
MASSON (J. M.) – 1971 – Thèse de Docteur-Ingénieur. Université des Sciences et Techniques du Languedoc, Montpellier.

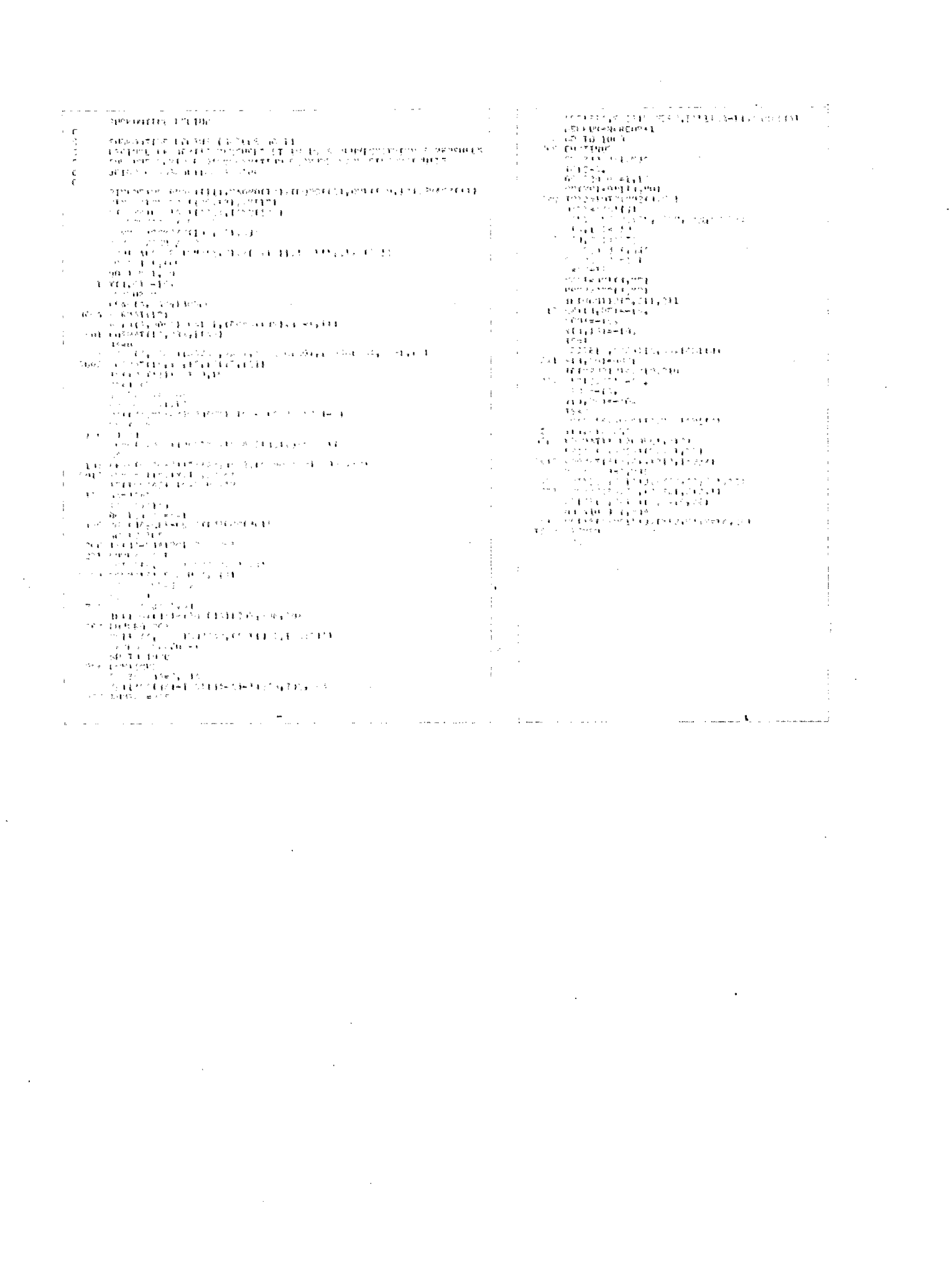












[illegible]

1	[illegible]	[illegible]
1	[illegible]	[illegible]
2	[illegible]	[illegible]
3	[illegible]	[illegible]
4	[illegible]	[illegible]
5	[illegible]	[illegible]
6	[illegible]	[illegible]
7	[illegible]	[illegible]
8	[illegible]	[illegible]
9	[illegible]	[illegible]
10	[illegible]	[illegible]
11	[illegible]	[illegible]
12	[illegible]	[illegible]
13	[illegible]	[illegible]

COEFFICIENTS [illegible]

[illegible]

[illegible]

COEFF OF [illegible]

TEST [illegible]

[illegible]

VAR [illegible]

[illegible]

[illegible]

COEFF OF [illegible]

TEST [illegible]

[illegible]

VAR [illegible]

1 [illegible]

2 [illegible]

OBS	[illegible]	[illegible]	[illegible]	[illegible]
1	[illegible]	[illegible]	[illegible]	[illegible]
2	[illegible]	[illegible]	[illegible]	[illegible]
3	[illegible]	[illegible]	[illegible]	[illegible]
4	[illegible]	[illegible]	[illegible]	[illegible]
5	[illegible]	[illegible]	[illegible]	[illegible]
6	[illegible]	[illegible]	[illegible]	[illegible]
7	[illegible]	[illegible]	[illegible]	[illegible]
8	[illegible]	[illegible]	[illegible]	[illegible]
9	[illegible]	[illegible]	[illegible]	[illegible]
10	[illegible]	[illegible]	[illegible]	[illegible]
11	[illegible]	[illegible]	[illegible]	[illegible]
12	[illegible]	[illegible]	[illegible]	[illegible]
13	[illegible]	[illegible]	[illegible]	[illegible]
14	[illegible]	[illegible]	[illegible]	[illegible]
15	[illegible]	[illegible]	[illegible]	[illegible]
16	[illegible]	[illegible]	[illegible]	[illegible]
17	[illegible]	[illegible]	[illegible]	[illegible]
18	[illegible]	[illegible]	[illegible]	[illegible]

[illegible]

VARIABLE [illegible]

VARIABLE [illegible]

[illegible]