JOURNAL DE LA SOCIÉTÉ STATISTIQUE DE PARIS

ROGER CONGARD

Régression unilatérale et régression mutuelle

Journal de la société statistique de Paris, tome 92 (1951), p. 284-302

<http://www.numdam.org/item?id=JSFS_1951__92__284_0>

© Société de statistique de Paris, 1951, tous droits réservés.

L'accès aux archives de la revue « Journal de la société statistique de Paris » (http://publications-sfds.math.cnrs.fr/index.php/J-SFdS) implique l'accord avec les conditions générales d'utilisation (http://www.numdam.org/conditions). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
http://www.numdam.org/

III

VARIÉTÉS

Régression unilatérale et régression mutuelle

Le Français est volontiers incrédule car il manque d'enthousiasme. Il juge sa condition sociale très mauvaise par rapport à celle de ses amis d'Outre-Atlantique mais dès qu'on lui propose un moyen d'améliorer sa condition il se récrie et affiche un pessimisme de principe. Exerçant sans cesse son esprit critique il dénigre toute innovation (en matière sociale) avant même d'avoir eu la possibilité d'en juger les résultats, la condamnant ainsi par avance à un échec certain. Mais s'il est une discipline à l'égard de laquelle le Français moyen fait preuve d'un pessimisme malveillant c'est bien la statistique. N'est-il pas courant d'entendre dire qu'elle est la forme la plus élevée du mensonge?

Chaque Français s'estime en droit de porter un jugement sur chaque chose,

(1) Dans ces comptes, les importateurs peuvent inscrire les devises étrangères représentant 10 à 15 % du montant de leurs importations. Ces devises E. F. A. C. sont cessibles et peuvent donner lieu à des achats de marchandises sans licence d'importation.

même sur celles qu'il ne connaît pas et il trouve souvent des arguments pour étayer *a posteriori* une opinion adoptée à la légère. La chose est facile surtout quand il s'agit de la statistique. N'existe-t-il pas en effet des statistiques fantaisistes établies tout spécialement pour appuyer une thèse *a priori* et même quand il s'agit de statistiques élaborées avec un grand souci d'objectivité, les résultats ne diffèrent-ils pas sensiblement selon qu'ils sont obtenus par tel service ou tel autre?

Dans ces conditions il apparaît plus commode de rejeter en bloc la statistique en lui déniant toute utilité. Mais l'adoption d'une telle position constitue une preuve de paresse intellectuelle imputable à celui qui refuse de faire l'effort de distinguer la vraie statistique de la recherche d'arguments chiffrés et qui, pour défendre son point de vue attribue à la statistique un caractère de précision auquel elle n'a jamais prétendu.

La statistique n'a d'autre but que de permettre de dégager des tendances, des variations, des mouvements de phénomènes économiques ou autres et des relations entre ces phénomènes. Elle nous indique des « ordres de grandeur » et pas davantage mais elle n'en est pas moins nécessaire car la connaissance imparfaite qu'elle permet d'acquérir nous donne la possibilité d'orienter sciemment notre action vers la poursuite du but visé au lieu de nous laisser mener à l'aveuglette vers les conséquences inconnues de nos actes.

Dans le domaine économique les phénomènes interfèrent constamment les uns sur les autres et la théorie pure ne peut tenir compte de toutes ces interférences. Ce n'est donc que par l'empirisme raisonné de la statistique, apprécié par référence à un modèle théorique qui nous livre des normes de jugement, que la science économique peut progresser.

Cet empirisme de la statistique se manifeste tout particulièrement quand il s'agit de mettre sous une forme synthétique les rapports de co-variation des phénomènes économiques observés. Tel est pourtant le but principal de la statistique économique puisque seul il peut conduire à la connaissance des éléments invariables et des liaisons constantes existant dans le monde économique. La recherche de ces rapports de co-variation qui, le plus souvent, recouvrent des liens de causalité, peut être conduite de différentes manières. La plus connue et la plus usitée est la méthode dite de Régression (1). Cette méthode consiste à substituer à une liaison stochastique une liaison fonctionnelle s'en rapprochant le plus possible. Autrement dit, en raisonnant en termes de géométrie, le but de la méthode sera de substituer à une série de points représentant graphiquement les observations, une ligne régulière et continue passant le plus près possible de tous ces points, qui sera la ligne de régression du phénomène y par rapport au phénomène x.

La ligne brisée que l'on peut constituer par la jonction des différents points d'observation (x_i, y_i) ne nous donne qu'une représentation discontinue de la liaison existant entre y et x. Les irrégularités de cette ligne brisée peuvent

(1) La dénomination « régression » vient des travaux de Galton sur l'hérédité. Galton mesurait un caractère, par exemple la taille, chez le père (résultat x_i) et le fils (résultat y_i). La représentation graphique dans le plan x, y de ces deux séries de résultats montrait que généralement, lorsque la taille du père était supérieure à la moyenne, celle du fils l'était aussi mais plus faiblement; il y avait régression.

être considérées comme le résultat d'erreurs d'observation ou comme étant dues à l'intervention de facteurs accidentels, secondaires par rapport aux phénomènes étudiés. Elles sont en opposition avec le sentiment intuitif que l'on a d'un phénomène ou d'un rapport de phénomènes variant de façon continue.

La notion de continuité du phénomène conduit à éliminer des observations tout ce que l'on peut raisonnablement considérer comme étranger au phénomène. Pour cela on substitue aux nombres observés de nouveaux nombres. C'est ce qu'on appelle procéder à un ajustement. Suivant la façon dont ces nouveaux nombres sont obtenus on peut distinguer :

— La méthode graphique qui consiste à tracer à la main une courbe continue, aussi régulière que possible, située le plus près possible des points d'observation. Cette méthode est moins arbitraire qu'elle ne le paraît, mais elle offre l'inconvénient de faire dépendre la courbe ajustée des échelles respectives qui ont été adoptées pour représenter graphiquement les deux phénomènes liés stochastiquement.

— La méthode « mécanique », qui consiste à remplacer les nombres observés par d'autres qui en sont déduits par application mécanique d'une formule.

Nous nous bornerons ici à étudier la méthode « mécanique », la première ne pouvant donner lieu qu'à une description, et nous limiterons nos développements à la seule méthode d'*ajustement analytique*.

Cette dernière méthode consiste à « ajuster » la série des données observées par une fonction du type $y = f(x)$. Le choix de ce type de fonction comporte une grande part d'arbitraire et de plus, une fois ce type de fonction défini, la détermination précise de la fonction ajustée est soumise à l'arbitraire du choix d'un procédé d'ajustement.

I. — *Choix du type de fonction.*

Il est toujours possible de trouver une fonction $y = f(x)$ vérifiée par tous les points d'observations : il suffirait, par exemple, n étant le nombre d'observations, de choisir un polynôme de degré $n - 1$. Un etelle solution algébrique est cependant sans intérêt, puisqu'il s'agit non de remplacer la ligne brisée représentative des observations par une courbe compliquée passant par tous les points, mais par une courbe simple, dépendant de quelques paramètres en nombre aussi petit que possible, épousant au mieux l'ensemble de la ligne des observations et laissant de côté ce qu'on estime être des variations accidentelles.

Dans certains cas, on sera seulement guidé dans la recherche d'une telle courbe par l'allure générale de la ligne brisée ou d'une courbe simple tracée à l'estime au milieu de l'ensemble des points d'observation. Parmi les courbes algébriques correspondant à cette allure générale, on choisira de préférence la plus simple, c'est-à-dire celle qui peut être caractérisée par le plus petit nombre de paramètres (droite, parabole, exponentielle) (1).

II. — *Procédés d'ajustement.*

Le type de courbe étant choisi, il convient de rechercher la valeur numérique des coefficients telle que la courbe s'ajuste aussi bien que possible aux obser-

(1) « Il est plus important en science d'avoir une loi simple qu'une loi vraie » H. Schultz, « The theory and measurement of demand », p. 53.

vations. Mais quand pourra-t-on dire qu'une courbe s'ajuste bien uo mal aux observations? Pour apprécier la qualité de l'ajustement il nous faut un critérium, et le choix de celui-ci ne peut être qu'arbitraire, au moins dans une certaine mesure.

En effet, nous pouvons prendre comme critérium aussi bien la petitesse des déviations absolues des points à la courbe ou la petitesse des déviations relatives, et nous pouvons « travailler » avec la 1re puissance, la 2e puissance ou la $n^{ième}$ puissance de ces déviations.

Par ailleurs, nous pouvons mesurer les déviations par des surfaces ou par des segments de droite et, dans ce dernier cas, nous pouvons mesurer les déviations parallèlement à l'un ou l'autre des axes de coordonnées ou à un axe quelconque. Les différentes méthodes ne conduisent pas nécessairement à la même conclusion. Il semble donc, comme l'a fait remarquer H. Schultz, que « la valeur de l'ajustement repose en dernier ressort sur des considérations esthétiques » (1).

Aussi nous laisserons-nous guider par le souci de simplicité (2), ce qui nous conduira à choisir :

— la ligne droite comme courbe d'ajustement,

— la direction de l'axe des x ou des y pour la mesure des déviations (appréciées par la longueur de segments de droites),

— la méthode de minimisation des carrés des déviations (nous prenons les carrés, car nous avons ainsi des valeurs toujours positives, ce qui facilite les calculs, tout en ayant une précision supérieure numériquement) et enfin nous choisirons de préférence la mesure des déviations absolues à celle des déviations relatives.

Toutefois, ces choix effectués uniquement par référence au critère de simplicité ne seront pas maintenus quand d'autres raisons militeront en faveur d'un procédé différent.

Quoi qu'il en soit, l'imprécision qui résulte de l'adoption du critère de simplicité nous met en garde dès l'abord contre la tendance que nous pourrions avoir à considérer les résultats statistiques obtenus comme devant traduire avec exactitude les liaisons réelles pouvant exister entre les phénomènes économiques étudiés. Mais il n'en reste pas moins que, dans le cadre d'un but plus modeste assigné à la statistique, il devient possible, grâce en particulier à la régression, de dégager les tendances générales de co-variation de ces phénomènes.

Il est d'usage, lorsqu'on se sert de régression, de sous-entendre qu'il s'agit de régression unilatérale et même plus particulièrement de régression linéaire. Mais en fait il ne s'agit là que d'un cas très particulier de régression qu'il convient de replacer dans un cadre plus général, sous peine de lui attribuer un domaine de validité qu'il ne couvre pas.

---

(1) Schultz, *op. cité*, p. 136 et suivantes.

(2) « L'abandon d'une théorie, nous dit M. L. Schwartz, peut résulter de la complication trop grande qu'elle introduit dans l'interprétation des phénomènes. Henri Poincaré en a donné comme exemple la théorie selon laquelle la terre serait immobile, théorie qui n'implique pas de contradiction, mais qui introduit une complication inouïe dans les calculs astronomiques. »

L. Schwartz « La notion de loi dans le domaine scientifique ». *La Revue Internationale*, avril 1946, p. 352.

A côté de la régression linéaire, on trouve autant de types de régression qu'il peut y avoir de types de fonctions d'ajustement. De même, à côté du mode de régression que nous qualifions d'unilatérale et qui correspond à l'idée d'imputation de l'erreur à une seule variable, nous ferons une place aux modes de régression qui sont établis sur l'hypothèse d'une imputation de l'erreur aux deux variables que nous désignerons par modes de régression mutuelle. En fait il existera autant de modes de régression mutuelle qu'il peut exister d'hypothèses différentes concernant l'imputation de l'erreur. On peut très bien concevoir par exemple l'imputation d'un certain pourcentage de l'erreur sur une variable et du reste sur l'autre variable, si des raisons particulières permettent de connaître les degrés de précision respectifs des séries d'observation mises en rapport. Dans le cadre de cette étude, nous nous en tiendrons à l'hypothèse d'une répartition égale de l'erreur entre les deux variables, ce qui apparaît plus logique en l'absence de considérations de fait concernant la précision des données statistiques utilisées.

Nous reprenons ci-dessous les deux termes de cette distinction : régression unilatérale et régression mutuelle.

I — RÉGRESSION UNILATÉRALE

Définition : Si nous cherchons par exemple à déterminer une courbe statique de demande (exprimant uniquement la relation entre le prix et la quantité achetée), la question se pose de savoir laquelle des deux régressions, de celle des prix par rapport aux quantités ou de celle des quantités par rapport aux prix doit être prise comme courbe de demande. Le statisticien qui procède empiriquement est à peine au courant de cette question. Il retient arbitrairement une des variables, généralement le prix, comme variable dépendante et détermine les paramètres de sa courbe de manière que la somme des carrés des déviations des prix observés aux prix calculés soit minima. Il ne réalise cependant pas que sa méthode implique que la non-concordance entre le point d'observation et le point correspondant de la courbe de demande n'est imputable uniquement qu'à « l'erreur » ou déviation de la variable dépendante y (prix), une déviation de la variable indépendante x (consommation) n'étant même pas envisagée par lui.

Une telle méthode aboutit à accorder aux chiffres de consommation un poids infiniment plus grand qu'aux observations des prix.

Nous désignerons du nom de « régression unilatérale » celle qui est obtenue en formulant l'hypothèse au moins implicite selon laquelle une seule variable est « sujette à erreur », par opposition à la « régression mutuelle » qui n'implique « en principe » aucune hypothèse quant à l'imputation de l'erreur.

Nous étudierons les principales « méthodes simples » de régression, c'est-à-dire la régression linéaire et la régression parabolique.

a) Régression linéaire.

La régression linéaire peut être totale si elle exprime la liaison stochastique existant entre la variable dépendante et une seule variable indépendante.

Elle sera partielle si elle exprime la liaison stochastique entre la variable dépendante et plusieurs variables indépendantes.

1° Régression a une seule variable ou régression totale.

Désignons par

$$x_1,\ x_2,\ x_3\ .\ .\ .\ .\ .\ x_n$$
$$y_1,\ y_2,\ y_3\ .\ .\ .\ .\ .\ y_n$$

les deux suites de nombres en correspondance. A chaque couple $(x_i\ y_i)$ correspond, dans le plan un point représentatif d'abscisse x_i et d'ordonnée y_i. Pour déterminer ces coordonnées nous choisirons un « système d'axes cartésien ».

Appliquons la méthode des moindres carrés pour ajuster une droite à la série des « points d'observation » en supposant que la variable x n'est pas sujette à erreur et en « comptant » les déviations parallèlement à l'axe des y.

L'équation générale d'une droite est $y = ax + b$. Les déviations des observations par rapport à cette droite comptées parallèlement à l'axe des y sont de la forme $y - ax - b$. Exprimons que la somme de leurs carrés est minima :

(1) $$\Sigma\ (y - a\ x - b)^2 \text{ minima.}$$

Pour que cette expression soit minima il faut que sa dérivée soit nulle. Comme nous cherchons la valeur des coefficients a et b qui rend minima cette expression, nous exprimerons la condition précédente en écrivant que chacune des dérivées partielles de l'expression (1) par rapport à a et b respectivement est nulle. Nous obtenons ainsi un système de deux équations normales :

(2) $$\Sigma\ x\ (y - ax - b) = 0$$

(3) $$\Sigma\ (y - ax - b) = 0 \qquad \text{soit :}$$

(4) $$\Sigma\ x\ y - a\ \Sigma\ x^2 - b\ \Sigma\ x = 0 \quad \text{et}$$

(5) $$\Sigma\ y - a\ \Sigma\ x - n\ b = 0$$

(n étant le nombre d'observations).

a) *Détermination de coefficient* a.

Pour obtenir la valeur de a nous éliminerons b entre les deux équations. Il suffira pour cela de multiplier (4) par n et (5) par $\Sigma\ x$.

Il vient :

(6) $$n\ \Sigma\ x\ y - a\ n\ \Sigma\ x^2 - b\ n\ \Sigma\ x = 0$$

(7) $$\Sigma\ x\ \Sigma\ y - a\ (\Sigma\ x)^2 - b\ n\ \Sigma\ x = 0$$

d'où l'on tire en retranchant (7) de (6)

$$n\ \Sigma\ x\ y - \Sigma\ x\ \Sigma\ y = a\ [n\ \Sigma\ x^2 - (\Sigma\ x)^2] \quad \text{d'où :}$$

(8) $$a = \frac{n\ \Sigma\ x\ y - \Sigma\ x\ \Sigma\ y}{n\ \Sigma\ x^2 - (\Sigma\ x)^2} = \frac{\Sigma\ x\ y - \dfrac{\Sigma\ y}{n}\ \Sigma\ x}{\Sigma\ x^2 - \dfrac{(\Sigma\ x)^2}{n}}$$

en désignant par $\bar{x}$ la moyenne des x $\left(\bar{x} = \frac{\Sigma x}{n}\right)$ et $\bar{y}$ la moyenne des y, on a :

$$a = \frac{\Sigma x y - \bar{y} \Sigma x}{\Sigma x^2 - \bar{x} \Sigma x}.$$

b) *Détermination du coefficient* b.

Multiplions l'équation (4) par Σx et l'équation (5) par Σx^2. Il vient :

(9) $$\Sigma x y \Sigma x - a \Sigma x^2 \Sigma x - b (\Sigma x)^2 = 0$$

(10) $$\Sigma y \Sigma x^2 - a \Sigma x \Sigma x^2 - n b \Sigma x^2 = 0$$

en retranchant (10) de (9) on obtient :

$$\Sigma x y \Sigma x - \Sigma y \Sigma x^2 = b [(\Sigma x)^2 - n \Sigma x^2], \quad \text{soit :}$$

$$b = \frac{\Sigma x y \Sigma x - \Sigma y \Sigma x^2}{(\Sigma x)^2 - n \Sigma x^2}$$

divisons par $-n$ le numérateur et le dénominateur. L'expression devient :

(11) $$b = \frac{\frac{\Sigma y}{n} \Sigma x^2 - \frac{\Sigma x}{n} \Sigma x y}{\Sigma x^2 - \frac{\Sigma x}{n} \cdot \Sigma x} = \frac{\bar{y} \Sigma x^2 - \bar{x} \Sigma x y}{\Sigma x^2 - \bar{x} \Sigma x}.$$

La droite de régression a donc pour équation :

(12) $$y = \frac{\Sigma x y - \bar{y} \Sigma x}{\Sigma x^2 - \bar{x} \Sigma x} \cdot x + \frac{\bar{y} \Sigma x^2 - \bar{x} \Sigma x y}{\Sigma x^2 - \bar{x} \Sigma x}.$$

Remarque. Si nous transportons l'origine au « centre des moyennes distances » défini par les coordonnées

$$y_m = \frac{\Sigma y}{n}, \quad x_m = \frac{\Sigma x}{n}$$

par rapport au nouveau système d'axe, nous aurons :

$$\frac{\Sigma x}{n} = 0 \quad \frac{\Sigma y}{n} = 0 \quad \text{soit } \Sigma x = \Sigma y = 0$$

en transposant ces valeurs dans (12) on constate que le coefficient b s'annule et que l'équation se réduit à :

(13) $$y = \frac{M_y}{M_{xx}}$$

en désignant par M_{xy} la nouvelle expression de $\Sigma x y$ et M_{xx} la nouvelle expression de Σx^2.

2° Régression partielle.

La régression partielle exprime une liaison stochastique entre la variable dépendante et plusieurs variables indépendantes.

a) *Régression à deux variables.* Nous ajusterons ici aux données une fonction de la forme $Z = a X + b Y + c$. Les paramètres a, b, c seront déterminés

par la condition que la somme des carrés des écarts des valeurs observées Z aux valeurs ajustées (a X + b Y + c) soit minima. Soit :

$$\sum_{1}^{n} (Z - a\,X - b\,Y - c)^2 \text{ minima.}$$

En écrivant, comme précédemment que les dérivées partielles de cette expression par rapport à a, b, c, sont nulles nous obtenons le système d'équations normales :

$$\begin{aligned} \Sigma\, X\,(Z - a\,X - b\,Y - c) &= 0 \\ \Sigma\, Y\,(Z - a\,X - b\,Y - c) &= 0 \\ \Sigma\,(Z - a\,X - b\,Y - c) &= 0 \end{aligned}$$

ce qui s'écrit également :

$$\begin{aligned} 1 &- \Sigma\, X\, Z - a\,\Sigma\, X^2 - b\,\Sigma\, X\, Y - c\,\Sigma\, X = 0 \\ 2 &- \Sigma\, Y\, Z - a\,\Sigma\, Y\, X - b\,\Sigma\, Y^2 - c\,\Sigma\, Y = 0 \\ 3 &- \Sigma\, Z - a\,\Sigma\, X - b\,\Sigma\, Y - c\, n = 0 \end{aligned}$$

Si nous transportons notre origine au « centre des moyennes distances » défini par les coordonnées : $X_m = \frac{\Sigma\, X}{n}$, $Y_m = \frac{\Sigma\, Y}{n}$, $Z_m = \frac{\Sigma\, Z}{n}$, nous pouvons en remplaçant X, Y, Z, par X_m, Y_m, Z_m, dans les 3 équations précédentes, éliminer les termes du 1° degré en X, Y, Z de ces équations.

Celles-ci se réduisent alors à deux équations :

$$\begin{aligned} 4 &- \Sigma\, X_m\, Z_m - a\,\Sigma\, X_m^2 - b\,\Sigma\, X_m\, Y_m = 0 \\ 5 &- \Sigma\, Y_m\, Z_m - a\,\Sigma\, Y_m\, X_m - b\,\Sigma\, Y_m^2 = 0. \end{aligned}$$

Désignons les moments des variables de la manière suivante :

$$\begin{aligned} &M\,xx = \Sigma\, X_m^2, \quad M\,yy = \Sigma\, Y_m^2, \quad M\,zz = \Sigma\, Z_m^2 \\ &M\,xy = \Sigma\, X_m\, Y_m, \quad M\,xz = \Sigma\, X_m\, Z_m, \quad M\,yz = \Sigma\, Y_m\, Z_m. \end{aligned}$$

Les équations 4 et 5 s'écrivent :

$$\begin{aligned} 6 &- M\,xz - a\,M\,xx - b\,M\,xy = 0 \\ 7 &- M\,yz - a\,M\,yx - b\,M\,yy = 0. \end{aligned}$$

En multipliant l'équation (6) par M yy et (7) par M xy il vient :

$$\begin{aligned} 8 &- M\,xz\,M\,yy - a\,M\,xx\,M\,yy - b\,M\,xy\,M\,yy = 0 \\ 9 &- M\,yz\,M\,xy - a\,M\,yx\,M\,xy - b\,M\,yy\,M\,xy = 0. \end{aligned}$$

En retranchant (9) de (8) nous obtenons :

$$M\,xz\,M\,yy - M\,yz\,M\,xy = a\,[M\,xx\,M\,yy - M\,xy^2]$$

d'où l'on tire :

$$10 \qquad a = \frac{M\,xz \times M\,yy - M\,yz \times M\,xy}{M\,xx\,M\,yy - M\,xy^2}.$$

En multipliant l'équation (6) par M xy et l'équation (7) par M xx, ces deux équations deviennent :

$$\begin{aligned} 11 &- M\,xz\,M\,xy - a\,M\,xx\,M\,xy - b\,M\,xy^2 = 0 \\ 12 &- M\,yz\,M\,xx - a\,M\,xy\,M\,xx - b\,M\,yy\,M\,xx = 0. \end{aligned}$$

Retranchons (12) de (11). Il vient :

$$M\,xz\,M\,xy - M\,yz\,M\,xx = b\,[M\,xy^2 - M\,yy\,M\,xx]$$

d'où l'on tire :

$$13 \qquad b = \frac{M\,yz\,M\,xx - M\,xz\,M\,xy}{M\,xx\,M\,yy - M\,xy^2}.$$

L'équation de la droite de régression sera :

$$Z_m = \frac{M\,xz \cdot M\,yy - M\,yz\,M\,xy}{M\,xx\,M\,yy - M\,xy^2} X_m + \frac{M\,yz\,M\,xx - M\,xz\,M\,xy}{M\,xx\,M\,yy - M\,xy^2} Y_m.$$

b) *Relations entre la régression partielle à deux variables et la régression totale.*

Dans un article paru dans *Econometrica* (1) MM. Ragnar Frisch et Frederick V. Waugh ont montré l'identité des méthodes de régression partielle et multiple.

« Il y a d'ordinaire, nous disent-ils, deux méthodes de traitement du trend linéaire dans l'analyse de la corrélation des séries de données temporelles : la première est de baser l'analyse sur les déviations des trends ajustés séparément à chacune des séries originelles, et la seconde est de baser l'analyse sur les séries originelles sans élimination du trend, mais en introduisant le temps lui-même comme variable supplémentaire dans une analyse de corrélation multiple.

La première méthode peut être appelée : « méthode du trend individuel » et la dernière : « méthode de la régression partielle ».

« Il y a certaines conceptions erronées quant à la valeur relative des deux méthodes et quant au genre de résultats statistiques qu'elles permettent d'obtenir. »

Et ils ajoutent plus loin :

« La méthode de régression partielle du trend ne peut en vérité jamais accomplir ce que la méthode du trend individuel ne peut, parce que les deux méthodes conduisent par définition à des résultats identiques. Elles diffèrent seulement dans la technique de calcul utilisée pour arriver au résultat. »

C'est ce que nous allons démontrer en reprenant et en complétant les calculs présentés par ces deux auteurs de manière à en faciliter la compréhension. Nous nous permettrons également d'en modifier légèrement le mode de notations pour les rendre comparables avec celles qui nous ont conduit à la détermination des coefficients de régression partielle.

Nous formulerons le « théorème » à démontrer de la façon suivante :

Le coefficient de régression partielle de la variable dépendante par rapport à l'une des variables indépendantes est égal au coefficient de régression totale des déviations de la variable dépendante, comptées à partir de la droite de régression totale de cette même variable, par rapport aux déviations de la variable indépendante considérée, comptées à partir de la droite de régression totale de cette même variable.

---

(1) « Partial time Regressions as compared with individual trends » by R. Frisch et F. V. Waugh. Econometrica octobre 1933, n° 4, p. 387.

Soient z, x, y les trois variables, z étant la variable dépendante, x et y, les variables indépendantes. En prenant pour origine des coordonnées le centre des moyennes distances, la droite de régression totale entre z et x a pour équation :

$$z = \frac{\Sigma\, z\, x}{\Sigma\, x^2} \cdot x$$

La droite de régression totale entre z et y a pour équation :

$$z = \frac{\Sigma\, z\, y}{\Sigma\, y^2} \cdot y$$

La droite de régression totale de y en x a pour équation :

$$y = \frac{\Sigma\, y\, x}{\Sigma\, x^2} \cdot x$$

et enfin, la droite de régression totale de x en y est représentée par :

$$x = \frac{\Sigma\, x\, y}{\Sigma\, y^2} \cdot y$$

Si nous exprimons par x', y', z', les déviations comptées à partir de ces droites de régression totale nous pouvons écrire :

(1) $$z' = z - \frac{\Sigma\, z\, x}{\Sigma\, x^2} \cdot x$$

(2) $$z' = z - \frac{\Sigma\, z\, y}{\Sigma\, y^2} \cdot y$$

(3) $$y' = y - \frac{\Sigma\, y\, x}{\Sigma\, x^2} \cdot x$$

(4) $$x' = x - \frac{\Sigma\, x\, y}{\Sigma\, y^2} \cdot y$$

Si nous nous reportons à la relation (10) de l'étude de la corrélation partielle, nous y trouvons la valeur du coefficient de régression a.

Notre théorème signifie que le coefficient a est égal au coefficient de régression totale de z' par rapport à x' soit :

$$a = \frac{\Sigma\, z'\, x'}{\Sigma\, x'^2} \quad \text{et que même } b = \frac{\Sigma\, z'\, y'}{\Sigma\, y'^2}.$$

Nous allons démontrer ces égalités.

1° *Démonstration de l'égalité :*

$$a = \frac{\Sigma\, z'\, x'}{\Sigma\, x'^2}.$$

Nous avons :

$$x' = x - \frac{\Sigma\, x\, y}{\Sigma\, y^2} \cdot y$$

nous prendrons pour valeur de z' :

$$z' = z - \frac{\Sigma\, z\, y}{\Sigma\, y^2} \cdot y.$$

Désignons respectivement par N et par D le numérateur et le dénominateur de la fraction $\frac{\Sigma z' x'}{\Sigma x'^2}$.

Nous aurons :

$$N = \Sigma\left(z - \frac{\Sigma yz}{\Sigma y^2} \cdot y\right)\left(x - \frac{\Sigma xy}{\Sigma y^2} \cdot y\right)$$

ou en développant :

$$N = \Sigma\left[zx - zy\frac{\Sigma xy}{\Sigma y^2} - xy\frac{\Sigma yz}{\Sigma y^2} + y^2\frac{\Sigma xy\,\Sigma yz}{(\Sigma y^2)^2}\right]$$

ou, en reprenant nos notations précédentes : $M\,xx = \Sigma x^2$, $M\,yy = \Sigma y^2$ $M\,zz = \Sigma z^2$, $M\,xy = \Sigma xy$, $M\,xz = \Sigma xz$, $M\,yz = \Sigma yz$ (ce qui nous est permis puisque x, y, z sont comptés à partir du centre des moyennes distances) nous obtenons :

$$N = M\,zx - M\,zy \cdot \frac{M\,xy}{M\,yy} - M\,xy\frac{M\,yz}{M\,yy} + M\,yy \cdot \frac{M\,xy\,M\,yz}{M\,yy^2}$$

soit, après simplifications :

$$N = \frac{M\,zx\,M\,yy - M\,zy\,M\,xy}{M\,yy}. \tag{5}$$

Calculons maintenant le dénominateur :

$$D = \Sigma\left(x - \frac{\Sigma xy}{\Sigma y^2} \cdot y\right)^2$$

en développant il vient :

$$D = \Sigma\left[x^2 - 2xy\frac{\Sigma xy}{\Sigma x^2} + y^2\frac{(\Sigma xy)^2}{(\Sigma y^2)^2}\right]$$

soit :

$$D = M\,xx - 2\,M\,xy\frac{M\,xy}{M\,yy} + M\,yy \cdot \frac{(M\,xy)^2}{(M\,yy)^2}$$

et après simplifications :

$$D = \frac{M\,xx\,M\,yy - M\,xy^2}{M\,yy} \tag{6}$$

En divisant (5) par (6) nous obtenons :

$$\frac{N}{D} = \frac{M\,xz \cdot M\,yy - M\,zy\,M\,xy}{M\,xx\,M\,yy - M\,xy^2}$$

expression qui est identique à celle du coefficient a de la relation (10).

2° *Démonstration de l'égalité :*

$$b = \frac{\Sigma z'y'}{\Sigma y'^2}$$

Nous avons :

$$z' = z - \frac{\Sigma zx}{\Sigma x^2} \cdot x$$

$$y' = y - \frac{\Sigma yx}{\Sigma x^2} \cdot x$$

en désignant par N et D les numérateur et dénominateur de la fraction $\frac{\Sigma z' y'}{\Sigma y'^2}$; nous pouvons écrire :

$$N = \Sigma \left(z - \frac{M\,z\,x}{M\,x\,x} \cdot x\right)\left(y - \frac{M\,y\,x}{M\,x\,x} \cdot x\right)$$

soit en développant :

$$N = \Sigma \left[z\,y - z\,x \frac{M\,y\,x}{M\,x\,x} - x\,y \frac{M\,z\,x}{M\,x\,x} + x^2 \frac{M\,z\,x\,M\,y\,x}{M\,x\,x^2}\right] =$$

$$M\,z\,y - M\,z\,x \frac{M\,y\,x}{M\,x\,x} - M\,x\,y \frac{M\,z\,x}{M\,x\,x} + M\,x\,x \frac{M\,z\,x\,M\,y\,x}{M\,x\,x^2} =$$

$$(7) \qquad \frac{M\,z\,y\,M\,x\,x - M\,x\,z\,M\,x\,y}{M\,x\,x}$$

Calculons maintenant le dénominateur :

$$D = \Sigma \left(y - \frac{M\,y\,x}{M\,x\,x} \cdot x\right)^2$$

$$D = \Sigma \left[y^2 - 2\,x\,y \frac{M\,y\,x}{M\,x\,x} + x^2 \frac{(M\,y\,x)^2}{(M\,x\,x)^2}\right] =$$

$$M\,y\,y - 2 \frac{M\,x\,y^2}{M\,x\,x} + M\,x\,x \frac{M\,x\,y^2}{M\,x\,x^2} =$$

$$(8) \qquad \frac{M\,y\,y\,M\,x\,x - M\,x\,y^2}{M\,x\,x}.$$

En divisant (7) par (8) nous trouvons :

$$\frac{N}{D} = \frac{M\,z\,y\,M\,x\,x - M\,x\,z\,M\,x\,y}{M\,y\,y\,M\,x\,x - M\,x\,y^2}$$

expression qui est bien égale à celle de b donnée plus haut.

Le théorème énoncé se trouve ainsi démontré.

c) *La régression partielle à n variables.*

Nous avons vu que pour obtenir le système d'équations normales il suffisait de prendre les dérivées partielles par rapport aux paramètres a, b, c,... des variables. Si nous avons n variables, nous aurons n paramètres donc n équations normales. Il suffit de résoudre ce système de n équations à n inconnues pour trouver la solution du problème, c'est-à-dire les valeurs a, b, c,... k, des paramètres qui rendent minima la somme des carrés des déviations.

b) Régression parabolique.

Nous distinguerons la régression parabolique du 2e degré et la régression « parabolique » de degré quelconque n.

1° *Régression parabolique du 2e degré.*

La formule d'ajustement sera donnée par la fonction :

$$(1) \qquad Z = a + b\,X + c\,X^2.$$

Il nous suffira donc de poser $X^2 = Y$ pour transformer la relation (1) en une relation linéaire à deux variables (1).

$$Z = a + b\,X + c\,Y \qquad (2)$$

ce qui nous ramène au cas précédemment étudié.

Notons avec M. Dugé de Bernonville que si les deux variables X et Y de la nouvelle relation (2) ne sont pas indépendantes, celà n'a aucune importance car la condition d'indépendance n'est pas nécessaire pour l'application de la méthode de régression linéaire.

2° *Régression « parabolique » de degré n.*

Il suffit d'appliquer le même raisonnement que précédemment pour ramener cette régression à une régression linéaire à n variables.

Notons que si la « fonction parabolique » d'ajustement se réduit à $Y = A\,X^{\alpha}$ il devient très facile de l'ajuster sur un graphique à double échelle logarithmique puisque nous avons $\text{Log } Y = \text{Log } A + \alpha \text{ Log } X$. De même si nous avions la fonction :

$$Z = A\,X^{\alpha}\,Y^{\beta} \ldots\ldots$$

nous aurions à résoudre un problème du même genre, l'ajustement se ramenant à celui d'une fonction linéaire à 2, 3... n variables. Ainsi pour :

$$A\,X^{\alpha}\,Y^{\beta}\,e^{\gamma\,\delta\,+\,\varepsilon}$$

notre formule d'ajustement linéaire sur le graphique à double échelle logarithmique sera :

$$\text{Log}_e\,Z = \text{Log}_e\,A + \alpha\,\text{Log}_e\,X + \beta\,\text{Log}_e\,Y + \gamma\,\delta + \varepsilon$$

en prenant les logarithmes népériens.

c) La régression par formule « exponentielle »

La formule d'ajustement sera de la forme $Y = A\,B^{x}$. En prenant les logarithmes des deux membres nous la ramenons à :

$$\text{Log } Y = \text{Log } A + X \text{ Log } B$$

en posant $\begin{cases} \text{Log } Y = y, & X = x \\ \text{Log } A = b, & \text{Log } B = a \end{cases}$

elle devient : $y = ax + b$.

---

(1) *cf.* M. Léo Dugé de Bernonville, *Initiation à l'analyse statistique*, p. 113.
Pour les exemples pratiques d'application des différentes formules ainsi que pour la méthode d'ajustement par une fonction périodique, nous renvoyons le lecteur à cet excellent ouvrage, p. 108 à 116.

II — RÉGRESSION MUTUELLE

1° **Justification de la méthode.** — Nous avons signalé précédemment l'arbitraire consistant, dans la méthode de régression unilatérale à compter les déviations parallèlement à l'un des axes, celui de la variable dépendante. La méthode de régression unilatérale implique en effet l'hypothèse selon laquelle les déviations de N paires d'observations par rapport à l'équation de régression peuvent être attribuées entièrement aux variations de la variable dépendante pour des valeurs, données par l'observation, des variables indépendantes, celles-ci étant supposées être connues très exactement.

Si nous étudions par exemple la relation prix-quantités demandées, l'équation de notre courbe de régression sera différente suivant que l'erreur aura été imputée uniquement au prix (hypothèse où la quantité est la variable indépendante) ou uniquement à la quantité, celle-ci étant prise comme variable dépendante.

Mais, dans ces conditions, chaque régression n'a-t-elle pas son domaine propre? Ne devons-nous pas prendre la régression dans laquelle le prix est la variable dépendante quand notre but est d'expliquer les facteurs affectant le prix, et la régression dans laquelle la quantité est la variable dépendante quand l'explication des facteurs affectant la consommation nous intéresse davantage? Il n'y aurait aucune objection sérieuse à formuler à l'encontre de ce procédé ; toutefois, il nous faut reconnaître qu'en pratique, hormis le cas d'un marché de monopole, toutes les observations tant des séries de quantité que de prix sont sujettes aux erreurs.

Il est évident par conséquent que, dans la grande majorité des cas, un meilleur ajustement, ou une courbe de demande plus probable serait obtenue en prenant en considération les deux types d'erreurs dans le procédé d'ajustement. Malheureusement cette nouvelle méthode que nous appellerons « méthode de régression mutuelle » ne peut être appliquée facilement qu'à l'ajustement des fonctions linéaires. Les mathématiciens n'ont pas réussi à l'étendre aux fonctions non linéaires.

Toutefois, comme nous l'avons montré au cours de l'étude de la régression unilatérale, la « forme linéaire » permet indirectement la représentation de la « forme parabolique (1) » et de la « forme exponentielle » et ces trois formes simples suffisent amplement pour la représentation d'une majorité de trends.

Dans son ouvrage *Statistical laws of demand and Supply*, Henry Schultz a procédé à la détermination de la ligne de régression mutuelle, pour laquelle les déviations sont comptées parallèlement à la normale de la ligne de régression elle-même. Utilisant le procédé des moindres carrés, il a minimisé la somme des carrés des déviations perpendiculaires des points d'observation à la droite (dans l'hypothèse d'une seule variable indépendante).

L'extension de cette méthode aux cas où il y a plusieurs variables indépen-

(1) Nous entendons ici par « forme parabolique » uniquement la forme $Y = A X^a$. L'ajustement d'une parabole du second degré de la forme $Y = a + bX + cX^2$ étant extrêmement difficile à effectuer. Cf. G. Pietra, *Interpolating Plane Curves*, Metron III, 1923-1924, p. 311, et du même auteur *Dell' interpolazione parabolica nel caso in cui entrambi i valori delle variabili sono affetti da errori accidentali*, Metron III et IV, 1932, p. 77-85.

dantes ne peut se faire sans formuler d'hypothèses concernant les rapports des coefficients de toutes les variables indépendantes, au coefficient de la variable dépendante.

(Il convient de faire remarquer qu'ici le terme de variable dépendante ou indépendante ne préjuge plus d'un sens à donner à la liaison stochastique. En fait, si l'on veut employer une terminologie rigoureuse il n'y a plus de variables dépendantes ou indépendantes, il y a des variables tout court. Il nous suffira alors de remplacer dans l'expression précédente « toutes les variables indépendantes » par « toutes les variables sauf une », et « variable dépendante » par « variable restante ») (1).

Nous étudierons ici seulement l'une des méthodes de régression mutuelle à deux variables (c'est-à-dire dont la formule d'ajustement est représentée par l'équation $Y = a + b X$.), et pour laquelle le « rapport d'imputation » de l'erreur est supposé égal à 1 (c'est-à-dire que l'on suppose que les deux variables sont sujettes à erreur exactement dans la même mesure, cette hypothèse est d'ailleurs la seule que l'on puisse formuler *a priori* dans une étude générale, seule l'étude d'un cas particulier pourrait permettre de fixer un co-rapport d'imputation différent de 1). La méthode qui nous a semblé la plus simple est celle de la minimisation des aires (cf. *Revue Econometrica* « The method of minimized aeras as a basis for correlation analysis » (2).

2° **La méthode des « moindres aires ».** — Les deux équations de régression unilatérale : de Y par rapport à X et de X par rapport à Y ne sont pas compatibles.

Soit en effet l'équation de régression totale (1) $Y = \frac{M\,xy}{M\,xx} \cdot X$ de Y par

(1) Pour un ajustement à trois variables,

Cf. H. Schultz, *Statistical laws of demand and Supply*, p. 178-186.

Schultz prend soin d'appliquer sa méthode seulement aux variables exprimées en termes de chaînes relatives ou de rapports de tendance afin d'éviter que les résultats obtenus soient dépendants des unités qui ont servi à mesurer les variables.

Sur cette dernière question cf. Charles Frederick Roos, *A general Invariant Criterion of Fit for Lines and Planes Where all Variables Are Subject to Error*, Metron VIII, n° 1, 1937, p. 3 à 20 et Herbert Jones, *Some Geometrical considerations in the General Theory of Fitting Lines and Planes*, Metron VIII, n° 1 (1937) p. 21 à 30.

(2) Article cité. *Econometrica*. Janv. 1941, p. 38, par M. Elliot, B. Wooley.

Sur les autres méthodes employées cf. :

K. Pearsons « On lines and Planes of closest Fit » *Philosophical Magazine*, 1901 et dans la même revue (1927) :

E. Rhodes.

M. J. Van Uven « Adjustment of N points (in *n*-dimensional Space) to the Best linear (*n*—1) dimensional Space » koninklijke Akademie van Wetenshapen te Amsterdam, *Proceedings of the Section of Science*, vol. 23, 1930, p. 143.

T. Koopmans « Linear Regression Analysis of Economic Time series » Haarlem, 1937.

C. Gini « Sull' interpolazione di una retta quando i valore della variabile independente sono affetti di errori accidentali » *Metron*, vol. III, fév. 1924.

R. G. D. Allen « The assumptions of Linear Regression » *Economica New series*, vol. 6, mai 1939, p. 191.

W. E. Deming « Statistical Adjustment of Data » New York, 1943.

A. Wald « The fitting of Straight lines if both variables are subject to error » *Annals of Mathematical Statistics*, vol. XI, sept. 1940, p. 284.

Gerhard Tintner « An application of the variable Difference Method to multiple Regression » *Econometrica*, avril 1944.

rapport à X, et l'équation totale de régression de X par rapport à Y : $X_c = \frac{M\,xy}{M\,yy} Y$ (2).

Si nous tirons X de (1) nous obtenons :

$$X = \frac{M\,xx}{M\,xy} Y \text{ et d'après (2) } \quad X_c = \frac{M\,xy}{M\,yy} Y.$$

Pour que ces deux expressions de X soient égales, il faut que :

$$\frac{M\,xx}{M\,xy} = \frac{M\,xy}{M\,yy}$$

c'est-à-dire que l'on ait $M\,xy^2 = M\,xx\ M\,yy$
en désignant par r^2 le quotient :

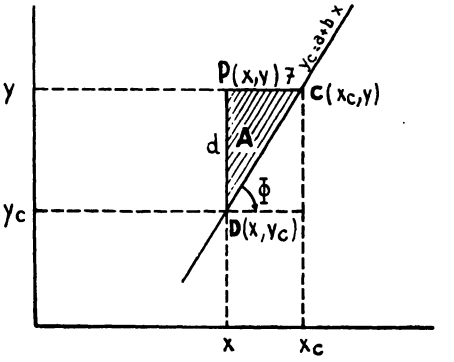
$$\frac{M\,xx\ M\,yy}{M\,xy^2}$$

nous dirons que les deux équations de régression ne sont compatibles que si $r^2 = 1$.

La méthode de régression mutuelle que nous allons présenter nous permettra d'obtenir pour X et X_c des valeurs identiques.

A. — *Régression linéaire de* Y *par rapport à* X.

Le problème qui se pose est celui de l'évaluation des paramètres de $Y = a + b\,X$ de manière à ce que la somme de N aires, limitées chacune par la courbe de liaison et les lignes parallèles aux axes de coordonnées (rectangulaires) passant par chaque couple d'observation soit minimisée.



Soit A la surface du triangle défini ci-dessus (surface hachurée) L'aire de déviation (A) est définie comme suit :

$$A = \frac{df}{2}$$

mais l'on a :

$$tg\ \Phi = b = \frac{d}{f}$$

d'où l'on tire :

$$f = \frac{d}{b}$$

et, en portant cette valeur dans A :

$$A = \frac{d^2}{2\,b}$$

En faisant la somme des n observations nous aurons :

$$\Sigma\, A = \frac{\Sigma\, d^2}{2\,b} \tag{1}$$

Y_c désignant les ordonnées des joints de la droite nous avons :

$$d = Y - Y_c = Y - a - b\,X$$

et par suite :

$$\Sigma\, A = \frac{\Sigma\, d^2}{2\,b} = \frac{\Sigma\,(y - a - b\,X)^2}{2\,b} \tag{2}$$

Il s'agit de déterminer a et b de manière que Σ A soit minimum. Pour cela il suffira d'annuler les dérivées partielles de Σ A par rapport à a et b.

Nous trouverons ainsi deux équations normales.

1° équation normale :

$$\frac{d\,\Sigma\, A}{d\,a} = \frac{d\,\Sigma}{d\,a}\left[\frac{(Y - a - b\,X)^2}{2\,b}\right] = -\,\Sigma\,\frac{2(Y - a - b\,X)}{2\,b}$$

$$= -\,\Sigma\,\frac{(Y - a - b\,X)}{b} = o \tag{3}$$

Pour que (3) soit nul il faut que son numérateur le soit :

$$\Sigma\,(Y - a - b\,X) = o \text{ ou } \Sigma\,(Y - Y_c) \text{ nul.}$$

Condition qui s'écrit :

$$\Sigma\, Y = \Sigma\, Y_c = \Sigma\,(a + b\,X) = N\,a + b\,\Sigma\, X$$

ce qui donne :

$$a = \frac{\Sigma\, Y - b\,\Sigma\, X}{N}$$

2° équation normale :

$$\frac{d\,\Sigma\, A}{d\,b} = \frac{d\,\Sigma}{d\,b}\left[\frac{(Y - a - b\,X)^2}{2\,b}\right] =$$

$$\frac{\Sigma}{2}\left[\frac{2\,(Y - a - b\,X) - (X \cdot b) - (Y - a - b\,X)^2}{b^2}\right] = o. \tag{4}$$

en appliquant la formule de dérivation d'un quotient $\dfrac{u'\,v - u\,v'}{v^2} = \left(\dfrac{u}{v}\right)'$.

Pour que la relation (4) soit vérifiée il faut que son numérateur soit nul; or celui-ci correspond à la fin du développement du carré :

$$[(Y - a - b\,X) + b\,X]^2$$

Nous pouvons l'écrire :

$$-\Sigma\left[[(Y-a-bX)+bX]^2-b^2X^2\right]=\Sigma\left[(Y-a)^2-b^2X^2\right]=0$$

ce qui peut être mis sous la forme :

$$\Sigma(Y-a-bX)(Y-a+bX)=\Sigma(Y-a-bX)[(Y-a)+bX]=$$
$$\Sigma(Y-a)(Y-a-bX)+\Sigma bX(Y-a-bX)=0$$

ce qui s'écrit :

$$\Sigma bX(Y-a-bX)=-\Sigma(Y-a)(Y-a-bX)=$$
$$-\Sigma[Y(Y-Y_c)-a(Y-Y_c)]=-\Sigma Y(Y-Y_c)+a\Sigma(Y-Y_c)$$

Or en vertu de la relation :

$$\Sigma(Y-Y_c)=0 \tag{3}$$

nous avons donc :

$$\Sigma bX(Y-Y_c)=-\Sigma Y(Y-Y_c).=$$
$$-\Sigma Y(Y-a-bX) \tag{5}$$

En développant l'expression (5) nous obtenons :

$$b\Sigma XY-abX-b^2\Sigma X^2=-\Sigma Y^2+a\Sigma Y+b\Sigma XY$$

en remplaçant a par sa valeur tirée de (3) il vient :

$$ab\Sigma X+b^2\Sigma X^2=\Sigma Y^2-a\Sigma Y$$

équivalent à :

$$b\Sigma X\frac{(\Sigma Y-b\Sigma X)}{N}+b^2\Sigma X^2=\Sigma Y^2-\Sigma Y\frac{(\Sigma Y-b\Sigma X)}{N}$$

$$\frac{b\Sigma X\Sigma Y}{N}-\frac{b^2(\Sigma X)^2}{N}+b^2\Sigma X^2=\Sigma Y^2-\frac{(\Sigma Y)^2}{N}+\frac{b\Sigma X\Sigma Y}{N}$$

d'où :

$$b^2\left[\Sigma X^2-\frac{(E\ X)^2}{N}\right]=\Sigma Y^2-\frac{(E\ Y)^2}{N}$$

en divisant les deux membres de cette dernière égalité par N on obtient :

$$b^2\left[\frac{\Sigma X^2}{N}-\left(\frac{\Sigma X}{N}\right)^2\right]=\frac{\Sigma Y^2}{N}-\left(\frac{\Sigma Y}{N}\right)^2$$

le coefficient de b^2 est ce que l'on appelle la variance ou fluctuation de la série de X (c'est-à-dire le carré de la standard déviation ou écart type ou écart quadratique moyen σ_x). Le second terme de l'égalité représente la fluctuation de la série des Y soit σ_y^2.

on a donc $b^2=\frac{\sigma_y^2}{\sigma_x^2}$ d'où $b=\pm\frac{\sigma_y}{\sigma_x}$ le signe de b sera déterminé par le signe de $\Sigma\, xy$.

B. — *Régression linéaire de* X *par rapport à* Y.

Si nous prenons comme formule d'ajustement l'équation linéaire $X_c = k + eY$, le même raisonnement que précédemment nous conduit à :

$$\sigma y^2\, e^2 = \sigma x^2$$

d'où l'on tire :

$$e = \frac{\sigma x}{\sigma y}$$

on a donc : $e\,b = 1$.

Par analogie à la relation (3) nous aurions également

$$\Sigma X = N k + e \Sigma Y$$

comme par ailleurs nous avons :

$$\Sigma Y = N a + b \Sigma X \quad (3)$$

il vient :

$$= N a + b N k + b e \Sigma Y; \quad \text{comme } b e = 1 :$$

$$+ b N k + \Sigma Y; \quad N a = - b N k, \quad k = -\frac{b}{a}.$$

$+ e Y_c$ peut alors s'écrire :

$$X c = k + e (a + b X) = k + ea + eb X = -\frac{b}{a} + \frac{a}{b} + X$$

on a bien $X_c = X$.

Ainsi les deux équations sont des fonctions compatibles entre elles, exprimant par conséquent la même relation linéaire entre X et Y (1).

Roger CONGARD

***