



La GREAT formation

Groupe de recherche en économie appliquée et théorique

N° 002

"Réfléchir à changer"

Février 2015

Introduction à l'économétrie



Massa Coulibaly

Table des matières

Introduction	1
1. Régression linéaire	3
1.1. Les données	4
1.2. Les relations	4
1.2.1. Types de relation	4
1.2.2. Formes de relations	4
1.3. Les variables	6
1.4. Les méthodes de régression	6
1.4.1. Méthode des moments	6
1.4.2. Méthode du maximum de vraisemblance	7
1.4.3. Méthodes des moindres carrés	7
2. Modèle linéaire simple	10
2.1. Hypothèses	11
2.2. Estimation	12
2.3. Propriétés	14
2.4. Inférence statistique	22
2.5. Coefficient de corrélation	25
2.6. Analyse de variance	31
2.7. Prévision	33
3. Extensions du modèle linéaire simple par les moindres carrés	39
3.1. Les singularités	40
3.2. Relation non linéaire	40
3.3. Modèle à 3 variables	45
3.3.1. Régression	47
3.3.2. Coefficient de corrélation multiple	49
3.3.3. Coefficients de corrélation partiels	51
3.3.4. Résumé des calculs	54

Introduction



Les propositions mathématiques
sont reçues comme vraies parce que
personne n'a intérêt qu'elles soient fausses

Montesquieu

C'est parce qu'il n'est pas prudent de raisonner sans modèle qu'il faut modéliser, en évitant ainsi de faire des propositions pas toujours réalisables ni même réalistes. On utilise les modèles pour:

- proposer des politiques économiques
- présenter des arguments raisonnables afin de justifier les politiques économiques
- expliquer et influencer les stratégies des entreprises
- fournir des conseils aux ménages
- planifier et allouer les ressources.

Pour jouer autant de rôles, un modèle doit

- faire ressortir les caractéristiques essentielles de la réalité économique
- aider à cerner la nature des chocs et réformes économiques
- faire ressortir les liens entre agents économiques (canaux de transmission)
- assurer l'existence d'un équilibre de long terme (ou état stationnaire).

En économie, un **modèle** est une représentation simplifiée de la réalité économique ou d'une partie de celle-ci. Sa construction commence par la constitution d'une banque de données i.e. de l'ensemble des séries statistiques nécessaires. Le cadre comptable est emprunté à la comptabilité nationale et consiste en un Tableau économique d'ensemble (TEE) qui retrace les ressources et les emplois des différents agents pour les différentes opérations.

Les méthodes sont nombreuses et variées

Modélisation mathématique	Données quantitatives		
		Sans	Avec
	Sans	Economie littéraire	Statistique
	Avec	Economie mathématique	Econométrie

Des applications plus précises des mathématiques sont requises :

- matrice de coefficients techniques pour les équilibres
- recherche opérationnelle (partie de la théorie d'optimisation) pour la gestion
- théorie des jeux pour les relations entre plusieurs décideurs
- statistiques pour l'économétrie.

Quel que soit le type de modèle retenu (macroéconométrique, VAR, MEGC, RBC), le passage du modèle théorique au modèle appliqué se fait par étapes successives dont:

- la construction du modèle théorique
- le chiffrage qui se fait soit par estimation économétrique soit par calibrage
- l'utilisation du modèle.

On distingue deux types d'équations du modèle

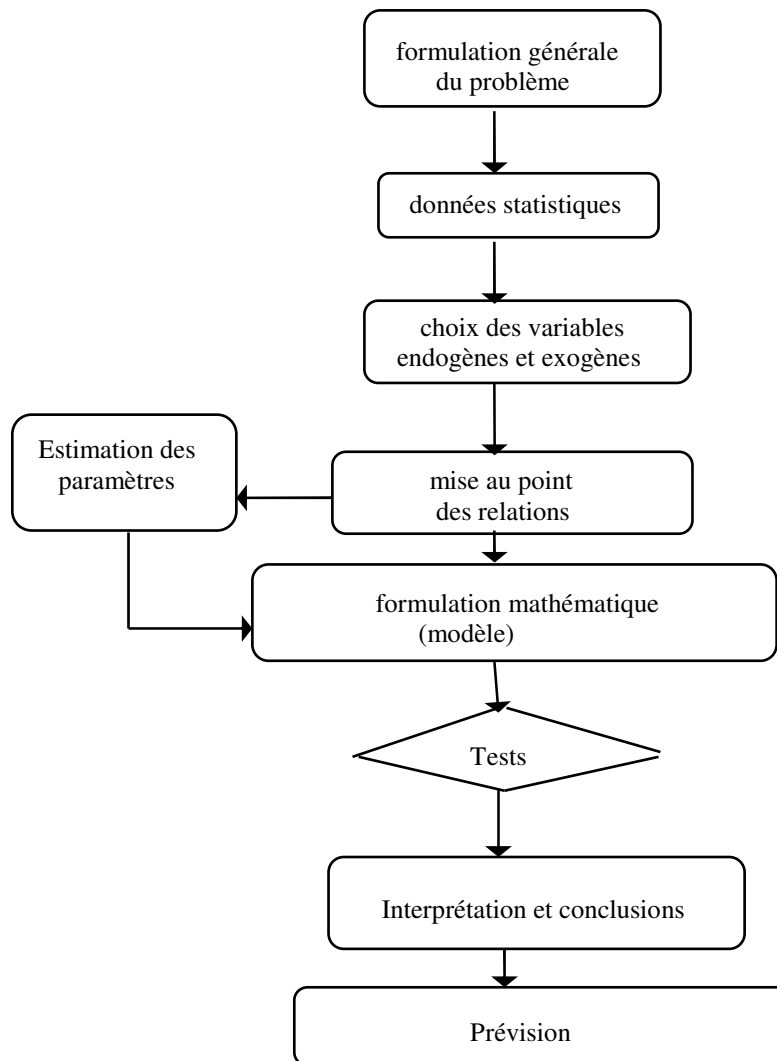
- équations de comportement
- équations comptables.

Pour la spécification des comportements, on distingue la forme structurelle, où apparaissent les relations de comportements et les formes réduites qui résultent de la résolution d'un ensemble de relations structurelles (courbe IS ou courbe de demande, courbe d'offre, courbe LM) et permettent de comprendre les mécanismes économiques de la première.

Dès que le modèle est achevé (après tests de robustesse et de validité des résultats), les équations sont rassemblées dans un logiciel de résolution qui permet d'obtenir la solution numérique du modèle sur sa période d'estimation pour voir dans quelle mesure il reproduit correctement la réalité e.g. GAMS, etc.

1. Régression linéaire

La modélisation en économétrie comportera les différents pas successifs suivants :



Nous distinguons aussi les éléments:

- ☐ données
- ☐ relations
- ☐ variables

1.1. Les données

Les données qu'on observe en économie sont:

- soit des observations individuelles sur des unités individuelles en un moment précis (e.g. le salaire d'une catégorie socioprofessionnelle)
- soit des séries chronologiques sur la même unité (e.g. le PNB annuel)
- soit des données de panel, combinant les deux précédentes catégories.

1.2. Les relations

1.2.1. Types de relation

Elles sont de 3 types:

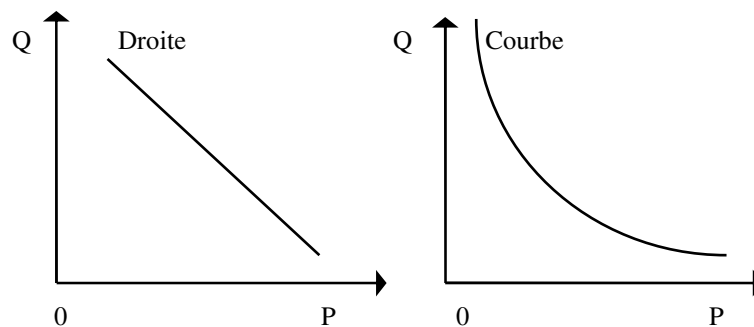
- **Une simple équation:** Une variable est déterminée à l'aide d'une ou de plusieurs variables indépendantes, e.g.: la demande d'un bien fonction de son prix $q = f(p)$ ou la consommation fonction du revenu, de l'offre et du niveau des prix $C = f(y, q, p)$
- **Equations multiples:** Deux ou plusieurs équations distinctes qui peuvent être traitées soit séparément (comme dans le cas d'une simple équation), soit globalement, e.g.: $C_i = f(y_i, q_i, p_i) \quad i = \overline{1, n}$
- **Equations simultanées:** Deux ou plusieurs variables sont déterminées simultanément à partir d'un nombre de variables indépendantes, e.g. l'équilibre sur un marché: $D = f(p, x)$ et $S = f(p, z)$ avec à l'équilibre: $D = S$

1.2.2. Formes de relations

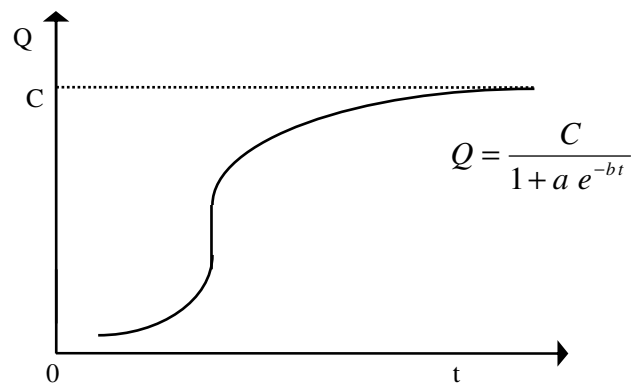
Quant aux *formes de ces relations*, elles sont multiples. Ainsi par exemple si la théorie économique nous dit que la demande est fonction du prix, elle ne précise pas toutefois la forme de cette relation qui peut être:

- linéaire $Q = \alpha + \beta P$
- logarithmique $\ln Q = \alpha + \beta \ln P$ ou $Q = e^{\alpha} P^{\beta}$
- semi-logarithmique $\ln Q = \alpha + \beta P$ ou $Q = e^{\alpha + \beta P}$

La forme à retenir peut être déduite de l'observation même des données en nous servant par exemple de la représentation graphique:



Dans certains cas, aucune forme simple de relation peut n'être satisfaisante, e.g. les phases de lancement d'un nouveau produit (fonction logistique).



Note A cause des erreurs d'observations et de certaines corrélations (partant la colinéarité) la forme de relation qu'on retient ne sera généralement pas exacte.

1.3. Les variables

On distingue:

- variables endogènes ou variables expliquées (déterminées par le modèle)
- variables exogènes ou variables explicatives (indépendantes ou prédéterminées)
- variable erreur (pour synthétiser l'effet de tous les facteurs non pris en compte) ou résidu (lorsqu'il s'agit d'erreurs sur la variable endogène).

1.4. Les méthodes de régression

1.4.1. Méthode des moments

Pour déterminer les paramètres, on égalise les moments (e.g. moyenne et variance) d'échantillon aux moments correspondants de la population.

Ex.1 A la v.a. X on associe la densité de répartition:

$$f(X, a, b) = \frac{1}{b^p \Gamma(p)} \left(\frac{X-a}{b} \right)^{p-1} \exp\left(-\frac{X-a}{b}\right) \quad \text{avec } X \geq a, \quad b, p > 0$$

1) Montrer que les deux premiers moments sont tels que:

$$m_1 = \int_a^{\infty} X f(X, a, b) dX = \frac{bp + a}{b^{p-1}}$$

$$m_2 = \int_a^{\infty} X^2 f(X, a, b) dX = \frac{p(p+1)b^2 + 2apb + a^2}{b^{p-1}}$$

2) Pour $p=2$ et 2 observations ($X_1=1$ et $X_2=3$) montrer que les estimateurs de la méthode des moments sont: $\hat{a} = 0$ et $\hat{b} = \frac{5}{6}$

1.4.2. Méthode du maximum de vraisemblance

Elle consiste à choisir comme estimation $\hat{\theta}$ du paramètre θ qui maximise la fonction de vraisemblance $L(X_1, X_2, \dots, X_n; \theta)$ par Likelihood.

$\hat{\theta}$ est solution de: $\frac{\partial L(\cdot)}{\partial \theta} = 0$ et $\frac{\partial^2 L(\cdot)}{\partial \theta^2} < 0$ qu'on remplace par cet autre système:

$$\frac{\partial \ln L}{\partial \theta} = 0 \text{ et } \frac{\partial^2 \ln L}{\partial \theta^2} < 0 \text{ si } f(X, \theta) > 0 \text{ (f densité}$$

de probabilité)

Ex.2. On considère un échantillon de n observations indépendantes de la v.a. X définie sur $\mathbb{R}^+$, de densité:

$$f(X) = \frac{1}{\Gamma(p)} e^{-\frac{X}{a}} \frac{X^{p-1}}{a^p} ; \quad a > 0, \quad p \text{ donné}$$

1) Montrer que:

$$E(X) = \int_0^{\infty} X f(X) dX = ap \text{ et } V(X) = E(X^2) - E^2(X) = a^2 p$$

2) Déterminer l'estimateur $\hat{a}$ du maximum de vraisemblance de a et vérifier qu'il est sans biais et convergent.

1.4.3. Méthodes des moindres carrés

On considère un modèle de régression linéaire $Y = aX + b$ à travers un nuage de points.

L'ajustement est jugé satisfaisant lorsque la distance moyenne entre les points et la droite est la plus réduite possible. Ce qui pose 2 problèmes:

- 1) Comment mesurer la distance entre un point et la droite
- 2) Comment calculer la moyenne des distances relatives aux différents points.

Le deuxième problème trouve sa solution dans l'adoption de la moyenne quadratique. D'où la recherche de la droite telle que la somme des carrés soit minime. C'est le principe de la méthode des moindres carrés.

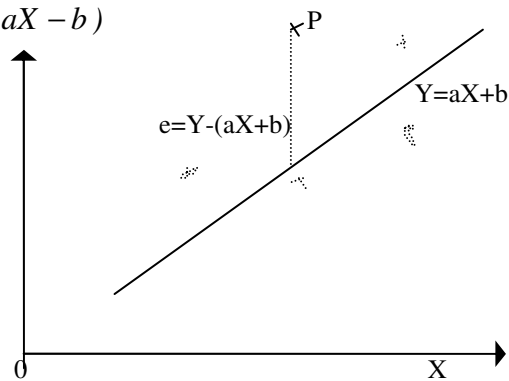
On résout le premier problème de deux manières

- soit par la régression simple
- soit par la régression orthogonale.

1.4.3.1. Régression simple

La méthode des moindres carrés consiste à estimer a et b tels que:

$$\min \sum e^2 = \min \sum (Y - aX - b)$$



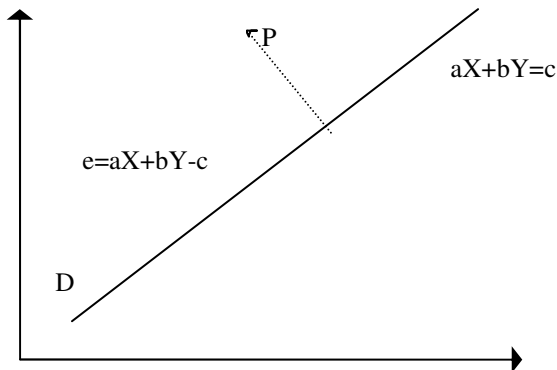
$$\begin{cases} \frac{\partial \sum e^2}{\partial a} = 0 \\ \frac{\partial \sum e^2}{\partial b} = 0 \end{cases} \text{ et telles que les dérivées secondes } \frac{\partial^2 \sum e^2}{\partial a^2} > 0 \text{ et } \frac{\partial^2 \sum e^2}{\partial b^2} > 0$$

$$\text{d'où: } \begin{cases} \hat{a} = \frac{\sum xy}{\sum x^2} \\ \hat{b} = \bar{Y} - \hat{a} \bar{X} \end{cases} \text{ où } \begin{cases} x = X - \bar{X} \\ y = Y - \bar{Y} \end{cases}$$

1.4.3.2. Régression orthogonale

La distance se mesurant perpendiculairement à la droite, elle a pour expression:

$$d(p, D) = \frac{|aX + bY - c|}{\sqrt{a^2 + b^2}} = |e| \quad \text{ssi} \quad a^2 + b^2 = 1$$



$$\min \sum e^2 = \min \sum (aX + bY - c)^2 \text{ sous contrainte } a^2 + b^2 = 1$$

$$\mathcal{L} = \sum (aX + bY - c)^2 - \lambda (a^2 + b^2 - 1), \text{ le Lagrangien}$$

Les conditions du premier ordre conduisent à:

$$c = a\bar{X} + b\bar{Y} \quad \text{d'où} \quad \begin{cases} a\left(\sum x^2 - \lambda\right) + b\sum xy = 0 \\ a\sum xy + b\left(\sum y^2 - \lambda\right) = 0 \end{cases}$$

Ce système n'a de solution non nulle que si son déterminant est nul i.e:

$$\lambda^2 - \left(\sum x^2 + \sum y^2\right)\lambda + \sum x^2 \sum y^2 - \left(\sum xy\right)^2 = 0 \text{ avec } \lambda \text{ inconnu}$$

$\Rightarrow$ la plus petite valeur de λ est:

$$\lambda = \frac{1}{2} \left(\sum x^2 + \sum y^2 - \sqrt{\left(\sum x^2 - \sum y^2\right)^2 + 4\left(\sum xy\right)^2} \right)$$

$$\text{d'où} \quad \frac{a}{b} = \frac{\lambda - \sum y^2}{\sum xy} = \frac{1}{2} \left(\frac{\sum x^2 - \sum y^2 - \sqrt{\left(\sum x^2 - \sum y^2\right)^2 + 4\left(\sum xy\right)^2}}{\sum xy} \right)$$

Connaissant le rapport a/b on connaît la droite d'ajustement:

$$aX + bY = c$$

$$a\bar{X} + b\bar{Y} = c$$

$$ax + by = 0 \Rightarrow y = -\frac{a}{b}x \quad d'où \quad Y = -\frac{a}{b}(X - \bar{X}) + \bar{Y}$$

Ex.3 La droite de régression $DY(X)$ est donnée par l'équation $X \cos \alpha + Y \sin \alpha = \beta$

Montrer que les estimateurs des moindres carrés de α et β sont donnés par:

$$\begin{cases} \hat{\beta} = \bar{X} \cos \alpha + \bar{Y} \sin \alpha \\ \hat{\alpha} = \frac{1}{2} \arctg \frac{2 \sum xy}{\sum x^2 - \sum y^2} \end{cases}$$

Partant de $\min_{\alpha} \beta(X \cos \alpha + Y \sin \alpha) / \alpha = E$ sachant que $\cos \alpha \beta + \sin \alpha \beta = 1$

2. Modèle linéaire simple

Le modèle le plus simple est celui de la relation linéaire entre deux variables X et Y /:

$$(1) \quad Y = \alpha + \beta X + u$$

L'introduction de u dans (1) se justifie pour plusieurs raisons:

- Y serait entièrement expliqué si l'on savait tous les facteurs pouvant l'influencer et si l'on disposait de toutes les données nécessaires. En fait, le nombre de facteurs peut s'étendre à l'infini et beaucoup d'entre eux seront difficilement quantifiables. Puisqu'on choisit de représenter Y comme une fonction explicite d'un petit nombre de facteurs (jugés les plus importants), on considère alors que u condense l'effet de toutes les variables omises dans le modèle
- La seconde justification de l'introduction de u est qu'au-delà des facteurs, il existera toujours un élément imprévisible d'aléa dans les phénomènes sociaux où l'homme est le moteur avec tout le subjectivisme qui le caractérise
- La troisième source d'erreur réside dans l'incertitude des observations/mesures sur Y

- Une quatrième raison serait liée à l'imperfection de la spécification de la fonction f .

En résumé Y et u sont des variables aléatoires, X est supposée non aléatoire et la fonction f est linéaire.

Si nous avons n observations sur X et Y , alors nous avons le modèle

$$Y_i = \alpha + \beta X_i + u_i \quad \text{avec} \quad i = \overline{1, n}$$

Pour alléger l'écriture, nous omettrons volontairement l'indice i en écrivant tout simplement

$$(2) \quad Y = \alpha + \beta X + u$$

Nous cherchons à estimer les paramètres inconnus α et β (coefficients de régression) ainsi que σ_u^2 . Pour ce faire, il existe plusieurs méthodes dont le choix dépend des hypothèses émises sur u . L'économétrie privilégie la méthode des moindres carrés, qui sous certaines hypothèses garantit des propriétés désirables aux estimateurs.

2.1. Hypothèses

H₁. Les variables X et Y sont des grandeurs observées sans erreur, Y est aléatoire

H₂. (*Hypothèse fondamentale*) $E(u_i) = 0 \quad \forall \quad i = \overline{1, n}$

H₃. (*Homoscédasticité*) L'erreur u_i suit une distribution indépendante de i et de X

$$\forall i, \quad V(u_i) = \sigma_u^2$$

H₄. (*Non corrélation des erreurs*) $\forall i, j \quad \text{avec} \quad i \neq j \quad E(u_i u_j) = 0$

H₅. (*Indépendance*) La loi de probabilité des erreurs est indépendante de X

H₆. Moyenne et variance finies

$$\lim_{n \rightarrow \infty} \bar{X} = X_0 \quad \text{finie}$$

$$\lim_{n \rightarrow \infty} \sigma_X^2 = S^2 > 0, \quad \text{finie}$$

H₇. Les erreurs suivent une loi normale ou de LAPLACE-GAUSS: $u \mapsto N(0, \sigma_u^2)$

Cette hypothèse qui n'est pas nécessaire pour l'estimation des paramètres a son importance pour justifier certains tests associés à la méthode des moindres carrés.

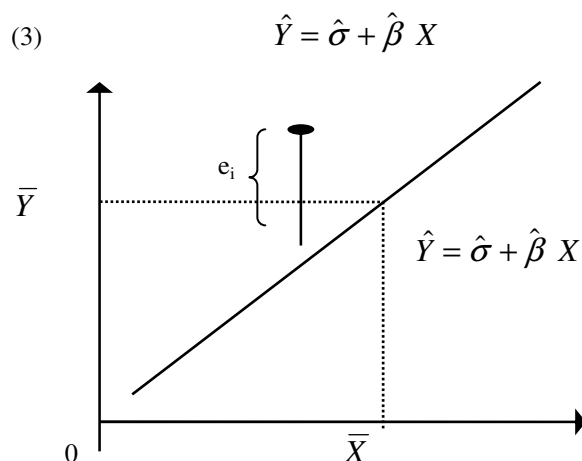
2.2. Estimation

Soient les observations

$$X_1, X_2, \dots, X_n \text{ avec } \bar{X} = \frac{1}{n} \sum X_i$$

$$Y_1, Y_2, \dots, Y_n \text{ avec } \bar{Y} = \frac{1}{n} \sum Y_i$$

Soit la droite de régression:



Les résidus (e_i) seront soit positifs soit négatifs selon qu'ils sont au-dessus ou au dessous de la droite de régression. Si on les élève au carré et qu'on les additionne, le résultat sera non négatif et devra varier directement avec le

nuage des points, à partir de la droite. Ce résultat est bien sûr fonction des différentes valeurs des estimateurs $\hat{\alpha}$ et $\hat{\beta}$

$$\sum e_i^2 = f(\hat{\alpha}, \hat{\beta})$$

Le principe de la méthode des moindres carrés est que $\hat{\alpha}$ et $\hat{\beta}$ doivent être choisis tel que $\sum e_i^2$ soit le plus petit possible.

Soit à minimiser: $\sum e^2 = \sum (Y - \hat{Y})^2 = \sum (Y - \hat{\alpha} - \hat{\beta} X)^2 = \text{RSS}$ (pour Residual Sum of Squares)

$$\begin{cases} \frac{\partial \sum e^2}{\partial \hat{\alpha}} = 0 \\ \frac{\partial \sum e^2}{\partial \hat{\beta}} = 0 \end{cases} \Leftrightarrow \begin{cases} -2 \sum (X - \hat{\alpha} - \hat{\beta} X) = 0 \\ -2 \sum X (Y - \hat{\alpha} - \hat{\beta} X) = 0 \end{cases}$$

Ce qui conduit aux équations normales $\begin{cases} \sum Y = n\hat{\alpha} + \hat{\beta} \sum X \\ \sum XY = \hat{\alpha} \sum X + \hat{\beta} \sum X^2 \end{cases}$

En divisant les 2 termes de la première équation par n, on obtient: $\bar{Y} = \hat{\alpha} + \hat{\beta} \bar{X}$

ce qui montre que la droite passe bien par le point $(\bar{X}, \bar{Y})$

$$\begin{aligned} \hat{Y} &= \hat{\alpha} + \hat{\beta} X \\ \bar{Y} &= \hat{\alpha} + \hat{\beta} \bar{X} \\ \hat{Y} - \bar{Y} &= \hat{\beta} (X - \bar{X}) \end{aligned}$$

Soit:

$$(3') \quad \hat{y} = \hat{\beta} x \quad \text{où} \quad \begin{aligned} X - \bar{X} &= x \\ Y - \bar{Y} &= y \\ \hat{Y} - \bar{Y} &= \hat{y} \end{aligned}$$

Avec ces changements de variables, le résidu s'écrit :

$$e = Y - \hat{Y} = Y - \bar{Y} - (\hat{Y} - \bar{Y}) = y - \hat{y} = y - \hat{\beta} x$$

d'où
$$\sum e^2 = \sum (y - \hat{\beta}x)^2$$

$$\frac{\partial \sum e^2}{\partial \hat{\beta}} = 0 \Leftrightarrow -2 \sum (y - \hat{\beta}x)x = 0 \Leftrightarrow \sum xy = \hat{\beta} \sum x^2$$

(4)
$$\hat{\beta} = \frac{\sum xy}{\sum x^2} = \frac{S_{xy}}{S_{xx}}$$

On vérifie aisément la condition du second ordre :

$$\frac{\partial^2 \sum e^2}{\partial \hat{\beta}^2} = 2 \sum x^2 > 0 \Rightarrow \text{minimum}$$

Connaissant $\hat{\beta}$, on détermine aisément $\hat{\alpha}$, sachant que:

(5)
$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} \quad \text{partant de} \quad \bar{Y} = \hat{\alpha} + \hat{\beta}\bar{X}$$

Ex. 4. Sur le prix (X) et la consommation (Y) de sucre, on dispose de cinq observations réparties entre deux échantillons tels que:

Echantillon	n	$\sum X$	$\sum Y$	$\sum x^2$	$\sum y^2$	$\sum xy$
I	2	4	14	2	2	-2
II	3	21	6	8	2	-2

avec $x = X - \bar{X}$ et $y = Y - \bar{Y}$

Pour l'échantillon global de cinq observations:

1. estimer, par la méthode des moindres carrés, les paramètres du modèle: $DY(X): Y = aX + b + u$ où $u \rightarrow N(0, \sigma_u^2)$
2. montrer que la droite de régression $DX(Y)$ est telle que: $x + y = 0$
3. tracer dans un même repère ayant pour origine $(\bar{X}, \bar{Y})$ les droites $DY(X)$ et $DX(Y)$ et trouver l'angle θ qu'elles forment entre elles

2.3. Propriétés

Pour établir les propriétés des estimateurs, on suppose les X fixés constants en prenant des échantillons répétés de taille n. D'un échantillon à l'autre les

valeurs de Y varieront comme conséquences des variations de u . En appliquant les formules de calcul des paramètres $\hat{\alpha}$ et $\hat{\beta}$ à chaque échantillon, les observations généreront une série de valeurs $\hat{\alpha}$ et $\hat{\beta}$. La distribution de ces valeurs est celle de l'échantillonnage correspondant aux estimateurs des moindres carrés. Ce qui nous intéresse ici ce sont les moyennes et variances de ces distributions ainsi que la comparaison entre ces distributions et celles d'autres estimateurs possibles.

P₁. Estimateurs linéaires

Les estimateurs sont fonctions linéaires des observations Y

De

$$(4) \quad \hat{\beta} = \frac{\sum xy}{\sum x^2} \text{ on peut écrire:}$$

$$\hat{\beta} = \frac{\sum x(Y - \bar{Y})}{\sum x^2} = \frac{\sum xY}{\sum x^2} - \underbrace{\frac{\bar{Y} \sum x}{\sum x^2}}_0 = \frac{\sum xY}{\sum x^2}$$

$$(6) \quad \hat{\beta} = \sum wY \quad \text{où} \quad w = \frac{x}{\sum x^2}$$

Ces w sont fixés constants et satisfont :

$$1) \sum w = 0$$

$$2) \sum w^2 = \sum \frac{x^2}{(\sum x^2)^2} = \frac{\sum x^2}{(\sum x^2)^2} = \frac{1}{\sum x^2}$$

$$3) \sum wx = \sum x \frac{x}{\sum x^2} = \frac{\sum x^2}{\sum x^2} = 1 \text{ et}$$

$$\sum wx = \sum w(X - \bar{X}) = \underbrace{\sum wX}_1 + \underbrace{\bar{X} \sum w}_0 = \sum wX = 1$$

De (5) $\hat{\alpha} = \bar{Y} - \hat{\beta} \bar{X} = \frac{1}{n} \sum Y - \bar{X} \sum wY$ soit:

$$(7) \quad \hat{\alpha} = \sum \left(\frac{1}{n} - \bar{X} w \right) Y$$

P₂. Estimateurs sans biais

Calculons les moyennes ou espérances de $\hat{\beta}$ respectivement $\hat{\alpha}$

D'après (6) :

$$\begin{aligned} \hat{\beta} &= \sum wY \\ &= \sum w(\alpha + \beta X + u) = \underbrace{\alpha \sum w}_0 + \beta \underbrace{\sum wX}_1 + \sum wu \end{aligned}$$

$$(8) \quad \hat{\beta} = \beta + \sum wu$$

$$\begin{aligned} \Rightarrow E(\hat{\beta}) &= E(\beta + \sum wu) \\ &= E(\beta) + E(\sum wu) \\ &= \beta + \sum w \underbrace{E(u)}_0 \end{aligned}$$

$E(\hat{\beta}) = \beta$; $\hat{\beta}$ est un estimateur sans biais de β

$$\begin{aligned} \text{De (7) : } \hat{\alpha} &= \sum \left(\frac{1}{n} - \bar{X} w \right) Y = \sum \left(\frac{1}{n} - \bar{X} w \right) (\alpha + \beta X + u) \\ &= \sum \left(\frac{1}{n} \alpha + \frac{1}{n} \beta X + \frac{1}{n} u - \alpha \bar{X} w - \beta \bar{X} wX - \bar{X} wu \right) \\ &= \alpha + \beta \bar{X} + \frac{1}{n} \sum u - \alpha \bar{X} \underbrace{\sum w}_0 - \beta \bar{X} \underbrace{\sum wX}_1 - \bar{X} \sum wu \end{aligned}$$

$$= \alpha + \beta \bar{X} - \beta \bar{X} + \frac{1}{n} \sum u - \bar{X} \sum w u$$

$$(9) \quad \hat{\alpha} = \alpha + \sum \left(\frac{1}{n} - \bar{X} w \right) u$$

$$\begin{aligned} \Rightarrow E(\hat{\alpha}) &= E\left(\alpha + \sum \left(\frac{1}{n} - \bar{X} w \right) u\right) \\ &= E(\alpha) + \sum \left(\frac{1}{n} - \bar{X} w \right) \underbrace{E(u)}_0 \end{aligned}$$

$E(\hat{\alpha}) = \alpha$; $\hat{\alpha}$ est un estimateur sans biais de α

P₃. Estimateurs efficaces

Calculons les variances des estimateurs $\hat{\beta}$ respectivement $\hat{\alpha}$

$$\text{De (8) : } \hat{\beta} = \beta + \sum w u \Rightarrow \hat{\beta} - \beta = \sum w u \quad \text{où } \beta = E(\hat{\beta})$$

$$\begin{aligned} V(\hat{\beta}) &= E(\hat{\beta} - E(\hat{\beta}))^2 \\ &= E(\hat{\beta} - \beta)^2 \\ &= E(\sum w u)^2 \\ &= E(w_1 u_1 + w_2 u_2 + \dots + w_n u_n)^2 \\ &= E(w_1^2 u_1^2 + \dots + w_n^2 u_n^2 + 2w_1 w_2 u_1 u_2 + \dots + 2w_{n-1} w_n u_{n-1} u_n) \\ &= E(w_1^2 u_1^2 + \dots + w_n^2 u_n^2) + 2E(w_1 w_2 u_1 u_2 + \dots + w_{n-1} w_n u_{n-1} u_n) \\ &= \sum w_i^2 \underbrace{E(u_i^2)}_{\sigma_u^2} + 2 \sum w_i w_{i-1} \underbrace{E(u_i u_{i-1})}_0 \\ &= \sigma_u^2 \sum w^2 = \frac{\sigma_u^2}{\sum x^2} \end{aligned}$$

$$(10) \quad V(\hat{\beta}) = \frac{\sigma_u^2}{\sum x^2}$$

$$\hat{\alpha} = \alpha + \sum \left(\frac{1}{n} - \bar{X}w \right) u$$

De (9) on a

$$\hat{\alpha} - \alpha = \sum \left(\frac{1}{n} - \bar{X}w \right) u$$

$$\begin{aligned} V(\hat{\alpha}) &= E(\hat{\alpha} - \alpha)^2 = E \left(\sum \left(\frac{1}{n} - \bar{X}w \right) u \right)^2 \\ &= E \left(\sum \left(\frac{1}{n} - \bar{X}w \right)^2 u^2 \right) = \sum \left(\frac{1}{n} - \bar{X}w \right)^2 E(u)^2 \end{aligned}$$

$$\begin{aligned} V(\hat{\alpha}) &= \sigma_u^2 \sum \left(\frac{1}{n} - \bar{X}w \right)^2 = \sigma_u^2 \sum \left(\frac{1}{n^2} + \bar{X}^2 w^2 - \frac{2}{n} \bar{X}w \right) \\ &= \sigma_u^2 \left(\frac{1}{n} + \bar{X}^2 \underbrace{\sum w^2}_{\frac{1}{\sum x^2}} - \frac{2\bar{X}}{n} \underbrace{\sum w}_0 \right) \end{aligned}$$

$$(11) \quad V(\hat{\alpha}) = \sigma_u^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum x^2} \right) = \sigma_u^2 \left(\frac{\sum x^2 + n\bar{X}^2}{n\sum x^2} \right) = \sigma_u^2 \frac{\sum X^2}{n\sum x^2}$$

De même, on peut calculer la covariance entre les estimateurs:

$$\text{cov}(\hat{\alpha}, \hat{\beta}) = E(\hat{\alpha} - \alpha)(\hat{\beta} - \beta)$$

$$\begin{aligned}
&= E\left(\sum \left(\frac{1}{n} - \bar{X}w\right)u \sum wu\right) = E\left(\left(\frac{1}{n} \sum u - \bar{X} \sum wu\right) \sum wu\right) \\
&= E\left(\frac{1}{n} \sum u \cdot \sum wu - \bar{X} (\sum wu)^2\right) \\
&= \frac{1}{n} E[(u_1 + u_2 + \dots + u_n) \cdot (w_1 u_1 + \dots + w_n u_n)] - \bar{X} E(\sum wu)^2 \\
&= \frac{1}{n} \sigma_u^2 \underbrace{\sum w}_0 - \bar{X} V(\hat{\beta}) = -\bar{X} V(\hat{\beta})
\end{aligned}$$

$$(12) \quad \text{cov}(\hat{\alpha}, \hat{\beta}) = -\sigma_u^2 \frac{\bar{X}}{\sum x^2}$$

Il faut ensuite établir que les variances calculées (d'après 10 et 11) sont les plus faibles possibles, donc que $\hat{\alpha}, \hat{\beta}$ sont les meilleurs estimateurs linéaires sans biais.

Soit un estimateur linéaire arbitraire: $\tilde{\beta} = \sum cY$ avec $c = w + d$ où d constante arbitraire

Pour que $\tilde{\beta}$ soit un estimateur sans biais de β , d doit satisfaire certaines conditions :

$$\tilde{\beta} = \sum cY = \sum c(\alpha + \beta X + u) = \alpha \sum c + \beta \sum cX + \sum cu$$

$$E(\tilde{\beta}) = \alpha \sum c + \beta \sum cX = \beta \quad \text{ssi}$$

$$1) \sum c = 0 \quad \text{et} \quad 2) \sum cX = 1$$

Ces 2 restrictions plus les propriétés de w et en tenant compte de la relation $c = w + d$, donnent les conditions que doit satisfaire d , à savoir :

$$1) \quad \sum c = 0 \Leftrightarrow \sum (w + d) = \underbrace{\sum w}_0 + \sum d \Rightarrow \sum d = 0$$

$$\begin{aligned}
2) \quad \sum cX = 1 &\Leftrightarrow \sum (w+d)X = \sum wX + \sum dX = \sum w(x + \bar{X}) + \sum dX \\
&= \sum wx + \bar{X} \underbrace{\sum w}_0 + \sum dX = 1
\end{aligned}$$

$$\Rightarrow \sum dX = \sum d(x + \bar{X}) = \sum dx + \bar{X} \underbrace{\sum d}_0 = \sum dx = 0$$

Sous toutes ces conditions, la variance de $\tilde{\beta}$ devient:

$$\begin{aligned}
V(\tilde{\beta}) &= E(\tilde{\beta} - \beta)^2 = E(\sum cu)^2 = \sigma_u^2 \sum c^2 \\
&= \sigma_u^2 \sum (w+d)^2 = \sigma_u^2 \sum (w^2 + d^2 + 2wd) \\
&= \sigma_u^2 \left(\frac{1}{\sum x^2} + \sum d^2 + 2 \frac{\sum xd}{\sum x^2} \right) = \sigma_u^2 \left(\frac{1}{\sum x^2} + \sum d^2 \right)
\end{aligned}$$

$V(\tilde{\beta}) = V(\hat{\beta}) + \sigma_u^2 \sum d^2$ où $\sum d^2$ est nécessairement une quantité non négative, par conséquent

$V(\tilde{\beta}) \geq V(\hat{\beta})$ et $\sum d^2 = 0$ uniquement si tous les d_i sont nuls.

On obtient un résultat similaire pour $V(\tilde{\alpha})$. Il suffit de poser :

$$\tilde{\alpha} = \sum cY \quad \text{avec} \quad c = \frac{1}{n} - \omega \bar{Y} + d \quad \text{avec} \quad \sum d = 0 \quad \text{et} \quad \sum dX = \sum dx = 0$$

$$\begin{aligned}
V(\tilde{\alpha}) &= \sigma_u^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum x^2} + \sum d^2 - 2\bar{X} \frac{\sum dx}{\sum x^2} \right) \\
&= \sigma_u^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum x^2} \right) + \sigma_u^2 \sum d^2 = V(\hat{\alpha}) + \sigma_u^2 \sum d^2
\end{aligned}$$

d'où $V(\tilde{\alpha}) \geq V(\hat{\alpha})$

Les variances de $\hat{\beta}$ et de $\hat{\alpha}$: $V(\hat{\beta}) = \frac{\sigma_u^2}{\sum x^2}$ respectivement $V(\hat{\alpha}) = \frac{\sigma_u^2 \sum X^2}{n \sum x^2}$
 dépendent de la variance des résidus (σ_u^2) qu'il faut au préalable déterminer:

$$\begin{aligned}\sum e^2 &= RSS = \sum (Y - \hat{\alpha} - \hat{\beta} \bar{X})^2 \\ &= \sum (Y - \bar{Y} + \hat{\beta} (\bar{X} - X))^2 = \sum (y - \hat{\beta} x)^2\end{aligned}$$

En remplaçant $\hat{\beta}$ par sa valeur, $\hat{\beta} = \frac{\sum xy}{\sum x^2}$ on obtient:

$$(13) \quad RSS = \sum y^2 - \frac{(\sum xy)^2}{\sum x^2} = n \left(S_y^2 - \frac{S_{xy}^2}{S_x^2} \right)$$

L'estimation de la variance de la variable erreur, σ_u^2 se fait sur la base de RSS.

$Y = \alpha + \beta X + u$, il s'en suit :
 $\bar{Y} = \alpha + \beta \bar{X} + \bar{u}$
 Soit: $Y - \bar{Y} = \beta (X - \bar{X}) + (u - \bar{u})$
 $y = \beta x + (u - \bar{u})$
 Or: $RSS = \sum (y - \hat{\beta} x)^2 = \sum e^2$ avec $y = \beta x + (u - \bar{u})$, alors:

$$\begin{aligned}RSS &= \sum (\beta x + (u - \bar{u}) - \hat{\beta} x)^2 \\ &= \sum (-(\hat{\beta} - \beta)x + (u - \bar{u}))^2 \\ &= \sum [(\hat{\beta} - \beta)^2 x^2 + (u - \bar{u})^2 - 2(\hat{\beta} - \beta)x(u - \bar{u})] \\ &= (\hat{\beta} - \beta)^2 \sum x^2 + \sum (u - \bar{u})^2 - 2(\hat{\beta} - \beta) \sum x(u - \bar{u})\end{aligned}$$

Pour obtenir la variance de l'erreur, on calcule l'espérance mathématique de $\sum e^2$.

$$\begin{aligned}
E(\sum e^2) &= E[(\hat{\beta} - \beta)^2 \sum x^2] + E\left[\sum (u - \bar{u})^2\right] - 2E[(\hat{\beta} - \beta) \sum x(u - \bar{u})] \\
&= \sum x^2 V(\hat{\beta}) + E\left(\sum \left(u^2 - \frac{(\sum u)^2}{n}\right)\right) - 2E\left(\sum \omega u \left(\sum ux - \underbrace{\bar{u} \sum x}_0\right)\right) \\
&= \sum x^2 \frac{\sigma_u^2}{\sum x^2} + E(\sum u^2) - \frac{E(\sum u^2)}{n} - 2E(\sum \omega u \sum ux) \\
&= \sigma_u^2 + n\sigma_u^2 - \sigma_u^2 - 2 \frac{1}{\sum x^2} (\sum x^2) \sigma_u^2 = (n-2)\sigma_u^2
\end{aligned}$$

Finalement donc, on aura en tout. $E(\sum e^2) = (n-2)\sigma_u^2$

d'où l'estimateur sans biais de σ_u^2 :

$$(14) \quad \hat{\sigma}_u^2 = \frac{\sum e^2}{n-2}$$

On peut estimer les paramètres α et β par la méthode du maximum de vraisemblance et comparer ces estimateurs à ceux des moindres carrés ordinaires.

2.4. Inférence statistique

Pour obtenir les LSE, nous n'avons à émettre aucune hypothèse particulière sur la distribution de probabilité des résidus. Les estimateurs obtenus ont des propriétés optimales sous les hypothèses H_1 , H_2 , H_3 et H_4 retenues.

Par contre si l'on veut obtenir des intervalles de confiance et tester certaines hypothèses sur les estimateurs, nous devons supposer que u a une distribution normale (H_7).

Si $u \mapsto N(0, \sigma_u^2)$ il s'ensuit:

$$1) \hat{\alpha} \mapsto N\left(\alpha, \frac{\sigma_u^2 \sum X^2}{n \sum x^2}\right)$$

$$2) \hat{\beta} \mapsto N\left(\beta, \frac{\sigma_u^2}{\sum x^2}\right)$$

$$3) \frac{RSS}{\sigma_u^2} \mapsto \chi^2(n-2)$$

alors $\hat{\sigma}_u^2 = \frac{RSS}{n-2}$ est un estimateur sans biais de σ_u^2

4) En remplaçant σ_u^2 par $\hat{\sigma}_u^2$ dans $V(\hat{\alpha})$ et $V(\hat{\beta})$ on obtient les variances estimées dont les racines carrées nous donnent les erreurs types, notées SE (pour Standard Error).

Alors $\frac{\hat{\alpha} - \alpha}{SE(\hat{\alpha})}$ et $\frac{\hat{\beta} - \beta}{SE(\hat{\beta})}$ ont chacune une t-distribution à (n-2) ddl

Ces distributions sont utilisées pour faire les intervalles de confiance de α , β , σ_u^2 , les 3 paramètres inconnus de notre modèle linéaire.

Nous avons donc :

$$i) \text{ pour } \hat{\alpha}: \quad \hat{\alpha} \mapsto N\left(\alpha, \frac{\hat{\sigma}_u^2 \sum X^2}{n \sum x^2}\right) \Rightarrow \frac{\hat{\alpha} - \alpha}{SE(\hat{\alpha})} \mapsto t(n-2)$$

$$(15) \quad t(\hat{\alpha}) = \frac{\hat{\alpha} - \alpha}{SE(\hat{\alpha})} = (\hat{\alpha} - \alpha) \frac{\sqrt{n \sum x^2}}{\hat{\sigma}_u \sqrt{\sum X^2}}$$

Pour tester l'hypothèse $\alpha = \alpha_0$, on remplace α par α_0 dans (15) et on voit si t qui en résulte tombe à l'intérieur d'une région critique appropriée, région déterminée par la t -distribution à $(n-2)$ ddl:

Au seuil ε ou pour obtenir un intervalle de confiance $(1-\varepsilon)\%$, on pose:

$$P\left(-t_{\frac{\varepsilon}{2}} < \frac{\hat{\alpha} - \alpha}{SE(\hat{\alpha})} < t_{\frac{\varepsilon}{2}}\right) = 1 - \varepsilon$$

d'où les limites:

$$(16) \quad \alpha = \hat{\alpha} \pm t_{\frac{\varepsilon}{2}} SE(\hat{\alpha}) = \hat{\alpha} \pm t_{\frac{\varepsilon}{2}} \frac{\hat{\sigma}_u \sqrt{\sum X^2}}{\sqrt{n \sum x^2}}$$

$$\text{ii) analogue pour } \hat{\beta}: \quad \hat{\beta} \mapsto N\left(\beta, \frac{\hat{\sigma}_u^2}{\sum x^2}\right) \Rightarrow \frac{\hat{\beta} - \beta}{SE(\hat{\beta})} \mapsto t(n-2)$$

$$(17) \quad t(\hat{\beta}) = \frac{\hat{\beta} - \beta}{SE(\hat{\beta})} = (\hat{\beta} - \beta) \frac{\sqrt{\sum x^2}}{\hat{\sigma}_u}$$

$$(18) \quad \beta = \hat{\beta} \pm t_{\frac{\varepsilon}{2}} SE(\hat{\beta}) = \hat{\beta} \pm t_{\frac{\varepsilon}{2}} \frac{\hat{\sigma}_u}{\sqrt{\sum x^2}}$$

3) les tests sur σ_u^2 sont obtenus en utilisant la χ^2 distribution :

$$\frac{\sum e^2}{\sigma_u^2} = \frac{(n-2)\hat{\sigma}_u^2}{\sigma_u^2} \mapsto \chi^2(n-2)$$

$$(19) \quad \chi^2 = \frac{(n-2)\hat{\sigma}_u^2}{\sigma_u^2}$$

$$P\left(\chi_{\frac{\varepsilon}{2}}^2 < \frac{(n-2)\hat{\sigma}_u^2}{\sigma_u^2} < \chi_{1-\frac{\varepsilon}{2}}^2\right) = 1 - \varepsilon$$

d'où:

$$(20) \quad \frac{(n-2)\hat{\sigma}_u^2}{\chi_{1-\varepsilon/2}^2} < \sigma_u^2 < \frac{(n-2)\hat{\sigma}_u^2}{\chi_{\varepsilon/2}^2}$$

Note: Si nous avons n observations et estimons 2 paramètres α et β , nous voyons à partir des équations normales que les résidus satisfont 2 restrictions linéaires :

$$(S) \quad \begin{cases} \sum \hat{u} = 0 = \sum e \\ \sum X\hat{u} = 0 = \sum Xe \end{cases}$$

en essence, il y a seulement $(n-2)$ résidus libres de varier, car étant données leurs valeurs, les 2 $(= n - (n-2))$ autres peuvent s'obtenir en résolvant (S). Nous dirons donc qu'il y a $n-2$ valeurs arbitraires possibles ou encore $n-2$ degrés de liberté (ddl).

Ex.5. On dispose d'une série de 10 observations sur l'offre (X) et le prix (Y) d'un bien:

X	10	7	10	4	8	8	6	7	9	11
Y	12	11	14	6	10	7	9	11	10	10

1) Par la méthode des moindres carrés, estimer les paramètres du modèle linéaire $DY(X)$

2) Au seuil $\varepsilon = 5\%$, vérifier sur les tables appropriées que:

$$t_{\varepsilon/2}(n-2) = 2.306, \quad \chi_{\varepsilon/2}^2(n-2) = 2.18 \quad \text{et} \quad \chi_{1-\varepsilon/2}^2(n-2) = 17.53$$

3) Montrer que:

$$P(-1.93 < \alpha < 10.73) = 0.95$$

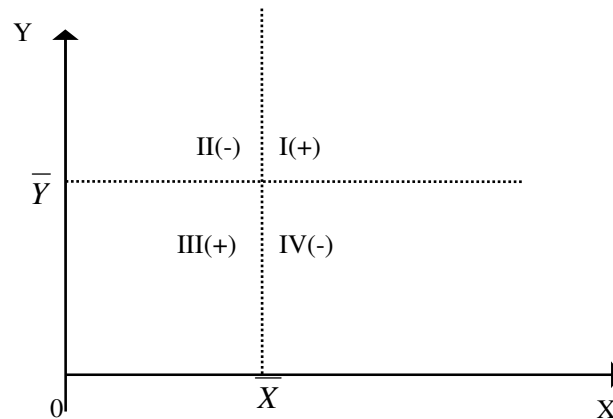
$$P(0.013 < \beta < 1.387) = 0.95$$

$$P(1.62 < \sigma_u^2 < 13.03) = 0.95$$

2.5. Coefficient de corrélation

Le coefficient de corrélation mesure en quelque sorte l'intensité de la liaison entre deux variables.

Soit le schéma suivant, représentant les nuages de points entre les variables X et Y :



Partant de $\bar{X}$ et $\bar{Y}$, le graphique se divise en 4 cadrans

Le terme $\sum xy$ peut servir à mesurer l'association entre X et Y. S'il :

- est positif, alors la plupart des points se trouvent en I et III
- est négatif, alors la plupart des points se trouvent en II et IV
- tend vers zéro, les points sont équitablement dispersés à travers tous les cadrans; il n'existe alors pas de liaison entre X et Y.

Mais cette mesure qui nous renseigne sur la liaison entre les 2 variables X et Y présente deux inconvénients:

- 1) sa valeur peut être arbitrairement élevée si l'on ajoute davantage d'observations
- 2) elle est fonction des unités de mesure choisies pour X et Y.

On y remédie en exprimant les déviations, x et y, en unités standard et en prenant la moyenne pour obtenir le coefficient de corrélation (encore appelé **Pearsonien**).

(21)

$$r = \frac{\sum \frac{x}{\sigma_x} \frac{y}{\sigma_y}}{n} = \frac{\sum xy}{n \sigma_x \sigma_y} = \frac{\sum xy}{\sqrt{\sum n \sigma_x^2 \sum n \sigma_y^2}}$$

$$= \frac{\sum xy}{\sqrt{\sum x^2} \sqrt{\sum y^2}} = \frac{\sum XY - n\bar{X}\bar{Y}}{\sqrt{\sum X^2 - n\bar{X}^2} \sqrt{\sum Y^2 - n\bar{Y}^2}}$$

Si nous tenons compte de l'ajustement $\hat{Y} = \hat{\alpha} + \hat{\beta} X$, nous obtenons des relations couramment utilisées en économétrie telles que:

A partir de la définition de la droite des moindres carrés, on a:

$$e = Y - \hat{Y} = \underbrace{Y - \bar{Y}}_{\hat{y}} - \underbrace{(\hat{Y} - \bar{Y})}_{\hat{y}} \Rightarrow y = \hat{y} + e$$

alors:

$$\begin{aligned} \sum y^2 &= \sum (\hat{y} + e)^2 = \sum (\hat{y}^2 + 2e\hat{y} + e^2) \\ &= \sum \hat{y}^2 + 2\sum e\hat{y} + \sum e^2 \\ \sum y^2 &= \sum \hat{y}^2 + 2\hat{\beta}\sum ex + \sum e^2 \\ &= \sum \hat{y}^2 + 2\hat{\beta}\sum x(y - \hat{\beta}x) + \sum e^2 \\ &= \sum \hat{y}^2 + 2\underbrace{\frac{\sum xy}{\sum x^2} \sum xy - 2\frac{(\sum xy)^2}{(\sum x^2)^2} \sum x^2}_0 + \sum e^2 \end{aligned}$$

$$(22) \quad \sum y^2 = \sum \hat{y}^2 + \sum e^2 = Q_1 + Q_2$$

La relation (22), sur laquelle se fonde l'analyse de variance, signifie que la variation (ou variabilité) totale des valeurs de Y, sur leur moyenne, est une mesure de 2 variations:

- 1) la variabilité expliquée par (ou due à) l'influence linéaire de X
- 2) la variabilité résiduelle (ou inexpliquée) par rapport à la droite des moindres carrés.

A l'aide de cette même relation (22), on peut déterminer le poids spécifique de la variabilité expliquée dans la variabilité totale de Y:

$$\frac{\sum \hat{y}^2}{\sum y^2} = \frac{\sum (\hat{\beta} x)^2}{\sum y^2} = \hat{\beta}^2 \frac{\sum x^2}{\sum y^2} = r^2 \frac{\sigma_y^2}{\sigma_x^2} \cdot \frac{\sigma_x^2}{\sigma_y^2} = r^2$$

$$(23) \quad r^2 = \frac{\sum \hat{y}^2}{\sum y^2}$$

Ainsi le carré du coefficient de corrélation (encore appelé coefficient de détermination - pour la raison de l'égalité (23)) est égal à la proportion de la variabilité de Y expliquée par l'influence linéaire de X.

En combinant (22) et (23) on écrira :

$$\begin{aligned} \sum y^2 &= \sum \hat{y}^2 + \sum e^2 \\ \Rightarrow 1 &= \frac{\sum \hat{y}^2}{\sum y^2} + \frac{\sum e^2}{\sum y^2} = r^2 + \frac{\sum e^2}{\sum y^2} \\ (23') \quad r^2 &= 1 - \frac{\sum e^2}{\sum y^2} \end{aligned}$$

d'où:

1) le maximum de r^2 est égal à 1 ($r = \pm 1$) $\sum e^2 = 0 \Rightarrow$ chaque e_i est nul, cela voudrait dire que tous les points du graphique sont situés sur la droite d'ajustement

2) le minimum de r^2 est égal à zéro $\Leftrightarrow \sum e^2 = \sum y^2$

$\sum \hat{y}^2 = 0 \Leftrightarrow \sum (\hat{Y} - \bar{Y})^2 = 0$ ssi $\hat{Y} = \bar{Y}$ i.e. la droite de régression est $\hat{Y} = \bar{Y}$

Notes: Pour des variables ordinales, ne suivant pas une loi normale, on fait recours aux coefficients des rangs puisqu'on n'est plus intéressé par les valeurs des variables mais le rang de l'une par rapport à celui de l'autre (ordonné lui du premier au dernier)

Soit la classification suivante de n éléments en X et en Y.

	E ₁	E ₂	...	E _i	...	E _n
X	1	2	...	i		n
Y	1 _p	2 _p	...	i _p	...	n _p

$$\bar{X} = \frac{\sum i}{n} = \frac{n(n+1)}{2n} = \frac{n+1}{2},$$

Les i_p n'étant que des permutations des i: $\bar{Y} = \frac{\sum i_p}{n} = \frac{\sum i}{n} = \frac{n+1}{2}$

Nous avons donc: $\bar{X} = \bar{Y} = \frac{n+1}{2}$

$$\sigma_X^2 = \frac{\sum (X - \bar{X})^2}{n} = \frac{\sum X^2}{n} - \bar{X}^2 = \frac{\sum i^2}{n} - \bar{X}^2$$

$$\sum i^2 = 1^2 + 2^2 + \dots + n^2 = \frac{n(n+1)(2n+1)}{6}$$

$$\sigma_X^2 = \frac{(n+1)(2n+1)}{6} - \frac{(n+1)^2}{4} = \frac{n+1}{12} (2(2n+1) - 3(n+1)) = \frac{n^2 - 1}{12}$$

$$\sigma_X = \sqrt{\frac{n^2 - 1}{12}} = \sigma_Y$$

$$r = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{n \sigma_X \sigma_Y} = \frac{\sum xy}{n \sigma_X \sigma_Y}$$

De d = X - Y et $\bar{X} = \bar{Y}$, il vient d = x - y

$$\begin{aligned} \sum d^2 &= \sum (x - y)^2 = \sum x^2 + \sum y^2 - 2 \sum xy \\ \Rightarrow \sum xy &= \frac{1}{2} (\sum x^2 + \sum y^2 - \sum d^2) \end{aligned}$$

$$\sum x^2 = \sum y^2 = n \sigma_x^2 = \frac{n(n^2 - 1)}{12}$$

$$\sum xy = \frac{1}{2} \left(2 \frac{n(n^2 - 1)}{12} - \sum d^2 \right) = \frac{n(n^2 - 1)}{12} - \frac{\sum d^2}{2}$$

$$\text{d'où: } r = \frac{\frac{n(n^2 - 1)}{12} - \frac{\sum d^2}{2}}{\frac{n(n^2 - 1)}{12}} = \frac{n(n^2 - 1) - 6 \sum d^2}{n(n^2 - 1)}$$

$$r = 1 - \frac{6 \sum d^2}{n(n^2 - 1)} = \text{Sr coefficient de corrélation des}$$

rangs de Spearman

Comme pour le coefficient des rangs de Spearman, pour Kendall, la première variable est triée et les rangs de la seconde sont mis en regard de ceux de la première. En ne s'intéressant qu'à la seconde variable, pour chacune de ses observations, il y aura des rangs supérieurs (p) et des rangs inférieurs (q). On fait le solde de ces rangs supérieurs et inférieurs, par soustraction (p-q). Par exemple, dans une population de 10 éléments, il y a 9 rangs supérieurs au premier et 0 rangs inférieurs, le solde est égal 9.

Si la seconde variable est également ordonnée, en faisant la somme de tous les soldes, on aura au maximum un solde total de:

$$S_{\max} = (n-1) + (n-2) \cdots + 2 + 1 = 1 + 2 + \cdots + (n-1) = \frac{(n-1)n}{2} \text{ soit la somme des } (n-1) \text{ premiers naturels}$$

$$\text{Si l'ordre est parfaitement inversé, le solde total sera: } S_{\max} = -\frac{n(n-1)}{2}$$

Ainsi, le coefficient de Kendall est défini comme étant le rapport de la somme des soldes observés (S) sur le solde maximum tel que défini ci-dessus ($S_{\max}$):

$$K_r = \frac{S}{S_{\max}} = \frac{S}{\frac{n(n-1)}{2}} = \frac{2S}{n(n-1)}$$

Le coefficient de Kendall traduit beaucoup plus un indice de concordance qu'une corrélation, ce qui nous renseigne quand même sur la liaison entre 2 phénomènes ; on pourrait l'interpréter en terme de : "il y a $K_r\%$ de chance qu'un élément $p^{\text{ème}}$ en X le soit aussi en Y e.g. le 1^{er} en X a $K_r\%$ de chance de l'être également en Y". On peut développer un petit programme Excel pour le calcul de K_r , en utilisant la fonction SI(NB.SI()) e.g.

$$=SI(NB.SI(B\$2:B2;">1")=0;MAX(\$B\$2:\$B\$11)-B3;MAX(\$A\$2:\$A\$11)-B3-NB.SI(B\$2:B2;">1"))$$

2.6. Analyse de variance

Nous avons vu que $\hat{\beta} \mapsto N\left(\beta, \frac{\sigma_u^2}{\sum x^2}\right)$

$$\rightarrow \frac{\hat{\beta} - \beta}{\sigma_u / \sqrt{\sum x^2}} \mapsto N(0,1)$$

$$\left. \begin{array}{l} D'où \frac{(\hat{\beta} - \beta)^2}{\sigma_u^2 / \sum x^2} \mapsto \chi^2(1) \\ Nous savons que \frac{\sum e^2}{\sigma_u^2} \mapsto \chi^2(n-2) \end{array} \right\} \text{ Il en résulte que}$$

$$\frac{(\hat{\beta} - \beta)^2}{\sum e^2 / (n-2) \sum x^2} = F \mapsto F(1, n-2)$$

Pour $\beta = 0$, F devient:

$$(24) \quad F = \frac{\hat{\beta}^2 \sum x^2}{\sum e^2 / (n-2)} = \frac{\sum \hat{\beta}^2 x^2}{\sum e^2 / (n-2)} = \frac{\sum \hat{y}^2}{\sum e^2 / (n-2)} = \frac{Q_1}{Q_2 / (n-2)}$$

L'hypothèse de non existence de liaison entre X et Y ($\beta = 0$) peut donc être testée à l'aide de (24). On rejettera l'hypothèse au seuil de signification $\epsilon\%$ si $F > F_\epsilon$

Tout le calcul se fait comme l'indique le tableau ci-après :

Sources de variabilité	Somme des carrés	ddl	Carrés moyens
X	$Q_1 = \sum \hat{y}^2$	1	$Q_1/1$
résiduelle	$Q_2 = \sum e^2 = \sum y^2 - \sum \hat{y}$	n - 2	$Q_2/(n - 2)$
Totale	$\sum y^2 = Q_1 + Q_2$	n - 1	

La valeur de F est fournie par le rapport des deux éléments de la dernière colonne du tableau.

S'il existe une liaison entre X et Y, on fait un test sur la signification du

coefficient de corrélation partant de: $t = \frac{(\hat{\beta} - \beta)\sqrt{\sum x^2}}{\hat{\sigma}_u}$ et on acceptait

l'estimation pour $|t| < t_{\frac{\epsilon}{2}}$

On peut tout aussi bien tester l'absence de liaison entre X et Y, hypothèse qu'on accepterait pour $|t| < t_{\frac{\epsilon}{2}}$

$$\begin{aligned}
 \beta = 0 \Rightarrow t &= \frac{\hat{\beta}\sqrt{\sum x^2}}{\hat{\sigma}_u} = \frac{\hat{\beta}\sqrt{\sum x^2}}{\sqrt{\sum e^2/(n-2)}} = \frac{\hat{\beta}\sqrt{\sum x^2}}{\sqrt{\sum y^2(1-r^2)/(n-2)}} \\
 t &= \frac{\hat{\beta}\sqrt{\sum x^2}\sqrt{(n-2)}}{\sqrt{\sum y^2}\sqrt{1-r^2}} = \underbrace{\hat{\beta}\frac{\sigma_x}{\sigma_y}}_r \frac{\sqrt{n-2}}{\sqrt{1-r^2}} \\
 (25) \quad t &= r \frac{\sqrt{n-2}}{\sqrt{1-r^2}}
 \end{aligned}$$

$$\Rightarrow t^2 = r^2 \frac{n-2}{1-r^2} = \frac{r^2}{(1-r^2)/(n-2)}$$

$$t^2 = \frac{\sum \hat{y}^2 / \sum y^2}{(\sum e^2 / \sum y^2) / (n-2)} = \frac{Q_1}{Q_2 / n-2}$$

$$(26) \quad t^2 = \frac{Q_1}{Q_2 / (n-2)} = F \quad \text{ainsi} \quad t^2 \mapsto F(1, n-2)$$

Aussi les tests de corrélation et de variance sont-ils équivalents au test de l'hypothèse de non relation linéaire entre X et Y.

2.7. Préviation

Si la modélisation permet d'accroître notre compréhension des phénomènes économiques étudiés, elle doit également nous aider à la prise de décision, à faire des prévisions.

On pose le problème de prédire la valeur moyenne de Y correspondant à une certaine valeur donnée de X, soit X_0 . pouvant ne pas être à l'intérieur des observations de l'échantillon considéré $X_1, X_2, \dots, X_n$.

La prévision peut prendre la forme d'une valeur particulière ou d'un intervalle de prévision (en terme de fourchette).

i. Estimation de Y_0 , X_0 donné

Soit :

$$(27) \quad \hat{Y}_0 = \hat{\alpha} + \hat{\beta}X_0$$

$$(28) \quad E(\hat{Y}_0) = E(\hat{\alpha}) + X_0 E(\hat{\beta}) = \alpha + \beta X_0 = E(Y_0 / X_0)$$

$$V(\hat{Y}_0) = V(\hat{\alpha}) + 2X_0 \text{COV}(\hat{\alpha}, \hat{\beta}) + X_0^2 V(\hat{\beta}) = \sigma_u^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum x^2} - 2X_0 \frac{\bar{X}}{\sum x^2} + \frac{X_0^2}{\sum x^2} \right)$$

$$= \sigma_u^2 \left(\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum x^2} \right)$$

$$(29) \quad V(\hat{Y}_0) = \sigma_u^2 \left(\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum x^2} \right)$$

La variance est donc fonction de l'écart entre la valeur X_0 et la moyenne des valeurs employées ($\bar{X}$) à estimer α et β

De (27) il ressort que $\hat{Y}_0$ est une fonction linéaire de $\hat{\alpha}$ et $\hat{\beta}$ qui ont une distribution normale, par conséquent il aura une distribution normale dont les paramètres sont donnés par (28), respectivement (29) :

$$\hat{Y}_0 \mapsto N \left(\alpha + \beta X_0; \sigma_u^2 \left(\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum x^2} \right) \right)$$

$$\frac{\hat{Y}_0 - E(Y_0 / X_0)}{SE(\hat{Y}_0)} = \frac{\hat{Y}_0 - E(Y_0 / X_0)}{\hat{\sigma}_u \sqrt{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum x^2}}} \mapsto t(n-2)$$

Par conséquent à $100(1-\varepsilon)\%$, $E(Y_0 / X_0)$ aura pour intervalle de confiance:

$$(30) \quad E(Y_0 / X_0) = \alpha + \beta X_0 = (\hat{\alpha} + \hat{\beta} X_0) \pm t_{\varepsilon/2} \hat{\sigma}_u \sqrt{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum x^2}}$$

ii. Modification du modèle

Quelques fois, il importe d'obtenir un intervalle de confiance pour Y_0 associée à X_0 , ce qui modifie quelque peu le problème puisqu'on dispose cette fois de (X_0, Y_0) ; on vérifie si la même structure linéaire se maintient ou pas.

Si la relation linéaire se maintient, on a :

$$\begin{aligned}
Y_0 &= \alpha + \beta X_0 + u_0 \\
\hat{Y}_0 &= \hat{\alpha} + \hat{\beta} X_0 \\
\underbrace{Y_0 - \hat{Y}_0}_z &= u_0 - (\hat{\alpha} - \alpha) - (\hat{\beta} - \beta) X_0 \\
(31) \quad z &= u_0 - (\hat{\alpha} - \alpha) - (\hat{\beta} - \beta) X_0 = u_0 - ((\hat{\alpha} - \alpha) + (\hat{\beta} - \beta) X_0) \\
(32) \quad E(z) &= 0 \\
V(z) = E(z^2) &= E(u_0^2) + E\left((\hat{\alpha} - \alpha) + (\hat{\beta} - \beta) X_0\right)^2 \\
E(z^2) &= \sigma_u^2 + E\left((\hat{Y}_0 - E(\hat{Y}_0 / X_0))^2\right) = \sigma_u^2 + \sigma_u^2 \left(\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum x^2} \right) \\
(33) \quad V(Y_0 - \hat{Y}_0) &= \sigma_u^2 \left(1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum x^2} \right)
\end{aligned}$$

Nous avons: $\frac{(Y_0 - \hat{Y}_0)}{\hat{\sigma}_u \sqrt{\left(1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum x^2}\right)}} \mapsto t(n-2)$

A 100 (1-ε)%, l'intervalle de prévision de Y_0 sera :

$$(34) \quad Y_0 = \hat{\alpha} + \hat{\beta} X_0 \pm t_{\varepsilon/2} \hat{\sigma}_u \sqrt{\left(1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum x^2}\right)}$$

iii. Méthode Fieller

Nous pouvons être également intéressés par la prévision de X , Y_0 étant donné. La prévision en question peut se ramener au cas précédent si l'on fait fait la régression de X par rapport à Y . Mais il est tout aussi possible de partir du même modèle $DY(X)$ pour prévoir X_0 associé à Y_0 donné et en obtenir l'intervalle de prévision.

De: $\hat{y} = \hat{\beta} x \Rightarrow \hat{x}_0 = \frac{y_0}{\hat{\beta}}$ où $y_0 = Y_0 - \bar{Y}$ et $\hat{x}_0 = \hat{X}_0 - \bar{X}$

$$(35) \quad \hat{X}_0 = \bar{X} + \hat{x}_0 = \bar{X} + \frac{y_0}{\hat{\beta}} = \bar{X} + \frac{Y_0 - \bar{Y}}{\hat{\beta}}$$

On utilise la méthode dite de Fieller pour obtenir les limites de confiance de $\hat{X}_0$ étant donné que y_0 et $\hat{\beta}$ ont des distributions normales

On pose $\frac{y_0}{\hat{\beta}} = \theta$

La variable $y_0 - \hat{\beta} \theta \mapsto N\left(0, \sigma_u^2 \left(1 + \frac{1}{n} + \frac{\theta^2}{\sum x^2}\right)\right)$

$$\frac{y_0 - \hat{\beta} \theta}{SE(y_0 - \hat{\beta} \theta)} \mapsto t(n-2)$$

qu'on utilise pour construire un intervalle de confiance pour θ

A 100(1- ϵ)% on trouve une valeur de $t_{\epsilon/2}$ (à partir de la table) à (n-2) ddl.

$$P\left(\frac{(y_0 - \hat{\beta} \theta)^2}{\hat{\sigma}_u^2 \left(1 + \frac{1}{n} + \frac{\theta^2}{\sum x^2}\right)} \leq t_{\epsilon/2}^2\right) = 1 - \epsilon$$

Pour trouver les limites de θ , on résoud l'équation :

$$(36) \quad (y_0 - \hat{\beta} \theta)^2 - t_{\epsilon/2}^2 \hat{\sigma}_u^2 \left(1 + \frac{1}{n} + \frac{\theta^2}{\sum x^2}\right) = 0$$

On trouve θ_1 et θ_2 qui sont les limites de $\hat{x}_0$

d'où celles de X_0 sont $\bar{X} \pm \theta_i$ ($i = 1, 2$)

Note Si β n'est pas significativement différent de zéro, θ_1 et θ_2 seront des complexes. Nous concluons dans ce cas qu'aucune valeur $\hat{X}_0$ n'est acceptable.

Les limites θ_1 et θ_2 sont encore appelées limites de tolérance et l'intervalle, intervalle de tolérance.

EX. 6. Pour étudier la liaison entre deux indicateurs économiques, X et Y, on a d'abord fait un changement de variables tel que:

X	1	2	3	4	5	6
t	- 5	- 3	- 1	1	3	5

avec les résultats $\sum Y = 30$ $\sum Y^2 = 166$ $\sum Yt = 28$

- 1) Par la méthode des moindres carrés, estimer les droites DY(t) et Dt(Y)
- 2) Partant de la relation linéaire entre t et X en déduire DY(X) et DX(Y)
- 3) Au seuil $\varepsilon = 5\%$
 - prévoir Y pour $X_0 = 10$ à l'aide de DY(X)
 - pour $X_0 = 10$, on observe $Y_0 = 1$, cet résultat est-il compatible avec DY(X)
 - prévoir Y pour $X_0 = 7$ à l'aide de DX(Y).

$$DY(t): Y = \alpha + \beta t + u \quad DY(t): Y = 5 + 0.4t$$

$$Dt(Y): t = a + bY + v \quad Dt(Y): t = -8.75 + 1.75Y$$

$$Dt(X): t = \gamma + \lambda X + \varepsilon \quad Dt(X): t = -7 + 2X$$

$$DY(X): Y = 5 + 0.4t = 5 + 0.4(-7 + 2X) = 2.2 + 0.8X$$

$$DX(Y): -7 + 2X = -8.75 + 1.75Y \Leftrightarrow X = \frac{1}{2}(-1.75 + 1.75Y) = \frac{1.75}{2}(-1 + Y)$$

$$R^2 = 0.8 \left(\frac{1.75}{2} \right) = 0.7$$

	X	t	Y	X^2	t^2	Y^2	Xt	Yt	
	1	-5		1	25		-5		
	2	-3		4	9		-6		
	3	-1		9	1		-3		
	4	1		16	1		4		
	5	3		25	9		15		
	6	5		36	25		30		
Total	21	0	30	91	70	166	35	28	
moyenne	3.5	0.0	5.0						
$\sum x^2$				17.5					
$\sum t^2$					70				
$\sum y^2$						16			
$\sum xt$							35		
$\sum yt$								28	
$\hat{\alpha}$			5.00						$DY(t): Y = \alpha + \beta t + u$
$\hat{\beta}$							0.40		
$\hat{a}$		-8.75							$Dr(Y): t = a + bY + v$
$\hat{b}$							1.75		
r^2								0.70	
$\hat{\gamma}$		-7.00							$Dr(X): t = \gamma + \lambda X + \varepsilon$
$\hat{\lambda}$							2.00		

$$Y_0 = \hat{\alpha} + \hat{\beta} X_0 \pm t_{\xi/2} \hat{\sigma}_u \sqrt{\left(1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum x^2}\right)}$$

$$\hat{\sigma}_u^2 = \frac{\sum e^2}{n-2} = \frac{(1-R^2)\sum y^2}{n-2} = \frac{(1-0.7)16}{6-2} = 1.2$$

$$t_{\varepsilon/2}(n-2) = t_{0.025}(4) = 2.776$$

$$Y_0 = 2.2 + 0.8(10) \pm 2.776 \sqrt{1.2 \left(1 + \frac{1}{6} + \frac{(10-3.5)^2}{17.5}\right)} = 10.2 \pm 2.776(2.073) = 10.2 \pm 5.75$$

$$P(4.45 < Y_0 < 15.95) = 0.95$$

Ainsi, le résultat $Y_0 = 1$ pour $X_0 = 10$ n'est pas dans l'intervalle de prévision ci-dessus et donc le modèle n'aurait pas pu prévoir un tel résultat sans modification.

Par la méthode de Fieller, on résout l'équation:

$$(x_0 - \hat{b}\theta)^2 - t_{\varepsilon/2}^2 \hat{\sigma}_v^2 \left(1 + \frac{1}{n} + \frac{\theta^2}{\sum y^2} \right) = 0$$

$$x_0 = X_0 - \bar{X} = 7 - 3.5 = 3.5$$

$$\hat{\sigma}_v^2 = \frac{\sum e_*^2}{n-2} = \frac{(1-R^2)\sum x^2}{n-2} = \frac{(1-0.7)17.5}{6-2} = 1.3125$$

$$\hat{b} = \frac{1.75}{2}$$

Soit l'équation: $35\theta^2 - 600\theta - 207 = 0$ et soient donc les solutions:
 $\theta_1 = 8.51$ et $\theta_2 = 16.79$

D'où les prévisions de Y: $Y_0 = \bar{Y} + \theta_1$ et $Y_0 = \bar{Y} + \theta_2$ avec $\bar{Y} = 5$, soit finalement l'intervalle de prévision: $P(13.51 < Y_0 < 21.79) = 0.95$

3. Extensions du modèle linéaire simple par les moindres carrés

L'analyse des moindres carrés est très importante et riche d'informations pour la régression linéaire. En effet, analyser la différence $Y - \hat{Y}$ permet de savoir:

- ❖ ☐ S'il n'y a pas de singularité dans les données
- ❖ ☐ Si la relation fonctionnelle choisie est bonne ou pas
- ❖ ☐ S'il n'y a pas de variables omises
- ❖ ☐ Si les hypothèses faites sur les erreurs sont valables ou non i.e.:
 - ✓ Si les erreurs sont indépendantes ou corrélées
 - ✓ Si la variable σ_u^2 est constante ou non
 - ✓ si les erreurs suivent une loi normale ou non.

Les réponses à ces questions (sur les hypothèses - voir le chapitre consacré aux problèmes économétriques) constituent des extensions à l'analyse du modèle linéaire à 2 variables.

3.1. Les singularités

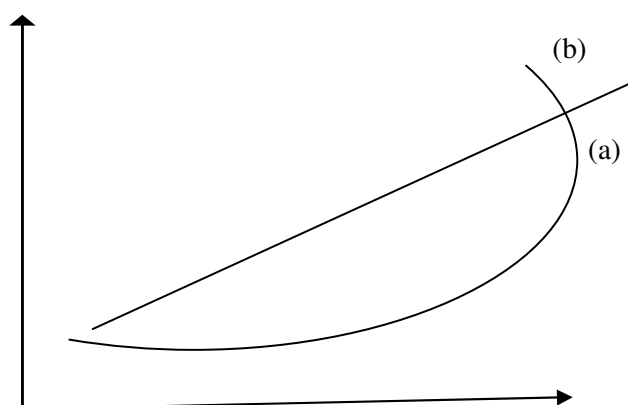
Une valeur singulière est une donnée qui se distingue nettement du reste des observations. Des fois, nous pouvons expliquer pourquoi de telles observations se particularisent des autres. Dans tous les cas, la procédure usuelle est de les omettre et de réestimer la régression ou de procéder à un lissage de la série e.g. Par extrapolation linéaire ou encore d'estimer les valeurs qu'on aurait obtenu en l'absence des circonstances particulières qui ont produit les singularités. du point de vue statistique, il y a une différence selon que les singularités sont écartées après une analyse préliminaire ou pas du tout; les résultats pourraient être ou peu précis ou pas du tout significatifs.

Dans la pratique, il est préférable de choisir la procédure qui donne des résultats significatifs avec moins de précision que celle conduisant à des résultats précis mais peu (ou pas du tout) significatifs.

3.2. Relation non linéaire

Il arrive qu'au-delà des singularités, l'adoption d'une mauvaise relation fonctionnelle conduise à des résultats peu significatifs.

Soit le graphique:



Si choisir ici une relation linéaire (a) conduit à de mauvais résultats, une fonction possible serait par exemple:

(b) $Y = \alpha e^{\beta x}$ courbe exponentielle

Ainsi en économie, la relation entre 2 variables peut être mieux représentée de façon non linéaire. En dehors de toute spécification théorique, la seule représentation graphique peut monter l'inadéquation d'une fonction linéaire.

On adjoint alors aux données initiales la relation non linéaire appropriée qu'on transforme ensuite en relation linéaire pour permettre l'usage des techniques ci-dessus non plus sur les données de départ mais sur leurs transformations.

La transformation peut se faire aussi longtemps que l'additivité des résidus est garantie. Cependant il ne sera pas toujours possible d'user de procédures simples de recherche de transformées. Nous pourrions ainsi être amenés à faire directement une minimisation non linéaire pour estimer les paramètres du problème e.g.:

$$Y = a + \frac{b}{(x+c)^d} + u$$

$$\min \left(Y - a + \frac{b}{(x+c)^d} \right)^2$$

Evidemment, il ne doit pas y avoir de difficulté lorsque la fonction est exprimée sous forme polynomiale e.g. $Y = a + bX + cX^2$

$$\begin{aligned} & \min \sum (Y - a - bX - cX^2)^2 \\ \rightarrow & \begin{cases} \sum Y = na + b\sum X + c\sum X^2 \\ \sum XY = a\sum X + b\sum X^2 + c\sum X^3 \\ \sum X^2Y = a\sum X^2 + b\sum X^3 + c\sum X^4 \end{cases} \end{aligned}$$

Il existe une variété de transformations possibles. Les plus usitées sont les fonctions logarithmiques ou semi-logarithmiques et les fonctions réciproques.

La difficulté est qu'en opérant les transformations il peut arriver que certaines hypothèses de départ ne soient plus respectées. Il faut y prendre garde en faisant les tests, qui ne sont généralement valables que lorsque certaines de ces hypothèses sont maintenues.

a) Soit la fonction $Z = ka^T v$ (a.1)

où T et Z sont les 2 variables

k et a les paramètres du modèle et

v - la variable erreur.

En prenant les logarithmes des 2 membres de (a.1) on obtient:

$$\underbrace{\log(Z)}_Y = \underbrace{\log(k)}_\alpha + \underbrace{\log(a)}_\beta \underbrace{T}_X + \underbrace{\log(v)}_u$$

Soit donc le modèle linéaire: $Y = \alpha + \beta X + u$ (a.2)

Ainsi l'erreur qui est multiplicative en (a.1) devient additive en (a.2)

Les coefficients α et β sont estimés par la méthode des moindres carrés et à l'aide des transformations on peut estimer à leur tour les paramètres k et a : $k = 10^\alpha$ et $a = 10^\beta$

b) Soit $Z = k T^a v$

$$\underbrace{\ln Z}_Y = \underbrace{\ln k}_\alpha + \underbrace{a}_\beta \underbrace{\ln T}_X + \underbrace{\ln v}_u$$

(b.1)

$$Y = \alpha + \beta X + u \rightarrow \hat{k} = e^{\hat{\alpha}} \text{ et } \hat{a} = \hat{\beta}$$

Il faut noter que le paramètre (a) désigne le coefficient d'élasticité théorique de Z par rapport à T :

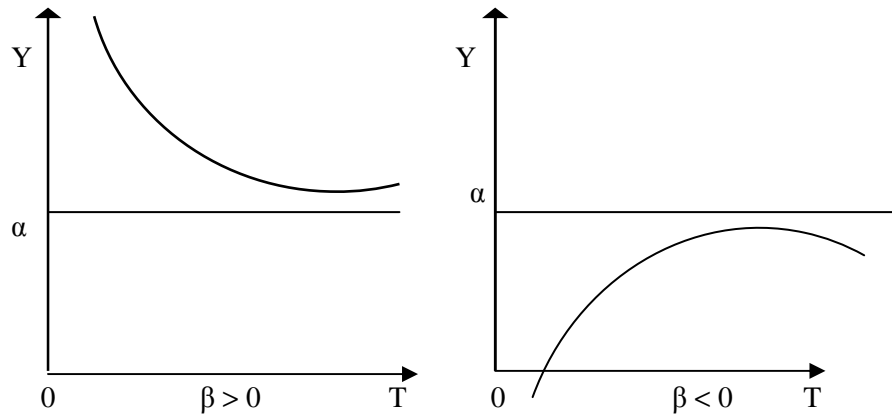
$$\frac{\partial Z}{\partial T} \bigg/ \frac{Z}{T} = \frac{a k T^{a-1} v}{k T^a v / T} = a$$

c) Soit la fonction hyperbolique

$$(c.1) \quad Y = \alpha + \frac{\beta}{T} + u$$

$$\text{On pose } \frac{1}{T} = X$$

d'où: $Y = \alpha + \beta X + u$



d) $Z = e^{\alpha - \beta/T}$ (d.1) avec $\beta > 0$

De telles fonctions utilisent 2 types de transformations.

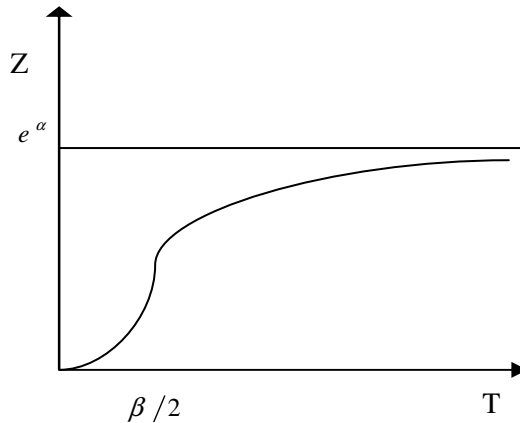
$$\log Z = \alpha - \beta/T \rightarrow Y = \alpha - \beta X \quad \text{où } X = 1/T$$

Pour $T = 0$, Y n'est pas défini mais $\lim_{T \rightarrow 0} Z = 0$

$$\frac{\partial Z}{\partial T} = e^{\alpha - \beta/T} \left(\frac{\beta}{T^2} \right) > 0$$

$$\frac{\partial Z}{\partial T} = e^{\alpha - \beta/T} \left(\frac{\beta^2}{T^4} - \frac{2\beta}{T^3} \right) = 0 \Leftrightarrow \beta(\beta - 2T) = 0$$

soit le point d'inflexion $T = \beta/2$



De telles courbes sont fréquentes dans l'analyse des budgets des ménages. Malheureusement l'application des moindres carrés (ou du maximum de vraisemblance) est souvent laborieuse.

Ex.7. Modèle semi-logarithmique

Au lieu du modèle logarithmique, on peut être amené à construire un modèle semi-logarithmique, par exemple si les valeurs de la variable dépendante sont positives et négatives. Naturellement lorsqu'il y a des valeurs négatives dans la série, on ne peut prendre le logarithme de la variable dans le cadre de la linéarisation d'un modèle exponentiel du genre:

$$(1) \quad Y = a X^{b_2} M^{b_1} K^{b_1} e^u \text{ où}$$

Y le solde global de la balance de paiements (tantôt négatif, tantôt positif)

X la valeur des exportations

M la valeur des importations

K le total des capitaux non monétaires de long terme.

Le modèle semi-logarithmique s'écrit:

$$(2) \quad Y = \ln a + \beta_2 \ln X + \beta_3 \ln M + \beta_4 \ln K + u \quad \Leftrightarrow$$

$$(2') \quad Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + u$$

Il est aisé d'appliquer la méthode des moindres carrés sur (2'), vérifier les hypothèses, faire les tests nécessaires (e.g. autocorrélation des erreurs,

stationnarité, validité du modèle, signification des régresseurs, etc), interpréter les résultats, faire de la simulation ou de la prévision.

Il faut cependant faire attention à la signification économique des coefficients.

Ainsi, dans le modèle (1), les b_1 sont des coefficients d'élasticités, faciles à établir et à interpréter. Mais comme ce modèle (1) n'est pas celui à établir, ses coefficients doivent être déduits à partir du modèle (2').

Le modèle (2') est équivalent à:

$$(3) \quad e^Y = a X^{\beta_2} M^{\beta_3} K^{\beta_4} e^u = e^Y = a Z_2^{\beta_2} Z_3^{\beta_3} Z_4^{\beta_4} e^u$$

On sait que:

$$\beta_j = \frac{de^Y}{dZ_j} \frac{Z_j}{e^Y} = \frac{e^Y dY Z_j}{dZ_j e^Y} = \frac{dY}{dZ_j} Z_j$$

Or de (1), on a:

$$b_j = \frac{dY}{dZ_j} \frac{Z_j}{Y} \Rightarrow \frac{dY}{dZ_j} Z_j = b_j Y$$

D'où:

$$\beta_j = \frac{dY}{dZ_j} Z_j = b_j Y \Leftrightarrow b_j = \frac{\beta_j}{Y}$$

Cette relation est fondamentale pour l'interprétation économique des paramètres qu'on utilise par la suite, permettant de ne pas se tromper d'indicateurs. Dans les applications pratiques, Y peut être remplacé par sa moyenne arithmétique, calculée sur la période d'estimation.

$$\text{Soit finalement} \quad \hat{\beta}_j = b_j \bar{Y} \Leftrightarrow b_j = \frac{\hat{\beta}_j}{\bar{Y}}$$

3.3. Modèle à 3 variables

Soit

$$(1) \quad Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u$$

Le problème est d'estimer les paramètres β_j et σ_u^2

On adopte les notations suivantes (dues à Yule):

$a_{1.23}$ au lieu de $\hat{\beta}_1$

$b_{12.3}$ au lieu de $\hat{\beta}_2$

$b_{13.2}$ au lieu de $\hat{\beta}_3$

On écrira ainsi

$$(2) \quad \hat{Y} = a_{1.23} + b_{12.3} X_2 + b_{13.2} X_3$$

Ces notations ont l'avantage d'indiquer le nombre de variables (1 se réfère à Y, 2 à X_2 et 3 X_3) Après le point on indique les autres variables prises dans la régression. Ainsi $a_{1.23}$ est le coefficient de régression de Y par rapport à X_2 et X_3 tandis que $b_{13.2}$ est le coefficient de Y par rapport à X_3 , l'autre variable considérée étant X_2 .

Nous pouvons étendre les coefficients de régression étudiés dans le cas du modèle linéaire à 2 variables, sachant que nous avons ici 2 variables explicatives.

Dans le modèle simple le coefficient de corrélation était tel que:

$$r^2 = 1 - \frac{\sum e^2}{\sum y^2}$$

Ici on définira le coefficient de corrélation multiple $R_{1.23}$ tel que:

$$R_{1.23}^2 = 1 - \frac{\sum e^2}{\sum y^2} \quad \text{qu'on écrira encore}$$

$$(3) \quad R_{1.23}^2 = 1 - \frac{S_{1.23}^2}{S_1^2} \quad \text{où} \quad \begin{cases} S_{1.23}^2 = \frac{1}{n} \sum (Y - \hat{Y})^2 \\ S_1^2 = \frac{1}{n} \sum y^2 \end{cases}$$

A côté de ce coefficient, nous allons introduire le coefficient de corrélation partiel. Nous nous demanderons par exemple si la corrélation entre Y et X_2 est due au fait que chacune de ces 2 variables est influencée par X_3 . Le coefficient de corrélation partiel entre Y et X_2 suppose que l'on fasse abstraction de l'influence de X_3 . On étudie donc la corrélation entre 2 variables, ceteris paribus. Notons ces coefficients:

$r_{12.3}$ corrélation partielle entre Y et X_2 , X_3 étant constant

$r_{13.2}$ corrélation partielle entre Y et X_3 , X_2 étant constant

$r_{23.1}$ corrélation partielle entre X_2 et X_3 , Y étant constant

Les coefficients de corrélation simples sont:

$$r_{12} = \frac{\sum yx_2}{\sqrt{\sum y^2 \sum x_2^2}} \quad r_{13} = \frac{\sum yx_3}{\sqrt{\sum y^2 \sum x_3^2}} \quad r_{23} = \frac{\sum x_2x_3}{\sqrt{\sum x_2^2 \sum x_3^2}}$$

On définit de même les coefficients de régression simples

$$b_{12} = \frac{\sum yx_2}{\sum x_2^2} \quad b_{13} = \frac{\sum yx_3}{\sum x_3^2}$$

En fait de compte, l'on a toute une hiérarchie de coefficients qu'il faut à présent établir.

3.3.1. Régression

$$RSS = \sum e^2 = \sum (Y - a_{1.23} - b_{12.3}X_2 - b_{13.2}X_3)^2 \min$$

$$\frac{\partial RSS}{\partial a_{1.23}} = 0 \Leftrightarrow \sum Y = na_{1.23} + b_{12.3} \sum X_2 + b_{13.2} \sum X_3$$

$$(4) \quad \bar{Y} = a_{1.23} + b_{12.3} \bar{X}_2 + b_{13.2} \bar{X}_3$$

(2) - (4) conduit à:

$$(5) \quad \hat{y} = b_{12.3}x_2 + b_{13.2}x_3$$

On a donc $\sum (Y - \hat{Y})^2 = \sum (Y - \bar{Y} - (\hat{Y} - \bar{Y}))^2 = \sum (y - \hat{y})^2$

$$\sum (y - \hat{y})^2 = \sum (y - b_{12.3}x_2 - b_{13.2}x_3)^2 = \sum e^2$$

$$(6) \quad \begin{cases} \frac{\partial \sum e^2}{\partial b_{12.3}} = 0 \\ \frac{\partial \sum e^2}{\partial b_{13.2}} = 0 \end{cases} \Leftrightarrow \begin{cases} b_{12.3} \sum x_2^2 + b_{13.2} \sum x_2 x_3 = \sum x_2 y \\ b_{12.3} \sum x_2 x_3 + b_{13.2} \sum x_3^2 = \sum x_3 y \end{cases}$$

En combinant (4) et (6) on obtient les estimateurs:

$$a_{1.23} = \bar{Y} - b_{12.3} \bar{X}_2 - b_{13.2} \bar{X}_3$$

$$b_{12.3} = \frac{\sum yx_2 \sum x_3^2 - \sum yx_3 \sum x_2 x_3}{\sum x_2^2 \sum x_3^2 - (\sum x_2 x_3)^2}$$

$$b_{13.2} = \frac{\sum yx_3 \sum x_2^2 - \sum yx_2 \sum x_2 x_3}{\sum x_2^2 \sum x_3^2 - (\sum x_2 x_3)^2}$$

En prenant en compte les coefficients de corrélation simples, les estimateurs peuvent encore s'écrire, en divisant tous les termes par $\sum x_2^2 \sum x_3^2$:

$$b_{12.3} = \frac{\frac{\sum yx_2}{\sum x_2^2} - \frac{\sum yx_3 \sum x_2 x_3}{\sum x_2^2 \sum x_3^2}}{1 - \frac{(\sum x_2 x_3)^2}{\sum x_2^2 \sum x_3^2}}$$

$$= \frac{\frac{\sum yx_2}{\sqrt{\sum y^2 \sum x_2^2}} \sqrt{\frac{\sum y^2}{\sum x_2^2}} - \frac{\sum yx_3}{\sqrt{\sum y^2 \sum x_3^2}} \frac{\sum x_2 x_3}{\sqrt{\sum x_2^2 \sum x_3^2}} \sqrt{\frac{\sum y^2}{\sum x_2^2}}}{1 - \left(\frac{\sum x_2 x_3}{\sqrt{\sum x_2^2 \sum x_3^2}} \right)^2}$$

$$b_{12.3} = \frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2} \frac{S_1}{S_2}$$

où

$$S_1 = \sqrt{S_1^2} = \sqrt{\frac{1}{n} \sum y^2} \quad \text{et} \quad S_2 = \sqrt{S_2^2} = \sqrt{\frac{1}{n} \sum x_2^2}$$

Finalement on écrira:

$$(7) \quad \begin{cases} a_{1.23} = \bar{Y} - b_{12.3} \bar{X}_2 - b_{13.2} \bar{X}_3 \\ b_{12.3} = \frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2} \frac{S_1}{S_2} \\ b_{13.2} = \frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2} \frac{S_1}{S_3} \end{cases}$$

3.3.2. Coefficient de corrélation multiple

$$R_{1.23}^2 = 1 - \frac{\sum e^2}{\sum y^2} = 1 - \frac{S_{1.23}^2}{S_1^2}$$

qu'on va aussi chercher à exprimer en fonction des coefficients simples

$$\begin{aligned} \sum e^2 &= \sum e \cdot e = \sum e(y - \hat{y}) = \sum e(y - b_{12.3}x_2 - b_{13.2}x_3) \\ &= \sum ey - b_{12.3} \sum ex_2 - b_{13.2} \sum ex_3 \end{aligned}$$

Mais comme:

$$\begin{cases} \frac{\partial \sum e^2}{\partial b_{12.3}} = 0 \\ \frac{\partial \sum e^2}{\partial b_{13.2}} = 0 \end{cases} \Rightarrow \begin{cases} \sum ex_2 = 0 \\ \sum ex_3 = 0 \end{cases}$$

(C'est à dire que l'erreur n'est pas corrélée aux variables explicatives)

$$\begin{aligned} \text{D'où } \sum e^2 &= \sum ey \\ &= \sum y(y - b_{12.3}x_2 - b_{13.2}x_3) \end{aligned}$$

$$(8) \quad RSS = \sum y^2 - b_{12.3} \sum yx_2 - b_{13.2} \sum yx_3$$

Ainsi:

$$R_{1.23}^2 = 1 - \frac{\sum e^2}{\sum y^2} = 1 - \frac{\sum y^2 - b_{12.3} \sum yx_2 - b_{13.2} \sum yx_3}{\sum y^2}$$

$$R_{12.3}^2 = b_{12.3} \frac{\sum yx_2}{\sum y^2} + b_{13.2} \frac{\sum yx_3}{\sum y^2}$$

$$= b_{12.3} \frac{\sum yx_2}{\sqrt{\sum y^2 \sum x_2^2}} \sqrt{\frac{\sum x_2^2}{\sum y^2}} + b_{13.2} \frac{\sum yx_3}{\sqrt{\sum y^2 \sum x_3^2}} \sqrt{\frac{\sum x_3^2}{\sum y^2}}$$

$$R_{1.23}^2 = b_{12.3} r_{12} \frac{S_2}{S_1} + b_{13.2} r_{13} \frac{S_3}{S_1}$$

En remplaçant $b_{12.3}$ et $b_{13.2}$ par leurs expressions de (9.7), il vient:

$$R_{1.23}^2 = \frac{1}{1 - r_{23}^2} \left((r_{12} - r_{13}r_{23}) \frac{S_1}{S_2} r_{12} \frac{S_2}{S_1} + (r_{13} - r_{12}r_{23}) \frac{S_1}{S_3} r_{13} \frac{S_3}{S_1} \right)$$

$$= \frac{1}{1-r_{23}^2} \left[r_{12}^2 - r_{12}r_{13}r_{23} + r_{13}^2 - r_{13}r_{12}r_{23} \right]$$

$$(9) \quad R_{1.23}^2 = \frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}$$

Si les variables explicatives sont indépendantes des erreurs, il y va de même pour $\hat{y}$:

$$\sum e\hat{y} = b_{12.3} \underbrace{\sum ex_2}_0 + \underbrace{\sum ex_3}_0 = 0$$

Ainsi puisque $y = \hat{y} + e$, alors

$$(10) \quad \begin{aligned} \sum y^2 &= \sum \hat{y}^2 + \sum e^2 \Leftrightarrow Q = Q_1 + Q_2 \\ &\Rightarrow \sum \hat{y}^2 = \sum y^2 - \sum e^2 \\ &= \sum y^2 - \left(\sum y^2 - b_{12.3} \sum x_2 y - b_{13.2} \sum x_3 y \right) \end{aligned}$$

d'où:

$$(11) \quad \sum \hat{y}^2 = b_{12.3} \sum x_2 y + b_{13.2} \sum x_3 y$$

$$\text{Mais} \quad \sum \hat{y}^2 = \sum y^2 - \sum e^2 \Rightarrow \frac{\sum \hat{y}^2}{\sum y^2} = 1 - \frac{\sum e^2}{\sum y^2} = R_{1.23}^2$$

3.3.3. Coefficients de corrélation partiels

Pour calculer le coefficient de corrélation entre Y et une des 2 variables, X_2 , X_3 , il faut éliminer l'influence linéaire de l'autre variable explicative sur Y et sur la variable considérée. Ainsi pour déterminer $r_{12.3}$, nous partons de la régression linéaire de Y sur X_3 , donnée par:

$$Y = a_{1.3} + b_{13} X_3 + u \Rightarrow y = b_{13} x_3 + u \Rightarrow$$

$$(12) \quad u = y - b_{13} x_3$$

Analogue, le résidu des X_2 , après élimination de l'influence de X_3 est défini par:

$$(13) \quad v = x_2 - b_{23}x_3$$

Partant, le coefficient de corrélation partiel entre Y et X_2 , avec X_3 maintenu constant, sera défini comme coefficient de corrélation simple entre u et v

$$r_{12.3} = \frac{\sum uv}{\sqrt{\sum u^2} \sqrt{\sum v^2}}, \text{ avec } E(u) = 0 \text{ et } E(v) = 0$$

$$r_{12.3} = \frac{\sum yx_2 - b_{23} \sum yx_3 - b_{13} \sum x_2x_3 + b_{12} \sum x_3^2}{\sqrt{\sum y^2 (1 - r_{13}^2)} \sqrt{\sum x_2^2 (1 - r_{23}^2)}}$$

$$r_{12.3} = \frac{\sum yx_2 - r_{13} \frac{S_1}{S_3} \sum x_2x_3 - r_{23} \frac{S_2}{S_3} \sum yx_3 + r_{13}r_{23} \frac{S_1}{S_3} \frac{S_2}{S_3} \sum x_3^2}{\sqrt{\sum y^2} \sqrt{\sum x_2^2} \sqrt{1 - r_{13}^2} \sqrt{1 - r_{23}^2}}$$

soit finalement:

$$(14) \quad r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - r_{13}^2} \sqrt{1 - r_{23}^2}} \text{ et par analogie}$$

$$r_{13.2} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{1 - r_{12}^2} \sqrt{1 - r_{23}^2}}$$

$$r_{23.1} = \frac{r_{23} - r_{12}r_{13}}{\sqrt{1 - r_{12}^2} \sqrt{1 - r_{13}^2}}$$

A partir de ces relations et en les comparant aux calculs précédents des coefficients de régression partiels, on aboutit à des expressions fort utiles

$$b_{12.3} = \frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2} \frac{S_1}{S_2} \text{ qu'on peut encore écrire:}$$

$$b_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\underbrace{\sqrt{1-r_{13}^2}\sqrt{1-r_{23}^2}}_{r_{12.3}}} \frac{S_1\sqrt{1-r_{13}^2}}{\underbrace{S_2\sqrt{1-r_{23}^2}}_{S_{1.3}/S_{2.3}}}$$

$$(15) \quad b_{12.3} = r_{12.3} \frac{S_{1.3}}{S_{2.3}} \text{ et analogue}$$

$$b_{13.2} = r_{13.2} \frac{S_{1.2}}{S_{3.2}}$$

Ces 2 dernière relations sont équivalents à $\hat{\beta} = r \frac{S_y}{S_x}$ du cas simple de 2 variables

Enfin il est très pratique et utile d'exprimer les coefficients de corrélation partiels en fonction du coefficient multiple et des coefficients simples

De (14) on a:

$$r_{12.3}^2 = \frac{(r_{12} - r_{13}r_{23})^2}{(1-r_{13}^2)(1-r_{23}^2)} = \frac{r_{12}^2 + r_{13}^2 r_{23}^2 - 2r_{12}r_{13}r_{23}}{(1-r_{13}^2)(1-r_{23}^2)}$$

qu'on mettra sous la forme:

$$r_{12.3}^2 = \frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23} - r_{13}^2 + r_{13}^2 r_{23}^2}{(1-r_{13}^2)(1-r_{23}^2)}$$

$$\text{Or d'après (9) } R_{1.23}^2 = \frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1-r_{23}^2}$$

$$\text{d'où } r_{12.3}^2 = \frac{R_{1.23}^2(1-r_{23}^2) - r_{13}^2(1-r_{23}^2)}{(1-r_{13}^2)(1-r_{23}^2)}$$

$$(16) \quad r_{12.3}^2 = \frac{R_{1.23}^2 - r_{13}^2}{1 - r_{13}^2} \text{ et analogue}$$

$$r_{13.2}^2 = \frac{R_{1.23}^2 - r_{12}^2}{1 - r_{12}^2}$$

$r_{12.3}^2$ et $r_{13.2}^2$ sont appelés rendements partiels. Ils mesurent la proportion de la variabilité de Y qu'explique l'addition de X_2 (respectivement de X_3), variabilité non due à X_3 (respectivement de X_2). Ce sont des gains relatifs de

précision $\frac{\sum e_j^2 - \sum e^2}{\sum e_j^2}$

Finalement, il en résulte les quantités suivantes:

1. $S_1^2 r_{13}^2$ variabilité de Y qu'explique X_3
2. $S_1^2 r_{12}^2$ variabilité de Y qu'explique X_2
3. $S_1^2 R_{1.23}^2$ variabilité de Y qu'explique X_2 et X_3
4. $S_1^2 (R_{1.23}^2 - r_{13}^2)$ accroissement dans la variabilité de Y due à l'introduction d'une seconde variable explicative, X_2

3.3.4. Résumé des calculs

1. Partant des données, on calcule:

$$\bar{Y}, \quad \bar{X}_2, \bar{X}_3$$

$$S_1, \quad S_2, \quad S_3$$

$$r_{12}, \quad r_{13}, \quad r_{23}$$

2. On détermine:

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - r_{13}^2} \sqrt{1 - r_{23}^2}} \text{ et } r_{13.2} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{1 - r_{12}^2} \sqrt{1 - r_{23}^2}}$$

$$S_{1.2} = S_1 \sqrt{1 - r_{12}^2} \quad S_{2.3} = S_2 \sqrt{1 - r_{23}^2}$$

$$S_{1.3} = S_1 \sqrt{1 - r_{13}^2} \quad S_{3.2} = S_3 \sqrt{1 - r_{23}^2}$$

3. On estime les coefficients de régression:

$$a_{1.23} = \bar{Y} - b_{12.3} \bar{X}_2 - b_{13.2} \bar{X}_3$$

$$b_{12.3} = r_{12.3} \frac{S_{1.3}}{S_{2.3}}$$

$$b_{13.2} = r_{13.2} \frac{S_{1.2}}{S_{3.2}}$$

4. Enfin on a: $R_{1.23}^2 = \frac{b_{12.3} \sum yx_2 + b_{13.2} \sum yx_3}{\sum y^2}$

Quant aux problèmes d'inférence statistique et de prévision, ils seront mieux traités dans le cadre du modèle linéaire général avec l'introduction du calcul matriciel.

Ex.8. L'offre Y d'un bien est supposée être une fonction linéaire de son prix X_2 et du taux de salaire X_3 dans l'industrie de production de ce bien: $Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u$

On considère un n échantillon:

Y	2	4	3	5	6	7	8	9
X_2	1	2	2	3	4	4	4	8
X_3	9	8	5	7	4	5	7	5

- 1) Par la méthode des moindres carrés, estimer le modèle $DY(X_2, X_3)$
- 2) Calculer le coefficient de corrélation corrigé
- 3) Faire une analyse de variance détaillée
- 4) Au seuil $\xi = 5\%$, tester les effets direct et additionnel de chacune des variables explicatives et calculer les gains de précision puis ensuite en déduire les coefficients de corrélation partiels

Microsoft Excel - DESS_Elcel méthodes quantitatives										
Fichier Edition Affichage Insertion Format Outils Données Fenêtre Outils de l'étudiant										
L11										
	A	B	C	D	E	F	G	H	I	J
1	Y	X ₂	X ₃		RAPPORT DÉTAILLÉ					
2	2	1	9							
3	4	2	8		Statistiques de la régression					
4	3	2	5		Coefficient de détermination multiple	0.900				
5	5	3	7		Coefficient de détermination R ²	0.811				
6	6	4	4		Coefficient de détermination corrigé $\bar{R}^2$	0.735				
7	7	4	5		Erreur-type	1.261				
8	8	4	7		Observations	8				
9	9	8	5							
10					ANALYSE DE VARIANCE					
11						ddl	S. carrés	M. carrés	F	P(F)
12					Régression	2	34.046	17.023	10.702	0.016
13					Résidus	5	7.954	1.591		
14					Total	7	42.000			
15										
16						Coefficients	SE	t	P	Inf P(0.95)
17					Constante	1.635	2.766	0.591	0.580	-5.476
18					Variable X2	1.046	0.272	3.850	0.012	0.348
19					Variable X3	0.032	0.332	0.098	0.926	-0.820

$$\bar{R}^2 = 1 - \frac{\sum e^2 / n - 3}{\sum y^2 / n - 1} = 1 - \frac{n - 1}{n - 3} \frac{\sum e^2}{\sum y^2}$$

Au regard du tableau d'analyse de variance, le modèle est validé:

$$F = 10.702 > F_{\text{tabulé}} = 5.79 \text{ avec un risque } \varepsilon < 5\% = 1.6\%$$

Pour les tests d'effets direct et additionnel, on suppose les régressions DY(X2) et DY(X3) au préalable effectuées:

Microsoft Excel - DESS_Elcel méthodes quantitatives						
Fichier Edition Affichage Insertion Format Outils Données Fenêtre Outils de l'étudiant						
I20						
	A	B	C	D	E	F
1	RAPPORT DÉTAILLÉ					
2						
3	Statistiques de la régression					
4	Coefficient de détermination multiple	0.900				
5	Coefficient de détermination R ²	0.810				
6	Coefficient de détermination corrigé $\bar{R}^2$	0.779				
7	Erreur-type	1.152				
8	Observations	8				
9						
10	ANALYSE DE VARIANCE					
11		ddl	S. carrés	M. carrés	F	P(F)
12	Régression	1	34.031	34.031	25.624	0.002
13	Résidus	6	7.969	1.328		
14	Total	7	42			
15						
16		Coefficients	SE	t	P	Inf P(0.95)
17	Constante	1.891	0.821	2.302	0.061	-0.119
18	Variable X2	1.031	0.204	5.062	0.002	0.533

Microsoft Excel - DESS_Excel méthodes quantitatives						
Fichier Edition Affichage Insertion Format Outils Données Fenêtre Outils de l'étudiant ?						
J20 fx						
	A	B	C	D	E	F
1	RAPPORT DÉTAILLÉ					
2						
3	<i>Statistiques de la régression</i>					
4	Coefficient de détermination multiple	0.499				
5	Coefficient de détermination R^2	0.249				
6	Coefficient de détermination corrigé $\bar{R}^2$	0.124				
7	Erreur-type	2.293				
8	Observations	8				
9						
10	ANALYSE DE VARIANCE					
11		<i>ddl</i>	<i>S. carrés</i>	<i>M. carrés</i>	<i>F</i>	<i>P(F)</i>
12	Régression	1	10.465	10.465	1.991	0.208
13	Résidus	6	31.535	5.256		
14	Total	7	42			
15						
16		<i>Coefficients</i>	<i>SE</i>	<i>t</i>	<i>P</i>	<i>Inf P(0.95)</i>
17	Constante	9.860	3.195	3.087	0.021	2.043
18	Variable X3	-0.698	0.494	-1.411	0.208	-1.907
						0.512

Au regard de ces deux tableaux de résultats, l'effet direct de X_2 est significatif mais pas celui de X_3 . Pour les effets additionnels, on construit le tableau:

Source de variabilité	Somme des carrés	ddl	Carrés moyens
X_2	$r_2^2 \sum y^2$	1	$r_2^2 \sum y^2 / 1$
$+X_3$	$(R^2 - r_2^2) \sum y^2$	1	$(R^2 - r_2^2) \sum y^2 / 1$
X_3	$r_3^2 \sum y^2$	1	$r_3^2 \sum y^2 / 1$
$+X_2$	$(R^2 - r_3^2) \sum y^2$	1	$(R^2 - r_3^2) \sum y^2 / 1$
X_2 et X_3	$R^2 \sum y^2$	2	$R^2 \sum y^2 / 2$
Résidu	$(1 - R^2) \sum y^2$	n-3	$(1 - R^2) \sum y^2 / n - 3$
Total	$\sum y^2$	n-1	

$$F_{+X_3} = \frac{(R^2 - r_2^2) / 1}{(1 - R^2) / n - 3} = (n - 3) \frac{R^2 - r_2^2}{1 - R^2} = 0.010 < F_{0.05}(1, 5) = 6.61$$

effet additionnel non significatif

$$F_{+X_2} = \frac{(R^2 - r_3^2)/1}{(1 - R^2)/n - 3} = (n - 3) \frac{R^2 - r_3^2}{1 - R^2} = 14.824 > F_{0.05}(1, 5) = 6.61$$

effet additionnel significatif

Les gains de précision sont définis par:

$$G_{+X_3} = \frac{\sum e_2^2 - \sum e^2}{\sum e_2^2} = \frac{R^2 - r_2^2}{1 - r_2^2} = 0.2\% \quad \text{et}$$

$$G_{+X_2} = \frac{\sum e_3^2 - \sum e^2}{\sum e_3^2} = \frac{R^2 - r_3^2}{1 - r_3^2} = 74.8\%$$

Les coefficients de corrélation partiels sont tels que: $G_{+X_3} = r_{13.2}^2$

et $G_{+X_2} = r_{12.3}^2$

Soient: $r_{12.3} = 0.865$ et $r_{13.2} = 0.044$