**Université de Ghardaïa**

**Faculté des Sciences Economiques De Gestion et Commerciales**

**Département : sciences financières et comptabilité**

# COURS

# INTRODUCTION A L'ECONOMETRIE

**À l'usage des étudiants inscrits en :**

**3eme année Licence (S5)**

**Enseignant : Dr. Bendjouad Messaoud**

**Année 2022/2023**

# Table Des Matières

## جدول المحتويات

## 1. Introduction

La recherche quantitative vise à expliquer les phénomènes par une investigation empirique systémique des phénomènes observables par la **collecte** de données numériques, **analysées** à travers des méthodes fondées sur des techniques mathématique, statistique ou informatique.

La recherche quantitative implique la collecte et l'analyse des données qui soient quantifiables.

Dans une recherche quantitative, la question de la mesure est essentielle car elle permet l'observation empirique et sa connexion avec dimension conceptuelle de la recherche.

Toutes les données quantitatives sont des données sous forme numérique telles que des statistiques, des pourcentages, etc. obtenus par des sondages, des questionnaires et des enquêtes, ou en manipulant des données statistiques pré-existantes.

Les données recueillies auprès des personnes sont appelées « variables ».

Les variables sont les caractéristiques de l'unité d'observation qui nous intéresse que l'on cherche à recueillir (par exemple la taille d'une entreprise, le chiffre d'affaires, l'âge d'un employé, etc.).

L'étiquette «variable» se réfère au fait que les données différent en fonction des unités observées.

Dans une recherche quantitative, les sujets sont généralement observés plusieurs fois (avant et après traitement, etc.).

Il existe trois principaux types de recherches quantitatives :

1. Descriptif
2. Expérimental (Quelle est la cause ?)
3. Ex post facto / causalité (Quelles sont les causes possibles ?).

Le but d'une recherche quantitative est de déterminer la relation générale entre une chose (la variable indépendante) vis-à-vis d'une autre (variable indépendante) dans une population.

Les recherches quantitatives sont un moyen pour les chercheurs de généraliser des données observées sur un échantillon.

Le but d'une recherche quantitative est d'obtenir des informations qui peuvent être déduites (ou généralisée) de cet échantillon à de larges populations d'unités.

En terme de méthodes, «**Econométrie**», est un terme générique englobant les méthodes et approches, y compris les tests classiques statistiques (test t, analyse de la variance, etc.), la régression, la modélisation par équations structurelles, les modèles l hiérarchique linéaire (l'analyse multi-niveaux), etc.

Des logiciels spécialisés (SAS, Stata, SPSS, R, EVIEWS, etc.) sont généralement utilisés pour effectuer les différents tests et les analyses statistiques.

### 1.1. Définition de l'Econométrie :

L'Econométrie est un ensemble des méthodes de recherche, utilisant des outils d'analyse mathématiques et statistiques, en vue de décrire, d'expliquer et prédire des phénomènes par le biais de concepts opérationnalisés sous forme de variables mesurables.

L'économétrie désigne l'ensemble des techniques destinées à mesurer des grandeurs économiques.

Dans ce cadre, elle remplit trois fonctions :

- La **mesure** de grandeurs préalablement définies par l'économie (emploi, croissance, valeur ajoutée, etc.);
- La **vérification** empirique de relations entre ces grandeurs prédites par des modèles issus de l'économie mathématique ;
- L'**étude** a priori de relations entre grandeurs mathématiques indépendamment d'un modèle économique sous-jacent.

L'économétrie est ainsi la branche de la statistique appliquée à l'économie.

### 1.2. La notion de modèle

Il est délicat de fournir une définition unique de la notion de modèle1. Dans le cadre de l'économétrie, nous pouvons considérer qu'un modèle consiste en une *présentation formalisée d'un phénomène* sous forme d'équations dont les variables sont des grandeurs

économiques. L'objectif du modèle est de représenter les traits les plus marquants d'une réalité qu'il cherche à styliser. Le modèle est donc l'outil que le modélisateur utilise lorsqu'il cherche à comprendre et à expliquer des phénomènes. Pour ce faire, il émet des hypothèses et explicite des relations.

- *Pourquoi des modèles ?*
- *Nombreux sont ceux – sociologues, économistes ou physiciens – qui fondent leurs analyses ou leurs jugements sur des raisonnements construits et élaborés. Ces constructions refèrent implicitement à des modèles ; alors pourquoi ne pas expliciter clairement les hypothèses et les relations au sein d'un modèle ?*

Le modèle est donc une présentation schématique et partielle d'une réalité naturellement plus complexe. Toute la difficulté de la modélisation consiste à ne retenir que la ou les représentations intéressantes pour le problème que le modélisateur cherche à expliciter. Ce choix dépend de la nature du problème, du type de décision ou de l'étude à effectuer. La même réalité peut ainsi être formalisée de diverses manières en fonction des objectifs.

### 1.3. La construction des modèles en économétrie

Dans les sciences sociales, et particulièrement en économie, les phénomènes étudiés concernent le plus souvent des comportements afin de mieux comprendre la nature et le fonctionnement des systèmes économiques. L'objectif du modélisateur est, dans le cadre de l'économétrie et au travers d'une mesure statistique, de permettre aux agents économiques (ménages, entreprises, État...) d'intervenir de manière plus efficace. La construction d'un modèle comporte un certain nombre d'étapes qui sont toutes importantes. En effet, en cas de faiblesse d'un des « maillons », le modèle peut se trouver invalidé pour cause d'hypothèses manquantes, de données non représentatives ou observées avec des erreurs, etc. Examinons les différentes étapes à suivre lors de la construction d'un modèle, ceci à partir de l'exemple du modèle keynésien simplifié.

### 1.3.1. Référence à une théorie

Une théorie s'exprime au travers d'hypothèses auxquelles le modèle fait référence. Dans la théorie keynésienne, quatre propositions sont fondamentales :

- la consommation et le revenu sont liés ;
- le niveau d'investissement privé et le taux d'intérêt sont également liés ;
- il existe un investissement autonome public ;
- enfin, le produit national est égal à la consommation plus l'investissement privé et public.

### 1.3.2.Formalisation des relations et choix de la forme des fonctions

À partir des propositions précédentes, nous pouvons construire des relations :

- la consommation est fonction du revenu : $C = f(Y)$ avec $f' > 0$ ;
- l'investissement privé dépend du taux d'intérêt : $I = g(r)$ avec $g' < 0$ ;
- il existe un investissement autonome public : $\overline{I}$ ;
- enfin, le produit national (ou le revenu national) est égal à la consommation plus l'investissement : $Y \equiv C + I + \overline{I}$.

### 1.3.3.Sélection et mesure des variables

Le modèle étant spécifié, il convient de collecter les variables représentatives des phénomènes économiques. Ce choix n'est pas neutre et peut conduire à des résultats différents, les questions qu'il convient de se poser sont par exemple :

- *Faut-il raisonner en euros constants ou en euros courants ?*
- *Les données sont-elles brutes ou CVS*1 ?
- *Quel taux d'intérêt faut-il retenir (taux au jour le jour, taux directeur de la Banque Centrale Européenne,...) ?* etc.

### 1.3.4.Décalages temporels

Dans le cadre de modèle spécifié en séries temporelles, les relations entre les variables ne sont pas toujours synchrones mais peuvent être décalées dans le temps. Nous pouvons concevoir que la consommation de l'année *t* est expliquée par le revenu de l'année (*t-1*) et non celui de l'année *t* . Pour lever cette ambiguïté, il est d'usage d'écrire

le modèle en le spécifiant à l'aide d'un indice de temps : $C_t = a_0 + a_1 Y_{t-1}$ . La variable $Y_{t-1}$ est appelée « variable endogène retardée ».

*On appelle « variable exogène » une variable dont les valeurs sont prédéterminées, et « variable endogène » une variable dont les valeurs dépendent des variables exogènes.*

### 1.3.5.Validation du modèle

La dernière étape est celle de la validation du modèle :

- *Les relations spécifiées sont-elles valides ?*
- *Peut-on estimer avec suffisamment de précision les coefficients ?*
- *Le modèle est-il vérifié sur la totalité de la période ?*
- *Les coefficients sont-ils stables ? Etc.*

À toutes ces questions, les techniques économétriques s'efforcent d'apporter des réponses.

## 1.4. Le rôle de l'économétrie

### 1.4.1. L'économétrie comme validation de la théorie

L'économétrie est un outil à la disposition de l'économiste qui lui permet d'infirmer ou de confirmer les théories qu'il construit. Le théoricien postule des relations ; l'application de méthodes économétriques fournit des estimations sur la valeur des coefficients ainsi que la précision attendue.

Une question se pose alors : pourquoi estimer ces relations, et les tester statistiquement ?

Plusieurs raisons incitent à cette démarche : tout d'abord cela force l'individu à établir clairement et à estimer les interrelations sous-jacentes. Ensuite, la confiance aveugle dans l'intuition peut mener à l'ignorance de liaisons importantes ou à leur mauvaise utilisation. De plus, des relations marginales mais néanmoins explicatives, qui ne sont qu'un élément d'un modèle global, doivent être testées et validées afin de les mettre à leur véritable place.

Enfin, il est nécessaire de fournir, en même temps que l'estimation des relations, une mesure de la confiance que l'économiste peut avoir en celles-ci, c'est-à-dire la précision que l'on peut en attendre. Là encore, l'utilisation de méthodes purement qualitatives exclut toute mesure quantitative de la fiabilité d'une relation.

### 1.4.2.L'économétrie comme outil d'investigation

L'économétrie n'est pas seulement un système de validation, mais également un outil d'analyse. Nous pouvons citer quelques domaines où l'économétrie apporte une aide à la modélisation, à la réflexion théorique ou à l'action économique par :

- la mise en évidence de relations entre des variables économiques qui n'étaient pas *a priori* évidentes ou pressenties ;
- l'induction statistique ou l'inférence statistique consiste à inférer, à partir des caractéristiques d'un échantillon, les caractéristiques d'une population. Elle permet de déterminer des intervalles de confiance pour des paramètres du modèle ou de tester si un paramètre est significativement1 inférieur, supérieur ou simplement différent d'une valeur fixée ;
- la simulation qui mesure l'impact de la modification de la valeur d'une variable sur une autre $(\Delta Ct = a_1 \Delta Yt)$ ;
- la prévision1, par l'utilisation de modèles économétriques, qui est utilisée par les pouvoirs publics ou l'entreprise afin d'anticiper et éventuellement de réagir à l'environnement économique.

Dans cet ouvrage, nous nous efforcerons de montrer, à l'aide d'exemples, les différentes facettes de l'utilisation des techniques économétriques dans des contextes et pour des objectifs différents.

## 2. Variables Statistiques et lois de probabilité

### 2.1. Les Variables statistiques.

- Population : Ensemble que l'on observe et qui sera soumis à une analyse statistique. Chaque élément de cet ensemble est un individu ou unité statistique.
- Echantillon : C'est un sous ensemble de la population considérée. Le nombre d'individus dans l'échantillon est la taille de l'échantillon.
- Caractère : C'est la propriété ou l'aspect singulier que l'on se propose d'observer dans la population ou l'échantillon. Un caractère qui fait le sujet d'une étude porte aussi le nom de variable statistique.

Différents types de variables statistiques :

• Lorsque la variable ne se prête pas à des valeurs numériques, elle est dite **qualitative** (exemple : opinions politiques, couleurs des yeux...) .Elle peut être ordonnée ou non, dichotomique ou non.

• Lorsque la variable peut être exprimée numériquement, elle est dite **quantitative** ( ou mesurable). Dans ce cas, elle peut être discontinue ou continue.

♦ Elle est **discontinue** si elle ne prend que des valeurs isolées les unes des autres. Une variable discontinue qui ne prend que des valeurs entières est dite discrète (exemple : nombre d'enfants d'une famille).

♦ Elle est dite **continue** lorsqu'elle peut prendre toutes les valeurs d'un intervalle fini ou infini (exemple : diamètre de pièces, salaires...).

| **Variables** | | | |
|---|---|---|---|
| **Variables qualitatives** | | **Variables quantitatives** | |
| **Nominale** | **Ordinale** | **Discrète** | **Continue** |
| Une **variable** est dite **qualitative nominale** quand ses valeurs sont des éléments d'une catégorie type nom non hiérarchique. En d'autres termes, ses éléments ne peuvent pas se ranger dans une gradation logique, selon une hiérarchie naturelle. La donnée qualitative nominale ne peut donc être appréhendée qu'à travers des modalités entre lesquelles il n'existe aucune relation d'ordre. | Une **variable** qualitative **ordinale** possède toutes les propriétés de la variable qualitative nominale avec en plus la possibilité de positionner et de hiérarchiser les individus entre eux selon la valeur attachée à leur caractère. En d'autres termes, il sera possible de ranger dans une gradation logique, selon une hiérarchie naturelle, les individus de la population étudiée pour le caractère retenu. EXAMPLE - Modalités : très | Une **variable** est dite **discrète** quand elle prendre un **nombre fini** ou dénombrable de valeurs. En d'autres termes, le passage d'une modalité à une autre est « brutal », sans continuité, sans glissement progressif. C'est typiquement le cas des variables qualitatives nominales et ordinales pour lesquelles la transition entre modalités se réalise sans nuance, abruptement. EXAMPLE : - Nombre des | Une **variable continue** peut, à l'inverse de la variable discrète, prendre un **nombre infini** ou non dénombrable de valeurs. Il n'y a, de ce fait, plus de modalité ou plutôt une infinité de modalités car entre deux valeurs données toutes les nuances de transitions sont possibles. Le cas « continu » ne concerne donc que les variables dites quantitatives pour lesquelles il peut y |

| | | | |
|---|---|---|---|
| EXAMPLE<br>- Couleur des fleurs<br>- Forme des graines (lisse – ridée) | souvent, souvent, parfois, rarement, jamais. | voitures.<br>- Nombre des étudiants dans la salle . | avoir autant de modalités qu'il y a d'individus.<br>EXAMPLE :<br>- Temps de vie d'une ampoule (245.546 h) |

## 2.2. Les lois de probabilité

Il est toujours possible d'associer à une variable aléatoire une probabilité et définir ainsi une loi de probabilité. Lorsque le nombre d'épreuves augmente indéfiniment, les fréquences observées pour le phénomène étudié tendent vers les probabilités et les distributions observées vers les distributions de probabilité ou loi de probabilité.

Identifier la loi de probabilité suivie par une variable aléatoire donnée est essentiel car cela conditionne le choix des méthodes employées pour répondre à une question donnée.

### 2.2.1. Lois discrètes

❖ **Loi uniforme :**

Une distribution de probabilité suit une **loi uniforme** lorsque toutes les valeurs prises par la variable aléatoire sont **équiprobables**. Si *n* est le nombre de valeurs différentes prises par la variable aléatoire,

$$\forall\, i,\ P(X = xi) = \frac{1}{n}$$

Dans le cas particulier d'une **loi discrète uniforme** où les valeurs de la variable aléatoire *X* correspondent au rang $xi = i$ ($\forall\, i \in [1, n]$)

**Espérance et variance :** $\boldsymbol{E(X) = \frac{n+1}{2}}$ $\boldsymbol{V(X) = \frac{n2-1}{12}}$

❖ **Loi de Bernoulli**

Soit un univers Ω constitué de **deux éventualités**, *S* pour succès et *E* pour échec $\Omega = \{E, S\}$

sur lequel on construit une variable aléatoire discrète, « *nombre de succès* » telle que au cours d'**un**e épreuve, si *S* est réalisé, $X = 1$ si *E* est réalisé, $X = 0$

On appelle **variable de Bernoulli** ou variable *indicatrice*, la variable aléatoire *X* telle que : $X : \Omega \rightarrow R$ $X(\Omega) = \{0,1\}$

La **loi de probabilité** associée à la variable de Bernoulli *X* telle que,

$$P(X = 0) = q$$

$$P(X = 1) = p \text{ avec } p+q = 1$$

est appelée **loi de Bernoulli notée B(1, *p*)**

**Espérance et variance ;** $E(X) = p$ $V(X) = pq$

### ❖ Loi binomiale

Décrite pour la première fois par Isaac **Newton** en 1676 et démontrée pour la première fois par le mathématicien suisse Jacob **Bernoulli** en 1713, la **loi binomiale** est l'une des distributions de probabilité les plus fréquemment rencontrées en statistique appliquée.

Soit l'application $S_n : \Omega_n \rightarrow R_n$

avec $S_n = X_1 + X_2 + \ldots + X_i + \ldots + X_n$ où $Xi$ est une variable de **Bernoulli**

La **variable binomiale**, *Sn* , représente le **nombre de succès** obtenus lors de la répétition de *n* épreuves **identiques et indépendantes**, chaque épreuve ne pouvant donner que deux résultats possibles.

Ainsi la loi de probabilité suivie par **la somme de *n* variables de Bernoulli** où la probabilité associée au succès est *p,* est la **loi binomiale** de paramètres *n* et *p*.

$$Sn : \Omega n \rightarrow IR^n$$

$$Sn = \sum_{i=1}^{n} Xi \rightarrow B(n,p)$$

La probabilité que $Sn = k$, c'est à dire l'obtention de *k* succès au cours de *n* épreuves indépendantes est :

$$P(S_n = k) = C_n^k p^k q^{n-k}$$

Il est facile de démontrer que l'on a bien une loi de probabilité car :

$$\sum_{k=0}^{n} P(S_n = k) = \sum_{k=0}^{n} C_n^k p^k q^{n-k} = (p+q)^n = 1 \quad \text{car } p+q = 1$$

**Espérance et variance : $E(Sn) = np$ $V(Sn) = npq$**

❖ **Loi de Poisson**

La **loi de Poisson** découverte au début du XIXe siècle par le magistrat français **Siméon-Denis Poisson** s'applique souvent aux phénomènes accidentels où la probabilité $p$ est très faible ($p < 0{,}05$). Elle peut également dans certaines conditions être définie comme **limite d'une loi Binomiale**.

Une variable aléatoire $X$ à valeurs dans R suit une **loi de Poisson de paramètre** $\lambda$ ($\lambda > 0$) si
les réels $p$k sont donnés par

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

on note : $X \longrightarrow P(\lambda)$

**Espérance et variance : $E(X) = \lambda$ $V(X) = \lambda$**

❖ **Loi géométrique**

. Lorsque le nombre de succès $n$ est égal à 1, la loi de la variable aléatoire discrète $X$ porte
le nom de loi de **Pascal** ou **loi géométrique** de paramètre $p$ telle que :
$P(X = k) = pq^{k-1}$ avec $k \in \mathrm{N}^*$

**Espérance et variance : $E(X) = \frac{1}{p}$ $V(X) = \frac{q}{p2}$**

**2.2.2.Lois continues :**

❖ **Loi uniforme**

. La variable aléatoire $X$ suit une **loi uniforme** sur le segment $[a,b]$ avec $a < b$ si sa
**Densité de probabilité** est donnée par :

$$f(x) = \frac{1}{b-a} \quad \text{si } x \in [a,b]$$

$$f(x) = 0 \quad \text{si } x \notin [a,b]$$

**Espérance et variance : $E(X) = \frac{b+a}{2}$ $V(X) = \frac{(b-a)2}{12}$**

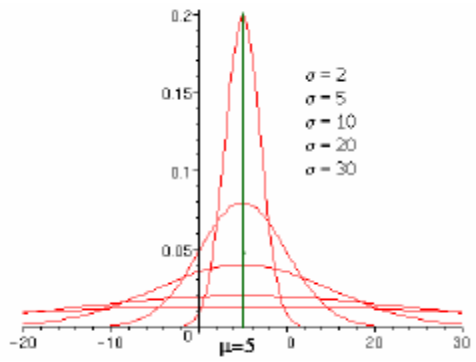
❖ **Loi normale ou loi de Laplace-Gauss.**

Une **variable aléatoire absolument continue** $X$ suit une loi normale de paramètres ($\mu$ , $\sigma$) si
sa **densité de probabilité** est donnée par : $f$ : R → R

$$x \mapsto f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad \text{avec } \mu \in \mathbb{R} \text{ et } \sigma \in \mathbb{R}^+$$

*Notation :* $X \to \mathcal{N}(\mu , \sigma)$

| $x$ | $\mu$ | $\mu + \sigma$ | $+\infty$ |
|---|---|---|---|
| $f''(x)$ | - | 0 | + |
| $f'(x)$ | 0 | - | |
| $f(x)$ | $\frac{1}{\sigma\sqrt{2\pi}}$ | $\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}}$ | 0 |



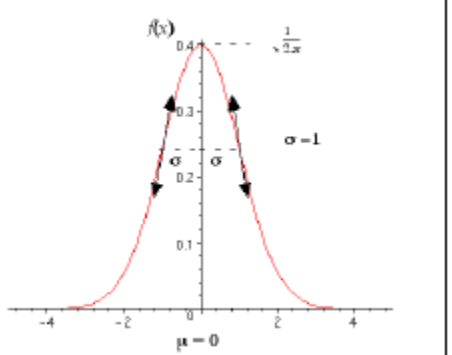

**Espérance et variance : $E(X) = \mu$ $V(X) = \sigma^2$**

### ❖ Loi normale réduite

Une variable aléatoire continue $X$ suit une **loi normale réduite** si sa densité de probabilité
est donnée par : $f$ : R → R

$$x \mapsto f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

| $x$ | 0 | 1 | $+\infty$ |
|---|---|---|---|
| $f''(x)$ | + | 0 | - |
| $f'(x)$ | 0 | - | |
| $f(x)$ | $\frac{1}{\sqrt{2\pi}}$ | $\frac{1}{\sqrt{2\pi e}}$ | 0 |




**Espérance et variance : $E(X) = 0$ $V(X) = 1$**

### ❖ Loi du $\chi^2$ de Pearson :

La **loi de Pearson ou loi de $\chi^2$ (Khi deux)** trouve de nombreuses applications dans le cadre de la comparaison de proportions, des tests de conformité d'une distribution observée à une distribution théorique et le

test d'indépendance de deux caractères qualitatifs. Ce sont le test **du khi-deux**.

Soit $X_1$, $X_2$, …, $X_i$,…, $X_n$, $n$ variables **normales centrées réduites**, on appelle $\chi^2$ la variable aléatoire définie par :

$$\chi^2 = X_1^2 + X_2^2 + \ldots + X_i^2 + \ldots + X_n^2 = \sum X_i^2$$

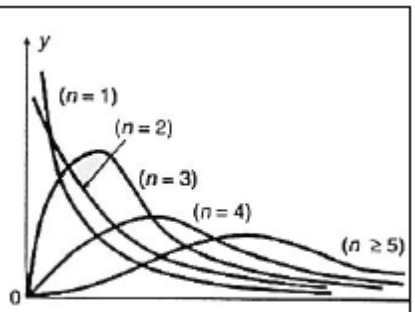
On dit que $\chi 2$ suit une **loi de Pearson** à $n$ **degrés de liberté** (d.d.l.).

Pour $\chi^2 > 0$, la fonction densité de probabilité est de la forme :

$$f(\chi^2) = C(n)\chi^{\frac{n}{2}-1}e^{-\frac{\chi^2}{2}} \text{ avec } C(1) = \frac{1}{\sqrt{2\pi}}$$

Pour $\chi^2 \leq 0$, $f(\chi^2) = 0$

Pour $n > 1$, on utilise la table du Khi 2




**Espérance et variance : $E(X) = n$** **$V(X) = 2n$**

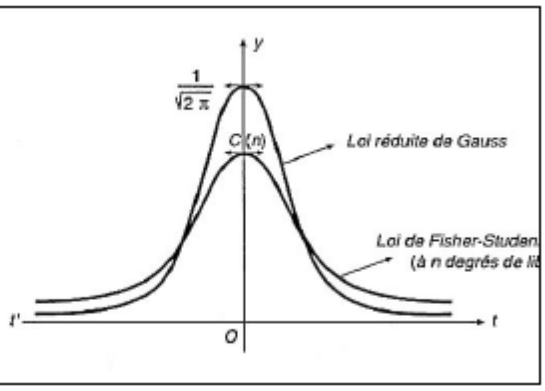
### ❖ Loi de student

Soit $U$ une variable aléatoire suivant une **loi normale réduite** N(0,1) et $V$ une variable aléatoire suivant une **loi de Pearson** à $n$ degrés de liberté $\chi 2n$, $U$ et $V$ étant **indépendantes**, on dit alors que $T_n = U/\sqrt{V/n}$ suit une **loi de Student** à $n$ degrés de liberté.

La fonction densité de probabilité est de la forme :

$$f(T) = C(n)\left(1+\frac{T^2}{n}\right)^{-\frac{(n+1)}{2}}$$

Calcul des probabilités avec la table de la loi de Student




**Espérance et variance : $E(T) = 0$ si $n>1$** **$V(T) = \frac{n}{n-2}$ si $n>2$**

### ❖ Loi de *Fisher-Snedecor*

La **loi de Fisher-Snedecor** est utilisée pour comparer deux variances observées et sert surtout dans les très nombreux tests d'analyse de variance et de covariance.

Soit $U$ et $V$ deux variables aléatoires indépendantes suivant une **loi de Pearson** respectivement à $n$ et $m$ degrés de liberté.

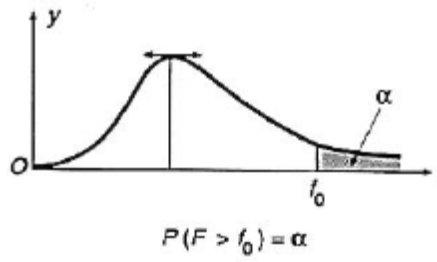
On dit que $F = \frac{U/n}{V/m}$ suit une loi de **Fisher-Snedecor** à $\begin{Bmatrix} n \\ m \end{Bmatrix}$ degrés de liberté.

Pour $F > 0$, la fonction densité de probabilité est de la forme :

$f(F) = C_{(n,m)} F^{\frac{n}{2}-1} (m+nF)^{-\left(\frac{n+m}{2}\right)}$

Pour $F \leq 0, f(F) = 0$

Utilisation des tables de Fisher-Snedecor pour le calcul des probabilités.




**Espérance et variance : $E(F) = \frac{m}{m-2}$ *si m>2*** **$V(F) = \frac{2m2(n+m-2)}{n(m-2)2(m-4)}$ *si m>4***

## 3. Analyse bivariée : mesure des liaisons entre deux variables

### 3.1. Présentation générale

Lorsque deux phénomènes ont une évolution commune, nous disons qu'ils sont «corrélés». La corrélation simple mesure le degré de liaison existant entre ces deux phénomènes représentés par des variables.

Si nous cherchons une relation entre trois variables ou plus, nous ferons appel alors à la notion de corrélation multiple.

Nous pouvons distinguer la corrélation linéaire, lorsque tous les points du couple de valeurs *(x,y)* des deux variables semblent alignés sur une droite, de la corrélation non linéaire lorsque le couple de valeurs se trouve sur une même courbe d'allure quelconque.

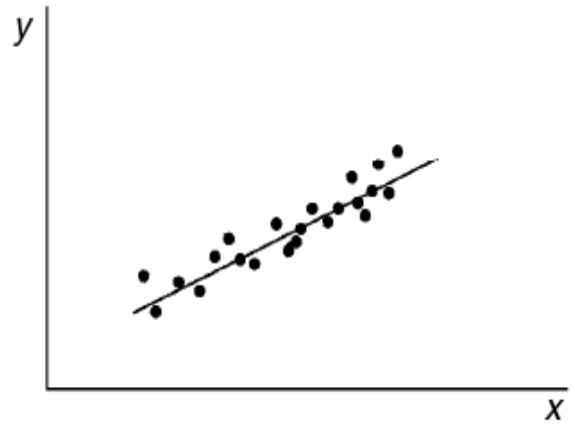
Deux variables peuvent être :

- en corrélation positive ; on constate alors une augmentation (ou diminution, ou constance) simultanée des valeurs des deux variables ;
- en corrélation négative, lorsque les valeurs de l'une augmentent, les valeurs de l'autre diminuent ;
- non corrélées, il n'y a aucune relation entre les variations des valeurs de l'une des variables et les valeurs de l'autre.

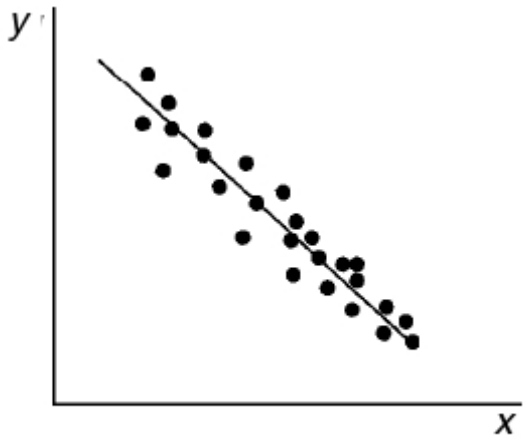
Le tableau 1, en croisant les critères de linéarité et de corrélation, renvoie à une représentation graphique.

***Tableau 1 – Linéarité et corrélation***

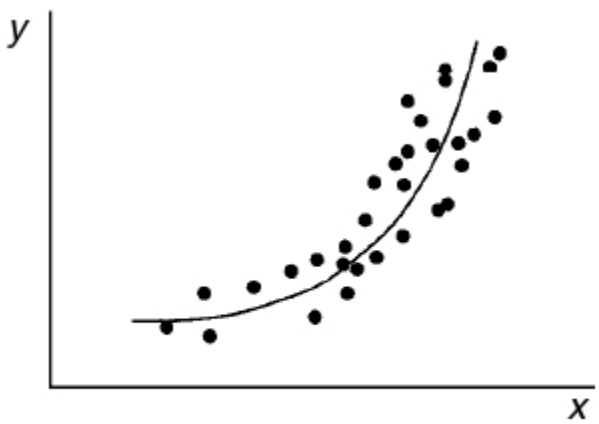
| | ***Corrélation positive*** | ***Corrélation négative*** | ***Absence de Corrélation*** |
|---|---|---|---|
| ***Relation linéaire*** | ***Graphe 1*** | ***Graphe 2*** | ***Graphe 5*** |
| ***Relation non linéaire*** | ***Graphe 3*** | ***Graphe 4*** | ***Graphe 5*** |



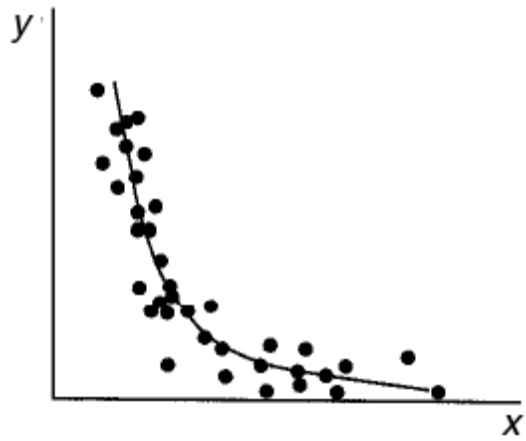

*Graphe 1*



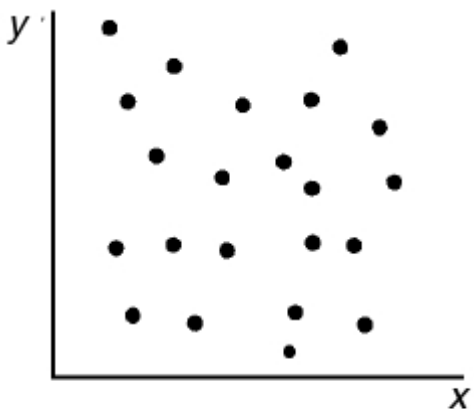

*Graphe 2*




*Graphe 3*




*Graphe 4*




*Graphe 5*

## 3.2. Mesure et limite du coefficient de corrélation

### 3.2.1.Le coefficient de corrélation linéaire

La représentation graphique ne donne qu'une « impression » de la corrélation entre deux variables sans donner une idée précise de l'intensité de la liaison, c'est pourquoi nous calculons une statistique appelée *coefficient de corrélation linéaire simple*, noté $r_{x,y}$ *ou bien* $p_{xy}$. Il est égal à :

$$r_{x,y} = \frac{\text{Cov}\ (x,y)}{\sigma_x\ \sigma_y} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

avec :

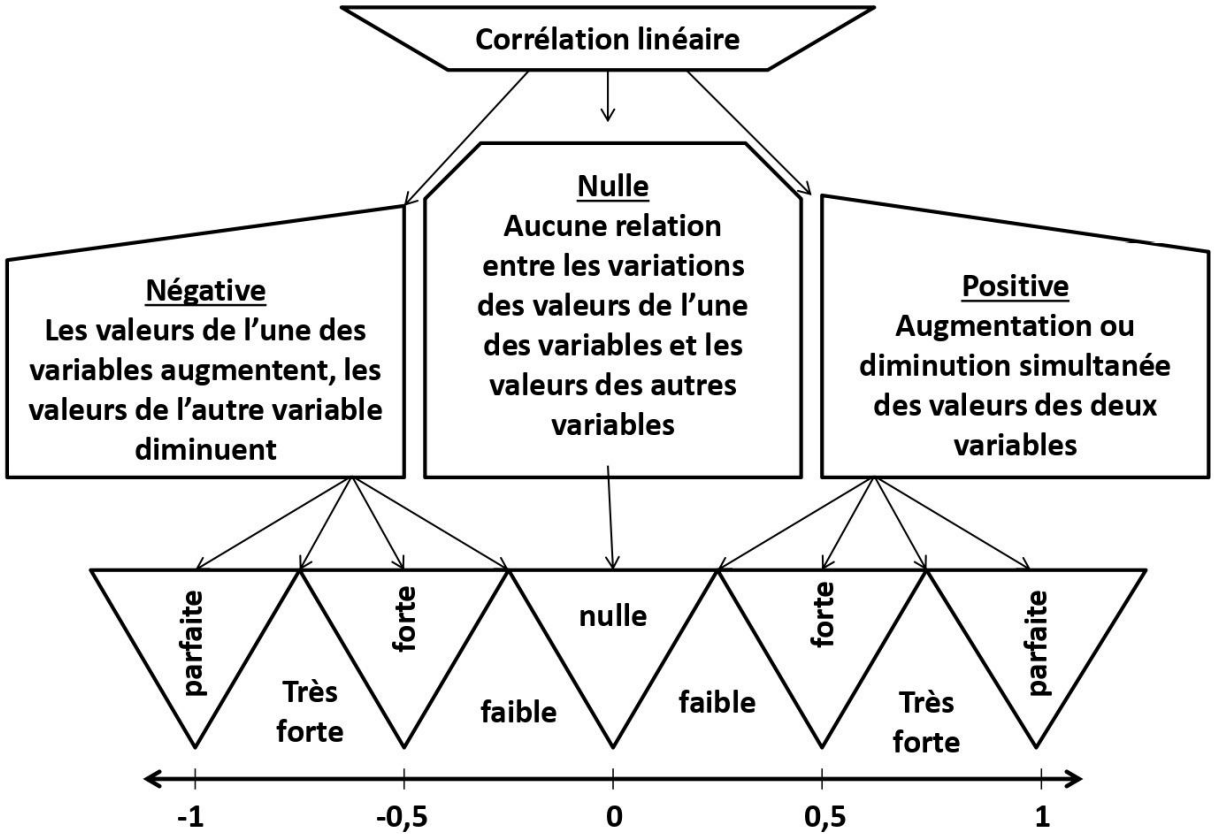
- Cov *(x,y)* = covariance entre *x* et *y* ;
- $\sigma x$ et $\sigma y$ = écart type de *x* et écart type de *y* ;
- *n* = nombre d'observations.

En développant la formule précédente, il vient :

$$\rho = \rho_{XY} = \frac{\sum_{t=1}^{n}(x_t - \overline{x})(y_t - \overline{y})}{\sqrt{\sum_{t=1}^{n}(x_t - \overline{x})^2}\sqrt{\sum_{t=1}^{n}(y_t - \overline{y})^2}} = \frac{\sum_{t=1}^{n}x_t y_t - n\,\overline{x}\,\overline{y}}{\sqrt{\sum_{t=1}^{n}x_t^2 - n\,\overline{x}^2}\sqrt{\sum_{t=1}^{n}y_t^2 - n\,\overline{y}^2}}$$

On peut démontrer que, par construction ce coefficient reste compris entre −1 et 1 :

- proche de 1, les variables sont corrélées positivement ;
- proche de −1, les variables sont corrélées négativement ;
- proche de 0, les variables ne sont pas corrélées.




### 3.3. Test de significative de coefficient de corrélation :

Dans la pratique, le coefficient est rarement très proche de l'une de ces trois bornes et il est donc difficile de proposer une interprétation fiable à la simple lecture de ce coefficient.

Ceci est surtout vrai en économie où les variables sont toutes plus au moins liées entre elles. De plus, il n'est calculé qu'à partir d'un échantillon d'observations et non pas sur l'ensemble des valeurs. On appelle $\rho_{x,y}$ ce coefficient empirique qui est une estimation du coefficient vrai $r_{x,y}$ . La théorie des tests statistiques nous permet de lever cette indétermination.

Soit à tester l'hypothèse $H_0$ : $p_{x,y} = 0$ , contre l'hypothèse $H_1$ : $p_{x,y} \neq 0$ . Sous l'hypothèse $H_0$, nous pouvons démontrer que $\frac{\rho x,y}{\sqrt{\frac{(1-\rho 2x,y)}{n-2}}}$ suit une loi de Student à ($n$ – 2) degrés de liberté. Nous calculons alors une statistique, appelé le $t$ de Student empirique :

$$t^* = \frac{|\rho_{x,y}|}{\sqrt{\frac{(1-\rho_{x,y}^2)}{n-2}}}$$

Si $t* > t^{\alpha/2}_{n-2}$ *valeur* lue dans une table de Student2 au seuil $\alpha$ = 0,05 (5 %) à ($n$–2) degrés de liberté[1], nous rejetons l'hypothèse H0, le coefficient de corrélation est donc significativement différent de 0 ; dans le cas contraire, l'hypothèse d'un coefficient de corrélation nul est acceptée. La loi de Student étant symétrique, nous calculons la valeur absolue du $t$ empirique et nous procédons au test par comparaison avec la valeur lue directement dans la table.

**Exercice :**

– **Calcul d'un coefficient de corrélation**

Un agronome s'intéresse à la liaison pouvant exister entre le rendement de maïs $x$ (en quintal) d'une parcelle de terre et la quantité d'engrais $y$ (en kilo). Il relève 10 couples de données consignés dans le tableau 2

***Tableau 2 – Rendement de maïs et quantité d'engrais***

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Rendement $x$ | 16 | 18 | 23 | 24 | 28 | 29 | 26 | 31 | 32 | 34 |
| Engrais $y$ | 20 | 24 | 28 | 22 | 32 | 28 | 32 | 36 | 41 | 41 |

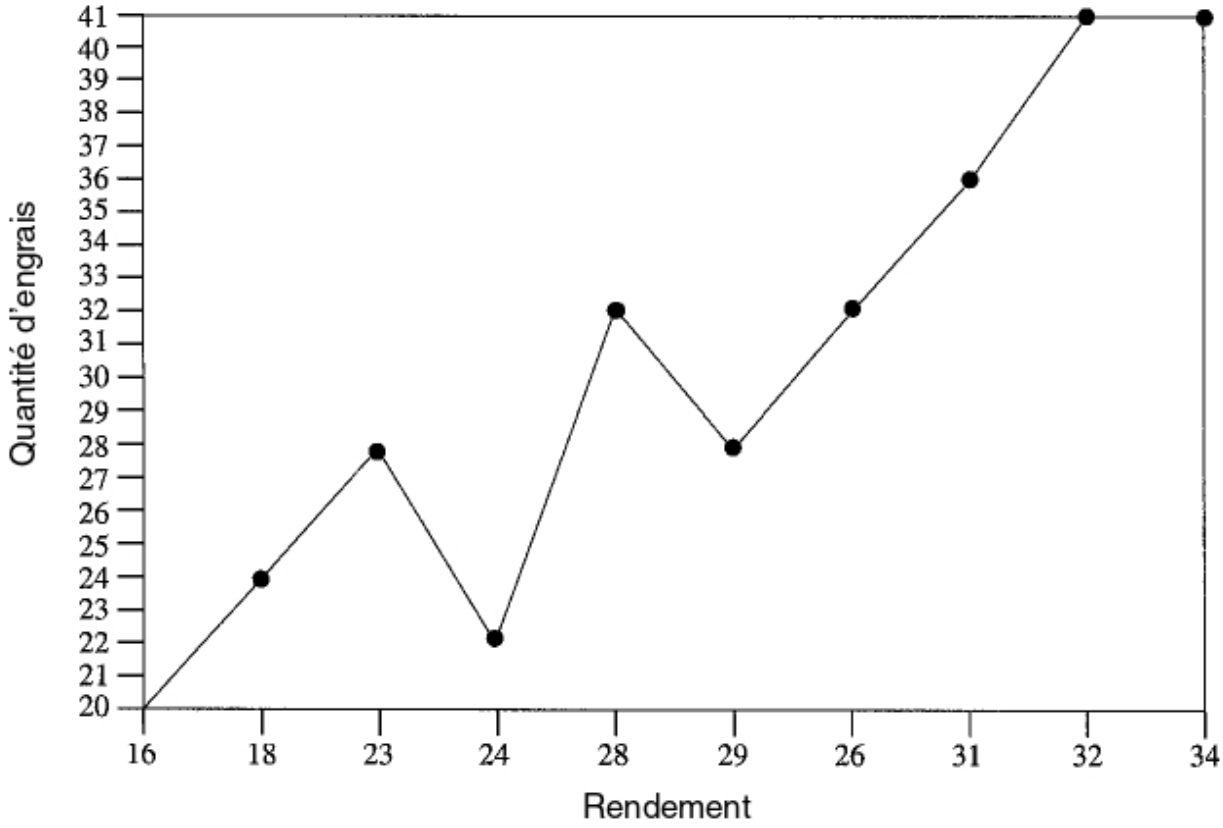
**1)** Tracer le nuage de points et le commenter.

**2)** Calculer le coefficient de corrélation simple et tester sa signification par rapport à 0, pour un seuil $\alpha$ = 0,05.

– **Solution**

**1)** Le nuage de points (graphique 1) indique que les couples de valeurs sont approximativement alignés : les deux variables semblent corrélées positivement.

***Graphique 1 – Nuage du couple de valeurs : rendement-quantité d'engrais***




**2)** Afin d'appliquer la formule, nous dressons le tableau de calcul 3.

***Tableau 3 – Calcul d'un coefficient de corrélation***

| | $x$ | $y$ | $x^2$ | $y^2$ | $xy$ |
|---|---|---|---|---|---|
| | 16 | 20 | 256 | 400 | 320 |
| | 18 | 24 | 324 | 576 | 432 |
| | 23 | 28 | 529 | 784 | 644 |
| | 24 | 22 | 576 | 484 | 528 |
| | 28 | 32 | 784 | 1 024 | 896 |
| | 29 | 28 | 841 | 784 | 812 |
| | 26 | 32 | 676 | 1 024 | 832 |
| | 31 | 36 | 961 | 1 296 | 1 116 |
| | 32 | 41 | 1 024 | 1 681 | 1 312 |
| | 34 | 41 | 1 156 | 1 681 | 1 394 |
| Somme | 261 | 304 | 7 127 | 9 734 | 8 286 |

$$\rho_{x,y} = \frac{(10)(8\,286) - (261)(304)}{\sqrt{(10)(7\,127) - 261^2}\,\sqrt{(10)(9\,734) - 304^2}} = \frac{3\,516}{(56,11)(70,17)}$$

soit $\rho_{x,y} = 0,89$ et $\rho^2_{x,y} = 0,79$

Le $t$ de Student empirique (d'après [3]) est égal à :

$$t^* = \frac{|\rho_{x,y}|}{\sqrt{\frac{(1-\rho^2_{x,y})}{n-2}}} = \frac{0,89}{0,1620} = 5,49 > t_8^{0,025} = 2,306$$

Le coefficient de corrélation entre *x* et *y* est significativement différent de 0.

### 3.4. Limites de la notion de corrélation

- **La relation testée est linéaire**

L'application de la formule [1] ou [2] ne permet de déterminer que des corrélations linéaires entre variables. Un coefficient de corrélation nul indique que la covariance entre la variable *x* et la variable *y* est égale à 0. C'est ainsi que deux variables en totale dépendance peuvent avoir un coefficient de corrélation nul.

- **Corrélation n'est pas causalité**

Le fait d'avoir un coefficient de corrélation élevé entre deux variables ne signifie pas qu'il existe un autre lien que statistique. En d'autres termes, une covariance significativement différente de 0 n'implique pas une liaison d'ordre économique, physique ou autre. Nous appelons *corrélation fortuite* ce type de corrélation que rien ne peut expliquer.

L'exemple le plus fameux concerne la forte corrélation existante entre le nombre de taches solaires observées et le taux de criminalité aux États-Unis.

Cela ne signifie pas qu'il existe une relation entre les deux variables, mais qu'une troisième variable, l'évolution de long terme (la tendance) ici, explique conjointement les deux phénomènes. La théorie de la cointégration traite de ce problème.

## 4. Le modèle de régression linéaire simple

### 4.1. Concepts et définition

**La régression linéaire** est une modélisation linéaire qui permet d'établir des estimations dans le futur à partir d'informations provenant du passé. Dans ce modèle de régression linéaire, on a plusieurs variables dont une qui est une variable explicative et les autres qui sont des variables expliquées. Cet outil est utilisé pour les analyses techniques boursières mais aussi pour la gestion de budgets. Elle est souvent calculée avec la méthode des moindres carrés qui permet de réduire les erreurs en ajoutant de l'information.

**Les objectifs de la régression linéaire simple**

- Description d'une éventuelle relation de cause à effet entre deux variables (études non-expérimentales) ;
- Explications et confrontations des hypothèses en se basant sur des études expérimentales contrôlées ;
- Prédiction d'une variable à partir de l'autre.

**Exemple Introductif**

Soit la fonction de consommation Keynésienne :

$$C = a_0 + a_1 R$$

Avec,

- C : Consommation par habitant
- R : revenu
- *a1 :* propension marginale à consommer
- *a0 :* consommation autonome ou incompressible

On a :

La consommation (C) est une variable « à expliquer » et le revenu R est une variable « explicative ». $a_1$ et $a_0$ sont les paramètres du modèle ou coefficients de la régression linéaire simple.

**Spécification**

**Modèle en série temporelle :** par exemple, la consommation et le revenu annuel pour *l'algerie* de 2000 à 2013.

$$C_t = a_0 + a_1 R_t \qquad t = 2000, \ldots, 2013$$

C*t* : Consommation au temps *t* .

R*t* : revenu au temps *t* .

**Modèle en coupe instantanée :** par exemple, la consommation et le revenu pour 15 pays en 2013 (date fixe)

$$C_i = a_0 + a_1 R_i \qquad i = 1, \ldots, 15$$

C*i* : Consommation relative au payé *i* en 2013

R*i* : revenu relatif au payé *i* en 2013

**Modèle en panel :** par exemple, la consommation et le revenu pour 15 pays de 2000 à 2013

$$C_{i,t} = a_0 + a_1 R_{i,t} \qquad i = 1, \ldots, 15 \; ; t = 2000, \ldots, 2013$$

### 4.2. Rôle du terme aléatoire

Le revenu est-il *l'unique* variable explicative de la consommation ?

**Sûrement NON !**

d'où, l'ajout du terme $\varepsilon$ qui résumera toutes les fluctuations non observables attribuables à un ensemble de facteurs ou de variables non prises en compte dans le modèle :

$$C_t = a_0 + a_1 R_t + \varepsilon_t \quad ou \quad C_i = a_0 + a_1 R_i + \varepsilon_i$$

La variable aléatoire $\varepsilon t$ (ou $\varepsilon i$) regroupe trois types d'erreur :

– Erreur de spécification

– Erreur de mesure

– Erreur de fluctuation d'échantillonnage

## Exemple Introductif

Le tableau 1 présente le revenu moyen par habitant sur 10 ans exprimé en dollars pour un pays.

***Tableau 4 – Évolution du revenu moyen par habitant en dollars***

| | *Revenue* |
|---|---|
| *1* | ***8000*** |
| *2* | ***9000*** |
| *3* | ***9500*** |
| *4* | ***9500*** |
| *5* | ***9800*** |
| *6* | ***11000*** |
| *7* | ***12000*** |
| *8* | ***13000*** |
| *9* | ***15000*** |
| *10* | ***16000*** |

*Sachant que la propension marginale à consommer est de 0,8 et que la consommation incompressible est 1 000, on demande :*
*1) de calculer la consommation théorique sur les 10 ans ;*
*2) considérant que notre erreur d'observation suit une loi normale de moyenne 0 et de variance 20 000, de générer cette variable aléatoire et de calculer une consommation observée tenant compte de cette erreur.*
*Solution*
*Les calculs des questions 1) et 2) sont présentés dans le tableau 5.*
*La consommation théorique (colonne 3) (Tableau 5) est calculée par application directe de la formule : Ct = 1 000 + 0,8 Yt .*
*La génération de la variable aléatoire $\varepsilon t$ ( $\varepsilon t \rightarrow N(0 ; 20\,000)$) ne pose pas de difficulté particulière ; bien entendu il en existe une infinité, un exemple en est présenté en colonne 4.*
*La consommation « observée » (colonne 5) (Tableau 5) est donc égale à Ct = 1 000 +0,8 Yt + $\varepsilon t$,*
*soit la somme de la colonne 3 et de la colonne 4 (Tableau 5).*

***Tableau 5 – Calcul de la consommation observée***

| *(1) années* | *(2) revenue disponible* | *(3) consommation théorique* | *(4) alea $\varepsilon_t$* | *(5) consommation observée* |
|---|---|---|---|---|
| *1* | *8000* | *7400* | *-10,01* | *7389,99* |
| *2* | *9000* | *8200* | *-30,35* | *8169,65* |
| *3* | *9500* | *8600* | *231,71* | *8831,71* |
| *4* | *9500* | *8600* | *52,84* | *8652,84* |
| *5* | *9800* | *8840* | *-51,92* | *8788,08* |
| *6* | *11000* | *9800* | *-183,79* | *9616,21* |
| *7* | *12000* | *10600* | *-6,55* | *10593,45* |
| *8* | *13000* | *11400* | *-213,89* | *11186,11* |
| *9* | *15000* | *13000* | *-241,91* | *12758,09* |
| *10* | *16000* | *13800* | *69,62* | *13869,62* |
| | | *moyenne :* | *-38,425* | |
| | | *Ecart tupe:* | *137,248601* | |

Nous observons que la moyenne de $\varepsilon_t$, $\bar{\varepsilon} = -38{,}42$ et la variance de $\varepsilon_t$, Var $(\varepsilon_t)$ =18 834,81 sont légèrement différentes des valeurs théoriques.

Cela est la conséquence du tirage particulier d'un échantillon de taille assez faible (dix observations).

❖ **Conséquence du terme aléatoire**

En général, les coefficients *a*0 et *a*1 sont inconnues et on les estime par échantillonnage. On pose :

$$\hat{a}_0 \text{ estimateur de } a_0$$
$$\hat{a}_1 \text{ estimateur de } a_1$$

$\hat{a}_0$ *et* $\hat{a}_{1\ sont}$ des variables aléatoires ***qui suivent les mêmes loi de probabilité*** que celle de $\varepsilon$t (les erreurs sont supposées indépendantes et identiquement distribuées par une loi normale)

### 4.3. Modèle et hypothèses

Le modèle théorique de régression simple s'écrit : pour $t = 1, \ldots, n$

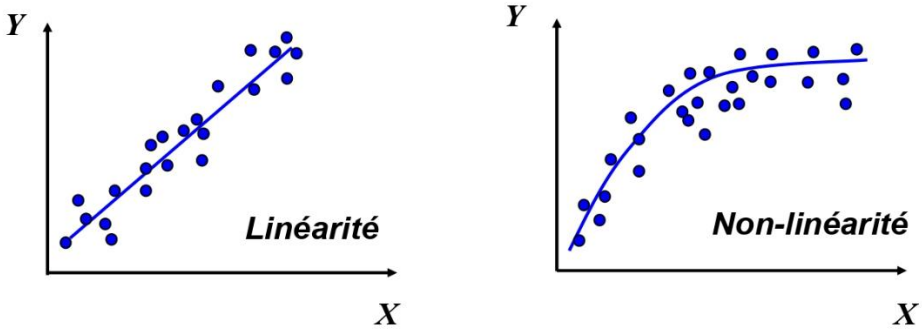
$$y_t = a_0 + a_1 x_t + \varepsilon_t$$

| | | |
|---|---|---|
| *n* | : | Nombre d'observations (taille de l'échantillon) |
| *yt* | : | Variable à expliquer au temps *t*, variable dépendante ou variable endogène. Elle est entachée d'une erreur additive εt |
| *xt* | : | Variable certaine explicative au temps *t*, variable indépendante ou variable exogène |
| *a1* | : | Paramètre du modèle, c'est le *coefficient de régression.* Il représente la pente de la droite (variation de *Y* due à une variation unitaire de *X*) |
| *a0* | : | Paramètre du modèle, c'est *l'ordonnée à l'origine.* |
| εt | : | Erreur de spécification de nature aléatoire et inconnue (différence entre le modèle vrai et le modèle spécifié), appelée encore *bruit blanc* ou *facteur de perturbation* cette erreur et restera inconnue. |

Les hypothèses suivantes permettent de déterminer les estimateurs des coefficients du modèle ayant de bonnes propriétés et de construire des tests statistiques (tests et intervalles de confiance).

**(H1) : Le modèle est linéaire en $X_t$ ou** f($x_t$)

On suppose l'existence d'une relation linéaire entre *X* et *Y*




**(H2) :Les valeurs sont observées sans erreur ( non aléatoire)**

$X_t$ est certaine et connue sans erreur : $x_t$ *mesurée* = $x_t$ *vraie*

**(H3) : l'espérance mathématique de l'erreur est nulle.**

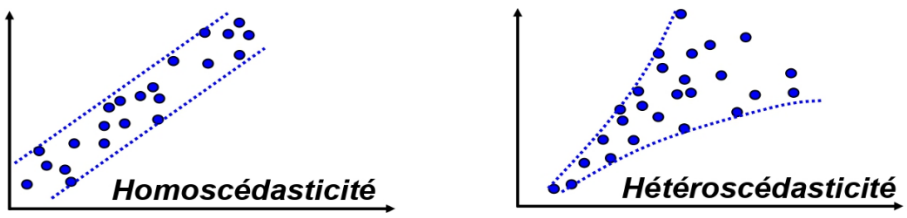
L'hypothèse principale $E(\varepsilon_t) = 0$

$y_t$ est entachée d'une erreur additive $y_t$ mesurée = $y_t$ vrais + $\varepsilon_t$

Mais en moyenne le modèle est bien spécifié et donc l'erreur moyenne est nulle

**(H4) : Hypothèse d'homoscédasticité**

$$E(\varepsilon_t^2) = \sigma_\varepsilon^2 = cste$$




La variance de l'erreur est constante : le risque de l'amplitude de l'erreur est le même quelle que soit la période.

**(H5) : Non-auto corrélation des erreurs**

$$E(\varepsilon_t \varepsilon_{t'}) = 0 \quad si\ t \neq t'$$

Les erreurs sont non corrélées (ou encore indépendantes). Une erreur à l'instant *t* n'a pas d'influence sur les erreurs suivantes.

**(H6) : l'erreur est indépendante de la variable explicative**

$$Cov(x_t, \varepsilon_t) = 0$$

**(H7) Supplémentaire : Normalité des erreurs**

$$\varepsilon_t \equiv N(0, \sigma_\varepsilon^2)$$

### 4.4. Différentes écritures du modèle (erreur & résidu)

- Modèle *théorique* spécifié par *un économiste* avec **l'erreur** inconnue $\varepsilon_t$ : $y_t = a_0 + a_1 x_t + \varepsilon_t$

- Modèle *empirique* estimé par *un économètre* à partir d'un échantillon d'observations avec **le résidu** observé : $y_t = \hat{a}_0 + \hat{a}_1 x_t + e_t = \hat{y}_t + e_t$

Avec $\hat{a}_0$ *et* $\hat{a}_1$ sont les estimateurs de $a_0$ et $a_1$ respectivement.

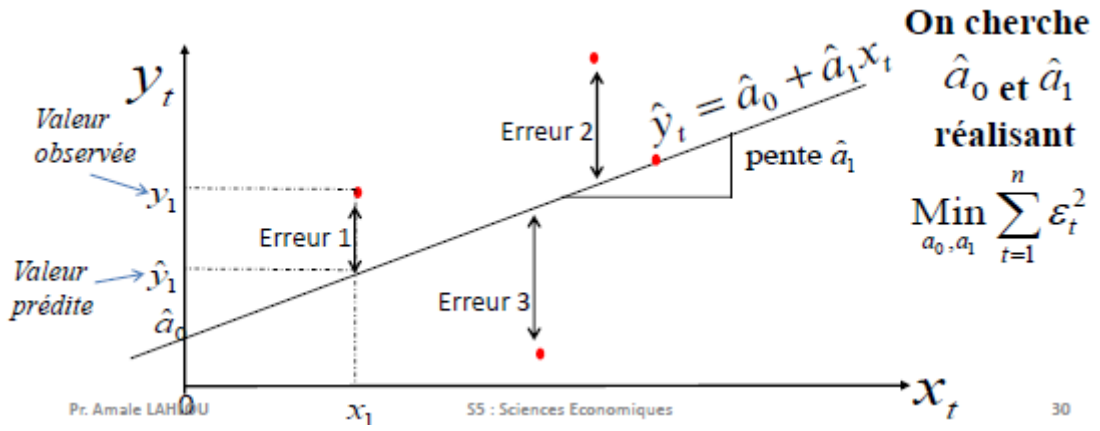
**N.B. :**

« erreur » est définie dans la spécification du modèle théorique ;

« résidu » est définie comme erreur observée sur les données.

### 4.5. Formulation des estimateurs

**Méthode des Moindres Carrés Ordinaires** (MCO) (on peut utiliser la méthode de maximum de vraisemblance) : à l'aide des données de l'échantillon nous estimerons les paramètres $a_0$ et $a_1$ du modèle de régression de façon à minimiser la somme des carrés des erreurs (des différences entre les valeurs observées $y_t$ et les valeurs estimées $\hat{y}_t$ par la droite) :




Pour déterminer les estimateurs des Moindres Carrés Ordinaires on doit minimiser analytiquement la quantité

$$\text{Min} \sum_{t=1}^{t=n} \varepsilon_t^2 = \text{Min} \sum_{t=1}^{t=n} (y_t - a_0 - a_1 x_t)^2 = \text{Min}\, S$$

Le coefficient représentant la pente de la droite ou la propension marginale est donné par :

$$\hat{a}_1 = \frac{\sum_{t=1}^{n} x_t y_t - n\bar{x}\,\bar{y}}{\sum_{t=1}^{n} x_t^2 - n\bar{x}^2} = \frac{\sum_{t=1}^{n} (x_t - \bar{x})(y_t - \bar{y})}{\sum_{t=1}^{n} (x_t - \bar{x})^2} = \frac{Cov(x,y)}{Var(x)}$$

Le coefficient représentant l'ordonnée à l'origine est donné par :

$$\hat{a}_0 = \bar{y} - \hat{a}_1 \bar{x}$$

**Exemple d'application**

Détermination des estimations des paramètres $\hat{a}_0$ et $\hat{a}_1$ de la droite de régression

| *années* | *($x_t$) revenue disponible* | *($y_t$) consommation théorique* |
|---|---|---|
| *1* | *8000* | *7400* |
| *2* | *9000* | *8200* |
| *3* | *9500* | *8600* |
| *4* | *9500* | *8600* |
| *5* | *9800* | *8840* |
| *6* | *11000* | *9800* |
| *7* | *12000* | *10600* |
| *8* | *13000* | *11400* |
| *9* | *15000* | *13000* |
| *10* | *16000* | *13800* |

Solution :

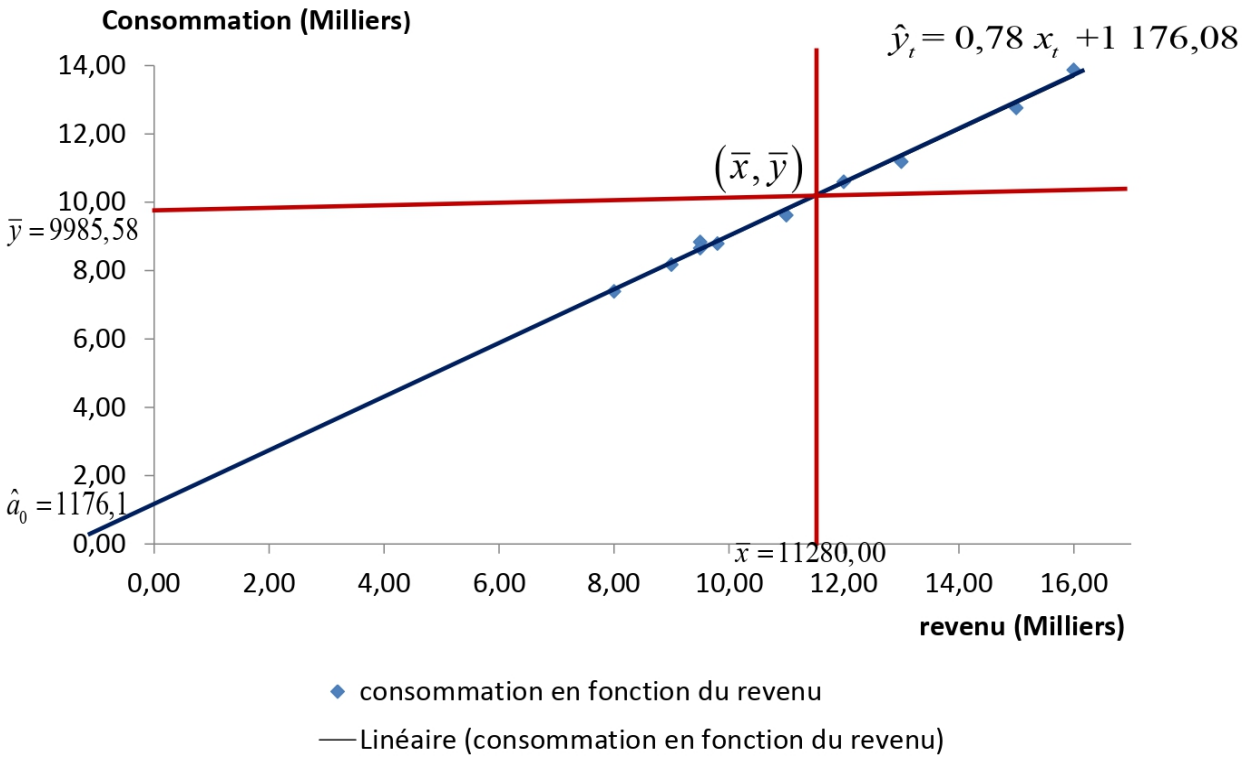
| *années* | *($x_t$)* | *($y_t$)* | *($x_t$-$\overline{x}$)* | *($y_t$-$\overline{y}$)* | *($x_t$-$\overline{x}$)²* | *($x_t$-$\overline{x}$)($y_t$-$\overline{y}$)* |
|---|---|---|---|---|---|---|
| *1* | *8000* | *7389,99* | *-3280* | *-2595,585* | *10758400* | *8513518,8* |
| *2* | *9000* | *8169,65* | *-2280* | *-1815,925* | *5198400* | *4140309* |
| *3* | *9500* | *8831,71* | *-1780* | *-1153,865* | *3168400* | *2053879,7* |
| *4* | *9500* | *8652,84* | *-1780* | *-1332,735* | *3168400* | *2372268,3* |
| *5* | *9800* | *8788,08* | *-1480* | *-1197,495* | *2190400* | *1772292,6* |
| *6* | *11000* | *9616,21* | *-280* | *-369,365* | *78400* | *103422,2* |
| *7* | *12000* | *10593,45* | *720* | *607,875* | *518400* | *437670* |
| *8* | *13000* | *11186,11* | *1720* | *1200,535* | *2958400* | *2064920,2* |
| *9* | *15000* | *12758,09* | *3720* | *2772,515* | *13838400* | *10313755,8* |
| *10* | *16000* | *13869,62* | *4720* | *3884,045* | *22278400* | *18332692,4* |
| *somme* | *112800* | *99855,75* | *0* | *0* | *64156000* | *50104729* |
| *moyenne* | ***11280*** | ***9985,575*** | ***0*** | ***0*** | | |

Calcul des estimations de $\hat{a}_0$ et $\hat{a}_1$

$$\hat{a}_1 = \frac{\sum_{t=1}^{10}(x_t - \overline{x})(y_t - \overline{y})}{\sum_{t=1}^{10}(x_t - \overline{x})^2} = \frac{50\ 104\ 729{,}00}{64\ 156\ 000{,}00} = 0{,}78$$

$$\hat{a}_0 = \overline{y} - \hat{a}_1\overline{x} = 9\ 985{,}58 - 0{,}78 \times 11\ 280{,}00 = 1\ 176{,}08$$

## Nuage de points et droite de régression linéaire




| | *Coefficients* |
|---|---|
| Constante | 1176,089634 |
| Variable X 1 | 0,780982745 |

Par calcul,

$\hat{a}_1 = 0{,}78$

$\hat{a}_0 = 1\ 176{,}08$

Ainsi, on peut alors prédire $y_t$ pour $x_t$ compris dans l'intervalle des valeurs de l'échantillon :

$$\hat{y}_t = \hat{a}_0 + \hat{a}_1\ x_t = 1\ 176{,}08 + 0{,}78\ x_t$$

**Extrait du rapport détaillé par une analyse sous EVIEWS.**

Dependent Variable: CONSOMMATION
Method: Least Squares
Date: 01/12/23 Time: 19:19
Sample: 1 10
Included observations: 10

| Variable | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| **REVENUE ($a_1$)** | **0.780983** | 0.017939 | 43.53518 | 0.0000 |
| **C ($a_0$)** | **1176.090** | 207.3921 | 5.670852 | 0.0005 |

| | | | |
|---|---|---|---|
| R-squared | 0.995797 | Mean dependent var | 9985.575 |
| Adjusted R-squared | 0.995271 | S.D. dependent var | 2089.553 |
| S.E. of regression | 143.6878 | Akaike info criterion | 12.95002 |
| Sum squared resid | 165169.4 | Schwarz criterion | 13.01054 |
| Log likelihood | -62.75009 | Hannan-Quinn criter. | 12.88363 |
| F-statistic | 1895.312 | Durbin-Watson stat | 1.881665 |
| Prob(F-statistic) | 0.000000 | | |

CONSOMMATION = 0.780982745184*REVENUE + 1176.08963433

Les estimateurs obtenus par la méthode des Moindres Carrés Ordinaires sont des estimateurs **linéaires** non biaisés convergents à variance minimale c'est-à-dire efficaces (**B**est **L**ineair **U**nbiased **E**stimators)

## 4.6. Les indicateur de calcul la qualité du modèle :

### 4.6.1.Coefficient de détermination $R^2$:

$R^2$ est un indicateur de la qualité de l'ajustement de la droite aux données. Autrement dit, il mesure l'adéquation entre le modèle et les données observées. Il nous indique le pourcentage de l'information restituée par le modèle par rapport à la qualité d'information initiale.

$$R^2 = \frac{SCE}{SCT} = \frac{\sum_{t=1}^{n}(\hat{y}_t - \bar{y})^2}{\sum_{t=1}^{n}(y_t - \bar{y})^2}$$

$$\mathbf{0 \leq R^2 \leq 1}$$

$$R^2 = 1 - \frac{SCR}{SCT} = 1 - \frac{\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}{\sum_{t=1}^{n}(y_t - \bar{y})^2} = 1 - \frac{\sum_{t=1}^{n} e_t^2}{\sum_{t=1}^{n}(y_t - \bar{y})^2}$$

Dans une régression, la somme totale des carrés permet d'exprimer la variation totale des y. Par exemple, vous collectez des données afin de déterminer un modèle expliquant les ventes totales en fonction de votre budget publicitaire.

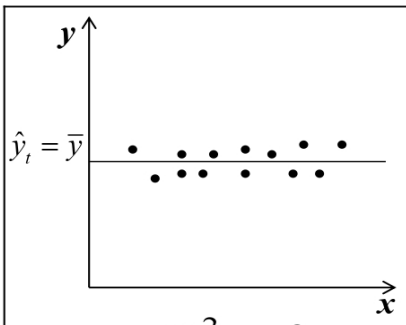
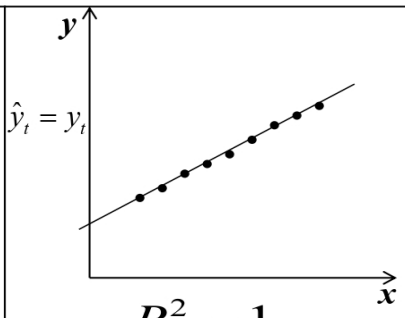
**Somme totale de carrés (SCT) = somme des carrés de la régression (SCE) + somme des carrés de l'erreur résiduelle (SCR)**

La somme des carrés de la régression est la variation attribuée à la relation entre les x et les y, en l'occurrence entre le budget publicitaire et les ventes. La somme des carrés de l'erreur résiduelle est la variation attribuée à l'erreur.

La comparaison de la somme des carrés de la régression à la somme totale des carrés indique la proportion de la variation totale expliquée par le modèle de régression (R2, coefficient de détermination). Plus cette valeur est élevée, meilleure est la relation expliquant les ventes en tant que fonction du budget publicitaire.

$$\sum_t (y_t - \bar{y})^2 = \sum_t (\hat{y}_t - \bar{\hat{y}})^2 + \sum_t e_t^2$$
$$SCT = SCE + SCR$$

| | |
|---|---|
| **SCT** | *La variabilité totale des (yt) c'est la somme des carrés des écarts des observations (yt) par rapport à la moyenne* $\bar{y}$ |
| **SCE** | *La variabilité expliquée par le modèle. (C'est la dispersion totale moins la dispersion résiduelle).* |
| **SCR** | *La variabilité résiduelle. C'est la somme des carrés des écarts des observations (yt) par rapport aux valeurs estimes par le modèle* $\bar{y}_t$ |

| | | |
|---|---|---|
|  <br>$R^2 = 0$<br>l'équation de la droite de régression détermine 0% de la distribution des points. Autrement dit, la droite de régression n'explique absolument pas la distribution des points. La variable explicative $x$ est donc inutile. | $0 \leq R^2 \leq 1$<br>Plus le $R^2$ se rapproche de 0, plus le nuage de points est diffusé autour de la droite de régression.<br>Plus le $R^2$ tend vers 1, plus le nuage de points se rapproche de la droite de régression. |  <br>$R^2 = 1$<br>l'équation de la droite de régression est capable de déterminer 100% de la distribution des points. Autrement dit, la droite de régression déterminée et les paramètres $a_0$ et $a_1$ calculés sont ceux qui déterminent parfaitement la distribution des points. |

Remarquons que le coefficient de corrélation linéaire simple s'écrit :

$$r = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\text{cov}(X,Y)}{\sigma_X^2} \frac{\sigma_X}{\sigma_y} = \hat{a}_1 \frac{\sigma_X}{\sigma_Y}$$

Ce qui implique :

$$\boxed{r = \hat{a}_1 \frac{\sigma_X}{\sigma_Y} \Rightarrow r^2 = \left( \hat{a}_1 \frac{\sigma_X}{\sigma_Y} \right)^2}$$

En plus :

$$\rho^2 = \frac{\hat{a}_1^2 \sum_{t=1}^{n} (x_t - \bar{x})^2}{\sum_{t=1}^{n} (y_t - \bar{y})^2} = \frac{\sum_{t=1}^{n} (\hat{a}_1 x_t - \hat{a}_1 \bar{x})^2}{\sum_{t=1}^{n} (y_t - \bar{y})^2} = \frac{\sum_{t=1}^{n} (\hat{y}_t - \bar{y})^2}{\sum_{t=1}^{n} (y_t - \bar{y})^2} = \frac{SCE}{SCT} = R^2$$

Donc, $\boxed{\rho^2 = R^2}$ et $\boxed{\rho = signe(\hat{a}_1)\, R}$

## Exemple : en garde le même exemple précédent

| années | $(x_t)$ | $(y_t)$ | $(x_t-\bar{x})$ | $(y_t-\bar{y})$ | $(x_t-\bar{x})^2$ | $(x_t-\bar{x})(y_t-\bar{y})$ | $(y_t-\bar{y})^2$ | $\hat{y}_t$ | $\varepsilon = (y_t - \hat{y})$ | $\varepsilon^2$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 8000 | 7389,99 | -3280 | -2595,585 | 10758400 | 8513518,8 | 6737061,49 | 7423,95159 | -33,961594 | 1153,38987 |
| 2 | 9000 | 8169,65 | -2280 | -1815,925 | 5198400 | 4140309 | 3297583,61 | 8204,93434 | -35,284339 | 1244,98458 |
| 3 | 9500 | 8831,71 | -1780 | -1153,865 | 3168400 | 2053879,7 | 1331404,44 | 8595,42571 | 236,284288 | 55830,265 |
| 4 | 9500 | 8652,84 | -1780 | -1332,735 | 3168400 | 2372268,3 | 1776182,58 | 8595,42571 | 57,4142885 | 3296,40052 |
| 5 | 9800 | 8788,08 | -1480 | -1197,495 | 2190400 | 1772292,6 | 1433994,28 | 8829,72054 | -41,640535 | 1733,93416 |
| 6 | 11000 | 9616,21 | -280 | -369,365 | 78400 | 103422,2 | 136430,503 | 9766,89983 | -150,689829 | 22707,4246 |
| 7 | 12000 | 10593,45 | 720 | 607,875 | 518400 | 437670 | 369512,016 | 10547,8826 | 45,567426 | 2076,39031 |
| 8 | 13000 | 11186,11 | 1720 | 1200,535 | 2958400 | 2064920,2 | 1441284,29 | 11328,8653 | -142,755319 | 20379,0811 |
| 9 | 15000 | 12758,09 | 3720 | 2772,515 | 13838400 | 10313756 | 7686839,43 | 12890,8308 | -132,740809 | 17620,1224 |
| 10 | 16000 | 13869,62 | 4720 | 3884,045 | 22278400 | 18332692 | 15085805,6 | 13671,8136 | 197,806446 | 39127,3901 |
| **somme** | **112800** | **99855,75** | **0** | **0** | **64156000** | **50104729** | **39296098,2** | | **0** | **165169,383** |
| **moyenne** | **11280** | **9985,575** | **0** | **0** | | | **SCT** | | | **SCR** |

Estimation de $\hat{a}_1$ = **0.780983**

$\hat{a}_0$ = **1176.090**

$\hat{y}_t = \hat{a}_1 x_t + \hat{a}_0 = 0.78x+1176.08$

**SCR =** **39296098,2**

**SCR =** **165169,383**

**SCR =** **39130928,8**

Coefficient de détermination $R^2$ **0.99579679**

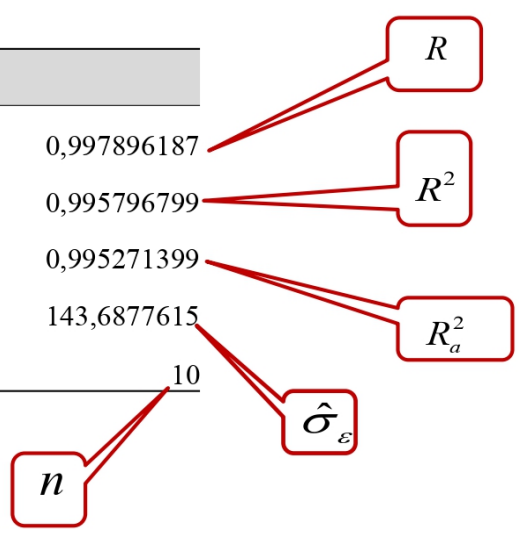
Coefficient de détermination $R^2$ ajusté **0.99527139**

Coefficient de corrélation linéaire $r_{xy}$ **0.99789618**

**SCR**= $\sum(y_t - \hat{y})^2 = \sum\varepsilon_t^2$ **SCT**= $\sum (y_t-\bar{y})^2$ **SCE**= $\sum(\hat{y}_t-\bar{y})^2$ $R^2 = \frac{SCE}{SCT}$ $R^2 = 1-\frac{\frac{SCR}{n-2}}{\frac{SCT}{N-1}}$ $r_{xy} = P_{xy} = \sqrt{R^2}$

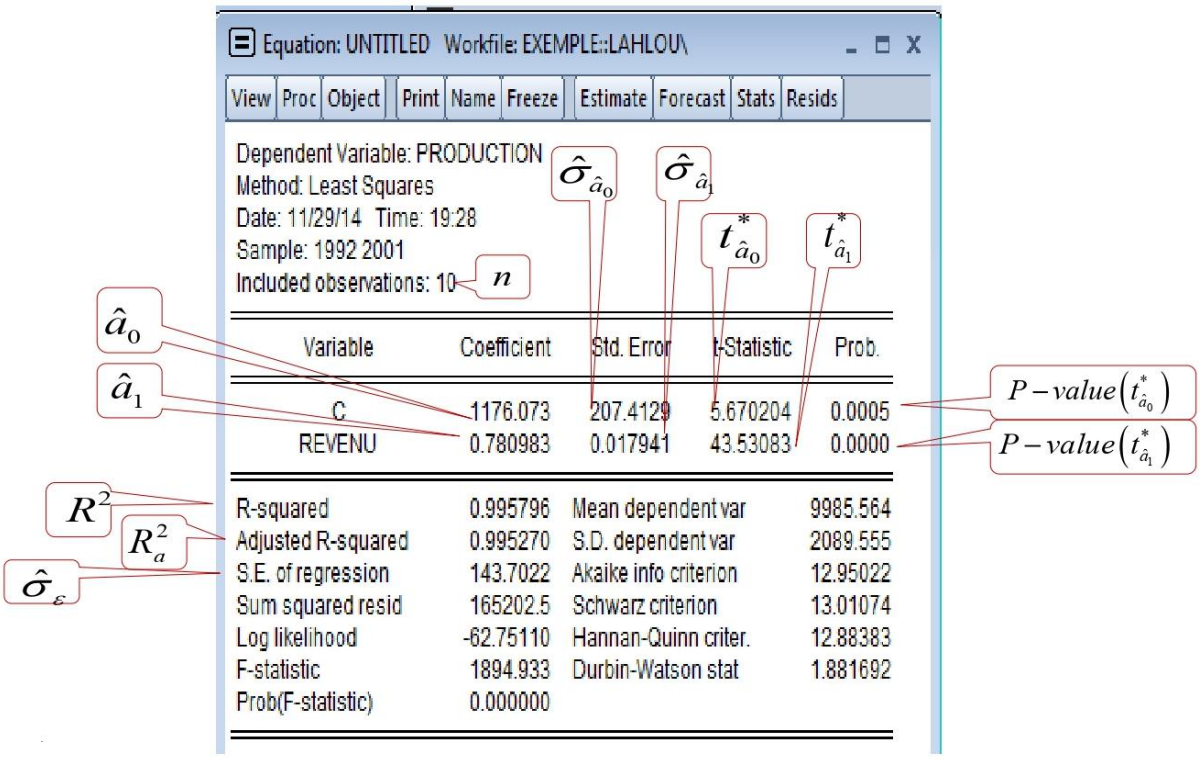
## Extrait du Rapport détaillé par une analyse sur Excel

| *Statistiques de la régression* | |
|---|---|
| Coefficient de détermination multiple R | 0,997896187 |
| Coefficient de détermination $R^2$ | 0,995796799 |
| Coefficient de détermination ajusté | 0,995271399 |
| Erreur-type | 143,6877615 |
| Observations | 10 |




99,58% de la variabilité dans la consommation peut s'expliquer par la variabilité du revenu. Seulement 0,42% restants s'expliquent très mal parfaite corrélation.

**Extrait du rapport détaillé par une analyse sous EVIEWS :**




### 4.6.2.Estimation de la variance de l'erreur :

Ainsi, l'estimateur de la variance de l'erreur $\sigma_\varepsilon^2$ , noté $\hat{\sigma}_\varepsilon^2$ est donné par la variation résiduelle :

$$\hat{\sigma}_\varepsilon^2 = \frac{\sum_{t=1}^{n} e_t^2}{n-2} = \frac{\sum_{t=1}^{n} (y_t - \hat{y}_t)^2}{n-2} = \frac{SCR}{n-2}$$

*Conséquence :* les estimateurs empiriques des variances des estimateurs $\hat{a}_1$ et $\hat{a}_0$ sont donnés par :

$$\hat{\sigma}^2_{\hat{a}_1} = \frac{\hat{\sigma}^2_\varepsilon}{\sum_t (x_t - \overline{x})^2}$$

$$\hat{\sigma}^2_{\hat{a}_0} = \hat{\sigma}^2_\varepsilon \left( \frac{1}{n} + \frac{\overline{x}^2}{\sum_t (x_t - \overline{x})^2} \right)$$

L'hypothèse de normalité des erreurs implique que :

$$\frac{\widehat{a}_1 - a_1}{\sigma_{\hat{a}_1}} \quad \text{et} \quad \frac{\widehat{a}_0 - a_0}{\sigma_{\hat{a}_0}}$$

suivent une loi normale centrée réduite $N(0, 1)$.

**Exemple explicatif**

| Année | $y_t$ | $x_t$ | $(y_t-\bar{y})$ | $(x_t-\bar{x})$ | $(y_t-\bar{y})(x_t-\bar{x})$ | $(y_t-\bar{y})^2$ | $(x_t-\bar{x})^2$ | $\hat{y}_t$ | $e_t=y_t-\hat{y}_t$ | $e_t^2$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 7 389,99 | 8 000,00 | -2 595,59 | -3 280,00 | 8 513 518,80 | 6 737 061,49 | 10 758 400,00 | 7 423,95 | -33,96 | 1 153,39 |
| 2 | 8 169,65 | 9 000,00 | -1 815,93 | -2 280,00 | 4 140 309,00 | 3 297 583,61 | 5 198 400,00 | 8 204,93 | -35,28 | 1 244,98 |
| 3 | 8 831,71 | 9 500,00 | -1 153,87 | -1 780,00 | 2 053 879,70 | 1 331 404,44 | 3 168 400,00 | 8 595,43 | 236,28 | 55 830,26 |
| 4 | 8 652,84 | 9 500,00 | -1 332,74 | -1 780,00 | 2 372 268,30 | 1 776 182,58 | 3 168 400,00 | 8 595,43 | 57,41 | 3 296,40 |
| 5 | 8 788,08 | 9 800,00 | -1 197,50 | -1 480,00 | 1 772 292,60 | 1 433 994,28 | 2 190 400,00 | 8 829,72 | -41,64 | 1 733,93 |
| 6 | 9 616,21 | 11 000,00 | -369,37 | -280,00 | 103 422,20 | 136 430,50 | 78 400,00 | 9 766,90 | -150,69 | 22 707,43 |
| 7 | 10 593,45 | 12 000,00 | 607,88 | 720,00 | 437 670,00 | 369 512,02 | 518 400,00 | 10 547,88 | 45,57 | 2 076,39 |
| 8 | 11 186,11 | 13 000,00 | 1 200,54 | 1 720,00 | 2 064 920,20 | 1 441 284,29 | 2 958 400,00 | 11 328,87 | -142,76 | 20 379,08 |
| 9 | 12 758,09 | 15 000,00 | 2 772,52 | 3 720,00 | 10 313 755,80 | 7 686 839,43 | 13 838 400,00 | 12 890,83 | -132,74 | 17 620,12 |
| 10 | 13 869,62 | 16 000,00 | 3 884,05 | 4 720,00 | 18 332 692,40 | 15 085 805,56 | 22 278 400,00 | 13 671,81 | 197,81 | 39 127,39 |
| **Somme** | 99 855,75 | 112 800,00 | | | 50 104 729,00 | **39 296 098,18** | 64 156 000,00 | | 0,00 | **165 169,38** |
| **Moyenne** | **9 985,58** | **11 280,00** | | | | SCT | | | | SCR |

$\bar{e} = \frac{1}{n}\sum_{t=1}^{n} e_t = 0$

$y_t$ est la valeur observée

$\hat{y}_t$ est la valeur projetée

$e_t$ est le résidu observé

Estimation de $\hat{a}_1 = 0,78$ ; $\hat{a}_0 = 1176,08$

$\hat{y}_t = \hat{a}_1 x_t + \hat{a}_0 = 0,78\,x + 1176,08$

| | |
|---|---|
| $\hat{\sigma}_\varepsilon^2$ | = 20646,1728 |
| $\hat{\sigma}_{\hat{a}_1}^2$ | = 0,0003 |
| $\hat{\sigma}_{\hat{a}_0}^2$ | = 43011,4655 |

| | |
|---|---|
| $\hat{\sigma}_\varepsilon$ | =143,6878 |
| $\hat{\sigma}_{\hat{a}_1}$ | = 0,0179 |
| $\hat{\sigma}_{\hat{a}_0}$ | = 207,3920 |

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{n-2}\sum_{t=1}^{n} e_t^2 = \frac{SCR}{n-2} \quad ; \quad \hat{\sigma}_{\hat{a}_1}^2 = \frac{\hat{\sigma}_\varepsilon^2}{\sum_{t=1}^{n}(x_t-\bar{x})^2} \quad ; \quad \hat{\sigma}_{\hat{a}_0}^2 = \hat{\sigma}_\varepsilon^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{t=1}^{n}(x_t-\bar{x})^2}\right)$$

### 4.6.3.Les tests statistiques :

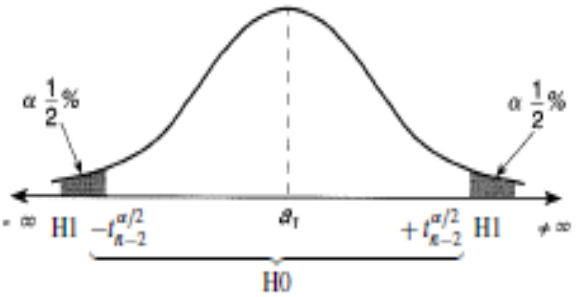
#### *4.6.3.1.* Test sur le coefficient de régression linéaire (pente de la droite de régression)

**Le Test sur le coefficient de régression linéaire (pente de la droite de régression) se fait au moyen de la statistique de student $T_{n-2}$**

$$\begin{cases} H_0 : a_1 = 0 \\ H_1 : a_1 \neq 0 \end{cases}$$

Sous $H_0$ et au risque : $\left|\frac{\hat{a}_1}{\hat{\sigma}_{\hat{a}_1}} \equiv T_{n-2}\right|$ est distribué selon une loi de Student à $n - 2$ degrés de liberté, la distribution d'échantillonnage sous H0 est donc représentée par le graphe suivant.

***Graphique : Distribution d'échantillonnage sous l'hypothèse H0***




La règle de décision pour un seuil $\alpha$ est alors la suivante :

On calcule le ratio empirique de Student $t_{\hat{a}1}^{*} = \dfrac{\hat{a}_1}{\hat{\sigma}_{\hat{a}1}}$ (rapport du coefficient sur son écart type)

– Si $t_{\hat{a}1}^{*}$ est inférieur à $-t_{n-2}^{\alpha/2}$ ou supérieur $+t_{n-2}^{\alpha/2}$ alors on rejette l'hypothèse H0 (nous sommes dans la zone hachurée H1), le coefficient $a_1$ est alors significativement différent de 0 (on accepte $a_1 \neq 0$) ; la variable explicative $R_t$ est donc contributive à l'explication de la variable $C_t$.

– Si $t_{\hat{a}1}^{*}$ est compris dans l'intervalle $\pm t_{n-2}^{\alpha/2}$, alors nous ne sommes pas en mesure de rejeter l'hypothèse H0 (donc on l'accepte), le coefficient $a_1$ n'est pas significativement différent de 0 (on accepte $a_1 = 0$) ; la variable explicative $R_t$ n'est donc pas explicative de la variable $C_t$.

Il est plus simple de profiter de la symétrie de la loi de Student et donc de calculer la valeur absolue du ratio de Student et de la comparer directement à la valeur lue dans la table.

La règle de décision pour un seuil $\alpha = 0{,}05$ est alors la suivante :

– si $t_{\hat{a}1}^{*} = \dfrac{|\widehat{a}_1|}{\widehat{\sigma}_{\hat{a}1}} > t_{n-2}^{0{,}025} \rightarrow$ on rejette l'hypothèse H0, le coefficient $a_1$ est alors significativement différent de 0 (on accepte $a_1 \neq 0$) ; la variable explicative $R_t$ est donc contributive à l'explication de la variable $C_t$ ;

– si $t_{\hat{a}1}^{*} = \dfrac{|\widehat{a}_1|}{\widehat{\sigma}_{\hat{a}1}} \leqslant t_{n-2}^{0{,}025} \rightarrow$ on accepte l'hypothèse H0, le coefficient $a_1$ n'est donc pas significativement différent de 0 (on accepte $a_1 = 0$) ; la variable explicative $R_t$ n'est donc pas contributive à l'explication de $C_t$.

Nous voyons l'importance que revêt ce test dans l'investigation économétrique ; en effet, il permet de tester la pertinence d'une variable explicative qui figure dans un modèle et sa contribution à l'explication du phénomène que l'on cherche à modéliser.

Dans notre exemple, nous calculons le ratio de Student :

$$t_{\hat{a}1}^{*} = \frac{|\widehat{a}_1|}{\widehat{\sigma}_{\hat{a}1}} = \frac{0{,}78}{0{,}017\,9} = 43{,}57 > t_{8}^{0{,}025} = 2{,}306^{1} \rightarrow a_1 \neq 0$$

### 4.6.3.2. Test de signification de la Somme des Carrés Expliqués

Le test **de signification de la Somme des Carrés Expliqués** se fait au moyen de la statistique de Fisher $F_{(1,n-2)}$

$$\begin{cases} H_0 & : & SCE = 0 \\ H_1 & : & SCE \neq 0 \end{cases}$$

Sous $H_0$ et au risque $\alpha$ $\boxed{\dfrac{SCE/1}{SCR/n-2} \equiv F_{(1,n-2)}}$

Tableau d'analyse de la variance pour un modèle de régression simple

| Source de variation | Degrés de liberté | Sommes des carrés | Moyenne des carrés | Fisher $(F^*)$ |
|---|---|---|---|---|
| Régression linéaire Variables explicatives | $k=1$ | $SCE = \sum_{t=1}^{n}(\hat{y}_t - \bar{y})^2$ | $MCE = \dfrac{SCE}{k}$ | $F^* = \dfrac{MCE}{MCR}$ |
| Résidu | $n-k-1$ | $SCR = \sum_{t=1}^{n} e_t^2 = \sum_{t=1}^{n}(y_t - \hat{y}_t)^2$ | $MCR = \dfrac{SCR}{n-k-1}$ | *k* nombre de facteurs. Pour la régression simple *k = 1* |
| Total | $n-1$ | $SCT = \sum_{t=1}^{n}(y_t - \bar{y})^2$ | | |

| Source de variation | Degrés de liberté | Sommes des carrés | Moyenne des carrés | Fisher $(F^*)$ |
|---|---|---|---|---|
| Régression linéaire Variable explicative $x$ | 1 | $\sum_{t=1}^{n}(\hat{y}_t - \bar{y})^2$ | $\dfrac{\sum_{t=1}^{n}(\hat{y}_t - \bar{y})^2}{1}$ | $F^* = \dfrac{\sum_{t=1}^{n}(\hat{y}_t - \bar{y})^2 / 1}{\sum_{t=1}^{n} e_t^2 / (n-2)}$ |
| Résidu | $n-2$ | $\sum_{t=1}^{n}(y_t - \hat{y}_t)^2 = \sum_{t=1}^{n} e_t^2$ | $\dfrac{\sum_{t=1}^{n} e_t^2}{n-2}$ | |
| Total | $n-1$ | $\sum_{t=1}^{n}(y_t - \bar{y})^2$ | | |

Il y a une seule variable explicative. D'où, le degré de liberté est : 1

Dans la variance $\sum_{t=1}^{n}(y_t - \bar{y})^2$

Il y a $n$ écart et une contrainte connue : $\sum_{t=1}^{n}(y_t - \bar{y}) = 0$

D'où, le degré de liberté est : $n-1$

Dans la variance $\sum_{t=1}^{n}(y_t - \hat{y}_t)^2 = \sum_{t=1}^{n} e_t^2$

Il y a $n$ écart et deux contraintes connues :

$\sum_{t=1}^{n} e_t = 0 \quad et \quad \sum_{t=1}^{n} e_t x_t = 0$

D'où, le degré de liberté est :

$n-2$

| Source de variation | Degrés de liberté | Sommes des carrées | Moyenne des carrées | Fisher $\left(F^*\right)$ |
|---|---|---|---|---|
| Régression Variable explicative ($x$) | 1 | $SCE$ | $MCE = \frac{SCE}{1}$ | $F^* = \frac{MCE}{MCR} = \frac{SCE/1}{SCR/(n-2)}$ |
| Résidus | $n-2$ | $SCR$ | $MCR = \frac{SCR}{n-2}$ | |
| Total ($y$) | $n-1$ | $SCT$ | | |

Fisher de degrés de liberté 1 et $n$-2

$$F^* = \frac{MCE}{MCR} = \frac{SCE/1}{SCR/n-2} = (n-2)\frac{SCE/SCT}{SCR/SCT} = (n-2)\frac{R^2}{1-R^2} = \frac{R^2}{(1-R^2)/(n-2)}$$

Si la variance expliquée par le modèle est significativement supérieure à la variance résiduelle, alors la variable $X$ est réellement explicative.

Le test de Fisher (analyse de la variance) permet d'intégrer la taille de l'échantillon $n$ dans l'appréciation de la qualité de la représentation. Soit le test d'hypothèses : $\begin{cases} H_0 : & SCE = SCR \\ H_1 : & SCE > SCR \end{cases}$

- Calculer le Fisher empirique : $F^* = \frac{SCE/1}{SCR/n-2}$
- Comparer $F^*$ avec $F^{\alpha}_{(1,n-2)}$, le Fisher tabulé à (1,$n$-2) degré de liberté et au seuil $\alpha$
- Conclure : si $F^* > F^{\alpha}_{(1,n-2)}$ ou la p-valeur associée est inférieur à α on rejette l'hypothèse nulle d'égalité des variances et donc la variable $X$ est significative et explicative de la variable $Y$

| Année | $y_t$ | $x_t$ | $(y_t-\bar{y})$ | $(x_t-\bar{x})$ | $(y_t-\bar{y})(x_t-\bar{x})$ | $(y_t-\bar{y})^2$ | $(x_t-\bar{x})^2$ | $\hat{y}_t$ | $e_t=y_t-\hat{y}_t$ | $e_t^2$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 7 389,99 | 8 000,00 | -2 595,59 | -3 280,00 | 8 513 518,80 | 6 737 061,49 | 10 758 400,00 | 7 423,95 | -33,96 | 1 153,39 |
| 2 | 8 169,65 | 9 000,00 | -1 815,93 | -2 280,00 | 4 140 309,00 | 3 297 583,61 | 5 198 400,00 | 8 204,93 | -35,28 | 1 244,98 |
| 3 | 8 831,71 | 9 500,00 | -1 153,87 | -1 780,00 | 2 053 879,70 | 1 331 404,44 | 3 168 400,00 | 8 595,43 | 236,28 | 55 830,26 |
| 4 | 8 652,84 | 9 500,00 | -1 332,74 | -1 780,00 | 2 372 268,30 | 1 776 182,58 | 3 168 400,00 | 8 595,43 | 57,41 | 3 296,40 |
| 5 | 8 788,08 | 9 800,00 | -1 197,50 | -1 480,00 | 1 772 292,60 | 1 433 994,28 | 2 190 400,00 | 8 829,72 | -41,64 | 1 733,93 |
| 6 | 9 616,21 | 11 000,00 | -369,37 | -280,00 | 103 422,20 | 136 430,50 | 78 400,00 | 9 766,90 | -150,69 | 22 707,43 |
| 7 | 10 593,45 | 12 000,00 | 607,88 | 720,00 | 437 670,00 | 369 512,02 | 518 400,00 | 10 547,88 | 45,57 | 2 076,39 |
| 8 | 11 186,11 | 13 000,00 | 1 200,54 | 1 720,00 | 2 064 920,20 | 1 441 284,29 | 2 958 400,00 | 11 328,87 | -142,76 | 20 379,08 |
| 9 | 12 758,09 | 15 000,00 | 2 772,52 | 3 720,00 | 10 313 755,80 | 7 686 839,43 | 13 838 400,00 | 12 890,83 | -132,74 | 17 620,12 |
| 10 | 13 869,62 | 16 000,00 | 3 884,03 | 4 720,00 | 18 332 692,40 | 15 085 805,56 | 22 278 400,00 | 13 671,81 | 197,81 | 39 127,39 |
| Somme | 99 855,75 | 112 800,00 | | | 50 104 729,00 | 39 296 098,18 | 64 156 000,00 | | 0,00 | 165 169,38 |
| Moyenne | 9 985,58 | 11 280,00 | | | | SCT | | | | SCR |

Estimation de $\hat{a}_1 = 0{,}780982745$
$\hat{a}_0 = 1176{,}089634$

$\hat{y}_t = \hat{a}_1 x_t + \hat{a}_0 = 0{,}78\,x + 1176{,}08$

| Source de variation | ddl | Sommes des carrées | Moyenne des carrées | Fisher |
|---|---|---|---|---|
| $x$ | 1 | 39130928,80 | 39130928,80 | 1895,311501 |
| Résidus | 8 | 165169,38 | 20646,17282 | |
| Total | 9 | 39296098,18 | | |

F théorique (1,8) (risque α=5 %) : $F^{\alpha}_{(1,8)} = 5{,}317655063$

$F^* = 1895{,}311501 > F^{\alpha}_{(1,8)}$ On rejette ($H_0$ : $a_1 = 0$) La variable $X_t$ est significative

L'analyse de la variance confirme que la variance expliquée est significativement plus élevée que la résiduelle.

105

# 5. Le modèle de régression multiple

## 5.1. Le modèle linéaire général

### 5.1.1. Présentation

Lors du chapitre précédent, nous avons considéré qu'une variable endogène est expliquée à l'aide d'une seule variable exogène. Cependant, il est extrêmement rare qu'un phénomène économique ou social puisse être appréhendé par une seule variable. Le modèle linéaire général est une généralisation du modèle de régression simple dans lequel figurent plusieurs variables explicatives :

$$yt = a_0 + a_1 x_{1t} + a_2 x_{2t} + \ldots + a_k x_{kt} + \varepsilon t$$

$$pour\ t = 1,\ldots, n$$

avec :

$y_t$ = variable à expliquer à la date t ;
$x_{1t}$ = variable explicative 1 à la date t ;
$x_{2t}$ = variable explicative 2 à la date t ;
. . .
*xkt* = variable explicative k à la date t ;
*a0, a1,. . . , ak* = paramètres du modèle ;

$\varepsilon_t$ = erreur de spécification (différence entre le modèle vrai et le modèle spécifié), cette erreur est inconnue et restera inconnue ;

*n* = nombre d'observations.

### 5.1.2.Forme matricielle

L'écriture précédente du modèle est d'un maniement peu pratique. Afin d'en alléger l'écriture et de faciliter l'expression de certains résultats, on a habituellement recours aux notations matricielles. En écrivant le modèle, observation par observation, nous obtenons :

$$y1 = a0 + a1\ x11 + a2\ x21 + \ldots + ak\ xk1 + \varepsilon 1$$
$$y2 = a0 + a1\ x12 + a2\ x22 + \ldots + ak\ xk2 + \varepsilon 2$$
$$\ldots$$
$$yt = a0 + a1\ x1t + a2\ x2t + \ldots + ak\ xkt + \varepsilon t$$
$$\ldots$$
$$yn = a0 + a1\ x1n + a2\ x2n + \ldots + ak\ xkn + \varepsilon n$$

Soit, sous forme matricielle :

$$Y = Xa + \varepsilon$$

Avec :

$$\begin{pmatrix} y_1 \\ \\ y_i \\ \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & & x_{1p} \\ 1 & & & \\ 1 & x_{i1} & x_{ij} & x_{ip} \\ 1 & & & \\ 1 & x_{n1} & & x_{np} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \\ \\ a_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \\ \varepsilon_i \\ \\ \varepsilon_n \end{pmatrix}$$

Nous remarquons la première colonne de la matrice X, composée de 1, qui correspond au coefficient $a_0$ (coefficient du terme constant). La dimension de la matrice X est donc de n lignes et (k + *1*) colonnes (k étant le nombre de variables explicatives réelles, c'est-à-dire constante exclue).

L'écriture sous forme matricielle rend plus aisée la manipulation du modèle linéaire général, c'est pourquoi nous l'adoptons par la suite.

## 5.2. Estimation des coefficients de régression :

Soit le modèle sous forme matricielle à k variables explicatives et n observations :

$$Y = Xa + \varepsilon$$

Afin d'estimer le vecteur a composé des coefficients a0, a1 . . . ak , nous appliquons la méthode des Moindres Carrés Ordinaires (MCO) qui consiste à minimiser la somme des carrés des erreurs, soit :

$$\boxed{\hat{a} = \left(X^T X\right)^{-1} X^T Y}$$

Comme $X^T X$ la matrice carrée d'ordre $(k+1)$ des produits croisés des variables explicatives est symétrique semi-définie positive (pas de colinéarité parfaite entre deux variables explicatives), alors elle est inversible.

- $\hat{a}_0$ étant l'ordonnée à l'origine (toute les valeurs $x_t$ sont nulles)
- $\hat{a}_p$ étant la variation de y suite à une variation unitaire de la variable $xp$ tandis que les autres variables sont maintenues constantes (c'est une propension marginale)

Soit la matrice symétrique ($X^T X$) donnée comme suit :

$$X^T X = \begin{pmatrix} 1 & 1 & 1 & 1 \\ x_{11} & x_{12} & x_{1t} & x_{1n} \\ x_{21} & x_{22} & x_{2t} & x_{2n} \\ & & & \\ x_{k1} & x_{k2} & x_{kt} & x_{kn} \end{pmatrix} \begin{pmatrix} 1 & x_{11} & x_{21} & x_{k1} \\ 1 & x_{12} & x_{22} & x_{k2} \\ & & & \\ 1 & x_{1t} & x_{2t} & x_{kt} \\ & & & \\ 1 & x_{1n} & x_{2n} & x_{kn} \end{pmatrix}$$

$$X^T X = \begin{pmatrix} n & \sum x_{1t} & \sum x_{2t} & \sum x_{kt} \\ \sum x_{1t} & \sum x_{1t}^2 & \sum x_{1t}x_{2t} & \sum x_{1t}x_{kt} \\ \sum x_{2t} & \sum x_{2t}x_{1t} & \sum x_{2t}^2 & \sum x_{2t}x_{kt} \\ & & & \\ \sum x_{kt} & \sum x_{kt}x_{1t} & \sum x_{kt}x_{2t} & \sum x_{kt}^2 \end{pmatrix}$$

De plus,

$$X^T Y = \begin{pmatrix} 1 & 1 & 1 & 1 \\ x_{11} & x_{12} & x_{1t} & x_{1n} \\ x_{21} & x_{22} & x_{2t} & x_{2n} \\ & & & \\ x_{k1} & x_{k2} & x_{kt} & x_{kn} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \\ y_t \\ \\ y_n \end{pmatrix} = \begin{pmatrix} \sum y_t \\ \sum x_{1t} y_t \\ \sum x_{2t} y_t \\ \\ \sum x_{kt} y_t \end{pmatrix}$$

D'où, $X^T X \hat{a} = X^T Y$

$$\begin{pmatrix} n & \sum x_{1t} & \sum x_{2t} & \sum x_{kt} \\ \sum x_{1t} & \sum x_{1t}^2 & \sum x_{1t} x_{2t} & \sum x_{1t} x_{kt} \\ \sum x_{2t} & \sum x_{1t} x_{2t} & \sum x_{2t}^2 & \sum x_{2t} x_{kt} \\ & & & \\ \sum x_{kt} & \sum x_{1t} x_{kt} & \sum x_{2t} x_{kt} & \sum x_{kt}^2 \end{pmatrix} \begin{pmatrix} \hat{a}_0 \\ \hat{a}_1 \\ \hat{a}_2 \\ \\ \hat{a}_k \end{pmatrix} = \begin{pmatrix} \sum y_t \\ \sum x_{1t} y_t \\ \sum x_{2t} y_t \\ \\ \sum x_{kt} y_t \end{pmatrix}$$

**Exercice** :

Soit le modèle à trois variables explicatives :

$$y_t = a_0 + a_1 x_{1t} + a_2 x_{2t} + a_3 x_{3t} + \varepsilon_t$$

Nous disposons des données du tableau 1.

- Mettre le modèle sous forme matricielle en spécifiant bien les dimensions de chacune des matrices.
- Estimer les paramètres du modèle.
- Calculer l'estimation de la variance de l'erreur ainsi que les écarts types de chacun des coefficients.
- Calculer le *R*2 et le *R*2 corrigé.

***Tableau 8 – Valeurs observées de y , x1 , x2 et x3***

| $t$ | $y_t$ | $x_1$ | $x_2$ | $x_3$ |
|---|---|---|---|---|
| 1 | 12 | 2 | 45 | 121 |
| 2 | 14 | 1 | 43 | 132 |
| 3 | 10 | 3 | 43 | 154 |
| 4 | 16 | 6 | 47 | 145 |
| 5 | 14 | 7 | 42 | 129 |
| 6 | 19 | 8 | 41 | 156 |
| 7 | 21 | 8 | 32 | 132 |
| 8 | 19 | 5 | 33 | 147 |
| 9 | 21 | 5 | 41 | 128 |
| 10 | 16 | 8 | 38 | 163 |
| 11 | 19 | 4 | 32 | 161 |
| 12 | 21 | 9 | 31 | 172 |
| 13 | 25 | 12 | 35 | 174 |
| 14 | 21 | 7 | 29 | 180 |

***Solution*** :

- **Forme matricielle** : Mettre le modèle sous forme matricielle en spécifiant bien les dimensions de chacune des matrices

Nous disposons de 14 observations et trois variables explicatives, le modèle peut donc s'écrire :

$$Y = Xa + \varepsilon \quad \text{avec :}$$

$$Y = \begin{pmatrix} 12 \\ 14 \\ 10 \\ \vdots \\ 21 \end{pmatrix} ; \; X = \begin{pmatrix} 1 & 2 & 45 & 121 \\ 1 & 1 & 43 & 132 \\ 1 & 3 & 43 & 154 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 7 & 29 & 180 \end{pmatrix} ; \; a = \begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{pmatrix} ; \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_t \\ \vdots \\ \varepsilon_{14} \end{pmatrix}$$

Dimensions :

(14,1) (14,4) (4,1) (14,1)

**2)** Estimation des paramètres

Nous savons d'après [3] que $\widehat{a} = (X' X)^{-1} X' Y$.

Calcul de $X' X$ et de $(X' X)^{-1}$

$$\underset{X'}{\begin{pmatrix} 1 & 1 & 1 & \dots & 1 \\ 2 & 1 & 3 & \dots & 7 \\ 45 & 43 & 43 & \dots & 29 \\ 121 & 132 & 154 & \dots & 180 \end{pmatrix}} \underset{X}{\begin{pmatrix} 1 & 2 & 45 & 121 \\ 1 & 1 & 43 & 132 \\ 1 & 3 & 43 & 154 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 7 & 29 & 180 \end{pmatrix}} =$$

$$= \begin{pmatrix} 14 & 85 & 532 & 2\,094 \\ 85 & 631 & 3\,126 & 13\,132 \\ 532 & 3\,126 & 20\,666 & 78\,683 \\ 2\,094 & 13\,132 & 78\,683 & 317\,950 \end{pmatrix}$$

$$(X'\,X)^{-1} = \begin{pmatrix} 20,16864 & 0,015065 & -0,23145 & -0,07617 \\ 0,015065 & 0,013204 & 0,001194 & -0,00094 \\ -0,23145 & 0,001194 & 0,003635 & 0,000575 \\ -0,07617 & -0,00094 & 0,000575 & 0,000401 \end{pmatrix}$$

Calcul de $X'\,Y$

$$\underset{X'}{\begin{pmatrix} 1 & 1 & 1 & \dots & 1 \\ 2 & 1 & 3 & \dots & 7 \\ 45 & 43 & 43 & \dots & 29 \\ 121 & 132 & 154 & \dots & 180 \end{pmatrix}} \underset{Y}{\begin{pmatrix} 12 \\ 14 \\ 10 \\ \vdots \\ 21 \end{pmatrix}} = \begin{pmatrix} 248 \\ 1\,622 \\ 9\,202 \\ 37\,592 \end{pmatrix}$$

Calcul de $\widehat{a}$

$$\underset{(X'\,X)^{-1}}{\begin{pmatrix} 20,16864 & 0,015065 & -0,23145 & -0,07617 \\ 0,015065 & 0,013204 & 0,001194 & -0,00094 \\ -0,23145 & 0,001194 & 0,003635 & 0,000575 \\ -0,07617 & -0,00094 & 0,000575 & 0,000401 \end{pmatrix}} \underset{X'\,Y}{\begin{pmatrix} 248 \\ 1\,622 \\ 9\,202 \\ 37\,592 \end{pmatrix}} = \widehat{a}$$

$$= \begin{pmatrix} 32,89132 \\ 0,801900 \\ -\ 0,38136 \\ -\ 0,03713 \end{pmatrix}$$

Soit $\widehat{a}_0 = 32,89$ ; $\widehat{a}_1 = 0,80$ ; $\widehat{a}_2 = -0,38$ ; $\widehat{a}_3 = -0,03$

Les calculs que nous venons de développer sont longs et fastidieux et mettent en évidence l'intérêt incontestable d'utiliser un ordinateur.

**3)** Calcul de $\widehat{\sigma}_\varepsilon^2$ et de $\widehat{\sigma}_{\widehat{a}}^2$

D'après [6] $\widehat{\sigma}_\varepsilon^2 = \dfrac{e'\,e}{n-k-1}$, nous devons donc calculer le résidu $e$.

$$\begin{aligned} e &= Y - \widehat{Y} = Y - X\widehat{a} \\ \text{Soit} \quad e_t &= y_t - (\widehat{a}_0 + \widehat{a}_1 x_{1t} + \widehat{a}_2 x_{2t} + \widehat{a}_3 x_{3t}) \\ e_t &= y_t - 32,89 - 0,80\,x_{1t} + 0,38\,x_{2t} + 0,03\,x_{3t} \end{aligned}$$

Par exemple pour $e_1$ :

$$e_1 = y_1 - 32,89 - 0,80\,x_{11} + 0,38\,x_{21} + 0,03\,x_{31}$$

$$e_1 = 12 - 32,89 - 0,80 \times 2 + 0,38 \times 45 + 0,03 \times 121 = -0,84$$

Le tableau 2 présente l'ensemble des résultats.

Par construction, la somme des résidus est bien nulle.

$$\widehat{\sigma}_\varepsilon^2 = \frac{e'\,e}{n-k-1} = \frac{\sum_{t=1}^{t=14} e_t^2}{14-3-1} = \frac{67,45}{10} = 6,745$$

***Tableau 9 – Calcul du résidu***

| $t$ | $y_t$ | $\widehat{y}_t$ | $e_t$ | $e_t^2$ |
|---|---|---|---|---|
| 1 | 12 | 12,84 | – 0,84 | 0,71 |
| 2 | 14 | 12,39 | 1,61 | 2,58 |
| 3 | 10 | 13,18 | – 3,18 | 10,11 |
| 4 | 16 | 13,39 | 1,61 | 2,58 |
| 5 | 14 | 17,70 | – 3,70 | 13,67 |
| 6 | 19 | 17,88 | 1,12 | 1,26 |
| 7 | 21 | 22,20 | – 1,20 | 1,44 |
| 8 | 19 | 18,86 | 0,14 | 0,02 |
| 9 | 21 | 16,51 | 4,49 | 20,14 |
| 10 | 16 | 18,76 | – 2,76 | 7,63 |
| 11 | 19 | 17,92 | 1,08 | 1,17 |
| 12 | 21 | 21,90 | – 0,90 | 0,81 |
| 13 | 25 | 22,71 | 2,29 | 5,27 |
| 14 | 21 | 20,76 | 0,24 | 0,06 |
| | | Somme | 0 | 67,45 |

La matrice des variances et covariances estimées des coefficients nous est donnée par [7], soit :

$$\widehat{\Omega}_{\hat{a}} = \widehat{\sigma}_{\varepsilon}^2 (X'\ X)^{-1}$$

$$\widehat{\Omega}_{\hat{a}} = 6{,}745 \times \begin{pmatrix} 20{,}16864 & 0{,}015065 & -0{,}23145 & -0{,}07617 \\ 0{,}015065 & 0{,}013204 & 0{,}001194 & -0{,}00094 \\ -0{,}23145 & 0{,}001194 & 0{,}003635 & 0{,}000575 \\ -0{,}07617 & -0{,}00094 & -0{,}000575 & 0{,}000401 \end{pmatrix}$$

Les variances des coefficients de régression se trouvent sur la première diagonale :

$$\begin{aligned} \widehat{\sigma}_{\hat{a}_0}^2 &= 6{,}745 \times 20{,}17 = 136{,}04 \rightarrow \widehat{\sigma}_{\hat{a}_0} = 11{,}66 \\ \widehat{\sigma}_{\hat{a}_1}^2 &= 6{,}745 \times 0{,}013 = 0{,}087 \rightarrow \widehat{\sigma}_{\hat{a}1} = 0{,}29 \\ \widehat{\sigma}_{\hat{a}2}^2 &= 6{,}745 \times 0{,}0036 = 0{,}024 \rightarrow \widehat{\sigma}_{\hat{a}2} = 0{,}15 \\ \widehat{\sigma}_{\hat{a}3}^2 &= 6{,}745 \times 0{,}0004 = 0{,}0026 \rightarrow \widehat{\sigma}_{\hat{a}3} = 0{,}05 \end{aligned}$$

**4)** Le calcul du $R^2$ est effectué à partir de la formule [9].

Nous connaissons $e'\ e = 67{,}45$, il convient de calculer $\sum_t (y_t - \overline{y})^2 = 226{,}86$.

$$R^2 = 1 - \frac{\sum_t e_t^2}{\sum_t (y_t - \overline{y})^2} = 1 - \frac{67{,}45}{226{,}86} = 0{,}702$$

Le $\overline{R}^2$ corrigé est donné par [11] :

$$\overline{R}^2 = 1 - \frac{n-1}{n-k-1}(1 - R^2) = 1 - \frac{14-1}{14-4}\ (1 - 0{,}702) = 0{,}613$$

Nous observons la baisse du coefficient de détermination lorsque nous le corrigeons par le degré de liberté.

# 6. ANNEXS :

## Lois de probabilités d'une variable discrète

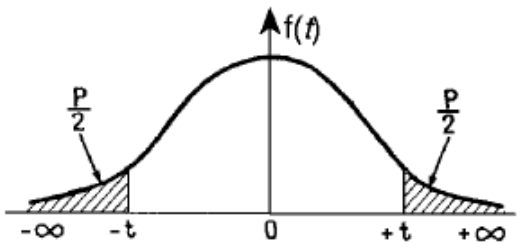
| Dénomination | Loi | Espérance | Variance |
|---|---|---|---|
| Loi de Bernoulli $X \hookrightarrow \mathcal{B}(1,p)$ | $X(\Omega)=\{0,1\}$ $P(X=1)=p$ $P(X=0)=q$ | $\mathbb{E}(X)=p$ | $\mathbb{V}(X)=pq$ |
| Loi Binomiale $X \hookrightarrow \mathcal{B}(n,p)$ | $X(\Omega)=[\![0,n]\!]$ $P(X=k)=C_n^k p^k q^{n-k}$ | $\mathbb{E}(X)=np$ | $\mathbb{V}(X)=npq$ |
| Loi Uniforme $X \hookrightarrow \mathcal{U}([\![1,n]\!])$ | $X(\Omega)=[\![1,n]\!]$ $P(X=k)=\frac{1}{n}$ | $\mathbb{E}(X)=\frac{n+1}{2}$ | $\mathbb{V}(X)=\frac{n^2-1}{12}$ |
| Loi Géométrique $X \hookrightarrow \mathcal{G}(p)$ | $X(\Omega)=[\![1,+\infty[\![$ $P(X=k)=pq^{k-1}$ | $\mathbb{E}(X)=\frac{1}{p}$ | $\mathbb{V}(X)=\frac{q}{p^2}$ |
| Loi de Poisson $X \hookrightarrow \mathcal{P}(\lambda)$ | $X(\Omega)=[\![0,+\infty[\![$ $P(X=k)=e^{-\lambda}\frac{\lambda^k}{k!}$ | $\mathbb{E}(X)=\lambda$ | $\mathbb{V}(X)=\lambda$ |

## Lois de probabilités d'une variable discrète

| Loi et Symbole $X \rightsquigarrow$ | Densité | Espérance | $Var(X)$ | Fonction caractéristique $\phi_X(t)=\mathbb{E}(e^{itX})$ |
|---|---|---|---|---|
| Loi Uniforme $\mathcal{U}[a,b]$ | $f_X(x)=\frac{1}{b-a}\mathbb{1}_{[a,b]}(x)$ | $\frac{a+b}{2}$ | $\frac{(b-a)^2}{12}$ | $\frac{e^{itb}-e^{ita}}{it(b-a)}$ |
| Loi Normale $\mathcal{N}(m,\sigma^2)$ | $f_X(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-m)^2}{2\sigma^2}}\mathbb{1}_{\mathbb{R}}(x)$ | $m$ | $\sigma^2$ | $e^{itm-\frac{\sigma^2t^2}{2}}$ |
| Loi Exponentielle $\mathcal{E}xp(\lambda)=\mathcal{G}(1,\lambda)$ | $f_X(x)=\lambda e^{-\lambda x}\mathbb{1}_{\mathbb{R}_+}(x)$ | $\frac{1}{\lambda}$ | $\frac{1}{\lambda^2}$ | $(1-\frac{it}{\lambda})^{-1}$ |
| Loi Gamma $\mathcal{G}(\alpha,\lambda)$ | $f_X(x)=\frac{\lambda^\alpha}{\Gamma(\alpha)}e^{-\lambda x}x^{\alpha-1}\mathbb{1}_{\mathbb{R}_+^*}(x)$ | $\frac{\alpha}{\lambda}$ | $\frac{\alpha}{\lambda^2}$ | $(1-\frac{it}{\lambda})^{-\alpha}$ |
| Loi du Chi-deux $\chi_n^2=G(\frac{n}{2},\frac{1}{2})$ | $f_X(x)=\frac{2^{-\frac{n}{2}}}{\Gamma(\frac{n}{2})}e^{-\frac{x}{2}}x^{\frac{n}{2}-1}\mathbb{1}_{\mathbb{R}_+}(x)$ | $n$ | $2n$ | $(1-2it)^{-\frac{n}{2}}$ |
| Première loi de Laplace | $f_X(x)=\frac{1}{2}e^{-\lvert x\rvert}\mathbb{1}_{\mathbb{R}}(x)$ | 0 | 2 | $\frac{1}{1+t^2}$ |

## 2. TABLE DE LA LOI DE STUDENT

Valeurs de $T$ ayant la probabilité $P$ d'être dépassées en valeur absolue

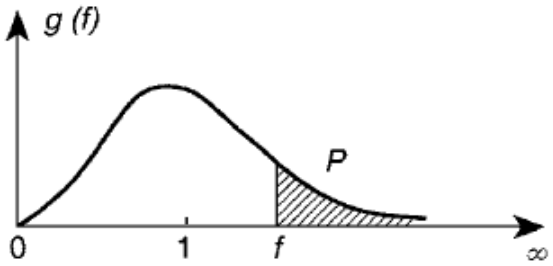



| $\nu$ | $P = 0{,}90$ | 0,80 | 0,70 | 0,60 | 0,50 | 0,40 | 0,30 | 0,20 | 0,10 | 0,05 | 0,02 | 0,01 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0,158 | 0,325 | 0,510 | 0,727 | 1,000 | 1,376 | 1,963 | 3,078 | 6,314 | 12,706 | 31,821 | 63,657 |
| 2 | 0,142 | 0,289 | 0,445 | 0,617 | 0,816 | 1,061 | 1,386 | 1,886 | 2,920 | 4,303 | 6,965 | 9,925 |
| 3 | 0,137 | 0,277 | 0,424 | 0,584 | 0,765 | 0,978 | 1,250 | 1,638 | 2,353 | 3,182 | 4,541 | 5,841 |
| 4 | 0,134 | 0,271 | 0,414 | 0,569 | 0,741 | 0,941 | 1,190 | 1,533 | 2,132 | 2,776 | 3,747 | 4,604 |
| 5 | 0,132 | 0,267 | 0,408 | 0,559 | 0,727 | 0,920 | 1,156 | 1,476 | 2,015 | 2,571 | 3,365 | 4,032 |
| 6 | 0,131 | 0,265 | 0,404 | 0,553 | 0,718 | 0,906 | 1,134 | 1,440 | 1,943 | 2,447 | 3,143 | 3,707 |
| 7 | 0,130 | 0,263 | 0,402 | 0,549 | 0,711 | 0,896 | 1,119 | 1,415 | 1,895 | 2,365 | 2,998 | 3,499 |
| 8 | 0,130 | 0,262 | 0,399 | 0,546 | 0,706 | 0,889 | 1,108 | 1,397 | 1,860 | 2,306 | 2,896 | 3,355 |
| 9 | 0,129 | 0,261 | 0,398 | 0,543 | 0,703 | 0,883 | 1,100 | 1,383 | 1,833 | 2,262 | 2,821 | 3,250 |
| 10 | 0,129 | 0,260 | 0,397 | 0,542 | 0,700 | 0,879 | 1,093 | 1,372 | 1,812 | 2,228 | 2,764 | 3,169 |
| 11 | 0,129 | 0,260 | 0,396 | 0,540 | 0,697 | 0,876 | 1,088 | 1,363 | 1,796 | 2,201 | 2,718 | 3,106 |
| 12 | 0,128 | 0,259 | 0,395 | 0,539 | 0,695 | 0,873 | 1,083 | 1,356 | 1,782 | 2,179 | 2,681 | 3,055 |
| 13 | 0,128 | 0,259 | 0,394 | 0,538 | 0,694 | 0,870 | 1,079 | 1,350 | 1,771 | 2,160 | 2,650 | 3,012 |
| 14 | 0,128 | 0,258 | 0,393 | 0,537 | 0,692 | 0,868 | 1,076 | 1,345 | 1,761 | 2,145 | 2,624 | 2,977 |
| 15 | 0,128 | 0,258 | 0,393 | 0,536 | 0,691 | 0,866 | 1,074 | 1,341 | 1,753 | 2,131 | 2,602 | 2,947 |
| 16 | 0,128 | 0,258 | 0,392 | 0,535 | 0,690 | 0,865 | 1,071 | 1,337 | 1,746 | 2,120 | 2,583 | 2,921 |
| 17 | 0,128 | 0,257 | 0,392 | 0,534 | 0,689 | 0,863 | 1,069 | 1,333 | 1,740 | 2,110 | 2,567 | 2,898 |
| 18 | 0,127 | 0,257 | 0,392 | 0,534 | 0,688 | 0,862 | 1,067 | 1,330 | 1,734 | 2,101 | 2,552 | 2,878 |
| 19 | 0,127 | 0,257 | 0,391 | 0,533 | 0,688 | 0,861 | 1,066 | 1,328 | 1,729 | 2,093 | 2,539 | 2,861 |
| 20 | 0,127 | 0,257 | 0,391 | 0,533 | 0,687 | 0,860 | 1,064 | 1,325 | 1,725 | 2,086 | 2,528 | 2,845 |
| 21 | 0,127 | 0,257 | 0,391 | 0,532 | 0,686 | 0,859 | 1,063 | 1,323 | 1,721 | 2,080 | 2,518 | 2,831 |
| 22 | 0,127 | 0,256 | 0,390 | 0,532 | 0,686 | 0,858 | 1,061 | 1,321 | 1,717 | 2,074 | 2,508 | 2,819 |
| 23 | 0,127 | 0,256 | 0,390 | 0,532 | 0,685 | 0,858 | 1,060 | 1,319 | 1,714 | 2,069 | 2,500 | 2,807 |
| 24 | 0,127 | 0,256 | 0,390 | 0,531 | 0,685 | 0,857 | 1,059 | 1,318 | 1,711 | 2,064 | 2,492 | 2,797 |
| 25 | 0,127 | 0,256 | 0,390 | 0,531 | 0,684 | 0,856 | 1,058 | 1,316 | 1,708 | 2,060 | 2,485 | 2,787 |
| 26 | 0,127 | 0,256 | 0,390 | 0,531 | 0,684 | 0,856 | 1,058 | 1,315 | 1,706 | 2,056 | 2,479 | 2,779 |
| 27 | 0,127 | 0,256 | 0,389 | 0,531 | 0,684 | 0,855 | 1,057 | 1,314 | 1,703 | 2,052 | 2,473 | 2,771 |
| 28 | 0,127 | 0,256 | 0,389 | 0,530 | 0,683 | 0,855 | 1,056 | 1,313 | 1,701 | 2,048 | 2,467 | 2,763 |
| 29 | 0,127 | 0,256 | 0,389 | 0,530 | 0,683 | 0,854 | 1,055 | 1,311 | 1,699 | 2,045 | 2,462 | 2,756 |
| 30 | 0,127 | 0,256 | 0,389 | 0,530 | 0,683 | 0,854 | 1,055 | 1,310 | 1,697 | 2,042 | 2,457 | 2,750 |
| ∞ | 0,12566 | 0,25335 | 0,38532 | 0,52440 | 0,67449 | 0,84162 | 1,03643 | 1,28155 | 1,64485 | 1,95996 | 2,32634 | 2,57582 |

Nota. — $\nu$ est le nombre de degrés de liberté.

## 4. TABLE DE LA LOI DE FISHER-SNEDECOR

Valeurs de $F$ ayant la probabilité $P$ d'être dépassées ($F = s_1^2/_2^2$)

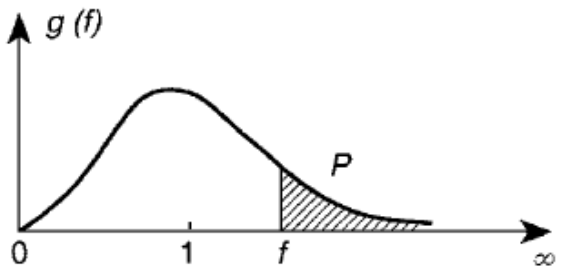



| $\nu_2$ | $\nu_1 = 1$ | | $\nu_1 = 2$ | | $\nu_1 = 3$ | | $\nu_1 = 4$ | | $\nu_1 = 5$ | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $P = 0{,}05$ | $P = 0{,}01$ | $P = 0{,}05$ | $P = 0{,}01$ | $P = 0{,}05$ | $P = 0{,}01$ | $P = 0{,}05$ | $P = 0{,}01$ | $P = 0{,}05$ | $P = 0{,}01$ |
| 1 | 161,4 | 4 052 | 199,5 | 4 999 | 215,7 | 5 403 | 224,6 | 5 625 | 230,2 | 5 764 |
| 2 | 18,51 | 98,49 | 19,00 | 99,00 | 19,16 | 99,17 | 19,25 | 99,25 | 19,30 | 99,30 |
| 3 | 10,13 | 34,12 | 9,55 | 30,81 | 9,28 | 29,46 | 9,12 | 28,71 | 9,01 | 28,24 |
| 4 | 7,71 | 21,20 | 6,94 | 18,00 | 6,59 | 16,69 | 6,39 | 15,98 | 6,26 | 15,52 |
| 5 | 6,61 | 16,26 | 5,79 | 13,27 | 5,41 | 12,06 | 5,19 | 11,39 | 5,05 | 10,97 |
| 6 | 5,99 | 13,74 | 5,14 | 10,91 | 4,76 | 9,78 | 4,53 | 9,15 | 4,39 | 8,75 |
| 7 | 5,59 | 12,25 | 4,74 | 9,55 | 4,35 | 8,45 | 4,12 | 7,85 | 3,97 | 7,45 |
| 8 | 5,32 | 11,26 | 4,46 | 8,65 | 4,07 | 7,59 | 3,84 | 7,01 | 3,69 | 6,63 |
| 9 | 5,12 | 10,56 | 4,26 | 8,02 | 3,86 | 6,99 | 3,63 | 6,42 | 3,48 | 6,06 |
| 10 | 4,96 | 10,04 | 4,10 | 7,56 | 3,71 | 6,55 | 3,48 | 5,99 | 3,33 | 5,64 |
| 11 | 4,84 | 9,65 | 3,98 | 7,20 | 3,59 | 6,22 | 3,36 | 5,67 | 3,20 | 5,32 |
| 12 | 4,75 | 9,33 | 3,88 | 6,93 | 3,49 | 5,95 | 3,26 | 5,41 | 3,11 | 5,06 |
| 13 | 4,67 | 9,07 | 3,80 | 6,70 | 3,41 | 5,74 | 3,18 | 5,20 | 3,02 | 4,86 |
| 14 | 4,60 | 8,86 | 3,74 | 6,51 | 3,34 | 5,56 | 3,11 | 5,03 | 2,96 | 4,69 |
| 15 | 4,54 | 8,68 | 3,68 | 6,36 | 3,29 | 5,42 | 3,06 | 4,89 | 2,90 | 4,56 |
| 16 | 4,49 | 8,53 | 3,63 | 6,23 | 3,24 | 5,29 | 3,01 | 4,77 | 2,85 | 4,44 |
| 17 | 4,45 | 8,40 | 3,59 | 6,11 | 3,20 | 5,18 | 2,96 | 4,67 | 2,81 | 4,34 |
| 18 | 4,41 | 8,28 | 3,55 | 6,01 | 3,16 | 5,09 | 2,93 | 4,58 | 2,77 | 4,25 |
| 19 | 4,38 | 8,18 | 3,52 | 5,93 | 3,13 | 5,01 | 2,90 | 4,50 | 2,74 | 4,17 |
| 20 | 4,35 | 8,10 | 3,49 | 5,85 | 3,10 | 4,94 | 2,87 | 4,43 | 2,71 | 4,10 |
| 21 | 4,32 | 8,02 | 3,47 | 5,78 | 3,07 | 4,87 | 2,84 | 4,37 | 2,68 | 4,04 |
| 22 | 4,30 | 7,94 | 3,44 | 5,72 | 3,05 | 4,82 | 2,82 | 4,31 | 2,66 | 3,99 |
| 23 | 4,28 | 7,88 | 3,42 | 5,66 | 3,03 | 4,76 | 2,80 | 4,26 | 2,64 | 3,94 |
| 24 | 4,26 | 7,82 | 3,40 | 5,61 | 3,01 | 4,72 | 2,78 | 4,22 | 2,62 | 3,90 |
| 25 | 4,24 | 7,77 | 3,38 | 5,57 | 2,99 | 4,68 | 2,76 | 4,18 | 2,60 | 3,86 |
| 26 | 4,22 | 7,72 | 3,37 | 5,53 | 2,98 | 4,64 | 2,74 | 4,14 | 2,59 | 3,82 |
| 27 | 4,21 | 7,68 | 3,35 | 5,49 | 2,96 | 4,60 | 2,73 | 4,11 | 2,57 | 3,78 |
| 28 | 4,20 | 7,64 | 3,34 | 5,45 | 2,95 | 4,57 | 2,71 | 4,07 | 2,56 | 3,75 |
| 29 | 4,18 | 7,60 | 3,33 | 5,42 | 2,93 | 4,54 | 2,70 | 4,04 | 2,54 | 3,73 |
| 30 | 4,17 | 7,56 | 3,32 | 5,39 | 2,92 | 4,51 | 2,69 | 4,02 | 2,53 | 3,70 |
| 40 | 4,08 | 7,31 | 3,23 | 5,18 | 2,84 | 4,31 | 2,61 | 3,83 | 2,45 | 3,51 |
| 60 | 4,00 | 7,08 | 3,15 | 4,98 | 2,76 | 4,13 | 2,52 | 3,65 | 2,37 | 3,34 |
| 120 | 3,92 | 6,85 | 3,07 | 4,79 | 2,68 | 3,95 | 2,45 | 3,48 | 2,29 | 3,17 |
| $\infty$ | 3,84 | 6,64 | 2,99 | 4,60 | 2,60 | 3,78 | 2,37 | 3,32 | 2,21 | 3,02 |

Nota. — $s_1^2$ est la plus grande des deux variances estimées, avec $\nu_1$ degrés de liberté.

## 5. TABLE DE LA LOI DE FISHER-SNEDECOR (suite)

Valeurs de $F$ ayant la probabilité $P$ d'être dépassées ($F = s_1^2/_2^2$)




| $\nu_2$ | $\nu_1 = 6$ | | $\nu_1 = 8$ | | $\nu_1 = 12$ | | $\nu_1 = 24$ | | $\nu_1 = \infty$ | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $P = 0{,}05$ | $P = 0{,}01$ | $P = 0{,}05$ | $P = 0{,}01$ | $P = 0{,}05$ | $P = 0{,}01$ | $P = 0{,}05$ | $P = 0{,}01$ | $P = 0{,}05$ | $P = 0{,}01$ |
| 1 | 234,0 | 5 859 | 238,9 | 5 981 | 243,9 | 6 106 | 249,0 | 6 234 | 254,3 | 6366 |
| 2 | 19,33 | 99,33 | 19,37 | 99,36 | 19,41 | 99,42 | 19,45 | 99,46 | 19,50 | 99,50 |
| 3 | 8,94 | 27,91 | 8,84 | 27,49 | 8,74 | 27,05 | 8,64 | 26,60 | 8,53 | 26,12 |
| 4 | 6,16 | 15,21 | 6,04 | 14,80 | 5,91 | 14,37 | 5,77 | 13,93 | 5,63 | 13,46 |
| 5 | 4,95 | 10,67 | 4,82 | 10,27 | 4,68 | 9,89 | 4,53 | 9,47 | 4,36 | 9,02 |
| 6 | 4,28 | 8,47 | 4,15 | 8,10 | 4,00 | 7,72 | 3,84 | 7,31 | 3,67 | 6,88 |
| 7 | 3,87 | 7,19 | 3,73 | 6,84 | 3,57 | 6,47 | 3,41 | 6,07 | 3,23 | 5,65 |
| 8 | 3,58 | 6,37 | 3,44 | 6,03 | 3,28 | 5,67 | 3,12 | 5,28 | 2,93 | 4,86 |
| 9 | 3,37 | 5,80 | 3,23 | 5,47 | 3,07 | 5,11 | 2,90 | 4,73 | 2,71 | 4,31 |
| 10 | 3,22 | 5,39 | 3,07 | 5,06 | 2,91 | 4,71 | 2,74 | 4,33 | 2,54 | 3,91 |
| 11 | 3,09 | 5,07 | 2,95 | 4,74 | 2,79 | 4,40 | 2,61 | 4,02 | 2,40 | 3,60 |
| 12 | 3,00 | 4,82 | 2,85 | 4,50 | 2,69 | 4,16 | 2,50 | 3,78 | 2,30 | 3,36 |
| 13 | 2,92 | 4,62 | 2,77 | 4,30 | 2,60 | 3,96 | 2,42 | 3,59 | 2,21 | 3,16 |
| 14 | 2,85 | 4,46 | 2,70 | 4,14 | 2,53 | 3,80 | 2,35 | 3,43 | 2,13 | 3,00 |
| 15 | 2,79 | 4,32 | 2,64 | 4,00 | 2,48 | 3,67 | 2,29 | 3,29 | 2,07 | 2,87 |
| 16 | 2,74 | 4,20 | 2,59 | 3,89 | 2,42 | 3,55 | 2,24 | 3,18 | 2,01 | 2,75 |
| 17 | 2,70 | 4,10 | 2,55 | 3,79 | 2,38 | 3,45 | 2,19 | 3,08 | 1,96 | 2,65 |
| 18 | 2,66 | 4,01 | 2,51 | 3,71 | 2,34 | 3,37 | 2,15 | 3,00 | 1,92 | 2,57 |
| 19 | 2,63 | 3,94 | 2,48 | 3,63 | 2,31 | 3,30 | 2,11 | 2,92 | 1,88 | 2,49 |
| 20 | 2,60 | 3,87 | 2,45 | 3,56 | 2,28 | 3,23 | 2,08 | 2,86 | 1,84 | 2,42 |
| 21 | 2,57 | 3,81 | 2,42 | 3,51 | 2,25 | 3,17 | 2,05 | 2,80 | 1,81 | 2,36 |
| 22 | 2,55 | 3,76 | 2,40 | 3,45 | 2,23 | 3,12 | 2,03 | 2,75 | 1,78 | 2,31 |
| 23 | 2,53 | 3,71 | 2,38 | 3,41 | 2,20 | 3,07 | 2,00 | 2,70 | 1,76 | 2,26 |
| 24 | 2,51 | 3,67 | 2,36 | 3,36 | 2,18 | 3,03 | 1,98 | 2,66 | 1,73 | 2,21 |
| 25 | 2,49 | 3,63 | 2,34 | 3,32 | 2,16 | 2,99 | 1,96 | 2,62 | 1,71 | 2,17 |
| 26 | 2,47 | 3,59 | 2,32 | 3,29 | 2,15 | 2,96 | 1,95 | 2,58 | 1,69 | 2,13 |
| 27 | 2,46 | 3,56 | 2,30 | 3,26 | 2,13 | 2,93 | 1,93 | 2,55 | 1,67 | 2,10 |
| 28 | 2,44 | 3,53 | 2,29 | 3,23 | 2,12 | 2,90 | 1,91 | 2,52 | 1,65 | 2,06 |
| 29 | 2,43 | 3,50 | 2,28 | 3,20 | 2,10 | 2,87 | 1,90 | 2,49 | 1,64 | 2,03 |
| 30 | 2,42 | 3,47 | 2,27 | 3,17 | 2,09 | 2,84 | 1,89 | 2,47 | 1,62 | 2,01 |
| 40 | 2,34 | 3,29 | 2,18 | 2,99 | 2,00 | 2,66 | 1,79 | 2,29 | 1,51 | 1,80 |
| 60 | 2,25 | 3,12 | 2,10 | 2,82 | 1,92 | 2,50 | 1,70 | 2,12 | 1,39 | 1,60 |
| 120 | 2,17 | 2,96 | 2,01 | 2,66 | 1,83 | 2,34 | 1,61 | 1,95 | 1,25 | 1,38 |
| ∞ | 2,09 | 2,80 | 1,94 | 2,51 | 1,75 | 2,18 | 1,52 | 1,79 | 1,00 | 1,00 |

Nota. — $s_1^2$ est la plus grande des deux variances estimées, avec $\nu_1$ degrés de liberté.

## 7. Bibliographie


1. CHAREMZA W.W. et DEADMAN D.F. « New Directions in Econometric Practice » Edward Elgar , 1992
2. CUTHBERTSON K. , HALL S.G. et TAYLOR M.P. « Applied Econometric Techniques » Philip Allan 1992
3. DESAI M. « Testing Monetarism »Frances Pinter , Londres 1981
4. DODGE Y. « Statistique . Dictionnaire Encyclopédique » Dunod , 1993
5. DOUGHERTY « Introduction to Econometrics » Oxford University Press 1992
6. ENGLE R.F. et GRANGER C.W.J. (eds) « Long Run Economic Relations : Readings in Cointegration « Oxford University Press 1991
7. GHOSH S.K. « Econometrics » Prentice Hall 1991
8. GRANGER , C.W. J. et NEWBOLD P. « Forecasting Economic Time Series » Academic Press , 1986
9. GRIFFITHS W.E. , HILL R.C. et JUDGE G.G. « Learning and Practicing Econometrics » John Wiley & Sons 1993 GUJARATI D.N. « Basic Econometrics » Mac Graw Hill 1988
10. HARVEY « The Econometric Analysis of Time Series « Philip Allan 1990
11. JUDGE, G.G. , HILL, R.C. , GRIFFITHS W.E. , LUTKEPOHL H. et LEE , T.C. « Introduction to the Theory and Practice of Econometrics « John Wiley & Sons , 1988
12. LUCAS R.E. « Economic Policy Evaluation : A Critique « in Brunner K. et Metzler A.M. (eds) « The Phillips Curve and Labor Markets « Carnegie Rochester Conference Series on Public Policy , vol. 1 , North Holland (1976)
13. MACKINNON , J.G. « Critical Values for Cointegration Tests « in R.F. Engle et C.W.J Granger (eds) 1991.
14. PINDYCK S.R. et RUBINFIELD D.L. « Econometric Models & Econometric Forecasts « Mc Graw Hill 1991

15. SALVATORE « Econométrie et Statistique Appliquées » Mc Graw Hill 1985 .
16. SIMS « Money Income and Causality « American Economic Review 1972 .
17. SIMS C.A. « Macroeconomics and Reality « Econometrica , vol .48, 1980 .
18. WONNACOTT T.H et R . J « Statistique « Economica 1988 .
19. WONNACOTT T.H et R.J « Econometrics » John Wiley 1979 .