UNIVERSITÉ JOSEPH FOURIER UFR IMA
M1 MIAGE

COURS
DE
CALCULS FINANCIERS
ET
STATISTIQUE

Serge Dégerine

4 octobre 2007

INTRODUCTION

Ce document comporte trois parties consacrées à deux thèmes très indépendants : les *Calculs Financiers* et la *Statistique*. Le point commun entre ces deux thèmes, dans la gestion des entreprises, est le recours à des techniques numériques et graphiques faisant appel à des notions mathématiques.

Les *Calculs Financiers* constituent la première partie de ce cours. Les notions introduites forment la base indispensable pour comprendre et analyser les produits bancaires ordinaires. Le premier chapitre présente la gestion d'un livret d'épargne. On y indique les règles communes à la plupart des livrets permettant de *calculer les intérêts*. Le chapitre suivant introduit les notions fondamentales d'*intérêts simples* et d'*intérêts composés* qui régissent la plupart des calculs financiers. Les notions liées sont celles de *taux proportionnels* et *taux équivalents*, *valeur acquise* et *valeur actuelle* et l'*équivalence de capitaux*. La *mesure de l'inflation*, qui fait l'objet du troisième chapitre, est une illustration des calculs à intérêts composés. On détermine, à partir de l'*indice des prix à la consommation*, les différents *taux d'inflation*. On revient sur l'*actualisation* et on précise les notions d'*euros constants* et d'*euros courants*. Enfin, le dernier chapitre traite le problème du *remboursement d'un emprunt*, avec la construction du *tableau d'amortissement*. On évoque aussi les *frais de dossier*, l'*assurance* afin de dégager le *taux effectif global*.
Les éléments mathématiques nécessaires à cette première partie sont les suites arithmétiques et les suites géométriques. Les calculs seront organisés avec le tableur Excel.

Les deux autres parties de ce cours concernent le thème *Statistique*.

La *Régression linéaire* est présentée en deuxième partie. Il s'agit de l'étude d'une variable statistique que l'on cherche à expliquer, souvent à des fins de prévision, à l'aide d'autres variables. L'exemple classique consiste à donner une fourchette de poids raisonnable pour un individu dont on connaît la taille. Le chapitre 5 traite les *aspects descriptifs* de la *régression linéaire simple*,

dans laquelle il y a une seule variable explicative. On introduit la *droite de régression* associée aux *estimateurs des moindres carrés*. Le *coefficient de corrélation linéaire* permet de mesurer l'importance de la relation entre les deux variables. L'*analyse descriptive des résidus* constitue une première approche visuelle de la validité du modèle. Les *aspects inductifs* de la régression linéaire simple font l'objet du chapitre suivant. Ils reposent sur un *modèle probabiiliste*, faisant appel à la *loi normale*, qui permet de préciser les propriétés des estimateurs en termes de *biais* et *variance*. L'*inférence statistique* nécessite d'introduire deux nouvelles lois de probabilités issues de la loi normale : la *loi du chi-deux* et la *loi de Student*. Ceci permet de définir les *estimateurs studentisés* et la notion de *p-valeur* afin de quantifier la pertinence de la régression. Les *résidus studentisés* précisent la validation du modèle. Il est alors possible d'effectuer une *prévision* sous forme d'*intervalle de confiance*. Enfin, la *régression linéaire multiple*, dans laquelle il y a plusieurs variables explicatives, est présentée au chapitre 7. On retrouve l'ensemble des notions introduites en régression linéaire simple dans ce cadre plus général. En particulier, le *coefficient de corrélation linéaire multiple* mesure l'importance de la dépendance entre la variable d'intérêt et l'ensemble des variables explicatives. Le *test de Fisher* et la p-valeur associée sont encore là pour juger de la pertinence de cette régression.
Les éléments mathématiques nécessaires à cette seconde partie relèvent de l'analyse de base, pour ce qui concerne la régression linéaire simple, et de l'algèbre linéaire pour la régression linéaire multiple. Plus précisément, ce dernier point fait appel au calcul matriciel. Cette difficulté est surmontée à l'aide des logiciels. En effet, les calculs de la régression linéaire simple seront menés dans un premier temps sous Excel. Nous les retrouverons alors dans le cadre du logiciel statistique R. L'extension à la régression linéaire multiple sera alors immédiate sous R.

Ce cours se termine avec une troisième partie consacrée à l'étude des *Séries Chronologiques*. Une série chronologique, ou série temporelle, est constituée d'observations effectuées régulièrement au cours du temps. Le tableau de bord de toute entreprise regorge de ce type de données, ne serait-ce que son chiffre d'affaires mensuel. L'objectif est de dégager une tendance, dans l'évolution de la grandeur étudiée, mais aussi un éventuel effet saisonnier, souvent à des fins de prévision. Le chapitre 8 se réduit aux *Généralités* donnant le cadre et le vocabulaire attachés à ce type d'étude. On y précise les notions de *tendance* et d'*effet saisonnier*, en particulier grâce à des représentations graphiques pertinentes. La distinction entre modèle additif et modèle multiplicatif est également discutée. Le chapitre 9, intitulé *Modèle de Byus-Ballot et Prévision*, se place dans le cadre du modèle additif. La chronique est

constituée de la somme de trois composantes : une *tendance linéaire*, un effet saisonnier périodique matérialisé par des *coefficients saisonniers* et un terme d'erreur. Le principe des moindres carrés, utilisé en régression linéaire, est appliqué ici de façon analogue. On retrouve ainsi les estimateurs des moindres carrés, leurs versions studentisées et les p-valeurs associées, pour juger de la pertinence de la tendance et/ou des coefficients saisonniers, les résidus studentisés, pour la validation du modèle, la prévision par intervalle de confiance et enfin le test de Fisher, pour juger de la présence ou non de l'effet saisonnier dans son ensemble. Le dernier chapitre, *Lissage et Série CVS*, se place, comme le précédent, dans le cadre du modèle additif. La différence est que la tendance n'est plus linéaire et est alors estimée par lissage. L'objectif n'est plus la prévision, mais l'estimation de l'effet saisonnier afin de le neutraliser pour constituer la *série CVS* (Corrigée des Variations Saisonnières). Les données économiques sont souvent exprimées ainsi, car elles sont plus pertinentes. Par exemple, le chômage exprimé en données brutes peut augmenter en été, alors qu'il diminue en données CVS.
Les éléments mathématiques nécessaires à cette dernière partie restent au niveau de l'analyse de base. Les calculs seront effectués sous Excel.

Le thème *Statistique* présenté ici n'est pas une entrée en la matière dans ce domaine, sans toutefois exiger de pré-requis solides. Des rudiments de *Statistique descriptive* et de *Calcul des Probabilités* figurent maintenant dans les programmes du lycée.

Bibliographie

Il est clair que le moyen le plus simple d'obtenir des compléments d'information est de faire une recherche sur le web à partir des mots-clef (Google). Certains cites sont d'ailleurs indiqués dans la partie consacrée aux calculs financiers. De plus, quelques ouvrages, pouvant être consultés à la bibliothèque, figurent en bibliographie.

Remarque

À la fin de chaque chapitre, une rubrique intitulée *En résumé* indique les éléments essentiels à retenir.

PREMIÈRE PARTIE

CALCULS FINANCIERS

Chapitre 1

GESTION D'UN LIVRET D'ÉPARGNE

1.1 Introduction

Nous commençons cette première partie par l'étude de la gestion du produit financier le plus populaire : le *livret d'épargne*. Le Tableau 1.1 indique de façon sommaire les principales caractéristiques des livrets d'épargne classiques. Pour un complément d'information, on pourra effectuer une recherche "livret d'épargne" sur le site http ://www.service-public.fr/. Ces caractéristiques sont en effet fixées par décrets gouvernementaux et ne dépendent pas de l'organisme bancaire gérant le produit. Par exemple, le taux du livret A a été fixé à 3,00% à partir du 1^{er} août 2007. Auparavant, il était de 2,75 %.

Produit	**Taux annuel**	**Capital**	**Conditions**
Livret A des caisses d'épargne	3%	1,5 € $\leq C \leq$ 15300 €	sans
Livret B des caisses d'épargne	libre	$C \geq$ 1,52 €	fiscalisé
Livret d'épargne entreprise	2,25%	15,24 € $\leq C \leq$ 4600 €	sans
Livret d'épargne populaire	livret A + 1%	30 € $\leq C \leq$ 7700 €	$IR \leq$ 722 €
Livret jeune	$\geq$ livret A	15,24 € $\leq C \leq$ 1600 €	12-25 ans

TAB. 1.1 – Caractéristiques de quelques livrets d'épargne classiques

Nous présentons ci-après les principes de la gestion d'un livret d'épargne, puis nous proposons une méthode pour organiser le calcul des intérêts.

1.2 Les principes de la gestion d'un livret

La gestion d'un livret d'épargne, du type de ceux présentés dans le Tableau 1.1, est basée sur les principes suivants :

Quinzaines : L'année civile est découpée en *quinzaines* qui débutent le 1^{er} et le 16 de chaque mois. Elles se terminent donc le 15 ou le dernier jour du mois et le nombre de jours variable de ces quinzaines n'est pas pris en compte.

Versement : Un *versement* produit des intérêts à partir de la quinzaine qui suit immédiatement la date de versement. Ceci vaut également pour le versement initial à l'ouverture du livret. Ainsi, un versement effectué le 1^{er} du mois ne prendra effet qu'à partir du 16 de ce même mois.

Retrait : Un *retrait* est décompté du capital productif d'intérêts dès le début de la quinzaine qui recouvre la date de ce retrait. Ainsi, un retrait effectué le 15 du mois prend effet dès le 1^{er} de ce même mois.

Intérêts : Les *intérêts* produits au cours d'une même année civile sont capitalisés, c'est-à-dire produisent eux-mêmes des intérêts, dès le 1^{er} janvier de l'année suivante.

Une bonne gestion consistera donc à effectuer ses versements en fin de quinzaine alors que les retraits se feront en début de quinzaine. Attention, un versement d'une certaine somme, suivie de son retrait au cours d'une même quinzaine, a pour effet de produire des intérêts négatifs !

En vertu de ces principes, un *capital* C, maintenu constant sur un livret pendant n quinzaines ($n \leq 24$), ne recouvrant ni 1^{er} janvier, ni date de changement de taux, produit les intérêts I donnés par

$$I = C \times i \times \frac{n}{24},$$

où i est le *taux d'intérêt annuel* en vigueur pendant cette période. Depuis le 1^{er} juillet 2004, les règles de fixation des taux des livrets réglementés sont arrêtées. Par exemple, la banque de France détermine le 15 janvier et le 15 juillet de chaque année, le taux d'intérêt du Livret A en fonction du taux d'inflation et du taux interbancaire de la zone euro pour la rémunération des dépôts (*cf.* http ://www.cbanque.com/placement/taux_livreta.php). Le Tableau 1.2 indique l'évolution de ce taux depuis 2000.

Date	1/07/2000	1/08/2003	1/08/2004	1/08/2005	1/02/2006	1/08/2006	1/08/2007
Taux	3,00 %	2,25 %	2,25 %	2,00 %	2,25 %	2,75 %	3,00 %

TAB. 1.2 – Évolution du taux d'intérêt annuel du Livret A depuis 2000

1.3 Méthode de calcul des intérêts

Afin de présenter la méthode de *calcul des intérêts* que nous préconisons, nous proposons de considérer l'exemple d'un livret d'épargne populaire.

1.3.1 Historique du livret

- Ouverture le 5 février 2002 avec un dépôt initial de 200 € .
- Retrait de 50 € le 15 avril 2002.
- Dépôt de 150 € le 26 juin 2002.
- Retrait de 100 € le 16 décembre 2002.
- Dépôt de 250 € le 1^{er} mars 2003.
- Le taux d'intérêt annuel en vigueur sur la période est $i = 3,00\%$.

L'objectif est de calculer le capital productif d'intérêts et les intérêts non capitalisés à la date du 16 mai 2003.

1.3.2 Conventions et notations

- Le temps t est mesuré en quinzaines et l'origine, $t = 0$, est fixée au 1^{er} janvier de la première année concernant l'étude. Ainsi t symbolise à la fois une date associée à un début de quinzaine et le nombre de quinzaines écoulées entre le 1^{er} janvier d'origine et cette date t.
- $C(t)$ est le capital productif d'intérêts à la date t.
- $I(t)$ sont les intérêts non capitalisés à la date t.
- I_{aa} sont les intérêts produits au cours de l'année aa.

En ce qui concerne le livret étudié, les faits marquants, dans l'ordre chronologique, se présentent alors comme suit :

– origine :			1^{er} jan. 02	$\rightarrow$	$t = 0$		
– ouverture :	5 fév. 02	$\rightarrow$	16 fév. 02	$\rightarrow$	$t = 3$	$\rightarrow$	+200 €
– retrait :	15 avr. 02	$\rightarrow$	1^{er} avr. 02	$\rightarrow$	$t = 6$	$\rightarrow$	-50 €
– dépôt :	26 juin 02	$\rightarrow$	1^{er} juil. 02	$\rightarrow$	$t = 12$	$\rightarrow$	+150 €
– retrait :	16 déc. 02	$\rightarrow$	16 déc. 02	$\rightarrow$	$t = 23$	$\rightarrow$	-100 €
– intérêts 02 :		$\rightarrow$	1^{er} jan. 03	$\rightarrow$	$t = 24$	$\rightarrow$	$+I_{02}$
– dépôt :	1^{er} mars 03	$\rightarrow$	16 mars 03	$\rightarrow$	$t = 29$	$\rightarrow$	+250 €

Il est souhaitable de résumer ces informations sur un schéma permettant de visualiser l'historique du livret (*cf.* Figure 1.1). La date $t = 33$ correspond au 16 mai 03 pour indiquer le calcul de $C(33)$ et $I(33)$.

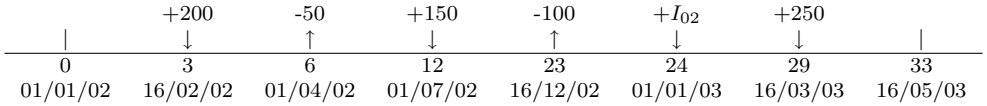

FIG. 1.1 – Historique du livret à l'étude

1.3.3 Calcul des intérêts chaque quinzaine

Il est particulièrement simple, à l'aide d'un tableur, de déterminer l'évolution du capital $C(t)$ et des intérêts $I(t)$, quinzaine par quinzaine, sur la période considérée. En effet le capital $C(t)$ évolue selon les dépôts et retraits et les intérêts $I(t)$ satisfont $I(t+1) = I(t) + C(t) \times \frac{i}{24}$. Il est cependant nécessaire d'arrondir, au centime le plus proche, les intérêts capitalisés le 1^{er} janvier. Ces résultats sont reportés dans le Tableau 1.3.

t	$C(t)$	$I(t)$	t	$C(t)$	$I(t)$	t	$C(t)$	$I(t)$
2	0	0	13	300	2,25	24	206,25	0
3	200	0	14	300	2,63	25	206,25	0,26
4	200	0,25	15	300	3,00	26	206,25	0,52
5	200	0,50	16	300	3,38	27	206,25	0,77
6	150	0,75	17	300	3,75	28	206,25	1,03
7	150	0,94	18	300	4,13	29	456,25	1,29
8	150	1,13	19	300	4,50	30	456,25	1,86
9	150	1,31	20	300	4,88	31	456,25	2,43
10	150	1,50	21	300	5,25	32	456,25	3,00
11	150	1,69	22	300	5,63	33	456,25	3,57
12	300	1,88	23	200	6,00			

TAB. 1.3 – Capital $C(t)$ et intérêts $I(t)$ au cours des quinzaines

Lorsque le capital $C(t)$ est constant, les intérêts $I(t)$ forment une progression arithmétique de raison $C(t) \times \frac{i}{24}$. Cette dernière quantité ne doit pas être arrondie dans les calculs. C'est la raison pour laquelle les intérêts $I(t)$, arrondis au centime le plus proche, qui figurent dans le Tableau 1.3 ne constituent pas exactement des suites arithmétiques sur les périodes où le

capital $C(t)$ reste constant.

En fait, la banque détermine, au début de chaque quinzaine, les intérêts $I_{aa}(t)$ qui seraient capitalisés au 1^{er} janvier de l'année suivante, si le livret restait en l'état jusqu'à cette date. Ces résultats sont reportés dans le Tableau 1.4.

t	$C(t)$	$I_{02}(t)$	t	$C(t)$	$I_{02}(t)$	t	$C(t)$	$I_{03}(t)$
2	0	0	13	300	7,63	24	206,25	6,19
3	200	5,25	14	300	7,63	25	206,25	6,19
4	200	5,25	15	300	7,63	26	206,25	6,19
5	200	5,25	16	300	7,63	27	206,25	6,19
6	150	4,13	17	300	7,63	28	206,25	6,19
7	150	4,13	18	300	7,63	29	456,25	12,13
8	150	4,13	19	300	7,63	30	456,25	12,13
9	150	4,13	20	300	7,63	31	456,25	12,13
10	150	4,13	21	300	7,63	32	456,25	12,13
11	150	4,13	22	300	7,63	33	456,25	12,13
12	300	6,38	23	200	6,25			

TAB. 1.4 – Capital $C(t)$ et intérêts annuels $I_{aa}(t)$ selon la banque de France

1.3.4 Calcul direct des intérêts

Il n'est évidemment pas nécessaire de réaliser de tels tableaux pour déterminer les intérêts annuels ou ceux produits à une date donnée. Il est cependant indispensable de calculer les intérêts annuels afin de les arrondir avant de les capitaliser. De même, en cas de changement de taux sur la période considérée, il faudra déterminer le montant arrondi des intérêts selon les différents taux. Pour le livret considéré ici, on calcul tout d'abord les intérêts I_{02} en se référant à l'historique de la Figure 1.1 :

$$\begin{aligned} I_{02} &= \{200 \times (24-3) - 50 \times (24-6) + 150 \times (24-12) - 100 \times (24-23)\} \\ &\times \frac{0,0300}{24} = 5000 \times \frac{0,0300}{24} = 6,25 \text{ €} . \end{aligned}$$

Le capital productif d'intérêts au 1^{er} janvier 2003 est alors :

$$C(24) = 200 - 50 + 150 - 100 + 6,25 = 206,25 \text{ €} .$$

Le capital et les intérêts non capitalisés au 16 mai 2003 sont :

$$\begin{aligned} C(33) &= 206,25 + 250 = 456,25 \text{ €} , \\ I(33) &= \{206,25 \times (33-24) + 250 \times (33-29)\} \times \frac{0,0300}{24} \\ &= 2856,25 \times \frac{0,0300}{24} = 3,57031125 \simeq 3,57 \text{ €} . \end{aligned}$$

1.3.5 À propos des arrondis

De manière générale, on ne doit pas arrondir le résultat d'un calcul intermédiaire, seules les informations communiquées au client doivent l'être. Ayant fait le choix d'une certaine précision, l'*arrondi* est effectué à la valeur la plus proche, ou à la valeur supérieure en cas d'égalité, sans tenir compte du signe. La valeur la plus proche relève du bon sens, bien qu'il existe des situations où le choix est différent (déclaration des revenus). Par contre, la valeur supérieure en cas d'égalité est due à une convention. Cette convention est celle retenue par les fonctions d'affichage en informatique (calculatrices ou logiciels). Un avantage, par rapport à la valeur inférieure, est qu'il suffit de connaître un chiffre de plus que la précision requise pour effectuer son arrondi sans erreur. En effet, en arrondissant au centime les sommes 4,2149 et 4,21501, on obtient 4,21 et 4,22 au vu des trois premiers chiffres après la virgule, alors que l'autre convention conduirait à 4,21 dans les deux cas. Par contre, cette convention ne permet pas d'arrondir de façon récursive : 4,2149 devient 4,215 à trois chiffres puis 4,22 au lieu de 4,21, mais c'est également le cas de l'autre : 4,21501 donne 4,215 puis 4,21 au lieu de 4,22.

En résumé Être capable de mener à bien tout calcul relatif à un livret d'épargne.

Chapitre 2

INTÉRÊTS SIMPLES ET INTÉRÊTS COMPOSÉS

2.1 Introduction

Nous présentons ici les deux modes de calcul d'intérêts : intérêts simples et intérêts composés. Ils font alors apparaître les notions de suites arithmétiques et suites géométriques. Ensuite, nous leur associons les notions de taux proportionnels et taux équivalents. Puis, nous déclinons plusieurs définitions, directement attachées aux calculs d'intérêts, plus particulièrement selon le mode composé qui reste le plus fréquemment utilisé : valeur actuelle, valeur acquise, équivalence de capitaux,...

2.2 Les deux conventions fondamentales

Notons tout d'abord qu'un *taux d'intérêt* est toujours relatif à une *période*. Par exemple, 3% est le taux annuel actuel du livret A. Le Tableau 2.1 précise le vocabulaire des taux usuels.

Un capital C, placé pendant une période au taux d'intérêt i par période, produit les intérêts $I_1 = Ci$, par définition du taux.

Un capital C, placé pendant n périodes ($n \geq 2$) au taux d'intérêt i par période, produit des intérêts I_n qui dépendent de la convention adoptée pour effectuer les calculs. Les deux conventions fondamentales sont les suivantes.

Période	**Taux**	**Période**	**Taux**
jour	journalier	semaine	hebdomadaire
quinzaine	bimensuel	mois	mensuel
deux mois	bimestriel	trois mois	trimestriel
six mois	semestriel	année	annuel
deux ans	biennal	trois ans	triennal
cinq ans	quinquennal	dix ans	décennal

TAB. 2.1 – Quelques taux usuels

2.2.1 Intérêts simples

Dans un calcul à *intérêts simples*, les intérêts produits au cours d'une période ne sont pas capitalisés (ne produisent pas d'intérêts) au cours des périodes suivantes. Dans ce cas, le capital reste constant et produit les intérêts Ci à chaque période. Les intérêts I_n produits au bout de n périodes sont donc :

$$I_n = Cin, \quad n = 0, 1, \ldots$$

La suite $I_n, n = 0, 1, \ldots$, est une *suite arithmétique* de premier terme $I_0 = 0$ et de raison Ci, car la différence $I_{n+1} - I_n$ est constamment égale à Ci. Notant $C_n = C + I_n$, la valeurs acquise par ce capital au bout de n périodes, on a :

$$C_n = C + I_n = C + Cin = C(1 + ni), \quad n = 0, 1, \ldots$$

La suite $C_n, n = 0, 1, \ldots$, est également arithmétique de premier terme $C_0 = C$ et de raison Ci, car $C_{n+1} - C_n = Ci$ pour tout n.

Notons que les points $\{n, I_n\}, n = 0, 1, \ldots$, se situent sur la droite de pente Ci passant par l'origine, puisque $I_n = Cin$. On dit que la croissance de I_n est linéaire. Il en est de même pour C_n, puisque les points $\{n, C_n\}, n = 0, 1, \ldots$, se situent sur la droite d'équation $C_n = Cin + C$.

2.2.2 Intérêts composés

Dans un calcul à *intérêts composés*, les intérêts produits au cours d'une période sont capitalisés (produisent eux-mêmes des intérêts) dès la période suivante. Dans ce cas, le capital (productif d'intérêts) augmente à chaque période. Notant $C_n, n \geq 0$, ce capital, on a $C_0 = C$, puis (raisonnement par récurrence) :

$$C_1 \quad = \quad C_0 + C_0 i = C_0(1 + i) = C(1 + i),$$

$$\begin{aligned} C_2 &= C_1 + C_1 i = C_1(1+i) = C(1+i)^2, \\ &\text{etc} \\ C_n &= C_{n-1} + C_{n-1} i = C_{n-1}(1+i) = C(1+i)^n. \end{aligned}$$

Ici, $C_n, n \geq 0$, est une *suite géométrique*, de premier terme $C_0 = C$ et de raison $1+i$, car le rapport $\frac{C_{n+1}}{C_n}$ est constamment égal à $1+i$. La suite des intérêts,

$$I_n = C_n - C = C(1+i)^n - C = C[(1+i)^n - 1], \quad n \geq 0,$$

n'a pas de structure particulière, si ce n'est qu'elle évolue comme une suite géométrique puisque $I_n = C_n - C$.

En écrivant C_n sous la forme,

$$C_n = \exp\{\log(C_n)\} = \exp\{n\log(1+i) + \log(C)\} = C\exp\{n\log(1+i)\},$$

on constate que les points $\{n, C_n\}, n = 0, 1, \ldots$, sont sur le graphe d'une fonction exponentielle. On dit que la croissance de C_n est exponentielle. On dit aussi que la croissance de I_n est exponentielle, mais dans ce cas les points $\{n, I_n\}, n = 0, 1, \ldots$, sont sur le graphe précédent, décalé vers le bas de la quantité C.

2.2.3 Cas d'un livret d'épargne

Dans le cas d'un livret d'épargne, on rencontre les deux modes de calcul : intérêts simple à l'échelle de la quinzaine au cours d'une même année civile et intérêts composés à l'échelle de l'année. En reprenant la formule de calcul des intérêts sous la forme,

$$I_n = C \times i \times \frac{n}{24} = C \times \frac{i}{24} \times n,$$

on reconnaît le placement de C à intérêts simples, pendant n quinzaines, au taux bimensuel $i_q = \frac{i}{24}$. On dit que i_q est le taux bimensuel proportionnel au taux annuel i. Ce mode de calcul à intérêts simples peut sembler défavorable au client. Cependant, si on procédait à intérêts composés, le taux bimensuel à appliquer ne serait pas le taux proportionnel i_q, mais le taux équivalent i'_q défini par

$$i'_q = \sqrt[24]{1+i} - 1 = (1+i)^{\frac{1}{24}} - 1.$$

En effet, ce taux provient de

$$C(1+i'_q)^{24} = C(1+i),$$

qui exprime qu'un capital C, placé sur un livret pendant un an au taux annuel i, produit les mêmes intérêts que placé, à intérêts composés, pendant 24 quinzaines au taux bimensuel i'_q.

Dans le cas du taux annuel $i = 3,00\%$ du livret A, les taux proportionnel i_q et équivalent i'_q sont donnés par :

$$\begin{aligned} i_q &= \frac{3,00\%}{24} = 0,125\%, \\ i'_q &= (1 + 3,00\%)^{\frac{1}{24}} - 1 = 0,11232375512\%. \end{aligned}$$

Notons qu'il ne faut pas arrondir ce type de taux, sous peine de résultats erronés là où ils seraient utilisés. On constate que i'_q est légèrement inférieur à i_q, ce qui était prévisible du fait de leur signification.

On peut avoir le sentiment que le fait d'effectuer les calculs à intérêts simples désavantage le client. En fait, l'avantage se fait en faveur de l'une ou de l'autre convention selon les situations.
Considérons le cas d'un livret A sur lequel on dépose 1000 € fin décembre

t	$C(t)$	$I(t)$	$K(t)$	$J(t)$	t	$C(t)$	$It)$	$K(t)$	$J(t)$
0	1000	0	1000	0	13	0	10,00	9,96	9,96
1	1000	0,83	1000,83	0,83	14	0	10,00	9,97	9,97
2	1000	1,67	1001,65	1,65	15	0	10,00	9,97	9,97
3	1000	2,50	1002,48	2,48	16	0	10,00	9,98	9,98
4	1000	3,33	1003,31	3,31	17	0	10,00	9,99	9,99
5	1000	4,17	1004,13	4,13	18	0	10,00	10,00	10,00
6	1000	5,00	1004,96	4,96	19	0	10,00	10,01	10,01
7	1000	5,83	1005,79	5,79	20	0	10,00	10,02	10,02
8	1000	6,67	1006,62	6,62	21	0	10,00	10,02	10,02
9	1000	7,50	1007,45	7,45	22	0	10,00	10,03	10,03
10	1000	8,33	1008,29	8,29	23	0	10,00	10,04	10,04
11	1000	9,17	1009,12	9,12	24	10,00	0	10,05	10,05
12	0	10,00	9,95	9,95					

TAB. 2.2 – Intérêts simples ou composés pour un livret

2004 et que l'on retire début juillet 2005. Le Tableau 2.2 donne l'évolution de ce livret, pour cette opération, au cours de l'année 2005, avec un taux

annuel de 2%. $C(t)$ et $I(t)$ représentent le capital (productif d'intérêts) et les intérêts non capitalisés selon le mode de calcul usuel à intérêts simples. $K(t)$ est le capital productif d'intérêts selon le calcul à intérêts composés et $J(t)$ la partie due aux intérêts dans ce capital. On constate une différence de 5 centimes d'Euros, en faveur de cette dernière formule, au 1^{er} janvier 2006. Par contre, au 1^{er} juillet 2005, les intérêts potentiels sont supérieurs de 5 centimes d'Euros pour la formule à intérêts simples. Ceci signifie qu'un dépôt de 1000 € fin juin 2005 rapportera 5 centimes d'Euros de plus, à intérêts simples, au 1^{er} janvier 2006. Par ailleurs, il faut noter que la différence est très faible : 5 centimes d'Euros maximum pour un capital de 1000 € et un taux de 2%.

2.3 Taux proportionnels et taux équivalents

Nous précisons ici ces notions, introduites ci-dessus pour un livret, dans un cadre général. Soit i un taux relatif à une période de durée d et i' un taux relatif à une période de durée d'. Les durées d et d' sont exprimées dans la même unité de temps : jours, semaines, quinzaines, mois, années. On place alors un capital C pendant une durée D, qui soit à la fois un multiple de d et de d' : $D = nd = n'd'$.

2.3.1 Taux proportionnels

On écrit que les intérêts I, produits par C pendant la durée D à intérêts simples, sont les mêmes en effectuant les calculs avec i ou i' :

$$I = C \times i \times n = C \times i' \times n' \quad \Longrightarrow \quad \frac{i'}{i} = \frac{n}{n'} = \frac{d'}{d} \quad \Leftrightarrow \quad i' = \frac{d'}{d}\, i.$$

On dit que i' et i sont des *taux proportionnels*. Par exemple, 0,5% est le taux mensuel proportionnel au taux annuel de 6% ($d' = 1$ mois, $d = 12$ mois). Mais aussi, 6% est le taux annuel proportionnel au taux mensuel de 0,5%.

2.3.2 Taux équivalents

On écrit ici que le devenir C_D du capital C, placé pendant la durée D à intérêts composés, est le même en effectuant les calculs avec i ou i' :

$$C_D = C(1+i)^n = C(1+i')^{n'} \quad \Longrightarrow \quad 1 + i' = (1+i)^{\frac{n}{n'}} = (1+i)^{\frac{d'}{d}}$$
$$\Leftrightarrow \quad i' = (1+i)^{\frac{d'}{d}} - 1.$$

On dit que i' et i sont des *taux équivalents*. Par exemple 0,4867550565% est le taux mensuel équivalent au taux annuel de 6% et 6% est le taux annuel équivalent au taux mensuel de 0,4867550565%. Notons que le taux annuel équivalent au taux mensuel de 0,5% est de 6,167781186%. Ceci souligne la nécessité d'être prudent dans la façon d'arrondir les taux.

2.4 Valeur acquise par un capital

Soit C un capital placé pendants t années au taux annuel i. On appelle *valeur acquise* par C la quantité $C_t = C + I_t$, somme du capital et des intérêts produits.

- Si t est entier, $C_t = C(1 + ti)$, à intérêts simples, et $C_t = C(1 + i)^t$, à intérêts composés, selon les résultats établis précédemment.
- Si t n'est pas entier, on peut le convertir en années, mois et jours, moyennant la convention simplificatrice que, dans n'importe quel mois, il y a 30 jours et donc 360 jours dans l'année. Ainsi t correspond à n années entières, m mois et p jours. On distingue alors les deux formes de placement.

2.4.1 Calcul à intérêts simples

$$C_t = C\left(1 + ni + m\frac{i}{12} + p\frac{i}{360}\right) = C\left(1 + \left[n + \frac{m}{12} + \frac{p}{360}\right] i\right),$$

où $\frac{i}{12}$ et $\frac{i}{360}$ sont respectivement les taux mensuel et journalier proportionnels au taux annuel i. Notons que $n + \frac{m}{12} + \frac{p}{360}$ est la durée t du placement, exprimée en années. On écrira donc,

$$C_t = C(1 + ti),$$

en notant bien que la durée t, pas nécessairement entière, est exprimée dans l'unité de temps définie par la période de référence du taux i.

2.4.2 Calcul à intérêts composés

$$C_t = C(1 + i)^n (1 + i)^{\frac{m}{12}} (1 + i)^{\frac{p}{360}} = C(1 + i)^{\left[n + \frac{m}{12} + \frac{p}{360}\right]},$$

où $(1+i)^{\frac{m}{12}}-1$ et $(1+i)^{\frac{p}{360}}-1$ sont respectivement les taux mensuel et journalier équivalents au taux annuel i. On peut alors écrire,

$$C_t = C(1+i)^t,$$

où t est la durée du placement exprimée dans l'unité de temps définie par la période de référence du taux i.

2.5 Valeur actuelle d'un capital

On raisonne désormais uniquement à intérêts composés. Ceci correspond en effet à la pratique courante dans les calculs financiers, le cas des livrets d'épargne restant l'exception la plus connue.

2.5.1 Valeur actuelle

Le temps t est exprimé en années, généralement sous forme décimale, par référence à une date d'origine, $t = 0$, faisant office de date actuelle liée au problème considéré. Ainsi t symbolise à la fois une date et le temps écoulé entre cette date et la date actuelle. Soit C_t un capital considéré à un instant $t > 0$. Pour un taux d'intérêt annuel i, ce capital représente la valeur acquise par un capital C_0, affecté à l'instant d'origine, si $C_t = C_0(1+i)^t$. On dit que C_0 est la *valeur actuelle* de C_t, qui est alors donnée par :

$$C_0 = C_t(1+i)^{-t}.$$

L'objectif de cette notion est de pouvoir évaluer maintenant (à la date actuelle) différentes sommes d'argent (recettes ou dépenses) programmées dans le futur. Dans ce cadre, i est appelé *taux d'actualisation*. Notons qu'une valeur acquise peut s'interpréter comme la valeur actuelle d'un capital considéré dans le passé (t est alors négatif). Nous donnons ci-dessous quelques notions liées aux calculs de valeurs actuelles dans le cadre de la gestion de projet.

2.5.2 Gestion de projet

On considère un projet d'investissement, d'une valeur I_0 à la date $t_0 = 0$ (fixée comme origine), générant des flux de trésorerie F_k aux dates $t_k, k = 1, \ldots, n$.

On appelle *valeur actuelle nette* (VAN) de ce projet d'investissement, la différence entre la valeur actuelle F_0 des flux générés et l'investissement

initial I_0 :

$$VAN = F_0 - I_0 = \sum_{k=1}^{n} F_k(1+i)^{-t_k} - I_0,$$

où i est le taux d'actualisation requis, supposé constant sur la période. L'investissement est rentable lorsque sa VAN est positive et, plus elle est élevée, plus la rentabilité est forte.

Mais la VAN dépend du taux d'actualisation i utilisé dans son calcul. On constate qu'elle est une fonction décroissante par rapport à ce taux. Le *taux interne de rentabilité* (TIR) de cet investissement est alors défini comme étant le taux d'actualisation i pour lequel la VAN est nulle. Ce taux doit être déterminé de façon numérique, car l'équation $F_0 = I_0$ n'admet pas de solution explicite. Plus le TIR est élevé, plus l'investissement est rentable.

On définit également l'*indice de profitabilité* comme le rapport F_0/I_0. Il doit être supérieur à 1 ($F_0/I_0 = 1 \Leftrightarrow VAN = 0$). C'est le facteur par lequel le projet multiplie l'investissement, pour le taux d'actualisation considéré. Enfin le *délai de récupération* est la date à laquelle les flux de trésorerie F_k, cumulés et actualisés, dépassent pour la première fois l'investissement I_0. C'est donc le premier instant t_k pour lequel on a :

$$F_1(1+i)^{-t_1} + F_2(1+i)^{-t_2} + \ldots + F_k(1+i)^{-t_k} \geq I_0.$$

Vous trouverez des compléments sur ce thème dans le chapitre 6 de l'ouvrage *Finance* de Zvi Bodie et Robert Merton à l'adresse :
http ://www.escp-eap.net/publications/bmt/plan.html

2.6 Équivalence de capitaux

L'objectif est de comparer deux capitaux C_1 et C_2, liés à deux dates différentes t_1 et t_2 (*cf.* Figure 2.1), exprimées en années par rapport à une certaine origine.

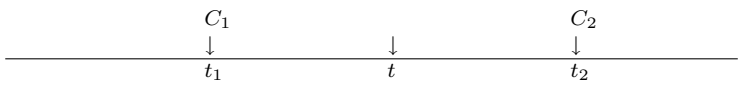

FIG. 2.1 – Équivalence de capitaux

On dit que C_1 et C_2 sont des *capitaux équivalents*, pour un taux d'actualisation (annuel) i, s'il existe une date t en laquelle ils ont la même valeur actuelle, ce qui s'écrit :

$$C_1(1+i)^{t-t_1} = C_2(1+i)^{t-t_2} \quad \Leftrightarrow \quad C_1(1+i)^{-t_1} = C_2(1+i)^{-t_2}$$

On constate que si cela est vrai pour une date t, ce sera encore le cas pour toute autre date. En supposant $t_2 > t_1$, le plus simple est d'écrire :

$$C_2 = C_1(1+i)^{t_2-t_1}.$$

Sous cette forme, C_2 apparaît comme le remboursement de C_1, prêté au taux i pendant la durée $t_2 - t_1$. On notera que C_2 est nécessairement supérieur à C_1, à moins d'admettre l'idée d'un taux négatif (déflation, par exemple).

On peut aussi s'interroger sur l'existence d'un taux i pour lequel les deux capitaux seraient équivalents. Ceci se traduit par :

$$\log(1+i) = \frac{1}{t_2-t_1}\log\left(\frac{C_2}{C_1}\right) \quad \Leftrightarrow \quad 1+i = \exp\left\{\frac{1}{t_2-t_1}\log\left(\frac{C_2}{C_1}\right)\right\}.$$

En acceptant la possibilité d'un taux négatif, cette équation a toujours une solution (unique), quelles que soient les capitaux (positifs) C_1 et C_2 et les dates t_1 et t_2 (différentes si $C_1 \neq C_2$).

Enfin, il est également possible d'exprimer la durée $t_2 - t_1$ en fonction des capitaux et du taux :

$$t_2 - t_1 = \frac{1}{\log(1+i)}\log\left(\frac{C_2}{C_1}\right).$$

En résumé Connaître les notions introduites (intérêts simples ou composés, taux proportionnels ou équivalents, valeur acquise ou actuelle, équivalence de capitaux) et savoir effectuer les calculs qui s'y rattachent.

Chapitre 3

MESURE DE L'INFLATION

3.1 Introduction

Le terme *inflation* évoque l'augmentation des prix et on parle de *déflation* lorsque les prix diminuent. La mesure de l'inflation est exprimée sous forme de taux, calculés à partir d'indices. Nous présentons ici l'indice des prix à la consommation usuel de l'INSEE. Nous examinerons ensuite les différents taux d'inflation attachés à cet indice : taux annuel glissant, calculé tous les mois, taux annuel et taux mensuel, taux moyens. Enfin, nous reviendrons sur l'actualisation en termes d'euros constants et d'euros courants.

3.2 Indice des prix à la consommation

L'*indice des prix à la consommation* (IPC) faisant référence, dans la communication des taux d'inflation par les médias, est celui déterminé par l'Institut National de la Statistique et des Études Économiques (INSEE, http ://www.insee.fr/). Cet indice est publié aux environs du 13 de chaque mois, pour le mois précédent. Il est basé sur 200 000 prix correspondant à plus de 1 000 types de biens et services, en respectant leur importance dans la consommation totale des ménages et actualisés chaque année. Pour cela, 160 000 prix sont relevés chaque mois dans près de 27 000 points de vente répartis dans 106 agglomérations de plus de 2 000 habitants, en métropole et dans les Dom. À ces prix relevés sur le terrain s'ajoutent près de 40 000 tarifs collectés directement auprès d'organismes nationaux ou régionaux tels qu'EdF, les opérateurs de télécommunications, la SNCF, les services publics locaux, ainsi que dans les catalogues de vente par correspondance (http ://www.insee.fr/fr/indicateur/indic_cons/info_ipc.htm).

Le Tableau 3.1 donne les valeurs de l'IPC de 1990 à août 2007, base 100 en 1998 (la moyenne de l'indice, sur les 12 mois de l'année 1998, est égale à 100).

année	J	F	M	A	M	J	J	A	S	O	N	D
2007	114,34	114,55	115,04	115,60	115,89	116,03	115,74	116,20				
2006	112,94	113,36	113,69	114,16	114,66	114,65	114,46	114,85	114,59	114,34	114,47	114,73
2005	110,7	111,3	112,0	112,2	112,3	112,5	112,3	112,7	113,2	113,1	112,9	113,0
2004	109,0	109,5	109,9	110,2	110,6	110,6	110,4	110,7	110,8	111,1	111,1	111,3
2003	106,9	107,6	108,1	107,9	107,8	108,0	107,9	108,1	108,5	108,8	108,9	109,0
2002	104,8	104,9	105,4	105,8	105,9	105,9	105,9	106,1	106,3	106,5	106,5	106,7
2001	102,5	102,8	103,2	103,7	104,4	104,4	104,2	104,2	104,4	104,5	104,2	104,3
2000	101,3	101,4	101,9	101,9	102,1	102,3	102,1	102,3	102,9	102,7	103,0	102,9
1999	99,7	100,0	100,4	100,6	100,6	100,6	100,4	100,5	100,7	100,8	100,8	101,3
1998	99,5	99,8	100,0	100,2	100,2	100,3	100,0	100,0	100,0	100,0	99,9	100,0
1997	98,9	99,1	99,2	99,2	99,3	99,3	99,2	99,4	99,6	99,6	99,7	99,8
1996	97,2	97,5	98,2	98,3	98,5	98,4	98,2	98,0	98,3	98,5	98,5	98,7
1995	95,3	95,6	95,9	96,0	96,1	96,2	96,0	96,4	96,8	96,9	97,0	97,0
1994	93,7	94,0	94,2	94,4	94,6	94,6	94,6	94,6	94,9	95,1	95,1	95,1
1993	92,0	92,3	92,8	92,9	93,0	93,0	93,1	93,1	93,4	93,5	93,6	93,6
1992	90,1	90,5	90,8	91,0	91,3	91,2	91,2	91,1	91,4	91,6	91,7	91,7
1991	87,8	88,0	88,2	88,4	88,7	88,9	89,2	89,3	89,4	89,9	90,2	89,9
1990	84,9	85,1	85,3	85,8	85,9	85,9	85,9	86,5	87,1	87,5	87,4	87,3

TAB. 3.1 – Indice INSEE des prix à la consommation de 1990 à août 2007

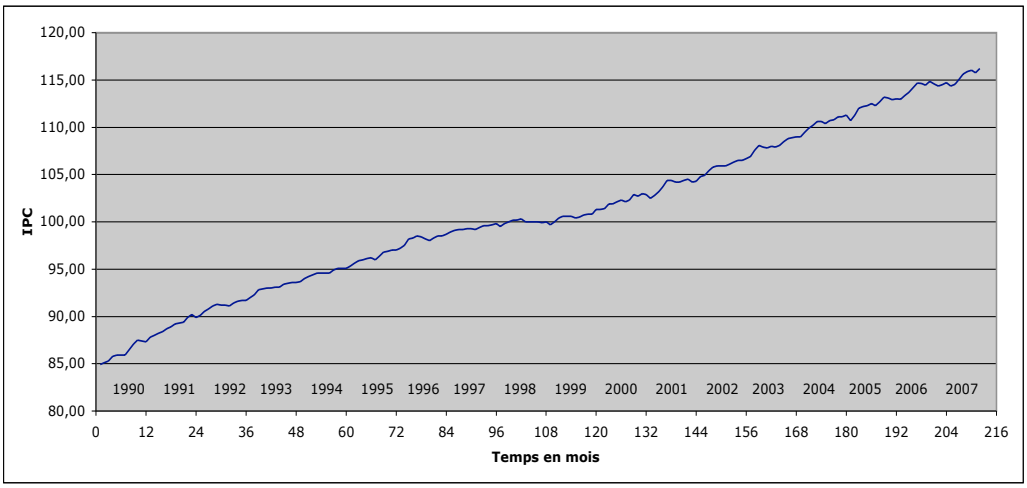

FIG. 3.1 – Évolution de l'indice des prix à la consommation de 1990 à août 2007

La Figure 3.1 permet de visualiser l'évolution de l'IPC de 1990 à août 2007. On constate une hausse régulière, avec un léger infléchissement, de 1990 à 1999, puis une reprise plus marquée à partir de l'année 2000.

L'INSEE publie de nombreux autres indices : l'IPC hors tabac, l'indice des prix à la consommation harmonisé (IPCH, indice européen), des indices

par famille de produits, ...

3.3 Taux d'inflation

Les *taux d'inflation* mesurent, sous différentes formes, les variations de l'IPC. Ils sont en effet plus expressifs pour le consommateur et ce sont eux qui sont communiqués par les médias.

3.3.1 Taux d'inflation annuel glissant

Le *taux d'inflation annuel glissant* indique chaque mois la variation relative des prix au cours des 12 derniers mois. En indexant le temps par la variable t, mesurée en mois, ce taux, noté $\tau_a(t)$, est donné par :

$$\tau_a(t) = \frac{I(t) - I(t-12)}{I(t-12)} = \frac{I(t)}{I(t-12)} - 1,$$

où $I(t)$ désigne l'IPC à la date t. La deuxième expression traduit la relation,

$$I(t) = I(t-12)[1 + \tau_a(t)],$$

à rapprocher d'un calcul à intérêts composés. Par exemple, le taux de janvier 2005 est calculé avec les indices de janvier 2004 et janvier 2005 :

$$\frac{110,7}{109,0} - 1 = 0,015596\ldots \simeq 1,56\%.$$

Le Tableau 3.2 donne les valeurs de ce taux de janvier 1991 à décembre 2005, dont l'évolution est représentée sur la Figure 3.2. Bien que l'information soit identique à celle contenue dans l'IPC, on constate que ce changement de variable conduit à un aspect visuel plus expressif. On observe une diminution plus ou moins régulière de l'inflation de 1991 jusqu'au milieu de l'année 1999, puis une forte reprise jusqu'en 2000, suivie d'une certaine stabilité.

3.3.2 Taux d'inflation annuel

Le *taux d'inflation annuel* mesure la variation des prix à l'échelle de l'année. Il correspond donc au taux glissant précédent relevé chaque mois de décembre. Ses valeurs sont reportées dans le Tableau 3.3 et représentées sur la Figure 3.3. On retrouve, de façon plus grossière, les grandes tendances observées sur le graphe du taux d'inflation annuel glissant.

année	J	F	M	A	M	J	J	A	S	O	N	D
2007	1,24	1,05	1,19	1,26	1,07	1,20	1,12	1,18				
2006	2,02	1,85	1,51	1,75	2,10	1,91	1,92	1,91	1,23	1,10	1,39	1,53
2005	1,56	1,64	1,91	1,81	1,54	1,72	1,72	1,81	2,17	1,80	1,62	1,53
2004	1,96	1,77	1,67	2,13	2,60	2,41	2,32	2,41	2,12	2,11	2,02	2,11
2003	2,00	2,57	2,56	1,98	1,79	1,98	1,89	1,89	2,07	2,16	2,25	2,16
2002	2,24	2,04	2,13	2,03	1,44	1,44	1,63	1,82	1,82	1,91	2,21	2,30
2001	1,18	1,38	1,28	1,77	2,25	2,05	2,06	1,86	1,46	1,75	1,17	1,36
2000	1,60	1,40	1,49	1,29	1,49	1,69	1,69	1,79	2,18	1,88	2,18	1,58
1999	0,20	0,20	0,40	0,40	0,40	0,30	0,40	0,50	0,70	0,80	0,90	1,30
1998	0,61	0,71	0,81	1,01	0,91	1,01	0,81	0,60	0,40	0,40	0,20	0,20
1997	1,75	1,64	1,02	0,92	0,81	0,91	1,02	1,43	1,32	1,12	1,22	1,11
1996	1,99	1,99	2,40	2,40	2,50	2,29	2,29	1,66	1,55	1,65	1,55	1,75
1995	1,71	1,70	1,80	1,69	1,59	1,69	1,48	1,90	2,00	1,89	2,00	2,00
1994	1,85	1,84	1,51	1,61	1,72	1,72	1,61	1,61	1,61	1,71	1,60	1,60
1993	2,11	1,99	2,20	1,98	1,86	1,97	2,08	2,20	2,19	2,07	2,07	2,07
1992	2,62	2,84	2,95	3,05	2,93	2,59	2,24	2,02	2,24	1,89	1,66	2,00
1991	3,42	3,41	3,40	3,03	3,26	3,49	3,84	3,24	2,64	2,74	3,20	2,98

TAB. 3.2 – Taux d'inflation annuel glissant de 1991 à août 2007

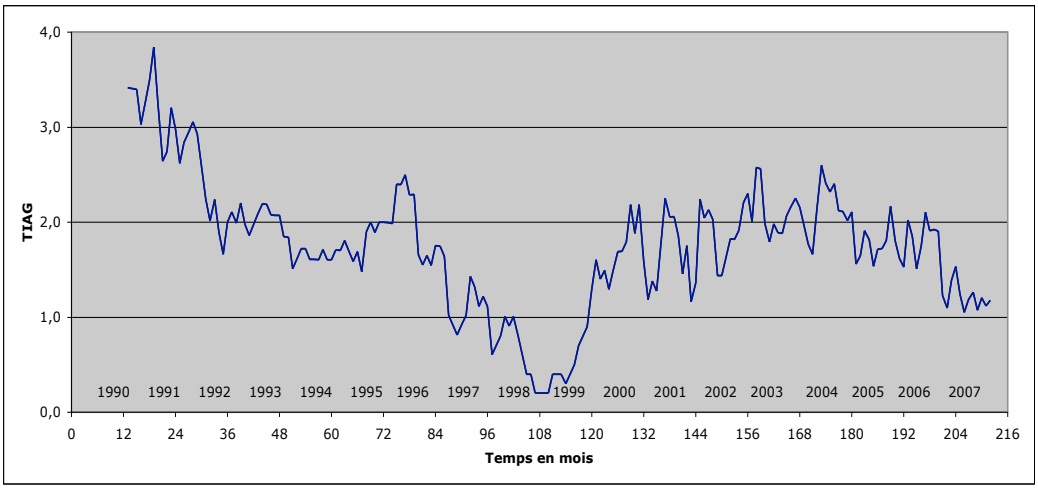

FIG. 3.2 – Évolution du taux d'inflation annuel glissant de 1991 à août 2007

année	1991	1992	1993	1994	1995	1996	1997	1998
taux	2,98	2,00	2,07	1,60	2,00	1,75	1,11	0,20
année	1999	2000	2001	2002	2003	2004	2005	2006
taux	1,30	1,58	1,36	2,30	2,16	2,11	1,53	1,53

TAB. 3.3 – Taux d'inflation annuel de 1991 à 2006

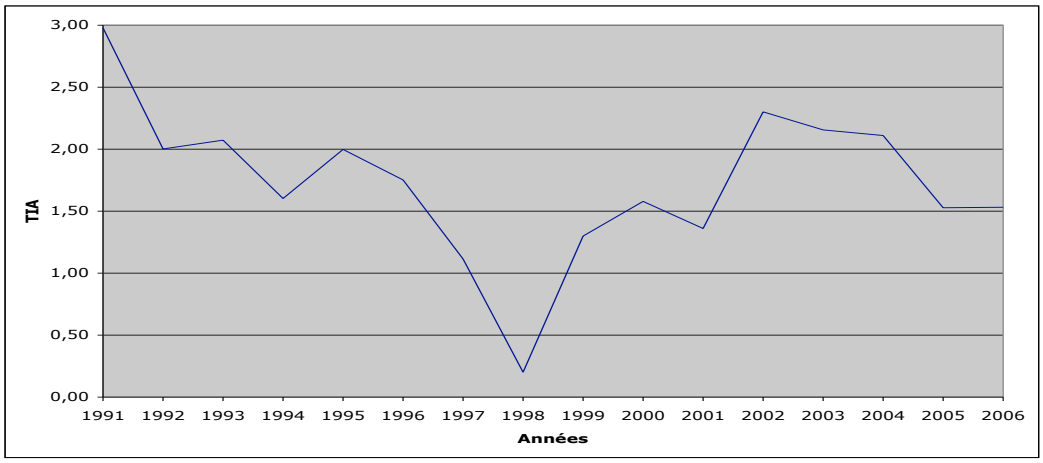

FIG. 3.3 – Évolution du taux d'inflation annuel de 1991 à 2006

3.3.3 Taux d'inflation mensuel

Le *taux d'inflation mensuel* suit l'évolution des prix chaque mois. C'est lui qui est communiqué au grand public, plutôt que l'indice des prix à la consommation dont il est issu, selon la relation,

$$\tau(t) = \frac{I(t) - I(t-1)}{I(t-1)} = \frac{I(t)}{I(t-1)} - 1,$$

faisant référence à :

$$I(t) = I(t-1)[1 + \tau(t)].$$

Par exemple, le taux de décembre 2005 est donné par :

$$\frac{113,0}{112,9} - 1 = 0,0008857\ldots \simeq 0,09\%.$$

Les valeurs du taux d'inflation mensuel, de février 1990 à août 2007, sont données dans le Tableau 3.4 et représentées sur la Figure 3.4. On notera qu'il peut être négatif (déflation).

année	J	F	M	A	M	J	J	A	S	O	N	D
2007	-0,34	0,18	0,43	0,49	0,25	0,12	-0,25	0,40				
2006	-0,05	0,37	0,29	0,41	0,44	-0,01	-0,17	0,34	-0,23	-0,22	0,11	0,23
2005	-0,54	0,54	0,63	0,18	0,09	0,18	-0,18	0,36	0,44	-0,09	-0,18	0,09
2004	0,00	0,46	0,37	0,27	0,36	0,00	-0,18	0,27	0,09	0,27	0,00	0,18
2003	0,19	0,65	0,46	-0,19	-0,09	0,19	-0,09	0,19	0,37	0,28	0,09	0,09
2002	0,48	0,10	0,48	0,38	0,09	0,00	0,00	0,19	0,19	0,19	0,00	0,19
2001	-0,39	0,29	0,39	0,48	0,68	0,00	-0,19	0,00	0,19	0,10	-0,29	0,10
2000	0,00	0,10	0,49	0,00	0,20	0,20	-0,20	0,20	0,59	-0,19	0,29	-0,10
1999	-0,30	0,30	0,40	0,20	0,00	0,00	-0,20	0,10	0,20	0,10	0,00	0,50
1998	-0,30	0,30	0,20	0,20	0,00	0,10	-0,30	0,00	0,00	0,00	-0,10	0,10
1997	0,20	0,20	0,10	0,00	0,10	0,00	-0,10	0,20	0,20	0,00	0,10	0,10
1996	0,21	0,31	0,72	0,10	0,20	-0,10	-0,20	-0,20	0,31	0,20	0,00	0,20
1995	0,21	0,31	0,31	0,10	0,10	0,10	-0,21	0,42	0,41	0,10	0,10	0,00
1994	0,11	0,32	0,21	0,21	0,21	0,00	0,00	0,00	0,32	0,21	0,00	0,00
1993	0,33	0,33	0,54	0,11	0,11	0,00	0,11	0,00	0,32	0,11	0,11	0,00
1992	0,22	0,44	0,33	0,33	0,22	-0,11	0,00	-0,11	0,33	0,22	0,11	0,00
1991	0,57	0,23	0,23	0,23	0,34	0,23	0,34	0,11	0,11	0,56	0,33	-0,33
1990		0,24	0,24	0,59	0,12	0,00	0,00	0,70	0,69	0,46	-0,11	-0,11

TAB. 3.4 – Taux d'inflation mensuel de février 1990 à août 2007

La représentation graphique du taux d'inflation mensuel donne plutôt l'impression d'une certaine stabilité sur toute la période, un peu comme celle de l'IPC. On voit ici l'intérêt que représente la considération d'une grandeur à différentes échelles.

3.3.4 Taux d'inflation moyen

À partir de l'IPC, il est possible de déterminer différents taux d'inflation. Par exemple, le *taux d'inflation trimestriel* du dernier trimestre 2005 est donné par :

$$\frac{113,0}{113,2} - 1 = -0,00176678\ldots \simeq -0,18\%.$$

Cependant, ce taux peut être calculé à partir des taux d'inflation mensuels de ce trimestre :

$$(1 - 0,0009)(1 - 0,0018)(1,0009) - 1 = -0,0018008\ldots \simeq -0,18\%.$$

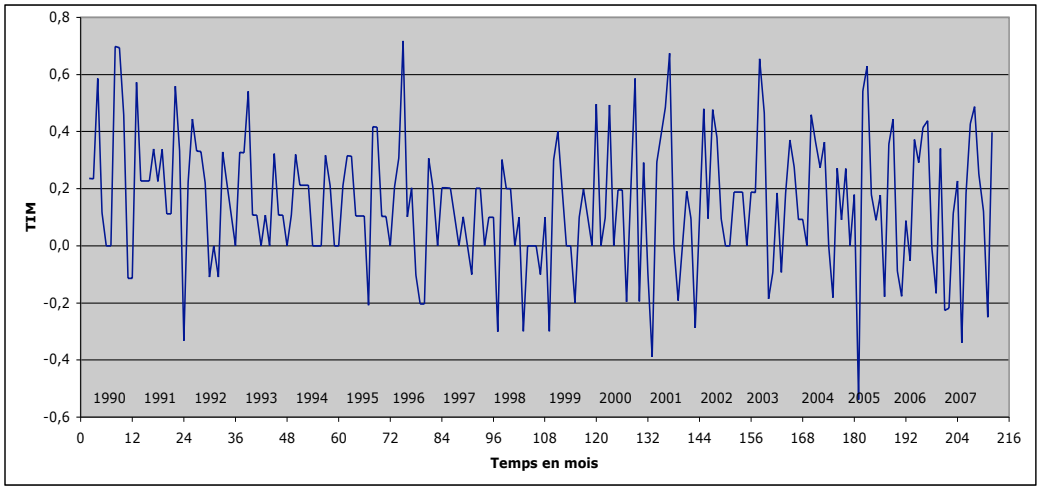

FIG. 3.4 – Évolution du taux d'inflation mensuel de février 1990 à août 2007

La différence avec la valeur précédente (valeurs non arrondies) est due au fait d'avoir arrondi les taux mensuels. Par contre le calcul,

$$-0,09\% - 0,18\% + 0,09\% = -0,18\%,$$

bien que donnant le bon résultat arrondi, n'est pas correct. Il correspond à un calcul à intérêts simples qui n'est pas conforme aux relations entre l'IPC et les taux mensuels. En notant $I(t)$ l'indice de décembre et τ_1, τ_2 et τ_3 les taux mensuels respectifs de décembre, novembre et octobre, on a :

$$I(t) = (1+\tau_1)I(t-1) = (1+\tau_1)(1+\tau_2)I(t-2) = (1+\tau_1)(1+\tau_2)(1+\tau_3)I(t-3).$$

De la même manière, on définit le *taux d'inflation mensuel moyen* du dernier trimestre 2005 par :

$$[(1-0,0009)(1-0,0018)(1,0009)]^{1/3} - 1 = -0,00060063\ldots \simeq -0,06\%.$$

Il ne s'agit donc pas de la moyenne arithmétique de τ_1, τ_2 et τ_3, mais de la valeur τ telle que $1+\tau$ soit la *moyenne géométrique* de $1+\tau_1, 1+\tau_2$ et $1+\tau_3$:

$$1 + \tau = [(1+\tau_1)(1+\tau_2)(1+\tau_3)]^{1/3}.$$

En récrivant cette égalité sous la forme,

$$(1+\tau)^3 = (1+\tau_1)(1+\tau_2)(1+\tau_3),$$

on remarque que le taux moyen τ de trois taux différents τ_1, τ_2 et τ_3 est le taux constant qui a le même effet que ces trois taux. C'est le principe même de la notion de moyenne, qui n'est donc pas nécessairement de nature arithmétique. Notons aussi que τ est le taux mensuel équivalent au taux trimestriel du trimestre considéré.

De la même manière, on peut considérer le taux d'inflation mensuel moyen de chaque année, ce qui correspond à des calculs de la forme,

$$1+\tau = [(1+\tau_1)(1+\tau_2)\dots(1+\tau_{12})]^{1/12},$$

à partir des taux mensuels $\tau_1, \tau_2, \dots, \tau_{12}$, de l'année, ou

$$1+\tau = (1+\tau_a)^{1/12},$$

à partir du taux annuel τ_a (taux équivalents), ou encore directement à partir des IPC,

$$1+\tau = \left[\frac{I(t)}{I(t-12)}\right]^{1/12},$$

pour les valeurs de t correspondant aux mois de décembre. Cette dernière approche est évidemment la plus précise, puisqu'elle n'utilise pas de résultats intermédiaires arrondis.

3.4 Actualisation

Actualiser une somme d'argent à l'instant t, c'est l'évaluer à cet instant, compte tenu d'un *taux d'actualisation* à définir selon l'usage que l'on fait de cette somme. Afin d'illustrer ceci, nous considérons une somme S ayant servi à l'achat de timbres poste en 1999. Nous cherchons à déterminer le taux d'actualisation (en 2006) dans les trois situations suivantes :

(i) S a été entièrement consacrée à l'achat de timbres ordinaires à 3 F ; timbre Marianne rouge actuel à 0,53 € (type A).

(ii) S a été entièrement consacrée à l'achat de timbres pour pli non urgent à 2,70 F ; timbre Marianne vert actuel à 0,48 € (type B).

(iii) Une proportion p de S a servi à l'achat de timbres A et le reste à l'achat de timbres B.

En utilisant le taux de conversion de l'Euro en Francs, égal à 6,55957, on obtient 0,46 € pour le timbre A et 0,41 € pour le timbre B en 1999. Le taux d'actualisation, pour les deux premières situations est donné par :

(i) $\frac{0,53}{0,46} - 1 = 0,152173913 \simeq 15,2\%$,

(ii) $\frac{0,48}{0,41} - 1 = 0,1707317073 \simeq 17,1\%$.

En utilisant directement le taux de conversion, sans calculer le prix des timbres de 1999 en Euros, on obtient :

(i) $\frac{0,53}{3} \times 6,55957 - 1 = 0,1588573667 \simeq 15,9\%$,

(ii) $\frac{0,48}{2,70} \times 6,55957 - 1 = 0,1661457778 \simeq 16,6\%$.

Cette solution est évidemment plus rigoureuse. Dans la troisième situation, on note n_A et n_B les quantités de timbres de chaque type,

$$n_A = \frac{pS}{3}, \quad n_B = \frac{(1-p)S}{2,70},$$

et le taux d'actualisation est donné par :

$$\begin{aligned} & \frac{n_A \times 0,53 + n_B \times 0,48}{S} \times 6,55957 - 1 \\ = \; & p\frac{0,53}{3} \times 6,55957 + (1-p)\frac{0,48}{2,70} \times 6,55957 - 1 \\ = \; & p\left[\frac{0,53}{3} \times 6,55957 - 1\right] + (1-p)\left[\frac{0,48}{2,70} \times 6,55957 - 1\right] \\ \simeq \; & p \times 15,9\% + (1-p) \times 16,6\%. \end{aligned}$$

Il s'agit d'une fonction affine de p, c'est-à-dire de la forme $ap + b$, qui varie de 15,9%, pour $p = 100\%$, à 16,6%, pour $p = 0\%$.

Ainsi, si une somme d'argent a servi à payer un ensemble de produits, on doit utiliser un taux d'actualisation qui tienne compte de la variation des prix de ces produits, mais aussi des quantités achetées. Un bon exemple de ce principe est fourni par le taux d'inflation défini précédemment. C'est lui que nous utiliserons en général, c'est-à-dire en l'absence d'information sur l'usage fait des sommes considérées.

3.5 Euros constants, euros courants

Voici un extrait du site http ://www.educnet.education.fr/insee/comext/combien/tendanceslonguessolde1.htm, permettant de comprendre la nécessité

d'introduire les notions d'*euros constants* et d'*euros courants*.

Une grandeur "à prix courants" ou "en euros courants" est calculée avec les prix de l'année considérée (on utilise également les expressions grandeur "nominale", grandeur "en valeur" ou grandeur "au prix de l'année courante"). L'influence des variartions des prix n'a pas été éliminée. Le montant de la grandeur dépend donc à la fois de la quantité de biens et services importés et exportés dans l'année par l'ensemble des agents et du niveau des prix constaté au cours de cette même année. Il est donc difficile de comparer ces valeurs d'une année sur l'autre étant donné que les prix varient chaque année. Les données "à prix constants" ou "en euros constants" au contraire n'enregistrent que "l'effet quantité", c'est-à-dire les variations liées à une augmentation ou une baisse de la quantité de biens et services importés et exportés par les agents et non celles liées à une fluctuation des prix. On utilise également les expressions grandeur "réelle", grandeur "en volume" ou grandeur "au prix d'une année de base".

L'opération qui consiste à passer une somme d'euros courants en euros constants, c'est-à-dire d'évacuer l'inflation, se dit *déflater* une somme. Les euros courants font référence à la date liée à la grandeur considérée, alors que les euros constants font référence à une année de base à choisir. En l'absence d'information complémentaire, le taux d'actualisation utilisé est celui de l'inflation.

Le Tableau 3.5 rappelle les taux d'inflation annuels de 1996 à 2003 et indique l'évolution d'une coupure de 100 Euros du 1^{er} janvier 1996 en euros courants des années 1997 à 2004 et en euros constants de 1996.

année	1996	1997	1998	1999	2000	2001	2002	2003	2004
taux	1,75	1,11	0,20	1,30	1,58	1,36	2,30	2,16	
courants	100	101,75	102,88	103,09	104,43	106,08	107,52	109,99	112,37
constants	100	98,28	97,20	97,01	95,76	94,27	93,01	90,92	88,99

TAB. 3.5 – Coupure de 100 Euros du 1^{er} janvier 1996 en euros courants des années 1997 à 2004 et en euros constants de 1996

En résumé Connaître les notions liées à l'inflation (indice des prix à la consommation, taux d'inflation, actualisation, euros constants et euros courants) et savoir effectuer les calculs qui s'y rattachent.

Chapitre 4

REMBOURSEMENT D'UN EMPRUNT

4.1 Introduction

Nous considérons ici les notions élémentaires liées au *remboursement d'un emprunt*. Après avoir fixé un choix de notations, nous précisons le principe de base d'un remboursement. Puis nous présentons deux modes de remboursement : par amortissements constants et par versements constants, ce dernier étant le plus usuel, en explicitant le tableau d'amortissement associé. Enfin, nous complétons cette étude avec quelques frais annexes attachés au remboursement d'emprunts : amortissement différé, assurance, frais de dossier et avec la notion de taux effectif global.

4.2 Notations et principe de base

Un capital C, emprunté à l'instant t_0, est remboursé à l'aide de n versements $v_1, v_2, \ldots, v_n$ effectués aux instants $t_1, t_2, \ldots, t_n$. Le temps est mesuré en mois et son origine a été fixée. La Figure 4.1 schématise ce remboursement.

FIG. 4.1 – Remboursement d'un emprunt

On suppose que la *période* qui s'écoule entre deux dates consécutives est constantes et égale à p mois : $t_k - t_{k-1} = p, k = 1, \ldots, n$. C'est généralement le cas avec, le plus souvent, des remboursements mensuels ($p = 1$), mais

aussi trimestriels ($p = 3$), semestriels ($p = 6$), voire annuels ($p = 12$). Pour simplifier, on suppose aussi que le *taux d'intérêt* par période, noté τ, est constant. C'est alors le taux équivalent au taux taux annuel τ_a fixé pour le prêt :

$$\tau = (1 + \tau_a)^{\frac{p}{12}} - 1.$$

Il existe des prêts à taux variable avec diverses options : garantie de taux plafond, passage à taux fixe, etc (faire "prêt à taux variable" dans Google).

Chaque *versement* v_k est la somme de deux termes,

$$v_k = i_k + m_k, \quad k = 1, 2, \ldots, n,$$

où :

- i_k représente la part de v_k affectée au règlement des *intérêts*,
- m_k, appelé *amortissement*, représente la part de v_k affectée au remboursement du capital.

Le principe de base, commun à tout type de prêt, est que le client est à jour d'intérêts lors de chaque *échéance*. En notant C_k, le *capital restant dû* juste avant le k^e versement, on a alors :

$$i_k = C_k \times \tau, \quad k = 1, 2, \ldots, n.$$

L'ensemble des informations concernant les modalités de remboursement d'un emprunt sont rassemblées dans un *tableau d'amortissement* selon le modèle présenté dans le Tableau 4.1.

Numéro	Échéance	Capital	Amortissement	Intérêts	Versement
1	05/03/06	C_1	m_1	i_1	v_1
2	05/04/06	C_2	m_2	i_2	v_2
k		C_k	m_k	i_k	v_k
n		C_n	m_n	i_n	v_n
Total			C	I	V

TAB. 4.1 – Éléments d'un tableau d'amortissement

Cependant, sans condition supplémentaire, il n'est pas possible de construire un tel tableau.

On peut, par exemple, choisir les amortissements $m_1, m_2, \ldots, m_n$ de façon quelconque, à condition de respecter la contrainte $\sum_{k=1}^{n} m_k = C$. En effet, on aura $C_1 = C$, puis $C_{k+1} = C_k - m_k, k = 1, \ldots, n-1$. Les intérêts sont alors donnés par le principe de base, $i_k = C_k \times \tau$, et les versements proviennent de $v_k = m_k + i_k$. On examine ci-après le cas particulier d'amortissements constants.

4.3 Amortissements constants

4.3.1 Principe

Les conséquences, sur le tableau d'amortissement, d'un remboursement à *amortissements constants* sont immédiates :

- $m_k = m = \frac{C}{n}, \quad k = 1, \ldots, n.$
- $C_k = C_{k-1} - m_{k-1} = C_{k-1} - \frac{C}{n} = C(1 - \frac{k-1}{n}), \quad k = 1, \ldots, n.$
- $i_k = C_k \times \tau = C\tau(1 - \frac{k-1}{n}), \quad k = 1, \ldots, n.$
- $v_k = i_k + m_k = C[(1 - \frac{k-1}{n})\tau + \frac{1}{n}], \quad k = 1, \ldots, n.$

On remarque que C_k, i_k et v_k forment trois suites arithmétiques décroissantes de raison $-\frac{C}{n}$, pour C_k, et $-\frac{C\tau}{n}$, pour i_k et v_k, et de premiers termes respectifs $C, C\tau$ et $C(\tau + \frac{1}{n})$. La décroissance est donc linéaire : les points $\{k, C_k\}, \{k, i_k\}$ et $\{k, v_k\}, k = 1, \ldots, n$, se situent sur des droites de pentes $-\frac{C}{n}$, pour C_k, et $-\frac{C\tau}{n}$, pour i_k et v_k.

Le *coût du crédit*, qui est la somme des intérêts, est donné par :

$$I = \sum_{k=1}^{n} i_k = C\tau \frac{n+1}{2}.$$

4.3.2 Illustration

On considère un prêt à la consommation de 4 500 €, remboursable en 3 ans par trimestrialités, au taux d'intérêt annuel de 6,30%, avec amortissements constants. La première échéance est fixée au 5 juin 2006.

L'amortissement constant est donné par :

$$m = \frac{C}{n} = \frac{4500}{12} = 375 \text{ €} ,$$

et le taux trimestriel équivalent vaut :

$$\tau = (1 + \tau_a)^{\frac{p}{12}} - 1 = (1,0630)^{3/12} - 1 = 1,539101508\%.$$

Le tableau d'amortissement est présenté dans le Tableau 4.2.

Numéro	Échéance	Capital	Amortissement	Intérêts	Versement
1	05/06/06	4 500,00	375,00	69,26	444,26
2	05/09/06	4 125,00	375,00	63,49	438,49
3	05/12/06	3 750,00	375,00	57,72	432,72
4	05/03/07	3 375,00	375,00	51,94	426,94
5	05/06/07	3 000,00	375,00	46,17	421,17
6	05/09/07	2 625,00	375,00	40,40	415,40
7	05/12/07	2 250,00	375,00	34,63	409,63
8	05/03/08	1 875,00	375,00	28,86	403,86
9	05/06/08	1 500,00	375,00	23,09	398,09
10	05/09/08	1 125,00	375,00	17,31	392,31
11	05/12/08	750,00	375,00	11,54	386,54
12	05/03/09	375,00	375,00	5,77	380,77
Total			4 500,00	(9) 450,18	(9) 4 950,18

TAB. 4.2 – Tableau d'amortissement d'un prêt à amortissements constants, $\tau_a = 6,30\%$

La valeur précise de la raison des suites arithmétiques i_k et v_k est :

$$-\frac{C\tau}{n} = -\frac{4500 \times 0,01539105508}{12} = -5,771645655 \quad (\simeq -5,77),$$

et le coût du crédit serait :

$$I = C\tau\frac{n+1}{2} = 4500 \times 0,01539105508 \times \frac{13}{2} = 450,1883611 \simeq 450,19.$$

Cependant, il est important de présenter au client un tableau cohérent au centime près. Pour cela, on est amené à remplacer 450,19, somme arrondie des intérêts non arrondis, par 450,18, qui est la somme des intérêts arrondis. En effet, il serait difficilement justifiable d'ajouter 1 centime aux intérêts, et par suite au versement, à une certaine échéance pour récupérer 450,19. On notera aussi que les suites i_k et v_k, arrondies au centime, ne sont pas rigoureusement arithmétiques, mais ceci ne devrait pas inquiéter le client !

4.4 Versements constants

4.4.1 Principe

En fait, le remboursement des prêts s'effectue habituellement par *versements constants*, selon le schéma de la Figure 4.2.

FIG. 4.2 – Remboursement d'un emprunt par versements constants

Pour ceci, il est nécessaire d'établir, dans un premier temps, le montant v de ce versement. En imaginant que le taux d'inflation annuel est constant et égal au taux d'intérêt annuel τ_a du prêt sur la durée du remboursement, on peut écrire que la somme des versements, actualisés à la date t_0, est égale au capital C emprunté :

$$C = v(1+\tau)^{-1} + v(1+\tau)^{-2} + \ldots + v(1+\tau)^{-n}.$$

Il s'agit de la somme des termes d'une suite géométrique, de premier terme $v(1+\tau)^{-1}$ et de raison $(1+\tau)^{-1}$:

$$C = v(1+\tau)^{-1}\frac{1-(1+\tau)^{-n}}{1-(1+\tau)^{-1}} = v\frac{1-(1+\tau)^{-n}}{\tau}.$$

D'où l'expression de v :

$$v = \frac{C\tau}{1-(1+\tau)^{-n}} = \frac{i_1}{1-(1+\tau)^{-n}} = C\,\frac{(1+\tau_a)^{\frac{p}{12}}-1}{1-(1+\tau_a)^{-\frac{np}{12}}}$$

Notons que $\frac{np}{12}$ représente la durée du remboursement, exprimée en années.

4.4.2 Loi des amortissements

Le bon sens précédent, ayant conduit à l'expression du versement v, est confirmé par la *loi des amortissements*. Celle-ci stipule que les amortissements $m_1, m_2, \ldots, m_n$, forment une suite géométrique de raison $(1+\tau)$. En effet, partant de l'hypothèse de versements constants,

$$v = m_k + i_k = m_{k+1} + i_{k+1},$$

on obtient :

$$m_{k+1} = m_k + i_k - i_{k+1} = m_k + C_k\tau - C_{k+1}\tau = m_k + m_k\tau = m_k(1+\tau).$$

Cette loi permet d'obtenir le premier amortissement m_1, en écrivant que la somme des amortissements est égale au capital emprunté :

$$C = m_1 + m_2 + \ldots + m_n = m_1 \frac{1-(1+\tau)^n}{1-(1+\tau)} = m_1 \frac{(1+\tau)^n - 1}{\tau},$$

soit,

$$m_1 = \frac{C\tau}{(1+\tau)^n - 1}.$$

On a alors :

$$v = i_1 + m_1 = C\tau + \frac{C\tau}{(1+\tau)^n - 1} = \frac{C\tau(1+\tau)^n}{(1+\tau)^n - 1} = \frac{C\tau}{1-(1+\tau)^{-n}}.$$

Notons que le *coût du crédit* se déduit simplement du versement v :

$$I = i_1 + i_2 + \ldots + i_n = nv - C = C\left[\frac{n\tau}{1-(1+\tau)^{-n}} - 1\right].$$

On peut également établir les expressions des termes m_k, C_k et i_k :

- $m_k = m_1(1+\tau)^{k-1} = \frac{C\tau(1+\tau)^{k-1}}{(1+\tau)^n - 1}, \quad k = 1, \ldots, n.$
- $C_k = C - m_1 - \ldots - m_{k-1} = C\,\frac{(1+\tau)^n - (1+\tau)^{k-1}}{(1+\tau)^n - 1}, \quad k = 1, \ldots, n.$
- $i_k = C_k\tau = C\tau\,\frac{(1+\tau)^n - (1+\tau)^{k-1}}{(1+\tau)^n - 1}, \quad k = 1, \ldots, n.$

Seule la suite $m_k, k = 1, \ldots, n$, est géométrique et les points $\{k, m_k\}, k = 1, \ldots, n$, se situent sur le graphe de la fonction exponentielle d'équation :

$$m_k = m_1(1+\tau)^{k-1} = \frac{m_1}{1+\tau}(1+\tau)^k = \frac{m_1}{1+\tau}e^{k\log(1+\tau)}.$$

La croissance de m_k est donc exponentielle. Par ailleurs, on dit que la décroissance de i_k est exponentielle car $i_k = v - m_k$. C'est également le cas de C_k que l'on peut écrire sous la forme :

$$C_k = C\,\frac{(1+\tau)^n}{(1+\tau)^n - 1} - C\,\frac{(1+\tau)^{-1}}{(1+\tau)^n - 1}e^{k\log(1+\tau)}.$$

4.4.3 Prêt immobilier

Pour des prêts importants, comme un *prêt immobilier*, la loi d'amortissement, due aux remboursements constants, a pour conséquence que les premières échéances sont essentiellement consacrées au paiement des intérêts.

Par suite, le capital restant dû diminue lentement. On a ainsi l'impression de payer plus d'intérêts que nécessaire, ce qui n'est évidemment pas le cas.

Notons $M_1, M_2, \ldots, M_K$, les amortissements annuels d'un prêt immobilier, remboursé par mensualités constantes sur K années. Le premier amortissement M_1 est donné par :

$$M_1 = m_1 + m_2 + \ldots + m_{12} = m_1 \frac{(1+\tau)^{12} - 1}{\tau} = m_1 \frac{\tau_a}{\tau} = \frac{C\tau_a}{(1+\tau_a)^K - 1}.$$

Il est facile d'établir que les amortissements annuels $M_1, M_2, \ldots, M_K$, forment une suite géométrique de raison $(1+\tau_a)$:

$$M_{k+1} = m_{12k+1} + \ldots + m_{12k+12} = (1+\tau_a)^k [m_1 + \ldots + m_{12}] = M_1 (1+\tau_a)^k.$$

Notons que le tableau d'amortissement, à l'échelle de l'année, n'est pas celui d'un remboursement annuel à annuités constantes, car les versements annuels, ici égaux à $12v$, sont inférieurs à ce que serait l'annuité :

$$12v = \frac{12C\tau}{1 - (1+\tau_a)^{-K}} < \frac{C\tau_a}{1 - (1+\tau_a)^{-K}},$$

puisque $12\tau < \tau_a$.

4.4.4 Illustrations

Prêt à la consommation. On reprend le prêt considéré précédemment, en remplaçant la contrainte d'amortissements constants par celle de versements constants. On calcule tout d'abord le montant du versement (après avoir calculé τ) :

$$v = \frac{C\tau}{1 - (1+\tau_a)^{-\frac{np}{12}}} = \frac{4500 \times 0,01539105508}{1 - (1,0630)^{-3}} = 413,5655163\ldots \simeq 413,57\ \text{€}\ .$$

Le tableau d'amortissement est présenté dans le Tableau 4.3. On peut construire ce tableau, ligne par ligne, à l'aide des relations $i_k = C_k\tau, m_k = v - i_k$ et $C_{k+1} = C_k - m_k$, ou utiliser la loi des amortissements pour la colonne m_k, puis $i_k = v - m_k$ et $C_{k+1} = C_k - m_k$. Quelle que soit la méthode utilisée, on calcule les valeurs sans arrondir, y-compris v, et on affiche les résultats au centime le plus proche. Ensuite, on modifie les résultats arrondis de façon à rendre les colonnes cohérentes au centime près. Pour cela, on ne change ni le versement v, ni les capitaux C_k. Dans un premier temps, on modifie m_k de sorte à satisfaire $C_{k+1} = C_k - m_k$. Puis on modifie i_k pour satisfaire

Numéro	Échéance	Capital	Amortissement	Intérêts	Versement
1	05/06/06	4 500,00	344,31	69,26	413,57
2	05/09/06	4 155,69	(1) 349,60	(6) 63,97	413,57
3	05/12/06	3 806,09	354,99	58,58	413,57
4	05/03/07	3 451,10	360,45	53,12	413,57
5	05/06/07	3 090,65	(6,00) 365,99	(7) 47,58	413,57
6	05/09/07	2 724,66	371,63	41,94	413,57
7	05/12/07	2 353,03	377,35	36,22	413,57
8	05/03/08	1 975,68	383,16	30,41	413,57
9	05/06/08	1 592,52	389,06	24,51	413,57
10	05/09/08	1 203,46	395,04	(2) 18,53	413,57
11	05/12/08	808,42	401,12	(4) 12,45	413,57
12	05/03/09	407,30	407,30	6,27	413,57
Total			4 500,00	(79) 462,84	(79) 4 962,84

TAB. 4.3 – Tableau d'amortissement d'un prêt à versements constants, $\tau_a = 6,30\%$

$v - i_k = m_k$. Enfin, on modifie le total avec $V = nv$ et $I = V - C$. Les résultats initiaux (avant cohérence) sont indiqués entre parenthèses (derniers chiffres modifiés) dans le Tableau 4.3. On constate que le coût du crédit est supérieur de 5 centimes à la valeur exacte ($12 \times 413,5655163\ldots - 4500 \simeq 462,79$).

Prêt immobilier. On considère un prêt immobilier de 200 000 € , au taux annuel de 3,70 %, remboursé en 25 ans par mensualités constantes. La mensualité est donnée par :

$$v = \frac{200000[1,0370^{\frac{1}{12}} - 1]}{1 - 1,0370^{-25}} = 1016,185861\ldots \simeq 1\ 016,19\ \text{€}\ ,$$

d'où le montant de l'annuité :

$$12 \times 1016,19 = 12\ 194,28\ \text{€}\ .$$

Le tableau d'amortissement annuel est donné dans le Tableau 4.4. Il a été construit en utilisant la loi d'amortissement, $M_{k+1} = M_k(1 + \tau_a)$, à partir de M_1 non arrondi, puis $C_{k+1} = C_k - M_k$ et $I_k = 12\ 194,28 - M_k$. On notera que les modifications de M_k puis I_k, pour assurer la cohérence, sont très importantes. Le coût du crédit est de 104 857 € , au lieu de 104 855,76 ($25 \times 12 \times v - C$), soit 1,24 € de plus. On observe bien, sur cet exemple, la prépondérance des intérêts sur les amortissements au cours des premières années. Il faut près de 16 ans pour amortir la moitié du capital.

Numéro	Capital	Amortissement	Intérêts	Annuité
1	200 000,00	4 999,67	(56) 7 194,61	12 194,28
2	195 000,33	(6) 5 184,65	(57) 7 009,63	12 194,28
3	189 815,68	5 376,49	(4) 6 817,79	12 194,28
4	184 439,19	5 575,42	(1) 6 618,86	12 194,28
5	178 863,77	5 781,71	(2) 6 412,57	12 194,28
6	173 082,06	5 995,63	(0) 6 198,65	12 194,28
7	167 086,43	6 217,47	(76) 5 976,81	12 194,28
8	160 868,96	(2) 6 447,51	(1) 5 746,77	12 194,28
9	154 421,45	(7) 6 686,08	(16) 5 508,20	12 194,28
10	147 735,37	6 933,46	(77) 5 260,82	12 194,28
11	140 801,91	(90,00) 7 189,99	(3) 5 004,29	12 194,28
12	133 611,92	7 456,03	(0) 4 738,25	12 194,28
13	126 155,89	7 731,90	(3) 4 462,38	12 194,28
14	118 423,99	8 017,98	(25)4 176,30	12 194,28
15	110 406,01	(5) 8 314,64	(59) 3 879,64	12 194,28
16	102 091,37	8 622,29	(4) 3 571,99	12 194,28
17	93 469,08	8 941,31	(2) 3 252,97	12 194,28
18	84 527,77	9 272,14	(09) 2 922,14	12 194,28
19	75 255,63	9 615,21	(2) 2 579,07	12 194,28
20	65 640,42	(7) 9 970,98	(26) 2 223,30	12 194,28
21	55 669,44	(90) 10 339,89	(3) 1 854,39	12 194,28
22	45 329,55	(7) 10 722,48	(76) 1 471,80	12 194,28
23	34 607,07	(1) 11 119,20	(2) 1 075,08	12 194,28
24	23 487,87	11 530,62	(1) 663,66	12 194,28
25	11 957,25	11 957,25	(6,98) 237,03	12 194,28
Total		200 000,00	(5,76) 104 857,00	304 857,00

TAB. 4.4 – Tableau d'amortissement annuel d'un prêt immobilier, $\tau_a = 3,70\%$

4.5 Frais annexes et taux effectif global

D'autres frais viennent parfois s'ajouter au remboursement d'un prêt : amortissement différé, frais de dossier et assurance. Ces derniers sont pris en compte dans le taux effectif global.

4.5.1 Amortissement différé

L'amortissement proprement dit du prêt débute une période (p mois) avant la première échéance, c'est-à-dire à la date t_0. En général, entre la date de remise des fonds, notée t_f, et le point de départ de l'amortissement, s'écoule un certain temps, appelé période d'*amortissement différé* (différé d'amortissement, différé de paiement) pour lequel on paie des intérêts calculés

au prorata du nombre de jours de cette période (*cf.* Figure 4.3).

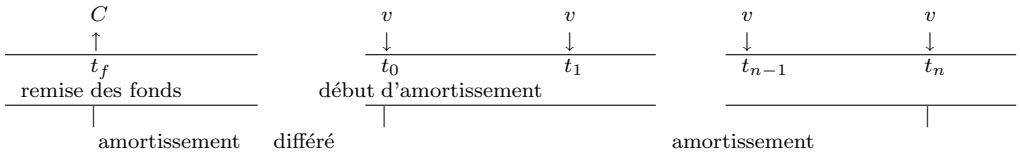

FIG. 4.3 – Amortissement et amortissement différé

Dans notre premier exemple d'illustration, la date t_0 est le 5 mars 2006. En supposant que les fonds ont été versés le 14 février 2006 (date t_f), les intérêts dus au titre de la période d'amortissement différé sont :

$$I_d = 4500 \left[(1,0630)^{\frac{21}{360}} - 1 \right] = 16,066075 \simeq 16,07 \text{ €} .$$

Ces intérêts devraient être réglés le 5 mars. En fait ils le seront en même temps que la première échéance, c'est-à-dire le 5 juin, mais augmentés des intérêts correspondants :

$$I_d^* = I_d(1+\tau) = 4500 \left[(1,0630)^{\frac{21}{360}} - 1 \right] (1,0630)^{\frac{3}{12}} = 16,2087\ldots \simeq 16,21 \text{ €} .$$

Le taux d'intérêt, pour la période d'amortissement différé, est parfois différent du taux d'emprunt. Cette période peut être également volontairement augmentée, par exemple 2 ans pour certains prêts immobiliers, afin de faciliter l'accès à la propriété. Dans ce cas, l'emprunteur règle, à chaque échéance, les intérêts constants $C\tau$, puisque le capital restant dû ne varie pas. Pour plus d'information sur le sujet, voir par exemple le site : http ://www.cbanque.com/credit/differe.php

4.5.2 Frais de dossier et assurance

Les *frais de dossier* s'élèvent en général à 1% du montant emprunté avec un minimun et un maximum suivant les établissements. Certains prêts sont proposés sans frais de dossier. Cependant, une différence a priori faible sur le taux d'intérêt peut s'avérer plus coûteuse. Considérons un prêt de 50 000 € sur 20 ans, remboursé par mensualités constantes, avec un taux annuel de 3,05% et 500 € de frais de dossier, ou au taux annuel de 3,25%, sans frais de dossier. Le coût du crédit pour ces deux situations est le suivant :

- *Avec frais de dossier*

$$v = \frac{50000 \times (1,0305^{1/12} - 1)}{1 - 1,0305^{-20}} = 277,503\ldots \simeq 277,50 \text{ €}$$

$$\begin{aligned} I &= 20 \times 12 \times 277,50 - 50000 = 16\,666 \text{ €} \\ \text{Coût} &= 16\,666 + 500 = 17\,166 \text{ €} \end{aligned}$$

- *Sans frais de dossier*

$$\begin{aligned} v &= \frac{50000 \times (1,0325^{1/12} - 1)}{1 - 1,0325^{-20}} = 282,396\ldots \simeq 282,40 \text{ €} \\ I &= 20 \times 12 \times 282,40 - 50000 = 17\,776 \text{ €} \\ \text{Coût} &= 17\,776 \text{ €} \end{aligned}$$

Il n'y a pas d'obligation légale pour le consommateur de souscrire une *assurance* emprunteur. Mais la majorité des organismes prêteurs n'accordent un prêt au demandeur que s'il peut présenter des garanties en cas de défaillance. L'emprunteur peut soit s'assurer directement auprès de sa banque, soit choisir son propre assureur. Souscrire auprès de l'assureur de son choix peut se révéler très intéressant (surtout pour les jeunes de 25 ans à 45 ans). En effet, le banquier propose un tarif moyen unique, couvrant toutes les tranches d'âge, qui désavantage beaucoup les personnes jeunes et celles exclues pour cause de problèmes médicaux. En souscrivant une assurance individuelle chez un assureur, l'emprunteur peut réaliser des économies (qui peuvent atteindre 50%). Aucune loi n'oblige l'emprunteur à souscrire l'assurance de prêt proposée par l'organisme financier auprès duquel il a contracté le crédit. Il n'est donc pas tenu d'accepter les conditions d'assurance proposées par son banquier qui peut seulement exiger de lui qu'il soit garanti en cas de décès. La vente liée est, en effet, totalement interdite depuis 1986.

Les frais d'assurance seront prélevés à chaque échéance. Le montant est un taux fixe appliqué soit au capital emprunté C, soit au capital restant dû C_k. Prenons l'exemple du prêt à la consommation de 4 500 € . En appliquant un taux de 0,05% au capital emprunté, l'échéance est augmentée de $4500 \times 0{,}05\% = 2{,}25$ € . On peut vérifier que ceci équivaut à pratiquer un taux de 6,67%, au lieu de 6,30 %, comme taux d'emprunt (utiliser l'outil "Valeur cible" d'Excel).

4.5.3 Taux effectif global

Le *taux effectif global* (TEG) est calculé à partir des caractéristiques d'un prêt et incorpore tous les éléments de coût du prêt : *taux nominal d'intérêt* (τ_a), frais de dossier, frais d'hypothèque, timbres fiscaux, etc. et coût de l'assurance. Le coût de l'assurance n'est pas compris dans le TEG, si vous souscrivez une assurance auprès d'un organisme tiers, en revanche si c'est la

banque qui vous fournit un contrat d'assurance collective, alors le TEG est exprimé coût d'assurance inclus.

Ce taux est obligatoirement indiqué en France, dans toutes les offres et tous les actes de prêt, et ce depuis de nombreuses années.

Le TEG a une autre fonction ; il permet de vérifier qu'en comptant tous les éléments qui s'ajoutent dans le coût d'un crédit, le taux ne dépasse pas le taux de l'usure, taux maximum défini par la loi pour chacun des types de crédits et publié tous les trimestres par la Banque de France. Pour plus d'informations, voir par exemple le site :
http ://www.guideducredit.com/HTMcorps/Fichiersfinancement/teg.htm

Reprenons l'exemple du prêt à la consommation de 4 500 € . Le taux nominal est $\tau_a = 6,30\%$, avec une échéance à 413,57 € , sans assurance. Avec l'assurance de 2,25 € , l'échéance devient 415,82 € , ce qui porte le TEG à 6,67%. On ajoute 1% de frais de dossier, soit 45 € . Pour déterminer le TEG, on imagine que les frais de dossier sont retirés du capital emprunté C et on cherche le taux d'intérêt annuel conduisant à la même échéance de 415,82 € , pour un capital de 4 465 € . Le TEG est alors de 7,20% (utiliser l'outil "Valeur cible" d'Excel).

Toutes les considérations faites dans ce chapitre sont effectuées hors inflation. Il est clair que l'inflation, sur la période de remboursement, a une incidence sur le coût réel d'un crédit. Celle-ci n'étant pas connue, il est difficile de la prendre en compte. Elle l'est cependant, sous forme de prévision, dans les taux proposés par les organismes bancaires.

En résumé Connaître les notions liées aux remboursements d'emprunts, essentiellement à versements constants, et savoir effectuer les calculs qui s'y rattachent (montant des échéance, coût du crédit, tableau d'amortissement, TEG).

DEUXIÈME PARTIE

RÉGRESSION LINÉAIRE

Chapitre 5

RÉGRESSION LINÉAIRE SIMPLE : APPROCHE DESCRIPTIVE

5.1 Introduction

La *régression linéaire simple* étudie la dépendance, sous forme linéaire, entre deux grandeurs. L'exemple classique du poids d'un individu en fonction de sa taille est illustré, sur la Figure 5.1, par un échantillon de 32 étudiants en distinguant les deux sexes. L'objectif est de prévoir le poids d'un individu dont on connaît la taille.

Ce chapitre est consacré à l'approche descriptive du problème, les aspects décisionnels étant reportés au chapitre suivant. Le premier paragraphe présente la droite de régression, à travers une illustration. Le coefficient de corrélation linéaire empirique, ainsi que le coefficient de détermination, qui permet de mesurer la dépendance linéaire entre les deux grandeurs, font l'objet du deuxième paragraphe. Le paragraphe suivant concerne l'analyse descriptive des résidus. Le chapitre se termine par une mise en garde sur l'utilisation de la méthode.

5.2 Droite de régression

5.2.1 Illustration

On étudie la vitesse coronarienne Y en fonction du poids X chez les individus. Une grande vitesse coronarienne est un indice de bon fonctionnement

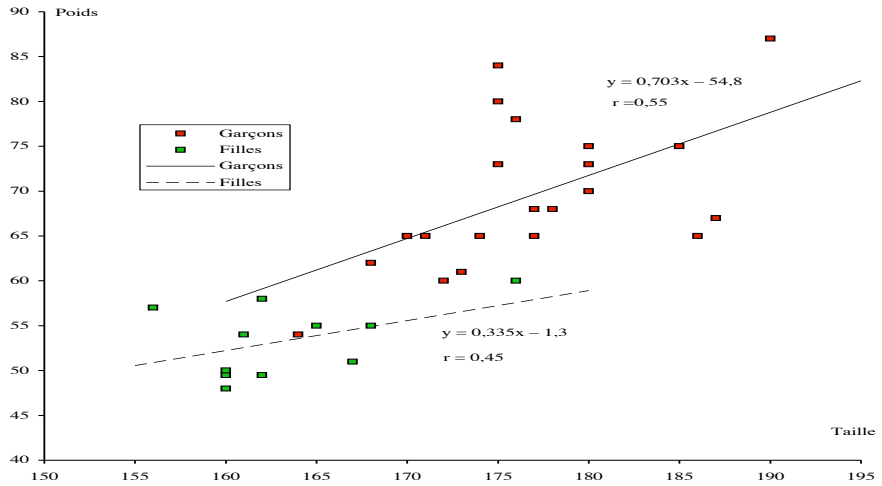

FIG. 5.1 – Régression du poids par rapport à la taille d'un ensemble d'étudiants

cardiaque. On a mesuré chez $n = 18$ patients, le poids x_i en kg et la vitesse coronarienne $y_i, i = 1, \ldots, n$. Les données indiquées dans le Tableau 5.1 sont représentées sur la Figure 5.2. Ce graphique fait apparaître une dépendance linéaire entre les deux variables. Les deux variables sont appréhendées de façon différente. La variable Y a un caractère incertain alors que X est supposée parfaitement connue. On dit que X est la *variable explicative* (variable indépendante, variable exogène, régresseur) et que Y est la *variable expliquée* (variable dépendante, variable endogène, régressante). L'objectif est de prévoir une plage de valeurs raisonnable (intervalle de confiance) pour la vitesse coronarienne d'un individu dont on connaît le poids.

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
x_i	45	48	50	50	52	53	56	58	63	66	66	69	72	74	79	79	84	89
y_i	75	77	78	77	77	72	72	72	70	71	69	69	68	66	64	66	62	61

TAB. 5.1 – Vitesse coronarienne y_i et poids x_i de 18 patients

Résumé numérique, caractéristiques empiriques et premiers résultats

À l'aide du résumé numérique,

$$n = 18; \Sigma x = 1153; \Sigma x^2 = 76903; \quad \Sigma y = 1266; \Sigma y^2 = 89508; \quad \Sigma xy = 79947,$$

on calcule les caractéristiques empiriques,

$$\overline{x} = 64,1; var(x) = 169,27; \quad \overline{y} = 70,3; var(y) = 25,89; \quad cov(x,y) = -63,74,$$

permettant d'obtenir les résultats considérés dans ce chapitre :

Droite de régression	$y = -0,377x + 94,5$
Coefficient de corrélation linéaire	$r = -0,96$
Coefficient de détermination	$r^2 = 93\%$

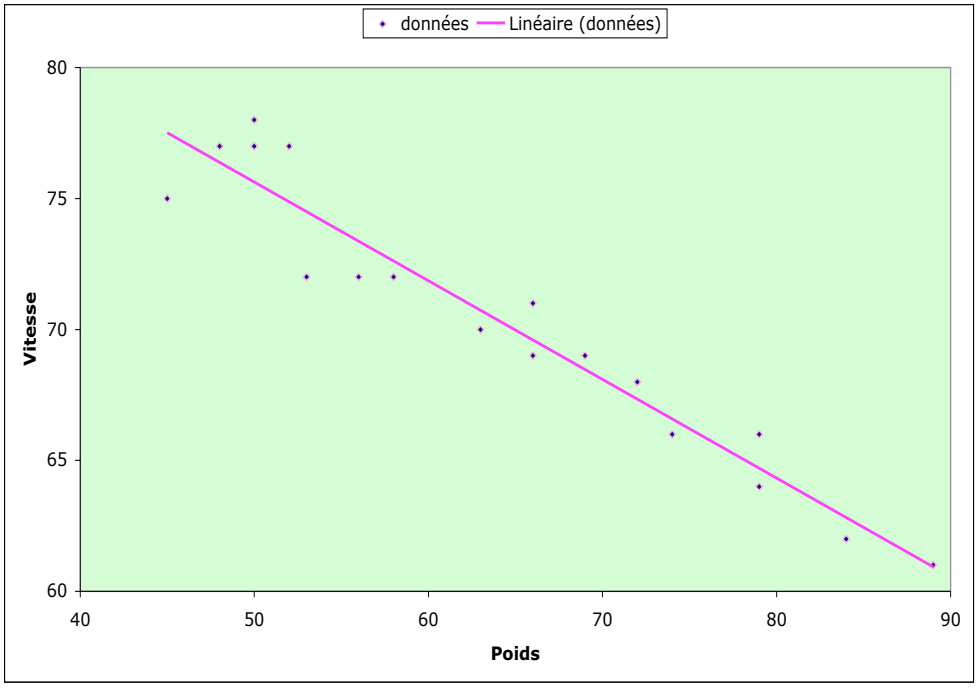

FIG. 5.2 – Vitesse coronarienne en fonction du poids dans l'espace des variables

5.2.2 Les hypothèses du modèle

On suppose que les mesures $(x_i, y_i), i = 1, \ldots, n$ sont telles que, pour chaque individu i, la valeur y_i est approximativement égale à $ax_i + b$, où a et b sont fixés, mais inconnus :

$$y_i = ax_i + b + \varepsilon_i, \quad i = 1, \ldots, n.$$

Ainsi, $\varepsilon_i, i = 1, \ldots, n$ sont des *erreurs* traduisant cette approximation. Elles sont de moyenne nulle, du même ordre de grandeur et indépendantes les unes des autres, pour tout ensemble d'individus choisis au hasard. Ce point sera précisé dans l'étude descriptive des résidus et formalisé au chapitre suivant. Les valeurs $x_1, x_2, \ldots, x_n$ constituent le *plan d'expérience*. Elles peuvent provenir d'observations effectuées au hasard, mais elles sont parfois fixées par l'expérimentateur. Enfin a et b sont les *paramètres* du modèle. Ce sont des quantités inconnues que l'on cherche à estimer.
L'*espace des variables* (ici $\mathbb{R}^2$) permet de visualiser les observations $(x_i, y_i), i = 1, \ldots, n$ dans un système d'axes orthogonaux sur lesquels sont mesurées les variables X et Y (*cf.* Figure 5.2). L'*espace des observations* (ici $\mathbb{R}^n$) permet de visualiser les variables sous forme vectorielle (*cf.* Figure 5.3). On peut en effet écrire :

$$y = ax + b\mathbb{1} + \varepsilon,$$

en considérant les vecteurs :

$$y = (y_1, \ldots, y_n)^T, \quad x = (x_1, \ldots, x_n)^T, \quad \mathbb{1} = (1, \ldots, 1)^T, \quad \varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T.$$

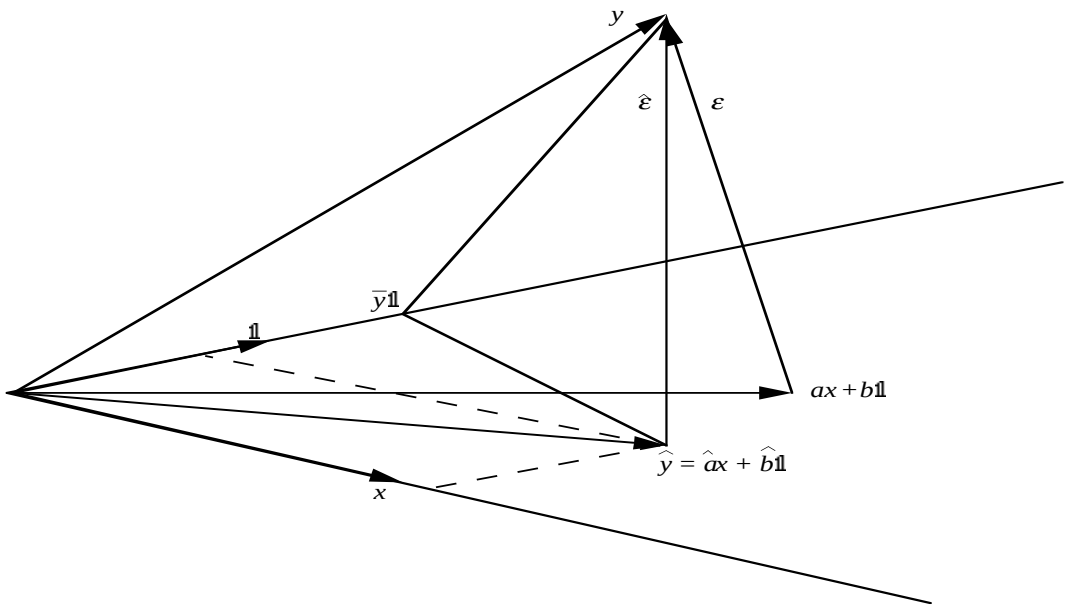

FIG. 5.3 – Espace des observations

Le sous-espace de $\mathbb{R}^n$ de dimension 2 engendré par x et $\mathbb{1}$ est appelé *espace des moyennes*. C'est l'espace dans lequel se situe le vecteur $ax + b\mathbb{1}$.

5.2.3 Estimateurs des moindres carrés

Si le paramètre (a, b) était connu, les erreurs ε_i seraient observables et données par $\varepsilon_i = y_i - ax_i - b, i = 1, \ldots, n$. L'estimation de ce paramètre par la *méthode des moindres carrés* consiste à retenir la valeur de (a, b) pour laquelle la moyenne des carrés de ces erreurs est minimum. Ainsi la *droite de régression* linéaire de y par rapport à x est la *droite des moindres carrés*, définie par la minimisation du critère :

$$D_{y/x}(a, b) = \frac{1}{n}\sum_{i=1}^{n}[y_i - ax_i - b]^2 = \frac{1}{n}||y - ax - b\mathbb{1}||^2.$$

$D_{y/x}(a, b)$ représente la moyenne des carrés des écarts verticaux entre les points (x_i, y_i) du nuage et ceux $(x_i, ax_i + b)$ de mêmes abscisses x_i situés sur la droite d'équation $y = ax + b$. On a :

$$D_{y/x}(a, b) = \overline{x^2}a^2 + b^2 + 2\overline{x}ab - 2\overline{xy}a - 2\overline{y}b + \overline{y^2},$$

où :

$$\overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2, \quad \overline{y^2} = \frac{1}{n}\sum_{i=1}^{n} y_i^2, \quad \overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i.$$

$D_{y/x}(a, b)$ est un polynôme du second degré en les variables a et b dont le graphe est un paraboloïde de révolution orienté vers le haut. Le minimum est réalisé par la solution du système d'équations obtenu en annulant les deux dérivées partielles :

$$\begin{aligned} \frac{\partial}{\partial a}D_{y/x}(a, b) &= 2\overline{x^2}a + 2\overline{x}b - 2\overline{xy} = 0, \\ \frac{\partial}{\partial b}D_{y/x}(a, b) &= 2b + 2\overline{x}a - 2\overline{y} = 0. \end{aligned}$$

La deuxième équation montre que la droite cherchée passe par le point moyen $(\overline{x}, \overline{y})$ puisqu'elle équivaut à $a\overline{x} + b = \overline{y}$. Le report de ce résultat dans la première équation conduit à :

$$var(x)a - cov(x, y) = 0,$$

où $cov(x, y)$ désigne la *covariance empirique* entre les variables X et Y :

$$cov(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \overline{xy} - \overline{x}\ \overline{y}.$$

Ainsi la droite des moindres carrés est définie par :

$$y = \hat{a}x + \hat{b}, \quad \hat{a} = \frac{cov(x,y)}{var(x)}, \quad \hat{b} = \overline{y} - \hat{a}\overline{x}.$$

On peut aussi obtenir ce résultat en introduisant le point moyen $(\overline{x}, \overline{y})$ dans l'expression du critère :

$$D_{y/x}(a,b) = \frac{1}{n}\sum_{i=1}^{n}[(y_i - \overline{y}) - a(x_i - \overline{x}) + (\overline{y} - a\overline{x} - b)]^2.$$

On effectue ensuite le développement :

$$D_{y/x}(a,b) = var(y) + a^2 var(x) - 2acov(x,y) + (\overline{y} - a\overline{x} - b)^2.$$

Pour a fixé, la valeur de b qui minimise le critère doit annuler le dernier terme, d'où la relation, $b = \overline{y} - a\overline{x}$, qui exprime que la droite passe par le point moyen. Puis $D_{y/x}(a, \overline{y} - a\overline{x})$ est un polynôme du second degré en a dont l'annulation de la dérivée, $2var(x)a - 2cov(x,y) = 0$, donne la condition sur a. Cette seconde approche établit qu'il s'agit bien d'un minimum.

On dit que les coefficients,

$$\hat{a} = \frac{cov(x,y)}{var(x)}, \quad \hat{b} = \overline{y} - \hat{a}\overline{x},$$

définissant cette droite sont les *estimateurs des moindres carrés*. Dans l'espace des variables, $y = \hat{a}x + \hat{b}$ est l'équation de la droite qui minimise la moyenne des carrés des écarts verticaux. Dans l'espace des observations, $\hat{y} = \hat{a}x + \hat{b}\mathbb{1}$ est la projection orthogonale de l'observation y sur l'espace des moyennes. Les composantes $\hat{y}_i = \hat{a}x_i + \hat{b}, i = 1, \ldots, n$ du vecteur $\hat{y}$ sont les *valeurs ajustées*. Le critère des moindres carrés est donc justifié par les hypothèses faites sur les erreurs $\varepsilon_i, i = 1, \ldots, n$. Les propriétés statistiques de l'estimateur $(\hat{a}, \hat{b})$ seront étudiées dans le chapitre suivant.

5.3 Coefficient de corrélation linéaire empirique

La variance empirique de Y se décompose sous la forme :

$$var(y) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{a}x_i - \hat{b})^2 + \frac{1}{n}\sum_{i=1}^{n}(\hat{a}x_i + \hat{b} - \overline{y})^2.$$

Ce résultat peut être vérifié directement. Il correspond, dans l'espace des observations, à la relation :

$$y - \overline{y}\mathbb{1} = (y - \hat{y}) \oplus (\hat{y} - \overline{y}\mathbb{1}) \quad \Longrightarrow \quad ||y - \overline{y}\mathbb{1}||^2 = ||y - \hat{y}||^2 + ||\hat{y} - \overline{y}\mathbb{1}||^2,$$

utilisant l'orthogonalité des vecteurs $(y - \hat{y})$ et $(\hat{y} - \overline{y}\mathbb{1})$ (théorème de pythagore). Le premier terme de cette décomposition, d'ailleurs égal à la valeur minimum $D_{y/x}(\hat{a}, \hat{b})$ du critère, mesure la dispersion des points autour de la droite alors que le second mesure la dispersion des points de mêmes abscisses situés sur la droite. En introduisant le *coefficient de corrélation linéaire empirique*,

$$r = \frac{cov(x, y)}{\sqrt{var(x)}\sqrt{var(y)}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2}},$$

cette décomposition s'écrit

$$var(y) = (1 - r^2)var(y) + r^2 var(y).$$

Ainsi r^2, appelé *coefficient de détermination*, représente la part de variance (empirique) de Y expliquée par la régréssion linéaire de y sur x. On a évidemment $0 \leq r^2 \leq 1$ et les variables sont dites *non linéairement corrélées* lorsque $r^2 = 0$. À l'opposé, l'égalité $r^2 = 1$ équivaut à ce que les points soient alignés (les variables sont linéairement liées, $y_i = \hat{a}x_i + \hat{b}, i = 1, \ldots, n$). On remarque que r est symétrique par rapport aux deux variables, mais les deux droites de régression (y par rapport à x et x par rapport à y) sont distinctes, sauf si $r^2 = 1$. Notons enfin que r représente la pente de la droite exprimée en fonction des variables centrées et réduites :

$$y = \hat{a}x + \hat{b} \quad \Leftrightarrow \quad \frac{y - \overline{y}}{\sqrt{var(y)}} = r\frac{x - \overline{x}}{\sqrt{var(x)}}.$$

Pour l'illustration concernant la vitesse coronarienne, on obtient $r = -0,96$ et $r^2 = 93\%$. Il s'agit d'une très forte corrélation. Cela est cohérent avec la proximité des points à la droite. Pour fixer les ordres de grandeur, on considère que la corrélation est faible, moyenne ou forte lorsque le module de r est inférieur à 0,5, compris entre 0,5 et 0,7 ou supérieur à 0,7. La part r^2 de variance expliquée est alors inférieure à 25%, comprise entre 25% et 50% ou plus grande que 50%. Lorsque r^2 dépasse 80% ($|r| > 0,9$), on peut parler de très forte corrélation. Cependant r^2 peut être très voisin de 1 sans pour autant que le modèle linéaire soit justifié. Nous illustrerons ce point en fin de chapitre. Dans l'exemple introductif, la corrélation entre le poids et la

taille est faible puisque r est de l'ordre de 0,5. La faible taille de l'échantillon ne permet pas d'apprécier à sa juste valeur la dépendance entre ces deux variables. Cependant, il est peu vraisemblable que la corrélation obtenue sur un échantillon plus important soit très forte. En fait, la corrélation entre le poids et la taille sera très forte si l'on considère le "poids moyen" en fonction de la "taille moyenne".

5.4 Analyse descriptive des résidus

Les erreurs estimées par $\hat{\varepsilon}_i = y_i - \hat{y}_i, i = 1, \ldots, n$ sont appelées *résidus*. Ils sont empiriquement centrés par construction,

$$\frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{a}x_i - \hat{b}) = \overline{y} - \hat{a}x - \hat{b} = 0.$$

Mais la représentation des $\hat{\varepsilon}_i$ en fonction des x_i peut révéler que le modèle est mauvais. Cette représentation traduit la disposition des points autour de la droite. On doit alors constater que cette disposition résulte du hasard. Plus exactement, il faut rejeter toute situation dans laquelle la disposition des points autour de la droite aurait un aspect structuré. La Figure 5.4 donne quelques exemples simples de situations structurées pour lesquelles les hypothèses du modèle linéaire ne sont certainement pas satisfaites.

Dans le même esprit, on peut considérer la représentation des résidus $\hat{\varepsilon}_i$ en fonction des valeurs ajustées, $\hat{y}_i = \hat{a}x_i + \hat{b}$, pour mettre en évidence, par exemple, une dispersion des résidus dépendante du niveau de la grandeur étudiée. On observe ainsi que les résidus sont sensiblement plus élevés lorsque la vitesse coronarienne est importante (*cf.* Tableau 5.2 et Figure 5.5). Dans le cas de la régression simple, cette représentation n'ajoute rien à la précédente (changement d'échelle). Elle est par contre très utile en régression linéaire multiple.

$\hat{y}_i$	60,9	62,8	64,7	64,7	66,6	67,3	68,5	69,6	69,6
$\hat{\varepsilon}_i$	0,06	-0,82	-0,71	1,29	-0,59	0,66	0,53	1,40	-0,60
$\hat{y}_i$	70,7	72,6	73,4	74,5	74,9	75,6	75,6	76,4	77,5
$\hat{\varepsilon}_i$	-0,73	-0,61	-1,37	-2,50	2,13	2,37	1,37	0,62	-2,51

TAB. 5.2 – Valeurs ajustées et résidus pour la vitesse coronarienne

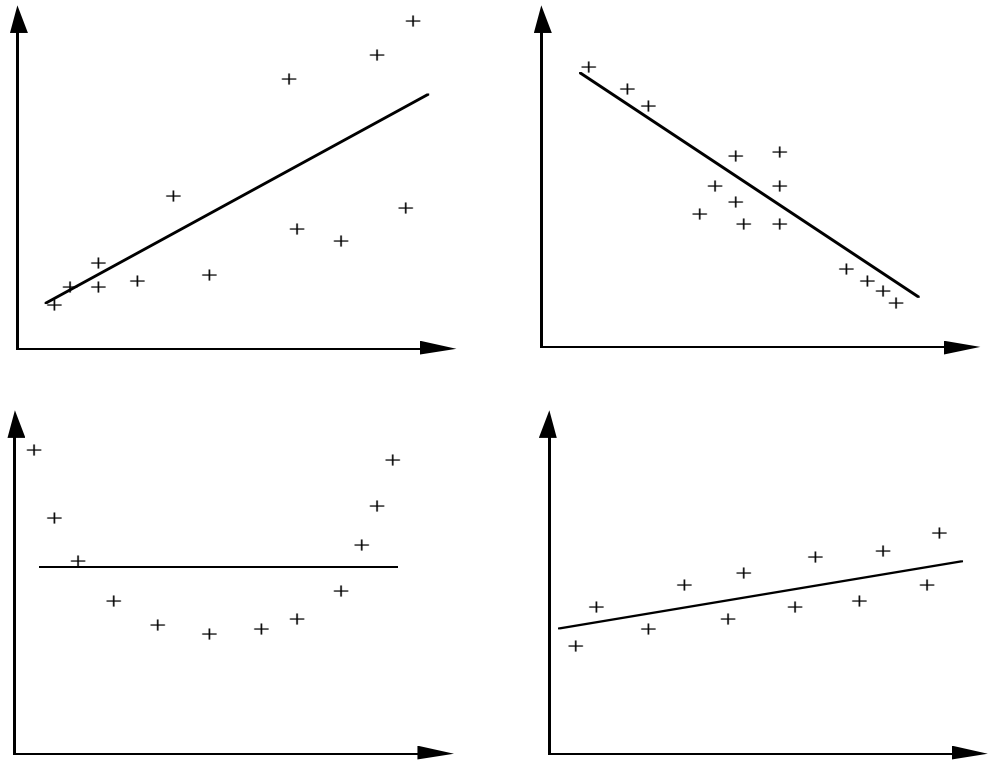

FIG. 5.4 – Exemples de données structurées

5.5 Remarque

L'exemple qui suit illustre une situation où la régression est un non-sens que les arguments statistiques de nature mathématique ne peuvent déceler. Le Tableau 5.3 indique, pour les années 1924 à 1937 en Grande Bretagne, le nombre relatif à 10 000 habitants de certificats de déficience mentale (variable Y), le nombre, en millions, de licences de récepteurs radio (variable X), ainsi que le prénom du Président des États Unis de l'époque (la variable Z est le nombre de lettres) (*cf.* Montgomery & Peck, page 37).

Résumé numérique, caractéristiques empiriques et résultats

$n = 14; \Sigma x = 65,262; \Sigma x^2 = 385,193966; \Sigma y = 208; \Sigma y^2 = 3490; \Sigma xy = 79947.$

$\overline{x} = 4,662; var(x) = 5,783607; \overline{y} = 14,9; var(y) = 28,55; cov(x,y) = 12,7481.$

$\Sigma z = 96; \Sigma z^2 = 668; \Sigma zy = 1485; \overline{z} = 6,9; var(z) = 0,69; cov(z,y) = 4,19.$

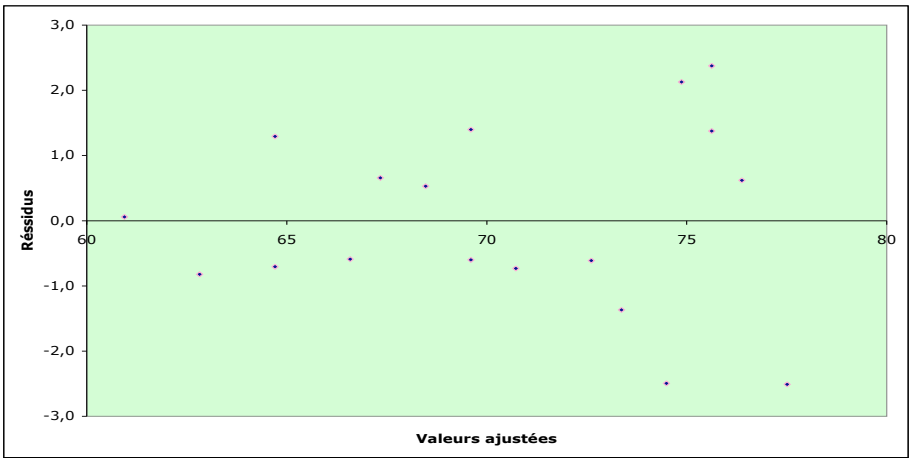

FIG. 5.5 – Résidus en fonction des valeurs ajustées pour la vitesse coronarienne

Année	*Déficience Y*	*Radio X*	*Prénom Z*
1924	8	1,350	Calvin 6
1925	8	1,960	Calvin 6
1926	9	2,270	Calvin 6
1927	10	2,483	Calvin 6
1928	11	2,730	Calvin 6
1929	11	3,091	Calvin 6
1930	12	3,647	Herbert 7
1931	16	4,620	Herbert 7
1932	18	5,497	Herbert 7
1933	19	6,260	Herbert 7
1934	20	7,012	Franklin 8
1935	21	7,618	Franklin 8
1936	22	8,131	Franklin 8
1937	23	8,593	Franklin 8

TAB. 5.3 – Exemple de non-sens

Droites de régression	$y = 2,20x + 4,6$	$y = 6,0z - 26,6$
Coefficients de corrélation linéaire	$r_{xy} = 0,992$	$r_{zy} = 0,94$
Coefficients de détermination	$r^2_{xy} = 98,4\%$	$r^2_{zy} = 89\%$

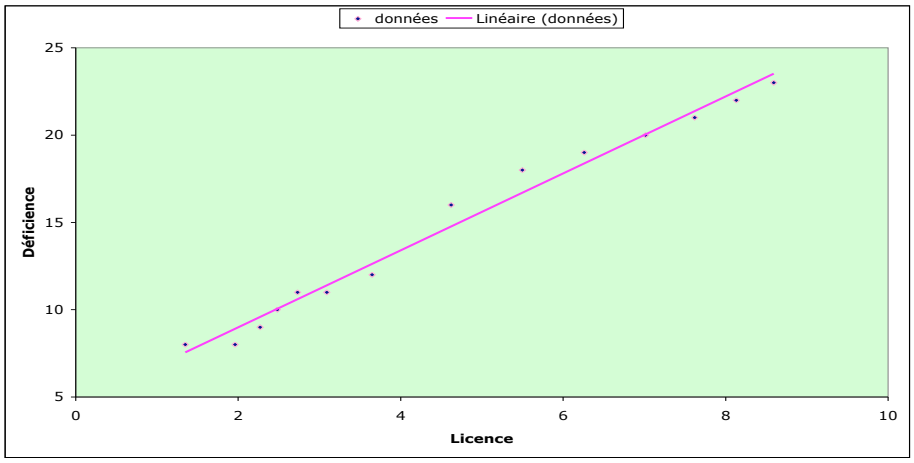

FIG. 5.6 – Déficience mentale en fonction des licences radio

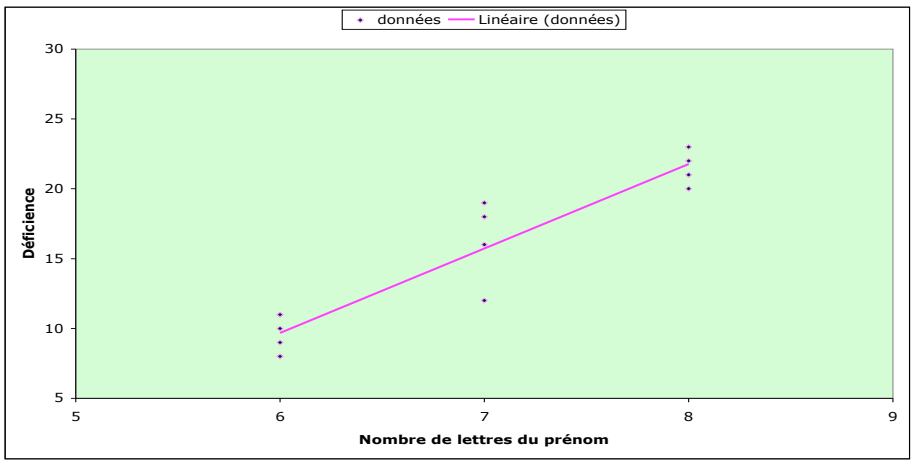

FIG. 5.7 – Déficience mentale en fonction du prénom du président

On considère la régression linéaire du nombre de déficients mentaux en fonction du nombre de licences radio, puis en fonction du nombre de lettres du prénom du Président. Dans les deux cas, le coefficient de détermination

est très élevé et les points sont bien répartis autour de la droite (cf. Figures 5.6 et 5.7). On imagine facilement le type de conclusion erronée que l'on pourrait faire dans la première situation. La deuxième renforce l'absurdité de ce type d'étude. Ce phénomène se produira chaque fois que deux variables, sans aucun lien *a priori*, sont très fortement linéairement corrélées à une même troisième variable externe (ici l'année).

Fonctions d'Excel. Nous indiquons ici les fonctions d'Excel relatives à la régression linéaire, avec les notations introduites dans le chapitre. Si x et y représentent le jeu de données, alors :

- $\overline{x} = \text{MOYENNE}(x)$
- $var(x) = \text{VAR.P}(x)$
- $cov(x, y) = \text{COVARIANCE}(x; y)$
- $\hat{a} = \text{PENTE}(y; x)$
- $\hat{b} = \text{ORDONNEE.ORIGINE}(y; x)$
- $r = \text{COEFFICIENT.CORRELATION}(y; x)$
- $r^2 = \text{COEFFICIENT.DETERMINATION}(y; x)$

Notons aussi que ce logiciel permet de représenter le nuage de points et de faire figurer la droite de régression (tendance).

En résumé. On retiendra le principe des moindres carrés, les expressions de la droite de régression, $y = \hat{a}x + \hat{b}$, et du coefficient de corrélation linéaire empirique r, l'interprétation du coefficient de détermination r^2 et l'utilité de la représentation des résidus $\hat{\varepsilon}_i, i = 1, \ldots, n$.

Chapitre 6

RÉGRESSION LINÉAIRE SIMPLE : APPROCHE INDUCTIVE

6.1 Introduction

Ce chapitre permet d'aller plus loin dans l'étude d'une régression linéaire, en se plaçant dans le cadre d'un modèle probabiliste. Nous pouvons ainsi préciser les propriétés des estimateurs et faire de l'inférence statistique : validation du modèle et de la dépendance entre les deux variables, intervalle de confiance pour la prévision. Le chapitre se termine par l'étude de la régression linéaire à l'aide du logiciel de statistique R.

6.2 Le modèle probabiliste

Il est nécessaire de rappeler la notion de loi normale pour présenter ce modèle.

6.2.1 Loi normale

On dit qu'une variable aléatoire X suit la *loi normale* (ou *gaussienne*) de moyenne m et de variance σ^2, ce que l'on note $X \sim \mathcal{N}(m, \sigma^2)$, lorsque sa fonction densité est donnée par :

$$f(x; m; \sigma) = f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-m)^2}{2\sigma^2}\right\}, \quad x \in \mathbb{R} \quad (\sigma > 0).$$

On montre en effet que l'on a :

- *moyenne* ou *espérance mathématique* : $\mathbb{E}(X) = \int_{-\infty}^{\infty} xf(x)dx = m$,
- *variance* : $Var(X) = \mathbb{E}\{[X - \mathbb{E}(X)]^2\} = \int_{-\infty}^{\infty}(x-m)^2 f(x)dx = \sigma^2$.

La *fonction densité* est symétrique par rapport à la moyenne m. L'aire sous la courbe est toujours égale à 1, mais la "cloche" est plus ou moins "pointue" selon la valeur de σ^2. Ceci est illustré par la Figure 6.1. Théoriquement X prend ses valeurs dans l'intervalle $]-\infty, +\infty[$, mais les résultats suivants montrent que, pratiquement, X varie entre $m - 3\sigma$ et $m + 3\sigma$ (σ est appelé *écart-type*) :

- $P(m - \sigma < X < m + \sigma) = 0,6826$,
- $P(m - 2\sigma < X < m + 2\sigma) = 0,9544$,
- $P(m - 3\sigma < X < m + 3\sigma) = 0,9973$.

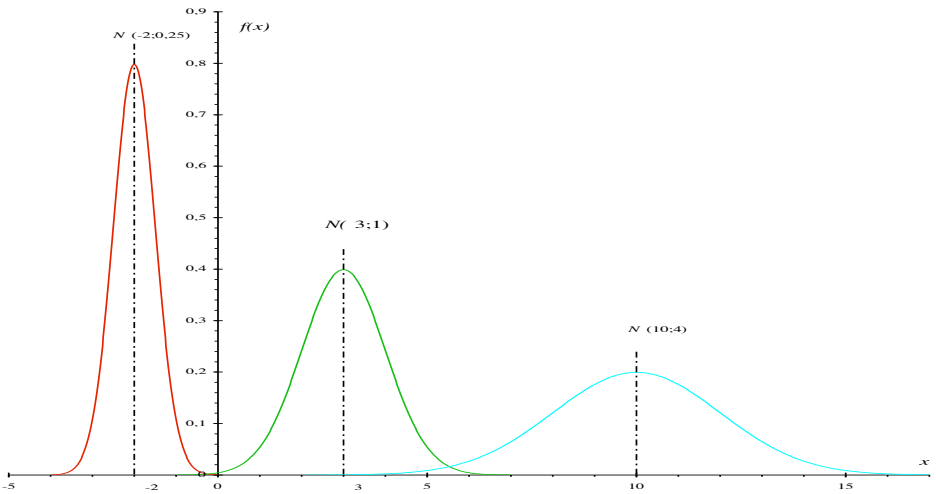

FIG. 6.1 – Densités de lois normales

La *fonction de répartition*,

$$F(x; m; \sigma) = F(x) = P\{X \leq x\} = \int_{-\infty}^{x} f(u)du, \quad x \in \mathbb{R},$$

n'a pas d'expression analytique. Les calculs de probabilités sont effectués à l'aide de la fonction de répartition de la *loi normale centrée réduite* $\mathcal{N}(0,1)$, notée Φ, et de la relation

$$F(x) = \Phi\left(\frac{x-m}{\sigma}\right)$$

provenant de $X = m + \sigma T$, où $T \sim \mathcal{N}(0,1)$. En effet, la fonction Φ est tabulée et figure dans les calculatrices ou logiciels adaptés. La relation entre Φ et la fonction densité de la loi $\mathcal{N}(0,1)$ est illustrée par les Figures 6.2 et 6.3.

Le logiciel Excel propose les fonctions suivantes (avec les notations ci-dessus) :

- $f(x;m;\sigma)$ =LOI.NORMALE($x;m;\sigma$;FAUX), $\quad x \in \mathbb{R}$
- $F(x;m;\sigma)$ =LOI.NORMALE($x;m;\sigma$;VRAI), $\quad x \in \mathbb{R}$
- $F^{-1}(\alpha;m;\sigma)$ =LOI.NORMALE.INVERSE($\alpha;m;\sigma$), $\quad 0 < \alpha < 1$
- $\Phi(x)$ =LOI.NORMALE.STANDARD(x), $\quad x \in \mathbb{R}$
- $\Phi^{-1}(\alpha)$ =LOI.NORMALE.STANDARD.INVERSE(α), $\quad 0 < \alpha < 1$

Signalons deux résultats utiles, le premier étant inclus dans la relation entre X et T ci-dessus.

- $X \sim \mathcal{N}(m;\sigma^2), \lambda \in \mathbb{R} \quad \Longrightarrow \quad Y = \lambda X \sim \mathcal{N}(\lambda m; \lambda^2\sigma^2)$
- $X \sim \mathcal{N}(m_X;\sigma_X^2)$ et $Y \sim \mathcal{N}(m_Y;\sigma_Y^2)$ indépendantes $\quad \Longrightarrow$
  $Z = X + Y \sim \mathcal{N}(m_X + m_Y; \sigma_X^2 + \sigma_Y^2)$

6.2.2 Le modèle

Le *modèle probabiliste* consiste à considérer que les données $y_1, \ldots, y_n$ sont les observations de variables aléatoires $Y_1, \ldots, Y_n$ satisfaisant :

$$Y_i = ax_i + b + \varepsilon_i, \quad i = 1, \ldots, n, \quad \varepsilon_i : i.i.d. \mathcal{N}(0, \sigma^2).$$

De façon précise, les erreurs $\varepsilon_1, \ldots, \varepsilon_n$ sont des variables aléatoires indépendantes et identiquement distribuées ($i.i.d.$) de loi normale, centrée, et de variance σ^2 ($\mathcal{N}(0,\sigma^2)$) :

$$\mathbb{E}(\varepsilon_i) = 0, \quad Var(\varepsilon_i) = \mathbb{E}(\varepsilon_i^2) = \sigma^2, \quad i = 1, \ldots, n.$$

Il s'en suit que les variables aléatoires $Y_1, \ldots, Y_n$ sont également indépendantes, de loi normale, de même variance σ^2, mais de moyennes distinctes données par :

$$\mathbb{E}(Y_i) = ax_i + b, \quad i = 1, \ldots, n.$$

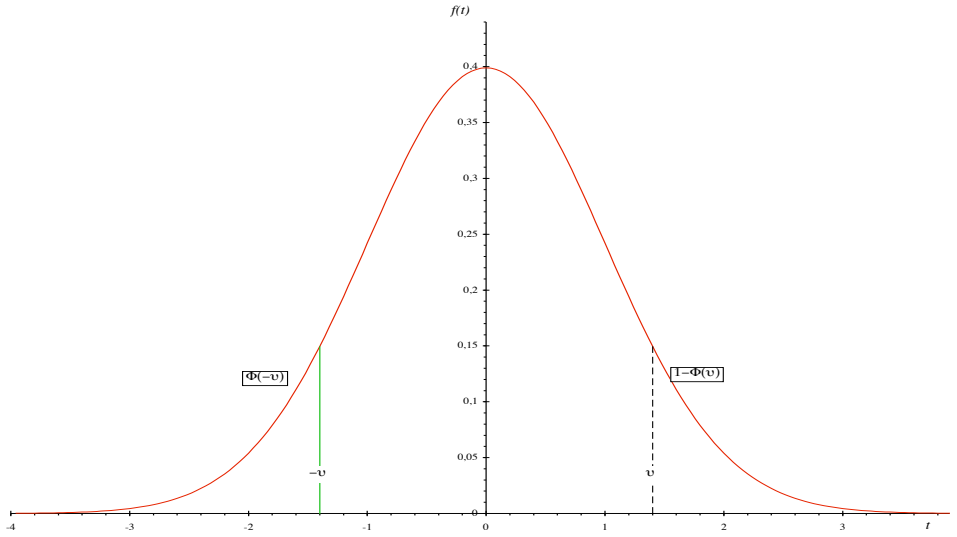

FIG. 6.2 – Densité de la loi normale centrée réduite

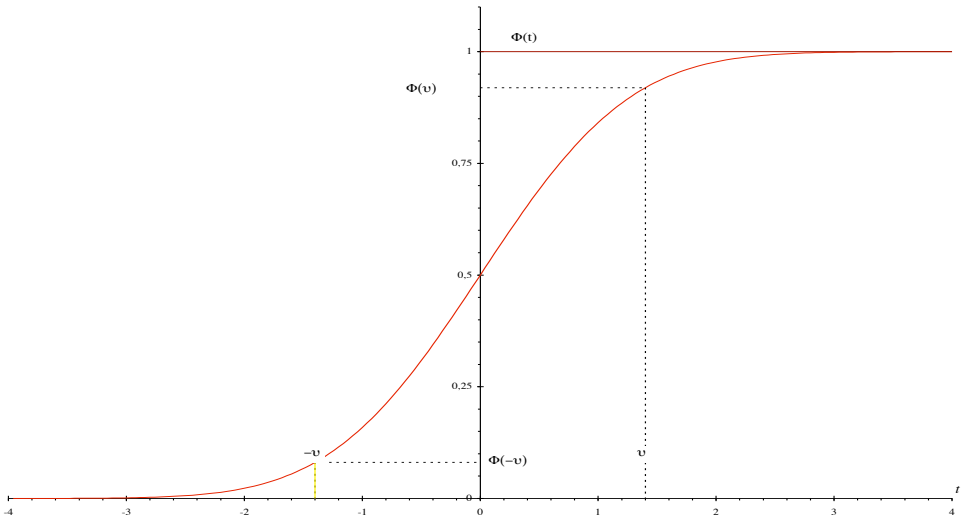

FIG. 6.3 – Fonction de répartition de la loi normale centrée réduite

6.3 Propriétés des estimateurs

6.3.1 Biais et variance

Les estimateurs $\hat{a}$ et $\hat{b}$ sont des combinaisons linéaires des variables $Y_i, i = 1, \ldots, n$:

$$\begin{aligned} \hat{a} &= \frac{cov(x,Y)}{var(x)} = \frac{1}{n\ var(x)} \sum_{i=1}^{n} (x_i - \overline{x}) Y_i, \\ \hat{b} &= \overline{Y} - \hat{a}\overline{x} = \frac{1}{n\ var(x)} \sum_{i=1}^{n} (\overline{x^2} - \overline{x} x_i) Y_i. \end{aligned}$$

En utilisant la linéarité de l'espérance $\mathbb{E}$, on montre que l'on a :

$$\mathbb{E}(\hat{a}) = a, \quad \mathbb{E}(\hat{b}) = b.$$

On dit que ces estimateurs sont *sans biais*, car ils reproduisent, en moyenne, les valeurs qu'ils sont censé estimer. Le calcul des variances repose aussi sur la linéarité de l'espérance, en tenant compte du fait que les variables Y_i sont non correlées :

$$Var(\hat{a}) = \frac{\sigma^2}{n\ var(x)}, \quad Var(\hat{b}) = \frac{\sigma^2 \overline{x^2}}{n\ var(x)}.$$

Ces estimateurs sont les meilleurs au sens où ce sont ceux dont la variance est minimum parmi tous les estimateurs sans biais obtenus par combinaisons linéaires des variables $Y_i, i = 1, \ldots, n$ (propriété de Gauss-Markov).

On montre enfin que $\hat{a}$ et $\hat{b}$ sont également des variables aléatoires de loi normale. Ceci repose sur l'indépendance et la loi normale des variables Y_i.

6.3.2 Estimateur de la variance de l'erreur

On a constaté que les variances de $\hat{a}$ et $\hat{b}$ sont proportionnelles à la variance σ^2 des erreurs ε_i. Il est nécessaire d'estimer ce nouveau paramètre. On montre que

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \frac{1}{n-2} \sum_{i=1}^{n} (Y_i - \hat{a} x_i - \hat{b})^2$$

est un estimateur sans biais de σ^2 : $\mathbb{E}(\hat{\sigma}^2) = \sigma^2$. Notons que si a et b étaient connus, on disposerait des erreurs $\varepsilon_i = Y_i - a x_i - b, i = 1, \ldots, n$ qui satisfont :

$$\mathbb{E}\left(\frac{1}{n} \sum_{i=1}^{n} \varepsilon_i^2 \right) = \sigma^2.$$

Pour obtenir l'estimateur sans biais de σ^2, on a remplacé, dans l'expression ci-dessus, les erreurs ε_i par les résidus $\hat{\varepsilon}_i = Y_i - \hat{a}x_i - \hat{b}$ qui ont nécessité l'estimation de 2 paramètres. Ce manque d'information est compensé en divisant par $n-2$ au lieu de n, ce qui a pour effet d'augmenter le résultat. Dans le même esprit, la variable

$$\sum_{i=1}^{n} \left(\frac{\varepsilon_i}{\sigma}\right)^2$$

suit la loi du chi-deux à n degrés de liberté. On montre que la variable

$$\frac{(n-2)\hat{\sigma}^2}{\sigma^2} = \sum_{i=1}^{n} \left(\frac{\hat{\varepsilon}_i}{\sigma}\right)^2$$

suit la loi du chi-deux à $(n-2)$ degrés de liberté et est indépendante de $\hat{a}$ et $\hat{b}$. Ceci est au centre des résultats du paragraphe suivant.

6.4 Inférence statistique

L'objectif est de décider, au vu de $\hat{a}$ et $\hat{b}$, si les vraies valeurs inconnues a et b sont différentes de zéro. Si c'est le cas pour a, cela signifie que la dépendance linéaire entre Y et x est pertinente. Dans le cas de b, cela permet éventuellement d'utiliser un modèle dans lequel la droite passe par l'origine (situation plutôt rare). Une étude des résidus $\hat{\varepsilon}_i$ est également souhaitable. Après être ainsi rassuré sur la validité du modèle, celui-ci est utilisé pour prédire, sous forme d'un intervalle de confiance, la valeur Y d'un individu pour lequel x est connu. Tout ceci repose sur la loi de Student.

6.4.1 Loi du chi-deux

La *loi du chi-deux* à 1 degré de liberté, notée χ^2_1, est la loi de T^2 lorsque T suit la loi $\mathcal{N}(0;1)$. La loi du chi-deux à n degrés de liberté, notée χ^2_n, est la loi de la somme de n variables indépendantes de loi χ^2_1. Pour une variable Z de loi χ^2_n, on montre que $\mathbb{E}(Z) = n$ et $Var(Z) = 2n$. La fonction densité est donnée par (*cf.* Figure 6.4) :

$$f(z) = \frac{1}{2\Gamma(n/2)} \left(\frac{z}{2}\right)^{\frac{n}{2}-1} e^{-\frac{z}{2}}, \quad z \in \mathbb{R}^{+*}.$$

La fonction gamma $\Gamma(\cdot)$ est une fonction spéciale satisfaisant $\Gamma(x+1) = x\Gamma(x)$ et $\Gamma(n+1) = n!$. La loi du chi-deux ne sera pas directement utilisée ici. Elle sert en fait à définir la loi de Student.

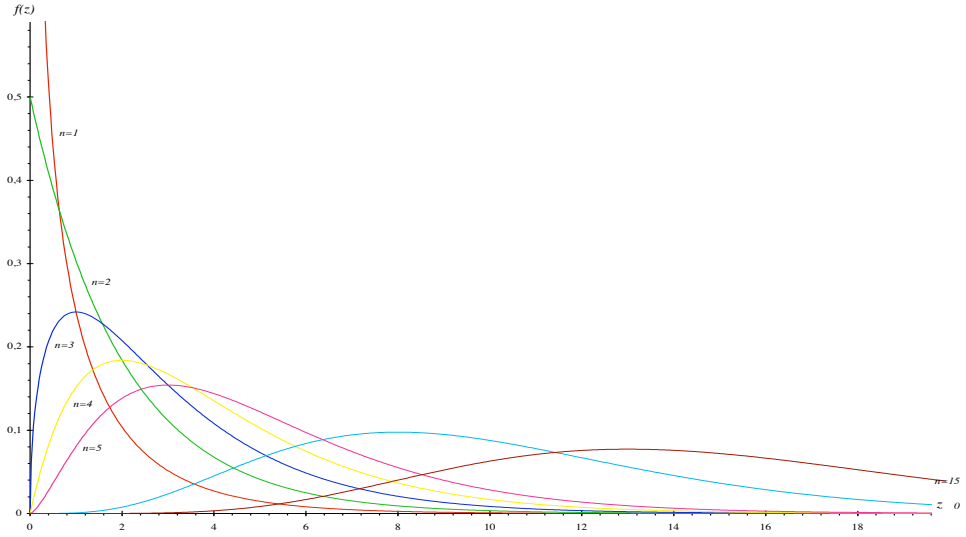

FIG. 6.4 – Densités de lois du chi-deux pour $n = 1, 2, 3, 4, 5, 10$ et 15

6.4.2 Loi de Student

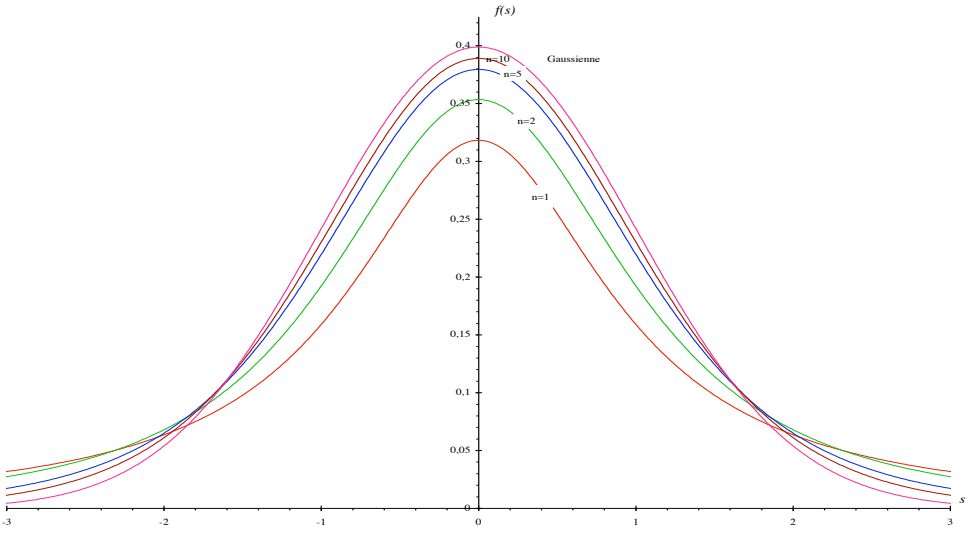

FIG. 6.5 – Densités de lois de Student pour $n = 1, 2, 5$ et 10

La *loi de Student* à n degrés de liberté, notée $\mathcal{S}_n$, est la loi du quotient d'une variable $\mathcal{N}(0,1)$ par la racine carrée d'une variable χ^2_n, préalablement divisée par ses degrés de liberté, les deux variables étant de plus indépendantes :

$$T \sim \mathcal{N}(0,1) \textit{ indépendante de } Z_n \sim \chi^2_n \quad \Rightarrow \quad S_n = \frac{T}{\sqrt{Z_n/n}} \sim \mathcal{S}_n.$$

La fonction densité est donnée par :

$$f_{\mathcal{S}_n}(s) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\sqrt{n\pi}} \left(1 + \frac{s^2}{n}\right)^{-\frac{n+1}{2}}, \quad s \in \mathbb{R}.$$

Cette fonction est symétrique par rapport à l'origine. Son graphe est analogue à celui de la loi $\mathcal{N}(0,1)$. Il devient moins "plat" lorsqque n augmente et tend à se confondre avec celui de la loi $\mathcal{N}(0,1)$ (*cf.* Figure 6.5). La moyenne et la variance de S_n n'existent pas pour les premières valeurs de n :

$$\mathbb{E}(S_n) = 0 \text{ pour } n > 1, \quad Var(S_n) = \frac{n}{n-2} \text{ pour } n > 2.$$

La fonction de répartition est également tabulée et donnée par les calculatrices et logiciels adaptés. Pour $n = 1$, la loi est aussi appelée *loi de Cauchy*.

Le logiciel Excel propose les fonctions suivantes :

- LOI.STUDENT.INVERSE$(\alpha; n) = t_{n;\alpha}$
- LOI.STUDENT$(t_{n;\alpha}; n; 2) = \alpha$
- LOI.STUDENT$(t_{n;\alpha}; n; 1) = \alpha/2$

où $P(|S_n| > t_{n;\alpha}) = \alpha$.

6.4.3 Estimateurs studentisés et p-valeurs

L'idée est d'apprécier les observations de $\hat{a}$ et $\hat{b}$ dans une échelle universelle, c'est-à-dire indépendante des échelles de mesure des variables Y et x. En utilisant les résultats précédents, on a :

$$\frac{\hat{a} - a}{\sqrt{\sigma^2/n\ var(x)}} \div \sqrt{\frac{(n-2)\hat{\sigma}^2}{\sigma^2}/(n-2)} = \frac{\sqrt{n\ var(x)}}{\hat{\sigma}}(\hat{a} - a) \sim \mathcal{S}_{n-2}.$$

Ainsi, si $a = 0$, la variable $\hat{a}^S$ définie par

$$\hat{a}^S = \frac{\sqrt{n\ var(x)}}{\hat{\sigma}}\hat{a},$$

suit la loi de Student à $(n-2)$ degrés de liberté. Son comportement ne dépend plus des données du problème, mais uniquement du nombre n d'observations. On dit que $\hat{a}^S$ est la *version studentisée* de $\hat{a}$. En notant $\hat{a}^S_{obs}$ la valeur observée de la variable $\hat{a}^S$, la probabilité

$$P\left(\left|\hat{a}^S\right| > \left|\hat{a}^S_{obs}\right|\right) = 2\left(1 - F_{\mathcal{S}_{n-2}}\left(\left|\hat{a}^S_{obs}\right|\right)\right),$$

où $F_{\mathcal{S}_{n-2}}$ désigne la fonction de répartition de la loi $\mathcal{S}_{n-2}$, est la *p-valeur* associée à $\hat{a}^S_{obs}$. Elle représente la probabilité que $|\hat{a}^S|$, qui évalue a en un certain sens, s'éloigne de zéro d'au moins $|\hat{a}^S_{obs}|$ alors que a est nul. Si cette p-valeur est très faible, il est naturel de dire que l'hypothèse "$a = 0$" n'est pas réaliste. C'est le principe du test de l'hypothèse H_0 : "$a = 0$" contre l'alternative H_1 : "$a \neq 0$". La *règle de décision* d'un tel test est la suivante :

- si $|\hat{a}^S| > c$, on rejette H_0 (on décide H_1),
- si $|\hat{a}^S| < c$, on accepte H_0,

où la *valeur critique* c est associée à un choix de α par :

$$P(|\hat{a}^S| > c | a = 0) = \alpha \quad \Leftrightarrow \quad c = F^{-1}_{\mathcal{S}_{n-2}}(1 - \alpha/2).$$

La quantité α est la probabilité de rejeter à tort H_0. C'est le *risque de première espèce* fixé par l'expérimentateur (en général on prend $\alpha = 5\%$). C'est aussi le *niveau de signification* du test qui, de façon générale, est le maximum du risque de première espèce. Le *risque de deuxième espèce* est la probabilité de rejeter à tort H_1, ou encore d'accepter à tort H_0. Il dépend de a, mais aussi de σ. Son maximum est ici égal à $1 - \alpha$, car a peut être très proche de zéro sans être nul. Cependant, lorsque a s'éloigne de zéro, ce risque diminue.
Plutôt que de déterminer la valeur critique c en fonction du niveau de signification α, il suffit de comparer la p-valeur à α : on rejette H_0 lorsque la p-valeur est inférieure à α. Notons qu'une p-valeur très faible rassure sur le fait que a est vraiment différent de zéro. Par contre, une p-valeur proche de un ne renforce pas le fait que a soit nul, bien que l'on accepte cette hypothèse.

On détermine de la même manière la version studentisée de $\hat{b}$ et la p-valeur associée :

$$\hat{b}^S = \frac{\sqrt{n\ var(x)}}{\sqrt{\overline{x^2}}\hat{\sigma}}\hat{b}, \quad P\left(\left|\hat{b}^S\right| > \left|\hat{b}^S_{obs}\right|\right) = 2\left(1 - F_{\mathcal{S}_{n-2}}\left(\left|\hat{b}^S_{obs}\right|\right)\right).$$

Illustration

Les résultats concernant la vitesse coronarienne sont présentés dans le Tableau 6.1. l'*erreur standard* d'un estimateur est l'estimation de son écart-type. La version studentisée est donc le rapport entre l'estimation et l'erreur standard associée. Les calculs utilisent les résultats intermédiaires suivants :

$$\sum_{i=1}^{n} \hat{\varepsilon}_i^2 = 33,9692\ldots; \quad \hat{\sigma}^2 = 2,1230\ldots; \quad \hat{\sigma} = 1,4570\ldots.$$

Notons que $\sum_{i=1}^{n} \hat{\varepsilon}_i^2$ peut être obtenu par $\sum_{i=1}^{n} \hat{\varepsilon}_i^2 = n\ (1 - r^2)var(y)$, sans arrondir r^2. Les p-valeurs montrent que a et b sont non nuls sans aucune

Paramètre	*Estimation*	*Erreur standard*	*Studentisation*	*p-valeur*
Pente a	$\hat{a} = -0,3766$	$\hat{\sigma}(\hat{a}) = 0,0264$	$\hat{a}^S_{obs} = -14,2651$	$1,6E-10$
Ordonnée b	$\hat{b} = 94,4536$	$\hat{\sigma}(\hat{b}) = 1,7254$	$\hat{b}^S_{obs} = 54,7435$	$1,2E-19$

TAB. 6.1 – Inférence statistique pour la vitesse coronarienne

ambiguïté.

Sous Excel, l'écart-type estimé $\hat{\sigma}$ est donné par :

$$\hat{\sigma} = \text{ERREUR.TYPE}(y; x)$$

6.4.4 Résidus standardisés

Les résidus $\hat{\varepsilon}_i, i = 1, \ldots, n$ sont également des variables gaussiennes, comme combinaisons linéaires des variables $Y_i, i = 1, \ldots, n$. On montre qu'ils sont centrés, $\mathbb{E}(\hat{\varepsilon}_i) = 0$, mais aussi de façon empirique, puisque $\frac{1}{n}\sum_{i=1}^{n} \hat{\varepsilon}_i = 0$. Ils sont corrélés entre eux et satisfont :

$$Var(\hat{\varepsilon}_i) = \frac{\sigma^2}{n}\left[n - 1 - \frac{(x_i - \overline{x})^2}{var(x)}\right], i = 1, \ldots, n.$$

On appelle *résidus standardisés* (ou *résidus studentisés*, bien que $\hat{\varepsilon}_i$ ne soit pas indépendant de $\hat{\sigma}^2$), notés $\hat{\varepsilon}_i^S$, les variables normalisées par l'estimation de l'écart-type,

$$\hat{\varepsilon}_i^S = \frac{\hat{\varepsilon}_i}{\frac{\hat{\sigma}}{\sqrt{n}}\sqrt{n - 1 - \frac{(x_i - \overline{x})^2}{var(x)}}}, \quad i = 1, \ldots, n.$$

La variable $\hat{\varepsilon}_i^S$ suit approximativement la loi $\mathcal{S}_{n-2}$. La comparaison de $|\hat{\varepsilon}_i^S|$ avec $t_{n-2;\alpha} = F^{-1}_{\mathcal{S}_{n-2}}(1-\alpha/2), i = 1, \ldots, n$, permet éventuellement de rejeter le modèle. On peut représenter les résidus standardisés en fonction des valeurs ajustées, en faisant figurer la bande délimitée par les seuils $\pm t_{n-2;\alpha}$. Il ne s'agit pas d'une véritable bande de confiance. Cependant, si le nombre de points au dehors de cette bande est largement supérieur à $n\alpha$, le modèle doit être rejeté.

Illustration

Les résidus et leur version standardisée pour la vitesse coronarienne sont donnés dans le Tableau 6.2. La Figure 6.6 montre que les résidus standardisés se situent bien à l'intérieur de la "bande de confiance".

i	1	2	3	4	5	6	7	8	9
$\hat{\varepsilon}_i$	-2,5	0,6	2,4	1,4	2,1	-2,5	-1,4	-0,6	-0,7
$\hat{\varepsilon}_i^S$	-1,9	0,5	1,7	1,0	1,5	-1,8	-1,0	-0,4	-0,5
i	10	11	12	13	14	15	16	17	18
$\hat{\varepsilon}_i$	1,4	-0,6	0,5	0,7	-0,6	-0,7	1,3	-0,8	-2,5
$\hat{\varepsilon}_i^S$	1,0	-0,4	0,4	0,5	-0,4	-0,5	1,0	-0,6	0,0

TAB. 6.2 – Résidus et résidus standardisés pour la vitesse coronarienne

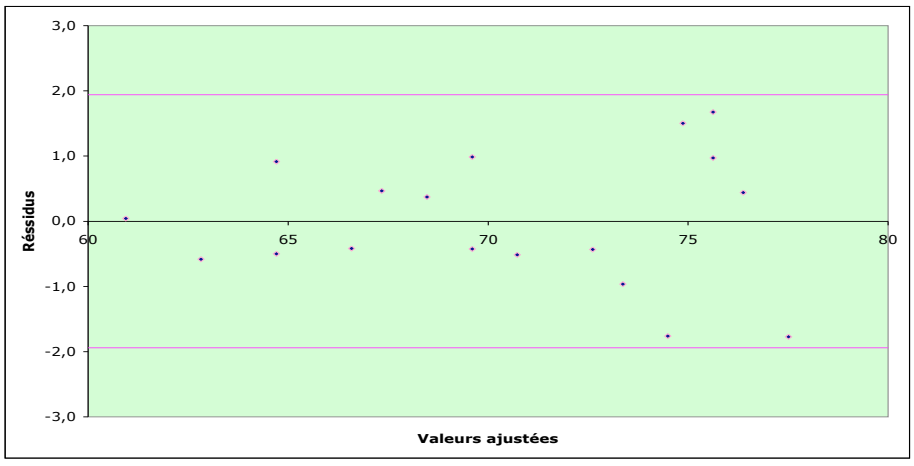

FIG. 6.6 – Résidus standardisés en fonction des valeurs ajustées ; $t_{n-2;\alpha} = 2,1$

6.4.5 Prévision

Lorsque le modèle est validé, il peut être utilisé pour faire une *prévision*. On considère un nouvel individu pour lequel la variable explicative x est connue et égale à x_0. Il est naturel de lui prévoir la valeur $\hat{y}_0 = \hat{a}x_0 + \hat{b}$ pour la variable expliquée Y. La vraie valeur Y_0 satisfait

$$Y_0 = ax_0 + b + \varepsilon_0, \quad \varepsilon_0 \sim \mathcal{N}(0, \sigma^2),$$

où ε_0 est une erreur indépendante des variables $Y_i, i = 1, \ldots, n$ et donc de $\hat{a}, \hat{b}$ et $\hat{\sigma}^2$. La statistique $\hat{Y}_0 = \hat{a}x_0 + \hat{b}$ est un estimateur sans biais de $\mathbb{E}(Y_0) = ax_0 + b$, mais aussi de la valeur inconnue Y_0 (car la meilleure prévision de ε_0 est 0). On montre que :

$$Var(\hat{Y}_0) = \frac{\sigma^2}{n} \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - x_0)^2}{var(x)} = \frac{\sigma^2}{n}\left[1 + \frac{(x_0 - \overline{x})^2}{var(x)}\right].$$

Il est nécessaire de tenir compte de la variance de l'erreur ε_0 pour définir l'intervalle de confiance, appelé ici *intervalle de prévision*. Sous l'hypothèse gaussienne, la variable $Y_0 - \hat{Y}_0$ est normale, centrée, et sa variance est donnée par :

$$Var(Y_0 - \hat{Y}_0) = Var(Y_0) + Var(\hat{Y}_0) = \frac{\sigma^2}{n}\left[n + 1 + \frac{(x_0 - \overline{x})^2}{var(x)}\right].$$

L'intervalle de prévision, avec une confiance de $(1 - \alpha)$, est alors :

$$IC(Y_0; 1 - \alpha) = \hat{a}x_0 + \hat{b} \pm t_{n-2;\alpha}\frac{\hat{\sigma}}{\sqrt{n}}\sqrt{n + 1 + \frac{(x_0 - \overline{x})^2}{var(x)}},$$

où $t_{n-2;\alpha}$ désigne le quantile d'ordre $1 - \alpha/2$: $P\{|S_{n-2}| > t_{n-2;\alpha}\} = \alpha$. La quantité $(1 - \alpha)$ est la probabilité que cet intervalle aléatoire contienne la valeur inconnue Y_0 :

$$P\left\{|Y_0 - \hat{Y}_0| < t_{n-2;\alpha}\frac{\hat{\sigma}}{\sqrt{n}}\sqrt{n + 1 + \frac{(x_0 - \overline{x})^2}{var(x)}}\right\} = 1 - \alpha.$$

Notons que l'amplitude de cet intervalle est minimum lorsque $x_0 = \overline{x}$.

Illustration

Pour un individu de poids $x_0 = 65kg$, on prévoit une vitesse coronarienne de :

$$\hat{y}_0 = \hat{a}x_0 + \hat{b} = -0,377 \times 65 + 94,5 = 69,995 \simeq 70,0.$$

L'intervalle de prévision à 95% est :

$$70,0 \pm 2,1 \frac{1,5}{\sqrt{18}} \sqrt{18 + 1 + \frac{(65 - 64,1)^2}{169,27}} = 70,0 \pm 3,2 = [66,8; 73,2].$$

Sous Excel, la prévision est réalisée par :

$$\hat{y}_0 = \text{PREVISION}(x_0; y; x)$$

Cette fonction permet également de déterminer les valeurs ajustées :

$$\hat{y}_i = \text{PREVISION}(x_i; y; x)$$

6.5 Régression linéaire sous R

Le *logiciel R* est un logiciel de statistique gratuit accessible sur le site : http ://cran.cict.fr/ . Une documentation en français est accessible à : http ://cran.r-project.org/doc/contrib/Paradis-rdebuts_fr.pdf http ://math.univ-lille1.fr/ philippe/Anne-Philippe-cours-R.pdf

Nous ne prétendons pas assurer ici une formation à R. Nous indiquons simplement les commandes minimales pour réaliser une régression linéaire et exploiter les résultats. Pour cela, nous reprenons l'illustration concernant la vitesse coronarienne en fonction du poids.

6.5.1 Aspects numériques

- Enregistrement des données dans les vecteurs x et y
  > x <-c(45,48,50,50,52,53,56,58,63,66,66,69,72,74,79,79,84,89)
  > y <-c(75,77,78,77,77,72,72,72,70,71,69,69,68,66,64,66,62,61)
- Création de l'objet "vitcor" par la fonction lm(·) (linear model)
  > vitcor<-lm(y ∼ x)

Cet objet contient toutes les informations liées à la régression linéaire de y sur x. Nous donnons ci-après les commandes utiles, immédiatement suivies des réponses de R.

- Lecture des valeurs de $\hat{a}$ et $\hat{b}$
  > vitcor
  Call :
  lm(formula = $y \sim x$)

Coefficients :

(Intercept)	x
94.4536	-0.3766

- Bilan de statistique inductive

```
> summary(vitcor)
Call :
lm(formula = y ~ x)

Residuals :
     Min       1Q   Median       3Q      Max
 -2.5087  -0.7246  -0.2646   1.1351   2.3740
Coefficients :
             Estimate   Std. Error  t value  Pr(> |t|)
 (Intercept)  94.4536       1.7254    54.74   < 2e-16  ***
 x            -0.3765       0.0264   -14.27   1.62e-10 ***
—
Signif. codes : 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error : 1.457 on 16 degrees of freedom
Multiple R-Squared : 0.9271, Adjusted R-squared : 0.9225
F-statistic : 203.5 on 1 and 16 DF, p-value : 1.619e-10
```

On reconnaît, dans ce bilan, la plupart des éléments définis précédemment. La rubrique "Coefficients" donne l'inférence sur les paramètres a et b indiquée dans le Tableau 6.1. On a $\hat{\sigma} = 1,457$, avec 16 degrés de liberté pour la loi de Student de $(n-2)\hat{\sigma}^2/\sigma^2$, et $r^2 = 0,9271$. La valeur 0,9225 du r^2 ajusté ne concerne pas ce cours. Les résultats de "F-statistic" concernent le test de Fisher qui sera présenté au chapitre suivant. Ce test équivaut ici au test de Student pour l'hypothèse $a = 0$ $(203,5 \simeq (-14,27)^2)$. En ce qui concerne les résidus $\hat{\varepsilon}_i$, la rubrique "Residuals" fournit quelques informations : minimum, maximum, 1^{er} et 3^e quartiles, ainsi que la médiane. En fait, il est possible de disposer de l'ensemble des valeurs des résidus avec la commande,

```
> vitcor$residuals
```

qui affiche :

1	2	3	4	5	6
-2.50874282	0.62091348	2.37401769	1.37401769	2.12712189	-2.49632601
7	8	9	10	11	12
-1.36666971	-0.61356550	-0.73080500	1.39885131	-0.60114869	0.52850761
13	14	15	16	17	18
0.65816392	-0.58873188	-0.70597137	1.29402863	-0.82321087	0.05954964

ou, sous forme arrondie :

```
> round(vitcor$residuals,2)
    1     2     3     4     5     6     7     8     9    10    11    12
-2.51  0.62  2.37  1.37  2.13 -2.50 -1.37 -0.61 -0.73  1.40 -0.60  0.53
   13    14    15    16    17    18
 0.66 -0.59 -0.71  1.29 -0.82  0.06
```

L'accès direct aux différents résultats est possible. Notons par exemple l'équivalence des commandes suivantes :

```
>round(vitcor$coef[2],3)        >round(vitcor$coef["x"],3)
    x
-0.377
```

- Valeurs ajustées $\hat{y}_i, i = 1, \ldots, n$ arrondies

```
> round(fitted(vitcor),1)
   1    2    3    4    5    6    7    8    9   10   11   12
77.5 76.4 75.6 75.6 74.9 74.5 73.4 72.6 70.7 69.6 69.6 68.5
  13   14   15   16   17   18
67.3 66.6 64.7 64.7 62.8 60.9
```

- Prévision $\hat{y}_0$ et intervalle de prévision arrondi

```
> predict(vitcor,data.frame(x=65))
[1] 69.9777
> round(predict(vitcor, data.frame(x=65),interval="prediction"),1)
      fit  lwr  upr
[1,]   70 66.8 73.2
```

Citons aussi trois commandes utiles :

- Liste des objets disponibles dans la session

```
> ls()
[1] "vitcor" "x" "y"
```

- Paramètres associés à un objet

```
> names(vitcor)
 [1] "coefficients"  "residuals"  "effects"  "rank"  "fitted.values"
 [6] "assign"        "qr"         "df.residual"  "xlevels"  "call"
[11] "terms"         "model"
```

- Autres paramètres

```
> names(summary(vitcor))
[1] "call"          "terms"         "residuals"     "coefficients"  "sigma"
[6] "df"            "r.squared"     "adj.r.squared" "fstatistic"    "cov.unscaled"
```

En particulier, on reconnaît $\hat{\sigma}$ accessible par :

```
> summary(vitcor)$sigma
[1] 1.457078
```

6.5.2 Aspects graphiques

Le logiciel R propose de nombreux résultats sous forme graphique.

- Représentation du nuage de points avec la droite de régression associée (*cf.* Figure 6.7)
  > plot(x,y,main="Vitesse coronarienne en fonction du poids", xlab="poids", ylab="vitesse")
  > abline(vitcor)

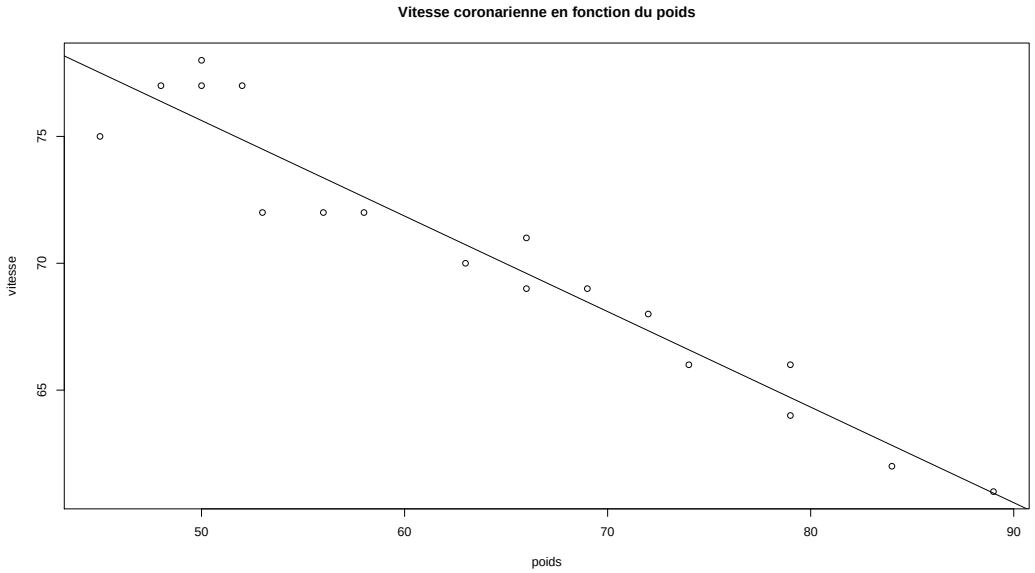

FIG. 6.7 – Nuage de points et droite de régression sous R

- Représentation des résidus en fonction des valeurs ajustées (*cf.* Figure 6.8)
  > plot(vitcor$fitted.values,vitcor$residuals,main="Résidus en fonction des valeurs ajustées",xlab="valeurs ajustées",ylab="résidus")

- Représentations diverses liées à l'inférence statistique du modèle linéaire
  > par(mfrow=c(2,2))
  > plot(vitcor)

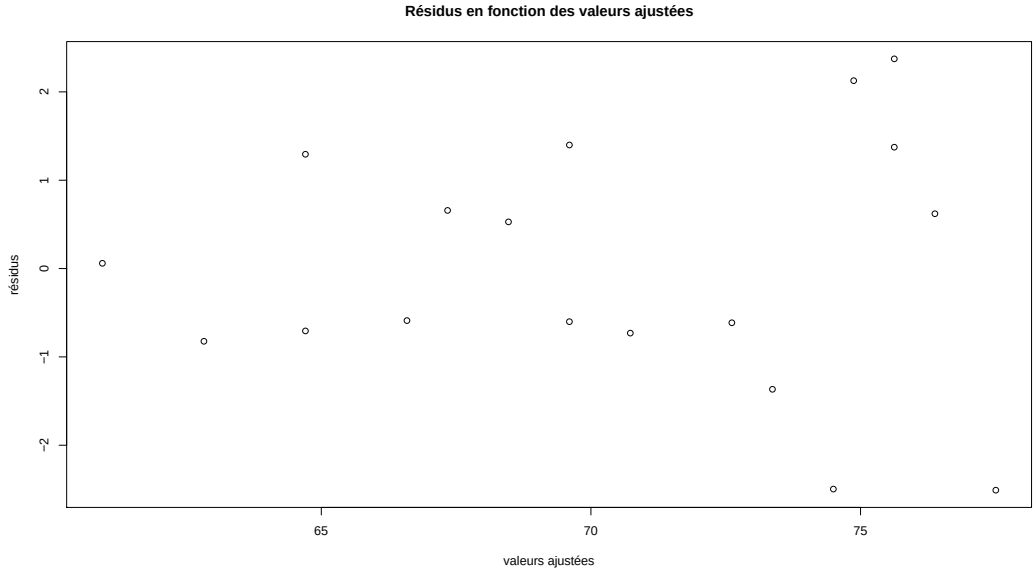

FIG. 6.8 – Résidus versus valeurs ajustées sous R

La Figure 6.9 présente quatre graphiques :

- Le premier (Residuals vs Fitted) donne les résidus en fonction des valeurs ajustées. Il correspond à la représentation de la Figure 6.8 déjà effectuée directement.
- Le second (Normal Q-Q plot) représente les quantiles empiriques des résidus standardisés $\hat{\varepsilon}_i^S$ en fonction des quantiles théoriques de la loi $\mathcal{N}(0;1)$. C'est un test graphique concernant l'hypothèse faite sur les erreurs ε_i. Les points doivent sembler alignés sur la diagonale principale du rectangle.
- Le troisième (Scale-Location plot) est le graphe des points $(\hat{y}_i, \sqrt{\hat{\varepsilon}_i^S}), i = 1, \ldots, n$. Il permet de détecter les "points douteux", ici ceux associés aux indices 1, 3 et 6. Notons que ces points ne sortent pas de la "bande de confiance" dans la Figure 6.6.
- Le dernier (Cook's distance plot) mesure l'influence de chaque point sur les paramètres de la droite de régression : plus la distance de Cook est importante, plus le point est influent, c'est-à-dire que sa suppression dans le jeu de données modifie de façon sensible la position de la droite. On retrouve les points d'indice 1, 3 et 6. Notons que ces point sont également signalés sur les deux premiers graphes.

Remarque L'accès aux valeurs des résidus standardisés est possible en chargeant préalablement la librairie MASS par la commande "library(MASS)". Les résidus sont alors donnés par la commande "stdres(vitcor)".

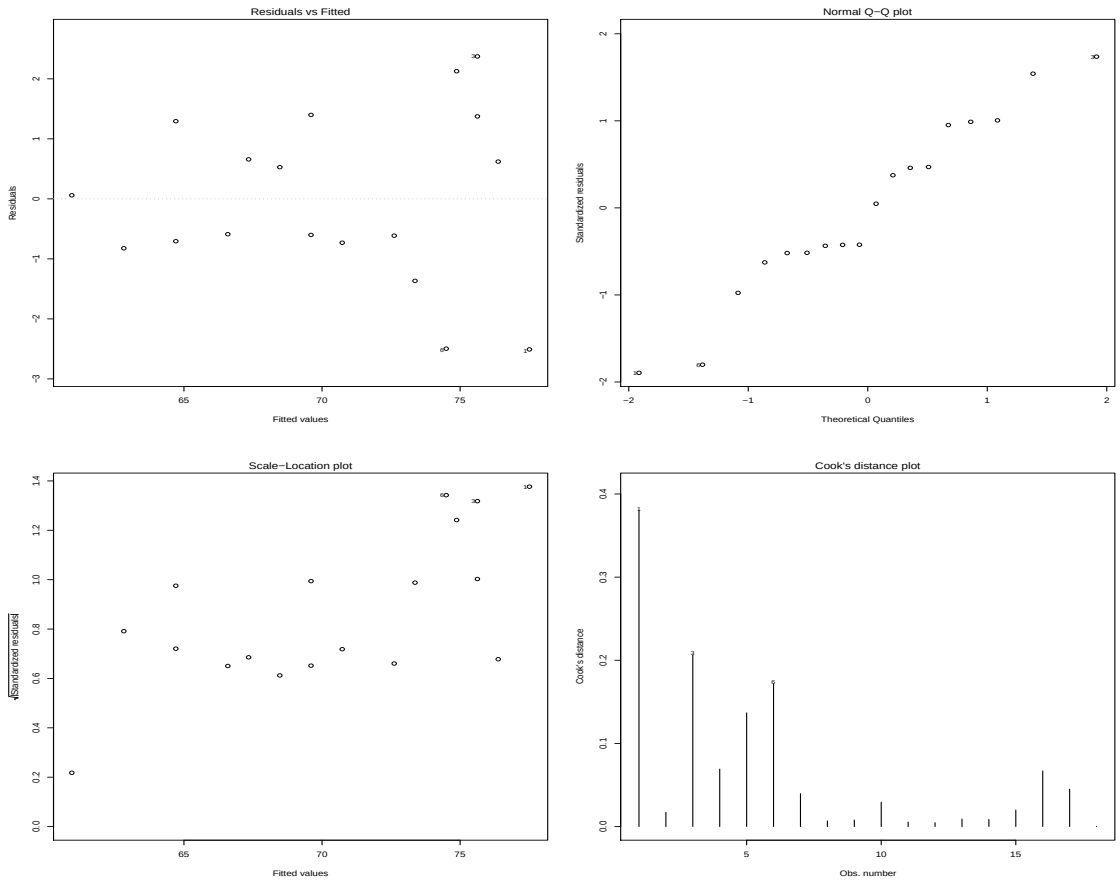

FIG. 6.9 – Inférence statistique du modèle linéaire sous R

En résumé On retiendra comment les versions studentisées des résidus, celles des estimateurs des paramètres de la tendance et les p-valeurs associées permettent de justifier l'utilisation du modèle à des fins de prévision, qui peut alors être exprimée sous forme d'un intervalle de confiance. On doit également être capable de réaliser une régression linéaire simple sous R et d'exploiter les résultats disponibles.

Chapitre 7

RÉGRESSION LINÉAIRE MULTIPLE

7.1 Introduction

La *régression linéaire multiple* est une extension naturelle de la régression linéaire simple faisant appel à plusieurs variables explicatives. La conséquence fondamentale est que l'expression des différents résultats (estimateurs, valeurs ajustées, résidus, ...) nécessite le recours au *calcul matriciel*. Cependant, nous ne détaillerons pas cet aspect qui se cache derrière les résultats numériques fournis par le logiciel R. En effet, ce logiciel permet la mise en œuvre de la méthode et l'interprétation des résultats de façon très semblable à ce qui a été présenté dans le cadre de la régression linéaire simple. Le chapitre se présente donc comme le traitement d'un exemple avec le logiciel R pour lequel nous donnerons, sans justification, la symbolique des écritures matricielles.

7.2 Les hypothèses du modèle

7.2.1 Illustration

On étudie le salaire d'enseignants-chercheurs américains, dans les années 70, en fonction de leur ancienneté et de leur mérite. Les données, ainsi que les valeurs ajustées et les résidus, sont présentées dans le Tableau 7.1. La *variable expliquée* Y est le salaire annuel en centaines de dollars. Les *variables explicatives* X_1 et X_2 sont respectivement l'ancienneté, exprimée en nombre d'années après la thèse, et le mérite mesuré par le nombre de publications.

Indice	*Salaire* Y	*Ancienneté* X_1	*Mérite* X_2	*Valeur ajustée* $\hat{Y}$	*Résidu* $\hat{\varepsilon}$
i	Y_i	x_{i1}	x_{i2}	$\hat{Y}_i$	$\hat{\varepsilon}_i$
1	381	17	15	297	84
2	195	14	7	271	-76
3	208	2	10	223	-15
4	351	24	8	317	34
5	146	8	10	249	-103
6	275	9	13	259	16
7	229	4	6	225	4
8	192	5	6	230	-38
9	227	4	7	227	0
10	247	12	8	264	-17
11	238	5	3	225	13
12	340	5	7	231	109
13	205	4	5	224	-19
14	238	3	4	218	20
15	181	7	7	240	-59
16	268	20	6	296	-28
17	273	3	3	216	57
18	255	7	5	237	18
19	250	13	5	263	-13
20	302	20	16	312	-10
21	271	13	12	275	-4
22	337	24	19	335	2
23	148	3	2	214	-66
24	293	15	12	284	9
25	301	5	1	222	79

TAB. 7.1 – Données, valeurs ajustées et résidus des salaires d'enseignants-chercheurs

7.2.2 Notations

On écrit le modèle sous la forme,

$$Y_i = \theta_0 + \theta_1 x_{i1} + \theta_2 x_{i2} + \varepsilon_i = \theta_0 + \sum_{j=1}^{p} \theta_j x_{ij} + \varepsilon_i, \quad i = 1, \ldots, n,$$

où les *erreurs* ε_i sont des variables aléatoires indépendantes, de loi normale centrée et de variance σ^2 et $p = 2$ est le nombre de variables explicatives.

Cela signifie que les observations y_i sont celles de variables aléatoires Y_i indépendantes, de loi normale de moyenne $\theta_0 + \theta_1 x_{i1} + \theta_2 x_{i2}$ et de variance σ^2. On peut résumer ces écritures sous forme vectorielle :

$$\begin{aligned}
\begin{bmatrix} Y_1 \\ \vdots \\ Y_i \\ \vdots \\ Y_n \end{bmatrix} &= \theta_0 \begin{bmatrix} 1 \\ \vdots \\ 1 \\ \vdots \\ 1 \end{bmatrix} + \theta_1 \begin{bmatrix} x_{11} \\ \vdots \\ x_{i1} \\ \vdots \\ x_{n1} \end{bmatrix} + \theta_2 \begin{bmatrix} x_{12} \\ \vdots \\ x_{i2} \\ \vdots \\ x_{n2} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_i \\ \vdots \\ \varepsilon_n \end{bmatrix} \\
&= \theta_0 \begin{bmatrix} 1 \\ \vdots \\ 1 \\ \vdots \\ 1 \end{bmatrix} + \sum_{j=1}^{p} \theta_j \begin{bmatrix} x_{1j} \\ \vdots \\ x_{ij} \\ \vdots \\ x_{nj} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_i \\ \vdots \\ \varepsilon_n \end{bmatrix},
\end{aligned}$$

symbolisée par :

$$Y = \theta_0 \mathbb{1} + \theta_1 X_1 + \theta_2 X_2 + \varepsilon = \theta_0 \mathbb{1} + \sum_{j=1}^{p} \theta_j X_j + \varepsilon.$$

L'étape suivante est l'écriture matricielle :

$$\begin{aligned}
\begin{bmatrix} Y_1 \\ \vdots \\ Y_i \\ \vdots \\ Y_n \end{bmatrix} &= \begin{bmatrix} 1 & x_{11} & x_{12} \\ \vdots & \vdots & \vdots \\ 1 & x_{i1} & x_{i2} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_i \\ \vdots \\ \varepsilon_n \end{bmatrix} \\
&= \begin{bmatrix} x_{10} & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{i0} & x_{i1} & \cdots & x_{ip} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n0} & x_{n1} & \dots & x_{np} \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_i \\ \vdots \\ \varepsilon_n \end{bmatrix},
\end{aligned}$$

symbolisée par :

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

où l'on a posé $X_0 = \mathbb{1}$. La matrice X, de dimension $n \times (p+1)$, est appelée *plan d'expérience*, Y, de dimension n est le vecteur des observations, θ, de dimension $(p+1)$, est le vecteur des paramètres et ε, de dimension n, est le

vecteur d'erreur.

Les hypothèses probabilistes se formalisent comme suit. $\mathbb{E}(\varepsilon) = 0$, où 0 désigne ici le vecteur nul de dimension n, indique que les composantes ε_i de ε sont centrées, $\mathbb{E}(\varepsilon_i) = 0, i = 1, \ldots, n$. $Var(\varepsilon) = \sigma^2 I_n$, où I_n est la matrice identité d'ordre n,

$$I_n = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix},$$

est la matrice de variance-covariance du vecteur aléatoire ε. Les termes diagonaux, égaux à σ^2, traduisent $Var(\varepsilon_i) = \sigma^2, i = 1, \ldots, n$ et les termes hors diagonale, égaux à 0, indiquent que la covariance entre deux composantes distinctes est nulle, $Cov(\varepsilon_i, \varepsilon_k) = 0$ si $i \neq k$. Enfin, les variables ε_i étant normales et indépendantes, on dit que le vecteur aléatoire ε est normal, ou gaussien, de moyenne nulle et de matrice de variance-covariance $\sigma^2 I_n$, ce que l'on résume par $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$. Selon le même formalisme, on a $Y \sim \mathcal{N}(X\theta, \sigma^2 I_n)$ et en particulier $\mathbb{E}(Y) = X\theta$.

7.3 Estimateur des moindres carrés

7.3.1 Critère des moindres carrés

Le *critère des moindres carrés* s'écrit :

$$\begin{aligned} D_{Y/X}(\theta) &= \frac{1}{n}\sum_{i=1}^{n}[Y_i - \theta_0 - \theta_1 x_{i1} - \theta_2 x_{i2}]^2 = \frac{1}{n}\left\|Y - \theta_0 \mathbb{1} - \theta_1 X_1 - \theta_2 X_2\right\|^2 \\ &= \frac{1}{n}\left\|Y - \theta_0 \mathbb{1} - \sum_{j=1}^{p}\theta_j X_j\right\|^2 = \frac{1}{n}\left\|Y - \sum_{j=0}^{p}\theta_j X_j\right\|^2 = \frac{1}{n}\|Y - X\theta\|^2. \end{aligned}$$

Il n'est pas possible de visualiser ce critère dans l'espace des variables. Par contre, on retrouve la symbolique de l'*espace des observations* sur la Figure 7.1. Le paramètre θ étant inconnu, il en est de même du vecteur $\mathbb{E}(Y) = X\theta$. On sait cependant que ce vecteur se situe dans l'*espace des moyennes* $\mathcal{M}(X)$, c'est-à-dire qu'il est une combinaison linéaire des colonnes de la matrice X. Le critère des moindres carrés consiste à retenir, parmi tous ces vecteurs, celui qui est le plus proche de Y, il s'agit du vecteur $\hat{Y} = X\hat{\theta}$, projection orthogonale de Y sur $\mathcal{M}(X)$. Il se caractérise par le fait que la

différence $\hat{\varepsilon} = Y - \hat{Y} = Y - X\hat{\theta}$ est orthogonale aux colonnes de X :

$$ {}^tX[Y - X\hat{\theta}] = 0 \quad \Leftrightarrow \quad {}^tXX\hat{\theta} = {}^tXY \quad \Leftrightarrow \quad \hat{\theta} = ({}^tXX)^{-1}\,{}^tXY. $$

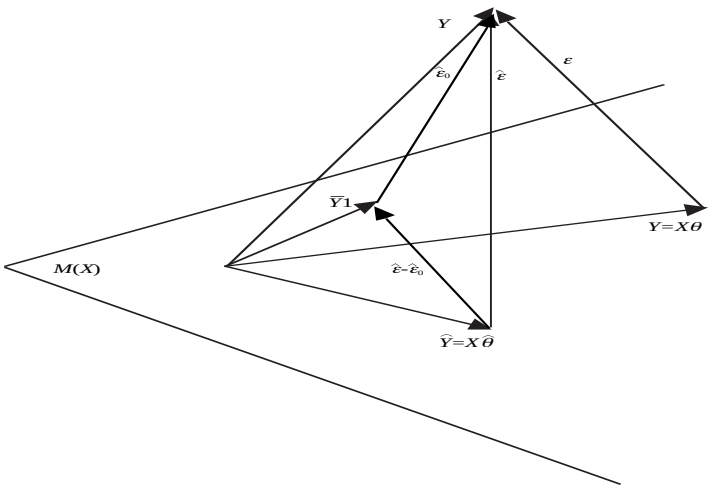

FIG. 7.1 – Principe des moindres carrés dans l'espace des observations

$\hat{\theta}$ est l'estimateur des moindres carrés de θ, les *valeurs ajustées* sont les composantes de $\hat{Y} = X\hat{\theta}$ et $\hat{\varepsilon}$ est le vecteur des résidus. La représentation des résidus en fonction des valeurs ajustées, c'est-à-dire des points $(\hat{y}_i, \hat{\varepsilon}_i), i = 1, \ldots, n$, permet de déceler une éventuelle anomalie dans le modèle : résidus structurés, points aberrants, ...

7.3.2 Illustration

La mise en œuvre, sous R, de la régression linéaire multiple s'effectue sans difficulté.

- Enregistrement des données dans les vecteurs y, x1 et x2
  > y < c(381,195,208,351,146,275,229,192,227,247,238,340,205,238,181,268, 273,255,250,302,271,337,148,293,301)
  > x1 <-c(17,14,2,24,8,9,4,5,4,12,5,5,4,3,7,20,3,7,13,20,13,24,3,15,5)
  > x2 <-c(15,7,10,8,10,13,6,6,7,8,3,7,5,4,7,6,3,5,5,16,12,19,2,12,1)
- Création de l'objet "salaire" par la fonction lm(·)
  > salaire<-lm(y ∼ x1+x2)

- Lecture des valeurs de $\hat{\theta}_0, \hat{\theta}_1$ et $\hat{\theta}_2$

  > salaire
  Call :
  lm(formula = $y \sim x1 + x2$)

  Coefficients :

(Intercept)	x1	x2
197.812	4.417	1.619

  L'équation de la régression s'écrit : $y = 198 + 4,42x_1 + 1,62x_2$.

- Valeurs ajustées $\hat{Y}$ arrondies

  > round(fitted(salaire))
  1 2 3 ... (*cf.* Tableau 7.1)
  297 271 223 ...

- Résidus $\hat{\varepsilon}$ arrondis

  > round(salaire$residuals)
  1 2 3 ... (*cf.* Tableau 7.1)
  84 -76 -15 ...

- Représentation des résidus en fonction des valeurs ajustées (*cf.* Figure 7.2)

  > plot(salaire$fitted.values,salaire$residuals,main="Résidus en fonction des valeurs ajustées",xlab="valeurs ajustées",ylab="résidus")

- Bilan de statistique inductive (*cf.* Section 7.4)

  > summary(salaire)
  Call :
  lm(formula = y ~ x1 + x2)

  Residuals :

Min	1Q	Median	3Q	Max
-103.3452	-18.5784	0.1827	18.1694	108.7653

  Coefficients :

	Estimate	Std. Error	t value	Pr(> |t|)	
(Intercept)	197.812	21.435	9.228	< 5.1e-09	***
x1	4.417	1.995	2.214	0.0375	*
x2	1.619	3.042	0.532	0.5998	

  —
  Signif. codes : 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

  Residual standard error : 51.6 on 22 degrees of freedom
  Multiple R-Squared : 0.343, Adjusted R-squared : 0.2832
  F-statistic : 5.742 on 2 and 22 DF, p-value : 0.009851

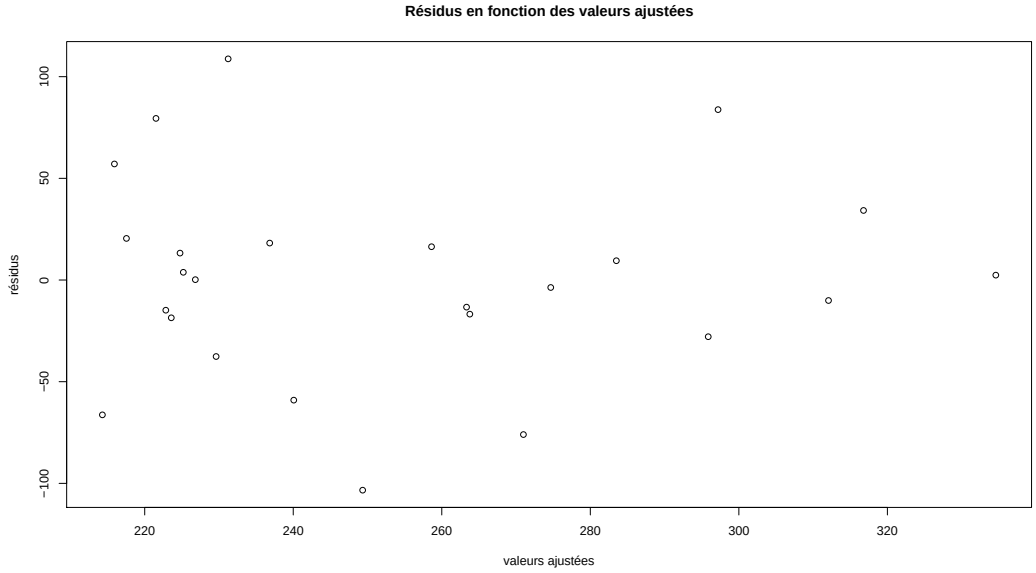

FIG. 7.2 – Résidus versus valeurs ajustées pour les salaires

7.4 Compléments statistiques

Nous donnons ci-dessous quelques explications sur les résultats statistiques fournis par le logiciel R.

7.4.1 Coefficient de corrélation linéaire multiple

Comme en régression linéaire simple, la variance empirique de Y se décompose sous la forme :

$$var(y) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2,$$

où $\hat{y}_i = \hat{\theta}_0 + \hat{\theta}_1 x_{i1} + \hat{\theta}_2 x_{i2}, i = 1, \ldots, n$. Ce résultat correspond, dans l'espace des observations, à la relation :

$$y - \overline{y}\mathbb{1} = (y - \hat{y}) \oplus (\hat{y} - \overline{y}\mathbb{1}) \quad \Longrightarrow \quad ||y - \overline{y}\mathbb{1}||^2 = ||y - \hat{y}||^2 + ||\hat{y} - \overline{y}\mathbb{1}||^2,$$

utilisant l'orthogonalité des vecteurs $(y - \hat{y})$ et $(\hat{y} - \overline{y}\mathbb{1})$ (théorème de pythagore). Notons que la moyenne empirique $\overline{y}$ de Y est égale à celle de $\hat{Y}$:

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n} \hat{y}_i \quad (\text{car } \frac{1}{n}\sum_{i=1}^{n} \hat{\varepsilon}_i = 0).$$

Le premier terme de cette décomposition est égal à la valeur minimum $D_{Y/X}(\hat{\theta})$ du critère. En introduisant le *coefficient de corrélation linéaire multiple*,

$$R = \frac{cov(y,\hat{y})}{\sqrt{var(y)}\sqrt{var(\hat{y})}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y})(\hat{y}_i-\overline{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i-\overline{y})^2}},$$

cette décomposition s'écrit

$$var(y) = (1-R^2)var(y) + R^2 var(y).$$

Notons que R est toujours positif et correspond à $|r|$ en régression linéaire simple. Il satisfait $0 \leq R \leq 1$ et son carré R^2, appelé *coefficient de détermination*, représente la part de variance de Y expliquée par la liaison linéaire entre Y et les variables explicatives $X_1, \ldots, X_p$.

Dans notre illustration, on a $R^2 = 0,343 = 34,3\%$ et $R = 0,586$, traduisant une corrélation moyenne.

7.4.2 Estimateur de la variance de l'erreur

L'estimateur $\hat{\sigma}^2$ de la variance σ^2 des erreurs $\hat{\varepsilon}_i$ est donné par :

$$\hat{\sigma}^2 = \frac{1}{n-(p+1)}\sum_{i=1}^{n}\hat{\varepsilon}_i^2 = \frac{\|\hat{\varepsilon}\|^2}{n-(p+1)}.$$

C'est un estimateur sans biais, $\mathbb{E}(\hat{\sigma}^2) = \sigma^2$, et la variable aléatoire

$$\frac{[n-(p+1)]\hat{\sigma}^2}{\sigma^2} = \frac{\|\hat{\varepsilon}\|^2}{\sigma^2},$$

suit la loi du chi-deux à $n-(p+1)$ degrés de liberté. Elle est indépendante de $\hat{\theta}$ et permet donc de construire les versions studentisées de ses composantes.

Dans notre illustration, on a $\hat{\sigma} = 51,6$ et $n-(p+1) = 25-(2+1) = 22$ degrés de liberté.

7.4.3 Estimateurs studentisés et p-valeurs

L'estimateur $\hat{\theta}$ de θ est sans biais, $\mathbb{E}(\hat{\theta}) = \theta$, ce qui signifie que cela est vrai pour chaque composante : $\mathbb{E}(\hat{\theta}_j) = \theta_j, j = 0, \ldots, p$. La matrice de variance-covariance de $\hat{\theta}$, donnée par

$$Var(\hat{\theta}) = \sigma^2({}^tXX)^{-1},$$

permet de construire les versions studentisées des estimateurs. Notant $\nu_0^2, \nu_1^2, \ldots, \nu_p^2$, les éléments diagonaux de la matrice $({}^tXX)^{-1}$, on a $Var(\hat{\theta}_j) = \sigma^2\nu_j^2$. L'estimateur $\hat{\theta}_j$ suit la loi $\mathcal{N}(\theta_j; \sigma^2\nu_j^2)$ et est indépendant de $\hat{\sigma}^2$. Selon la construction habituelle, la variable

$$\hat{\theta}_j^S = \frac{\hat{\theta}_j}{\hat{\sigma}\nu_j},$$

est la *version studentisée* de $\hat{\theta}_j$. Lorsque $\theta_j = 0$, cette variable suit la loi de Student à $n - (p+1)$ degrés de liberté et la probabilité,

$$P(|\hat{\theta}_j^S| > |\hat{\theta}_{j,obs}^S|) = 2(1 - F_{\mathcal{S}_{n-(p+1)}}(|\hat{\theta}_{j,obs}^S|)),$$

est la *p-valeur* associée à l'observation, notée $\hat{\theta}_{j,obs}^S$, de $\hat{\theta}_j^S$. Rappelons que cette p-valeur représente la probabilité que $\hat{\theta}_j^S$, qui évalue θ_j, s'éloigne de zéro d'au moins $|\hat{\theta}_{j,obs}^S|$ alors que $\theta_j = 0$. L'estimation $\hat{\sigma}(\hat{\theta}_j) = \hat{\sigma}\nu_j$ de l'écart-type $\sigma\nu_j$ de $\hat{\theta}_j$ est appelée *erreur standard* de $\hat{\theta}_j$.

Illustration

La commande "summary(salaire)" de R fournit les valeurs observées de $\hat{\theta}_j, \hat{\sigma}(\hat{\theta}_j), \hat{\theta}_j^S$, ainsi que les p-valeurs associées. Elles sont rappelées dans le Tableau 7.2. La constante est très significativement différente de zéro (p-valeur $= 5 \times 10^{-9}$). La dépendance par rapport à l'ancienneté est significative au niveau 5% ($0,0375 < 0,05$), par contre, elle ne l'est pas par rapport au mérite (p-valeur = 0,5998).

Paramètre	*Estimation*	*Erreur standard*	*Studentisation*	*p-valeur*
constante : θ_0	$\hat{\theta}_0 = 197,812$	$\hat{\sigma}(\hat{\theta}_0) = 21,435$	$\hat{\theta}_{0,obs}^S = 9,228$	$5,1E-0,9$
ancienneté : θ_1	$\hat{\theta}_1 = 4,417$	$\hat{\sigma}(\hat{\theta}_1) = 1,995$	$\hat{\theta}_{1,obs}^S) = 2,214$	0,0375
mérite : θ_2	$\hat{\theta}_2 = 1,619$	$\hat{\sigma}(\hat{\theta}_2) = 3,042$	$\hat{\theta}_{2,obs}^S = 0,532$	0,5998

TAB. 7.2 – Inférence statistique pour le salaire des enseignants-chercheurs

Paramètre	*Estimation*	*Erreur standard*	*Studentisation*	*p-valeur*
constante : θ_0	$\hat{\theta}_0 = 203,832$	$\hat{\sigma}(\hat{\theta}_0) = 17,923$	$\hat{\theta}_{0,obs}^S = 11,37$	$6,4E-11$
ancienneté : θ_1	$\hat{\theta}_1 = 5,102$	$\hat{\sigma}(\hat{\theta}_1) = 1,501$	$\hat{\theta}_{1,obs}^S) = 3,40$	0,00246

TAB. 7.3 – Inférence statistique pour le salaire en fonction de l'ancienneté

On considère alors la régression linéaire simple du salaire par rapport à l'ancienneté. Les résultats sont reportés dans le Tableau 7.3 et la Figure 7.3. Le coefficient de détermination diminue peu : $r^2 = 0,3345 < R^2 = 0,343$. Il est donc raisonnable de ne conserver que l'ancienneté comme variable explicative.

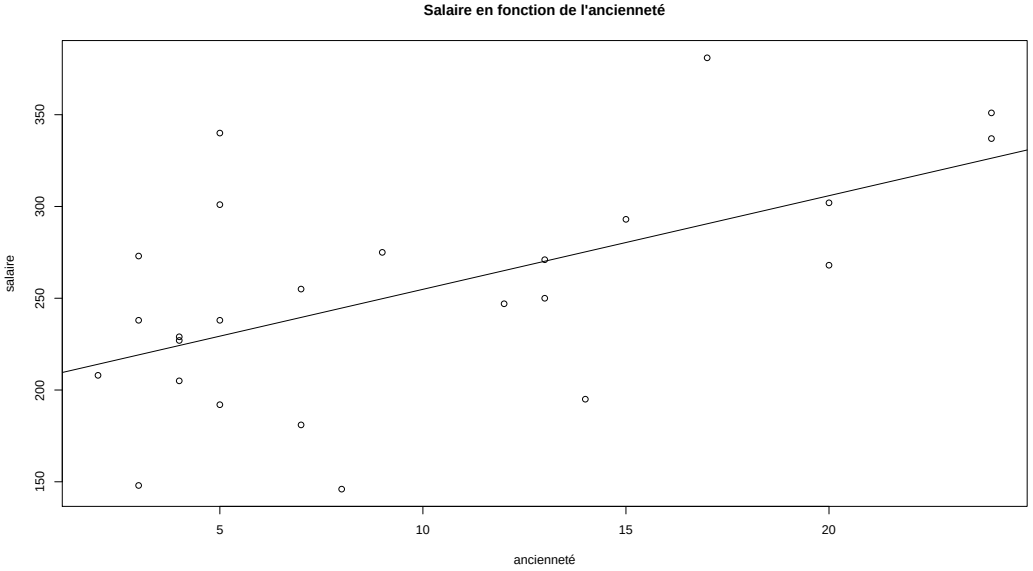

FIG. 7.3 – Régression du salaire en fonction de l'ancienneté

7.4.4 Résidus standardisés

Les *résidus standardisés* $\hat{\varepsilon}_i^S$ sont obtenus, selon le principe habituel, en divisant les résidus $\hat{\varepsilon}_i$ par leur écart-type estimé. Pour cela, il est nécessaire de connaître la matrice de variance-covariance du vecteur aléatoire $\hat{\varepsilon}$. On introduit la *matrice chapeau*, $H = X({}^tXX)^{-1}\,{}^tX$, qui traduit la projection orthogonale de Y sur $\mathcal{M}(X)$ au sens où $\hat{Y} = X\hat{\theta} = HY$. On a alors

$$\hat{\varepsilon} = Y - \hat{Y} = Y - HY = (I_n - H)Y, \quad Var(\hat{\varepsilon}) = \sigma^2(I_n - H).$$

En notant h_{ii} les éléments diagonaux de H, on a $Var(\hat{\varepsilon}_i) = \sigma^2(1 - h_{ii})$ et

$$\hat{\varepsilon}_i^S = \frac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}, \quad i = 1, \ldots, n.$$

Le logiciel R ne donne pas accès, de façon directe, aux résidus standardisés (utilisation de la librairie MASS). Par contre, il fournit la représentation

graphique des points $(\hat{y}_i, \sqrt{\hat{\varepsilon}_i^S}), i = 1, \ldots, n$ (*cf.* Figure 7.5). De façon approximative, $\hat{\varepsilon}_i^S$ suit la loi de Student à $n-(p+1)$ degrés de liberté. On peut utiliser l'égalité

$$P\left(\sqrt{|\hat{\varepsilon}_i^S|} > \sqrt{t_{n-(p+1);\alpha}}\right) = P(|\hat{\varepsilon}_i^S| > t_{n-(p+1);\alpha}) = 1-\alpha,$$

où $t_{n-(p+1);\alpha} = F^{-1}_{\mathcal{S}_{n-(p+1)}}(1-\alpha/2)$, pour apprécier l'ordre de grandeur des résidus. Notons que sous R, le quantile $t_{n-(p+1);\alpha}$ est donné par $t_{n-(p+1);\alpha} = \text{qt}(1-\alpha/2, n-(p+1))$.
Pour l'illustration, avec $\alpha = 5\%$, on a qt(0.975,22)=2.073873 et $\sqrt{2.073873} = 1,44$. Parmi les "points douteux" signalés sur le graphe de la Figure 7.5 (indices 1, 5 et 12), seuls ceux d'indice 5 et 12 sont clairement au delà de 1,44.

7.4.5 Test de Fisher

Le *test de Fisher* est un test global permettant de décider si la variable expliquée Y dépend de façon significative de l'ensemble des variables explicatives $X_1, X_2, \ldots, X_p$ (hors la constante X_0). L'hypothèse nulle H_0 se traduit par $\theta_1 = \theta_2 = \ldots = \theta_p = 0$ et l'alternative H_1 signifie donc qu'au moins l'un des θ_j est non nul. Le principe du test consiste à comparer $\hat{\sigma}^2$ avec un autre estimateur $\hat{\sigma}_1^2$ de σ^2 qui, sous H_0, est également sans biais et indépendant de $\hat{\sigma}^2$ (le rapport devrait alors être proche de 1) alors que, sous H_1, cet estimateur tend à être trop grand, car $\mathbb{E}(\hat{\sigma}_1^2) > \sigma^2$.

Loi de Fisher

Le test nécessite d'introduire la *loi de Fisher*. C'est la loi du rapport de deux chi-deux indépendants, divisés par leurs degrés de liberté :

$$X \sim \chi^2_n \text{ et } Y \sim \chi^2_m \text{ indépendants} \quad \Rightarrow \quad Z = \frac{X/n}{Y/m} \sim \mathcal{F}_{n,m},$$

où $\mathcal{F}_{n,m}$ désigne la loi de Fisher à n et m degrés de liberté. La fonction densité,

$$f_{n,m}(z) = n^{n/2} m^{m/2} \frac{\Gamma(\frac{n+m}{2})}{\Gamma(\frac{n}{2})\Gamma(\frac{m}{2})} z^{n/2-1}(m+nz)^{-(n+m)/2}, \quad z > 0,$$

est représentée sur la Figure 7.4 pour quelques valeurs des paramètres n et m. On établit les résultats suivants :

$$E(Z) = \frac{m}{m-2} \text{ si } m > 2, \quad Var(Z) = \frac{2m^2}{n} \frac{n+m-2}{(m-2)^2(m-4)} \text{ si } m > 4.$$

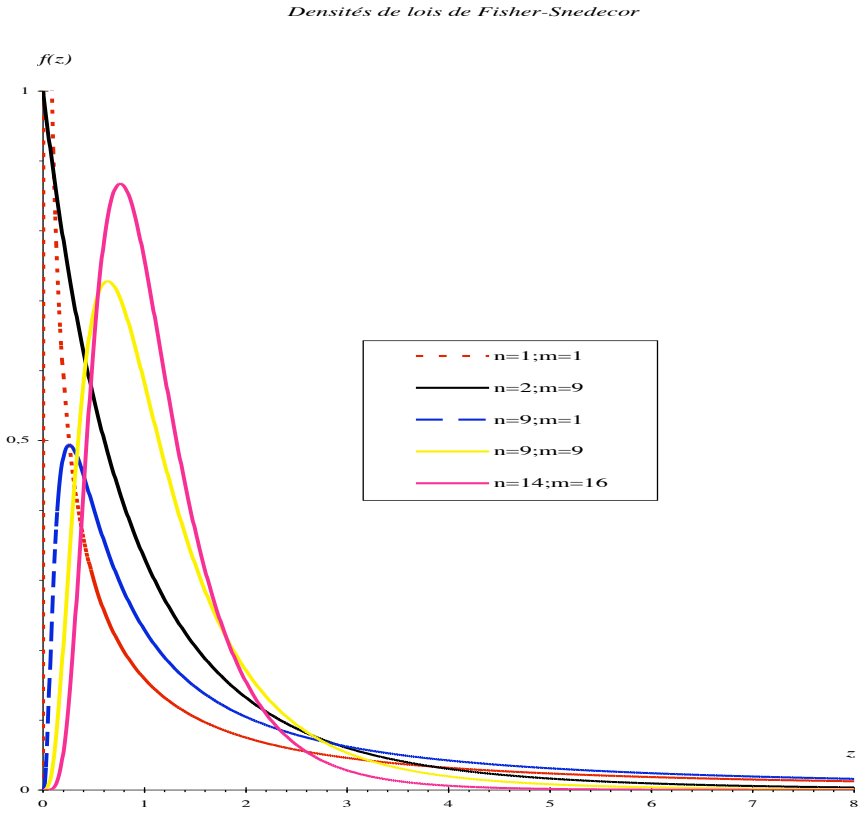

FIG. 7.4 – Exemples de densités de la loi de Fisher

Construction du test

Sous l'hypothèse nulle H_0, le modèle s'écrit :

$$Y = \theta_0 \mathbb{1} + \varepsilon, \quad \varepsilon \sim \mathcal{N}(\theta_0 \mathbb{1}; \sigma^2 I_n).$$

L'estimateur des moindres carrés de θ_0 est $\hat{\theta}_0 = \overline{Y}$. Il est sans biais, $\mathbb{E}(\hat{\theta}_0) = \theta_0$, et $Var(\hat{\theta}_0) = \sigma^2/(n-1)$. L'estimateur sans biais de σ^2 est

$$\hat{\sigma}_0^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \frac{\|Y - \overline{Y}\mathbb{1}\|^2}{n-1} = \frac{\|\hat{\varepsilon}_0\|^2}{n-1},$$

et $(n-1)\hat{\sigma}_0^2/\sigma^2$ suit la loi χ^2_{n-1}. Cependant, cet estimateur n'est pas indépendant de $\hat{\sigma}^2$ car $\|\hat{\varepsilon}_0\|^2 > \|\hat{\varepsilon}\|^2$. Par contre,

$$\hat{\sigma}_1^2 = \frac{\|\hat{\varepsilon}_0 - \hat{\varepsilon}\|^2}{p} = \frac{\|\hat{Y} - \overline{Y}\mathbb{1}\|^2}{p} = \frac{\|\hat{\varepsilon}_0\|^2 - \|\hat{\varepsilon}\|^2}{p},$$

est aussi un estimateur sans biais de σ^2, si H_0 est vraie. Dans ce cas, $p\hat{\sigma}_1^2/\sigma^2$ suit la loi χ^2_p et est indépendant de $\hat{\sigma}^2$. Ainsi, la statistique,

$$F = \frac{\hat{\sigma}_1^2}{\hat{\sigma}^2} = \frac{n-(p+1)}{p}\frac{\|\hat{\varepsilon}_0|^2 - \|\hat{\varepsilon}\|^2}{\|\hat{\varepsilon}\|^2},$$

suit la loi $\mathcal{F}_{p,n-(p+1)}$ si H_0 est vraie. Pour un niveau de signification α donné, on rejettera donc l'hypothèse nulle lorsque la valeur observée F_{obs} de F est supérieure à $f_{p,n-(p+1);\alpha} = F^{-1}_{p,n-(p+1)}(1-\alpha)$, où $F_{p,n-(p+1)}$ désigne la fonction de répartition de la loi $\mathcal{F}_{p,n-(p+1)}$, ce qui équivaut à constater que la p-valeur associée à F_{obs} est inférieure à α.

Notons que les relations

$$\|\hat{\varepsilon}\|^2 = (1-R^2)n\ var(Y), \quad \|\hat{\varepsilon}_0\|^2 = n\ var(y)$$

montrent que l'on a :

$$F = \frac{n-(p+1)}{p}\frac{R^2}{1-R^2}.$$

Illustration

Les résultats concernant le test de Fisher sont donnés par R dans "summary(salaire)" : $F_{obs} = 5,742$ avec 22 degrés de liberté, p-valeur = 0,009851. La dépendance est donc significative au niveau 1% (0,009851<0,01).

7.5 Compléments graphiques sous R

Les représentations graphiques, liées à l'inférence statistique du modèle linéaire, sont obtenues par :
> par(mfrow=c(2,2))
> plot(salaire)

La Figure 7.5 présente quatre graphes :

- Le premier (Residuals vs Fitted) donne les résidus en fonction des valeurs ajustées. Il correspond à la représentation de la Figure 7.2 déjà effectuée directement.
- Le second (Normal Q-Q plot) représente les quantiles empiriques des résidus standardisés $\hat{\varepsilon}_i^S$ en fonction des quantiles théoriques de la loi $\mathcal{N}(0;1)$. C'est un test graphique concernant l'hypothèse faite sur les erreurs ε_i. Les points doivent sembler alignés sur la diagonale principale du rectangle.
- Le troisième (Scale-Location plot) est le graphe des points $(\hat{y}_i, \sqrt{\hat{\varepsilon}_i^S}), i = 1, \ldots, n$. Il permet de détecter les "points douteux", ici ceux associés aux indices 1, 5 et 12.
- Le dernier (Cook's distance plot) mesure l'influence de chaque point sur les paramètres de la régression : plus la distance de Cook est importante, plus le point est influent, c'est-à-dire que sa suppression dans le jeu de données modifie de façon sensible les paramètres de la régression. On retrouve les points d'indice 1 et 5, mais pas 12. Par contre, le point d'indice 25 est signalé.

7.6 Prévision

7.6.1 Principe

Pour de nouvelles valeurs x_{0j} des variables explicatives X_j, la variable Y_0 satisfait :

$$Y_0 = \theta_0 + \theta_1 x_{01} + \ldots, + \theta_p x_{0p} + \varepsilon_0, \quad \varepsilon_0 \sim \mathcal{N}(0; \sigma^2).$$

Sous forme vectorielle, on écrit :

$$Y_0 = {}^t x_{0.}\theta + \varepsilon_0, \quad {}^t x_{0.} = {}^t(1, x_{01}, \ldots, x_{0p}).$$

La prévision $\hat{Y}_0$ de Y_0 est naturellement donnée par :

$$\hat{Y}_0 = \hat{\theta}_0 + \hat{\theta}_1 x_{01} + \ldots, + \hat{\theta}_p x_{0p} = {}^t x_{0.}\hat{\theta}$$

C'est une variable normale, indépendante de Y_0, avec :

$$\mathbb{E}(\hat{Y}_0) = \mathbb{E}(Y_0) = {}^t x_{0.}\theta, \quad Var(\hat{Y}_0) = \sigma^2 \, {}^t x_{0.}({}^t XX)^{-1} x_{0.}.$$

Ainsi, l'écart $Y_0 - \hat{Y}_0$ est une variable normale, centrée et de variance :

$$Var(Y_0 - \hat{Y}_0) = \sigma^2[1 + {}^t x_{0.}({}^t XX)^{-1} x_{0.}].$$

Elle est indépendante de $\hat{\sigma}^2$, d'où l'intervalle de confiance :

$$IC(Y_0; 1-\alpha) = \hat{Y}_0 \pm t_{n-(p+1);\alpha} \hat{\sigma} \sqrt{1 + {}^t x_{0.}({}^t XX)^{-1} x_{0.}},$$

avec le quantile $t_{n-(p+1);\alpha} = F^{-1}_{S_{n-(p+1)}}(1 - \alpha/2)$.

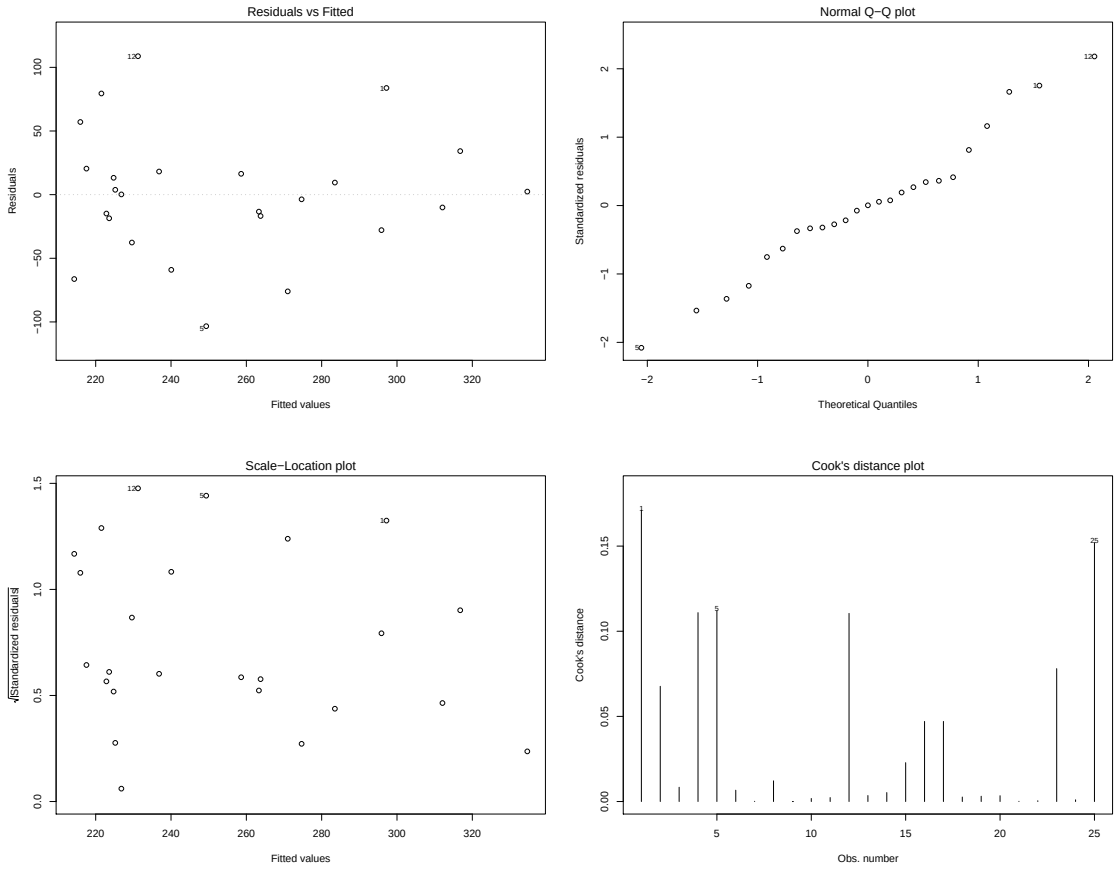

FIG. 7.5 – Inférence statistique sous R pour les salaires

7.6.2 Illustration

On considère un enseignant-chercheur ayant 10 ans d'ancienneté et 5 publications. La prévision du salaire, avec son intervalle de confiance à 95% sont obtenus par :

- round(predict(salaire,data.frame(x1=10,x2=5))) : 250 $
- round(predict(salaire,data.frame(x1=10,x2=5),interval="prediction")) : [139 ; 361]

En résumé On retiendra le principe de la régression linéaire multiple, sa mise en œuvre sous R, avec exploitation et interprétation des résultats.

TROISIÈME PARTIE

SÉRIES CHRONOLOGIQUES

Chapitre 8

GÉNÉRALITÉS

8.1 Introduction

Une série chronologique est constituée de l'ensemble des observations d'une grandeur effectuées à intervalles réguliers au cours du temps. Les exemples dans le monde économique et social sont donc nombreux : inflation, cours boursiers, chômage, productions, exportations, natalité, immigration, scolarisation, logement, etc. Ces grandeurs sont souvent mesurées à l'aide d'indices ou de taux publiés dans des revues spécialisées. Les exemples ne manquent pas non plus à l'intérieur même de l'entreprise, qu'elle soit à caractère industriel, commercial ou de services : chiffre d'affaires, stocks, ventes, prix, vie d'un produit, clientèle, etc. La plupart des disciplines scientifiques sont amenées à traiter des données temporelles : astronomie, météorologie, biologie, médecine, physique, etc.

La spécificité de l'analyse d'une série chronologique, qui la distingue d'autres analyses statistiques, est précisément dans l'importance accordée à l'ordre dans lequel sont effectuées les observations. Les méthodes statistiques classiques demandent souvent que les variables étudiées soient indépendantes et observées plusieurs fois (échantillon). Pour une série chronologique, la dépendance temporelle entre les variables constitue la source principale d'information. Celle-ci peut être entièrement contenue dans la valeur moyenne des variables, qui sont alors supposées indépendantes. Cependant l'ordre demeure essentiel, car cette moyenne représente l'évolution lente du phénomène (tendance) à laquelle s'ajoute parfois un effet périodique (mouvement saisonnier). Les contraintes sur la fonction moyenne sont nécessaires car chaque variable n'est observée qu'une seule fois.

Certaines données historiques célèbres sont utilisées par de nombreux auteurs afin de comparer les différentes approches dans l'analyse d'une série chronologique. C'est en particulier le cas de l'activité solaire depuis 1700 jusqu'à nos jours (*cf.* Figures 8.1 et 8.2), l'indice annuel du prix du blé en Europe de 1500 à 1869 construit par Beveridge (*cf.* Figure 8.3) et le nombre mensuel de passagers aériens internationaux aux Etats Unis de 1949 à 1960 (*cf.* Figure 8.4). On a également représenté un bruit blanc gaussien de variance unité (*cf.* Figure 8.5), c'est-à-dire une suite de variables aléatoires gaussiennes, centrées, de variance 1 et indépendantes. La marche aléatoire, obtenue en cumulant les valeurs de ce bruit, prend l'aspect d'un phénomène naturel (*cf.* Figure 8.6) alors qu'elle résulte uniquement du hasard.

On note $y_1, y_2, \ldots, y_T$ la séquence temporelle des données constituant la série chronologique étudiée. Sauf exception, il s'agira toujours d'une suite de valeurs numériques réelles. Ces valeurs sont considérées comme les observations d'une suite de variables aléatoires,

$$Y_t = g(t) + \varepsilon_t, \quad t = 1, \ldots, T,$$

dont la moyenne $g(t) = \mathbb{E}(Y_t)$ représente à elle seule la partie structurée de la grandeur observée. La séquence des erreurs $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_T$ est donc une suite de variables aléatoires centrées, non corrélées et de même variance σ^2 :

$$\mathbb{E}(\varepsilon_t) = 0, \quad Cov(\varepsilon_t, \varepsilon_s) = \mathbb{E}(\varepsilon_t \varepsilon_s) = \sigma^2 \delta_{ts}, \quad s, t = 1, \ldots, T,$$

où $\delta_{ts} = 1$ si $s = t, 0$ sinon, est le symbole de Kronecker. Cette hypothèse est en effet suffisante dans la mesure où les méthodes utilisées ne font intervenir que les deux premiers moments des variables, ce qui est souvent le cas en séries chronologiques. On ajoutera l'hypothèse gaussienne pour la construction d'intervalles de confiance ou de tests.

Ces hypothèses sont cependant encore trop générales pour espérer la moindre inférence statistique, sauf à estimer $g(t)$ par Y_t, puisque la moyenne se situe dans le même espace $\mathbb{R}^T$ que les observations. On introduit alors des contraintes sur la fonction $g(t)$. Celles-ci peuvent être de nature paramétrique et se traduisent par l'appartenance de $g(t), t = 1, \ldots, T$ à un sous espace vectoriel de $\mathbb{R}^T$. C'est le cadre de la régression linéaire. La description de ce sous espace permet de séparer la tendance des aspects périodiques dans la moyenne $g(t)$. L'approche non paramétrique consiste à lisser les observations, le plus souvent par le biais de moyennes mobiles, de façon à éliminer la partie résiduelle ε_t. Elle permet également de dégager les deux composantes de $g(t)$. Les deux approches peuvent être utilisées de façon complémentaire sur une

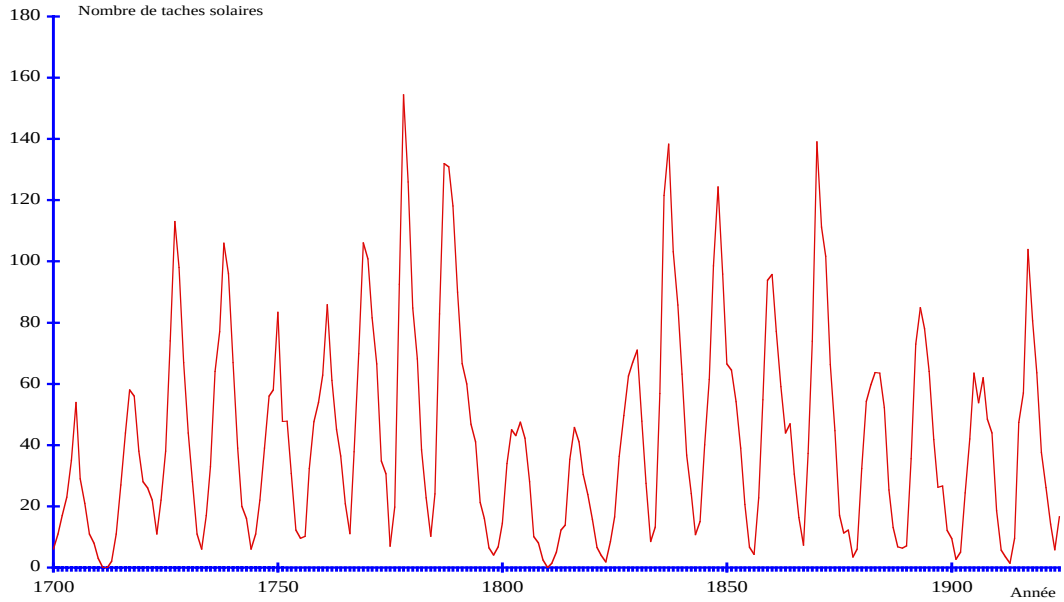

FIG. 8.1 – Nombre annuel de taches solaires selon Wolf de 1700 à 1924

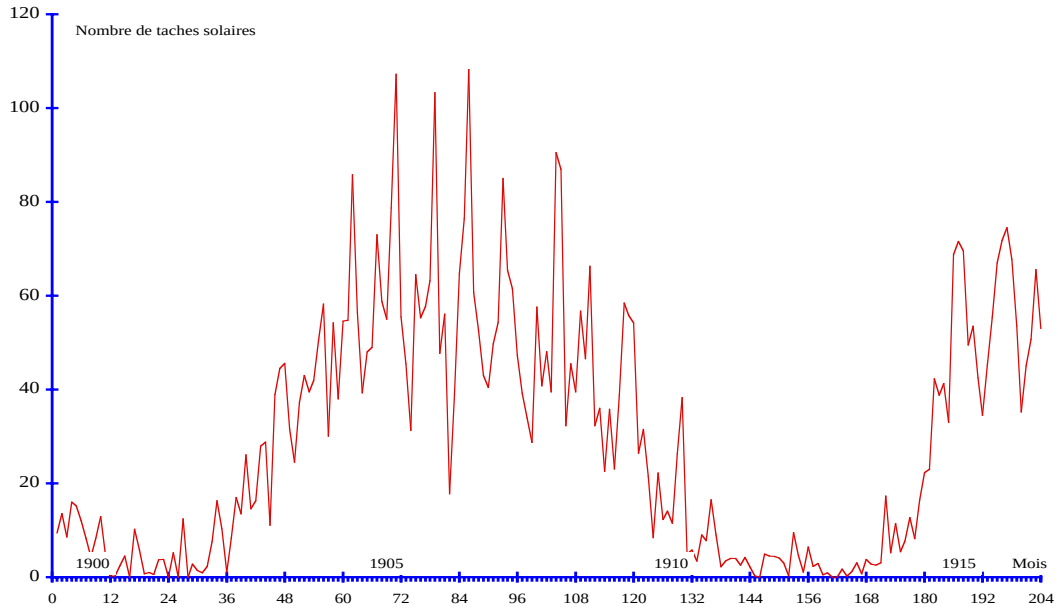

FIG. 8.2 – Nombre mensuel de taches solaires selon Wolf de 1900 à 1916

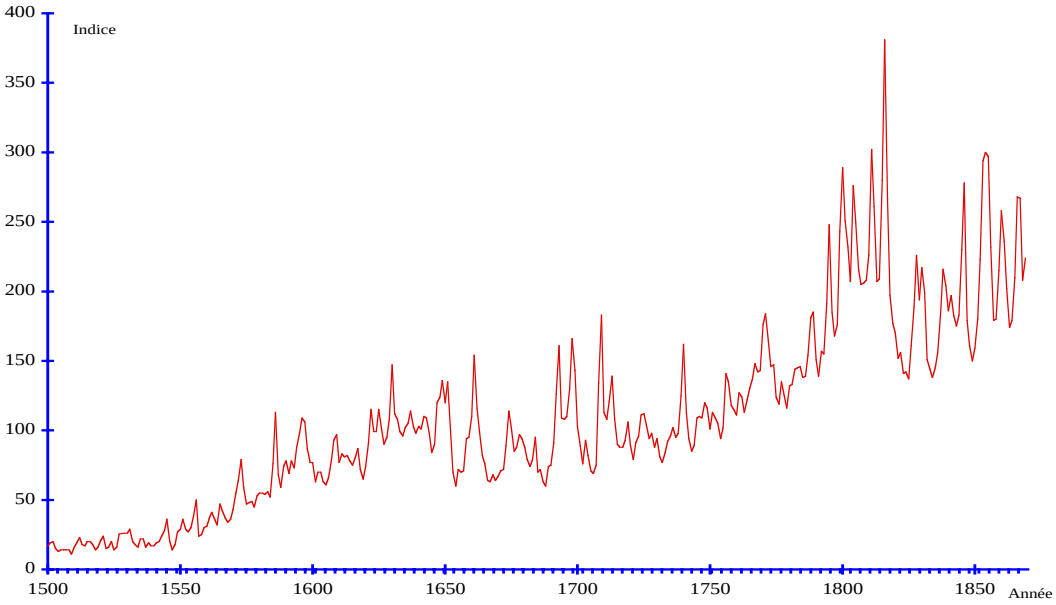

FIG. 8.3 – Indice annuel du prix du blé en Europe selon Beveridge de 1500 à 1869, base 100 : moyenne des années 1700 à 1745

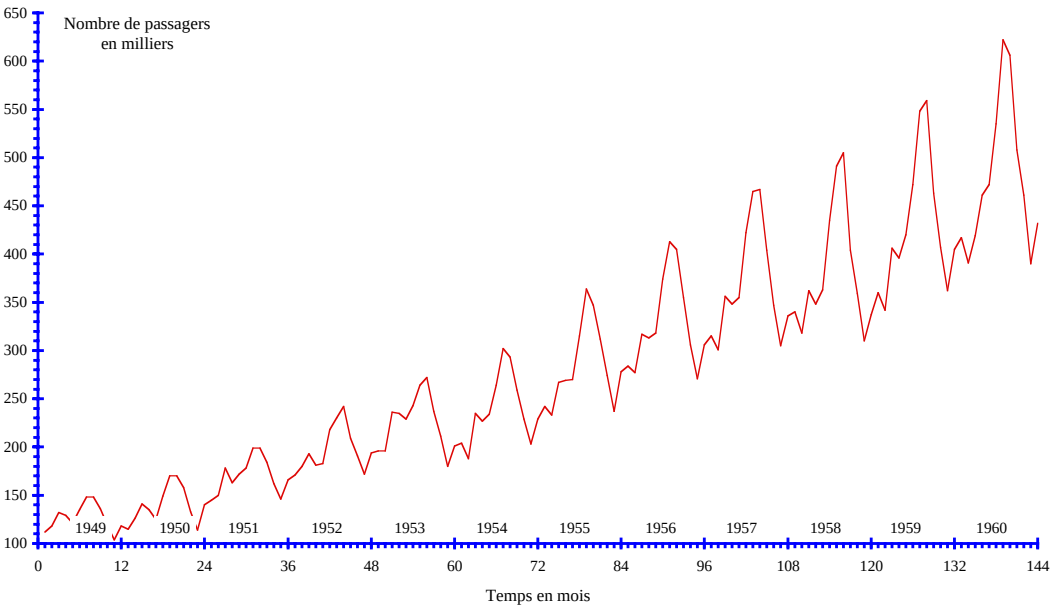

FIG. 8.4 – Nombre mensuel de passagers internationaux aux États Unis de 1949 à 1960

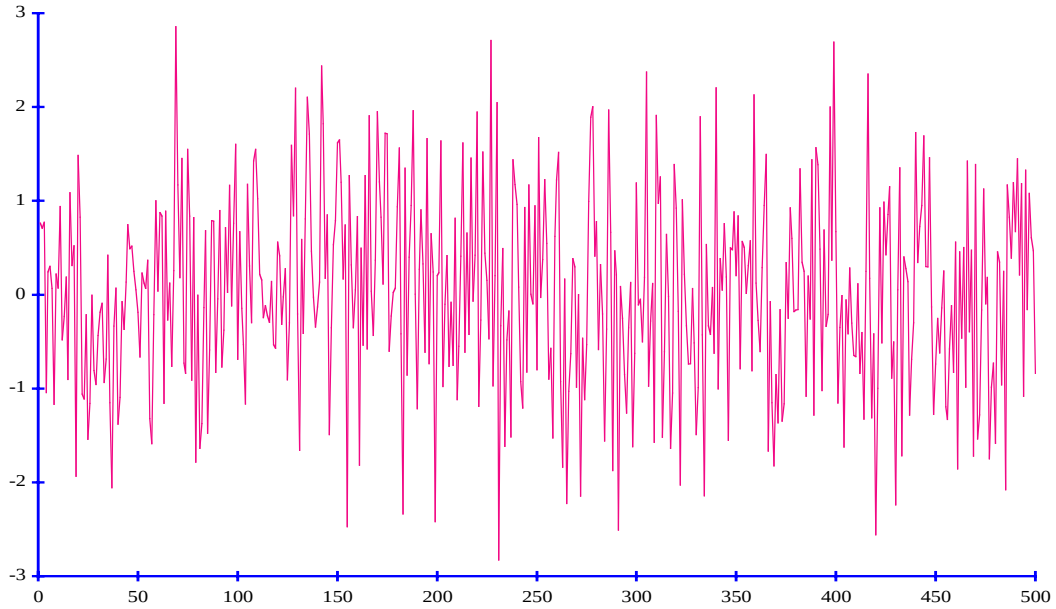

FIG. 8.5 – Bruit blanc gaussien de variance 1

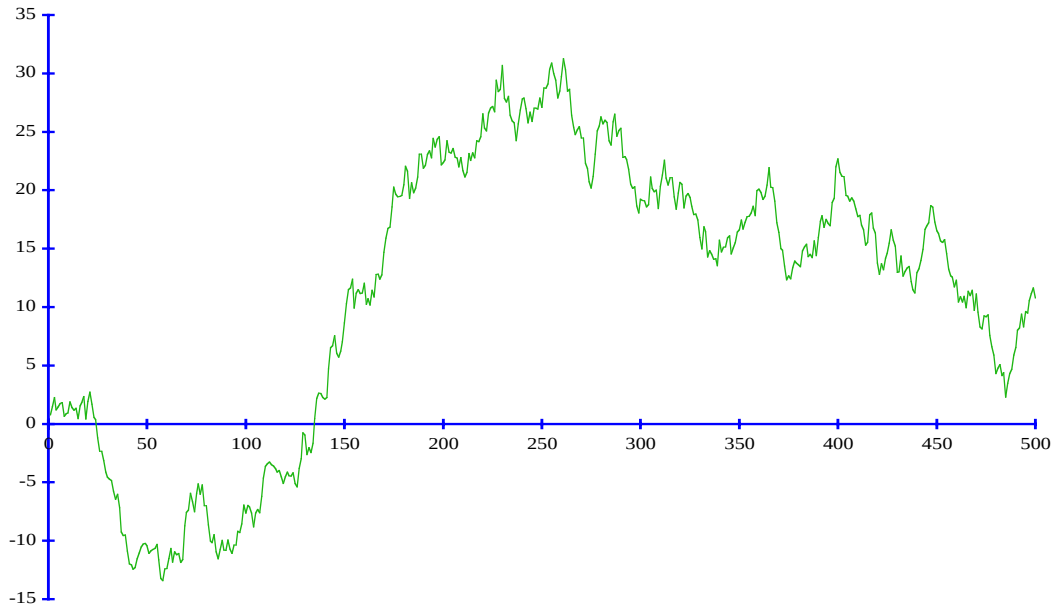

FIG. 8.6 – Marche aléatoire, valeurs cumulées du bruit blanc gaussien

même série, on parle de méthodes semi-paramétriques.

La suite de ce premier chapitre précise les différentes notions évoquées jusqu'ici : temps, tendance, effet saisonnier, composante résiduelle. Elle indique les premières démarches à effectuer sur une série brute avant de passer aux analyses statistiques développées dans les chapitres suivants. Le second chapitre est consacré aux méthodes basées sur la régression linéaire (modèle de Buys-Ballot). Il permet d'aborder, dans un cadre paramétrique, l'estimation de la tendance sous forme linéaire en présence ou non d'un effet saisonnier conduisant ainsi à des méthodes de prévision simples. Le dernier chapitre présente des méthodes de lissage (moyennes arithmétiques). La tendance et l'effet saisonnier sont estimés de façon non paramétrique par le biais de moyennes mobiles. L'objectif est de construire la série CVS (Corrigée des Variations Saisonnières), sans idée de prévision.

Le paragraphe suivant précise le rôle du temps dans la structure de la série. Les représentations graphiques mentionnées au second paragraphe constituent la première forme d'analyse destinée à éviter les erreurs grossières. Les composantes d'une série, tendance, effet saisonnier et erreurs ainsi que leur mode d'interaction font l'objet du dernier paragraphe.

8.2 Le temps

8.2.1 Définition d'une série chronologique

On appelle *série chronologique* (*série temporelle*, *chronique*) une suite d'observations numériques d'une grandeur effectuées à intervalles réguliers au cours du temps.

L'échelle de mesure et la variabilité de la grandeur sont telles que celle-ci sera toujours représentée par une variable continue à valeurs réelles. Le nombre mensuel de passagers aériens internationaux aux Etats Unis de 1949 à 1960, exprimé en milliers, varie entre 104 et 622 (*cf.* Figure 8.4). En général l'unité est choisie de sorte que la variable s'exprime avec 2 ou 3 chiffres significatifs.

La fréquence des observations peut être journalière, hebdomadaire, mensuelle, trimestrielle, annuelle ou autre. Dans bien des situations économiques, un effet saisonnier lié à une *période* connue est pressenti. Une chronique journalière sera observée pendant plusieurs semaines avec une périodicité de 5,

6 ou 7 jours selon le cas ; pour une chronique mensuelle (resp. trimestrielle) observée sur plusieurs années, la période est égale à 12 (resp. 4). Par la suite nous utiliserons systématiquement l'exemple mensuel dans nos commentaires.

La variable mesurée peut être l'état d'une grandeur à l'instant de mesure, on parle de *niveau* ou *stock*, ou le bilan d'une activité au cours de la dernière période écoulée à cet instant et on dit qu'il s'agit d'un *flux*. En météorologie la température est un niveau et la pluviométrie est un flux, de plus les relevés se font à heure fixe au cours du temps. En économie, l'indice des prix à la consommation est un niveau et le taux d'inflation correspondant est un flux ; dans ce cas l'indice fait référence au mois bien qu'il résulte de mesures pouvant être très étalées dans le temps. En finance un cours boursier est souvent considéré comme évoluant continûment, bien qu'on utilise sa valeur en ouverture de séance durant cinq jours par semaine pour déterminer les rentabilités journalières puis hebdomadaires ou mensuelles par sommation. L'analyse de ces trois séries, issues d'une même grandeur, est complémentaire et pas nécessairement redondante.

8.2.2 Quelques précautions élémentaires

Quel que soit le type de variable étudié, on s'assurera de respecter les points suivants.

Régularité des observations

Elle est parfaite dans le cas de certains relevés météorologiques, mais c'est déjà moins vrai pour beaucoup de variables économiques ou financières, puisque les mois ne comportent pas le même nombre de jours, en particulier de jours ouvrables. Une correction par simple proportionnalité peut être envisagée mais elle change la signification concrète des valeurs manipulées. En fait de légères entorses à la règle sont prises en compte dans la composante résiduelle ou dans la composante périodique lorsqu'elles sont systématiques (février, jours fériés de mai,...). Par contre une période de grève nécessite plus de précautions.

Stabilité des structures conditionnant le phénomène étudié

La plupart des chroniques étudiées concernent des grandeurs économiques et les techniques d'analyse cherchent à déterminer l'évolution lente du phénomène ainsi que ses variations saisonnières (pour une meilleure compréhension ou à des fins de prévision). Cela suppose une certaine stabilité qui, lorsqu'elle

n'est pas vérifiée, peut être obtenue en décomposant la chronique observée en plusieurs chroniques successives (empiriquement à l'aide d'une représentation graphique où à partir de la connaissance des modifications de l'environnement économique).

Permanence de la définition de la grandeur étudiée

Cette condition, qui paraît évidente, n'est parfois pas respectée. C'est en particulier le cas de certains indices économiques (changement du mode de calcul de l'indice).

Aspect périodique d'une partie de la grandeur observée

Cette condition est indispensable dans l'usage des techniques cherchant à déterminer des variations saisonnières. Elle suppose comparable deux observations relatives au même mois de deux années différentes. Elle n'exclut pas l'existence d'une évolution lente. Elle indique qu'une part du phénomène (la composante saisonnière) se répète de façon plus ou moins identique d'une année à l'autre. Dans ce cas il est souvent commode d'indexer la chronique à l'aide de deux indices : y_{ij} représente l'observation du j^e mois de la i^e année, et les données sont listées dans une table à double entrée.

8.2.3 Illustration

L'indice mensuel des prix à la consommation (INSEE) de 1970 à 1978 est donné dans le Tableau 8.1, le Tableau 8.2 indique le taux d'inflation correspondant et les graphes associés sont reproduits dans les Figures 8.7 et 8.8(*cf.* [GM90]).
Pour la période considérée, l'indice est de base 100 en juillet 1970. La méthode de calcul de cet indice est modifiée périodiquement pour tenir compte des changements de mode de consommation. Il est cependant préférable de maintenir durablement la définition de la grandeur étudiée plutôt que de l'adapter systématiquement aux modifications de l'environnement car alors la série n'a plus aucun sens. On remarque que l'indice I_t et le taux, $\tau_t = (I_t - I_{t-1})/I_{t-1}$, sont deux grandeurs qui se comportent très différemment bien que liées fonctionnellement. On distingue deux grandes périodes pour la variation de l'indice : le changement intervient à la fin de l'année 1973 et correspond à la première augmentation brutale du prix du pétrole. L'indice traduit l'évolution à moyen terme des prix. Cependant il est courant de considérer le taux d'inflation annuel, calculé sur les 12 derniers mois, pour mesurer cette évolution. Celle-ci fait apparaître quatre périodes : une croissance moyenne jusqu'à la

fin 1973, une forte croissance suivie d'une forte décroissance au cours des années 1974 et 1975 et une stabilité pour les années 1976 à 1978. La variation du taux d'inflation mensuel s'analyse différemment : on retrouve la croissance moyenne jusqu'à la fin 1973 où intervient une rupture, les années 1974 et 1975 se manifestent par une forte décroissance et la période de stabilité des années 1976 à 1978 est inchangée. L'analyse figurant dans [GM90] est différente : une période stable jusqu'au mois d'octobre 1973 avec un taux voisin de 0,4%, une croissance forte jusqu'au milieu de l'année 1974 suivie d'une décroissance pour atteindre une nouvelle période de stabilité à partir de 1976 avec un taux proche de 0,7%.

	mois	janv	fév	mars	avr	mai	juin	juil	août	sept	oct	nov	déc
	j	1	2	3	4	5	6	7	8	9	10	11	12
année	*i*												
1970	1	97,9	98,2	98,5	99,0	99,4	99,8	100,0	100,4	100,8	101,2	101,6	101,9
1971	2	102,5	103,0	103,4	104,0	104,7	105,1	105,6	106,0	106,5	107,1	107,5	108,0
1972	3	108,3	108,9	109,4	109,8	110,4	111,0	111,9	112,5	113,2	114,2	114,9	115,5
1973	4	115,5	115,8	116,4	117,2	118,3	119,2	120,2	121,0	122,1	123,4	124,5	125,3
1974	5	127,4	129,1	130,6	132,7	134,3	135,8	137,5	138,6	140,1	141,8	143,1	144,3
1975	6	145,9	147,0	148,2	149,5	150,6	151,7	152,8	153,8	155,1	156,3	157,3	158,2
1976	7	159,9	161,0	162,4	163,8	164,9	165,6	167,2	168,4	170,2	171,8	173,2	173,8
1977	8	174,3	175,5	177,1	179,4	181,1	182,5	184,1	185,1	186,7	188,2	188,9	189,4
1978	9	190,3	191,7	193,4	195,5	197,4	198,9	201,5	202,5	203,8	205,7	206,8	207,8

TAB. 8.1 – Indice mensuel des prix à la consommation, base 100 en juillet 1970, de 1970 à 1978

	mois	janv	fév	mars	avr	mai	juin	juil	août	sept	oct	nov	déc
	j	1	2	3	4	5	6	7	8	9	10	11	12
année	*i*												
1970	1		0,31	0,31	0,51	0,40	0,40	0,20	0,40	0,40	0,40	0,40	0,30
1971	2	0,59	0,49	0,39	0,58	0,67	0,38	0,48	0,38	0,47	0,56	0,37	0,47
1972	3	0,28	0,55	0,46	0,37	0,55	0,54	0,81	0,54	0,62	0,88	061	0,52
1973	4	0,00	0,26	0,52	0,69	0,94	0,76	0,84	0,67	0,91	1,06	0,89	0,64
1974	5	1,68	1,33	1,16	1,61	1,21	1,12	1,25	0,80	1,08	1,21	0,92	0,84
1975	6	1,11	0,75	0,82	0,88	0,74	0,73	0,73	0,65	0,85	0,77	0,64	0,57
1976	7	1,07	0,69	0,87	0,86	0,67	0,42	0,97	0,72	1,07	0,94	0,81	0,35
1977	8	0,29	0,69	0,91	1,30	0,95	0,77	0,88	0,54	0,86	0,80	0,37	0,26
1978	9	0,48	0,74	0,89	1,09	0,97	0,76	1,31	0,50	0,64	0,93	0,53	0,48

TAB. 8.2 – Taux mensuel des prix à la consommation de 1970 à 1978

8.3 Représentations graphiques

8.3.1 Représentation de la chronique

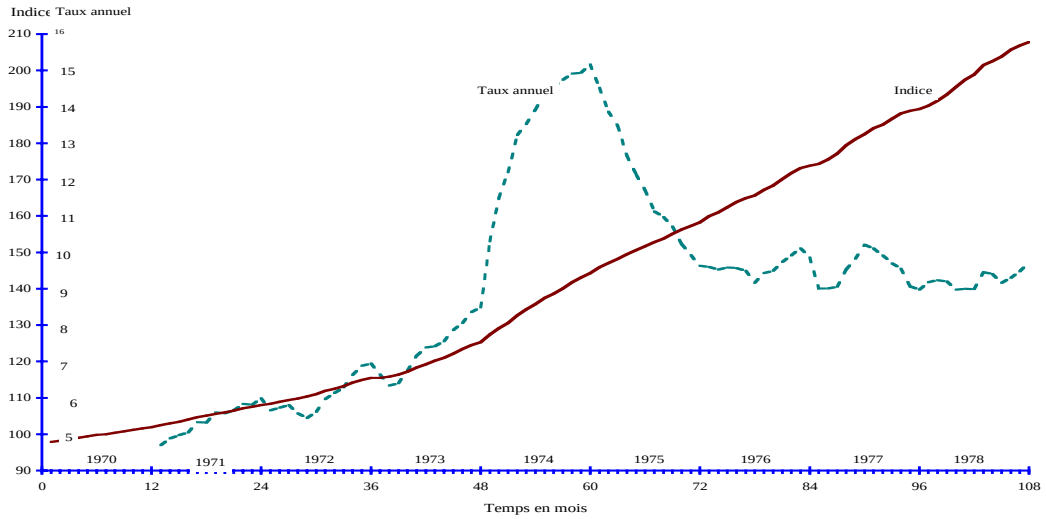

FIG. 8.7 – Indice mensuel des prix à la consommation, base 100 en juillet 1970, de 1970 à 1978

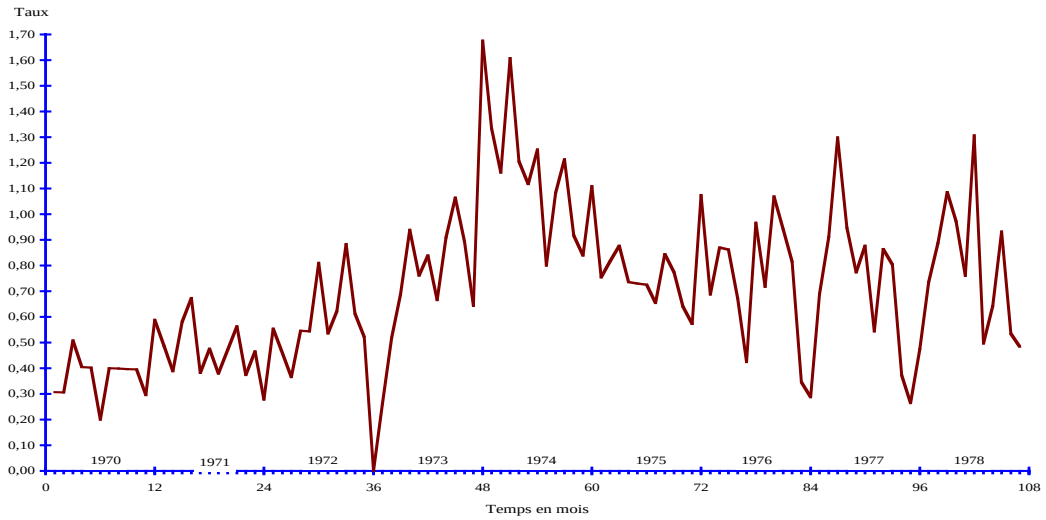

FIG. 8.8 – Taux d'inflation mensuel de 1970 à 1978

La représentation graphique des observations est une étape indispensable avant d'entreprendre une analyse plus technique de la chronique. Les points $(t, y_t), t = 1, \ldots, T$ sont représentés dans un système d'axes orthogonaux (échelles arithmétiques). Ils sont joints chronologiquement par des segments de droites pour faciliter la visualisation. Cette représentation permet d'apprécier l'évolution lente du phénomène (tendance), de dégager les périodes de stabilité. Elle suggère parfois d'opérer une transformation de la grandeur. C'est très nettement le cas pour le trafic aérien (*cf.* Figure 8.4) où la transformation logarithmique s'impose. Nous reviendrons sur ce point dans le prochain paragraphe.

L'interprétation de la tendance est délicate. Lorsqu'elle est naturellement liée au phénomène observé (trafic aérien, indice des prix,...), elle est reproductible dans un contexte similaire et sera considérée comme déterministe. Par contre une tendance apparente peut résulter du pur hasard comme le montre l'exemple de la marche aléatoire (*cf.* Figure 8.6). Dans ce cas elle est de nature stochastique et n'a pas d'interprétation autre que descriptive. L'analyse d'une série chronologique ne doit pas se faire au vu de ses seules valeurs numériques mais doit prendre en compte le contexte des observations. Cette représentation graphique est également utile pour le choix d'un modèle. L'aspect graphique est un indicateur sommaire permettant d'opérer un premier tri.

8.3.2 Représentation du mouvement saisonnier

La représentation de la chronique peut faire apparaître, en plus de la tendance, un aspect périodique plus ou moins marqué de période connue, 12 pour une chronique mensuelle, 4 pour une chronique trimestrielle. Il est très net dans le cas du trafic aérien (*cf.* Figure 8.4). L'évolution de la production industrielle française, considérée au Chapitre 10, présente un mouvement saisonnier exceptionnel directement observable à la lecture des données. En général l'effet saisonnier est moins spectaculaire. Une première appréciation de son importance est obtenue graphiquement en représentant, pour une chronique mensuelle, les courbes annuelles (sur 14 mois). La représentation polaire de la chronique a le même objectif, elle est plus originale mais moins lisible. L'étude quantitative du mouvement saisonnier est traitée dans les deux prochains chapitres.

8.3.3 Illustration

Nous avons repris le taux d'inflation mensuel introduit plus haut en ne considérant que les quatre dernières années. Les différentes représentations graphiques sont données dans les Figures 8.9, 8.10 et 8.11. Même en faisant abstraction de l'année 1975, qui n'appartient pas à la période de stabilité finale, l'effet saisonnier semble non négligeable mais est mal stabilisé dans le temps. L'approche numérique devrait permettre de préciser cette impression.

8.4 Les modèles pour la moyenne

8.4.1 Les composantes du modèle

Nous avons déjà évoqué à plusieurs reprises les notions de tendance, effet saisonnier et composante résiduelle. On distingue en effet généralement trois *composantes* dans une chronique $Y_t, t = 1, \ldots, T$.

Tendance

La *composante fondamentale* ou *tendance* (trend) traduit l'évolution à moyen terme du phénomène. On parle aussi de mouvement conjoncturel ou mouvement extra-saisonnier. La chronique correspondante, notée $f_t, t = 1, \ldots, T$, est une fonction à variation lente supposée déterministe dans cette approche. Elle sera estimée sous forme paramétrique (polynôme, exponentielle,...) ou comme le résultat d'une opération de lissage.

La composante saisonnière

La *composante saisonnière* ou *mouvement saisonnier* représente des effets périodiques de *période* connue p qui se reproduisent de façon plus ou moins identique d'une période sur l'autre. La chronique correspondante, également déterministe, est notée $S_t, t = 1, \ldots, T$. Elle est généralement supposée rigoureusement périodique : $S_{t+p} = S_t$ et les valeurs $S_j = S_{ij}, j = 1, \ldots, p$ d'une période sont appelées *coefficients saisonniers*. Le bilan de l'effet saisonnier sur une période doit être nul car il est pris en compte dans la tendance. La composante saisonnière permet simplement de distinguer, à l'intérieur d'une même période, une répartition stable dans le temps d'effets positifs ou négatifs qui se compensent sur l'ensemble de la période.

La composante résiduelle

La *composante résiduelle* ou *variations accidentelles* est la partie non structurée du phénomène. Elle est modélisée par une suite de variables aléa-

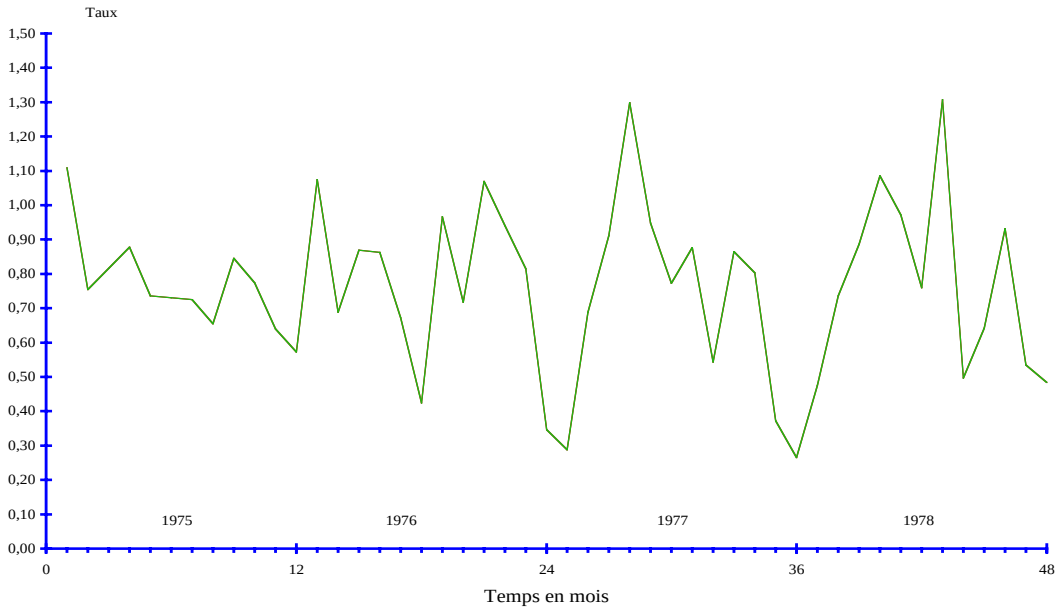

FIG. 8.9 – Taux d'inflation mensuel de 1975 à 1978

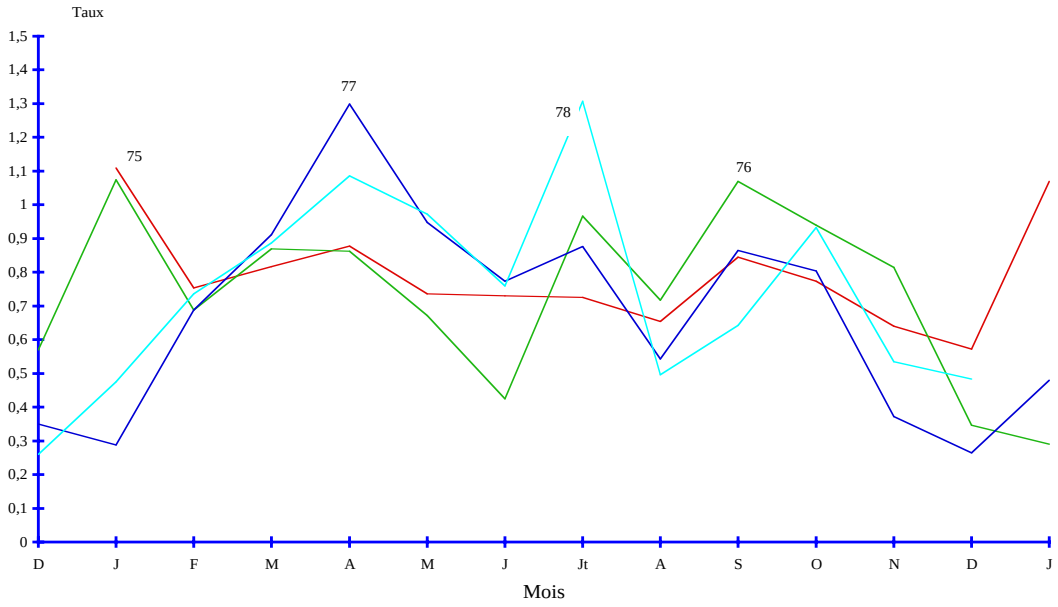

FIG. 8.10 – Mouvement saisonnier du taux d'inflation mensuel de 1975 à 1978

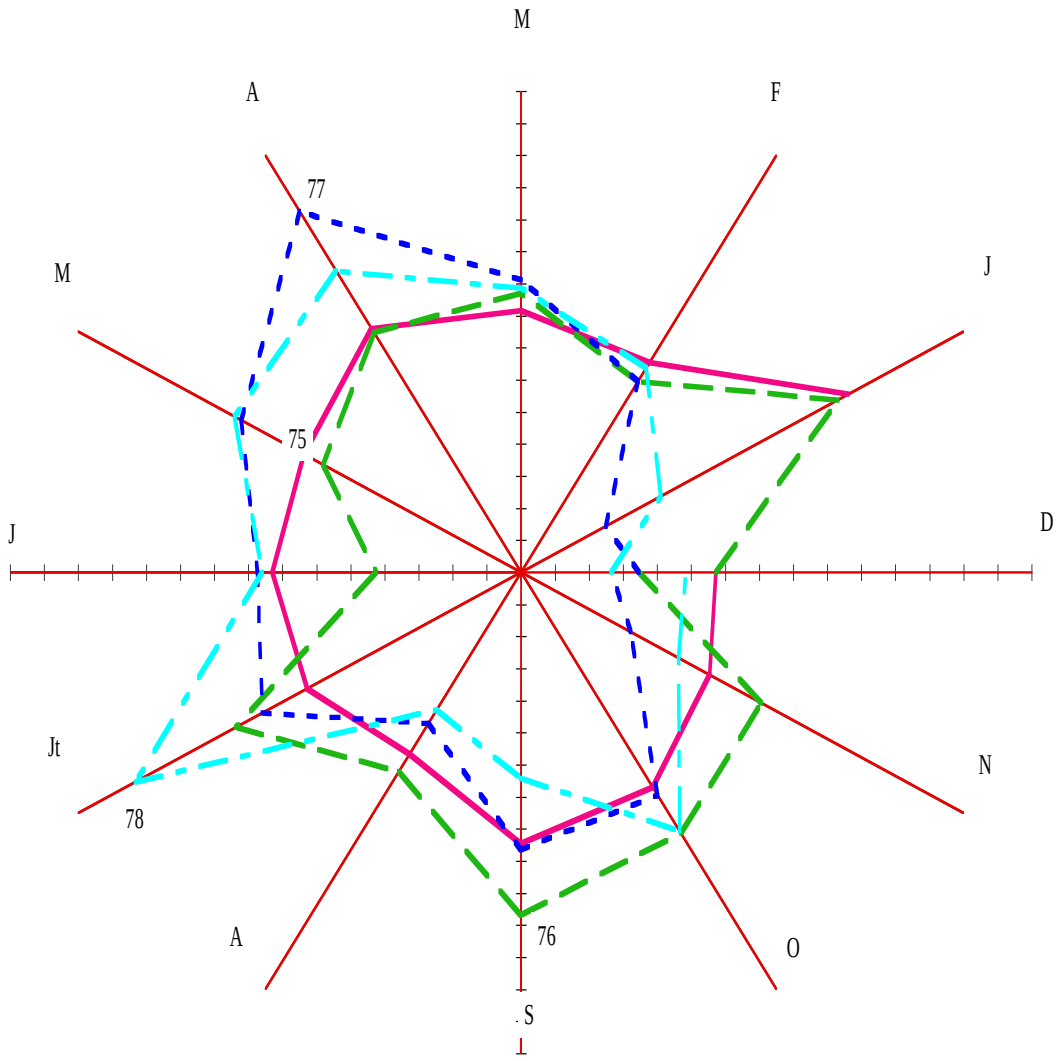

FIG. 8.11 – Représentation polaire du mouvement saisonnier du taux d'inflation mensuel de 1975 à 1978 (échelle polaire : pôle à 0%, graduation de 0,1%)

toires $\varepsilon_t, t = 1, \ldots, T$, centrées, non corrélées et de même variance, on parle de *bruit blanc*.

Certains phénomènes étudiés à très long terme présentent une *composante cyclique* (cycles d'activité) dont la période, de plusieurs années, est souvent mal définie. Un exemple célèbre est celui de l'activité solaire (*cf.* Figure 8.2), dont le cycle est d'environ 11 ans. Cette composante est prise en compte dans la tendance sur les séries de taille moyenne et ne sera pas étudiée en tant que telle ici.

On comprend aisément l'intérêt de connaître la tendance. La composante saisonnière sert également à analyser le phénomène étudié et participe à la volonté de prévision. Cependant bien des séries économiques sont publiées en données corrigées des variations saisonnières (*série CVS*). De telles séries, dites désaisonnalisées, sont obtenues en éliminant la composante saisonnière de la série initiale. À cette fin la composante saisonnière est estimée en général de façon non paramétrique. La série CVS gomme l'effet saisonnier en le répartissant de façon uniforme sur toute la période et conserve ainsi la grandeur observée dans son ensemble. En particulier elle contient la composante résiduelle et ne doit pas être confondue avec la tendance. Elle permet de comparer directement deux valeurs consécutives. Le chômage peut augmenter d'un mois sur l'autre en données brutes alors qu'il baisse en données corrigées des variations saisonnières.

On suppose généralement que la partie résiduelle a un ordre de grandeur très inférieur à celui des parties explicatives f_t et S_t. En fait il sera toujours possible de déceler la présence d'une partie déterministe, même de faible importance, à condition que la série observée soit suffisamment longue.

8.4.2 Les schémas de composition

Pour pouvoir séparer les trois composantes servant à décrire la série observée, il est nécessaire de préciser leur mode d'interaction. La plupart des séries chronologiques entrent dans l'un des schémas suivants :

- *Schéma additif* : $Y_t = f_t + S_t + \varepsilon_t$,
- *Schéma multiplicatif* : $Y_t = f_t \times S_t \times (1 + \varepsilon_t)$,
- *Schéma mixte* : $Y_t = f_t \times S_t + \varepsilon_t$.

En utilisant $(1+\varepsilon_t)$ dans le cas multiplicatif, on conserve la même signification et les mêmes propriétés à chacune des trois composantes f_t, S_t et ε_t dans les trois schémas de composition. Cependant il est nécessaire de supposer que $(1+\varepsilon_t)$ reste positif dans le modèle multiplicatif car la composante résiduelle

ne peut être responsable du signe de la grandeur observée. Notons que le shéma multiplicatif (pour une variable positive) se réduit au shéma additif par transformation logarithmique :

$$\tilde{Y}_t = ln(Y_t) = ln(f_t) + ln(S_t) + ln(1 + \varepsilon_t) = \tilde{f}_t + \tilde{S}_t + \tilde{\varepsilon}_t.$$

Les trois composantes $\tilde{f}_t, \tilde{S}_t$ et $\tilde{\varepsilon}_t$ conservent en effet la même signification : variation lente, périodicité et erreur. Dans les schémas multiplicatif et mixte les oscillations dues à l'effet saisonnier ont une amplitude proportionnelle à la valeur de la tendance. C'est précisément l'argument utilisé pour faire le choix entre le schéma additif et les deux autres schémas au vu de la représentation graphique de la chronique. La distinction entre le schéma multiplicatif et le schéma mixte peut également s'apprécier graphiquement selon le même principe. Elle peut aussi relever de considérations sur l'origine des erreurs : une erreur structurelle (de modélisation) a des chances d'être proportionnelle à la grandeur étudiée alors qu'une erreur de mesure pourrait ne pas en dépendre. Le trafic aérien aux Etats Unis entre 1949 et 1960 (*cf.* Figure 8.4) fournit un exemple où le schéma multiplicatif s'impose. Nous avons déjà remarqué que dans ce cas il fallait considérer le logarithme de la variable qui est alors expliqué par un schéma additif. Par la suite nous ne considérerons plus que le cas additif. Il peut être nécessaire d'appliquer plusieurs fois la transformation logarithmique. Notons que la relation entre l'indice des prix et le taux d'inflation correspondant, $I_t = I_{t-1}(1 + \tau_t)$, invite à considérer la variable $\ln(1 + \tau_t) = \ln I_t - \ln I_{t-1}$ plutôt que τ_t. La différence est négligeable pour des taux faibles. Les rentabilités d'un cours boursier sont définies de la même façon que le taux d'inflation. Le schéma mixte ne peut pas être transformé en shéma additif.

En résumé On retiendra l'intérêt de la représentation graphique d'une chronique (stabilité, tendance, choix de modèle, etc) et de celle de son mouvement saisonnier.

Chapitre 9

MODÈLE DE BUYS-BALLOT ET PRÉVISION

Le modèle de *Buys-Ballot* fournit un schéma additif simple que l'on peut traiter très complètement par des méthodes élémentaires. La tendance est représentée par une droite, l'effet saisonnier est rigoureusement périodique de période p connue et la partie résiduelle est une suite de variables aléatoires indépendantes identiquement distribuées de loi normale centrée et de variance σ^2 :

$$Y_t = \alpha t + \beta + S_t + \varepsilon_t, t = 1, \ldots, T; \quad S_t = S_{t+p}; \quad \varepsilon_t \sim i.i.d.\mathcal{N}(0; \sigma^2).$$

9.1 Aspectcs descriptifs

Les représentations graphiques de la série et de son mouvement saisonnier donnent une première idée de la tendance et de l'effet saisonnier. La méthode des moindres carrés est utilisée pour estimer la tendance et les coefficients saisonniers. Ceci permet d'obtenir la série ajustée qui, otée à la série initiale, fournit les résidus. La représentation de ces résidus est un premier indicateur de la validité du modèle qu'il faut considérer avant d'effectuer une prévision.

9.1.1 Estimations des moindres carrés

La simplicité des calculs est obtenue en supposant que la série est observée pendant n "années" de p "mois" :

$$Y_{ij} = \alpha[p(i-1)+j] + \beta + S_j + \varepsilon_{ij}; \quad j = 1, \ldots, p; i = 1, \ldots, n; \quad \sum_{j=1}^{p} S_j = 0.$$

Les coefficients saisonniers $S_j, j = 1, \ldots, p$, caractérisent la composante périodique S_t dont l'effet annuel moyen est nul. Pour mener à bien les calculs, on effectue un changement de variables :

$$\beta_j = \beta + S_j, j = 1, \ldots, p \Longleftrightarrow \beta = \frac{1}{p}\sum_{j=1}^{p} \beta_j, S_j = \beta_j - \beta, j = 1, \ldots, p.$$

Ainsi la partie déterministe du modèle est décrite par $p + 1$ paramètres linéairement indépendants : $\alpha, \beta_1, \ldots, \beta_p$. La *méthode des moindres carrés* consiste à chercher, parmi les chroniques $x_{ij}(a, b_1, \ldots, b_p) = a[p(i-1)+j]+b_j$, composées d'une tendance linéaire et d'un mouvement saisonnier périodique, celle qui est la plus proche de l'observation selon le critère :

$$\min_{a,b_1,\ldots,b_p} \frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{p}[y_{ij} - x_{ij}(a, b_1, \ldots, b_p)]^2.$$

En d'autres termes, elle retient les paramètres pour lesquels la moyenne des carrés des erreurs observées est minimum.

Notons plus simplement $x_{ij} = x_{ij}(a, b_1, \ldots, b_p)$ et introduisons les moyennes mensuelles :

$$\overline{y}_{.j} = \frac{1}{n}\sum_{i=1}^{n} y_{ij}, \quad \overline{x}_{.j} = \frac{1}{n}\sum_{i=1}^{n} x_{ij} = a[p(n-1)/2 + j] + b_j, \quad j = 1, \ldots, p.$$

Le critère à minimiser se scinde en deux parties :

$$\begin{aligned} \frac{1}{np}\sum_{i,j}[y_{ij} - x_{ij}]^2 &= \frac{1}{p}\sum_{j}\{\overline{y}_{.j} - a[p(n-1)/2 + j] - b_j\}^2 \\ + \quad \frac{1}{np}\sum_{i,j}[y_{ij} - \overline{y}_{.j}]^2 &+ \frac{1}{np}\sum_{i,j}[x_{ij} - \overline{x}_{.j}]^2 - \frac{2}{np}\sum_{i,j}[y_{ij} - \overline{y}_{.j}][x_{ij} - \overline{x}_{.j}]. \end{aligned}$$

Ainsi, pour une pente a fixée quelconque, le minimum par rapport à b_j est réalisé en annulant le premier terme et équivaut à écrire que les chroniques y_{ij} et x_{ij} ont mêmes moyennes mensuelles :

$$b_j = \overline{y}_{.j} - a[p(n-1)/2 + j] \quad \Longleftrightarrow \quad \overline{x}_{.j} = \overline{y}_{.j}, \quad j = 1, \ldots, p.$$

Soient $t_{ij} = p(i-1) + j$ les dates d'observation et introduisons les moyennes annuelles ainsi que les moyennes globales :

$$\overline{y}_{i.} = \frac{1}{p}\sum_{j=1}^{p} y_{ij}, \quad \overline{t}_{i.} = \frac{1}{p}\sum_{j=1}^{p} t_{ij} = p(i-1) + (p+1)/2, \quad i = 1, \ldots, n,$$

$$\overline{y}_{..} = \frac{1}{np}\sum_{i,j} y_{ij} = \frac{1}{n}\sum_{i} \overline{y}_{i.} = \frac{1}{p}\sum_{j} \overline{y}_{.j}, \quad \overline{t}_{..} = \frac{1}{np}\sum_{i,j} t_{ij} = (np+1)/2.$$

Tenant compte de la solution obtenue pour les variables b_j et de la relation $x_{ij} - \overline{x}_{.j} = a(\overline{t}_{i.} - \overline{t}_{..})$, il reste à minimiser par rapport à a la deuxième partie du critère :

$$\frac{1}{np}\sum_{i,j}[y_{ij} - \overline{y}_{.j}]^2 + a^2 var(\overline{t}_{i.}) - 2a cov(\overline{t}_{i.}, \overline{y}_{i.}),$$

où l'on a posé ($\sum_{i=1}^{n} i^2 = n(n+1)2n+1)/6$) :

$$var(\overline{t}_{i.}) = \frac{1}{n}\sum_{i}[\overline{t}_{i.} - \overline{t}_{..}]^2 = \frac{p^2(n^2-1)}{12}, \quad cov(\overline{t}_{i.}, \overline{y}_{i.}) = \frac{1}{n}\sum_{i}[\overline{t}_{i.} - \overline{t}_{..}][\overline{y}_{i.} - \overline{y}_{..}].$$

L'estimation de la pente α de la tendance est alors :

$$\hat{\alpha} = \frac{cov(\overline{t}_{i.}, \overline{y}_{i.})}{var(\overline{t}_{i.})} = \frac{12}{np(n^2-1)}[\sum_{i=1}^{n} i\overline{y}_{i.} - \frac{n(n+1)}{2}\overline{y}_{..}].$$

Son report dans la solution pour b_j donne les estimations des paramètres β_j :

$$\hat{\beta}_j = \overline{y}_{.j} - \hat{\alpha}[p(n-1)/2 + j], \quad j = 1, \ldots, p,$$

qui, par centrage, fournissent les estimations de l'ordonnée à l'origine β de la tendance ainsi que celles des coefficients saisonniers S_j :

$$\hat{\beta} = \frac{1}{p}\sum_{j=1}^{p} \hat{\beta}_j = \overline{y}_{..} - \hat{\alpha}\overline{t}_{..} = \overline{y}_{..} - \hat{\alpha}(np+1)/2,$$

$$\hat{S}_j = \hat{\beta}_j - \hat{\beta} = \overline{y}_{.j} - \overline{y}_{..} - \hat{\alpha}[j - (p+1)/2], \quad j = 1, \ldots, p.$$

En résumé on observe les résultats suivants :

- La tendance ne dépend que des moyennes annuelles, elle est la droite des moindres carrés construite sur les points $(\overline{t}_{i.}, \overline{y}_{i.}), i = 1, \ldots, n$, c'est-à-dire que $\hat{\alpha}$ et $\hat{\beta}$ sont solution du problème de minimisation :

$$\min_{a,b} \frac{1}{n}\sum_{i=1}^{n}[\overline{y}_{i.} - a\overline{t}_{i.} - b]^2.$$

- La composante saisonnière est définie par les moyennes mensuelles de la chronique privée de la tendance estimée :

$$\hat{S}_j = \frac{1}{n}\sum_{i=1}^{n}[y_{ij} - \hat{\alpha}[p(i-1) + j] - \hat{\beta}], \quad j = 1, \ldots, p.$$

9.1.2 Série ajustée, résidus et prévision

La *série ajustée* est la série la plus proche de la série observée au sens du critère des moindres carrés :

$$\hat{y}_{ij} = \hat{\alpha}[p(i-1)+j] + \hat{\beta} + \hat{S}_j; \quad j = 1, \dots p; i = 1, \dots, n.$$

La représentation de la chronique des *résidus*,

$$\hat{\varepsilon}_{ij} = y_{ij} - \hat{y}_{ij}; \quad j = 1, \dots p; i = 1, \dots, n,$$

permet d'évaluer graphiquement la validité du modèle. Lorsque le modèle est ainsi justifié, il peut être utilisé pour effectuer une *prévison* pour la période suivante en utilisant l'expression de la série ajustée avec $i = n + 1$:

$$\hat{y}_{n+1,j} = \hat{\alpha}[pn + j] + \hat{\beta} + \hat{S}_j, \quad j = 1, \dots p.$$

9.1.3 Illustration

Nous illustrons cette méthode sur la chronique mensuelle du chiffre d'affaires de la presse parisienne dans une petite ville de province ($\simeq$ 8000 h.) de 1981 à 1985. Les données, leurs différentes moyennes et les résultats sont présentés habituellement dans la *Table de Buys-Ballot*. Nous avons également fait figurer les valeurs de la *série ajustée* ainsi que celles de la *prévision* pour l'année 1986 (*cf.* Tableau 9.1).

La représentation de la chronique dans la Figure 9.1 montre une très nette tendance à la croissance.

La représentation du mouvement saisonnier dans la Figure 9.2 fait apparaître des pics d'activité aux mois de janvier, mars et surtout octobre. Par contre, on ne constate pas de creux d'activité vraiment marqués. Les valeurs numériques des coefficients saisonniers du Tableau 9.1 confirment le pic d'octobre et, dans une moindre mesure, ceux de janvier et mars en y ajoutant septembre. Ils indiquent aussi un creux en mai et novembre, voire avril.

Les résidus représentés dans la Figure 9.3 ont une forme incurvée invitant à ajuster une tendance parabolique. Ceci sort du cadre de ce cours.

En acceptant le modèle, malgré les réserves précédentes, nous avons représenté, dans la Figure 9.4, la série initiale, la série ajustée avec la tendance associée et la prévision pour l'année 1986. On constate que la série observée est au dessous de la série ajustée aux deux extrémités alors qu'elle est au

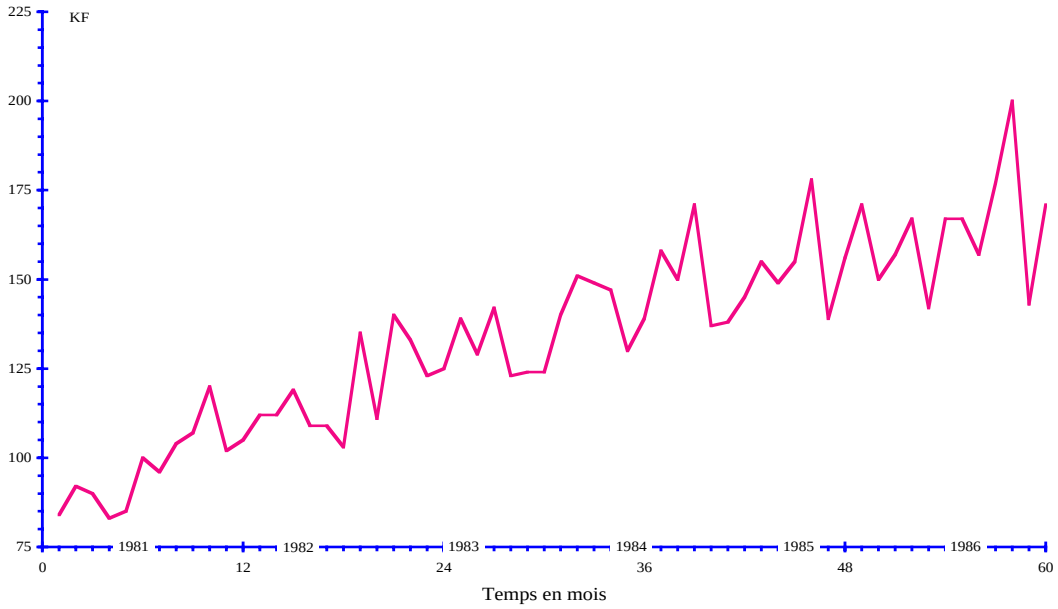

FIG. 9.1 – Chiffre d'affaires mensuel de la presse parisienne dans une petite ville de province : représentation de la chronique

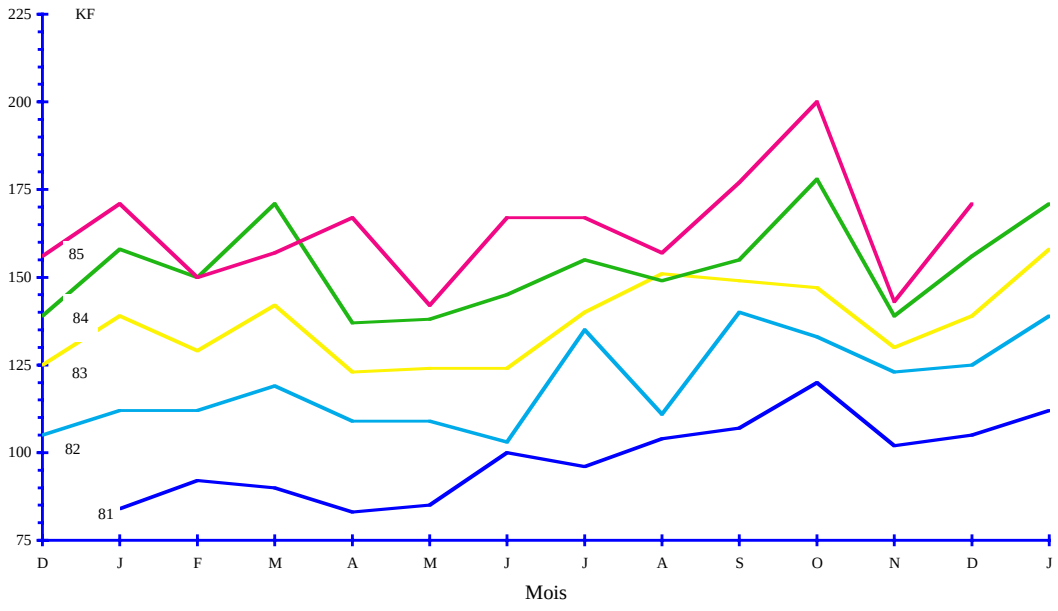

FIG. 9.2 – Mouvement saisonnier du chiffre d'affaires mensuel de la presse parisienne dans une petite ville de province

FIG. 9.3 – Chiffre d'affaires mensuel de la presse parisienne dans une petite ville de province : représentation des résidus

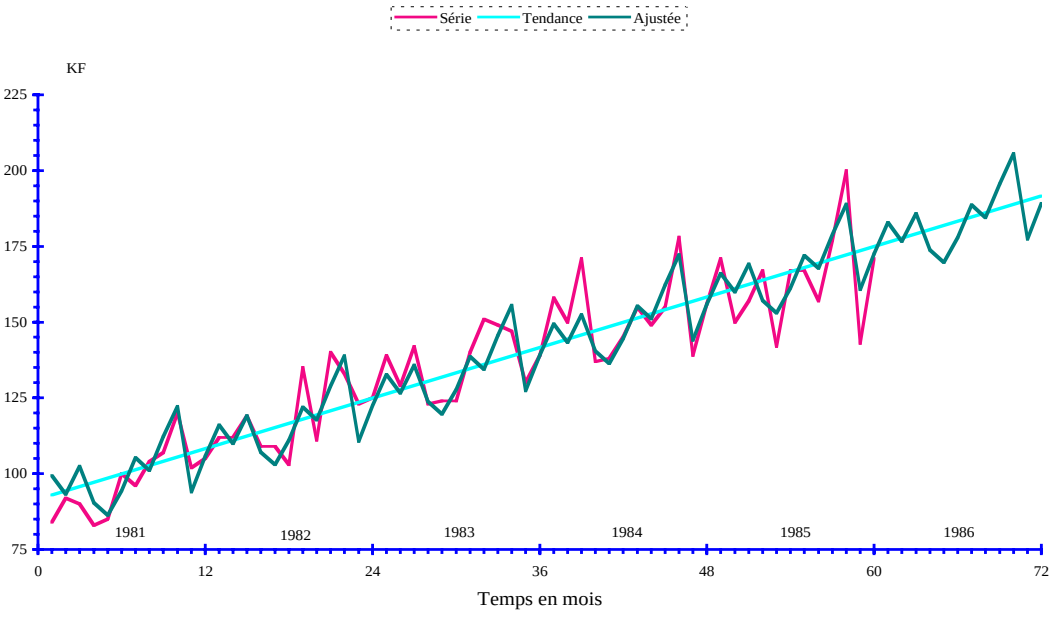

FIG. 9.4 – Chiffre d'affaires mensuel de la presse parisienne dans une petite ville de province

	mois	janv	fév	mars	avr	mai	juin	juil	août	sept	oct	nov	déc	
	j	1	2	3	4	5	6	7	8	9	10	11	12	
année	*i*													moy. an.
1981	1	84	92	90	83	85	100	96	104	107	120	102	105	97
	ajustée	99	93	102	90	86	94	105	101	112	122	94	106	
1982	2	112	112	119	109	109	103	135	111	140	133	123	125	119
	ajustée	116	110	119	107	103	111	122	118	129	139	111	123	
1983	3	139	129	142	123	124	124	140	151	149	147	130	139	136
	ajustée	133	127	136	124	120	128	139	134	146	156	127	139	
1984	4	158	150	171	137	138	145	155	149	155	178	139	156	153
	ajustée	149	143	152	140	136	144	155	151	162	172	144	156	
1985	5	171	150	157	167	142	167	167	157	177	200	143	171	164
	ajustée	166	160	169	157	153	161	172	168	179	189	161	173	
														moy. géné.
	moy. mens.	133	127	136	124	120	128	139	134	146	156	127	139	134
	coef. sais.	7	-1	7	-7	-12	-5	4	-2	8	17	-13	-2	
1986	prévues	183	177	186	174	170	178	189	184	196	206	177	189	

Tendance : pente = 1,39 ; ordonnée à l'origine = 92

TAB. 9.1 – Chiffre d'affaires mensuel de la presse parisienne dans une petite ville de province (unité : 1KF) : table de Buys-Ballot

dessus dans la partie centrale. Ceci est cohérent avec le constat effectué sur la représentation des résidus. Notons aussi que l'ordre de grandeur du mouvement saisonnier est très faible par rapport aux données (environ 10%) bien que la concordance entre les pics et les creux des deux séries soit assez bien respectée. L'étude du mouvement saisonnier dans le cadre de ce modèle (*cf.* Paragraphe 9.2.2) montre, grâce à l'approche numérique, qu'il est effectivement présent avec un mois d'octobre fort et les mois de mai et novembre faibles alors que la représentation graphique n'est pas aussi nette sur ce point.

9.2 Aspects inductifs

Dans l'approche descriptive, seule l'hypothèse de bruit blanc pour les erreurs ε_t est utilisée pour justifier la méthode des moindres carrés. Cette hypothèse peut être infirmée par la représentation des résidus $\hat{\varepsilon}_t$. L'approche inductive permet d'aller plus loin dans la validation du modèle et dans son utilisation.

9.2.1 Moyenne et variance des estimateurs

Les propriétés des estimateurs obtenus à la section précédente sont analogues à celles que nous avons obtenues dans le cadre de la régression linéaire simple. Ils sont sans biais et de variance minimum parmi les estimateurs sans biais qui sont linéaires en les observations. Nous sommes ici dans un

cadre de régression linéaire multiple. Cependant la particularité du modèle de Buys-Ballot permet de donner des expressions explicites de leurs variances.

Pour les paramètres de la tendance, on utilise le modèle de régression linéaire simple :

$$\overline{Y}_{i.} = \alpha \bar{t}_{i.} + \beta + \bar{\varepsilon}_{i.}, \quad i = 1, \ldots, n; \quad \bar{\varepsilon}_{i.} = \frac{1}{p} \sum_{j=1}^{p} \varepsilon_{ij} \sim i.i.d. \mathcal{N}(0; \sigma^2/p).$$

L'estimateur de la pente s'écrit :

$$\hat{\alpha} = \frac{1}{var(\bar{t}_{i.})} \frac{1}{n} \sum_{i=1}^{n} [\bar{t}_{i.} - \bar{t}_{..}] \overline{Y}_{i.} = \alpha + \frac{1}{var(\bar{t}_{i.})} \frac{1}{n} \sum_{i=1}^{n} [\bar{t}_{i.} - \bar{t}_{..}] \bar{\varepsilon}_{i.}$$

Il est sans biais et sa variance est donnée par :

$$Var(\hat{\alpha}) = \frac{\sigma^2}{p} \frac{1/n}{var(\bar{t}_{i.})} = \frac{\sigma^2}{np} \frac{12}{p^2(n^2-1)}.$$

L'estimateur de l'ordonnée à l'origine,

$$\hat{\beta} = \overline{Y}_{..} - \hat{\alpha}(np+1)/2 = \beta - (\hat{\alpha} - \alpha)\bar{t}_{..} + \bar{\varepsilon}_{..} = \beta + \frac{1}{n} \sum_{k=1}^{n} \left\{ 1 - \frac{[\bar{t}_{k.} - \bar{t}_{..}]\bar{t}_{..}}{var(\bar{t}_{i.})} \right\} \bar{\varepsilon}_{k.},$$

est également sans biais. En utilisant les expressions en fonction des $\bar{\varepsilon}_{i.}$, on obtient :

$$Var(\hat{\beta}) = \frac{\sigma^2}{np} \left[1 + \frac{3(np+1)^2}{p^2(n^2-1)} \right].$$

Pour les autres estimateurs, on introduit les erreurs mensuelles moyennes :

$$\bar{\varepsilon}_{.j} = \frac{1}{n} \sum_{i=1}^{n} \varepsilon_{ij} \sim i.i.d. \mathcal{N}(0, \sigma^2/n), \quad j = 1, \ldots, p,$$

qui satisfont :

$$Cov(\bar{\varepsilon}_{i.}, \bar{\varepsilon}_{.j}) = Cov(\bar{\varepsilon}_{..}, \bar{\varepsilon}_{.j}) = Cov(\bar{\varepsilon}_{..}, \bar{\varepsilon}_{i.}) = \frac{\sigma^2}{np}, \quad Cov(\bar{\varepsilon}_{.j} - \bar{\varepsilon}_{..}, \bar{\varepsilon}_{i.}) = 0.$$

Les estimateurs des coefficients saisonniers s'écrivent :

$$\hat{S}_j = S_j + (\bar{\varepsilon}_{.j} - \bar{\varepsilon}_{..}) - [j - (p+1)/2](\hat{\alpha} - \alpha), \quad j = 1, \ldots, p.$$

Ils sont donc sans biais. D'autre part $\overline{\varepsilon}_{.j} - \overline{\varepsilon}_{..}$ est non corrélé avec $\hat{\alpha}$, d'où :

$$Var(\hat{S}_j) = \frac{\sigma^2}{np}\left[(p-1) + \frac{12[j-(p+1)/2]^2}{p^2(n^2-1)}\right], \quad j = 1, \dots p.$$

Notons que $Var(\hat{S}_j)$ est symétrique par rapport au milieu de l'année, où elle est minimum, et augmente lorsque l'on s'en écarte.

La série ajustée et la prévision s'écrivent :

$$\hat{Y}_{ij} = \alpha t_{ij} + \beta + S_j + \overline{\varepsilon}_{.j} + p[i-(n+1)/2](\hat{\alpha}-\alpha), \quad j = 1, \dots, p, i = 1, \dots, n+1.$$

Ce sont des estimateurs sans biais de la valeur moyenne de la chronique pour chacune des dates considérées et la variance,

$$Var(\hat{Y}_{ij}) = \frac{\sigma^2}{np}\left[p + 12\frac{[i-(n+1)/2]^2}{(n^2-1)}\right],$$

ne dépend pas du mois, elle est symétrique par rapport à l'année centrale, où elle est minimum, et augmente lorsque l'on s'en éloigne.

Enfin l'erreur estimée résultant de ce modèle, appelée *résidu*,

$$\hat{\varepsilon}_{ij} = Y_{ij} - \hat{Y}_{ij} = \varepsilon_{ij} - \overline{\varepsilon}_{.j} - p[i-(n+1)/2](\hat{\alpha}-\alpha),$$

a pour variance

$$Var(\hat{\varepsilon}_{ij}) = \frac{\sigma^2}{np}\left[(n-1)p - 12\frac{[i-(n+1)/2]^2}{(n^2-1)}\right].$$

Tous ces estimateurs ont une variance proportionnelle à σ^2. On dispose d'un estimateur sans biais de σ^2,

$$\hat{\sigma}^2 = \frac{1}{np-p-1}\sum_{i,j}\hat{\varepsilon}_{ij}^2 = \frac{1}{np-p-1}\sum_{i,j}\left[Y_{ij} - \hat{\alpha}[p(i-1)+j] - \hat{\beta} - \hat{S}_j\right]^2,$$

qui est non corrélé avec les estimateurs $\hat{\alpha}, \hat{\beta}, \hat{S}_j$, et $\hat{Y}_{ij}$. Sans prétendre justifier ce résultat, notons que l'on a $\mathbb{E}(\Sigma_{ij}\varepsilon_{ij}^2) = np\sigma^2$. Le remplacement de ε_{ij} par $\hat{\varepsilon}_{ij}$ dans l'expression de $\hat{\sigma}^2$ a nécéssité l'estimation de $p+1$ paramètres indépendants $\alpha, \beta_1, \dots, \beta_p$. Ceci est à l'origine de la division par $np-(p+1)$ au lieu de np.

9.2.2 Inférence statistique

Tous les résultats précédents restent vrais lorsque l'on suppose simplement que les erreurs ε_t sont centrées, non corrélées et de même variance σ^2 (bruit blanc). L'hypothèse de normalité des erreurs implique celle des variables $\hat{\alpha}, \hat{\beta}, \hat{S}_j, \hat{Y}_{ij}$ et $\hat{\varepsilon}_{ij}$ et la non corrélation équivaut à l'indépendance. La variable $(np-p-1)\hat{\sigma}^2/\sigma^2$ suit la loi du chi-deux à $(np-p-1)$ degrés de liberté. Elle est indépendante des estimateurs $\hat{\alpha}, \hat{\beta}, \hat{S}_j$ et $\hat{Y}_{ij}$. On peut donc définir les versions studentisées de ces estimateurs, puis déterminer, en terme de p-valeur, si les paramètres correspondants sont significativement différents de zéro et construire des intervalles de confiance pour ces paramètres.

Résidus standardisés

Dans un premier temps, il est raisonnable de considérer les *résidus standardisés*, appelés aussi *résidus studentisés*, afin de valider le modèle. Ils sont définis par :

$$\hat{\varepsilon}^S_{ij} = \hat{\varepsilon}_{ij} \frac{\sqrt{np}}{\hat{\sigma}} \left[(n-1)p - 12 \frac{[i-(n+1)/2]^2}{(n^2-1)} \right]^{-1/2} ; \quad j = 1, \ldots, p; i = 1, \ldots, k.$$

La variable $\hat{\varepsilon}^S_{ij}$ suit approximativement la loi de Student à $np-p-1$ degrés de liberté. On représente donc la chronique de ces résidus en faisant figurer les seuils $\pm t_\alpha$ où t_α sera utilisé désormais dans ce chapitre pour désigner le quantile d'ordre $1-\alpha/2$ de la loi de Student à $np-p-1$ degrés de liberté : $P\{|S_{np-p-1}| > t_\alpha\} = \alpha$ (attention aux deux sens de la notation α : pente de la tendance ou niveau de signification). Bien qu'il ne s'agisse pas d'une véritable bande de confiance, le modèle doit être remis en cause si le nombre de points sortant de cette bande excède fortement $np\alpha$.

Estimateurs studentisés, p-valeurs et test d'égalité à zéro

Les *estimateurs studentisés* sont définis à l'aide des variances indiquées au Paragraphe 9.2.1, selon la normalisation habituelle consistant à diviser l'estimateur par son écart-type estimé. Les expressions obtenues sont donc les suivantes :

- $\hat{\alpha}^S = \hat{\alpha} \frac{\sqrt{np}}{\hat{\sigma}} \sqrt{\frac{p^2(n^2-1)}{12}}$,
- $\hat{\beta}^S = \hat{\beta} \frac{\sqrt{np}}{\hat{\sigma}} \left[1 + \frac{3(np+1)^2}{p^2(n^2-1)} \right]^{-1/2}$,
- $\hat{S}^S_j = \hat{S}_j \frac{\sqrt{np}}{\hat{\sigma}} \left[(p-1) + \frac{12[j-(p+1)/2]^2}{p^2(n^2-1)} \right]^{-1/2}, \quad j = 1, \ldots, p.$

Les *p-valeurs* associées aux observations de ces estimateurs studentisés permettent d'apprécier le caractère significatif ou non des paramètres correspondants. Par exemple, la probabilité $P\{|S_{np-p-1}| > |\hat{\alpha}^S_{obs}|\}$, qui est la p-valeur associée à l'observation $\hat{\alpha}^S_{obs}$ de $\hat{\alpha}^S$, sert à tester la présence ou non de la tendance. Plus cette p-valeur est faible, plus la présence d'une tendance non horizontale est significative. Le test de l'hypothèse nulle $\alpha = 0$ contre l'alternative $\alpha \neq 0$, avec un *niveau de signification* α, a pour *région critique* $\{|\hat{\alpha}^S| > t_\alpha\}$. Rejeter l'hypothèse $\alpha = 0$ équivaut donc à observer une p-valeur inférieure à α.

Intervalle de prévision

On peut construire un *intervalle de confiance* pour chaque paramètre, toujours selon le même principe basé sur la loi de Student. Par exemple, pour l'ordonnée à l'origine β, l'intervalle de confiance de niveau $(1-\alpha)$ est donné par :

$$IC(\beta; 1-\alpha) = \hat{\beta} \pm t_\alpha \frac{\hat{\sigma}}{\sqrt{np}} \sqrt{1 + \frac{3(np+1)^2}{p^2(n^2-1)}}.$$

Notons que zéro est dans cet intervalle si et seulement si $|\hat{\beta}^S| < t_\alpha$. On acceptera donc l'hypothèse nulle $\beta = 0$, pour un niveau de signification α, lorsque l'intervalle contient zéro.

Un intervalle de confiance est particulièrement intéressant pour la *prévision* de $Y_{n+1,j}$. La variable $\hat{Y}_{n+1,j}$ est un estimateur de la moyenne $\mathbb{E}(Y_{n+1,j})$ de $Y_{n+1,j}$ mais aussi de la valeur de $Y_{n+1,j}$. Pour l'intervalle de confiance, il faut tenir compte de la variance de l'erreur $\varepsilon_{n+1,j}$, ce qui équivaut à considérer la variable,

$$Y_{n+1,j} - \hat{Y}_{n+1,j} = (\alpha - \hat{\alpha})t_{n+1,j} + (\beta - \hat{\beta}) + (S_j - \hat{S}_j) + \varepsilon_{n+1,j}.$$

Cette variable suit la loi normale centrée, $\mathbb{E}(Y_{n+1,j} - \hat{Y}_{n+1,j}) = 0$, de variance,

$$Var(Y_{n+1;j} - \hat{Y}_{n+1,j}) = Var(Y_{n+1;j}) + Var(\hat{Y}_{n+1,j}) = \frac{\sigma^2}{np}\left[(n+1)\left[p + \frac{3}{n-1}\right]\right].$$

Elle est indépendante de $(np-p-1)\hat{\sigma}^2/\sigma^2$, qui suit la loi du chi-deux à $(np-p-1)$ degrés de liberté. La loi de Student, selon le shéma habituel, conduit à l'intervalle de confiance de niveau $(1-\alpha)$:

$$IC(Y_{n+1,j}; 1-\alpha) = \hat{Y}_{n+1,j} \pm t_\alpha \frac{\hat{\sigma}}{\sqrt{np}} \sqrt{(n+1)\left[p + \frac{3}{n-1}\right]}.$$

Nécéssité de l'effet saisonnier : test de Fisher

On peut évaluer chaque coefficient saisonnier S_j en considérant la p-valeur associée à l'observation de $\hat{S}_j^S$. Il est également possible de tester la présence du mouvement saisonnier dans son ensemble. Ceci est réalisé par un *test de Fisher*, dont le principe est de comparer deux estimateurs de σ^2 qui sont sans biais et indépendants sous l'hypothèse nulle H_0, c'est-à-dire en l'absence d'effet saisonnier, alors que l'un d'eux présente un biais positif sous l'alternative H_1.

Principe du test
Au départ, il s'agit d'effectuer un choix entre deux modèles dont l'un, associé à l'hypothèse nulle, est inclus dans l'autre :

- H_0 : $Y_t = \alpha_0 t + \beta_0 + \varepsilon_t$ absence de mouvement saisonnier
- H_1 : $Y_t = \alpha t + \beta + S_t + \varepsilon_t$ présence d'un mouvement saisonnier

Dans les deux cas, l'erreur ε_t est un bruit blanc gaussien de variance σ^2.

En l'absence du mouvement saisonnier, le modèle est limité à une tendance linéaire. Celle-ci est estimée par la droite des moindres carrés ajustée sur l'ensemble des observations mensuelles :

$$\hat{\alpha}_0 = \frac{cov(t_{ij}, Y_{ij})}{var(t_{ij})}, \quad \hat{\beta}_0 = \overline{Y}_{..} - \hat{\alpha}_0 \overline{t}_{..},$$

où

$$cov(t_{ij}, Y_{ij}) = \frac{1}{np}\sum_{t=1}^{np} tY_t - \frac{np+1}{2}\overline{Y}_{..},$$

$$var(t_{ij}) = \frac{1}{np}\sum_{t=1}^{np} t^2 - \left[\frac{np+1}{2}\right]^2 = \frac{n^2p^2-1}{12}.$$

Notons $\hat{Y}_0$ le vecteur constitué de la série ajustée dans le cadre de ce modèle, c'est-à-dire le vecteur de composantes $\hat{\alpha}_0 t + \hat{\beta}_0, t = 1\ldots, T$, et $\hat{\varepsilon}_0 = Y - \hat{Y}_0$, le vecteur des résidus correspondants. L'estimateur sans biais de σ^2, sous cette hypothèse, est

$$\hat{\sigma}_0^2 = \frac{1}{np-2}\sum_{t=1}^{np}\left[Y_t - \hat{\alpha}_0 t - \hat{\beta}_0\right]^2 = \frac{\|Y - \hat{Y}_0\|^2}{np-2} = \frac{\|\hat{\varepsilon}_0\|^2}{np-2}.$$

Le calcul de cet estimateur peut s'effectué sans determiner les composantes de $\hat{\varepsilon}_0$:

$$\hat{\sigma}_0^2 = \frac{np}{np-2}\left[var(Y_{ij}) - \frac{n^2p^2-1}{12}\hat{\alpha}_0^2\right], \quad var(Y_{ij}) = \frac{1}{np}\sum_{t=1}^{np}[Y_t - \overline{Y}_{..}]^2.$$

Cet estimateur n'est pas indépendant de $\hat{\sigma}^2$ car :

$$(np-2)\hat{\sigma}_0^2 = \|\hat{\varepsilon}_0\|^2 > (np-p-1)\hat{\sigma}^2 = \|\hat{\varepsilon}\|^2.$$

Par contre, il permet de construire un nouvel estimateur de σ^2, qui est sans biais et indépendant de $\hat{\sigma}^2$ sous H_0 :

$$\tilde{\sigma}^2 = \frac{\|\tilde{\varepsilon}\|^2}{(p-1)} = \frac{\|\hat{\varepsilon}_0\|^2 - \|\hat{\varepsilon}\|^2]}{(p-1)}, \quad \frac{(p-1)\tilde{\sigma}^2}{\sigma^2} \sim \chi^2_{p-1},$$

où $\tilde{\varepsilon} = \hat{\varepsilon}_0 - \hat{\varepsilon}$.

Aspect géométrique du test

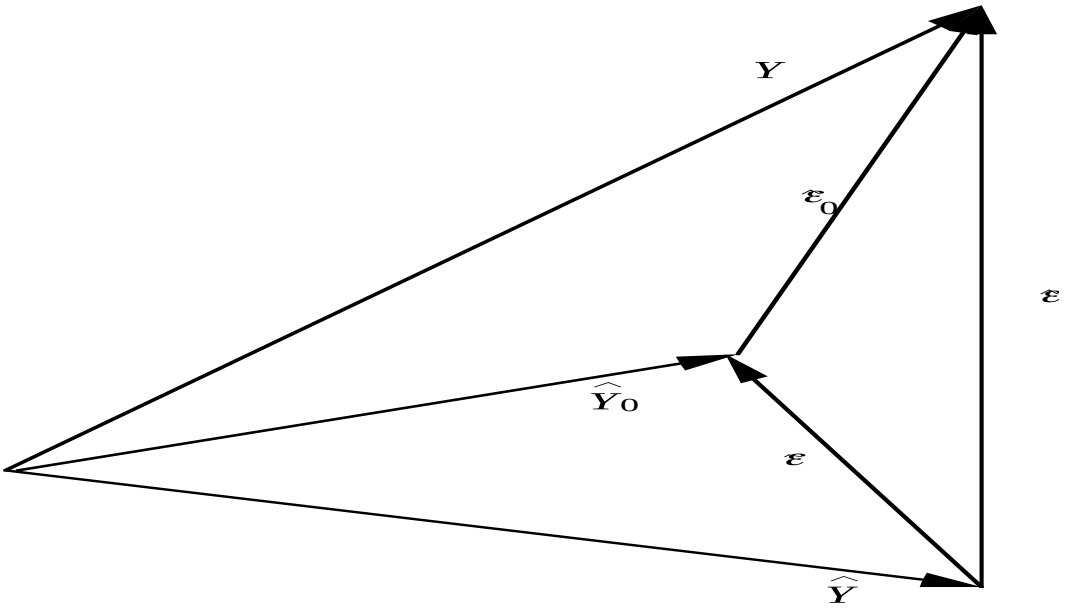

FIG. 9.5 – Aspect géométrique du test de Fisher

L'indépendance entre $\tilde{\sigma}^2$ et $\hat{\sigma}^2$ est liée à l'orthogonalité des vecteurs $\hat{\varepsilon}$ et $\tilde{\varepsilon}$ visible sur la Figure 9.5 qui traduit l'aspect géométrique du test de Fisher. Le vecteur Y, de composantes $Y_t, t = 1, \ldots, T$, évolue dans l'*espace des observations* $\mathbb{R}^T$ de dimension $T = np$. Sous l'hypothèse H_1, la série ajustée est représentée par le vecteur $\hat{Y}$, de composantes $\hat{\alpha}t + \hat{\beta} + \hat{S}_t, t = 1, \ldots, T$, qui est la projection orthogonale de Y sur le sous-espace de $\mathbb{R}^T$ constitué des chroniques réduites à une tendance linéaire, $\alpha t + \beta$, et un effet saisonnier S_t périodique de période p (sans erreur ε_t). Ce sous-espace est de dimension

$(p+1)$, égale au nombre de paramètres libres pour décrire de telle chroniques $(\alpha, \beta_1, \ldots, \beta_p)$. L'erreur correspondante, $\hat{\varepsilon} = Y - \hat{Y}$, évolue donc dans un sous-espace de dimension $(np - p - 1)$ et l'estimateur sans biais de σ^2 est $\hat{\sigma}^2 = \|\hat{\varepsilon}\|^2/(np - p - 1)$. Sous l'hypothèse H_0, la série ajustée est représentée par le vecteur $\hat{Y}_0$, de composantes $\hat{\alpha}_0 t + \hat{\beta}_0, t = 1, \ldots, T$, qui est la projection orthogonale de Y sur le sous-espace de $\mathbb{R}^T$ constitué des chroniques réduites à une tendance linéaire, $\alpha_0 t + \beta_0$ (sans effet saisonnier S_t ni erreur ε_t). Ce sous-espace est de dimension 2, égale au nombre de paramètres libres pour décrire de telle chroniques (α_0, β_0). L'erreur correspondante, $\hat{\varepsilon}_0 = Y - \hat{Y}_0$, évolue donc dans un sous-espace de dimension $(np-2)$ et l'estimateur sans biais de σ^2 est $\hat{\sigma}_0^2 = \|\hat{\varepsilon}_0\|^2/(np-2)$. Cet estimateur n'étant pas indépendant du précédent, on utilise la décomposition orthogonale $\hat{\varepsilon}_0 = \tilde{\varepsilon} \oplus \hat{\varepsilon}$ pour construire un troisième estimateur sans biais de σ^2 selon le même principe : $\tilde{\sigma}^2 = \|\tilde{\varepsilon}\|^2/(p-1)$. Ici $(p-1) = (np-2) - (np-p-1)$ est la dimension du sous-espace dans lequel evolue $\tilde{\varepsilon}$.

Sous l'hypothèse nulle, le rapport des deux estimateurs sans biais de σ^2,

$$F = \frac{\tilde{\sigma}^2}{\hat{\sigma}^2} = \frac{(np - p - 1)[\|\hat{\varepsilon}_0\|^2 - \|\hat{\varepsilon}\|^2]}{(p-1)\|\hat{\varepsilon}\|^2},$$

suit la *loi de Fisher* à $(p - 1, np - p - 1)$ degrés de liberté.

Région critique du test et p-valeur
Sous H_0, la *statistique de test* $F = \tilde{\sigma}^2/\hat{\sigma}^2$ suit la loi $\mathcal{F}_{p-1,np-p-1}$ avec $\mathbb{E}(F) = (np - p - 1)/(np - p - 3)$. Sous H_1, la loi de F est différente et sa moyenne $\mathbb{E}(F)$ est plus grande que $(np - p - 1)/(np - p - 3)$. On décidera donc de rejeter H_0 lorsqu'on observe de grandes valeurs pour F. La *région critique* du test, pour un *niveau de signification* α, est donnée par :

$$\{F > f_{p-1,np-p-1;\alpha}\}, \quad f_{p-1,np-p-1;\alpha} = F^{-1}_{\mathcal{F}_{p-1,np-p-1}}(1-\alpha),$$

où $F_{\mathcal{F}_{p-1,np-p-1}}$ désigne la fonction de répartition de la loi $\mathcal{F}_{p-1,np-p-1}$. Cette fonction, tout comme son inverse $F^{-1}_{\mathcal{F}_{p-1,np-p-1}}$, n'a pas d'expression explicite. Il existe des tables spécifiques à la réalisation du test de Fisher donnant $f_{n,m;\alpha}$ en fonction de quelques valeurs courantes de n, m et α. Elles sont maintenant avantageusement remplacées par les versions numériques de ces fonctions données par les logiciels adaptés (Excel) et même par les calculatrices scientifiques actuelles. Cela permet en particulier d'appréhender le test en terme de p-valeur associée à l'observation F_{obs} de la statistique de test :

$$p\text{-valeur} = P\{F > F_{obs}|H_0\} = 1 - F_{\mathcal{F}_{p-1,np-p-1}}(F_{obs}).$$

Illustration

Le Tableau 9.2 donne les résidus réels et standardisés de la chronique du chiffre d'affaires de la presse parisienne. Pour cela on a utilisé les résultats intermédiaires suivants :

$$\Sigma\hat{\varepsilon}_t^2 = 3516,37; \quad \hat{\sigma}^2 = 74,82; \quad \hat{\sigma} = 8,6.$$

Les résidus standardisés sont représentés sur la Figure 9.6, dans laquelle on a également indiqué la "bande de confiance" $\pm t_\alpha$ de niveau $(1-\alpha) = 95\%$ où t_α est obtenu sous Excel par :

$$t_\alpha\text{=LOI.STUDENT.INVERSE(0,05 ;47)=2,01.}$$

On retrouve la forme incurvée des résidus, mais le nombre de points au dehors de la bande reste faible.

année	*mois* / Résidus	janv	fév	mars	avr	mai	juin	juil	août	sept	oct	nov	déc
	Réels	-15	-1	-12	-7	-1	6	-9	3	-5	-2	8	-1
1981	Standardisés	-2,04	-0,16	-1,64	-0,98	-0,16	0,74	-1,22	0,39	-0,69	-0,29	1,05	-0,11
	Réels	-4	2	0	2	6	-8	13	-7	11	-6	12	2
1982	Standardisés	-0,53	0,27	-0,02	0,24	0,79	-1,05	1,70	-0,87	1,44	-0,77	1,60	0,32
	Réels	6	2	6	-1	4	-4	1	17	3	-9	3	0
1983	Standardisés	0,80	0,31	0,80	-0,10	0,57	-0,49	0,18	2,15	0,44	-1,11	0,34	-0,03
	Réels	9	7	19	-3	2	1	0	-2	-7	6	-5	0
1984	Standardisés	1,11	0,87	2,41	-0,45	0,22	0,07	-0,04	-0,27	-0,95	0,74	-0,66	0,02
	Réels	5	-10	-12	10	-11	6	-5	-11	-2	11	-18	-2
1985	Standardisés	0,64	-1,32	-1,61	1,30	-1,45	0,77	-0,66	-1,42	-0,26	1,46	-2,35	-0,21

TAB. 9.2 – Chiffre d'affaires mensuel de la presse parisienne dans une petite ville de province : résidus réels et standardisés

Le Tableau 9.3 confirme sans ambiguïté la nécéssité de la tendance. Le Tableau 9.4 indique que seuls les coefficients saisonniers des mois d'octobre, novembre, mai et septembre sont significatifs à 5%. Ceci tempère les commentaires inspirés par la Figure 9.2 ainsi que par les valeurs réels des coefficients. On a déjà remarqué que l'effet du mouvement saisonnier est relativement faible puisque, au vu des valeurs réelles des coefficients saisonniers, il ne représente environ que 10% des valeurs de la chronique. Cependant le test de Fisher, dont les éléments sont regroupés dans le Tableau 9.5, confirme la nécéssité de ce mouvement. Pour cela, on a utilisé les fonctions suivantes d'Excel :

$$f_{11,47;5\%} = \text{INVERSE.LOI.F(0,05 ;11 ;47)} = 1,9991\,;$$
$$p\text{-valeur} = \text{LOI.F(5,16 ;11 ;47)} = \text{3E-5.}$$

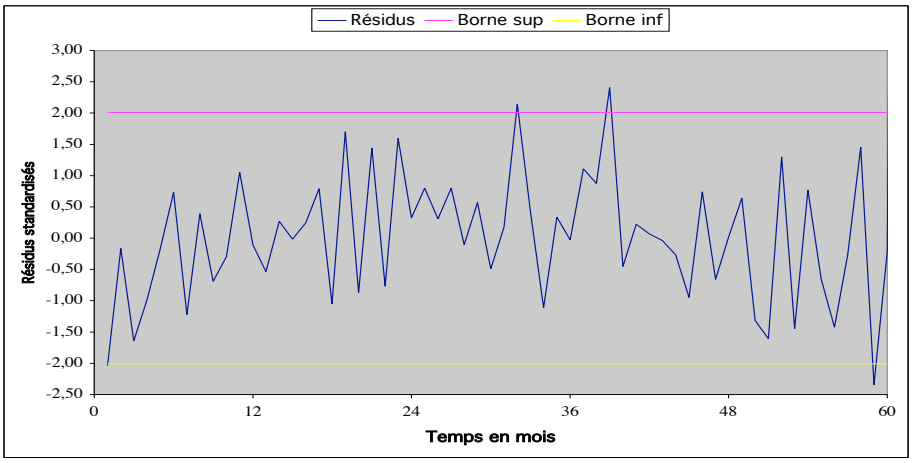

FIG. 9.6 – Chiffre d'affaires mensuel de la presse parisienne dans une petite ville de province : représentation des résidus standardisés

Enfin le Tableau 9.6 rappelle la prévision pour l'année 1986 en indiquant les bornes des "intervalles de confiance" à 95%. Notons que ces ces intervalles sont de la forme $\hat{Y}_{6,j} \pm 19,6$.

Paramètre	pente	ordonnée à l'origine
Valeur réelle	1,39	92
Valeur studentisée	21,13	39,85
p-valeur	1E-25	7E-38

TAB. 9.3 – Tendance du chiffre d'affaires mensuel de la presse parisienne dans une petite ville de province : aspect significatif des coefficients

Mois	janv	fév	mars	avr	mai	juin	juil	août	sept	oct	nov	déc
Coef. réels	6,5	-1,1	6,7	-6,7	-12,2	-5,4	4,0	-1,6	8,2	16,8	-12,8	-2,4
Coef. studentisés	1,8	-0,3	1,8	-1,8	-3,3	-1,5	1,1	-0,4	2,2	4,5	-3,4	-0,6
p-valeurs	0,087	0,773	0,076	0,079	0,002	0,149	0,289	0,664	0,032	4E-5	0,001	0,526

Seuil du test de Student de niveau 5% à 47 d.l. : 2,01

TAB. 9.4 – Chiffre d'affaires mensuel de la presse parisienne dans une petite ville de province : aspect significatif des coefficients saisonniers

Modèle	pente	ordonnée à l'origine	écart-type de l'erreur	variance de l'erreur
Avec effet saisonnier	1,390	91,5	8,6	74,82
Sans effet saisonnier	1,388	91,6	11,6	133,88

$var(Y) = 707,13$; $cov(t, Y) = 416,25$; statistique du test observée $= 5,16$; p-valeur = 3E-5

Seuil du test de Fisher-Snedecor de niveau 5% à 11 et 47 d.l. : 2,00

TAB. 9.5 – Test de l'effet saisonnier du chiffre d'affaires mensuel de la presse parisienne dans une petite ville de province

Mois	janv	fév	mars	avr	mai	juin	juil	août	sept	oct	nov	déc
Borne sup	202	196	205	193	189	197	208	204	215	225	197	209
Prévision	183	177	186	174	170	178	189	184	196	206	177	189
Borne inf	163	157	166	154	150	158	169	165	176	186	158	170

TAB. 9.6 – Chiffre d'affaires mensuel de la presse parisienne dans une petite ville de province : prévision 1986 avec intervalle de confiance à 95%

En résumé On retiendra comment le modèle de Buys-Ballot permet d'estimer une tendance linéaire et un effet saisonnier périodique, conduisant ainsi à la possibilité d'effectuer une prévision. On notera aussi la façon d'apprécier ces résultats grâce à l'apport de la statistique inductive (p-valeurs, tests et intervalles de confiance).

Chapitre 10

LISSAGE ET SÉRIE CVS

Ce chapitre aborde les mêmes points que le précédent, recherche de la tendance, de la composante saisonnière et prévision, dans un cadre non paramétrique. Nous conservons le principe du schéma additif :

$$Y_t = f_t + S_t + \varepsilon_t, \quad t = 1, \ldots, T, \quad \varepsilon_t \sim b.b.(\sigma^2).$$

La *composante saisonnière* S_t, lorsqu'elle est présente, est souvent supposée rigoureusement périodique de période p connue et définie par les *coefficients saisonniers* $S_j, j = 1, \ldots, p$, satisfaisant la contrainte $\Sigma_{j=1}^{p} S_j = 0$. La différence fondamentale avec le chapitre précédent est de ne pas structurer la *tendance* f_t. Celle-ci est une fonction non paramétrique à variation lente. Il n'est donc pas nécessaire d'imposer de fortes contraintes à la partie aléatoire ε_t car aucune étude statistique fine ne peut être envisagée dans ce cadre.

En fait, l'approche est ici beaucoup plus descriptive et l'un des buts principaux est de désaisonnaliser la série observée. Ceci consiste à éliminer l'effet saisonnier, ou plus exactement à le répartir de façon uniforme à l'intérieur de chaque période. Pour cela il suffit de retrancher à la série brute Y_t la composante saisonnière S_t, dont l'effet est en moyenne nul sur une période. La nouvelle série ainsi obtenue, dite *désaisonnalisée* ou *Corrigée des Variations Saisonnières* (*série CVS*) et notée Y_t^c, constitue une série artificielle sans effet saisonnier conservant tout le potentiel de la série initiale. Elle conserve en particulier la partie résiduelle ε_t et ne doit pas être confondue avec la tendance. L'intérêt est de rendre comparables deux valeurs consécutives. Le principe de la méthode consiste à retrancher la tendance, estimée par lissage de la série brute, puis à estimer la composante saisonnière sur la nouvelle série ainsi obtenue. Cette estimation est réalisée de façon paramétrique. On comprend qu'il n'est pas nécessaire, pour un tel objectif, de contraindre la tendance à une forme paramétrique. Par ailleurs son estimation sous forme

de fonction lisse permet de localiser les différentes phases du *mouvement conjoncturel* (croissance, stagnation,...).

Les opérations de lissage évoquées jusqu'ici sont réalisées par le biais de *moyennes mobiles*. Une chronique $Y_t, t = 1, \ldots, T$ est lissée en remplaçant chaque valeur Y_t par une moyenne pondérée des valeurs qui l'entourent :

$$\tilde{Y}_t = \sum_{j=-l}^{k} \gamma_j Y_{t-j}, \quad t = k+1, \ldots, T-l.$$

Cette écriture regroupe une très grande diversité dans la mesure où aucune restriction n'est faite sur les coefficients γ_j. Nous nous contenterons ici de simples *moyennes arithmétiques* qui s'avèrent suffisantes pour la construction de la séries CVS dans la plupart des situations ordinaires. Dans ce cas les coefficients sont simples, positifs et symétriques ($\gamma_{-j} = \gamma_j$).

L'inconvénient de l'approche non paramétrique est de ne pas pouvoir effectuer de prévisions. Un compromis est obtenu par les méthodes de *lissage exponentiel*. Celles-ci consistent à ajuster un modèle paramétrique de façon locale, c'est-à-dire dont les paramètres évoluent au cours du temps. Une prévision à très court terme est alors possible. Par exemple, dans le lissage exponentiel simple, la prévision à un pas est donnée par :

$$\hat{Y}_T(1) = (1-\gamma) \sum_{k=0}^{\infty} \gamma^k Y_{T-k},$$

où la *constante de lissage* γ satisfait $0 < \gamma < 1$ et $Y_t = 0$ pour $t \leq 0$ par convention. L'intérêt de la méthode réside dans la facilité de *mise à jour* de cette prévision lors de l'acquisition d'une nouvelle donnée Y_{T+1} :

$$\hat{Y}_{T+1}(1) = \hat{Y}_T(1) + (1-\gamma)\left[Y_{T+1} - \hat{Y}_T(1)\right].$$

La nouvelle prévision est égale à la précédente modifiée par la dernière erreur de prévision effectuée. Ceci est un exemple simple du célèbre *filtre de Kalman*. Le lissage exponentiel ne sera pas étudié ici.

Le premier paragraphe présente le lissage dans le cadre simple des moyennes arithmétiques. La série CVS est alors définie au paragraphe suivant.

10.1 Lissage : moyennes arithmétiques

Les moyennes arithmétiques conduisent à une classe particulière de moyennes mobiles. Celles-ci sont cependant très utilisées car elles sont à la fois de conception simple, faciles à mettre en œuvre et suffisantes dans bien des situations. C'est pourquoi le terme *moyenne mobile* fera toujours référence ici à une moyenne arithmétique.

10.1.1 Définitions et propriétés immédiates

Le lissage d'une chronique $Y_t, t = 1, \ldots, T$ par une *moyenne arithmétique d'ordre impair* $m = 2k + 1$ est défini pour $t = k + 1, \ldots, T - k$, par :

$$\tilde{Y}_t = M_m(Y_t) = \frac{1}{m}\{Y_{t-k} + \ldots + Y_t + \ldots + Y_{t+k}\} = \frac{1}{2k+1}\sum_{j=-k}^{k} Y_{t+j}.$$

Chaque valeur Y_t est donc remplacée par une simple moyenne arithmétique des valeurs qui l'entourent.

Afin de préserver la symétrie, la *moyenne mobile d'ordre pair* $m = 2k$ est définie aux mêmes instants $t = k + 1, \ldots, T - k$ et porte sur les mêmes valeurs mais avec un poids 0,5 aux deux extrémités :

$$\tilde{Y}_t = M_m(Y_t) = \frac{1}{m}\left[\frac{Y_{t-k}}{2} + \sum_{j=1-k}^{k-1} Y_{t+j} + \frac{Y_{t+k}}{2}\right] = \frac{1}{2k}\left[\sum_{j=-k}^{k} Y_{t+j} - \frac{Y_{t-k} + Y_{t+k}}{2}\right].$$

Elle est égale à la moyenne d'ordre 2 de deux moyennes arithmétiques d'ordre m consécutives :

$$\tilde{Y}_t = \frac{\tilde{Y}_{t-0,5} + \tilde{Y}_{t+0,5}}{2} = \frac{1}{2}\left[\frac{1}{m}\sum_{j=-k}^{k-1} Y_{t+j} + \frac{1}{m}\sum_{j=1-k}^{k} Y_{t+j}\right].$$

La série lissée est plus courte que l'originale puisque $[m/2]$ valeurs sont manquantes à chaque extrémité de la période d'observation. Un calcul récursif en temps de $M_m(Y_t)$ est facile à mettre en œuvre. Le point $(t, \tilde{Y}_t)$ est le centre de gravité des points (s, Y_s) qui ont servi à sa définition (avec un poids 0,5 aux extrémités dans le cas pair). En conséquence, à concavité constante, la série lissée reste d'un même côté par rapport à la série brute (sous-estimation ou surestimation de la tendance) et les points de retournement sont décalés dans les cas asymétriques (mauvaise localisation des changements de tendance) (*cf.*

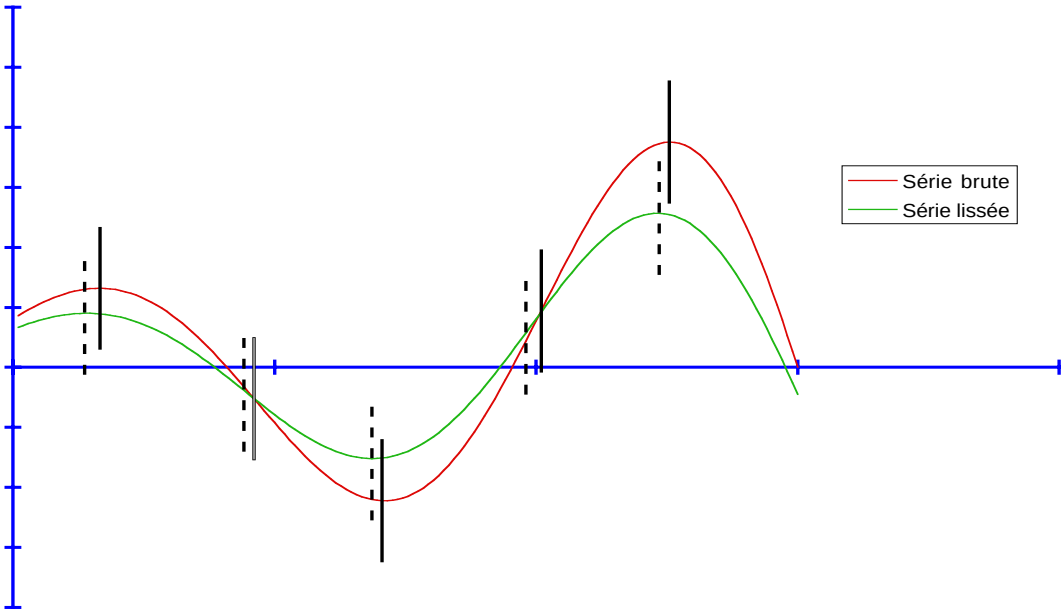

FIG. 10.1 – Le lissage par moyennes mobiles décale les changements de tendance (moyenne mobile d'ordre 51, séries représentées de taille 150)

Figure 10.1). En effet l'opération de moyenne mobile est clairement linéaire et dans le cas du modèle additif avec composante saisonnière, $Y_t = f_t + S_t + \varepsilon_t$, on a :

$$\tilde{Y}_t = M_m(Y_t) = M_m(f_t) + M_m(S_t) + M_m(\varepsilon_t).$$

En particulier une tendance linéaire n'est pas modifiée par moyenne mobile.

Illustration

On considère la chronique trimestrielle donnant l'évolution en pourcentage de la production industrielle française au cours des années 1981 à 1986 (source INSEE). Nous avons représenté (*cf.* Figure 10.2) la série brute ainsi que les séries lissées par les moyennes mobiles d'ordre 3, 4 et 5. Les données et les valeurs lissées sont dans le Tableau 10.1 . Il est clair que la moyenne mobile lisse la série puisqu'elle atténue les oscillations. Dans cet exemple, où l'effet saisonnier (période 4) est très prononcé, les moyennes mobiles d'ordre 3 et 5 ne doivent pas être utilisées sur la série brute car elles perturbent complètement l'effet saisonnier (les pics et les creux sont déplacés). Par contre la moyenne mobile d'ordre 4 est particulièrement bien adaptée. La section qui suit justifie ce constat. De façon générale, en présence d'un mouvement

saisonnier de période p, il ne faut pas effectuer de lissage avec une moyenne mobile d'odre proche de p ou qui soit un diviseurs ou un multiple de p.

		trimestre	JFM	AMJ	JAS	OND
année	i	j	1	2	3	4
1981	1	Y_t	-1,9	-2,9	-14,0	23,3
		$M3(Y_t)$		-6,3	2,1	2,5
		$M4(Y_t)$			1,1	1,3
		$M5(Y_t)$			0,5	0,5
		$Y_t - M4(Y_t)$			-15,1	22,0
		CVS	1,4	0,5	1,6	1,1
1982	2	Y_t	-1,9	-1,9	-17,6	23,8
		$M3(Y_t)$	6,5	-7,1	1,4	1,7
		$M4(Y_t)$	0,9	0,5	0,7	0,8
		$M5(Y_t)$	-2,4	5,1	0,3	0,3
		$Y_t - M4(Y_t)$	-2,8	-2,4	-18,3	23,0
		CVS	1,4	1,5	-2,0	1,6
1983	3	Y_t	-1,0	-1,9	-16,8	23,8
		$M3(Y_t)$	7,0	-6,6	1,7	3,0
		$M4(Y_t)$	0,9	1,0	1,4	1,2
		$M5(Y_t)$	-2,7	5,6	1,2	0,1
		$Y_t - M4(Y_t)$	-1,9	-2,9	-18,2	22,6
		CVS	2,3	1,5	-1,2	1,6
1984	4	Y_t	1,9	-6,6	-14,1	23,5
		$M3(Y_t)$	6,4	-6,3	0,9	2,5
		$M4(Y_t)$	0,9	1,2	0,7	0,7
		$M5(Y_t)$	-2,4	5,7	0,6	-0,4
		$Y_t - M4(Y_t)$	1,0	-7,8	-14,8	22,8
		CVS	5,2	-3,2	1,5	1,3
1985	5	Y_t	-1,9	-2,9	-12,0	20,5
		$M3(Y_t)$	6,2	-5,6	1,9	1,9
		$M4(Y_t)$	1,4	1,3	0,8	0,9
		$M5(Y_t)$	-1,5	5,4	0,2	0,4
		$Y_t - M4(Y_t)$	-3,3	-4,2	-12,8	19,6
		CVS	1,4	0,5	3,6	-1,7
1986	6	Y_t	-2,8	-1,0	-13,7	20,5
		$M3(Y_t)$	5,6	-5,8	1,9	
		$M4(Y_t)$	1,0	0,8		
		$M5(Y_t)$	-1,8	4,7		
		$Y_t - M4(Y_t)$	-3,8	-1,8		
		CVS	0,5	2,4	1,9	-1,7
		S'_j	-2,8	-2,9	-15,1	22,6
		S_j	-3,3	-3,4	-15,6	22,2

TAB. 10.1 – Évolution en pourcentage de la production industrielle française de 1981 à 1986 : tableau de résultats

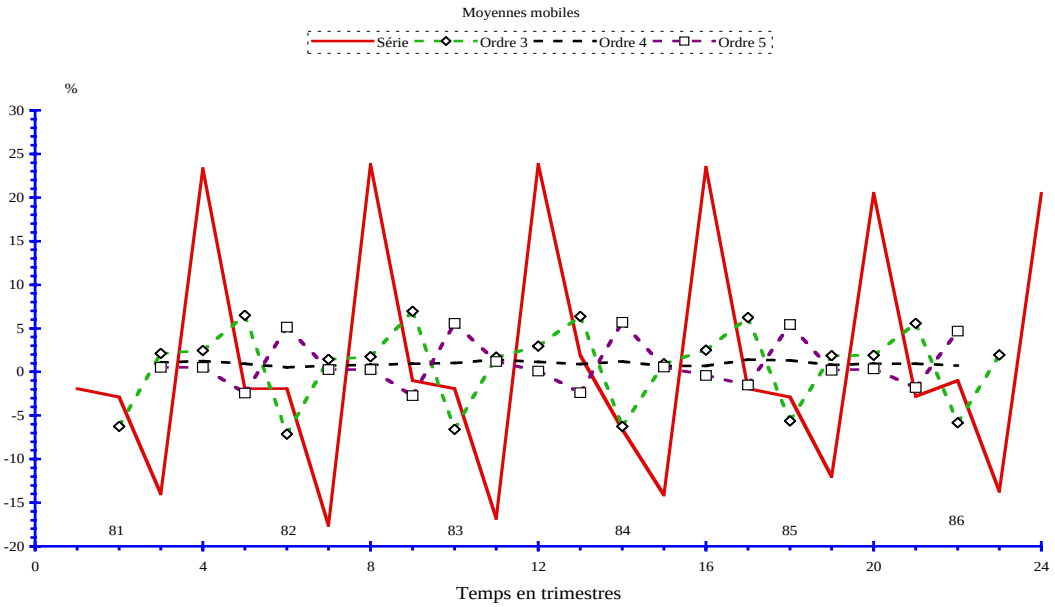

FIG. 10.2 – Évolution trimestrielle de la production industrielle française de 1981 à 1986 : moyennes mobiles d'ordre variable

10.2 Série corrigée des variations saisonnières (Série CVS)

Les moyennes mobiles permettent de construire des *séries désaisonnalisées* sans avoir à faire d'hypothèses contraignantes sur la série brute. On se place dans le cadre du modèle additif :

$$Y_t = f_t + S_t + \varepsilon_t, \quad t = 1, \ldots, T, \quad \varepsilon_t \sim b.b.(\sigma^2),$$

où le mouvement saisonnier est rigoureusement périodique de période p et défini par les *coefficients saisonniers* $S_j, j = 1, \ldots, p$ vérifiant $\Sigma_{j=1}^{p} S_j = 0$. Appliquons la moyenne mobile d'ordre p :

$$M_p(Y_t) = M_p(f_t) + M_p(S_t) + M_p(\varepsilon_t) = M_p(f_t) + M_p(\varepsilon_t).$$

En effet la moyenne mobile d'ordre p d'une série périodique de période p est clairement égale à la moyenne arithmétique des valeurs d'une période, elle est donc constante et $M_p(S_t) = 0$ compte tenu que les coefficients saisonniers "sont centrés" (de moyenne nulle). La tendance f_t étant une fonction lisse, la série $M_p(f_t)$ est peu différente de f_t. Enfin la partie résiduelle $M_p(\varepsilon_t)$ doit être voisine de 0 puisqu'elle représente la moyenne de p variables centrées non

corrélées. En résumé la série $M_p(Y_t)$ constitue une estimation de la tendance, $\hat{f}_t = M_p(Y_t)$, et l'estimateur est sans biais lorsque la tendance est linéaire.

Pour estimer l'effet saisonnier, on retire la tendance ainsi estimée à la série brute, $\Delta_t = Y_t - M_p(Y_t), t = [p/2] + 1, \ldots, T - [p/2]$. Pour chaque "mois" $j = 1, \ldots, p$ fixé, les différences Δ_{ij} observées sur les "années" $i = 1$ ou $2, \ldots, n-1$ ou n constituent n ou $n-1$ estimations du coefficient saisonnier S_j. Le nombre d'années étant en général faible, on retient les médianes S'_j de ces valeurs comme première estimation, car la moyenne est trop sensible aux valeurs extrêmes. Lorsque la médiane porte sur un nombre pair de valeurs, on prend la demi-somme des valeurs centrales. L'estimation définitive est obtenue en centrant ces médianes en accord avec la contrainte $\Sigma_{j=1}^{p} S_j = 0$:

$$\hat{S}_j = S'(j) - \frac{1}{p}\sum_{k=1}^{p} S'_k, \quad j = 1, \ldots, p.$$

La *série Corrigée des Variations Saisonnières*, notée Y_t^c, est définie sur toute la période d'observation, $t = 1, \ldots, T$, en retranchant à la série brute l'estimation de la composante saisonnière :

$$Y_{ij}^c = Y_{ij} - \hat{S}_j, \quad j = 1, \ldots, p, \quad i = 1, \ldots, n.$$

Illustration

Les calculs concernant l'évolution de la production industrielle figurent dans le Tableau 10.1. La tendance est stable au voisinage de 1% et la série CVS conserve les irrégularités dues à la composante résiduelle (*cf.* Figure 10.3). La représentation des mouvements saisonniers (*cf.* Figure 10.4) montrent clairement l'élimination de l'effet saisonnier.

Itération du procédé

Il se peut que la série CVS obtenue ci-dessus présente encore un effet saisonnier. Dans ce cas on peut estimer à nouveau la tendance en lissant la série par une moyenne mobile d'ordre plus faible que p (2 ou 3 pour une série trimestrielle, 5 ou 7 pour une série annuelle). On reprend alors l'estimation des coefficients saisonniers comme précédemment à l'aide de la série brute Y_t et de la nouvelle tendance.

En résumé On retiendra l'estimation de la tendance par lissage et la construction de la série CVS.

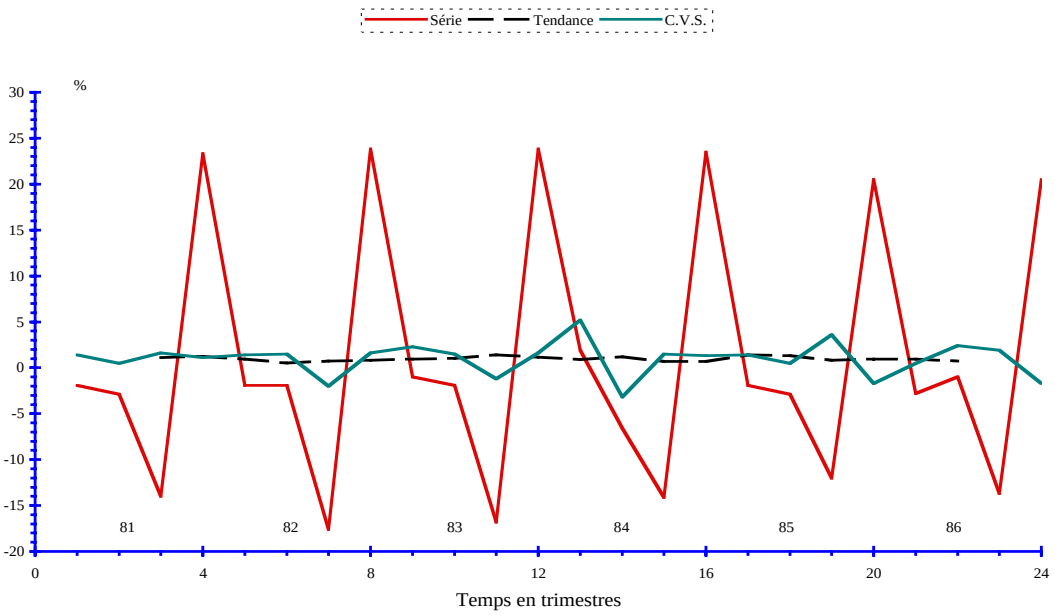

FIG. 10.3 – Évolution trimestrielle de la production industrielle française de 1981 à 1986 : chronique, tendance et série CVS

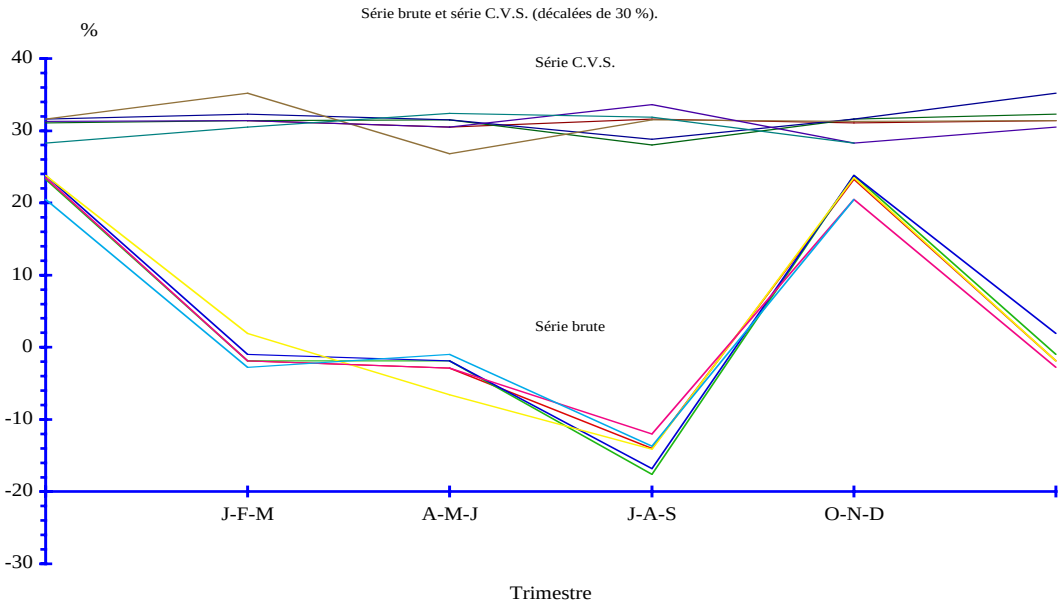

FIG. 10.4 – Évolution trimestrielle de la production industrielle française de 1981 à 1986 : mouvements saisonniers

Liste des tableaux

Table des figures

Bibliographie

[BAI89a] G. BAILLARGEON. *Méthodes statistiques de l'ingénieur*, volume 1. SMG, 3^e edition, 1989.

[BAI89b] G. BAILLARGEON. *Probabilités, statistique et techniques de régression*. SMG, 1989.

[BBT89] A. BENSAMER and B. BLEUSE-TRILLON. *Pratique des chroniques et de la prévision à court terme*. Masson, Paris, 1989.

[CAL65] G. CALOT. *Cours de statistique descriptive*. Dunod, Paris, 3^e edition, 1965. Collection Statistique et Programmes Économiques.

[CG00] M. COMTE and J. GADEN. *Statistiques et probabilités pour les sciences économiques et sociale*. Presses Universitaires de France, 2000.

[GM90] C. GOURIEROUX and A. MONFORT. *Séries temporelles et modèles dynamiques*. Economica, Paris, 1990.

[GRA98] B. GRAIS. *Méthodes statistiques*. Dunod, Paris, 1998.

[GRA00] B. GRAIS. *Statistique descriptive*. Dunod, Paris, 2000.

[JAF96] P. JAFFARD. *Initiation aux méthodes de la statistique et du calcul des probabilités*. Masson, 3^e edition, 1996.

[PY98] B. PY. *Statistique descriptive*. Economica, Paris, 1998.

[SCH97] D. SCHLACTHER. *Comprendre les mathématiques financières*. Hachette Supérieur, 1997.

Index

Table des matières