

LE MODÈLE LINÉAIRE EN STATISTIQUE

RÉGRESSION LINÉAIRE

Serge Dégerine

3 Décembre 2002

INTRODUCTION

La *régression linéaire* constitue une partie importante du *modèle linéaire*. De manière très générale, le modèle linéaire rend compte d'une situation dans laquelle la moyenne du vecteur aléatoire Y , que constituent les observations, dépend linéairement et de façon connue d'un paramètre vectoriel θ :

$$Y = X\theta + \varepsilon, \quad \mathbb{E}(Y) = X\theta, \quad \text{Var}(Y) = \text{Var}(\varepsilon) = \sigma^2 I$$

La partie aléatoire est donc représentée par une erreur additive ε dont les composantes sont centrées, de même variance σ^2 et non corrélées. Nous ajouterons l'hypothèse gaussienne pour les problèmes de tests ou d'intervalles de confiance. Les observations sont donc des variables aléatoires de même variance et non corrélées entre elles (indépendantes dans le cas gaussien). Leurs moyennes sont par contre différentes et représentées par les composantes du vecteur $X\theta$. Ce vecteur est une combinaison linéaire inconnue, dont les coefficients sont les composantes du paramètre θ , des colonnes de la matrice X qui, elle, est connue et constitue le *plan d'expérience*. La diversité des modèles provient de la nature de cette matrice. Dans le cas de la régression linéaire simple ou multiple, les colonnes de X sont des variables quantitatives comme Y , éventuellement observées, mais supposées parfaitement connues, c'est-à-dire non aléatoires. En *analyse de la variance*, les éléments de X sont à valeurs dans $\{0, 1\}$; ils indiquent la présence ou non, dans chaque observation, des différentes modalités de variables qualitatives externes. L'*analyse de la covariance* correspond à la juxtaposition dans X de ces deux types de variables. Les modèles *logit* et *probit* constituent un exemple de généralisation dans lequel une fonction connue de la moyenne dépend linéairement du paramètre.

L'intérêt du modèle linéaire est de fournir de façon simple, par la méthode des moindres carrés, les estimateurs des paramètres θ et σ^2 , leurs propriétés ainsi que les intervalles de confiance et les tests sous hypothèse gaussienne. De plus une représentation géométrique (*cf.* Figure 1) facilite la compréhension des résultats. Cependant le contexte de la régression linéaire est très différent

de celui de l'analyse de la variance ou de la covariance et les techniques de calcul sont finalement très spécifiques.

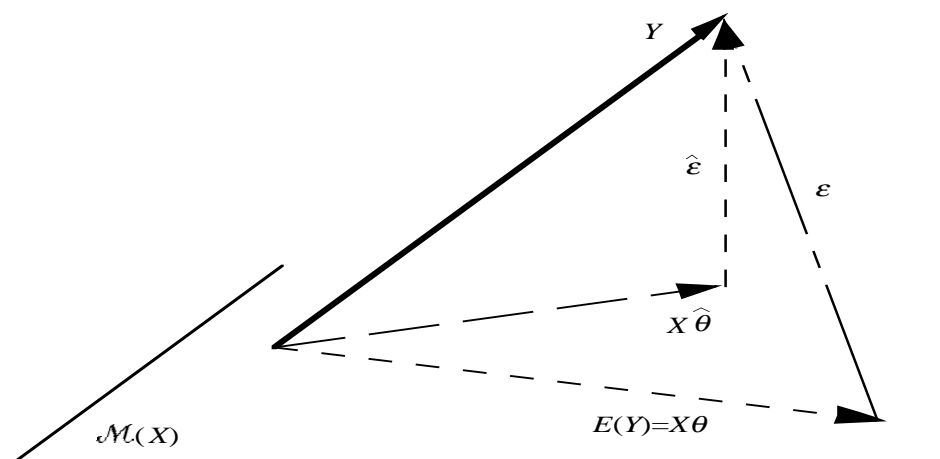


Figure 1: Aspect géométrique du modèle linéaire

La *régression linéaire simple*, présentée dans le premier chapitre, permet d'aborder le modèle linéaire de façon élémentaire. Dans ce cas la variable d'intérêt est confrontée à une seule autre variable, c'est par exemple l'étude du poids des individus en fonction de la taille. La *régression linéaire multiple*, au chapitre suivant, nécessite un formalisme plus important car la variable d'intérêt est analysée simultanément à travers plusieurs variables. La *régression poissonnienne*, ainsi que les modèles logit et probit, entrent dans le cadre du *modèle linéaire généralisé* présenté au troisième chapitre. L'*analyse de la variance* fait l'objet du quatrième chapitre. Elle a pour but d'étudier l'influence d'un ou plusieurs facteurs sur la grandeur étudiée. On mesure ainsi l'effet d'un ou plusieurs médicaments, à doses variables, pour le traitement d'une maladie. L'organisation des calculs est très spécifique. L'*analyse de la covariance*, au cinquième chapitre, permet de prendre en compte des facteurs qualitatifs dans un problème de régression. Par exemple on cherche à éliminer l'effet du sexe dans l'étude du poids en fonction de la taille chez les individus. Elle utilise simultanément les techniques de la régression et celles de l'analyse de la variance. Enfin le dernier chapitre est consacré à l'*analyse de la déviance*.

N.B. Le présent document contient uniquement les deux premiers chapitres.

Chapitre 1

RÉGRESSION LINÉAIRE SIMPLE

La *régression linéaire simple* a pour objectif d'étudier la dépendance, sous forme linéaire, entre deux grandeurs. L'exemple classique du poids d'un individu en fonction de sa taille est illustré ci-dessous par un échantillon de 32 étudiants.

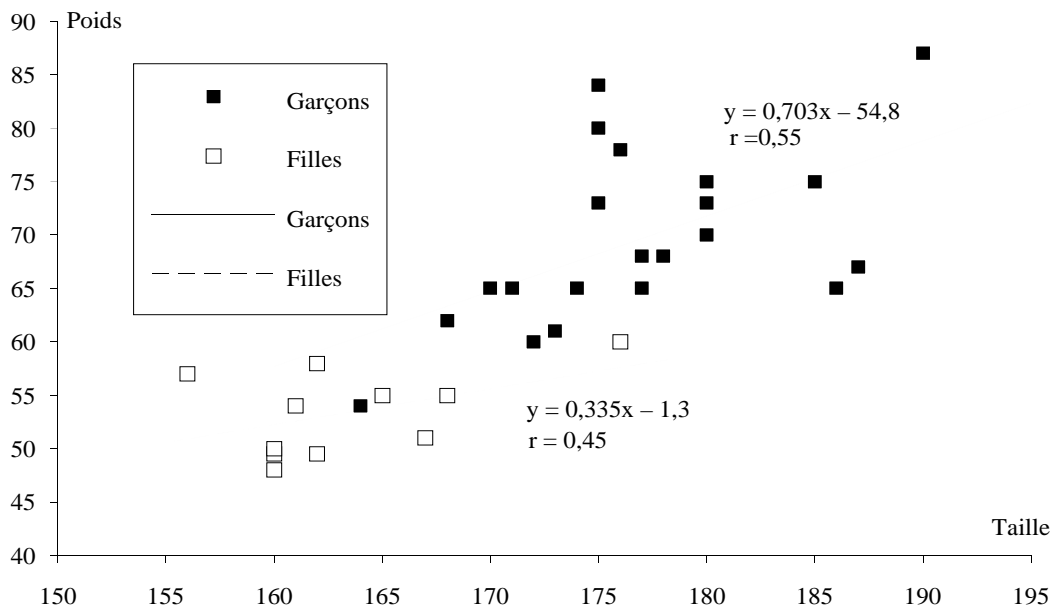


Figure 1.1: Régression du poids par rapport à la taille des étudiants de la promotion 97 d'IUP 3

1.1. DROITES ASSOCIÉES À UN ENSEMBLE DE POINTS DU PLAN 5

Nous commencerons par une approche descriptive du problème. Le premier paragraphe présente quatre principes différents permettant d'associer une droite à un ensemble de points du plan. Le contexte de la *droite de régression* est précisé au paragraphe suivant dans lequel est introduit le *coefficient de corrélation linéaire*. L'aspect descriptif est étendu au cadre de la *régression non linéaire* avec les *courbes de régression* et *rapports de corrélation* dans le troisième paragraphe. Les deux derniers paragraphes sont consacrés aux aspects de la statistique inductive dans ce contexte : étude des estimateurs, intervalles de confiance et tests.

1.1 DROITES ASSOCIÉES À UN ENSEMBLE DE POINTS DU PLAN

On se propose ici de montrer qu'il existe plusieurs façons assez naturelles d'associer une droite à un ensemble de points du plan. On se donne n points sous forme de leurs coordonnées $(x_i, y_i), i = 1, \dots, n$, dans un repère orthogonal. On cherche une droite d'équation $y = ax + b$ qui soit le plus proche possible de l'ensemble de ces points. Cela dépend de la manière de mesurer la proximité du point (x_i, y_i) à la droite lorsque le critère global est la valeur moyenne de ces n mesures.

1.1.1 Droites des moindres carrés

La *droite des moindres carrés* de y par rapport à x minimise la somme des carrés des distances des points à la droite mesurées verticalement. Le critère à minimiser se présente donc sous la forme

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

La dérivée partielle par rapport à b montre que la droite passe par le point moyen (\bar{x}, \bar{y}) :

$$\frac{\partial}{\partial b} D_{y/x}(a, b) = \frac{-2}{n} \sum_{i=1}^n [y_i - ax_i - b] = 0 \quad \Leftrightarrow \quad b = \bar{y} - a\bar{x}.$$

Le report de ce résultat dans la dérivée partielle par rapport à a donne la pente :

$$\frac{\partial}{\partial a} D_{y/x}(a, b) = \frac{-2}{n} \sum_{i=1}^n x_i [y_i - ax_i - b] = 0 \quad \Rightarrow \quad \text{cov}(x, y) - a \text{var}(x) = 0.$$

On peut aussi introduire le point moyen (\bar{x}, \bar{y}) dans l'expression du critère,

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

puis effectuer le développement,

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

Pour a fixé, la droite qui minimise le critère doit passer par le point moyen, d'où la relation : $b = \bar{y} - a\bar{x}$. Puis $D_{y/x}(a, \bar{y} - a\bar{x})$ est un polynôme du second degré en a dont l'annulation de la dérivée, $2a \text{var}(x) - 2 \text{cov}(x, y) = 0$, donne la condition sur a . Cette seconde approche établit qu'il s'agit bien d'un minimum. Dans le premier cas, on constate que la matrice des dérivées secondes (matrice hessienne), en la solution, est définie positive,

$$\frac{\partial^2 D_{y/x}(a, b)}{\partial \begin{pmatrix} a \\ b \end{pmatrix} \partial \begin{pmatrix} a & b \end{pmatrix}} = 2 \begin{bmatrix} \bar{x}^2 & \bar{x} \\ \bar{x} & 1 \end{bmatrix} > 0.$$

En résumé la droite des moindres carrés de y par rapport à x est définie par :

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

En inversant le rôle des variables, la droite des moindres carrés de x par rapport à y minimise la somme des carrés des distances des points à la droite mesurées, dans le même repère, horizontalement. Le critère s'écrit,

$$D_{x/y}(c, d) = \frac{1}{n} \sum_{i=1}^n [x_i - cx_i - d]^2,$$

et sa solution est donnée par :

$$x = \hat{c}y + \hat{d}, \quad \hat{c} = \frac{\text{cov}(x, y)}{\text{var}(y)}, \quad \hat{d} = \bar{x} - \hat{c}\bar{y}.$$

Ces deux droites se coupent au point moyen (\bar{x}, \bar{y}) , elles présentent le même sens de variation, donné par le signe de la covariance, et la première est moins inclinée que la seconde car le rapport des pentes est inférieur à un,

$$0 \leq \frac{\text{cov}^2(x, y)}{\text{var}(x)\text{var}(y)} \leq 1.$$

1.1.2 Droite des moindres distances

La *droite des moindres distances* accorde la même importance aux deux variables car la distance de (x_i, y_i) à la droite est mesurée selon le sens usuel, c'est-à-dire de façon perpendiculaire. Dans ce cas le repère doit être orthonormé. On parle aussi de *droite de régression orthogonale*. Le critère à minimiser est (on utilise la relation $\cos^2 \theta = \frac{1}{1+\tan^2 \theta}$) :

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

La droite passe par le point moyen (\bar{x}, \bar{y}) et par suite $b = \bar{y} - a\bar{x}$. Ainsi a doit minimiser la quantité,

$$D_{x,y}(a, \bar{y} - a\bar{x}) = \frac{1}{(1+a^2)} [var(y) + a^2 var(x) - 2acov(x, y)].$$

La dérivée par rapport à a ,

$$\frac{\partial}{\partial a} D_{x,y}(a, \bar{y} - a\bar{x}) = \frac{2}{(1+a^2)^2} \{a^2 cov(x, y) + a[var(x) - var(y)] - cov(x, y)\},$$

s'annule en les valeurs,

$$\frac{var(y) - var(x) \pm \sqrt{[var(y) - var(x)]^2 + 4cov^2(x, y)}}{2cov(x, y)}$$

qui, reportées partiellement dans le critère, donnent :

$$D_{x,y}(a, \bar{y} - a\bar{x}) = var(x) \pm \frac{-1}{(1+a^2)} \sqrt{[var(y) - var(x)]^2 + 4cov^2(x, y)}.$$

Ainsi la solution est :

$$\hat{a} = \frac{var(y) - var(x) + \sqrt{[var(y) - var(x)]^2 + 4cov^2(x, y)}}{2cov(x, y)}.$$

L'autre solution correspond à la droite orthogonale à celle-ci puisque le produit des racines est égal à -1. Dans ce cas la somme des carrés des distances est maximum sous la contrainte que la droite passe par le point moyen. Ces directions correspondent aux deux composantes de l'*analyse en composantes principales* (ACP) d'un nuage de points dans le plan. La droite de régression orthogonale donne le sous-espace de dimension un du plan sur lequel le nuage projeté a la plus grande dispersion, c'est-à-dire conserve la plus grande

information sur le couple (x, y) dans une seule variable.

La droite des moindres distances a le même sens de variation, donné par le signe de $cov(x, y)$, que les droites des moindres carrés. On montre qu'elle se situe dans l'angle aigu formé par ces deux droites. Pour cela il faut étudier la position des deux pentes de ces droites par rapport aux racines de l'équation du second degré de la situation orthogonale.

1.1.3 Droite des moindres rectangles

Comme dans le cas précédent, la *droite des moindres rectangles* prend en compte les deux variables de façon symétrique. La proximité de (x_i, y_i) à la droite est mesurée par la surface du rectangle défini de la façon suivante : ses côtés sont parallèles aux axes, (x_i, y_i) constitue un premier sommet et les deux sommets qui lui sont adjacents sont situés sur la droite. Le repère doit être orthonormé pour que la surface apparente soit bien celle que l'on minimise. Le critère est :

$$D_{xy}(a, b) = \frac{1}{n} \sum_{i=1}^n |y_i - ax_i - b| \times |x_i - \frac{1}{a}(y_i - b)| = \frac{1}{|a|} \times \frac{1}{n} \sum_{i=1}^n (y_i - ax_i - b)^2.$$

Là encore la droite passe par le point moyen (\bar{x}, \bar{y}) et donc $b = \bar{y} - a\bar{x}$. Puis a doit minimiser la quantité,

$$D_{xy}(a, b) = \frac{1}{|a|} [\text{var}(y) + a^2 \text{var}(x) - 2acov(x, y)].$$

Cette fonction présente deux minima locaux pour $a = \pm \sqrt{\frac{\text{var}(y)}{\text{var}(x)}}$ et le minimum global est obtenu lorsque a a le signe de $cov(x, y)$. Elle se situe également dans l'angle aigu formé par les deux droites des moindres carrés, sa pente étant la moyenne géométrique des deux autres pentes (en module),

$$\sqrt{\frac{\text{var}(y)}{\text{var}(x)}} = \left[\frac{\text{cov}(x, y)}{\text{var}(x)} \times \frac{\text{var}(y)}{\text{cov}(x, y)} \right]^{1/2}.$$

Elle est proche de la droite de régression orthogonale et coïncide avec elle lorsque $\text{var}(y) = \text{var}(x)$, la pente étant égale à ± 1 selon le signe de la covariance.

1.1.4 Optimisation numérique

Le critère obtenu en considérant la somme des carrés des écarts verticaux et horizontaux,

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

n'admet pas de solution explicite mais peut être résolu de façon numérique. L'introduction du point moyen (\bar{x}, \bar{y}) conduit à :

$$\begin{aligned} D_{x|y}(a, b) &= \text{var}(y) + a^2 \text{var}(x) - 2acov(x, y) + (\bar{y} - a\bar{x} - b)^2 \\ &+ \text{var}(x) + \frac{1}{a^2} \text{var}(y) - \frac{2}{a} cov(x, y) + \frac{1}{a^2} (\bar{y} - a\bar{x} - b)^2 \end{aligned}$$

La droite cherchée passe donc par le point moyen (\bar{x}, \bar{y}) et sa pente a doit minimiser la fonction $f(a)$ suivante :

$$D_{x|y}(a, b) - \text{var}(x) - \text{var}(y) = a^2 \text{var}(x) + \frac{1}{a^2} \text{var}(y) - 2acov(x, y) - \frac{2}{a} cov(x, y).$$

La minimisation numérique consiste à prendre le minimum de la parabole tangente au point d'abscisse a_0 , ce qui produit un point d'abscisse a_1 , puis d'itérer le procédé. On obtient ainsi l'algorithme suivant :

$$a_0 = \frac{1}{2} \left(\frac{cov(x, y)}{\text{var}(x)} + \frac{cov(x, y)}{\text{var}(y)} \right), \quad a_{k+1} = a_k - \frac{f'(a_k)}{f''(a_k)}, \quad k = 0, 1, \dots$$

On montre que la droite est également située dans l'angle aigu formé par les deux droites des moindres carrés. En utilisant le critère normalisé par les variances,

$$\frac{1}{\text{var}(y)} \times \frac{1}{n} \sum_{i=1}^n [y_i - ax_i - b]^2 + \frac{1}{\text{var}(x)} \times \frac{1}{n} \sum_{i=1}^n \left[x_i - \frac{1}{a}(y_i - b) \right]^2,$$

on retrouve la droite des moindres rectangles.

1.1.5 Exemple

Un ensemble de points à coordonnées entières, pour simplifier les calculs, permet d'illustrer les différentes droites introduites ci-dessus. Les données sont les suivantes :

Indice	i	1	2	3	4	5	6	7	8	9	10
Variable X	x_i	1	2	3	5	7	9	11	13	14	15
Variable Y	y_i	3	1	5	2	6	4	7	9	8	5

Tous les paramètres calculés ne dépendent des données qu'à travers le résumé numérique suivant :

$$n = 10; \quad \sigma x = 80; \quad \Sigma x^2 = 880; \quad \Sigma y = 50; \quad \sigma y^2 = 310; \quad \sigma xy = 489.$$

En particulier, les caractéristiques empiriques usuelles s'en déduisent immédiatement :

$$\bar{x} = 8; \quad \text{var}(x) = 24; \quad \bar{y} = 5; \quad \text{var}(y) = 6; \quad \text{cov}(x, y) = 8,9.$$

Les équations des différentes droites sont alors données par :

Type	Critère	Équation
moindres carrés	$D_{y/x}$	$y = 0,37x + 2,0$
moindres carrés	$D_{x/y}$	$y = 0,67x - 0,4$
moindres carrés	$D_{x y}$	$y = 0,57x + 0,4$
moindres distances	$D_{x,y}$	$y = 0,41x + 1,7$
moindres rectangles	D_{xy}	$y = 0,50x + 1,0$

N.B. L'équation $y = 0,67x - 0,4$ équivaut à $x = 1,48y + 0,6$.

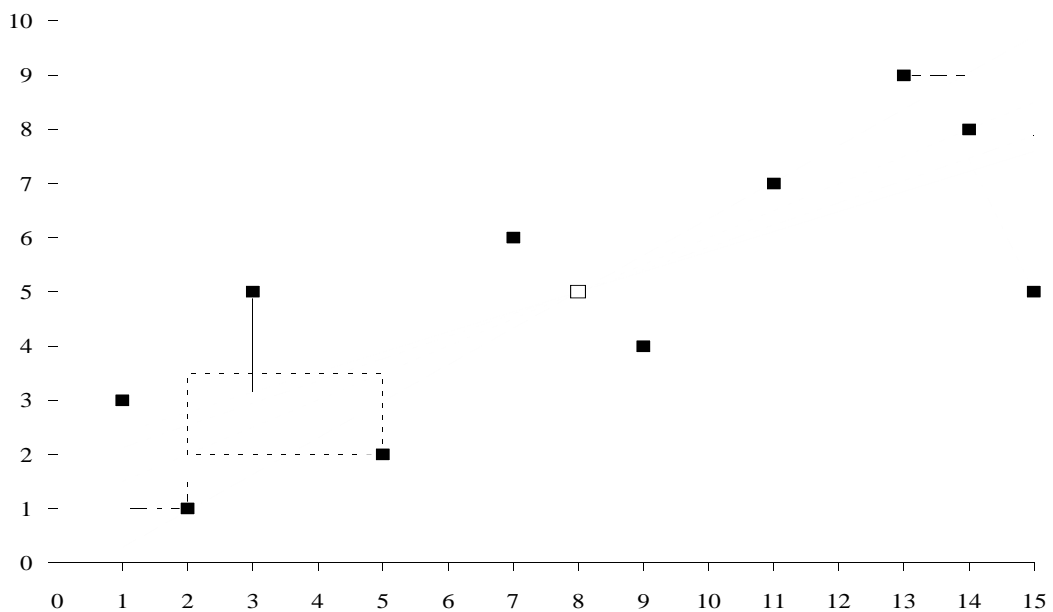


Figure 1.2: Droites associées à un ensemble de points selon différents critères

1.2 DROITE DE RÉGRESSION LINÉAIRE

Le contexte de la régression linéaire simple est présenté en première section à travers une illustration qui sera reprise tout au long de ce chapitre. Les hypothèses du modèle ne sont précisées qu'à la section suivante. La droite des moindres carrés constitue la base de l'estimation des paramètres et fait l'objet de la dernière section ainsi que la présentation du coefficient de corrélation linéaire empirique.

1.2.1 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 cardiaque. On a mesuré chez $n = 18$ patients, le poids x_i en kg et la vitesse coronarienne $y_i, i = 1, \dots, n$. Les données indiquées dans le Tableau 1.1 sont représentées sur la Figure 1.3. Ce graphique fait apparaître une dépendance linéaire entre les deux variables. Celle-ci est formalisée en postulant que la *fonction de régression* de Y à X est linéaire : $\mathbb{E}(Y|X = x) = ax + b$. Les deux variables sont donc appréhendées de façon différente. La variable Y conserve son caractère aléatoire alors que X est supposée parfaitement connue : on raisonne conditionnellement à X . 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 par exemple 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

Tableau 1.1: Vitesse coronarienne y_i et poids x_i de 18 patients

Résumé numérique, caractéristiques empiriques et premiers résultats

$$n = 18; \quad \sigma_x = 1153; \quad \Sigma x^2 = 76903; \quad \Sigma y = 1266; \quad \sigma_y^2 = 89508; \quad \sigma_{xy} = 79947.$$

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

<i>Droite de régression</i>	$y = -0,377x + 94,5$
<i>Coefficient de corrélation linéaire</i>	$r = -0,96$
<i>Coefficient de détermination</i>	$r^2 = 93\%$
<i>Variance estimée de l'erreur</i>	$\hat{\sigma}^2 = 2,12$

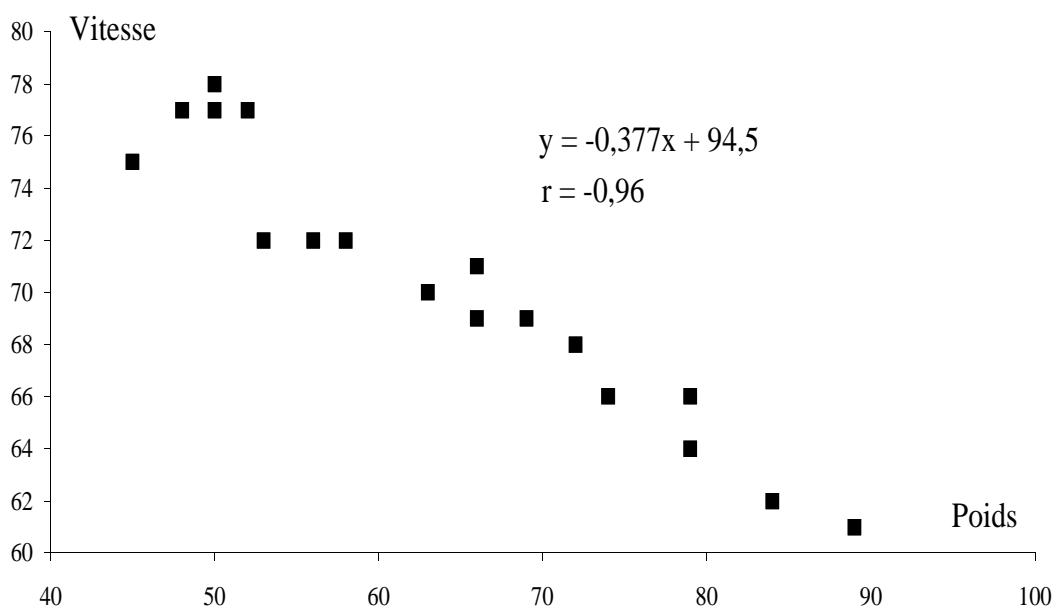


Figure 1.3: Vitesse coronarienne en fonction du poids chez 18 patients

1.2.2 Les hypothèses du modèle

Les données d'un modèle de régression linéaire simple sont constituées de n couples de réels (x_i, y_i) dont l'origine peut être très diverse selon la nature de la variable explicative X . Dans l'exemple ci-dessus X est aléatoire si les patients ont été sélectionnés au hasard mais il est possible qu'on ait veillé à choisir des patients de poids très différents. Plus généralement il est des situations où les valeurs de X sont fixées par l'expérimentateur, c'est pourquoi on dit que ces valeurs constituent le plan d'expérience. Par exemple x_1, x_2, \dots peuvent être les instants d'observation d'une grandeur Y dont on étudie l'évolution au cours du temps (série chronologique). Dans tous les cas on considère que les valeurs de X sont connues et non aléatoires. L'aspect aléatoire porte donc uniquement sur la variable expliquée Y . Ainsi y_1, y_2, \dots sont considérés comme les observations de variables aléatoires Y_1, Y_2, \dots de la forme,

$$Y_i = ax_i + b + \varepsilon_i, \quad i = 1, \dots, n,$$

où :

- le *plan d'expérience* x_1, x_2, \dots, x_n est fixé connu,
- les *paramètres* a et b sont inconnus,

- les *erreurs* $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ sont des variables aléatoires centrées, de même variance σ^2 et non corrélées entre elles :

$$\mathbb{E}(\varepsilon_i) = 0; \quad \text{Var}(\varepsilon_i) = \sigma^2; \quad \text{Cov}(\varepsilon_i, \varepsilon_j) = 0 \quad \text{si } i \neq j.$$

Pour les problèmes d'intervalles de confiance ou de tests, au dernier paragraphe, on fait l'hypothèse supplémentaire que les erreurs $\varepsilon_i, i = 1, \dots, n$ sont gaussiennes et donc indépendantes. Dans ce cas la loi de probabilité des observations associées au modèle linéaire est définie par la densité :

$$f(y_1, \dots, y_n; a, b, \sigma^2) = \frac{1}{(\sqrt{2\pi\sigma^2})^n} \exp \left\{ - \sum_{i=1}^n \frac{(y_i - ax_i - b)^2}{2\sigma^2} \right\}.$$

Les variables $Y_i, i = 1, \dots, n$ sont gaussiennes, indépendantes, de même variance σ^2 mais de moyennes différentes de la forme $\mathbb{E}(Y_i) = ax_i + b, i = 1, \dots, n$. On notera que le plan d'expérience x_1, x_2, \dots, x_n ne fait pas partie des paramètres du modèle.

On peut écrire le modèle sous forme vectorielle :

$$Y = ax + b\mathbb{I} + \varepsilon, \quad \mathbb{E}(\varepsilon) = 0, \quad \text{Var}(\varepsilon) = \sigma^2 I,$$

où

$$Y = (Y_1, \dots, Y_n)^T, \quad x = (x_1, \dots, x_n)^T, \quad \mathbb{I} = (1, \dots, 1)^T, \quad \varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^T,$$

ce qui se traduit sur Y par :

$$\mathbb{E}(Y) = ax + b\mathbb{I}, \quad \text{Var}(Y) = \sigma^2 I.$$

L'*espace des variables* (ici \mathbb{R}^2) permet de représenter les données $(x_i, y_i), i = 1, \dots, n$ dans les axes, choisis orthogonaux, associés aux variables X et Y (cf. Figure 1.3). L'*espace des observations* (ici \mathbb{R}^n) est celui dans lequel Y prend ses valeurs. Il permet de "représenter" l'observation $y = (y_1, \dots, y_n)^T$, le sous-espace engendré par x et \mathbb{I} dans lequel se situe la moyenne $\mathbb{E}(Y) = ax + b\mathbb{I}$, appelé *espace des moyennes*, ainsi que le terme d'erreur ε (cf. Figure 1.4).

1.2.3 Estimateur des moindres carrés

Si le paramètre (a, b) était connu, les erreurs ε_i seraient observables et données par $\varepsilon_i = y_i - ax_i - b, i = 1, \dots, n$. L'estimation de ce paramètre par la *méthode des moindres carrés* consiste à retenir la valeur de (a, b) pour

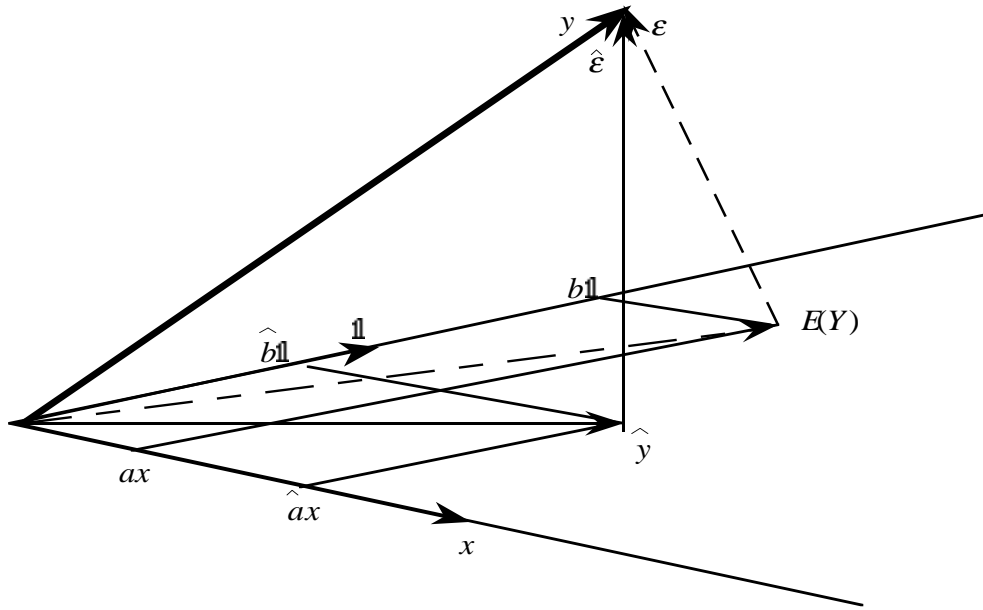


Figure 1.4: Régression linéaire simple dans l'espace des observations

laquelle la somme 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 au Paragraphe 1.1, associée au critère :

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

On dit que les coefficients,

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

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 somme 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. Le critère des moindres carrés est donc justifié par l'hypothèse d'erreurs centrées, de même variance, et non corrélées entre elles. Les propriétés statistiques de l'estimateur (\hat{a}, \hat{b}) seront étudiées par la suite.

1.2.4 Coefficient de corrélation linéaire empirique

La variance empirique de Y se décompose sous la forme :

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

Le premier terme, 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*,

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

cette décomposition s'écrit

$$\text{var}(y) = (1 - r^2)\text{var}(y) + r^2\text{var}(y).$$

Ainsi r^2 , appelé *coefficient de détermination*, représente la part de variance (empirique) de Y expliquée par la régression 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, $y_i = \hat{a}x_i + \hat{b}$, $i = 1, \dots, n$). On remarque que r est symétrique par rapport aux deux variables mais les deux droites de régression 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 :

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

Dans l'exemple de la section 1.2.1, 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 (*cf.* section 1.2.6). Dans l'exemple introductif, la corrélation entre le poids et la taille est faible puisque r est de l'ordre de 0,5. De plus la faible taille de l'échantillon ne permet pas d'apprécier à sa juste valeur la dépendance entre ces deux variables ; nous précisons ce point avec l'exemple des conscrits traité dans le Paragraphe 1.3.

1.2.5 Analyse descriptive des résidus

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

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

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. On donne ci-dessous quelques exemples simples de situations structurées pour lesquelles les hypothèses du modèle linéaire ne sont certainement pas satisfaites.

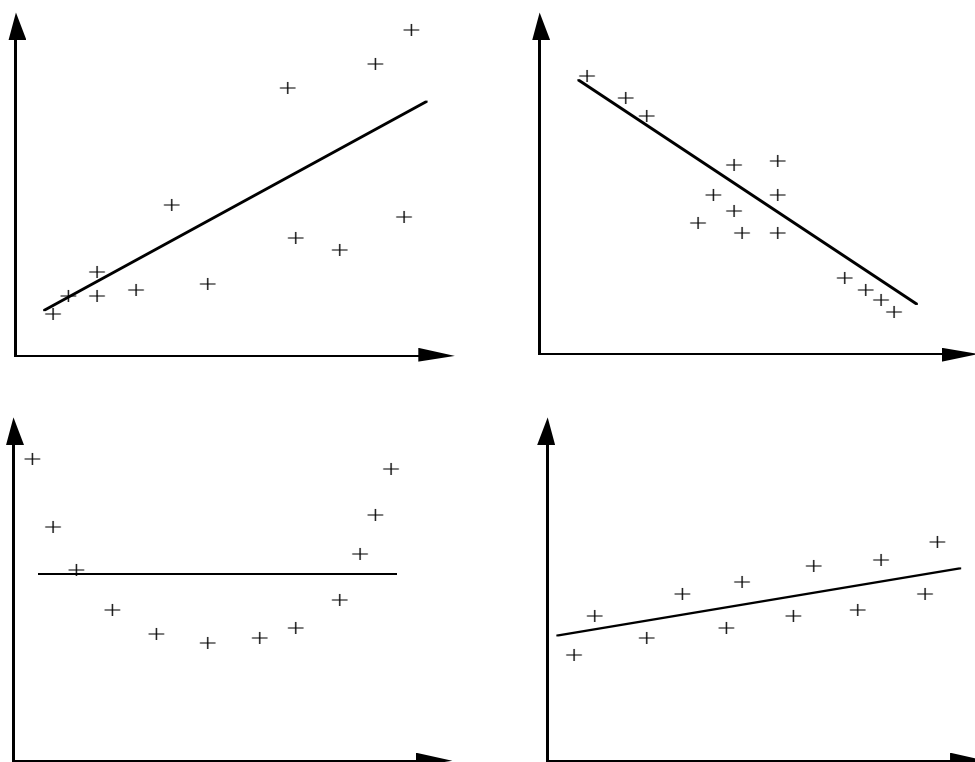


Figure 1.5: Exemples de données structurées

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 variance 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 1.2 et Figure 1.6). 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

Tableau 1.2: Valeurs ajustées et résidus pour la vitesse coronarienne

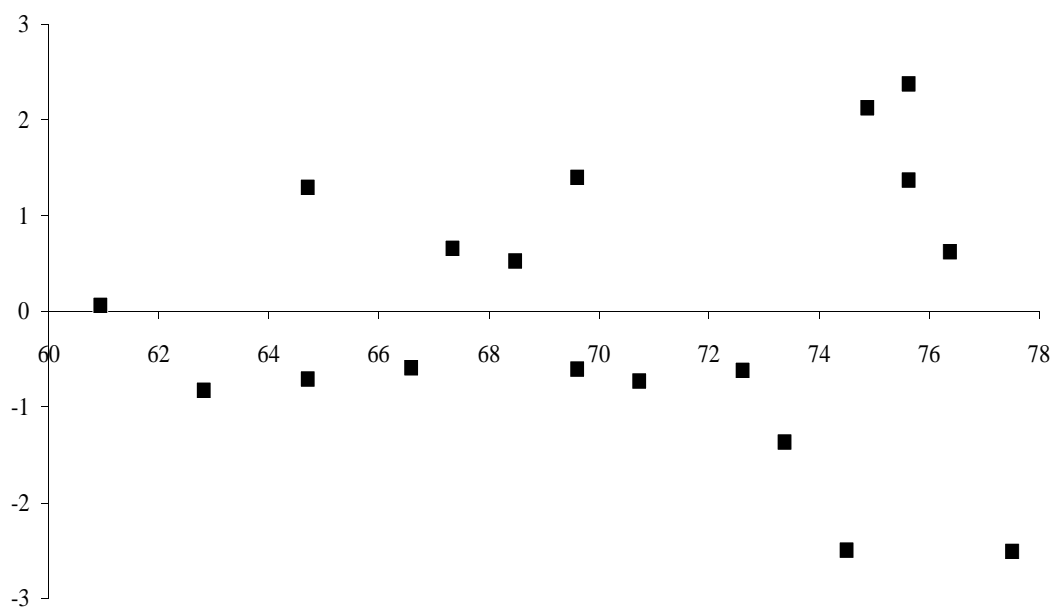


Figure 1.6: Résidus en fonction des valeurs ajustées pour la vitesse coronarienne

1.2.6 Exemples

Le Tableau 1.3 donne le chiffre d'affaires moyen et la marge moyenne des entreprises du secteur “réparation et commercialisation de l'automobile” selon le nombre de salariés au 31 décembre 1988 (source INSEE). Les sommes

sont exprimées en milliers de francs (KF). Les résultats concernant cet exemple sont regroupés dans le Tableau 1.4.

<i>Nombre de salariés</i>	<i>Chiffre d'affaires</i>	<i>Marge</i>
hors tranche	1462	319
0 salarié	612	213
1 salarié	784	343
2 salariés	1308	495
3 à 5 salariés	2070	765
6 à 9 salariés	4204	1551
10 à 19 salariés	10363	3066
20 à 49 salariés	37386	8029
50 à 99 salariés	95874	18736
100 à 199 salariés	242496	48768
200 salariés et plus	1822758	345893

Tableau 1.3: Chiffre d'affaires et marge des entreprises

En représentant toutes les entreprises, on ne distingue pas les points proche de l'origine, mais on constate l'alignement (*cf.* Figure 1.7). La représentation sans les quatre derniers points permet de montrer que le modèle n'est pas adapté car tous les premiers points sont au-dessous de la droite (*cf.* Figure 1.8). On doit donc distinguer les petites entreprises (de 0 à 9 salariés) (*cf.* Figure 1.9) des moyennes ou grandes (10 salariés ou plus) (*cf.* Figure 1.10). Par ailleurs la catégorie "hors tranche" apparaît comme un point aberrant. Malgré la taille très faible des échantillons, les données sont très significative car ce sont déjà des moyennes calculées sur l'ensemble des entreprises de chaque catégorie. Par contre la variance de l'erreur attachée à chaque point n'est sans doute pas constante. Elle est cependant si faible par rapport à la variation de la marge que l'interprétation des pentes (taux de marge) conserve tout son sens. Ce taux est proche de 40% dans les petites entreprises et de 20% dans les moyennes ou grandes.

De plus, devant l'interprétation de b , il est naturel d'ajuster dans chaque situation le modèle associé à une droite passant par l'origine, $y_i = ax_i + \varepsilon_i$, $i = 1, \dots, n$. Le critère s'écrit :

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

et la solution est donnée par :

$$\hat{a} = \frac{\overline{xy}}{\overline{x^2}} = \frac{\sigma_{xy}}{\sigma_{x^2}}.$$

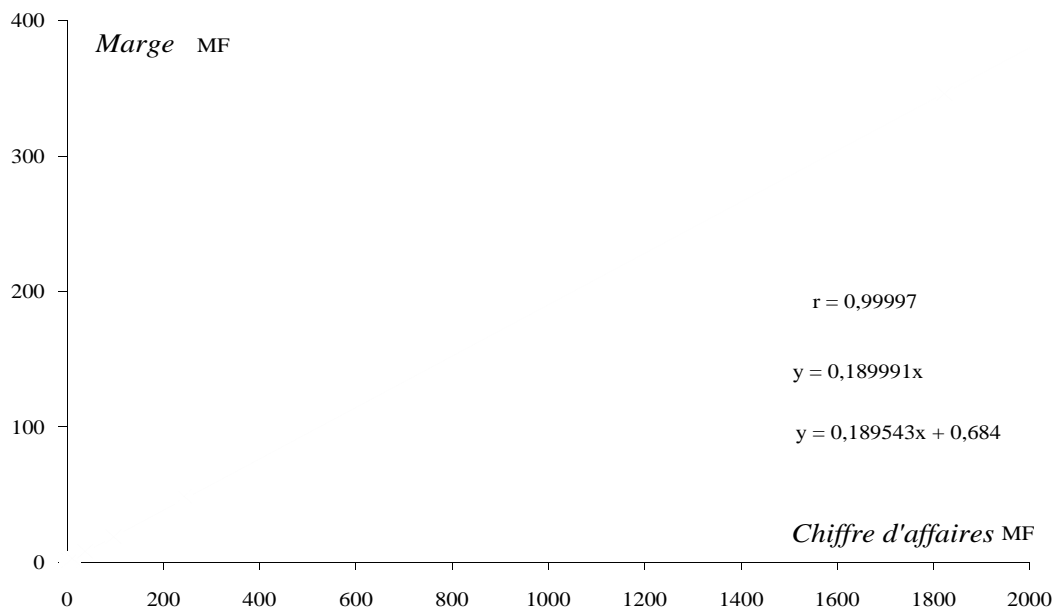


Figure 1.7: Marge en fonction du chiffre d'affaires de l'ensemble des entreprises, unité = 1MF

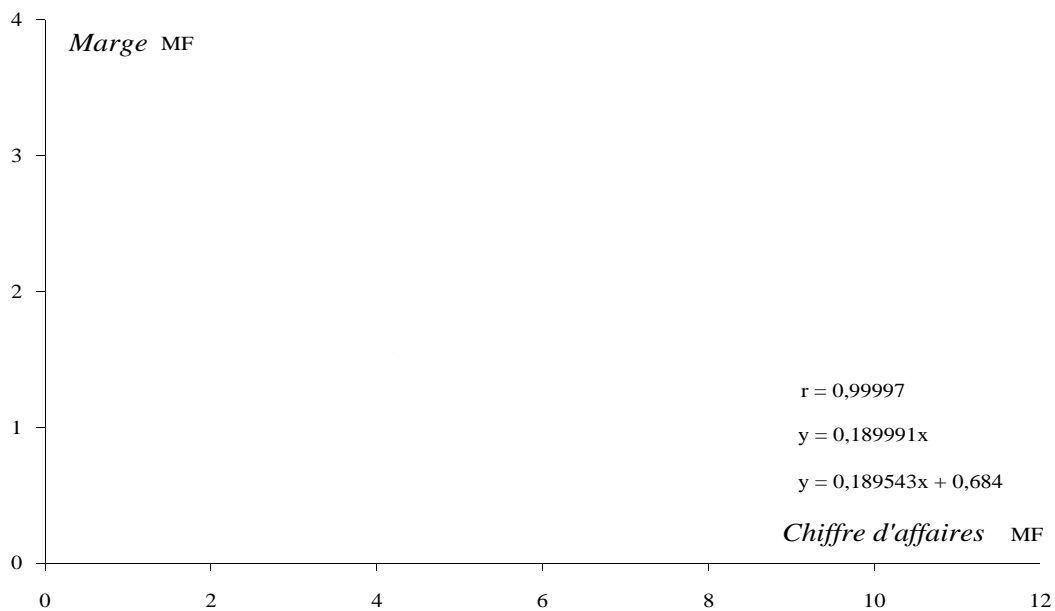


Figure 1.8: Zoom sur la marge en fonction du chiffre d'affaires de l'ensemble des entreprises, unité = 1MF

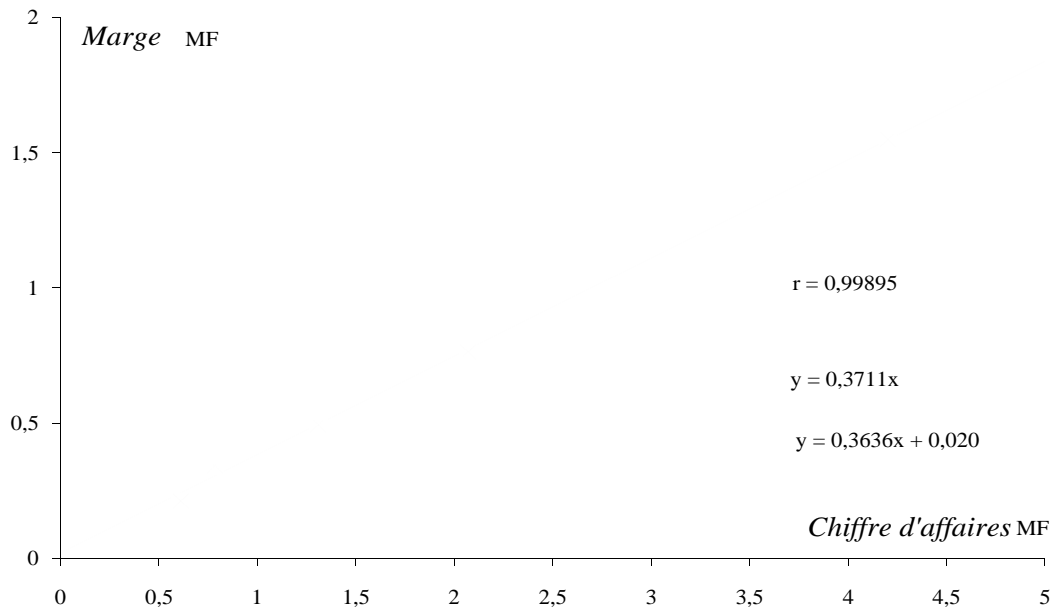


Figure 1.9: Marge en fonction du chiffre d'affaires des petites entreprises, unité = 1MF

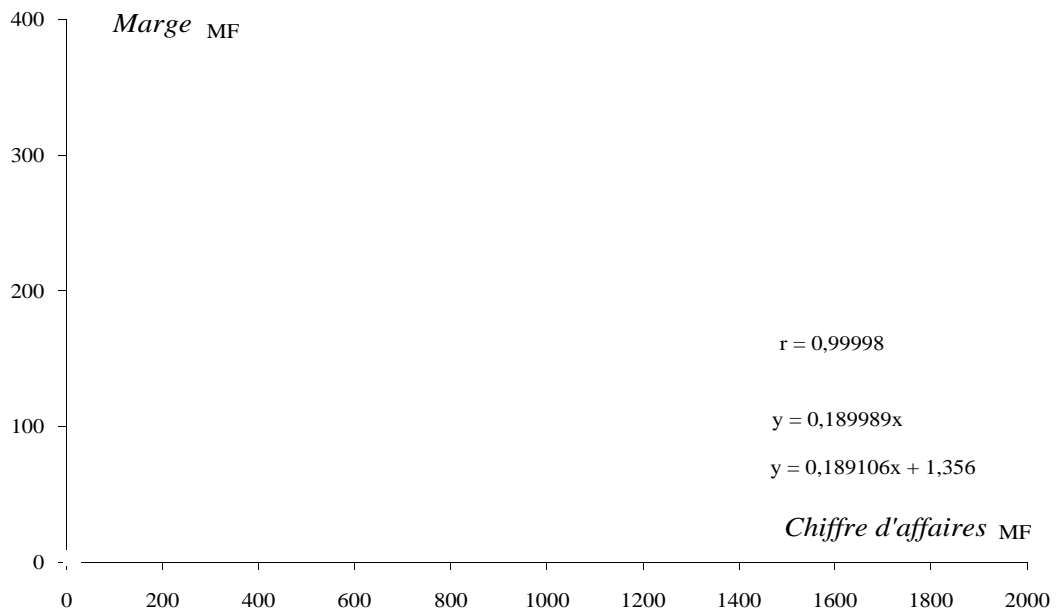


Figure 1.10: Marge en fonction du chiffre d'affaires des moyennes ou grandes entreprises, unité = 1MF

Résultat	Ensemble	Petites	Moyennes ou grandes
n	11	5	5
σx	2219317	8978	2208877
σx^2	$3,391974761 \times 10^{12}$	24658580	$3,391974761 \times 10^{12}$
σy	428178	3367	424492
σy^2	$1,224486887 \times 10^{11}$	3398869	$1,224451882 \times 10^{11}$
σxy	$6,444431352 \times 10^{11}$	9150682	$6,444335182 \times 10^{11}$
$y = \hat{a}x + \hat{b}$	$y = 0,189543x + 684$	$y = 0,3636x + 20$	$y = 0,189106x + 1356$
r	0,99997	0,99990	0,99998
$y = \hat{a}x$	$y = 0,189991x$	$y = 0,3711x$	$y = 0,189989x$

Tableau 1.4: Résultats sur les entreprises

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 1.5 indique, pour les années 1927 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).

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

Tableau 1.5: Exemple de non-sens

Résumé numérique, caractéristiques empiriques et résultats

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

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

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

$$\begin{array}{ll} \text{Droites de régression} & y = 2,20x + 4,6 \quad y = 6,0z - 26,6 \\ \text{Coefficients de corrélation linéaire} & r_{xy} = 0,992 \quad r_{zy} = 0,94 \end{array}$$

On considère la régression linéaire du nombre de déficients mentaux en fonction du nombre de licences radio, puis en fonction du nombre de lettres

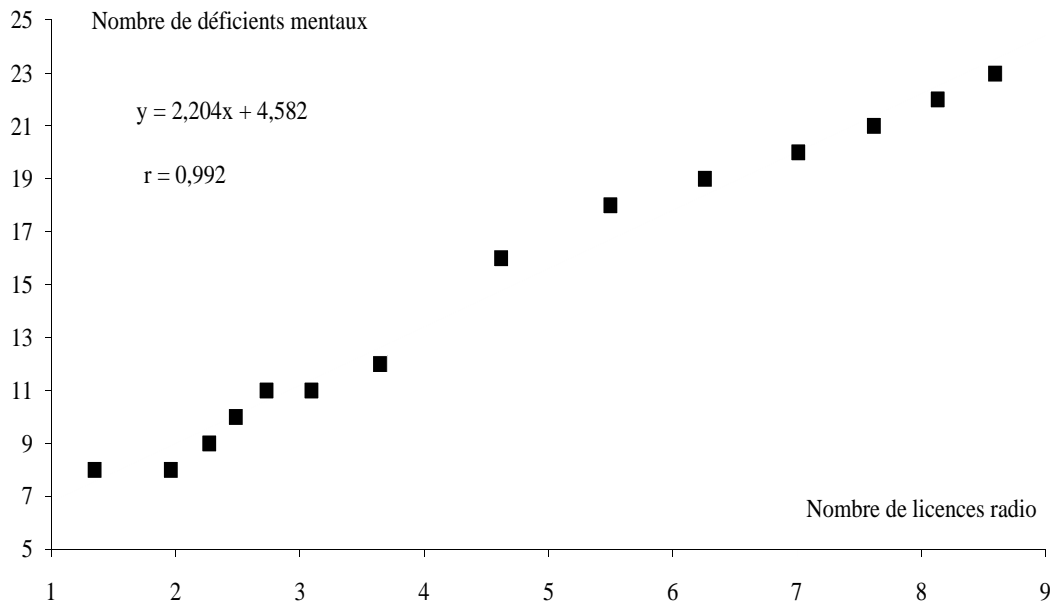


Figure 1.11: Déficience mentale en fonction des licences radio

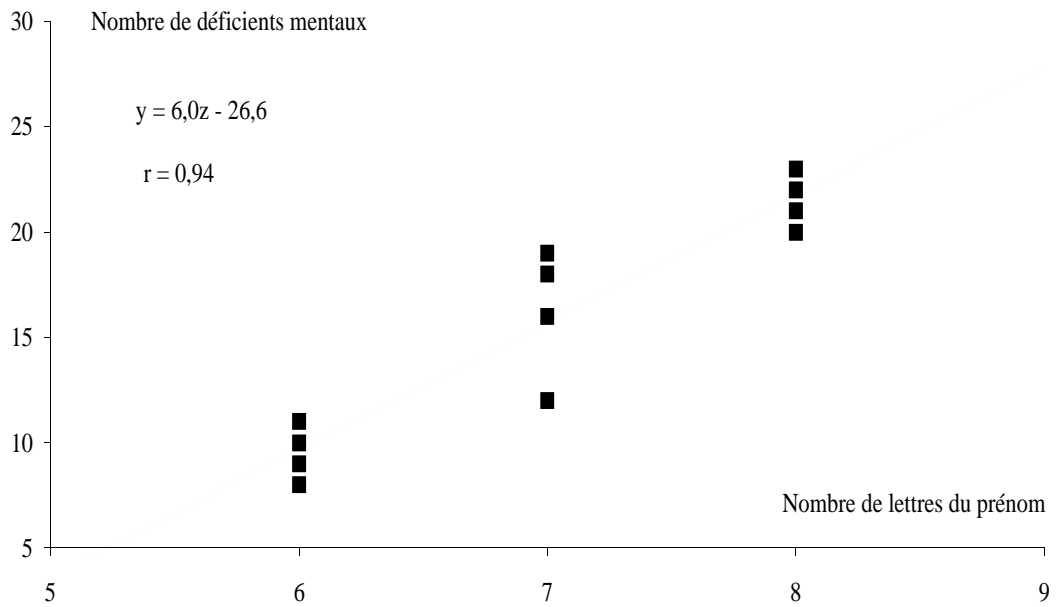


Figure 1.12: Déficience mentale en fonction du prénom du président

du prénom du Président. Dans les deux cas, le coefficient de corrélation est très élevé et les points sont bien répartis autour de la droite (cf. Figures 1.11 et 1.12).

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).

1.3 COURBES DE RÉGRESSION

Nous considérons ici une approche très ordinaire de la régression non linéaire dans un cadre purement descriptif. Il s'agit de situations dans lesquelles plusieurs observations de la variable d'intérêt Y sont associées, de façon naturelle ou conventionnelle, à une même valeur x de la variable explicative X . C'est par exemple le cas d'un plan d'expérience dans lequel, pour chaque niveau $x_i, i = 1, \dots, k$ du facteur X , sont effectuées les mesures $y_{ij}, j = 1, \dots, n_i$ de la grandeur Y . Plus généralement les données sont en nombre très important et forment une table de contingence $\{(x_i, y_j; n_{ij}), i = 1, \dots, k; j = 1, \dots, p\}$. Dans le cas de variables continues, les observations sont donc assimilées par convention aux centres des classes. On donne en dernière section trois exemples illustrant ces diverses situations. Notons que la première situation peut se formaliser dans le cadre d'une table de contingence avec $n_{ij} \in \{0, 1\}$.

1.3.1 Courbe des moindres carrés

On se place dans le cadre d'une *table de contingence* pour notre présentation.

On rappelle les notations des caractéristiques conditionnelles empiriques.

Pour $i = 1, \dots, k$, on définit les :

- *moyennes conditionnelles*

$$\bar{y}_i = \frac{1}{n_i} \sum_{j=1}^p n_{ij} y_j = \frac{1}{f_i} \sum_{j=1}^p f_{ij} y_j = \sum_{j=1}^p f_j^i y_j,$$

- *variances et écarts-types conditionnels*

$$\tilde{s}_i^2(y) = \frac{1}{n_{ij}} \sum_{j=1}^p n_{ij} (y_j - \bar{y}_i)^2 = \frac{1}{f_i} \sum_{j=1}^p f_{ij} (y_j - \bar{y}_i)^2 = \sum_{j=1}^p f_j^i (y_j - \bar{y}_i)^2,$$

X	Y	y_1		y_j		y_p	Ensemble
x_1		n_{11}		n_{1j}		n_{1p}	$n_{1.}$
x_i		n_{i1}		n_{ij}		n_{ip}	$n_{i.}$
x_k		n_{k1}		n_{kj}		n_{kp}	$n_{k.}$
Total		$n_{.1}$		$n_{.j}$		$n_{.p}$	n

Tableau 1.6: Table de contingence

$$\tilde{s}_i(y) = \sqrt{\tilde{s}_i^2(y)}, \quad (\tilde{s}_i^2(y) = \overline{y_i^2} - (\overline{y_i})^2).$$

On peut noter que le calcul des caractéristiques conditionnelles peut s'effectuer à partir des effectifs n_{ij} ou des fréquences f_{ij} de la distribution conjointe ou en utilisant les distributions conditionnelles f_j^i . La distribution marginale de Y est la moyenne, pondérée par la distribution marginale de X , des distributions conditionnelles de Y sachant $X = x_i, i = 1, \dots, k$:

$$f_{.j} = \sum_{i=1}^k f_i^j \times f_{i.}, \quad j = 1, \dots, p.$$

Ceci a pour conséquence les relations permettant de calculer les caractéristiques marginales en fonction des caractéristiques conditionnelles et de la loi marginale de la variable qui conditionne :

- La moyenne marginale \overline{y} de Y est égale à la moyenne, pondérée par la distribution marginale de X ($f_{i.}, i = 1, \dots, k$), des moyennes conditionnelles $\overline{y_i}$ de Y sachant $X = x_i, i = 1, \dots, k$:

$$\overline{y} = \frac{1}{n} \sum_{i=1}^k n_{i.} \overline{y_i} = \sum_{i=1}^k f_{i.} \overline{y_i}.$$

- La variance marginale $\tilde{s}^2(y)$ de Y est égale à la moyenne, pondérée par la distribution marginale de X ($f_{i.}, i = 1, \dots, k$), des variances conditionnelles $\tilde{s}_i^2(y)$ de Y sachant $X = x_i, i = 1, \dots, k$, augmentée de la variance, calculée selon la loi de pondération, des moyennes conditionnelles $\overline{y_i}, i = 1, \dots, k$:

$$\tilde{s}^2(y) = \sum_{i=1}^k f_{i.} \tilde{s}_i^2(y) + \sum_{i=1}^k f_{i.} (\overline{y_i} - \overline{y})^2.$$

Il s'agit de la décomposition de la variance d'un mélange de populations, lorsque la variable X est de nature qualitative, en *var intra* et *var inter* et que l'on retrouve en analyse de la variance.

On appelle *courbe de régression* de Y en x la courbe représentative des moyennes conditionnelles \bar{y}_i en fonction des valeurs $x_i, i = 1, \dots, k$, de la *variable de liaison* X .

La courbe se présente sous forme d'un ensemble de points (*cf.* Figure 1.13). Un tracé régulier joignant les points $(x_i, \bar{y}_i), i = 1, \dots, k$, souligne la variation de Y en fonction de x . Lorsque la variable X est continue il peut s'interpréter comme une "estimation" de la courbe là où elle n'est pas connue. Par contre il n'a plus d'interprétation lorsque la variable X est discrète.

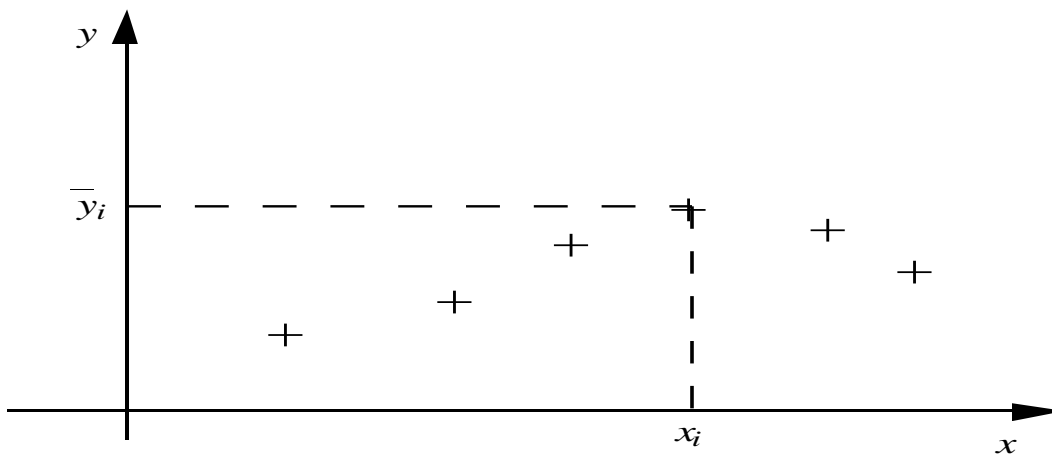


Figure 1.13: Courbe de régression

Cette courbe de régression s'interprète comme une courbe des moindres carrés. Il y a n_{ij} observations en le même point $M_{ij} = (x_i, y_j)$. La grosseur des points de la Figure 1.14 représente l'importance relative du nombre d'individus.

Considérons une valeur x_i fixée de la variable X . Le nombre d'observations présentant cette valeur est égal à n_i . Ils se répartissent, selon la variable Y , en :

$$n_{i1} \text{ en } y_1, \dots, n_{ij} \text{ en } y_j, \dots, n_{ip} \text{ en } y_p.$$

Les représentants M_{ij} sont sur la même verticale d'abscisse x_i . On cherche

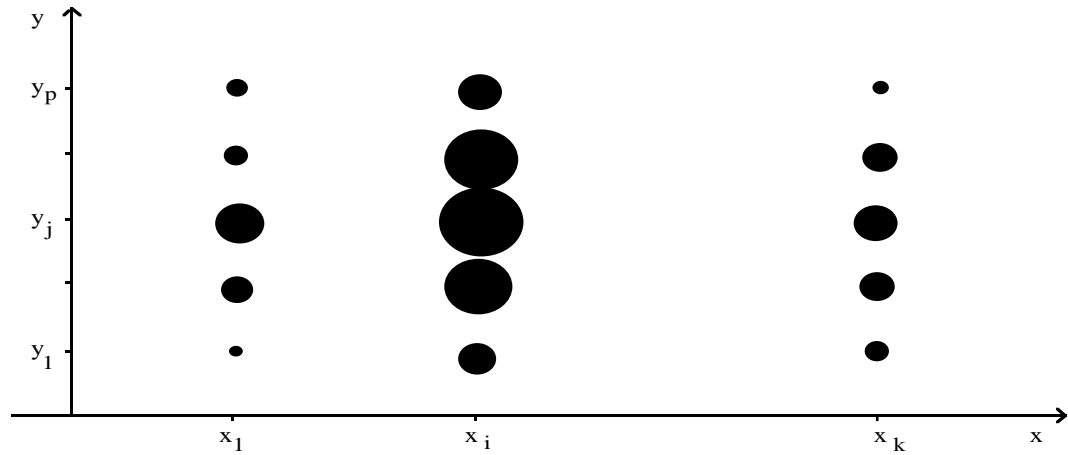


Figure 1.14: Courbe des moindres carrés

l'ordonnée y du point M , d'abscisse x_i , qui soit le plus proche d'eux selon le critère de l'écart quadratique moyen :

$$\min_y \frac{1}{n_i} \sum_{j=1}^p n_{ij} (y_j - y)^2.$$

On sait que la solution est donnée par la moyenne,

$$\bar{y}_i = \frac{1}{n_i} \sum_{j=1}^p n_{ij} y_j,$$

et que la valeur du minimum est égale à la variance,

$$\tilde{s}_i^2(y) = \frac{1}{n_i} \sum_{j=1}^p n_{ij} (y_j - \bar{y}_i)^2.$$

Cherchons la fonction $y = g(x)$ qui soit la plus proche de l'ensemble de la population selon ce même critère. Cette fonction est en fait définie par ses valeurs $g(x_1), \dots, g(x_i), \dots, g(x_k)$, en $x_1, \dots, x_i, \dots, x_k$. On doit ainsi trouver les valeurs $g(x_1), \dots, g(x_i), \dots, g(x_k)$ telles que l'écart quadratique moyen suivant soit minimum :

$$\frac{1}{n} \sum_{i=1}^k \sum_{j=1}^p n_{ij} [y_j - g(x_i)]^2 = \frac{1}{n} \sum_{i=1}^k \left\{ \frac{1}{n_i} \sum_{j=1}^p n_{ij} [y_j - g(x_i)]^2 \right\}.$$

Ce qui précède montre que la fonction la plus proche est la courbe de régression :

$$g(x_i) = \bar{y}_i, \quad i = 1, \dots, k.$$

En ce sens on peut dire que la courbe de régression est la courbe des moindres carrés.

Le minimum atteint est égal à :

$$\frac{1}{n} \sum_{i=1}^k n_i \left\{ \frac{1}{n_i} \sum_{j=1}^p n_{ij} [y_j - \bar{y}_{i.}]^2 \right\} = \sum_{i=1}^k f_i \tilde{s}_i^2(y) = \text{var intra}.$$

Il mesure la dispersion des observations autour de la courbe de régression. On peut imaginer l'utilisation d'autres critères :

- *écart absolu* $\rightarrow g(x_i) =$ médiane des observations associées à x_i ;
- *distance discrète* $\rightarrow g(x_i) =$ valeur correspondant à la plus grande fréquence des observations associées à x_i (mode si la distribution est unimodale ou celui associé à la plus grande fréquence).

Cependant il n'y a pas, pour ces critères, de décomposition de la variance de Y en la somme de la partie expliquée par X (*var inter*) et du reste (*var intra*).

Toutes ces notions se transposent immédiatement, avec des notations analogues, en inversant le rôle des variables. En général l'étude de la régression s'impose dans un seul sens. On considère ici les situations extrêmes, toujours d'un point de vue empirique, que sont l'indépendance et la liaison fonctionnelle.

- Cas de l'*indépendance*. On a les implications, réciproques ou non, suivantes :
 X et Y sont indépendantes \Leftrightarrow les distributions conditionnelles sont identiques (et donc égales à la distribution marginale de même nature) \Rightarrow les moyennes conditionnelles sont égales (et coïncident avec la moyenne marginale correspondante) \Leftrightarrow les courbes de régression sont des droites parallèles aux axes (et se coupent au point moyen (\bar{x}, \bar{y})) :
 $\bar{x}_{.j} = \bar{x}, j = 1, \dots, p; \bar{y}_{i.} = \bar{y}, i = 1, \dots, k.$

L'absence de réciproque à l'implication centrale ci-dessus est claire : les distributions conditionnelles peuvent avoir la même moyenne sans pour autant être identiques.

Lorsque les moyennes conditionnelles sont égales, on dit que la variable étudiée est *non corrélée* à la variable de liaison. Cela équivaut à ce que

la courbe de régression soit une droite horizontale, la variable de liaison étant placée en abscisse. Sous forme symbolique on a les définitions suivantes :

$$\begin{aligned} \bar{y}_i = \bar{y}, \quad i = 1, \dots, k &\Leftrightarrow Y \text{ est non corrélée avec } X, \\ \bar{x}_j = \bar{x}, \quad j = 1, \dots, p &\Leftrightarrow X \text{ est non corrélée avec } Y. \end{aligned}$$

La notion d'*absence de corrélation* n'est pas réciproque : Y peut ne pas être corrélée avec X alors que X est corrélée avec Y , la variable X peut même être en liaison fonctionnelle avec Y . L'indépendance est un cas particulier d'*absence réciproque de corrélation*.

- Cas de la *liaison fonctionnelle*. Lorsque Y est liée fonctionnellement à X ($X = x_i \Rightarrow Y = y_{\phi(i)}$), la courbe de régression de Y en x est confondue avec la courbe de liaison. Si la liaison fonctionnelle n'est pas réciproque, la courbe de régression de X en y est distincte de la précédente. Lorsqu'il y a *liaison fonctionnelle réciproque*, les courbes de liaison et les courbes de régression sont confondues en une seule courbe.

En pratique, la situation est souvent intermédiaire entre les deux cas extrêmes décrits ci-dessus, surtout si le nombre total d'observations est relativement faible. Cependant il est facile d'imaginer des situations d'indépendance : chez l'adulte, la taille ne dépend pas de l'âge. Par contre en période de croissance, on peut utiliser la courbe de régression de la taille (ou du poids) en fonction de l'âge pour surveiller la santé d'un nourrisson. En effet la régression permet d'établir, à l'aide d'un très grand nombre de données expérimentales, une relation vraie en moyenne entre deux grandeurs. Une relation fonctionnelle correspond plutôt à une courbe d'étalonnage dans un système complexe.

1.3.2 Rapports de corrélation

Lorsque la courbe de régression de Y en x n'est pas parallèle à l'axe des abscisses, on dit que la variable Y est *corrélée* avec la variable X ou Y est en corrélation avec X .

Considérons la décomposition de la variance empirique de Y :

$$\tilde{s}^2(y) = \sum_{i=1}^k f_i \tilde{s}_i^2(y) + \sum_{i=1}^k f_i (\bar{y}_i - \bar{y})^2.$$

La moyenne (pondérée) des variances conditionnelles, premier terme du second membre de l'égalité, que nous avons appelé *var intra* pour "variance intragroupe", représente la dispersion de Y "autour" de la courbe de régression. La variance des moyennes conditionnelles, deuxième terme du second membre de l'égalité, que nous avons appelé *var inter* pour "variance intergroupe", représente la dispersion de Y "le long" de la courbe de régression. Disons que la variance totale de Y , variance marginale $\tilde{s}^2(y)$, s'explique en partie par la variable X selon le terme de la "variance intergroupe",

$$\sum_{i=1}^k f_i (\bar{y}_i - \bar{y})^2,$$

variance due à des "observations moyennes" \bar{y}_i situées sur la courbe de régression en "nombre" $f_i, i = 1, \dots, k$. L'importance relative de ce terme est mesurée par le *rapport de corrélation* de Y en x , noté $\eta_{y/x}^2$,

$$\eta_{y/x}^2 = \frac{\sum_{i=1}^k f_i (\bar{y}_i - \bar{y})^2}{\tilde{s}^2(y)} = 1 - \frac{\sum_{i=1}^k f_i \tilde{s}_i^2(y)}{s^2(y)}.$$

Les deux expressions de $\eta_{y/x}^2$ permettent d'établir facilement les propriétés suivantes.

- Le rapport de corrélation est un nombre positif compris entre 0 et 1 :

$$0 \leq \eta_{y/x}^2 \leq 1.$$

Les deux situations extrêmes s'interprètent comme suit :

- $\eta_{y/x}^2 = 0 \Leftrightarrow Y$ est non corrélée avec X .
En effet $\eta_{y/x}^2 = 0$ si et seulement si $\bar{y}_i = \bar{y}, i = 1, \dots, k$, d'après l'expression de la variance intergroupe (la courbe de régression est une droite horizontale).
- $\eta_{y/x}^2 = 1 \Leftrightarrow Y$ est liée fonctionnellement à X .
En effet $\eta_{y/x}^2 = 1$ si et seulement si $\tilde{s}_i^2(y) = 0, i = 1, \dots, k$, d'après l'expression de la variance intragroupe. Or, pour chaque i de $1, \dots, k$, $\tilde{s}_i^2(y)$ ne peut être nul que si toutes les observations, associées à la valeur x_i de la variable X , présentent la même valeur $y_{\phi(i)}$ de la variable Y et alors $\bar{y}_i = y_{\phi(i)}$.

En échangeant les rôles des variables, on définit le rapport de corrélation de X en Y , noté $\eta_{x/y}^2$. Il n'a aucune raison d'être proche du précédent (il

peut être nul alors que l'autre est égal à 1).

Les variables peuvent être corrélées, sans être linéairement corrélées, puisque l'on a simplement les inégalités :

$$0 \leq r^2 \leq \min(\eta_{y/x}^2, \eta_{x/y}^2) \leq 1,$$

qui traduisent en particulier que la part de variance expliquée par la droite de régression est inférieure ou égale à celle expliquée par la courbe de régression (une droite est une courbe particulière et le critère est le même, *cf.* section suivante). Ainsi si Y est non corrélée avec X ($\eta_{y/x}^2 = 0$) ou si X est non corrélée avec Y ($\eta_{x/y}^2 = 0$), alors X et Y sont non linéairement corrélées ($r^2 = 0$). D'autre part, si les variables X et Y sont en liaison fonctionnelle réciproque selon une droite ($r^2 = 1$), alors Y est liée fonctionnellement à X ($\eta_{y/x}^2 = 1$) et X est liée fonctionnellement à Y ($\eta_{x/y}^2 = 1$).

1.3.3 Droites des moindres carrés pondérés

Les droites de régression définies au paragraphe précédent conservent tout leur sens ici. La présence de plusieurs observations n_{ij} au même point (x_i, y_j) ne modifie que l'écriture symbolique des différentes grandeurs considérées. Le critère définissant la droite de régression de Y par rapport à x s'écrit :

$$D_{Y/x}(a, b) = \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^p n_{ij} [y_j - ax_i - b]^2,$$

et la solution est donnée par :

$$y = \hat{a}x + \hat{b}, \quad \hat{a} = \frac{\text{cov}(x, y)}{s^2(x)}, \quad \hat{b} = \bar{y} - \hat{a}\bar{x},$$

où :

$$\begin{aligned} \bar{x} &= \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^p n_{ij} x_i = \frac{1}{n} \sum_{i=1}^k n_i x_i, \\ \bar{y} &= \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^p n_{ij} y_j = \frac{1}{n} \sum_{j=1}^p n_{.j} y_j = \frac{1}{n} \sum_{i=1}^k n_i \bar{y}_i, \\ \text{cov}(x, y) &= \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^p n_{ij} (x_i - \bar{x})(y_j - \bar{y}) = \frac{1}{n} \sum_{i=1}^k n_i (x_i - \bar{x})(y_i - \bar{y}), \\ \tilde{s}^2(y) &= \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^p n_{ij} (x_i - \bar{x})^2 = \frac{1}{n} \sum_{i=1}^k n_i (x_i - \bar{x})^2. \end{aligned}$$

On constate que la régression linéaire de Y par rapport à x est la droite des moindres carrés construite sur les points (x_i, \bar{y}_i) , $i = 1, \dots, k$, à condition de donner à chaque point une importance relative égale à f_i . C'est-à-dire que (\hat{a}, \hat{b}) est aussi la solution du problème :

$$\min_{a,b} \sum_{i=1}^k f_i [\bar{y}_i - ax_i - b]^2.$$

Introduisons encore le point moyen (\bar{x}, \bar{y}) , qui est aussi celui associé aux nouvelles données :

$$\bar{x} = \sum_{i=1}^k f_i x_i, \quad \bar{y} = \sum_{i=1}^k f_i \bar{y}_i.$$

On obtient :

$$\begin{aligned} \sum_{i=1}^k f_i [\bar{y}_i - ax_i - b]^2 &= \sum_{i=1}^k f_i [(\bar{y}_i - \bar{y}) - a(x_i - \bar{x}) + (\bar{y} - a\bar{x} - b)]^2 \\ &= \sum_{i=1}^k f_i (\bar{y}_i - \bar{y})^2 + a^2 \tilde{s}^2(x) - 2a \sum_{i=1}^k f_i (x_i - \bar{x})(\bar{y}_i - \bar{y}) + (\bar{y} - a\bar{x} - b)^2 \\ &= \eta_{y/x}^2 \tilde{s}^2(y) + a^2 s^2(x) - 2acov(x, y) + (\bar{y} - a\bar{x} - b)^2. \end{aligned}$$

La solution est donc celle annoncée et son report dans le critère donne :

$$\min_{a,b} \sum_{i=1}^k f_i [\bar{y}_i - ax_i - b]^2 = \sum_{i=1}^k f_i [\bar{y}_i - \hat{a}x_i - \hat{b}]^2 = [\eta_{y/x}^2 - r^2] \tilde{s}^2(y).$$

Ceci permet de décomposer la variance de Y en trois parties :

$$\tilde{s}^2(y) = [1 - \eta_{y/x}^2] \tilde{s}^2(y) + [\eta_{y/x}^2 - r^2] \tilde{s}^2(y) + r^2 \tilde{s}^2(y).$$

La première, propre à la variable Y , mesure la dispersion des points autour de la courbe de régression. Les deux suivantes traduisent la part expliquée par la variable X . Celle-ci est alors décomposée en la partie due à la régression linéaire (dernier terme) et le gain qu'apporte la courbe de régression par rapport à la droite (terme central). On peut mettre cette décomposition sous une forme plus explicite :

$$\sum_{i,j} f_{ij} (y_j - \bar{y})^2 = \sum_i f_i \sum_j f_j (y_j - \bar{y}_i)^2 + \sum_i f_i [\bar{y}_i - \hat{a}x_i - \hat{b}]^2 + \sum_i f_i [\hat{a}x_i + \hat{b} - \bar{y}]^2.$$

La droite de régression prend tout son sens lorsque la liaison entre les deux variables a un caractère linéaire ; c'est-à-dire que la courbe de régression doit elle-même être proche d'une droite et donc r^2 doit être voisin du rapport de

corrélation correspondant. L'égalité $\eta_{y/x}^2 = r^2$ équivaut à ce que la droite et la courbe de régression de Y en x soient confondues. En effet cette égalité dit que le gain de la courbe par rapport à la droite est nul, c'est-à-dire que le minimum ci-dessus est égal à zéro :

$$\sum_{i=1}^k f_i [\bar{y}_i - \hat{a}x_i - \hat{b}]^2 = 0 \quad \Leftrightarrow \quad \bar{y}_i = \hat{a}x_i + \hat{b}, \quad i = 1, \dots, k.$$

Une différence notable entre $\eta_{y/x}^2$ et r^2 permet donc de rejeter une hypothèse de liaison linéaire entre Y et X . La forme de la courbe de régression peut suggérer des transformations à effectuer sur les variables de façon à rendre la liaison linéaire, ce qui facilite les opérations de prévision (*cf.* Section 1.3.5).

La droite des moindres carrés calculée sur les couples (x_i, \bar{y}_i) , sans le système de pondération $(n_i$ ou $f_i)$, est différente de la droite de régression à l'exception des deux situations suivantes : les points (x_i, \bar{y}_i) sont alignés, c'est-à-dire que la courbe et la droite de régression sont confondues, ou les poids $(n_i$ ou $f_i)$ sont identiques. En pratique on peut être plus ou moins proche de ces situations et constater ainsi la proximité des deux droites. Seule la droite de régression est cependant à retenir bien que la droite des moindres carrés soit plus proche graphiquement des points (x_i, \bar{y}_i) formant la courbe de régression (le graphe ne fait pas apparaître l'importance relative des points).

1.3.4 Vecteurs aléatoires gaussiens

On donne ci-dessous quelques rappels, sans démonstrations, sur les vecteurs aléatoires gaussiens (*cf.* Barra, CH VII).

Soit X un vecteur aléatoire à valeurs dans \mathbb{R}^n , d'espérance $\mathbb{E}(X) = m$ et de matrice de variance-covariance $\mathbb{E}\{(X - m)(X - m)^T\} = \sigma^2$. On dit que X suit la loi normale (gaussienne, de Laplace-Gauss) $\mathcal{N}(m, \sigma^2)$ lorsque la fonction caractéristique est donnée par :

$$\mathbb{E}\{\exp(i \langle u, X \rangle)\} = \exp\{i \langle u, m \rangle - \frac{1}{2} {}^t u \Sigma^2 u\}, \quad u \in \mathbb{R}^n.$$

Une matrice de covariance est toujours symétrique définie non négative, $\sigma^2 \geq 0$, c'est-à-dire que ses valeurs propres sont réelles positives ou nulles. Lorsqu'elle est définie strictement positive, $\sigma^2 > 0$, la loi $\mathcal{N}(m, \sigma^2)$ admet une densité,

$$f(x) = (2\pi)^{-n/2} (\det \Sigma^2)^{-1/2} \exp\{-\frac{1}{2} {}^t (x - m) \sigma^{-2} (x - m)\}, \quad x \in \mathbb{R}^n.$$

C'est évidemment la situation la plus courante.

Toute transformation affine, $Y = AX + b$, où A est une matrice $p \times n$, et $b \in \mathbb{R}^p$, conserve le caractère gaussien :

$$X \sim \mathcal{N}(m, \sigma^2) \quad \Rightarrow \quad Y = AX + b \sim \mathcal{N}(Am + b, A\Sigma^2 A^T)$$

Les formes linéaires ${}^t aX$, $a \in \mathbb{R}^n$, sont donc des variables aléatoires gaussiennes. Réciproquement si ${}^t aX$ est gaussienne pour tout $a \in \mathbb{R}^n$, alors X est un *vecteur gaussien*. Par contre lorsque les composantes d'un vecteur aléatoire X sont gaussiennes, celui-ci n'est pas nécessairement gaussien, sauf si elles sont indépendantes. L'exemple suivant illustre ce problème. Soient X et Y deux variables $\mathcal{N}(0, 1)$ indépendantes, alors le vecteur aléatoire $(U, V)^T$ défini par $U = X$ et $V = \text{signe}(X)Y$ n'est pas gaussien alors que ses composantes sont $\mathcal{N}(0, 1)$. Sa densité est donnée par :

$$f(u, v) = \frac{1}{\pi} \exp\left\{-\frac{1}{2}(u^2 + v^2)\right\} \quad \text{si } uv \geq 0, \quad 0 \text{ sinon.}$$

Soient $X \sim \mathcal{N}(m, \sigma^2)$ et $V\Lambda^2 V^T = \sigma^2$ la décomposition spectrale de σ^2 dans laquelle les valeurs propres sont rangées par ordre décroissant dans Λ^2 . Si r est le rang de σ^2 et σ la matrice formée par les r premières colonnes de $V\Lambda$, on peut écrire X sous la forme $X = m + \sigma Y$ avec $Y \sim \mathcal{N}(0, I_r)$.

Soient X et Y deux vecteurs aléatoires à valeurs dans \mathbb{R}^p et \mathbb{R}^q auxquels on associe le vecteur aléatoire Z , à valeurs dans \mathbb{R}^{p+q} , avec les notations suivantes pour les deux premiers moments :

$$Z = \begin{pmatrix} X \\ Y \end{pmatrix}, \quad m_Z = \begin{pmatrix} m_X \\ m_Y \end{pmatrix}, \quad \sigma_Z^2 = \begin{bmatrix} \sigma_X^2 & \sigma_{XY}^2 \\ \Sigma_{YX}^2 & \sigma_Y^2 \end{bmatrix}.$$

Si X et Y sont indépendants et gaussiens alors Z est gaussien et la matrice de covariance $\sigma_{XY}^2 = \mathbb{E}\{(X - m_X)(Y - m_Y)^T\}$ est nulle. Réciproquement si Z est gaussien et que $\sigma_{XY}^2 = 0$, alors X et Y sont indépendants et gaussiens. Le résultat suivant éclaire le problème de la régression linéaire dans le cas de variables gaussiennes.

On suppose que la matrice de variance-covariance σ_X^2 de X est non singulière. Pour que Z soit gaussien il faut et il suffit que les conditions suivantes soient satisfaites :

- (i) X est gaussien,

- (ii) la loi de Y conditionnelle à X est gaussienne ; de plus la moyenne conditionnelle $\mathbb{E}(Y|X = x)$ est une fonction linéaire de x ,

$$\mathbb{E}(Y|X = x) = m_Y + \sigma_{YX}^2 \Sigma_X^{-2} (x - m_X),$$

et la matrice de variance conditionnelle $\sigma^2(Y|X = x)$ ne dépend pas de x ,

$$\sigma^2(Y|X = x) = \sigma_{YX}^2 \Sigma_X^{-2} \sigma_{XY}.$$

Dans le cas où X et Y sont les deux composantes d'un vecteur gaussien $Z = (X, Y)^T$, la fonction de régression $\mathbb{E}(Y|X = x)$ et la variance conditionnelle $\text{Var}(Y|X = x)$ se mettent sous la forme :

$$\mathbb{E}(Y|X = x) = m_Y + \text{Cov}(X, Y) \frac{(x - m_X)}{\text{Var}(X)}, \quad \text{Var}(Y|X = x) = \text{Var}(Y)[1 - \rho^2],$$

où ρ est le *coefficient de corrélation linéaire* entre les deux variables,

$$\rho = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}.$$

Lorsque le couple (X, Y) admet une densité,

$$f_{X,Y}(x, y) = \frac{\exp -\frac{1}{2(1-\rho^2)} \left\{ \left(\frac{x-m_X}{\sigma_X} \right)^2 + \left(\frac{y-m_Y}{\sigma_Y} \right)^2 - 2\rho \left(\frac{x-m_X}{\sigma_X} \right) \left(\frac{y-m_Y}{\sigma_Y} \right) \right\}}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}},$$

la division par la densité marginale de X ,

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma_X} \exp -\frac{1}{2} \left(\frac{x - m_X}{\sigma_X} \right)^2,$$

donne la densité de la loi conditionnelle,

$$f_{Y|X=x}(y) = \frac{\exp -\frac{1}{2(1-\rho^2)\sigma_Y^2} \left\{ y - m_Y - \rho \frac{\sigma_Y}{\sigma_X} (x - m_X) \right\}^2}{2\pi\sigma_Y\sqrt{1-\rho^2}}.$$

Ceci justifie les résultats des moments conditionnels indiqués ci-dessus.

Ainsi lorsque les observations $(x_i, y_i), i = 1, \dots, n$ sont celles d'un échantillon gaussien, la droite de régression est une estimation de la fonction de régression (linéaire) obtenue en remplaçant les paramètres inconnus par les éléments empiriques correspondants. De plus la fonction de régression empirique doit être proche de la droite de régression. Ceci justifie le rôle très important

de la régression linéaire. Nous verrons également à la section suivante que certaines transformations simples sur les variables permettent de se ramener au cas linéaire.

Il existe des méthodes non paramétriques plus sophistiquées pour l'estimation des fonctions de régression lorsque le nombre d'observations est important (lissage par noyau, ajustement de fonctions splines, utilisations des ondelettes). Ces techniques et surtout leur justification sortent du cadre de ce cours.

1.3.5 Transformations sur les variables

De façon générale l'étude de la liaison d'une variable d'intérêt Y avec une variable x se ramène à la situation linéaire lorsque l'on dispose de deux fonctions connues f et g pour lesquelles la variable $Z = f(Y)$ satisfait avec $t = g(x)$ les hypothèses du modèle de régression linéaire simple :

$$Z_i = at_i + b + \varepsilon_i, \quad i = 1, \dots, n, \quad \mathbb{E}(\varepsilon_i) = 0, \quad \text{Cov}(\varepsilon_i, \varepsilon_j) = \sigma^2 \delta_{ij}.$$

La fonction f doit être inversible pour obtenir la fonction de liaison $y = f^{-1}\{ag(x) + b\}$, mais cela n'est pas nécessaire pour g . Dans les transformations usuelles, le rôle des paramètres a et b est clairement identifié dans la liaison entre les variables initiales, mais celui de l'erreur ε_i , et par suite la qualité de l'approximation, sont souvent difficiles à interpréter. Notons qu'il n'y a aucun problème lorsque la transformation porte uniquement sur la variable x . Dans ce cas la fonction de liaison est aussi la fonction de régression, ce qui n'est pas vrai dans le cas général.

Fonctions exponentielles

La transformation la plus courante concerne les phénomènes de variation de type exponentiel. On ajuste une droite de régression sur le logarithme de la grandeur étudiée (*cf.* . exemple en section 1.3.6) :

$$Z_i = \log Y_i = ax_i + b + \varepsilon_i \Leftrightarrow Y_i = \beta e^{\alpha x_i} \times \eta_i, \quad i = 1, \dots, n,$$

où $\alpha = a, \beta = e^b$ et $\eta_i = e^{\varepsilon_i}$. On note que l'erreur est de type multiplicatif, son effet sur la variable Y est d'autant plus grand que celle-ci est grande. Lorsque Z suit la loi normale $\mathcal{N}(ax+b, \sigma^2)$ on dit que Y suit la loi log-normale de paramètres $ax + b$ et σ^2 . On montre les résultats suivants :

$$\mathbb{E}(Y) = e^{ax+b+\sigma^2}, \quad \text{Var}(Y) = e^{2(ax+b)+\sigma^2} (e^{\sigma^2} - 1).$$

La fonction de régression ne correspond donc pas à la fonction exponentielle ajustée par ce procédé.

La fonction $y = \beta e^{-\alpha x}$ traduit une décroissance vers 0 lorsque α et β sont positifs et est associée à l'équation différentielle $dy/dx = -\alpha y$. On rencontre aussi des décroissances ou croissances vers une valeur limite L connue modélisées par :

$$\begin{aligned} y = L + \beta e^{-\alpha x} &\rightarrow Z = \log(Y - L), \quad \frac{dy}{dx} = -\alpha(y - L), \\ y = L - \beta e^{-\alpha x} &\rightarrow Z = \log(L - Y), \quad \frac{dy}{dx} = -\alpha(L - y). \end{aligned}$$

Dans ces situations la vitesse de variation de y par rapport à x , dy/dx , est proportionnelle à la variation possible, $y, y - L$ et $L - y$.

Fonctions puissances

La variation relative de y est proportionnelle à celle de x :

$$\begin{aligned} y = \beta x^\alpha &\rightarrow Z = \log Y = at + b + \varepsilon, \quad \frac{dy}{y} = \alpha \frac{dx}{x}, \\ y = L + \beta x^\alpha &\rightarrow Z = \log(Y - L) = at + b + \varepsilon, \quad \frac{dy}{y-L} = \alpha \frac{dx}{x}, \end{aligned}$$

où $t = \log x, \alpha = a$ et $\beta = e^b$.

Fonctions hyperboliques

$$\begin{aligned} y = \frac{1}{\alpha x + \beta} &\rightarrow Z = \frac{1}{Y} = ax + b + \varepsilon, \quad \frac{dy/dx}{y} = \alpha y, \quad \alpha = a, \quad \beta = b, \\ y = \frac{x}{\alpha x + \beta} &\rightarrow Z = \frac{1}{Y} = at + b + \varepsilon, \quad \frac{dy}{y} = (1 - \alpha) \frac{dx}{x}, \quad \alpha = b, \quad \beta = a, \end{aligned}$$

où $t = \frac{1}{x}$.

Autres transformations

$$\begin{aligned} y = \frac{L}{1 - e^{-L(\alpha x + \beta)}} &\rightarrow Z = \frac{1}{L} \log \frac{Y}{Y-L} = ax + b + \varepsilon, \quad \frac{dy/dx}{y} = -\alpha(y - L), \\ y = \frac{L}{1 + e^{-L(\alpha x + \beta)}} &\rightarrow Z = \frac{1}{L} \log \frac{Y}{L-Y} = ax + b + \varepsilon, \quad \frac{dy/dx}{y} = \alpha(L - y), \end{aligned}$$

où $\alpha = a$ et $\beta = b$.

1.3.6 Exemples

Effet 519 sur décroissance D_A après αMT

On dispose de quatre lots de quatre rats chacun que l'on traite par αMT (alpha-Methyl-Tyrosine). Au bout d'un temps fixé t , variable pour chaque lot, on mesure la teneur y en D_A (Dépamide) dans le cerveau du rat. Les données et quelques résultats, avec $Z = \log Y$ sont présentées dans le Tableau 1.7.

La courbe de régression (non représentée) correspond à un tracé régulier passant par les valeurs moyennes. Elle apparaît légèrement plus incurvée que l'exponentielle ajustée par la méthode des moindres carrés (*cf.* Figure 1.15).

Temps en mn : t_i	30	60	120	180
y_{i1}	545	451	350	237
y_{i2}	588	400	320	230
y_{i3}	554	462	298	237
y_{i4}	525	470	292	276
Moyennes conditionnelles : \bar{y}_i	553	445,75	315	245
Variances conditionnelles : $\bar{s}_i^2(y)$	518,5	743,1875	517	328,5

Moyennes globales	$\bar{t} = 97,5$	$\bar{y} = 389,7$	$\bar{z} = 5,915$
Variances globales	$var(t) = 3318,75$	$var(y) = 14608,46$	$var(z) = 0,102345$
Covariances globales	$cov(t, y) = -6685,78$	$cov(t, z) = -17,9709$	
Décomposition de la variance	$var\ inter(y) = 14081,68$	$var\ intra(y) = 526,80$	$\eta_{y;t}^2 = 0,963$

Tableau 1.7: Données et quelques résultats pour la décroissance D_A

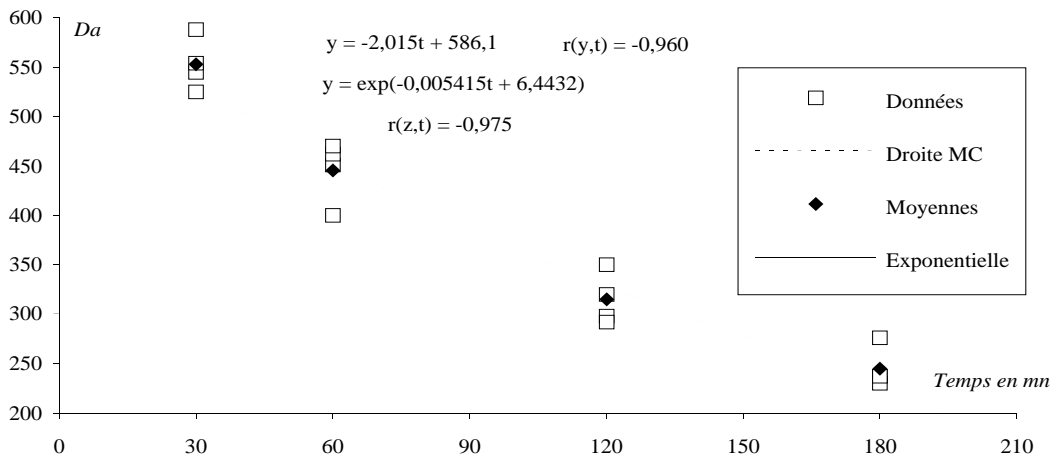


Figure 1.15: Décroissance D_A en fonction du temps

Poids en fonction de la taille d'un échantillon de conscrits

Une enquête a permis de répartir les 1000 conscrits d'un bureau de recrutement selon la taille X en cm et le poids Y en kg (*cf.* Tableau 1.8). Les caractéristiques conditionnelles sont données dans le Tableau 1.9 et quelques résultats figurent dans le Tableau 1.10. La courbe de régression est pratiquement confondue avec la droite de régression (*cf.* Figure 1.16). Ceci est cohérent avec la proximité entre le coefficient de détermination et le rapport de corrélation : $r^2 = 9,9\% < \eta_{y;x}^2 = 10,1\%$. Cela signifie que, en moyenne, le poids est bien une fonction linéaire de la taille mais, pour une taille donnée, le poids conserve une grande diversité.

Poids en kg Taille en cm	de 50 à 65	de 65 à 75	de 75 à 90	Ensemble
de 150 à 165	80	52	10	142
de 165 à 175	113	150	41	304
de 175 à 185	91	210	67	368
de 185 à 195	21	102	63	186
Total	305	514	181	1000

Tableau 1.8: Répartition des conscrits selon la taille et le poids

Taille en cm x_i	Moyenne \bar{y}_i	Variance $\bar{s}_i^2(y)$	Écart-type $\bar{s}_i(y)$	Fréquence f_i
157,5	63,8	61,06	7,8	14,2
170	67,0	70,39	8,4	30,4
180	69,2	66,42	8,1	36,8
190	72,8	62,60	7,9	18,6
Ensemble	68,5	73,54	8,6	100

Tableau 1.9: Caractéristiques conditionnelles du poids en fonction de la taille

Caractéristiques globales	$\bar{x} = 175,6$	$var(x) = 101,75$	$cov(x, y) = 27,16$
Décomposition de la variance	$var\ inter = 7,44$	$var\ intra = 66,16$	$\eta_{y;x}^2 = 10,1\%$

Tableau 1.10: Résultats concernant les conscrits

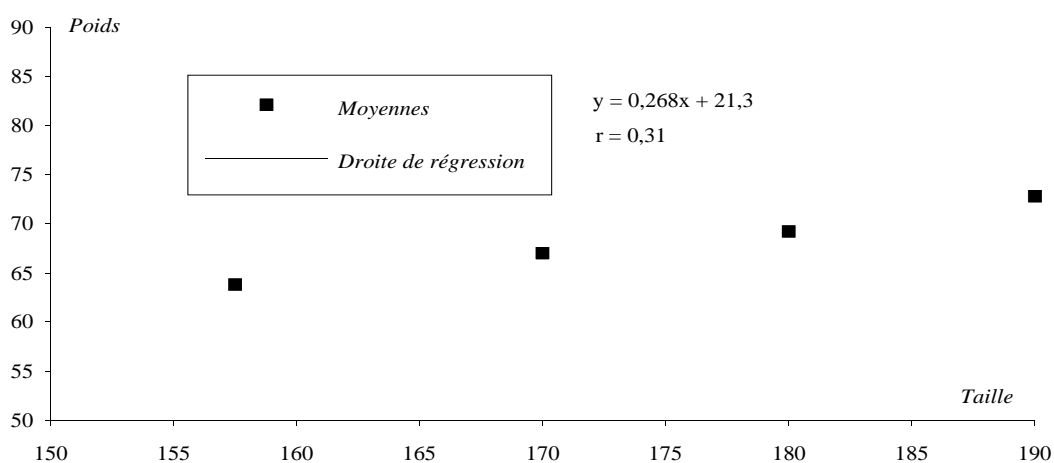


Figure 1.16: Courbe et droite de régression du poids en fonction de la taille d'un ensemble de conscrits

Ménages français en 1988 selon le nombre de personnes et le nombre de pièces habitables

Les données, exprimées en fréquences, sont présentées dans le Tableau 1.11. Les Tableaux 1.12 et 1.13 donnent les caractéristiques conditionnelles et les résultats sont reportés dans le Tableau 1.14. Les courbes et droites de régression (*cf.* Figure 1.17) sont très proches : $r^2 = 23\% < \eta_{y;x}^2 \simeq \eta_{x;y}^2 = 24\%$ et la corrélation entre les variables est plutôt faible.

Nombre de pièces Nombre de personnes	1	2	3	4	5	6	Ensemble
1	4,0	6,9	6,4	4,1	1,9	1,3	24,6
2	0,9	4,0	8,3	8,4	4,7	4,0	30,3
3	0,2	0,6	4,3	6,1	3,8	3,0	18,2
4	0,1	0,2	1,8	5,2	4,7	4,0	16,2
5	0,0	0,1	0,7	2,8	3,0	4,1	10,6
Total	5,2	12,0	21,5	26,7	18,1	16,4	100

Tableau 1.11: Répartition des ménages selon le nombre de personnes et de pièces habitables

Nombre de personnes x_i	Moyenne \bar{y}_i	Variance $\bar{s}_i^2(y)$	Écart-type $\bar{s}_i(y)$	Fréquence f_i
1	2,9	1,87	1,4	24,6
2	3,8	1,71	1,3	30,3
3	4,2	1,31	1,1	18,2
4	4,6	1,11	1,1	16,2
5	5,0	0,99	1,0	10,6
Ensemble	3,9	1,99	1,4	100

Tableau 1.12: Caractéristiques du nombre de pièces conditionnelles au nombre de personnes

1.4 PROPRIÉTÉS DES ESTIMATEURS

On reprend les hypothèses du modèle de régression linéaire simple en remplaçant le paramètre (a, b) par (α, β) :

$$Y_i = \alpha x_i + \beta + \varepsilon_i, \quad i = 1, \dots, n, \quad \mathbf{IE}(\varepsilon_i) = 0, \quad \mathbf{IE}(\varepsilon_i, \varepsilon_j) = \sigma^2 \delta_{ij}.$$

On ajoutera l'hypothèse gaussienne lorsque cela sera nécessaire.

Nombre de pièces y_j	Moyenne \bar{x}_j	Variance $\bar{s}_j^2(x)$	Écart-type $\bar{s}_j(x)$	Fréquence f_j
1	1,3	0,41	0,6	5,2
2	1,5	0,55	0,7	12,0
3	2,2	1,10	1,0	21,5
4	2,8	1,50	1,2	26,7
5	3,1	1,59	1,3	18,1
6	3,3	1,69	1,3	16,4
Ensemble	2,6	1,70	1,3	100

Tableau 1.13: Caractéristiques du nombre de personnes conditionnelles au nombre de pièces

Corrélation	$cov(x, y) = 0,89$	$r = 0,49$	$r^2 = 23\%$
Décomposition de $var(y)$	$var\ inter = 0,47$	$var\ intra = 1,50$	$\eta_{y;x}^2 = 24\%$
Décomposition de $var(x)$	$var\ inter = 0,40$	$var\ intra = 1,29$	$\eta_{x;y}^2 = 24\%$
Équations des droites	$y = 0,52x + 2,6$	$x = 0,45y + 0,8$	

Tableau 1.14: Résultats concernant les ménages

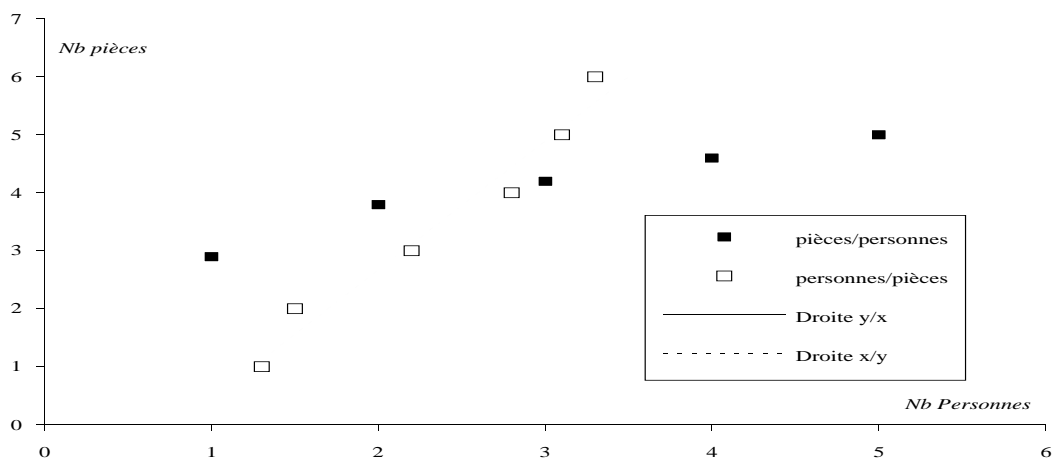


Figure 1.17: Courbes et droites de régression pour la répartition des ménages selon le nombre de personnes et le nombre de pièces habitables

1.4.1 Biais et variances

Le calcul des deux premiers moments (moyenne et variance) des estimateurs est facilité par les expressions suivantes :

- $\hat{\alpha} = \alpha + \frac{1}{\text{var}(x)} \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}) \varepsilon_i,$
- $\bar{Y} = \alpha \bar{x} + \beta + \bar{\varepsilon}, \quad \bar{\varepsilon} = \frac{1}{n} \sum_{i=1}^n \varepsilon_i,$
- $\hat{\beta} = \beta - (\hat{\alpha} - \alpha) \bar{x} + \bar{\varepsilon}.$

Les estimateurs sont donc sans biais et les variances sont obtenues en remarquant que $\hat{\alpha}$ et \bar{Y} (ou $\bar{\varepsilon}$) sont non corrélés :

$$\text{Var}(\hat{\alpha}) = \frac{\sigma^2}{n \text{var}(x)}; \quad \text{Var}(\hat{\beta}) = \frac{\sigma^2}{n} \left[\frac{\bar{x}^2}{\text{var}(x)} + 1 \right] = \frac{\sigma^2 \bar{x}^2}{n \text{var}(x)}.$$

Par contre $\hat{\alpha}$ et $\hat{\beta}$ sont corrélés et on a :

$$\text{Cov}(\hat{\alpha}, \hat{\beta}) = -\frac{\sigma^2 \bar{x}}{n \text{var}(x)}.$$

L'estimation de la variance σ^2 des erreurs utilise le vecteur des résidus :

$$\hat{\varepsilon} = (\hat{\varepsilon}_1, \dots, \hat{\varepsilon}_n)^T, \quad \hat{\varepsilon}_i = Y_i - \hat{\alpha} x_i - \hat{\beta}, \quad i = 1, \dots, n.$$

L'estimateur sans biais de σ^2 est donné par :

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n \hat{\varepsilon}_i^2 = \frac{1}{n-2} \sum_{i=1}^n [Y_i - \hat{\alpha} x_i - \hat{\beta}]^2.$$

La division par $n-2$ et non par le nombre de termes n est une conséquence de l'estimation des deux paramètres α et β . Un calcul direct permet également de montrer que $\hat{\sigma}^2$ est sans biais. L'orthogonalité de $\hat{\varepsilon}$ au sous-espace des moyennes donne :

$$\sum_{i=1}^n [Y_i - \alpha x_i - \beta]^2 = \sum_{i=1}^n [Y_i - \hat{\alpha} x_i - \hat{\beta}]^2 + \sum_{i=1}^n [\hat{\alpha} x_i + \hat{\beta} - \alpha x_i - \beta]^2.$$

En développant le dernier terme et en utilisant $\hat{\beta} - \beta = \bar{\varepsilon} - (\hat{\alpha} - \alpha) \bar{x}$, on obtient :

$$\sum_{i=1}^n [Y_i - \alpha x_i - \beta]^2 = \sum_{i=1}^n [Y_i - \hat{\alpha} x_i - \hat{\beta}]^2 + n \text{var}(x) (\hat{\alpha} - \alpha)^2 + n \bar{\varepsilon}^2.$$

Il suffit alors de prendre l'espérance des deux membres. On peut également vérifier directement que le vecteur des résidus $\hat{\varepsilon}$ est non corrélé avec le couple $(\hat{\alpha}, \hat{\beta})$:

$$\text{Cov}(\hat{\alpha}, \hat{\varepsilon}_i) = \text{Cov}(\hat{\beta}, \hat{\varepsilon}_i) = 0, \quad i = 1, \dots, n.$$

Les résidus sont des variables aléatoires centrées, $\mathbb{E}(\hat{\varepsilon}_i) = 0$, mais corrélées entre elles car $\sum_{i=1}^n \hat{\varepsilon}_i = 0$. Un calcul direct permet d'obtenir,

$$\text{Cov}(\hat{\varepsilon}_i, \hat{\varepsilon}_j) = \frac{\sigma^2}{n} \left[n\delta_{ij} - 1 - \frac{(x_i - \bar{x})(x_j - \bar{x})}{n \text{var}(x)} \right]$$

On peut aussi utiliser la non corrélation entre $\hat{\varepsilon}$ et le couple $(\hat{\alpha}, \hat{\beta})$ dans la relation,

$$\varepsilon = \hat{\varepsilon} + (x \ \mathbb{I}) \begin{pmatrix} \hat{\alpha} - \alpha \\ \hat{\beta} - \beta \end{pmatrix}, \quad \sigma_{\hat{\varepsilon}}^2 = \sigma^2 I_n - (x \ \mathbb{I}) \begin{pmatrix} 1 & -\bar{x} \\ -\bar{x} & \bar{x}^2 \end{pmatrix} \begin{pmatrix} {}^t x \\ {}^t \mathbb{I} \end{pmatrix},$$

1.4.2 Propriété de Gauss-Markov

Il s'agit d'une conséquence de résultats généraux établis dans le cadre du modèle linéaire. Elle ne dépend pas de la loi de probabilité des erreurs car elle ne fait référence qu'à des propriétés du second ordre (moyennes et variances).

Soit $u\alpha + v\beta$ une combinaison linéaire fixée quelconque des composantes du paramètre (α, β) . Alors $u\hat{\alpha} + v\hat{\beta}$ est l'unique estimateur de variance minimum de $u\alpha + v\beta$ parmi les estimateurs sans biais linéaires en Y .

Cette propriété est connue sous le nom d'*estimateur de Gauss-Markov*. Si $(\tilde{\alpha}, \tilde{\beta})$ désigne un autre estimateur sans biais de (α, β) , linéaire en Y , sa matrice de covariance est supérieure à celle de $(\hat{\alpha}, \hat{\beta})$ (la différence des deux matrices est une forme quadratique définie non-négative). On parle aussi de BLUE pour Best Linear Unbiased Estimator.

1.4.3 Lois de probabilités

On suppose désormais que les erreurs $\varepsilon_i, i = 1, \dots, n$ sont indépendantes et gaussiennes $\mathcal{N}(0, \sigma^2)$. Le vecteur ε est donc gaussien $\mathcal{N}(0, \sigma^2 I_n)$. Par suite le vecteur ${}^t(\hat{\alpha}, \hat{\beta}, \hat{\varepsilon})$, à valeurs dans \mathbb{R}^{n+2} , est gaussien puisqu'il est obtenu par transformation linéaire de ε :

- $\hat{\alpha} = \alpha + \frac{1}{n \text{var}(x)} {}^t(x - \bar{x}\mathbb{I})\varepsilon,$
- $\hat{\beta} = \beta - (\hat{\alpha} - \alpha)\bar{x} + \frac{1}{n} {}^t \mathbb{I} \varepsilon = \beta - \frac{1}{n \text{var}(x)} [{}^t(x - \bar{x})\mathbb{I} - \text{var}(x){}^t \mathbb{I}] \varepsilon,$

$$\bullet \hat{\varepsilon} = \varepsilon - (\hat{\alpha} - \alpha)x - (\hat{\beta} - \beta)\mathbb{I} = \left[I_n - \frac{x(x-\bar{x}\mathbb{I})^T + \mathbb{I}(\bar{x}^2\mathbb{I} - \bar{x}x)^T}{n \operatorname{var}(x)} \right] \varepsilon$$

Sa moyenne et sa matrice de covariance sont obtenues en reportant les résultats de la section précédente. En particulier $\hat{\varepsilon}$ et le couple $(\hat{\alpha}, \hat{\beta})$ sont indépendants puisque non corrélés. Alors $\hat{\sigma}^2$ et $(\hat{\alpha}, \hat{\beta})$ sont indépendants. On montre que $(n-2)\hat{\sigma}^2/\sigma^2$ suit la loi du khi-deux à $(n-2)$ degrés de liberté. Pour cela, il faut effectuer une transformation orthogonale M sur Y dans \mathbb{R}^n (changement de base) de sorte que les deux premières composantes du nouveau vecteur $Z = MY$ engendrent l'espace des moyennes, c'est-à-dire le plan défini par \mathbb{I} et x . Les autres composantes constituent alors le vecteur des résidus, elles sont indépendantes et de loi $\mathcal{N}(0, \sigma^2)$:

$$M = \begin{pmatrix} \frac{\mathbb{I}^T}{\sqrt{n}} \\ \frac{(x-\bar{x}\mathbb{I})^T}{\sqrt{n \operatorname{var}(x)}} \\ N \end{pmatrix}, \quad \mathbb{E}(Z) = \begin{pmatrix} \sqrt{(n)}(\alpha\bar{x} + \beta) \\ \sqrt{n \operatorname{var}(x)}\alpha \\ (0) \end{pmatrix}, \quad \operatorname{Var}(Z) = \sigma^2 I_n,$$

où la matrice N est choisie de sorte que $MM^T = I_n$.

1.4.4 Maximum de vraisemblance

La log-vraisemblance est le logarithme népérien de la densité :

$$L(y_1, \dots, y_n; \alpha, \beta, \sigma^2) = -\frac{n}{2} \log 2\pi - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \alpha x_i \beta)^2.$$

Maximiser cette fonction par rapport au couple (α, β) équivaut donc à minimiser le terme quadratique. L'estimateur des moindres carrés $(\hat{\alpha}, \hat{\beta})$ est aussi l'*estimateur de maximum de vraisemblance*. Ceci n'est plus vrai dans le cas de σ^2 . Les trois dérivées sont données par :

- $\frac{\partial}{\partial \alpha} L(y_1, \dots, y_n; \alpha, \beta, \sigma^2) = \frac{1}{\sigma^2} \sum_{i=1}^n x_i (y_i - \alpha x_i \beta),$
- $\frac{\partial}{\partial \beta} L(y_1, \dots, y_n; \alpha, \beta, \sigma^2) = \frac{1}{\sigma^2} \sum_{i=1}^n (y_i - \alpha x_i \beta),$
- $\frac{\partial}{\partial \sigma^2} L(y_1, \dots, y_n; \alpha, \beta, \sigma^2) = \frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{i=1}^n (y_i - \alpha x_i \beta)^2.$

L'estimateur de maximum de vraisemblance de $(\alpha, \beta, \sigma^2)$, obtenu en annulant les trois dérivées, donne $(\hat{\alpha}, \hat{\beta})$ et la variance empirique, c'est-à-dire l'estimateur biaisé, pour σ^2 :

$$\hat{\sigma}_{mv}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{\alpha} x_i - \hat{\beta})^2 = \frac{n-2}{n} \hat{\sigma}^2.$$

La *matrice d'information de Fisher* est l'opposé de l'espérance mathématique des dérivées secondes :

- $\frac{\partial^2}{\partial \alpha^2} L(\cdot) = -\frac{1}{\sigma^2} \sum_{i=1}^n x_i^2 = -\frac{n}{\sigma^2} \overline{x^2}$, $\frac{\partial^2}{\partial \beta^2} L(\cdot) = -\frac{n}{\sigma^2}$,
- $\frac{\partial^2}{\partial \alpha \partial \beta} L(\cdot) = -\frac{1}{\sigma^2} \sum_{i=1}^n x_i = -\frac{n}{\sigma^2} \overline{x}$,
- $\frac{\partial^2}{\partial (\sigma^2)^2} L(\cdot) = \frac{n}{2(\sigma^2)^2} - \frac{1}{(\sigma^2)^3} \sum_{i=1}^n (y_i - \alpha x_i \beta)^2$,
- $\frac{\partial^2}{\partial \alpha \partial \sigma^2} L(\cdot) = -\frac{1}{(\sigma^2)^2} \sum_{i=1}^n x_i (y_i - \alpha x_i \beta)$,
- $\frac{\partial^2}{\partial \beta \partial \sigma^2} L(\cdot) = -\frac{1}{(\sigma^2)^2} \sum_{i=1}^n (y_i - \alpha x_i \beta)$,
- $I(\alpha, \beta, \sigma^2) = \frac{n}{\sigma^2} \begin{bmatrix} \overline{x^2} & \overline{x} & 0 \\ \overline{x} & 1 & 0 \\ 0 & 0 & \frac{1}{2\sigma^2} \end{bmatrix}$.

Son inverse donne la *borne de Cramer-Rao*, c'est-à-dire la variance minimum des estimateurs sans biais :

$$I(\alpha, \beta, \sigma^2)^{-1} = \frac{\sigma^2}{n \operatorname{var}(x)} \begin{bmatrix} 1 & -\overline{x} & 0 \\ -\overline{x} & \overline{x^2} & 0 \\ 0 & 0 & 2\sigma^2 \operatorname{var}(x) \end{bmatrix}.$$

L'estimateur $(\hat{\alpha}, \hat{\beta})$ est *efficace* puisqu'il atteint la borne de Cramer-Rao. Ce n'est pas vrai pour $\hat{\sigma}^2$ car

$$\operatorname{Var}(\hat{\sigma}^2) = \frac{2\sigma^4}{n-2} > \frac{2\sigma^4}{n}.$$

Il est cependant asymptotiquement efficace.

1.5 INTERVALLES DE CONFIANCE ET TESTS

1.5.1 Intervalles de confiance

Les *intervalles de confiance* se construisent de façon analogue à ceux des paramètres d'une loi normale au vu d'un échantillon. Par exemple, dans le cas de α , les lois de $\hat{\alpha}$ et $\hat{\sigma}^2$ et l'indépendance des estimateurs conduisent à la *loi de Student* à $(n-2)$ degrés de liberté :

$$\begin{aligned} \hat{\alpha} &\sim \mathcal{N}\left(\alpha, \frac{\sigma^2}{n \operatorname{var}(x)}\right) \text{ et } \frac{n-2}{\sigma^2} \hat{\sigma}^2 \sim \chi_{n-2}^2 \text{ indépendantes} \\ &\implies \frac{\sqrt{n \operatorname{var}(x)}}{\hat{\sigma}} (\hat{\alpha} - \alpha) \sim \mathcal{S}_{n-2}^t. \end{aligned}$$

L'intervalle de confiance de niveau $(1 - \alpha)$ (ou de risque α) pour α (les deux α n'ont pas la même signification!) est donné par :

$$I(\alpha; 1 - \alpha) = \hat{\alpha} \pm t_{n-2; \alpha} \frac{\hat{\sigma}}{\sqrt{n \operatorname{var}(x)}}, \quad P\{|\mathcal{S}_{n-2}^t| > t_{n-2; \alpha}\} = \alpha.$$

Celui de β ,

$$I(\beta; 1 - \alpha) = \hat{\beta} \pm t_{n-2; \alpha} \frac{\sqrt{\bar{x}^2} \hat{\sigma}}{\sqrt{n \operatorname{var}(x)}}, \quad P\{|\mathcal{S}_{n-2}^t| > t_{n-2; \alpha}\} = \alpha,$$

n'est pas indépendant du précédent. C'est cependant une pratique courante plutôt que de chercher à construire une région de confiance pour le couple (α, β) (*cf.* Tassi, p. 346). Notons que \bar{Y} , qui est l'estimateur efficace de l'ordonnée $\gamma = \alpha \bar{x} + \beta$ de la droite du modèle en \bar{x} , est indépendant de $\hat{\alpha}$. L'intervalle de confiance,

$$I(\gamma; 1 - \alpha) = \bar{Y} \pm t_{n-2; \alpha} \frac{\hat{\sigma}}{\sqrt{n}}, \quad P\{|\mathcal{S}_{n-2}^t| > t_{n-2; \alpha}\} = \alpha,$$

n'est cependant pas indépendant de celui de α car ils sont tous deux fonction de $\hat{\sigma}$. Il sera tout de même préférable à celui de β pour construire une région de confiance approximative pour la droite. La région rectangulaire $I(\alpha; 1 - \alpha) \times I(\gamma; 1 - \alpha)$ a pour niveau approximatif $(1 - \alpha)^2$. De façon plus précise on peut construire un *ellipsoïde de confiance* de niveau $(1 - \alpha)$ en utilisant la loi de *Fisher-Snedecor*. On a

$$\begin{aligned} \frac{n}{\sigma^2} \{(\bar{Y} - \gamma)^2 + \operatorname{var}(x)(\hat{\alpha} - \alpha)^2\} &\sim \chi_2^2 \text{ et } \frac{n-2}{\sigma^2} \hat{\sigma}^2 \sim \chi_{n-2}^2 \text{ indépendantes} \\ \implies \frac{n}{2\hat{\sigma}^2} \{(\bar{Y} - \gamma)^2 + \operatorname{var}(x)(\hat{\alpha} - \alpha)^2\} &\sim \mathcal{F}_{2, n-2}. \end{aligned}$$

La région est l'intérieur de l'ellipse centrée en $(\bar{Y}, \hat{\alpha})$ définie par

$$\frac{n}{2\hat{\sigma}^2} \{(\bar{Y} - \gamma)^2 + \operatorname{var}(x)(\hat{\alpha} - \alpha)^2\} = f_{2, n-2; \alpha}, \quad P\{\mathcal{F}_{2, n-2} > f_{2, n-2; \alpha}\} = \alpha.$$

Chacun des intervalles précédents, considéré isolément, est optimal au sens où son amplitude est minimum pour la confiance imposée. Ce n'est pas le cas de l'intervalle de confiance usuel pour σ^2 ,

$$I(\sigma^2; 1 - \alpha) = \left[\frac{(n-2)\hat{\sigma}^2}{z_{n-2; \alpha/2}}, \frac{(n-2)\hat{\sigma}^2}{z_{n-2; 1-\alpha/2}} \right], \quad P\{\chi_{n-2}^2 > z_{n-2; p}\} = p.$$

Remarque

En pratique le calcul de $\hat{\sigma}^2$ ne s'effectue pas en fonction des résidus mais en utilisant la relation

$$\hat{\sigma}^2 = \frac{n}{n-2}(1-r^2)\text{var}(y) = \frac{n}{n-2}[\text{var}(y) - \hat{\alpha}^2\text{var}(x)].$$

1.5.2 Tests

Les tests unilatéraux ou bilatéraux sur les paramètres scalaires α, β (*tests de Student* et sur σ^2 (*tests du chi-deux*) sont semblables à ceux de la théorie des tests sur les paramètres de la loi normale au vu d'un échantillon. En particulier le test de $\alpha = 0$ contre $\alpha \neq 0$ permet de vérifier si Y dépend significativement (de façon linéaire) de la variable explicative X . Sa région critique, pour un test de niveau de signification α , est donnée par :

$$|\hat{\alpha}| > t_{n-2;\alpha} \frac{\hat{\sigma}}{\sqrt{n \text{var}(x)}}, \quad P\{|\mathcal{S}_{n-2}^t| > t_{n-2;\alpha}\} = \alpha.$$

Il équivaut à constater si l'intervalle de confiance au risque α contient ou non zéro. On introduit souvent l'*estimateur studentisé* consistant à normaliser l'estimateur initial en le divisant par son écart-type estimé,

$$\hat{\alpha}^S = \frac{\hat{\alpha} \sqrt{n \text{var}(x)}}{\hat{\sigma}}.$$

Le test équivaut donc à comparer $\hat{\alpha}^S$ avec $t_{n-2;\alpha}$. On peut aussi noter la loi d'une transformation de r sous l'hypothèse $\alpha = 0$:

$$\hat{\alpha}^S = \frac{\sqrt{n-2}r}{\sqrt{1-r^2}} \sim \mathcal{S}_{n-2}^t.$$

Une autre approche consiste à tester l'appartenance de la moyenne $\mathbb{E}(Y)$ au sous-espace $\{\beta\mathbb{I}, \beta \in \mathbb{R}\}$, c'est-à-dire $\alpha = 0$, par comparaison des estimateurs de la variance σ^2 à l'aide d'un *test de Fisher*. Dans la décomposition,

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (Y_i - \hat{\alpha}x_i - \hat{\beta})^2 + \sum_{i=1}^n (\hat{\alpha}x_i + \hat{\beta} - \bar{Y})^2 = (n-2)\hat{\sigma}^2 + n \text{var}(x)\hat{\alpha}^2,$$

les deux termes du second membre sont indépendants. Sous l'hypothèse nulle, $\alpha = 0$, la statistique $n \text{var}(x)\hat{\alpha}^2$ est aussi un estimateur sans biais de σ^2 et on a :

$$\begin{aligned} \frac{n-2}{\sigma^2}\hat{\sigma}^2 &\sim \chi_{n-2}^2 \text{ et } \frac{n \text{var}(x)}{\sigma^2}\hat{\alpha}^2 \sim \chi_1^2 \text{ indépendantes} \\ \implies \frac{n \text{var}(x)}{\hat{\sigma}^2}\hat{\alpha}^2 &\sim \mathcal{F}_{1,n-2}. \end{aligned}$$

Les deux tests sont équivalents puisque le carré d'une variable de loi \mathcal{S}_{n-2}^t suit la loi $\mathcal{F}_{1,n-2}$.

Illustration

Les résultats concernant l'exemple de la marge en fonction du chiffre d'affaires des entreprises sont les suivants.

- *Petites entreprises*

- $\hat{\sigma}^2 = 0,000791$; $\hat{\sigma} = 0,028MF$; $\hat{\alpha}^S = 37,772$; $\hat{\beta}^S = 0,954$.
- $I(\alpha; 95\%) = 0,3637 \pm 3,182 \times \frac{0,028}{\sqrt{5 \times 1,307}} = 0,3637 \pm 0,0306$.
- $I(\beta; 95\%) = 0,020 \pm 3,182 \times \frac{\sqrt{4,931716 \times 0,028}}{\sqrt{5 \times 1,307}} = 0,020MF \pm 0,068MF$.

- *Moyennes ou grandes entreprises*

- $\hat{\sigma}^2 = 1,074782$; $\hat{\sigma} = 1,037MF$; $\hat{\alpha}^S = 283,534$; $\hat{\beta}^S = 2,468$.
- $I(\alpha; 95\%) = 0,189106 \pm 3,182 \times \frac{1,037}{\sqrt{5 \times 695,143}} = 0,189106 \pm 0,002122$.
- $I(\beta; 95\%) = 1,356 \pm 3,182 \times \frac{\sqrt{678389,593 \times 1,037}}{\sqrt{5 \times 695,143}} = 1,356MF \pm 1,748MF$.

Dans les deux cas, la marge dépend de façon très significative du chiffre d'affaires. Par contre, β n'est pas significativement différent de zéro, mais il le devient au niveau 10% dans le cas des grandes entreprises ($t_{3;0,10} = 2,353$).

1.5.3 Étude des résidus

La représentation graphique des *résidus*, ou plus simplement la disposition des données $(x_i, y_i), i = 1, \dots, n$, autour de la droite de régression permet d'exclure les situations qui visiblement ne satisfont pas les hypothèses du modèle de régression. Par ailleurs les résidus $\hat{\varepsilon}_i, i = 1, \dots, n$, sont centrés, corrélés entre eux et de variance :

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

Sous l'hypothèse gaussienne $\hat{\varepsilon}_i$ est une variable normale. Bien qu'elle ne soit pas indépendante de $\hat{\sigma}^2$, on appelle *résidus studentisés* (ou *résidus standardisés*), notés $\hat{\varepsilon}_i^S$, les variables normalisées par l'estimation de l'écart-type,

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

La comparaison de $|\hat{\varepsilon}_i^S|$ avec $t_{n-2;\alpha}$, $i = 1, \dots, n$, permet éventuellement de rejeter le modèle.

Une étude plus rigoureuse du résidu en x_i consiste à estimer l'ensemble des paramètres du modèle sans l'observation y_i . On note $\hat{\alpha}_{(i)}, \hat{\beta}_{(i)}$ et $\hat{\sigma}_{(i)}^2$ ces estimateurs et $\hat{\varepsilon}_{(i)} = Y_i - \hat{\alpha}_{(i)}x_i - \hat{\beta}_{(i)} = Y_i - \hat{Y}_{(i)}$ le nouveau résidu. Les résultats généraux, obtenus dans le cadre du modèle linéaire (*cf.* CH II), montrent que l'on a :

$$\hat{\varepsilon}_{(i)} = \frac{\sigma^2 \hat{\varepsilon}_i}{\text{Var}(\hat{\varepsilon}_i)}, \quad (n-3)\hat{\sigma}_{(i)}^2 = (n-2-\hat{\varepsilon}_i^S)\hat{\sigma}^2, \quad \hat{\varepsilon}_i^V = \hat{\varepsilon}_i^S \sqrt{\frac{n-3}{n-2-\hat{\varepsilon}_i^S}} \sim \mathcal{S}_{n-3}^t,$$

où $\hat{\varepsilon}_i^V$ est le résidu $\hat{\varepsilon}_{(i)}$ studentisé, dit *résidu par validation croisée*. Lorsque n est grand, $\hat{\varepsilon}_i^V$ et $\hat{\varepsilon}_i^S$ sont sensiblement égaux à la normalisation "brutale" $\hat{\varepsilon}_i/\hat{\sigma}$ dont la loi approximative est $\mathcal{N}(0, 1)$.

1.5.4 Prévision

La *prévision* concerne l'estimation de la valeur de Y_0 associée à une nouvelle donnée x_0 de X . D'après le modèle, on a $Y_0 = \alpha x_0 + \beta + \varepsilon_0$ où ε_0 est une variable centrée de variance σ^2 non corrélée avec $Y_i, i = 1, \dots, n$, mais non observable. On se limite aux prédicteurs \tilde{Y}_0 linéaires en Y , sans biais, et on cherche à minimiser la variance de l'erreur de prédiction. On a :

$$\mathbb{E}(\tilde{Y}_0 - Y_0) = 0 \Rightarrow \mathbb{E}(\tilde{Y}_0) = \alpha x_0 + \beta; \quad \text{Var}(\tilde{Y}_0 - Y_0) = \text{Var}(\tilde{Y}_0) + \sigma^2.$$

On est donc ramené à chercher un estimateur sans biais de $\alpha x_0 + \beta$ qui soit linéaire en Y et de variance minimum. On a vu que la solution est donnée par $\hat{Y}_0 = \hat{\alpha}x_0 + \hat{\beta}$ et sa variance est égale à :

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

Il est nécessaire de tenir compte de la variance de l'erreur ε_0 pour définir l'intervalle de confiance sous hypothèse gaussienne (*intervalle de prévision*) :

$$\hat{\alpha}x_0 + \hat{\beta} \pm t_{n-2;\alpha} \frac{\hat{\sigma}}{\sqrt{n}} \sqrt{n+1 + \frac{(x_0 - \bar{x})^2}{\text{var}(x)}}, \quad P\{|\mathcal{S}_{n-2}^t| > t_{n-2;\alpha}\} = \alpha$$

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

Illustration

On considère la prévision en $x_0 = 1,462MF$, qui correspond à la catégorie “hors tranche”, dans le cadre des petites entreprises :

$$\hat{Y}_0 = \hat{\alpha}x_0 + \hat{\beta} = 0,3637 \times 1,462 + 0,020 = 0,552MF.$$

L'intervalle de confiance de niveau 95% est donné par :

$$0,552 \pm 3,182 \times \frac{0,028}{\sqrt{5}} \times \sqrt{6 + \frac{(1,462 - 1,796)^2}{1,707537}} = 0,552 \pm 0,098.$$

Il ne contient pas la valeur 0,319; la catégorie “hors tranche” serait donc un point aberrant parmi les petites entreprises. La différence $(0,319 - 0,552)$ est le résidu $\hat{\varepsilon}_{(0)}$ en $x_0 = 1,462$ du modèle incluant cette catégorie.

1.5.5 Prévision inverse

Dans certaines situations, il est naturel d'effectuer une *prévision inverse*, c'est-à-dire d'estimer x_0 au vu d'une observation de Y_0 . C'est par exemple le cas de la calibration d'un appareil de mesure : pour les valeurs connues x_1, \dots, x_n , de la grandeur mesurée X , on observe les valeurs de Y_1, \dots, Y_n que donne l'appareil. Ceci permet d'estimer les paramètres α, β et σ^2 . Un estimateur raisonnable de x_0 est obtenu en inversant la relation $y = \hat{\alpha}x + \hat{\beta}$, soit :

$$\hat{x}_0 = \frac{1}{\hat{\alpha}}(Y_0 - \hat{\beta}) = \bar{x} + \frac{1}{\hat{\alpha}}(Y_0 - \bar{Y}).$$

Cet estimateur est cependant biaisé et présente un écart quadratique moyen infini (cf. Montgomery & Peck, p. 402). La région de confiance, de niveau $(1 - \alpha)$, est construite à partir du résultat suivant :

$$P\{[Y_0 - \bar{Y} - \hat{\alpha}(x_0 - \bar{x})]^2 \leq t_{n-2;\alpha}^2 \frac{\hat{\sigma}^2}{n} \left[n + 1 + \frac{(x_0 - \bar{x})^2}{\text{var}(x)} \right]\} = 1 - \alpha,$$

où $P\{|\mathcal{S}_{n-2}^t| > t_{n-2;\alpha}\} = \alpha$. Il faut alors résoudre, par rapport à x_0 , l'inégalité,

$$A(x_0 - \bar{x})^2 - 2B(x_0 - \bar{x}) + C \leq 0,$$

où :

$$A = \left[\hat{\alpha}^2 - t_{n-2;\alpha}^2 \frac{\hat{\sigma}^2}{n \text{var}(x)} \right], \quad B = \hat{\alpha}(Y_0 - \bar{Y}),$$

et

$$C = \left[(Y_0 - \bar{Y})^2 - t_{n-2;\alpha}^2 \frac{(n+1)\hat{\sigma}^2}{n} \right].$$

On constate que le coefficient A de $(x_0 - \bar{x})^2$ est négatif si et seulement si l'hypothèse $\alpha = 0$ est acceptée par le test qui la concerne. Dans ce cas l'ensemble des x_0 est constitué soit de toute la droite \mathbb{R} , soit de deux demi-droites. Ces domaines sont sans intérêt, mais il n'est pas réaliste non plus d'estimer x_0 lorsque Y ne dépend pas de X . Par contre, lorsque le test rejette l'hypothèse $\alpha = 0$, la région de confiance pour x_0 prend la forme d'un intervalle :

$$\bar{x} + \hat{\alpha} \frac{(Y_0 - \bar{Y})}{A} \pm t_{n-2;\alpha} \frac{\hat{\sigma}}{\sqrt{nA}} \sqrt{n+1 + \frac{(Y_0 - \bar{Y})^2}{A \text{var}(x)}}.$$

On utilise parfois l'approximation

$$\hat{x}_0 \pm t_{n-2;\alpha} \frac{\hat{\sigma}}{\sqrt{n}|\hat{\alpha}|} \sqrt{n+1 + \frac{(Y_0 - \bar{Y})^2}{\hat{\alpha}^2 \text{var}(x)}}, \quad \hat{x}_0 = \bar{x} + \frac{(Y_0 - \bar{Y})}{\hat{\alpha}},$$

consistant à remplacer A par $\hat{\alpha}^2$. Notons également l'expression de l'intervalle en fonction de la statistique F du test de $\alpha = 0$,

$$\hat{x}_0 \pm t_{n-2;\alpha} \frac{1}{\sqrt{A}} \sqrt{\frac{n+1}{n} \hat{\sigma}^2 + \frac{(Y_0 - \bar{Y})^2}{F - t_{n-2;\alpha}^2}}, \quad F = \frac{\hat{\alpha}^2 n \text{var}(x)}{\hat{\sigma}^2}.$$

Ainsi l'amplitude de l'intervalle est d'autant plus petite que F est grand.

Chapitre 2

RÉGRESSION LINÉAIRE MULTIPLE

La régression linéaire multiple est une extension immédiate de la régression linéaire simple, en particulier dans son interprétation géométrique dans l'espace des observations. La différence essentielle réside dans le formalisme qui passe par des écritures matricielles des estimateurs et de leur variance. Sur le plan pratique, la mise en œuvre des méthodes statistiques obtenues nécessite donc, sauf dans des situations particulières, de recourir à des moyens informatiques pour l'inversion de matrices. On retrouve dans ce chapitre la plupart des éléments de base présentés dans le précédent et la justification de certains d'entre eux : hypothèses du modèle, méthode des moindres carrés, propriétés des estimateurs, étude des résidus, prévision et tests. Nous considérons également la spécificité des applications à la régression polynomiale et aux séries chronologiques. Nous illustrons notre propos par un exemple issu du livre d'Antoniadis et al. (voir aussi Montgomery & Peck) auquel nous renvoyons pour des compléments sur le sujet.

2.1 LES HYPOTHÈSES DU MODÈLE

2.1.1 Modèle standard

Le modèle de régression linéaire simple peut s'écrire sous forme matricielle :

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_i \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_i & 1 \\ \vdots & \vdots \\ x_n & 1 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_i \\ \vdots \\ \varepsilon_n \end{bmatrix} \Leftrightarrow Y = X\theta + \varepsilon,$$

où $X = [x, \mathbb{1}]$ et $\theta = (\alpha, \beta)^T$. La généralisation consiste à considérer une matrice X comportant p colonnes et par suite un paramètre θ de dimension p . De façon précise, le modèle de régression linéaire multiple suppose que y_1, \dots, y_n sont les observations de variables aléatoires Y_1, \dots, Y_n satisfaisant :

$$Y_i = \sum_{j=1}^p \theta_j x_{ij} + \varepsilon_i, \quad i = 1, \dots, n,$$

où :

- le *plan d'expérience* X est une matrice connue de dimension $n \times p$, avec $p < n$, et de rang p , appelée *matrice des régresseurs*,
- les paramètres $\theta_1, \dots, \theta_p$ sont réels et inconnus,
- les erreurs $\varepsilon_1, \dots, \varepsilon_n$ sont des variables aléatoires centrées, de même variance σ^2 et non corrélées entre elles :

$$\mathbb{E}(\varepsilon_i) = 0, \quad \text{Var}(\varepsilon_i) = \sigma^2, \quad \text{Cov}(\varepsilon_i, \varepsilon_j) = 0 \text{ si } i \neq j.$$

On peut synthétiser les écritures du modèle sous la forme $(Y, X\theta, \sigma^2 I)$:

$$Y = X\theta + \varepsilon, \quad \mathbb{E}(\varepsilon) = 0, \quad \text{Var}(\varepsilon) = \sigma^2 I,$$

ce qui implique :

$$\mathbb{E}(Y) = X\theta, \quad \text{Var}(Y) = \sigma^2 I,$$

en introduisant les vecteurs aléatoires $Y = (Y_1, \dots, Y_n)^T$ et $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^T$, ainsi que le vecteur des paramètres $\theta = (\theta_1, \dots, \theta_p)^T$. Notons également

l'écriture :

$$Y = \sum_{j=1}^p \theta_j X_j + \varepsilon, \quad X = [X_1, \dots, X_p] \text{ où } X_j = \begin{bmatrix} x_{1j} \\ \vdots \\ x_{ij} \\ \vdots \\ x_{nj} \end{bmatrix}, \quad j = 1, \dots, p,$$

montrant que $\mathbb{E}(Y)$ est une combinaison linéaire des variables explicatives $X_j, j = 1, \dots, p$, qui constituent les colonnes de X , dont les coefficients sont les paramètres du modèle. C'est pourquoi on parle de *régression linéaire* de Y sur les variables $X_j, j = 1, \dots, p$. Il faut cependant souligner que la linéarité évoquée dans le terme *modèle linéaire* signifie que les observations $Y_i, i = 1 \dots, n$ dépendent linéairement des paramètres, ce qui apparaît dans l'écriture initiale de chaque Y_i en fonction des $\theta_j, j = 1, \dots, p$. De façon précise la moyenne est une application linéaire connue du paramètre vectoriel $\theta : \mathbb{E}(Y) = X\theta$.

La variable explicative constante $\mathbb{1}$ ne fait pas nécessairement partie des colonnes de X , bien que ce soit souvent le cas. La régression linéaire simple est une situation particulière du cas $p = 2$ qui permet un traitement spécifique du fait de la présence d'une seule variable explicative autre que la constante $\mathbb{1}$. Dans le cas général, on ne fait pas de distinction entre les modèles qui présentent ou non la variable $\mathbb{1}$, excepté dans la définition du coefficient de détermination. L'hypothèse fondamentale pour la matrice des régresseurs X est qu'elle soit de rang $p < n$. Cela équivaut à ce que les p vecteurs X_1, \dots, X_p de \mathbb{R}^n soient linéairement indépendants. Le sous-espace de dimension p qu'ils engendrent est appelé *espace des moyennes* et est noté $\mathcal{M}(X)$.

Pour construire des intervalles de confiance et effectuer des tests d'hypothèse, on ajoutera l'hypothèse de normalité des erreurs et par suite des observations. Lorsque le nombre n d'observations est important, on peut appliquer les résultats obtenus sous hypothèse gaussienne en invoquant les propriétés asymptotiques des estimateurs. Dans ce cas il suffit que les erreurs soient indépendantes et identiquement distribuées (*i.i.d.*).

Dans le cas $p = 3$ avec la constante $\mathbb{1}$, il est encore possible de visualiser les observations dans l'*espace des variables* \mathbb{R}^3 ainsi que le plan solution au problème (*cf.* Figure 2.1).

Sinon seul l'*espace des observations* \mathbb{R}^n conserve une interprétation géométrique

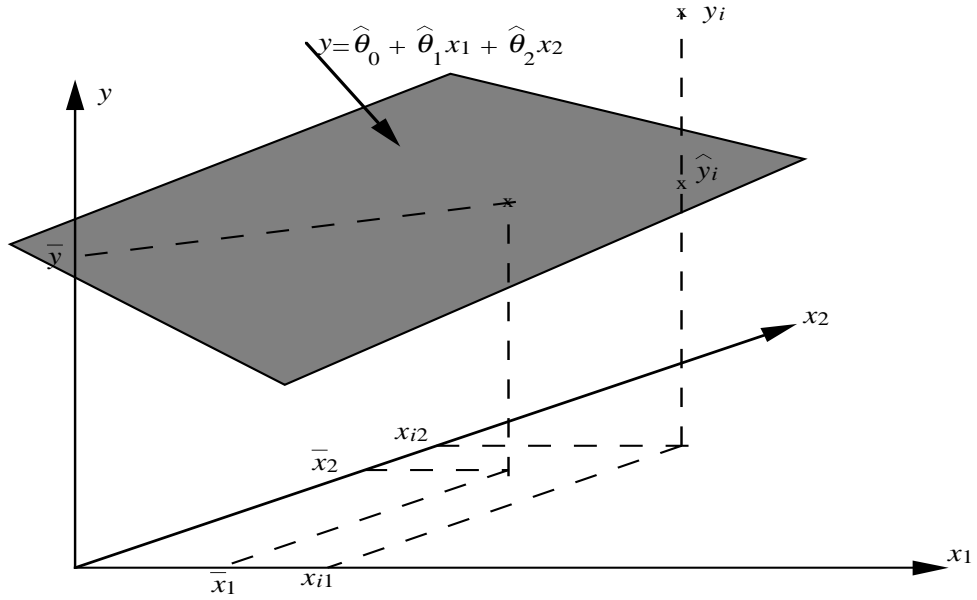


Figure 2.1: Espace des variables du modèle linéaire

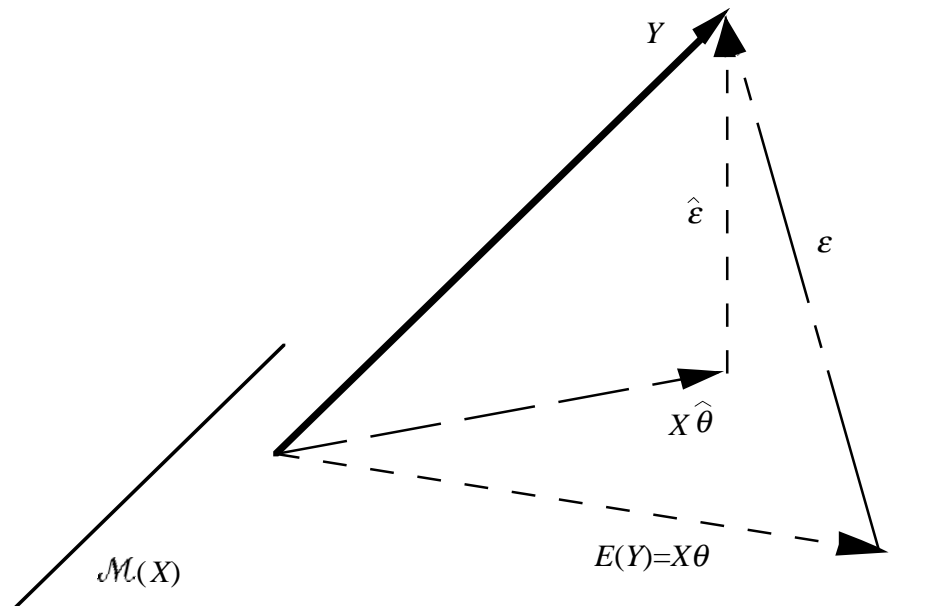


Figure 2.2: Espace des observations du modèle linéaire

intéressante (*cf.* Figure 2.2). On représente de façon symbolique le vecteur aléatoire Y des observations comme la somme de sa moyenne inconnue, $\mathbb{E}(Y) = X\theta$, dans l'espace des moyennes $\mathcal{M}(X)$ et du vecteur des erreurs ε non observable. On visualise également la projection $\hat{Y} = X\hat{\theta}$ et le vecteur des résidus $\hat{\varepsilon} = Y - \hat{Y} = Y - X\hat{\theta}$.

2.1.2 Illustration

Cette illustration concerne la maintenance d'un réseau de distributeurs de boissons (*cf.* Antoniadis et al. ou Montgomery & Peck). On dispose de $n = 25$ observations du temps en minutes (variable d'intérêt Y) pour approvisionner un réseau de distributeurs de boissons selon le nombre de caisses de bouteilles placées (variable X_1) et la distance parcourue en mètres (variable X_2). Le Tableau 2.1 présente l'ensemble des données, les valeurs ajustées, les différentes formes de résidus et les éléments diagonaux de la matrice de prédiction. Ces dernières notions, ainsi que d'autres résultats indiqués ci-après, seront introduits dans la suite du chapitre.

i	y_i	\mathbb{I}	x_{i1}	x_{i2}	\hat{y}_i	$\hat{\varepsilon}_i$	$\hat{\varepsilon}_i^S$	ε_i^V	h_{ii}
1	16,68	1	7	560	21,71	-5,03	-1,63	-1,70	0,1018
2	11,50	1	3	220	10,35	1,15	0,36	0,36	0,0707
3	12,03	1	3	340	12,08	-0,05	-0,02	-0,02	0,0987
4	14,88	1	4	80	9,96	4,92	1,58	1,64	0,0854
5	13,75	1	6	150	14,19	-0,44	-0,14	-0,14	0,0750
6	18,11	1	7	330	18,40	-0,29	-0,09	-0,09	0,0429
7	8,00	1	2	110	7,16	0,84	0,27	0,26	0,0818
8	17,83	1	7	210	16,67	1,16	0,37	0,36	0,0637
9	79,24	1	30	1460	71,82	7,42	3,21	4,31	0,4983
10	21,50	1	5	605	19,12	2,38	0,81	0,81	0,1963
11	40,33	1	16	688	38,09	2,24	0,72	0,71	0,0861
12	21,00	1	10	215	21,59	-0,59	-0,19	-0,19	0,1137
13	13,50	1	4	255	12,47	1,03	0,33	0,32	0,0611
14	19,75	1	6	462	18,68	1,07	0,34	0,33	0,0782
15	24,00	1	9	448	23,33	0,67	0,21	0,21	0,0411
16	29,00	1	10	776	29,66	-0,66	-0,22	-0,22	0,1659
17	15,35	1	6	200	14,91	0,44	0,14	0,13	0,0594
18	19,00	1	7	132	15,55	3,45	1,11	1,12	0,0963
19	9,50	1	3	36	7,71	1,79	0,58	0,57	0,0964
20	35,10	1	17	770	40,89	-5,79	-1,87	-2,00	0,1017
21	17,90	1	10	140	20,51	-2,61	-0,88	-0,87	0,1653
22	52,32	1	26	810	56,01	-3,69	-1,45	-1,49	0,3916
23	18,75	1	9	450	23,36	-4,61	-1,44	-1,48	0,0413
24	19,83	1	8	635	24,40	-4,57	-1,50	-1,54	0,1206
25	10,75	1	4	150	10,96	-0,21	-0,07	-0,07	0,0666

Tableau 2.1: Maintenance d'un réseau de distributeurs de boissons

Dans cet exemple, le plan d'expérience est $X = [\mathbb{I} \ X_1 \ X_2]$ et le paramètre est noté $\theta = (\theta_0 \ \theta_1 \ \theta_2)^T$. On utilise en effet l'indice 0 pour le coefficient de la

constante \mathbb{I} . On obtient les résultats suivants :

$${}^tXX = \begin{bmatrix} 25 & 219 & 10232 \\ 219 & 3055 & 133899 \\ 10232 & 133899 & 6725688 \end{bmatrix}, \quad [{}^tXX]^{-1} = \begin{bmatrix} 0,11321519 & -0,00444859 & -0,00008367 \\ -0,00444859 & 0,00274378 & -0,00004786 \\ -0,00008367 & -0,00004786 & 0,0000123 \end{bmatrix}$$

$${}^tXY = \begin{bmatrix} 559,60 \\ 7375,44 \\ 337071,69 \end{bmatrix}, \quad \hat{\theta} = \begin{bmatrix} 2,34123115 \\ 1,61590721 \\ 0,01438483 \end{bmatrix},$$

- Équation du plan de régression : $y = 2,341 + 1,6159x_1 + 0,014385x_2$
- Analyse de la variance : $SST = SS_{reg} + SSE \rightarrow 5784,5426 = 5550,8109 + 233,7317$
- Coefficient de détermination : $R^2 = 0,959593749 \simeq 96,0\%$
- Statistique de test : $F = \frac{(SST-SSE)/(p-1)}{SSE/(n-p)} = \frac{R^2/(p-1)}{(1-R^2)/(n-p)} = 261,24, \quad f_{2;22;5\%} = 3,44$

2.1.3 Modèle général

On étend facilement la situation du modèle standard à celle où les erreurs sont corrélées selon une structure connue :

$$Y = X\theta + \varepsilon, \quad \mathbb{E}(\varepsilon) = 0, \quad \text{Var}(\varepsilon) = \mathbb{E}(\varepsilon\varepsilon^T) = \sigma^2G,$$

où G est une matrice symétrique définie positive (matrice de covariance des erreurs au facteur d'échelle σ^2 près). Ce modèle, noté $(Y, X\theta, \sigma^2G)$, se ramène au modèle standard par toute transformation linéaire définie par une racine carrée $G^{1/2}$ de la matrice G . On a en effet :

$$Z = G^{-1/2}Y = G^{-1/2}X\theta + G^{-1/2}\varepsilon = U\theta + \eta, \quad \mathbb{E}(\eta) = 0, \quad \text{Var}(\eta) = \sigma^2I.$$

Une racine carrée de G est une matrice, notée $G^{1/2}$, satisfaisant $G^{1/2}G^{T/2} = G$ où $G^{T/2}$ désigne la matrice transposée de $G^{1/2}$. Si $G^{1/2}$ est une racine carrée de G , toute autre racine est de la forme $G^{1/2}V$, où V est une matrice orthogonale, $VV^T = I$. Parmi les racines les plus courantes, on distingue les racines triangulaires inférieure ou supérieure et la racine symétrique. Les racines triangulaires se calculent facilement par l'algorithme de Cholesky. La racine symétrique est associée à la décomposition spectrale $G = VD^2V^T$ en posant $G^{1/2} = VDV^T$. Ces trois racines sont uniques alors que par exemple la racine VD dépend de l'ordre des valeurs propres et du choix des vecteurs propres.

Un cas particulier du modèle $(Y, X\theta, \sigma^2G)$ est celui où G est diagonale. Les erreurs sont non corrélées mais présentent des variances différentes connues à un facteur près. C'est par exemple la situation où, pour chaque

valeur du plan d'expérience d'un modèle standard, plusieurs observations de la variable Y sont effectuées et que l'on considère le modèle associé aux valeurs moyennes de ces observations. On a vu dans le premier chapitre que ceci conduit à remplacer la droite des moindres carrés par celle des moindres carrés pondérés. On parle de *moindres carrés généralisés* pour le modèle $(Y, X\theta, \sigma^2 G)$. Cette situation ne doit pas être confondue avec la notion de modèle linéaire généralisé.

D'un point de vue géométrique, le modèle $(Y, X\theta, \sigma^2 G)$ équivaut à remplacer la norme euclidienne classique dans l'espace des observations \mathbb{R}^n , associée à la forme quadratique I_n , par celle définie par G^{-1} , $\langle x, y \rangle_{G^{-1}} = {}^t x G^{-1} y$.

Une autre généralisation correspond à la situation où la matrice du plan d'expérience X n'est pas de rang plein. Cela signifie que certaines colonnes de X s'expriment comme combinaisons linéaires des autres. Dans ce cas le paramètre θ n'est pas identifiable puisque $X\theta = X\tilde{\theta}$ pour certaines valeurs $\theta \neq \tilde{\theta}$ alors que $IE(y)$ est unique.

2.2 MÉTHODE DES MOINDRES CARRÉS

2.2.1 Estimateur des moindres carrés

La démarche est identique à celle utilisée en régression linéaire simple. De plus les écritures matricielles liées à l'aspect géométrique dans l'espace des observations montrent la simplicité du problème.

On appelle *estimateur des moindres carrés* du paramètre θ la statistique $\hat{\theta}(y)$ qui, pour toute observation y , minimise le critère :

$$D(\theta) = \sum_{i=1}^n \left[y_i - \sum_{j=1}^p x_{ij} \theta_j \right]^2 = \|y - X\theta\|^2 = {}^t [y - X\theta] [y - X\theta].$$

L'estimateur est noté plus simplement $\hat{\theta}$. Là encore le critère est strictement convexe et la solution est donnée par le système obtenu en annulant les dérivées partielles de $D(\theta)$ par rapport à chaque composante θ_k , $k = 1, \dots, p$:

$$\frac{\partial}{\partial \theta_k} D(\theta) = -2 \sum_{i=1}^n x_{ik} \left[y_i - \sum_{j=1}^p x_{ij} \theta_j \right] = 0, \quad k = 1, \dots, p.$$

Sous forme matricielle, on peut écrire directement :

$$\frac{\partial}{\partial \theta} [{}^t yy - 2{}^t \theta {}^t X y + {}^t \theta ({}^t X X) \theta] = -2{}^t X [y - X \theta] = 0.$$

La matrice hessienne est définie positive, ce qui prouve la stricte convexité du critère :

$$\frac{\partial^2}{\partial \theta^2} D(\theta) = 2{}^t X X > 0.$$

Il est équivalent de remarquer que $\hat{y} = X \hat{\theta}$ est la projection orthogonale de y sur $\mathcal{M}(X)$. Celle-ci se caractérise en écrivant que le résidu $\hat{\varepsilon} = y - X \hat{\theta}$ est orthogonal à $\mathcal{M}(X)$, c'est-à-dire aux colonnes de X . Les *équations normales* s'écrivent donc sous forme matricielle :

$${}^t X [y - X \hat{\theta}] = 0 \quad \Leftrightarrow \quad {}^t X X \hat{\theta} = {}^t X y.$$

La matrice ${}^t X X$ est carrée d'ordre p , symétrique et inversible car X est de rang p . La solution est donc :

$$\hat{\theta} = ({}^t X X)^{-1} {}^t X y.$$

2.2.2 Coefficient de détermination

Lorsque la constante \mathbb{I} est présente dans les variables explicatives (colonnes de X), on écrit plutôt le modèle avec θ_0 pour coefficient de la constante :

$$Y_i = \theta_0 + \sum_{j=1}^p x_{ij} \theta_j + \varepsilon_i, \quad i = 1, \dots, n.$$

L'annulation de la dérivée du critère par rapport à θ_0 montre que l'hyperplan solution passe par le point moyen $(\bar{y}, \bar{x}_1, \dots, \bar{x}_p)$:

$$\bar{y} = \hat{\theta}_0 + \bar{x}_1 \hat{\theta}_1 + \dots + \bar{x}_p \hat{\theta}_p.$$

On retrouve également la décomposition de la variance empirique :

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

qui résulte de l'orthogonalité entre $\hat{y} - \bar{y}\mathbb{I}$ et $y - \hat{y}$ dans l'égalité :

$$y - \bar{y}\mathbb{I} = (y - \hat{y}) + (\hat{y} - \bar{y}\mathbb{I}).$$

Ceci permet de définir le *coefficient de détermination* :

$$R^2 = \frac{\|\hat{y} - \bar{y}\mathbb{I}\|^2}{\|y - \bar{y}\|^2} = 1 - \frac{\|\hat{\varepsilon}\|^2}{\|y - \bar{y}\|^2},$$

qui a la même interprétation que r^2 en régression linéaire simple.

Remarques

- On définit également le *coefficient de détermination corrigé* \overline{R}^2 (Adjusted R-squared) qui consiste à remplacer, dans la définition de $1 - R^2$, les variances empiriques par les estimateurs sans biais de σ^2 (cf. Mélard, p. 188) :

$$1 - \overline{R}^2 = \frac{n-1}{n-p}(1 - R^2) = \frac{\frac{1}{n-p} \sum_{i=1}^n (y_i - \hat{y}_i)^2}{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2}.$$

On a alors :

$$\overline{R}^2 = R^2 - \frac{p-1}{n-p}(1 - R^2),$$

et ce coefficient, inférieur à R^2 , peut être négatif.

- Lorsque la constante $\mathbb{1}$ ne fait pas partie du plan d'expérience, c'est-à-dire que l'hyperplan passe par l'origine, le coefficient de détermination est défini par (cf. Antoniadis et al. p. 26) :

$$R^2 = \frac{\sum_{i=1}^n \hat{y}_i^2}{\sum_{i=1}^n y_i^2} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n y_i^2}.$$

Ceci équivaut à poser $\bar{y} = 0$ dans l'expression initiale de R^2 .

- Notons enfin que, dans ce type d'expressions, p désigne toujours la dimension de $\mathcal{M}(X)$ et non le nombre de variables x_j autres que la constante.

2.3 PROPRIÉTÉS DES ESTIMATEURS

2.3.1 Biais et variance

On utilise la même notation pour désigner l'estimateur $\hat{\theta}$ ou son observation et cet abus est également adopté pour le résidu $\hat{\varepsilon}$. Le calcul des deux premiers moments est immédiat.

$$\mathbb{E}(\hat{\theta}) = \mathbb{E}\{(^t X X)^{-1} {}^t X Y\} = (^t X X)^{-1} {}^t X \mathbb{E}(Y) = (^t X X)^{-1} {}^t X X \theta = \theta,$$

$$\text{Var}(\hat{\theta}) = [({}^t X X)^{-1} {}^t X] \text{Var}(Y) [({}^t X X)^{-1} {}^t X]^T = \sigma^2 ({}^t X X)^{-1}.$$

Soit $H = X({}^t X X)^{-1} {}^t X$ la matrice définissant, dans l'espace des observations \mathbb{R}^n , la projection orthogonale sur le sous-espace des moyennes $\mathcal{M}(X)$. On a en effet :

$$\hat{Y} = X\hat{\theta} = HY, \quad \hat{\varepsilon} = Y - \hat{Y} = (I - H)Y, \quad \mathbb{E}(\hat{Y}) = X\theta, \quad \mathbb{E}(\hat{\varepsilon}) = 0.$$

On dit que H est la *matrice "chapeau"* ou la *matrice de prédiction*. Les matrices H et $I - H$ sont symétriques et idempotentes ($H^2 = H$, $(I - H)^2 = I - H$). On en déduit immédiatement :

$$\text{Var}(\hat{Y}) = \sigma^2 H, \quad \text{Var}(\hat{\varepsilon}) = \sigma^2(I - H), \quad \text{Cov}(\hat{\varepsilon}, \hat{Y}) = 0, \quad \text{Cov}(\hat{\varepsilon}, \hat{\theta}) = 0.$$

L'expression de $\text{Var}(\hat{\varepsilon})$ montre qu'un estimateur sans biais de σ^2 est donné par :

$$\hat{\sigma}^2 = \frac{\|\hat{\varepsilon}\|^2}{n - p} = \frac{1}{n - p} \sum_{i=1}^n \hat{\varepsilon}_i^2 = \frac{1}{n - p} ({}^t Y Y - {}^t Y H Y),$$

car

$$\mathbb{E}\{\|\hat{\varepsilon}\|^2\} = \mathbb{E}\{{}^t \hat{\varepsilon} \hat{\varepsilon}\} = \mathbb{E}\{\text{tr}(\hat{\varepsilon} {}^t \hat{\varepsilon})\} = \sigma^2 \text{tr}(I - H) = (n - p)\sigma^2.$$

Pour le calcul de ${}^t \hat{\varepsilon} \hat{\varepsilon}$, on a le choix parmi les différentes expressions suivantes :

$$\begin{aligned} {}^t \hat{\varepsilon} \hat{\varepsilon} &= {}^t [Y - \hat{Y}] [Y - \hat{Y}] = {}^t [Y - X\hat{\theta}] [Y - X\hat{\theta}] = {}^t Y [Y - X\hat{\theta}] \\ &= {}^t Y Y - {}^t Y X \hat{\theta} = {}^t Y Y - {}^t Y X ({}^t X X)^{-1} {}^t X Y = {}^t Y Y - {}^t \hat{\theta} ({}^t X X) \hat{\theta}. \end{aligned}$$

Mais on peut surtout calculer $\hat{\sigma}^2$ en utilisant le coefficient de déterminant :

$$\hat{\sigma}^2 = (1 - R^2) \frac{n}{n - p} \text{var}(y).$$

2.3.2 Propriété de Gauss-Markov

L'estimateur $\hat{\theta}$ est sans biais et satisfait la propriété suivante :

Soit ${}^t w \theta$ une combinaison linéaire fixée quelconque des composantes du paramètre θ . Alors ${}^t w \hat{\theta}$ est l'unique estimateur de variance minimum de ${}^t w \theta$ parmi les estimateurs sans biais linéaires en Y .

On dit que $\hat{\theta}$ est un *estimateur de Gauss-Markov*. En effet un estimateur linéaire en Y est de la forme ${}^t c Y$ et peut également s'écrire ${}^t c Y = {}^t w \hat{\theta} + {}^t d Y$ puisque ${}^t w \hat{\theta}$ est linéaire en Y . Écrivons qu'il est sans biais :

$$\mathbb{E}\{{}^t w \hat{\theta} + {}^t d Y\} = {}^t w \theta + {}^t d X \theta = {}^t w \theta, \quad \forall \theta \in \mathbb{R}^p \quad \Leftrightarrow \quad {}^t d X = 0.$$

Calculons sa variance :

$$\begin{aligned} \text{Var}\{{}^t w \hat{\theta} + {}^t d Y\} &= \text{Var}\{{}^t w \hat{\theta}\} + \text{Var}\{{}^t d Y\} + 2\text{Cov}\{{}^t w \hat{\theta}, {}^t d Y\} \\ &= \text{Var}\{{}^t w \hat{\theta}\} + \sigma^2 {}^t d d + 2\sigma^2 {}^t w ({}^t X X)^{-1} {}^t X d \\ &= \text{Var}\{{}^t w \hat{\theta}\} + \sigma^2 {}^t d d. \end{aligned}$$

Elle est donc minimum si et seulement si $d = 0$.

Une conséquence immédiate de ce résultat est que, si W est une matrice $r \times p$ connue ($r \leq p$) et CY , un estimateur sans biais de $W\theta$, alors la matrice de variance-covariance de CY est supérieure ou égale à celle de $W\hat{\theta}$ (la différence des deux matrices est définie non-négative). C'est en particulier le cas pour $W = I$, c'est-à-dire que si $\tilde{\theta}$ est un estimateur linéaire (en Y) sans biais de θ , on a $Var(\tilde{\theta}) \geq Var(\hat{\theta})$.

2.3.3 Lois de probabilités

Lorsque les erreurs $\varepsilon_i, i = 1, \dots, n$ sont indépendantes et gaussiennes $\mathcal{N}(0, \sigma^2)$, le vecteur ε est gaussien $\mathcal{N}(0, \sigma^2 I_n)$ ainsi que le vecteur des observations Y dont la loi est $\mathcal{N}(X\theta, \sigma^2 I_n)$. Il s'en suit que les vecteurs $\hat{\theta}, \hat{Y}$ et $\hat{\varepsilon}$ sont gaussiens comme transformations linéaires de ε . Leurs moyennes et variances ont été données ci-dessus. Le vecteur formé de \hat{Y} et $\hat{\varepsilon}$ est également gaussien avec,

$$\mathbb{E} \left(\begin{bmatrix} \hat{Y} \\ \hat{\varepsilon} \end{bmatrix} \right) = \begin{bmatrix} X\theta \\ 0 \end{bmatrix}, \quad Var \left(\begin{bmatrix} \hat{Y} \\ \hat{\varepsilon} \end{bmatrix} \right) = \sigma^2 \begin{bmatrix} H & 0 \\ 0 & I - H \end{bmatrix}.$$

Les vecteurs \hat{Y} et $\hat{\varepsilon}$ sont donc indépendants. On considère une transformation orthogonale $Z = MY$ telle que les p premières composantes de Z appartiennent à $\mathcal{M}(X)$, les suivantes étant dans le sous-espace orthogonal. On a :

$$Z = MY = \begin{bmatrix} U \\ V \end{bmatrix}, \quad \mathbb{E} \left(\begin{bmatrix} U \\ V \end{bmatrix} \right) = MX\theta = \begin{bmatrix} \mathbb{E}(U) \\ 0 \end{bmatrix},$$

$$Var \left(\begin{bmatrix} U \\ V \end{bmatrix} \right) = \sigma^2 MM^T = \sigma^2 I_n = \sigma^2 \begin{bmatrix} I_p \\ I_{n-p} \end{bmatrix}.$$

L'égalité $\|V\|^2 = \|\hat{\varepsilon}\|^2 = (n-p)\hat{\sigma}^2$ montre que $(n-p)\hat{\sigma}^2/\sigma^2$ suit la loi du khi-deux à $(n-p)$ degrés de liberté.

2.3.4 Maximum de vraisemblance

Écrivons la log-vraisemblance :

$$L(y; \theta, \sigma^2) = -\frac{n}{2} \log 2\pi - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} {}^t[y - X\theta][y - X\theta].$$

Maximiser la log-vraisemblance par rapport à θ équivaut à minimiser le critère des moindres carrés. L'estimateur $\hat{\theta}$ est donc de *maximum de vraisemblance*. Par contre la maximisation par rapport à σ^2 conduit à l'estimateur biaisé,

$$\hat{\sigma}_{mv}^2 = \frac{1}{n} {}^t[y - X\theta][y - X\theta].$$

Le calcul des dérivées permet d'obtenir la *matrice d'information de Fisher*, puis son inverse, qui représente la *borne de Cramer-Rao*.

$$\frac{\partial}{\partial \theta} L(y; \theta, \sigma^2) = \frac{1}{\sigma^2} {}^t X[y - X\theta], \quad \frac{\partial}{\partial \sigma^2} L(y; \theta, \sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} {}^t[y - X\theta][y - X\theta],$$

$$\frac{\partial^2}{\partial \theta^2} L(y; \theta, \sigma^2) = -\frac{1}{\sigma^2} {}^t X X, \quad \frac{\partial^2}{\partial (\sigma^2)^2} L(y; \theta, \sigma^2) = \frac{n}{2(\sigma^2)^2} - \frac{1}{(\sigma^2)^3} {}^t[y - X\theta][y - X\theta],$$

$$\frac{\partial^2}{\partial \sigma^2 \partial \theta} L(y; \theta, \sigma^2) = -\frac{1}{(\sigma^2)^2} {}^t X[y - X\theta],$$

$$I(\theta, \sigma^2) = \frac{1}{\sigma^2} \begin{bmatrix} {}^t X X & 0 \\ 0 & \frac{n}{2\sigma^2} \end{bmatrix}, \quad I^{-1}(\theta, \sigma^2) = \sigma^2 \begin{bmatrix} ({}^t X X)^{-1} & 0 \\ 0 & \frac{2\sigma^2}{n} \end{bmatrix}.$$

Ainsi $\hat{\theta}$ est un *estimateur efficace*, puisque sa variance est égale à la borne de Cramer-Rao, alors que $\hat{\sigma}^2$ n'est qu'asymptotiquement efficace,

$$\text{Var}(\hat{\sigma}^2) = \frac{2\sigma^4}{n-p} > \frac{2\sigma^4}{n}.$$

Remarque

Dans le cas gaussien, l'estimateur $\hat{\theta}$ est donc de variance minimum parmi les estimateurs sans biais de θ puisqu'il atteint la borne de Cramer-Rao. Il n'est pas nécessaire de se restreindre à la classe des estimateurs linéaires (Gauss-Markov). On montre aussi que $\hat{\sigma}^2$ est un estimateur sans biais de variance minimum (en utilisant le Théorème de Lehmann-Scheffe).

2.4 INTERVALLES DE CONFIANCE ET TESTS

On se place dans le cas gaussien pour valider les résultats qui suivent. On peut aussi évoquer le point de vue asymptotique pour les utiliser dans le cas où les erreurs sont simplement indépendantes et identiquement distribuées.

2.4.1 Intervalles de confiance

Sur le plan pratique, il est utile de disposer d'*intervalles de confiance* pour chaque composante θ_j de θ :

$$I(\theta_j; 1 - \alpha) = \hat{\theta}_j \pm t_{n-p;\alpha} \hat{\sigma} \sqrt{({}^t X X)_{jj}^{-1}}, \quad P\{|\mathcal{S}_{n-p}^t| > t_{n-p;\alpha}\} = \alpha.$$

Ces intervalles ne sont pas indépendants, même si ${}^t X X$ est diagonale, car ils utilisent tous l'estimateur $\hat{\sigma}^2$. C'est cependant le caractère non diagonal de ${}^t X X$ qui crée la plus forte dépendance.

En toute rigueur, on peut construire un *ellipsoïde de confiance* de niveau $1 - \alpha$, pour le paramètre vectoriel θ , défini par l'inégalité,

$$t[\theta - \hat{\theta}] {}^t X X [\theta - \hat{\theta}] \leq p f_{p,n-p;\alpha} \hat{\sigma}^2, \quad P\{\mathcal{F}_{p,n-p} > f_{p,n-p;\alpha}\} = \alpha.$$

L'intervalle de confiance usuel pour σ^2 est donné par :

$$I(\sigma^2; 1 - \alpha) = \left[\frac{(n-p)\hat{\sigma}^2}{z_{n-p;\alpha/2}}, \frac{(n-p)\hat{\sigma}^2}{z_{n-p;1-\alpha/2}} \right], \quad P\{\chi_{n-p}^2 > z_{n-p;\phi}\} = \phi.$$

2.4.2 Tests

On peut utiliser le test de Student pour analyser l'influence d'un facteur X_j . Cela équivaut à vérifier si la valeur nulle est dans l'intervalle de confiance de θ_j , ou encore à considérer l'estimateur studentisé $\hat{\theta}_j^S$. Pour étudier l'influence de plusieurs facteurs simultanément, on utilise un *test de Fisher*. Il s'agit de tester que la moyenne de Y appartient au sous-espace \mathcal{W} de $\mathcal{M}(X)$ engendré par les colonnes de X associées aux θ_j non nuls. Pour cela on considère la décomposition orthogonale de Y (*cf.* Figure 2.3) :

$$Y = P_{\mathcal{W}}(Y) \oplus [P_{\mathcal{M}(X)}(Y) - P_{\mathcal{W}}(Y)] \oplus [Y - P_{\mathcal{M}(X)}(Y)],$$

où $P_{\mathcal{W}}$ et $P_{\mathcal{M}(X)}$ désignent les projecteurs orthogonaux sur les sous-espaces \mathcal{W} et $\mathcal{M}(X)$. Sous l'hypothèse nulle, les deux derniers vecteurs du second membre sont centrés, de matrice de covariance proportionnelle à l'identité, conjointement gaussiens et orthogonaux. Ils sont donc indépendants. On peut effectuer une transformation orthogonale,

$$Z = MY = \begin{bmatrix} W \\ U \\ V \end{bmatrix}, \quad \mathbb{E}(Z) = \begin{bmatrix} \mathbb{E}(W) \\ 0 \\ 0 \end{bmatrix}, \quad \text{Var}(Z) = \sigma^2 \begin{bmatrix} I_q & 0 & 0 \\ 0 & I_{p-q} & 0 \\ 0 & 0 & I_{n-p} \end{bmatrix},$$

où q est la dimension de \mathcal{W} . On a alors :

$$\|P_{\mathcal{M}(X)}(Y) - P_{\mathcal{W}}(Y)\|^2 = \|U\|^2, \quad \|Y - P_{\mathcal{M}(X)}(Y)\|^2 = \|V\|^2.$$

On en déduit que la statistique,

$$T^2 = \frac{n-p}{p-q} \frac{\|P_{\mathcal{M}(X)}(Y) - P_{\mathcal{W}}(Y)\|^2}{\|Y - P_{\mathcal{M}(X)}(Y)\|^2} = \frac{n-p}{p-q} \frac{\|\hat{\varepsilon}_{\mathcal{W}}\|^2 - \|\hat{\varepsilon}\|^2}{\|\hat{\varepsilon}\|^2} = \frac{n-p}{p-q} \frac{R^2 - R_{\mathcal{W}}^2}{1 - R^2}$$

suit, toujours sous l'hypothèse nulle, la loi de *Fisher-Snedecor* à $(p-q)$ et $(n-p)$ degrés de liberté. La région critique du test est donnée par :

$$T^2 > f_{p-q, n-p; \alpha}, \quad P\{\mathcal{F}_{p-q, n-p} > f_{p-q, n-p; \alpha}\} = \alpha.$$

La forme de cette région critique se justifie par le fait que, sous l'alternative, la statistique T^2 a tendance à être plus grande puisqu'elle l'est en moyenne avec :

$$\mathbf{IE}\{\|\hat{\varepsilon}_{\mathcal{W}}\|^2 - \|\hat{\varepsilon}\|^2\} = (p-q)\sigma^2 + {}^t\mathbf{IE}(\hat{\varepsilon}_{\mathcal{W}})\mathbf{IE}(\hat{\varepsilon}_{\mathcal{W}}).$$

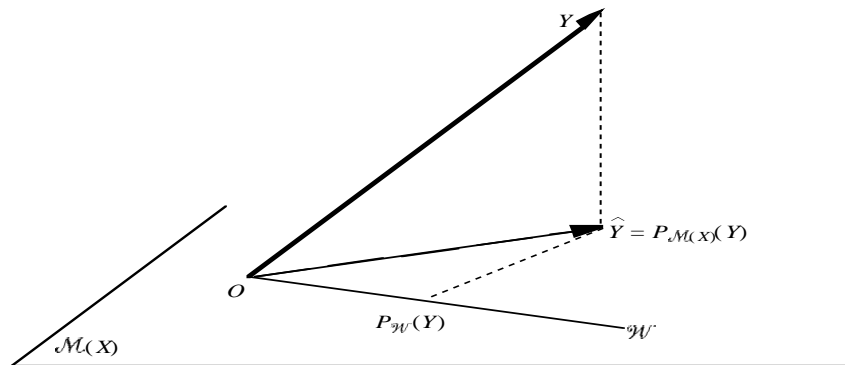


Figure 2.3: Aspect géométrique du test de Fisher

Dans le cas particulier où la constante $\mathbb{1}$ fait partie des colonnes de X , le test de non influence des régresseurs (autres que la constante) consiste à

considérer l'espace \mathcal{W} engendré par \mathbb{I} . La décomposition orthogonale de Y s'écrit :

$$Y - \bar{Y}\mathbb{I} = [P_{\mathcal{M}(X)}(Y) - \bar{Y}\mathbb{I}] \oplus [Y - P_{\mathcal{M}(X)}(Y)].$$

L'égalité portant sur le carré des normes est symbolisée comme suit :

$$\|Y - \bar{Y}\mathbb{I}\|^2 = \|P_{\mathcal{M}(X)}(Y) - \bar{Y}\mathbb{I}\|^2 + \|Y - P_{\mathcal{M}(X)}(Y)\|^2 \Leftrightarrow SST = SS_{reg} + SSE.$$

Elle donne lieu au tableau d'*analyse de la variance* :

Source de variation	Degrés de liberté	Somme des carrés	Carrés moyens
Régression	$p - 1$	SS_{reg}	$SS_{reg}/(p - 1)$
Résidus	$n - p$	SSE	$SSE/(n - p)$
Total	$n - 1$	SST	$SST/(n - 1)$

Tableau 2.2: Analyse de la variance du modèle de régression

La statistique de test est une transformation monotone du coefficient de détermination :

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

Illustration

On utilise les valeurs arrondies de la page 56. Les estimations de σ^2 et σ sont :

$$\hat{\sigma}^2 = \frac{SSE}{(n - p)} = \frac{233,7317}{25 - 3} = 10,624168, \quad \hat{\sigma} = 3,259,$$

et la lecture dans la table de Fisher-Snedecor donne $t_{n-p;\alpha} = t_{22;5\%} = 2,074$. Les intervalles de confiance à 95% pour chacun des coefficients de régression sont :

- $I(\theta_0; 95\%) = 2,341 \pm 2,074 \times 3,259\sqrt{0,11321519} = 2,341 \pm 2,274,$
- $I(\theta_1; 95\%) = 1,6159 \pm 2,074 \times 3,259\sqrt{0,00274378} = 1,6159 \pm 0,3541,$
- $I(\theta_2; 95\%) = 0,014385 \pm 2,074 \times 3,259\sqrt{0,00000123} = 0,014385 \pm 0,007496.$

Le test d'égalité à zéro de chacun de ces paramètres consiste à vérifier, pour un niveau de signification de 5%, que 0 est dans l'intervalle de confiance correspondant. À ce niveau, il apparaît que le temps (variable d'intérêt)

Y) dépend significativement du nombre de caisses (variable X_1), de la distance parcourue (variable X_2) en plus d'un effet constant (variable \mathbb{I}). Une meilleure appréciation de cette significativité est obtenue en considérant les versions studentisées de ces paramètres :

- $\theta_0^S = \frac{2,341}{3,259\sqrt{0,11321519}} = 2,134,$
- $\theta_1^S = \frac{1,6159}{3,259\sqrt{0,00274378}} = 9,4658,$
- $\theta_1^S = \frac{0,014385}{3,259\sqrt{0,00000123}} = 3,9799.$

Chacun de ces tests équivaut à tester le sous modèle correspondant par le *test de Fisher*. Par exemple le test du modèle sans la variable X_2 équivaut à celui de $\theta_2 = 0$. On a :

$$T^2 = \frac{n-p}{p-q} \times \frac{R^2 - r_1^2}{1 - R^2} = \frac{25-3}{3-2} \times \frac{0,959594 - 0,965^2}{1 - 0,959594} = 15,446,$$

à comparer avec $f_{p-q, n-p; \alpha} = f_{1, 22; 5\%} = 4,30$. On constate en effet que $3,980^2 = 15,840 \simeq 15,446$ et $2,074^2 = 4,3014 \simeq 4,30$.

2.4.3 Prévision

Les résultats de la régression linéaire simple s'étendent de façon analogue à la nouvelle situation. La *prévision* concerne la variable

$$Y_0 = {}^t x_0 \theta + \varepsilon_0, \quad {}^t x_0 = (x_{01}, \dots, x_{0p}),$$

où x_0 est donné et ε_0 est une erreur centrée, de variance σ^2 , non corrélée avec les variables $Y_i, i = 1, \dots, n$. Alors $\hat{Y}_0 = {}^t x_0 \hat{\theta}$ est l'unique estimateur linéaire sans biais pour lequel la variance de l'erreur de prévision est minimum. Elle est donnée par :

$$\text{Var}(Y_0 - \hat{Y}_0) = \text{Var}(Y_0) + \text{Var}(\hat{Y}_0) = \sigma^2 [1 + {}^t x_0 ({}^t X X)^{-1} x_0].$$

Sous l'hypothèse gaussienne, ou de façon asymptotique, l'intervalle de confiance $(1 - \alpha)$ pour la prévision (*intervalle de prévision*) est donc :

$$I(Y_0; 1 - \alpha) = {}^t x_0 \hat{\theta} \pm t_{n-p; \alpha} \hat{\sigma} \sqrt{1 + {}^t x_0 ({}^t X X)^{-1} x_0}, \quad P\{|S_{n-p}^t| > t_{n-p; \alpha}\} = \alpha.$$

2.4.4 Étude des résidus

Les *résidus* $\hat{\varepsilon}_i, i = 1, \dots, n$ sont centrés et corrélés entre eux. Leurs variances sont obtenues en fonction de la matrice de prédiction H ,

$$\text{Var}(\hat{\varepsilon}) = \sigma^2[I - H] \Leftrightarrow \text{Var}(\hat{\varepsilon}_i) = \sigma^2[1 - h_{ii}], \quad i = 1, \dots, n,$$

où $h_{ii} = {}^t x_i ({}^t X X)^{-1} x_i$ si x_i désigne la i^e ligne de la matrice X . Les *résidus standardisés* ou *résidus studentisés* sont définis par normalisation en utilisant l'estimation de σ^2 :

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

Dans le cas gaussien, $\hat{\varepsilon}_i$ est une variable normale centrée et $(n - p)\hat{\sigma}^2/\sigma^2$ suit la loi du chi-deux à $(n - p)$ degrés de liberté. Bien que ces deux variables ne soient pas indépendantes, on compare $\hat{\varepsilon}_i^S$ avec $t_{n-p;\alpha}$ pour déceler d'éventuelles données aberrantes ou, plus généralement, pour rejeter globalement le modèle.

De façon plus rigoureuse on définit les *résidus par validation croisée*,

$$\hat{\varepsilon}_i^V = \frac{Y_i - \hat{Y}_{(i)}}{\hat{\sigma}_{(i)} \sqrt{1 + {}^t x_i ({}^t X_{(i)} X_{(i)}^{-1}) x_i}},$$

où l'indice (i) signifie que la caractéristique correspondante est construite sans la i^e observation. Ainsi $\hat{Y}_{(i)}$ est la prévision de Y_i obtenue à partir des autres observations et ε_i^V est la version studentisée de l'erreur $Y_i - \hat{Y}_{(i)}$. Mais ici la loi de $\hat{\varepsilon}_i^V$ est la loi de Student à $(n - p - 1)$ degrés de liberté car l'estimateur $\hat{\sigma}_{(i)}^2$ est indépendant de $Y_i - \hat{Y}_{(i)}$. On montre que ces résidus sont obtenus en fonction des résidus studentisés initiaux par la relation :

$$\hat{\varepsilon}_i^V = \hat{\varepsilon}_i^S \sqrt{\frac{n - p - 1}{n - p - \hat{\varepsilon}_i^{S^2}}}, \quad i = 1, \dots, n.$$

Pour cela on utilise le lemme d'inversion,

$$(A + BCD)^{-1} = A^{-1} - A^{-1}B(DA^{-1}B + C^{-1})^{-1}DA^{-1},$$

qui, appliqué à la relation ${}^t X_{(i)} X_{(i)} = {}^t X X - x_i^t x_i$, donne :

$$\begin{aligned} ({}^t X_{(i)} X_{(i)})^{-1} &= ({}^t X X)^{-1} + \frac{({}^t X X)^{-1} x_i^t x_i ({}^t X X)^{-1}}{1 - {}^t x_i ({}^t X X)^{-1} x_i} \\ &= ({}^t X X)^{-1} + \frac{({}^t X X)^{-1} x_i^t x_i ({}^t X X)^{-1}}{1 - h_{ii}}. \end{aligned}$$

On établit ainsi les relations suivantes :

$$\hat{\theta}_{(i)} = \hat{\theta} - ({}^tXX)^{-1}x_i \frac{\hat{\varepsilon}_i}{1 - h_{ii}}, \quad (n - p - 1)\hat{\sigma}_{(i)}^2 = (n - p)\hat{\sigma}^2 - \frac{\hat{\varepsilon}_i^2}{1 - h_{ii}},$$

$$Y_i - \hat{Y}_{(i)} = \frac{\hat{\varepsilon}_i}{1 - h_{ii}}.$$

Il n'existe pas de différence sensible entre $\hat{\varepsilon}_i^S$ et $\hat{\varepsilon}_i^V$ lorsque n est grand et ces résidus sont souvent proches de la normalisation brutale $\hat{\varepsilon}_i/\hat{\sigma}$ assimilée à une loi $\mathcal{N}(0, 1)$. La différence est notable lorsque h_{ii} est proche de 1 et $\hat{\varepsilon}_i^V$ accentue alors les valeurs de $\hat{\varepsilon}_i^S$ anormalement grandes.

Illustration

Il n'est pas possible de visualiser les observations par rapport à l'hyperplan de régression. On peut représenter les chroniques des différentes formes de résidus. La Figure 2.4 est constituée des points $(i, \hat{\varepsilon}_i^V)$, $i = 1, \dots, n$ que l'on joint par des segments de droites. La Figure 2.5 donne les résidus $\hat{\varepsilon}_i^S$ en fonction des valeurs ajustées \hat{y}_i , $i = 1, \dots, n$ de la variable Y .

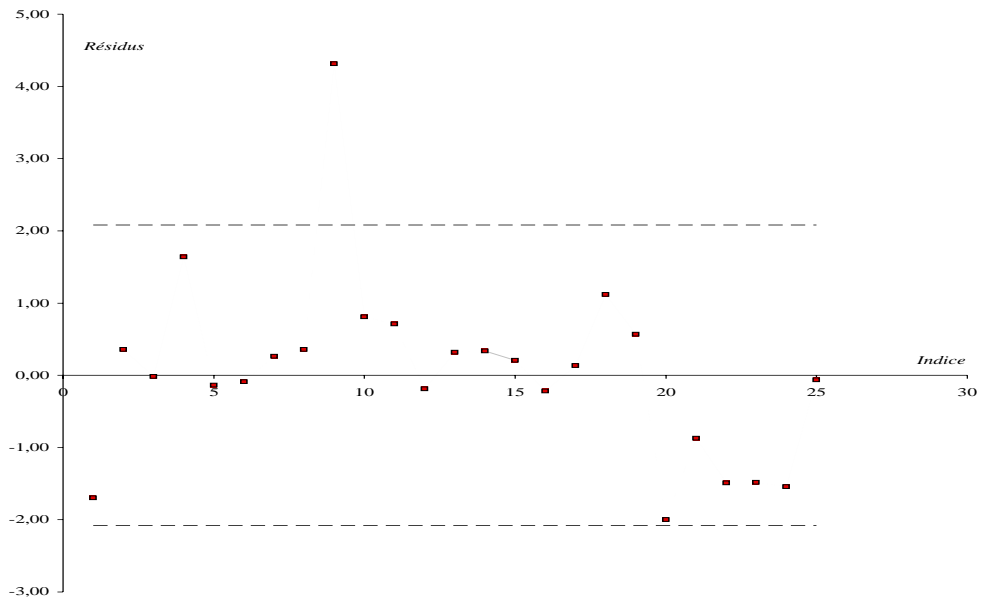


Figure 2.4: Chronique des résidus par validation croisée pour le réseau de distributeurs

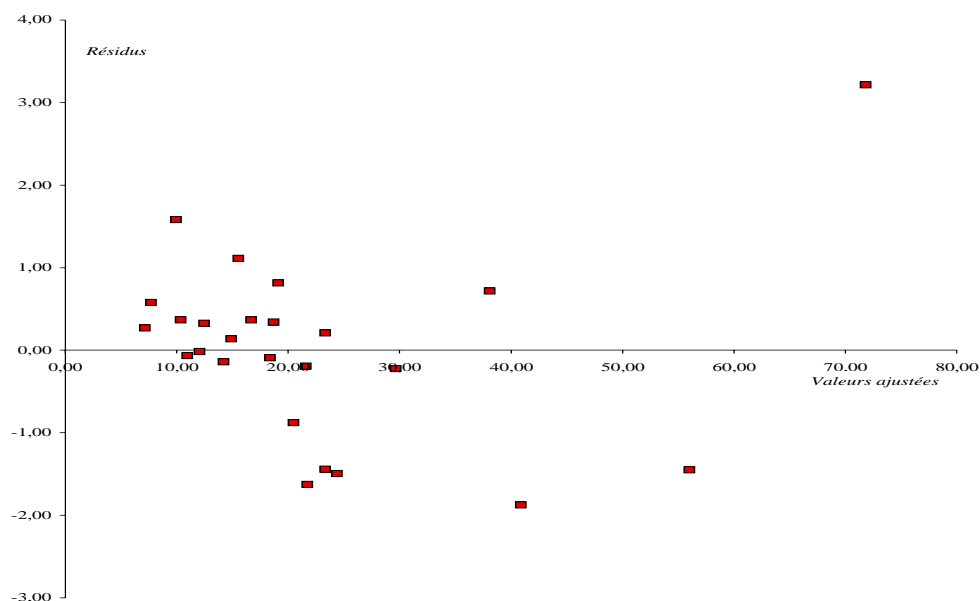


Figure 2.5: Résidus standardisés en fonction des valeurs ajustées pour le réseau de distributeurs de boissons

2.4.5 Étude des coefficients de régression

L'estimation de θ par le critère des moindres carrés est globale. Il est difficile d'apprécier la part de chaque variable explicative X_j à travers les coefficients de régression $\hat{\theta}_j$ dans la prévision de Y par $\hat{Y} = \sum_{j=1}^p \hat{\theta}_j X_j$. En effet les vecteurs X_j , dans l'espace des observations \mathbb{R}^n , sont de normes différentes et peuvent être fortement dépendants entre eux. La situation idéale serait qu'ils constituent une base orthonormée de $\mathcal{M}(X)$. Ce n'est pas le cas en pratique et la suppression de l'un d'entre eux dans le plan d'expérience ne détériore pas forcément de façon notable la prévision. Nous reviendrons sur ce point dans la section suivante.

Pour tout $j, 1 \leq j \leq p$, on note $X_{[j]}$ le plan d'expérience à $(p-1)$ variables obtenu en supprimant la colonne X_j dans X . On considère alors les différents éléments obtenus dans ce nouveau modèle :

$$\hat{Y}_{[j]} = X_{[j]} \hat{\theta}_{[j]}, \quad \hat{\varepsilon}_{[j]} = Y - \hat{Y}_{[j]}.$$

Le vecteur $\hat{\varepsilon}_{[j]}$, qui représente la "partie" de Y non expliquée linéairement par les $(p-1)$ variables autres que X_j , est appelé j^e variable de réponse ajustée. On effectue de même la régression de X_j sur $X_{[j]}$ et on note $e_{[j]}$

l'erreur correspondante. Ce vecteur est appelé j^e régresseur ajusté. Un simple résultat de projection donne :

$$\hat{Y} = \sum_{j=1}^p \hat{\theta}_j X_j = \hat{Y}_{[j]} \oplus \frac{\langle Y, e_{[j]} \rangle}{\|e_{[j]}\|^2} e_{[j]}.$$

Par identification des coefficients de X_j , on obtient :

$$\hat{\theta}_j = \frac{\langle Y, e_{[j]} \rangle}{\|e_{[j]}\|^2} = \frac{\langle \hat{\varepsilon}_{[j]}, X_j \rangle}{\|e_{[j]}\|^2}.$$

On en déduit les relations suivantes :

$$\hat{Y} = \hat{Y}_{[j]} \oplus \hat{\theta}_j e_{[j]}, \quad \hat{\varepsilon} = \hat{\varepsilon}_{[j]} - \hat{\theta}_j e_{[j]}, \quad \|\hat{\varepsilon}\|^2 = \|\hat{\varepsilon}_{[j]}\|^2 - \hat{\theta}_j^2 \|e_{[j]}\|^2.$$

On dispose donc immédiatement de l'apport de la variable X_j , sans effectuer un nouveau calcul de régression car $\|e_{[j]}\|^2 = 1/({}^t X X)_{jj}^{-1}$. Notons que le cosinus de l'angle entre les deux vecteurs $\varepsilon_{[j]}$ et $e_{[j]}$,

$$\pi_j = \frac{\langle \varepsilon_{[j]}, e_{[j]} \rangle}{\|\varepsilon_{[j]}\| \times \|e_{[j]}\|} = \hat{\theta}_j \frac{\|e_{[j]}\|}{\|\varepsilon_{[j]}\|},$$

est un indicateur de la liaison linéaire qui subsiste entre Y et X_j lorsque l'on a éliminé celle due aux autres variables. Si la constante \mathbb{I} fait partie du plan d'expérience, π_j est la corrélation empirique entre $\varepsilon_{[j]}$ et $e_{[j]}$. On dit qu'il s'agit de la *corrélation partielle* (empirique) entre Y et X_j (autre que \mathbb{I}) dans l'ensemble $\{Y, X_1, \dots, X_p\}$. Dans ce cas on peut noter la décomposition du coefficient de détermination sous la forme :

$$R^2 = R_{[j]}^2 + (1 - R_{[j]}^2) \pi_j^2.$$

Ainsi la proximité à zéro de π_j^2 va dans le sens d'exclure la variable X_j du plan d'expérience, ce que devrait confirmer le test de l'hypothèse $\theta_j = 0$. Par contre sa proximité à 1 ne signifie pas que cette variable joue un rôle important.

Illustration

Les Figures 2.6 et 2.7 représentent la régression du temps sur le nombre de caisses d'une part et sur la distance parcourue d'autre part, en supprimant donc l'autre variable. On constate que le coefficient de détermination est peu affecté : 93% et 80% au lieu de 96%. Les coefficients de corrélation partielle sont $\pi_1 = 0,896$ et $\pi_2 = 0,647$.

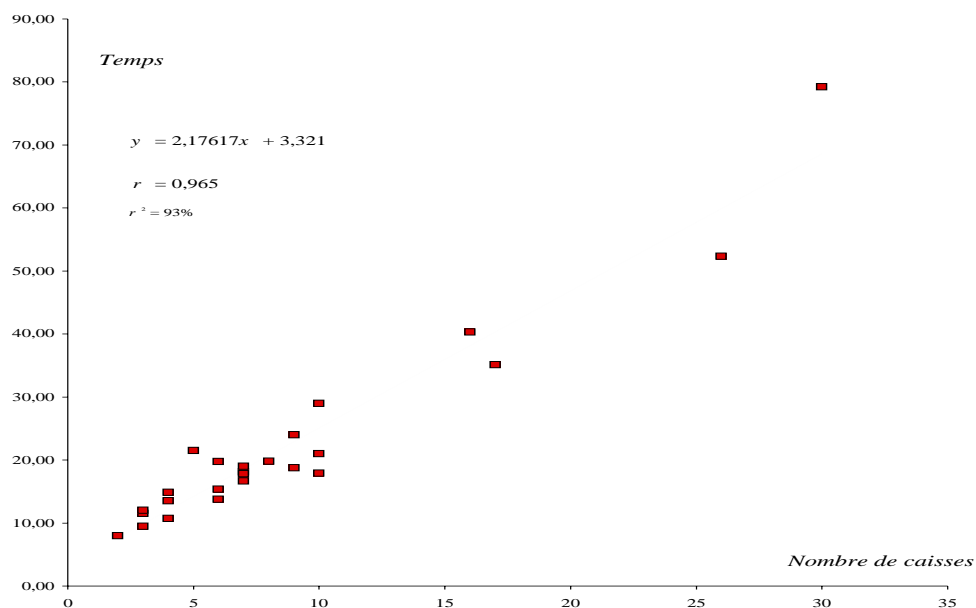


Figure 2.6: Temps en fonction du nombre de caisses pour le réseau de distributeurs de boissons

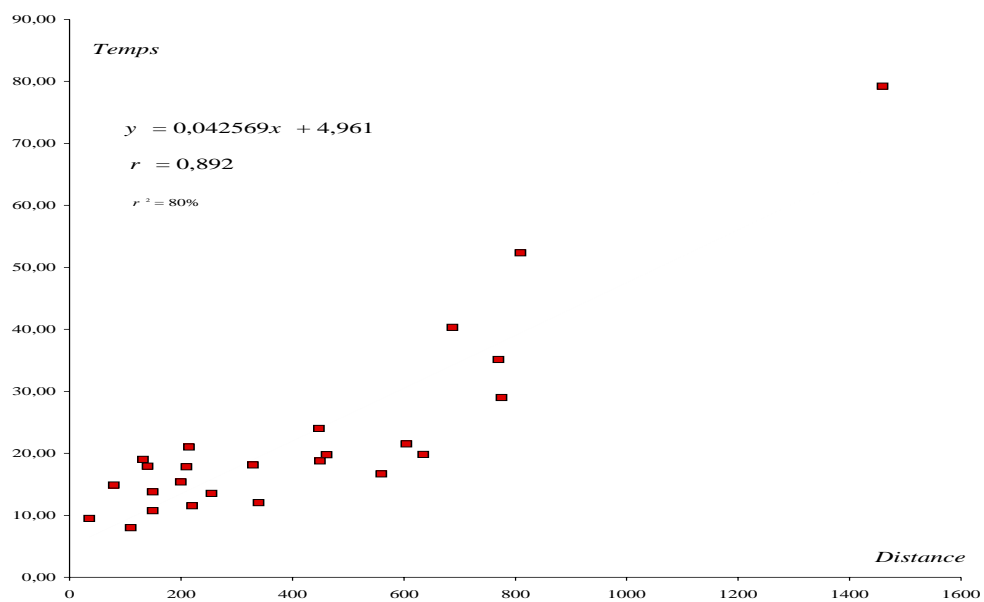


Figure 2.7: Temps en fonction de la distance pour le réseau de distributeurs de boissons

2.4.6 Étude du plan d'expérience

Le plan d'expérience idéal est celui pour lequel les colonnes de X forment une base orthonormée de $\mathcal{M}(X)$ puisque dans ce cas on a :

$$\hat{Y} = \sum_{j=1}^p \hat{\theta}_j X_j, \quad \hat{\theta}_j = \langle Y, X_j \rangle, j = 1, \dots, p, \quad \|\hat{Y}\|^2 = \sum_{j=1}^p \hat{\theta}_j^2.$$

Les variables n'interagissent pas dans la reconstitution de Y et il est immédiat d'apprécier l'influence de chacune d'elles. Il est toujours possible de se ramener à cette situation par changement de variables sans changer l'espace $\mathcal{M}(X)$. Mais les nouvelles variables, ainsi que les paramètres correspondants, peuvent perdre toute interprétation dans le contexte initial du problème. Ainsi en pratique il est fréquent de rencontrer des variables explicatives fortement dépendantes.

Illustration

La Figure 2.8 représente la régression du nombre de caisses placées sur la

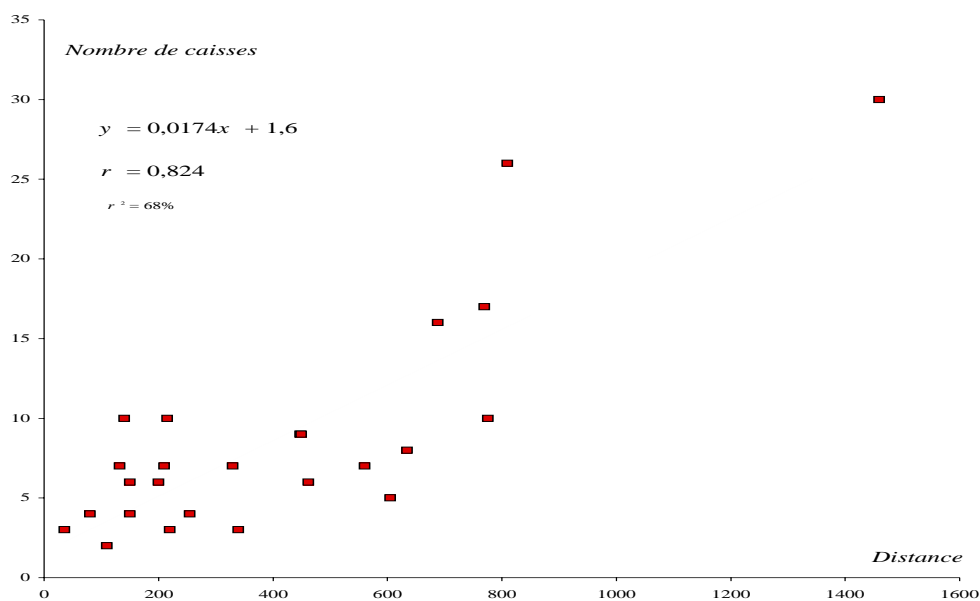


Figure 2.8: Nombre de caisses en fonction de la distance pour le réseau de distributeurs de boissons

distance parcourue. On constate que ces deux variables explicatives sont moyennement dépendantes ($r^2 = 68\%$).

La structure du plan d'expérience s'analyse à travers les éléments de la *matrice chapeau* $H = X({}^tXX)^{-1}{}^tX$ (cf. Antoniadis et al. p. 37). Les expressions,

$$\hat{Y}_i = \sum_{k=1}^n h_{ik}Y_k = h_{ii}Y_i + \sum_{k \neq i}^n h_{ik}Y_k, \quad i = 1, \dots, n,$$

montrent que h_{ii} représente le poids de l'observation Y_i dans sa propre prédiction \hat{Y}_i . On établit les propriétés suivantes :

$$\sum_{i=1}^n h_{ii} = p, \quad 0 \leq h_{ii} \leq 1, i = 1, \dots, n, \quad h_{ii} = 1 \text{ ou } 0 \Rightarrow h_{ik} = 0, k \neq i.$$

Ainsi h_{ii} devrait en moyenne être proche de p/n . Lorsque $h_{ii} > p/n$, l'observation correspondante est considérée comme *influyente*.

La *distance de Cook* (cf. Antoniadis et al. p. 40) est une mesure de cette influence. Rappelons l'ellipsoïde de confiance du paramètre θ :

$$I\{\theta; 1 - \alpha\} = \{\theta : {}^t(\theta - \hat{\theta})^tXX(\theta - \hat{\theta}) \leq pf_{p,n-p;\alpha}\hat{\sigma}^2\}.$$

La distance de Cook, notée C_i , mesure le décentrage de l'estimation de θ résultant de la suppression de la i^e observation :

$$C_i = \frac{{}^t(\theta_{(i)} - \hat{\theta})^tXX(\theta_{(i)} - \hat{\theta})}{p\hat{\sigma}^2}.$$

Elle se calcule dans le cadre du modèle complet par :

$$C_i = \frac{1}{p} \times \frac{h_{ii}}{1 - h_{ii}} \hat{\varepsilon}_i^2 = \frac{1}{p} \times \frac{h_{ii}}{(1 - h_{ii})^2 \hat{\sigma}^2} \hat{\varepsilon}_i^2.$$

Plus C_i est grande, plus la i^e observation a une influence simultanée sur l'ensemble des paramètres du modèle. Cook suggère de comparer C_i à $f_{p,n-p;\alpha}$ bien que la loi ne soit pas exacte.

2.5 RÉGRESSION POLYNOMIALE

2.5.1 Illustration

On illustre tout d'abord la *régression polynomiale* en construisant, par simulation, la parabole bruitée suivante :

$$y_i = 10 - x_i + 0,2x_i^2 + \varepsilon_i, \quad i = 1, \dots, 25,$$

i	y_i	\mathbb{I}	x_i	x_i^2	\hat{y}_i	$\hat{\varepsilon}_i$	$\hat{\varepsilon}_i^S$	ε_i^V	h_{ii}
1	10,39	1	1	1	7,04	3,35	1,44	1,47	0,1018
2	5,99	1	2	4	6,91	-0,92	-0,37	-0,37	0,0707
3	6,30	1	3	9	7,17	-0,87	-0,34	-0,33	0,0987
4	7,28	1	4	16	7,81	-0,53	-0,20	-0,20	0,0854
5	5,18	1	5	25	8,83	-3,65	-1,36	-1,39	0,0750
6	6,87	1	6	36	10,22	-3,35	-1,25	-1,26	0,0429
7	15,67	1	7	49	12,00	3,67	1,36	1,39	0,0818
8	11,57	1	8	64	14,17	-2,60	-0,96	-0,96	0,0637
9	18,59	1	9	81	16,71	1,88	0,70	0,69	0,4983
10	24,97	1	10	100	19,63	5,34	1,99	2,14	0,1963
11	25,61	1	11	121	22,93	2,68	1,00	1,00	0,0861
12	25,38	1	12	144	26,62	-1,24	-0,46	-0,45	0,1137
13	27,00	1	13	169	30,69	-3,69	-1,38	-1,41	0,0611
14	37,00	1	14	196	35,13	1,87	0,70	0,69	0,0782
15	37,30	1	15	225	39,96	-2,66	-0,99	-0,99	0,0411
16	46,67	1	16	256	45,17	1,50	0,56	0,55	0,1659
17	50,88	1	17	289	50,76	0,12	0,04	0,04	0,0594
18	52,42	1	18	324	56,73	-4,31	-1,60	-1,66	0,0963
19	64,67	1	19	361	63,09	1,58	0,59	0,58	0,0964
20	72,35	1	20	400	69,82	2,53	0,94	0,94	0,1017
21	75,68	1	21	441	76,94	-1,26	-0,47	-0,46	0,1653
21	75,68	1	21	441	76,94	-1,26	-0,47	-0,46	0,1653
22	83,33	1	22	484	84,43	-1,10	-0,42	-0,41	0,3916
23	94,52	1	23	529	92,31	2,21	0,86	0,85	0,0413
24	102,82	1	24	576	100,57	2,25	0,91	0,91	0,1206
25	106,42	1	25	625	109,21	-2,79	-1,19	-1,21	0,0666

Tableau 2.3: Parabole bruitée

considérée sur l'intervalle $[1; 25]$ avec un bruit uniforme sur $[-5; 5]$ ($\sigma^2 = 25/3$). Le Tableau 2.3 présente l'ensemble des données et les principaux résultats. Le plan d'expérience est $X = (\mathbb{I} \ x \ x^2)$ et le paramètre est noté $\theta = (\theta_0 \ \theta_1 \ \theta_2)^T$. On a les résultats suivants :

$${}^tXX = \begin{bmatrix} 25 & 325 & 5525 \\ 325 & 5525 & 105625 \\ 5525 & 105625 & 2153645 \end{bmatrix}, \quad [{}^tXX]^{-1} = \begin{bmatrix} 0,42434783 & -0,06652174 & -0,00217391 \\ -0,06652174 & 0,01332962 & -0,00048309 \\ -0,00217391 & -0,00048309 & 0,00001858 \end{bmatrix},$$

$${}^tXY = \begin{bmatrix} 1014,86 \\ 18727,19 \\ 378422,77 \end{bmatrix}, \quad \hat{\theta} = \begin{bmatrix} 7,54658261 \\ -0,6969098 \\ 0,19053233 \end{bmatrix}.$$

- Équation du polynôme ajusté : $y = 7,547 - 0,69691x + 0,190532x^2$
- Analyse de la variance : $SST = SS_{reg} + SSE \rightarrow 25684,7280 = 25511,7017 + 173,0263$
- Coefficient de détermination : $R^2 = 0,993263 \simeq 99,3\%$
- Statistique de test : $F = \frac{(SST - SSE)/(p-1)}{SSE/(n-p)} = \frac{R^2/(p-1)}{(1-R^2)/(n-p)} = 1621,77, \quad f_{2;22;5\%} = 3,44$

Les Figures 2.9 et 2.10 représentent les différentes régressions de la parabole bruitée Y en fonction de x et/ou de son carré x^2 . On constate, dans cette situation, que le carré de la variable x (et la constante) est suffisant pour représenter la variable d'intérêt Y . Ceci est cohérent avec le fait que le

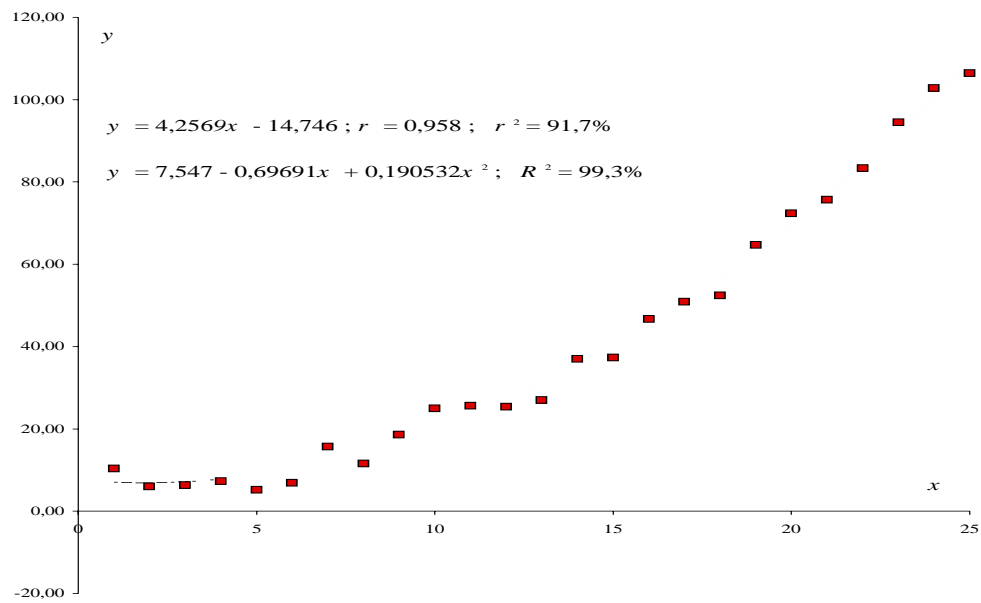


Figure 2.9: Régression simple et ajustement d'une parabole sur une parabole bruitée

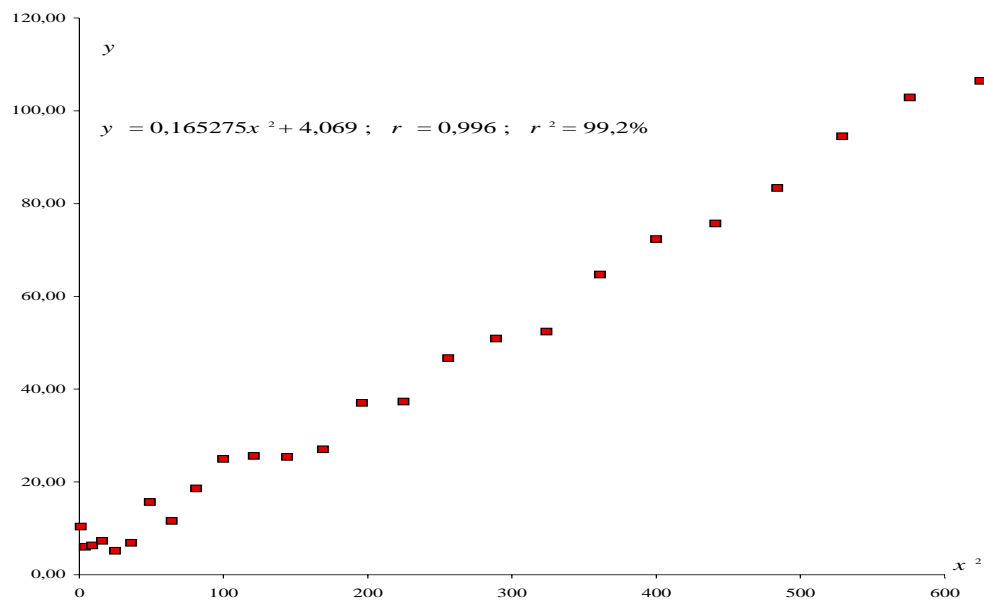


Figure 2.10: Régression simple sur le carré de x pour une parabole bruitée

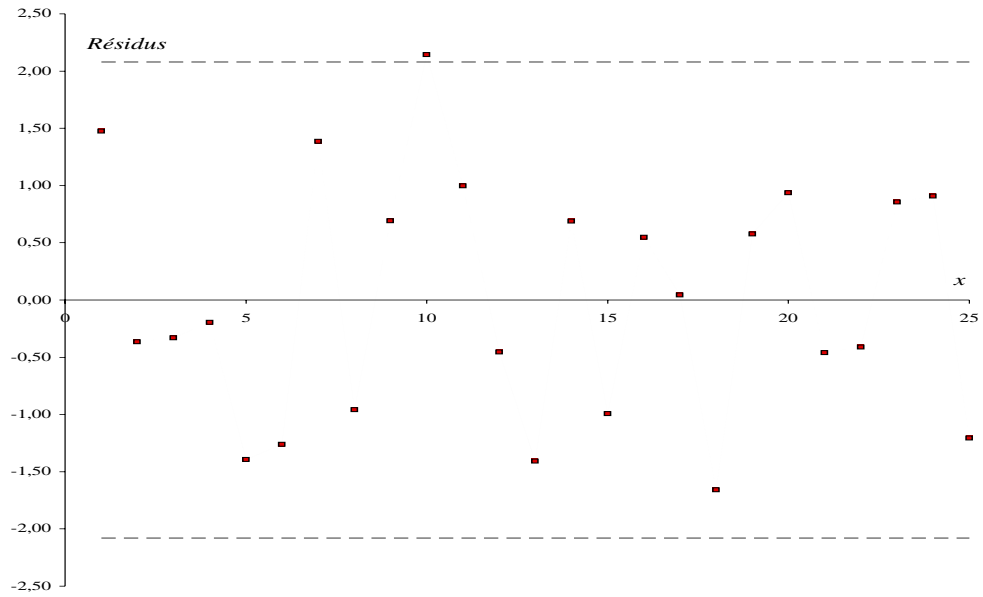


Figure 2.11: Résidus par validation croisée pour une parabole bruitée

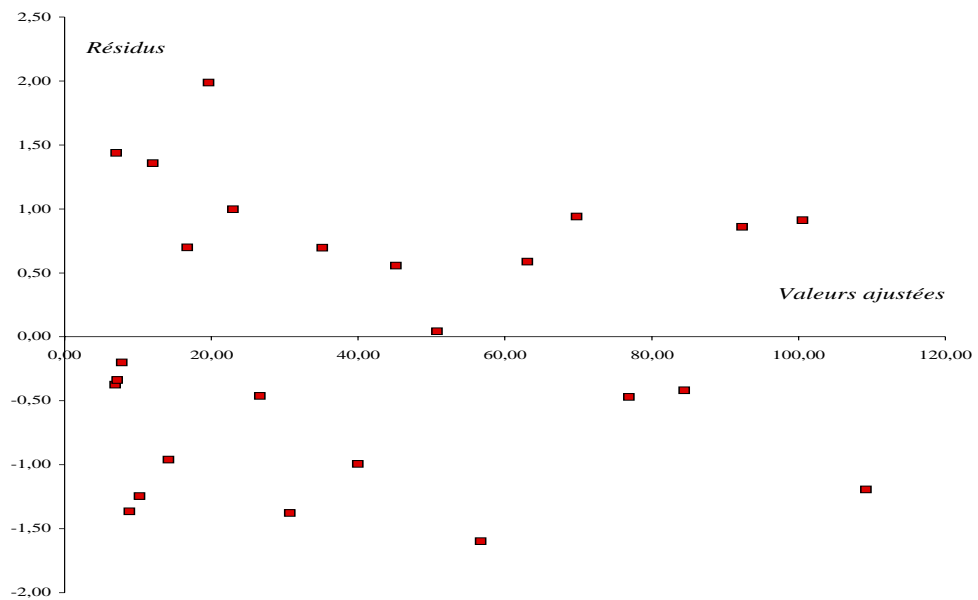
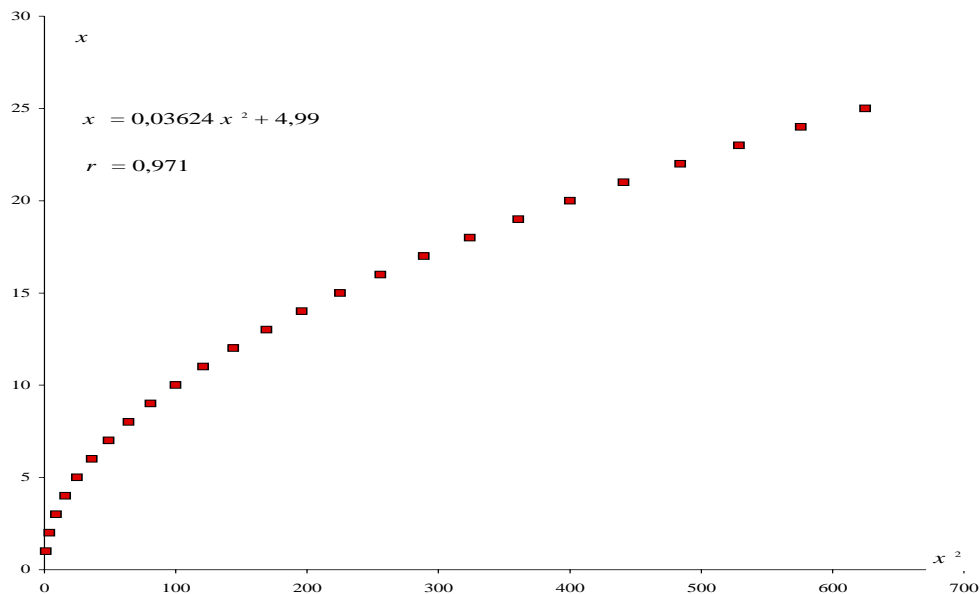


Figure 2.12: Résidus standardisés en fonction des valeurs ajustées pour une parabole bruitée

Figure 2.13: Régression de x en fonction de son carré x^2

coefficient θ_1 de x n'est pas très significativement différent de 0. On a en effet les résultats suivants :

$$\hat{\sigma}^2 = \frac{173,0263}{22} = 7,864832, \quad \hat{\sigma} = 2,804, \quad t_{22;5\%} = 2,074,$$

$$\theta_0^S = \frac{7,547}{2,804 \times \sqrt{0,424348}} = 4,132, \quad \theta_1^S = \frac{-0,69691}{2,804 \times \sqrt{0,0133296}} = -2,153,$$

$$\theta_2^S = \frac{0,190532}{2,804 \times \sqrt{0,00001858}} = 15,764.$$

Les résidus sont représentés sur les Figures 2.11 et 2.12 et la Figure 2.13 rend compte de la régression de x sur x^2 . Ces deux variables (sur cette plage de valeurs) sont fortement corrélées. Il est donc naturel qu'une seule, en l'occurrence x^2 suffise pour expliquer Y .

2.5.2 Généralités

On peut ajuster, dans le cadre du modèle linéaire, un polynôme de degré q en une variable,

$$y_i = \sum_{j=0}^q \theta_j x_i^j, \quad i = 1, \dots, n.$$

On peut aussi considérer la situation de plusieurs variables,

$$y_i = \theta_0 + \theta_1 x_i + \theta_2 z_i + \theta_3 x_i z_i + \varepsilon_i, \quad i = 1, \dots, n.$$

Il est clair que le degré des polynômes et le nombre de variables doivent rester faibles (principe de parcimonie) par rapport au nombre n d'observations. Un polynôme de degré $n - 1$ interpole exactement les observations! Il est préférable de se ramener à un degré faible par transformations sur les variables. Pour choisir le degré q du polynôme, on procède par étapes successives à l'aide de tests de Student sur les coefficients, soit de façon ascendante (sélection progressive) à partir du degré 1, soit en reculant à partir d'un degré plus élevé (élimination rétrograde). La procédure n'est cependant pas très rigoureuse bien qu'il soit possible de contrôler la probabilité de certaines erreurs (*cf.* section suivante). En général le degré retenu doit être faible (2 ou 3 maximum). La régression polynomiale est surtout une approximation de relations non linéaires. En particulier elle ne doit pas être utilisée pour faire de la prédiction en dehors du domaine d'observation des variables explicatives. Il est possible de procéder à des ajustements par des polynômes locaux (fonctions splines, *cf.* Montgomery & Peck).

2.5.3 Régression orthogonale

Les polynômes $1, x, x^2, \dots$, sont naturellement ordonnés par ordre croissant de leur degré mais ne sont pas orthogonaux entre eux dans \mathbb{R}^n . Le procédé d'*orthogonalisation de Gram-Schmidt* construit un nouveau système de polynômes, $\phi_0(x), \phi_1(x), \phi_2(x), \dots$, qui sont orthogonaux entre eux et normés (restreints aux valeurs x_1, \dots, x_n) :

$$\phi_0(x_i) = \frac{1}{\sqrt{n}}, \quad \phi_1(x_i) = \frac{x_i - \bar{x}}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}, \quad \dots, \phi_j(x_i) = \sum_{k=0}^j a(j, k) x_i^k, \dots$$

Le calcul numérique des coefficients $a(j, k)$ ne présente aucune difficulté. Notons \tilde{X} la matrice $n \times (r + 1)$ définie par les régresseurs $1, x, \dots, x_r$ et X , également $n \times (r + 1)$, celle des régresseurs orthonormés $\phi_0(x), \phi_1(x), \dots, \phi_r(x)$ associés. Soit A la matrice carrée d'ordre $(r + 1)$, triangulaire inférieure, obtenue en rangeant en lignes les coefficients des polynômes $\phi_j(x)$:

$$A(j + 1, k + 1) = a(j, k), \quad k = 0, \dots, j \quad j = 0, \dots, r.$$

On a alors :

$$\tilde{X}A^T = X, \quad A({}^t\tilde{X}\tilde{X})A^T = {}^tXX = I_{r+1}, \quad {}^t\tilde{X}\tilde{X} = A^{-1}A^{-T}, \quad ({}^t\tilde{X}\tilde{X})^{-1} = {}^tAA.$$

Le calcul de A passe ainsi par la *décomposition de Cholesky* de $({}^t\tilde{X}\tilde{X})^{-1}$, c'est le procédé souvent utilisé pour calculer l'inverse de cette matrice.

Le modèle ainsi restructuré s'écrit :

$$Y_i = \sum_{j=0}^q \alpha_j x_i^j + \varepsilon_i = \sum_{j=0}^q \theta_j \phi_j(x_i) + \varepsilon_i, \quad i = 1, \dots, n, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I_n).$$

L'intérêt de la méthode est que le sous-modèle faisant intervenir un polynôme de degré $q - 1$ est obtenu en posant $\theta_q = 0$ dans la deuxième écriture sans remettre en cause les valeurs de θ_j pour $j < q$. Ainsi on se fixe un degré maximal r pour lequel on détermine l'estimateur du paramètre vectoriel θ dans le modèle $Y = X\theta + \varepsilon$, puis on teste successivement les hypothèses $\theta_r = 0, \theta_{r-1} = 0, \dots$, jusqu'à ce que l'hypothèse nulle $\theta_q = 0$ soit rejetée et q est alors le degré du polynôme retenu. L'estimation de q est donc réalisée à travers un test d'hypothèses multiples.

Précisons les éléments de cette méthode. Pour une matrice X de rang $r + 1$, il suffit de reprendre les résultats du modèle linéaire en tenant compte de la relation ${}^tXX = I_{r+1}$. On a :

$$\hat{\theta} = {}^tXY, \quad \theta_j = \sum_{i=1}^n \phi_j(x_i)Y_i, \quad j = 0, \dots, r, \quad \hat{\sigma}_r^2 = \frac{1}{n - r - 1} \left[{}^tYY - \sum_{j=0}^r \hat{\theta}_j^2 \right],$$

où l'indice r dans $\hat{\sigma}_r^2$ rappelle qu'il s'agit de l'estimateur sans biais de σ^2 sous l'hypothèse d'une régression polynomiale de degré r . Pour $j = r, r - 1, \dots, 0$, on dispose du test de niveau de signification α_j , défini par la région critique ($Var(\hat{\theta}_j) = \sigma^2$)

$$\frac{|\hat{\theta}_j|}{\hat{\sigma}_j} > t_{n-j-1; \alpha_j}, \quad P\{|\mathcal{S}_{n-j-1}^t| > t_{n-j-1; \alpha_j}\} = \alpha_j,$$

pour tester l'hypothèse $\theta_j = 0$ contre l'alternative $\theta_j \neq 0$ dans le cadre du modèle où la régression est un polynôme de degré au plus j . Notons H_j l'hypothèse selon laquelle la régression est un polynôme de degré j . Cette hypothèse est retenue lorsque l'on a :

$$\frac{|\hat{\theta}_k|}{\hat{\sigma}_k} < t_{n-k-1; \alpha_k}, \quad k = r, r - 1, \dots, j + 1, \quad \frac{|\hat{\theta}_j|}{\hat{\sigma}_j} > t_{n-j-1; \alpha_j}.$$

La probabilité de décider que le polynôme est de degré j , alors qu'il est d'un degré inférieur, est donnée par :

$$p_j = P\{H_j | H_0 \cup H_1 \cup \dots \cup H_{j-1}\} = \alpha_j \prod_{k=r}^{j+1} (1 - \alpha_k), \quad j = r - 1, \dots, 0.$$

Celle de décider que le degré est supérieur ou égal à j , alors qu'il est inférieur, est la somme $p_j + p_{j+1} + \dots + p_r$. Notons que cette façon de procéder est équivalente à celle consistant à effectuer les tests sur la paramétrisation initiale.

2.6 APPLICATION AUX CHRONIQUES

Une chronique, ou série chronologique, est une suite d'observations $y_t, t = 1, \dots, T$, d'une grandeur scalaire effectuées à intervalles réguliers au cours du temps. Il existe une théorie statistique complète consacrée à ce type de données temporelles. Nous ne considérons ici que les aspects qui relèvent directement du modèle linéaire.

Les observations sont modélisées comme celles de variables aléatoires $Y_t, t = 1, \dots, T$ satisfaisant :

$$Y_t = f_t + S_t + \varepsilon_t, \quad \mathbb{E}(Y_t) = f_t + S_t, \quad \mathbb{E}(\varepsilon_t \varepsilon_s) = \delta_{ts} \sigma^2.$$

Ainsi seule la moyenne $\mathbb{E}(Y_t)$ traduit la partie structurée du phénomène. Elle se décompose en deux parties : la *tendance* ou *composante fondamentale* f_t représente l'évolution lente de la grandeur (polynôme de degré faible) et la *composante saisonnière* S_t prend en compte les aspects périodiques.

2.6.1 Modèle de Buys-Ballot

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 :

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

L'hypothèse gaussienne n'est utilisée que pour les aspects de statistique inférentielle (intervalles de confiance, tests, ...). Il est clair que ce modèle rentre dans le cadre de la régression linéaire multiple. Mais, en supposant que la série est observée pendant n "années" de p "mois", il est possible de conduire de façon directe l'essentiel de l'étude comme en régression linéaire simple. On pose :

$$Y_{ij} = \alpha[p(i-1) + j] + \beta_j + \varepsilon_{ij}, \quad \beta_j = \beta + S_j, \quad j = 1, \dots, p, i = 1, \dots, n,$$

avec la contrainte $\sum_{j=1}^p S_j = 0$. Les coefficients saisonniers $S_j, j = 1, \dots, p$, caractérisent la composante périodique S_t dont l'effet annuel moyen est nul. Ainsi la partie déterministe du modèle est décrite par $p + 1$ paramètres linéairement indépendants : $\alpha, \beta_1, \dots, \beta_p$. La méthode des moindres carrés consiste à chercher, parmi les chroniques $x_{ij}(a, b_1, \dots, b_p) = a[p(i-1) + j] + b_j$, composées d'une tendance linéaire et d'un mouvement saisonnier périodique, celle qui est la plus proche de l'observation selon le critère des moindres carrés :

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

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

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

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

Le critère à minimiser se scinde en deux parties :

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

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 :

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

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

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

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

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

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

où l'on a posé :

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

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

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

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

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

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

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

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

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 $(\bar{t}_{i.}, \bar{y}_{i.}), i = 1, \dots, n$, c'est-à-dire que $\hat{\alpha}$ et $\hat{\beta}$ sont solution du problème de minimisation :

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

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

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

année	mois	janv	fév	mars	avr	mai	juin	juil	août	sept	oct	nov	déc	moy. an.	
	j	1	2	3	4	5	6	7	8	9	10	11	12		
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

Tableau 2.4: Chiffre d'affaires mensuel de la presse parisienne dans une petite ville de province (unité : 1KF)

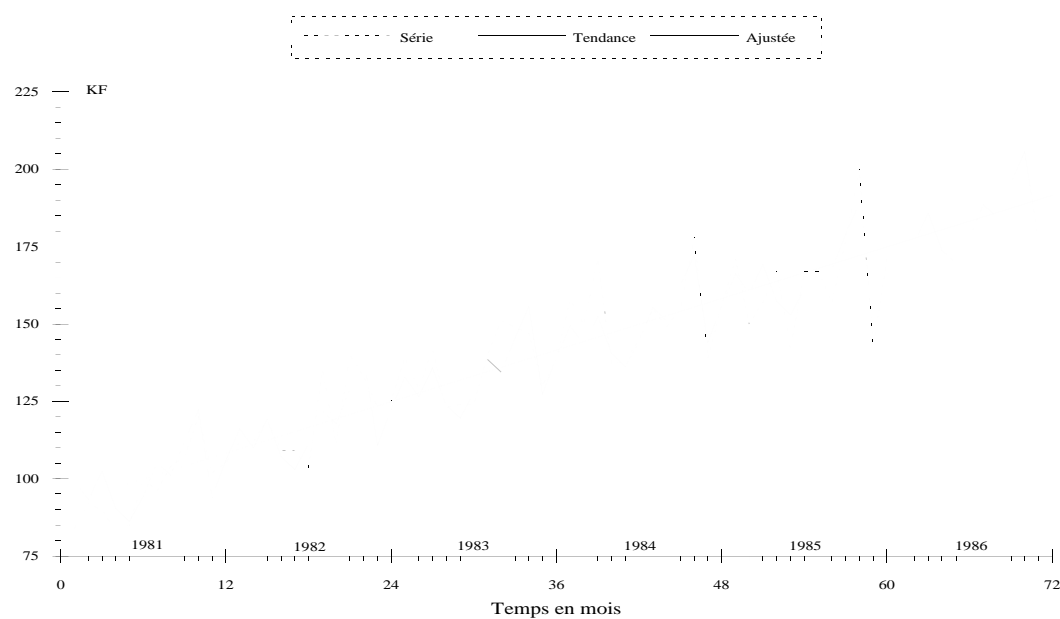


Figure 2.14: Chiffre d'affaires mensuel de la presse parisienne dans une petite ville de province

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 2.4). Celles-ci sont évidemment données par :

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

La série ajustée correspond aux années $i = 1, \dots, n$ et la prévision à l'année $i = n+1$. La représentation graphique de ces résultats permet de se faire une première idée de la capacité de ce modèle à rendre compte des observations (*cf.* Figure 2.14). On constate que la série observée est au dessous de la série ajustée aux deux extrémités alors qu'elle est au dessus dans la partie centrale. L'étude des résidus (*cf.* Figure 2.15) permet de mieux visualiser ce phénomène et invite à ajuster une tendance parabolique (*cf.* Paragraphe 2.6.4). 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 2.6.3) 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.

2.6.2 Moyenne et variance des estimateurs

Les propriétés des estimateurs obtenus à la section précédente résultent de celles du modèle linéaire général qui fera l'objet du Paragraphe 2.6.4. Ils sont sans biais et de variance minimum parmi les estimateurs sans biais qui sont linéaires en les observations (propriété de Gauss-Markov). Cependant la particularité du modèle de Buys-Ballot permet de donner des expressions plus explicites de leurs variances.

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

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

L'estimateur de la pente s'écrit :

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

Il est sans biais et sa variance est donnée par :

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

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

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

est également sans biais, il est corrélé avec $\hat{\alpha}$ et, en utilisant les expressions en fonction des $\bar{\varepsilon}_i$ ou la non corrélation entre $\hat{\alpha}$ et $\bar{y}_{..}$, on obtient :

$$\text{Var}(\hat{\beta}) = \frac{\sigma^2}{np} \left[1 + \frac{3(np + 1)^2}{p^2(n^2 - 1)} \right], \quad \text{Cov}(\hat{\alpha}, \hat{\beta}) = -\frac{\sigma^2}{np} \frac{6(np + 1)}{p^2(n^2 - 1)}.$$

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

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

qui satisfont :

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

Les estimateurs des coefficients saisonniers s'écrivent :

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

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

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

Notons que $\text{Var}(\hat{S}_j)$ est symétrique par rapport au milieu de l'année, où elle est minimum, et augmente lorsque l'on s'en écarte. Ces estimateurs sont évidemment corrélés entre eux puisque leur somme est nulle et on a :

$$\text{Cov}(\hat{S}_j, \hat{S}_k) = -\frac{\sigma^2}{np} \left[1 - \frac{12[j - (p + 1)/2][k - (p + 1)/2]}{p^2(n^2 - 1)} \right],$$

$$\begin{aligned} \text{Cov}(\hat{S}_j, \hat{\alpha}) &= -\frac{\sigma^2}{np} \frac{12[j - (p+1)/2]}{p^2(n^2-1)}, \\ \text{Cov}(\hat{S}_j, \hat{\beta}) &= \frac{\sigma^2}{np} \frac{6[j - (p+1)/2](np+1)}{p^2(n^2-1)}. \end{aligned}$$

La série ajustée et la prévision s'écrivent :

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

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

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

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. Plus généralement on a :

$$\text{Cov}(\hat{Y}_{ij}, \hat{Y}_{kl}) = \frac{\sigma^2}{np} \left[p\delta_{jl} + 12 \frac{[i - (n+1)/2][k - (n+1)/2]}{(n^2-1)} \right].$$

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

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

a pour variance

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

ce qui confirme la non corrélation entre $\hat{\varepsilon}_{ij}$ et \hat{Y}_{ij} en chaque (i, j) . En fait les vecteurs aléatoires $\hat{\varepsilon}$ et \hat{Y} formés par l'ensemble des np composantes sont non corrélés et vérifient $\text{Var}(\hat{\varepsilon}) = \text{Var}(Y) - \text{Var}(\hat{Y})$ où Y est le vecteur représentant les observations.

Tous ces estimateurs ont une variance proportionnelle à σ^2 . On peut vérifier avec les résultats ci-dessus que $\mathbf{IE}({}^t\hat{\varepsilon}\hat{\varepsilon}) = (np - p - 1)\sigma^2$. On dispose ainsi d'un estimateur sans biais de σ^2 ,

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

qui est non corrélé avec les estimateurs $\hat{\alpha}$, $\hat{\beta}$, \hat{S}_j , et \hat{Y}_{ij} .

2.6.3 Inférence statistique

Tous les résultats précédents restent vrais lorsque l'on suppose simplement que les erreurs sont centrées, non corrélées et de même variance σ^2 . 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 khi-deux à $(np - p - 1)$ degrés de liberté. La loi de Student permet donc de construire des intervalles de confiance pour les paramètres α, β et $S_j, j = 1, \dots, p$, ainsi que pour la prévision de $Y_{n+1,j}$. Dans ce dernier cas, $\hat{Y}_{n+1,j}$ représente l'estimation de la moyenne et il faut tenir compte de la variance de l'erreur pour construire l'intervalle. C'est ainsi que, pour un intervalle de confiance au niveau $(1 - \alpha)$, on obtient :

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

où $P\{|\mathcal{S}_{np-p-1}^t| > t_{np-p-1;\alpha}\} = \alpha$. Il est clair que la bande de confiance ainsi obtenue en faisant varier j ne constitue pas une région de confiance de niveau $(1 - \alpha)$.

Nous pourrions construire selon le même principe une bande de confiance pour la série elle-même afin d'apprécier la validité du modèle. L'observation de Y_{ij} sera dans la bande si et seulement si :

$$\left| \frac{\hat{\varepsilon}_{ij}}{\frac{\hat{\sigma}}{\sqrt{np}} \sqrt{\left[p(n+1) + 12 \frac{[i-(n+1)/2]^2}{(n^2-1)}\right]}} \right| \leq t_{np-p-1;\alpha}.$$

Cet intervalle n'est pas correct car Y_{ij} est utilisé dans sa construction. Une autre approche consiste à utiliser directement les résidus en négligeant le fait que $\hat{\varepsilon}_{ij}$ et $\hat{\sigma}^2$ ne sont pas indépendants :

$$\left| \frac{\hat{\varepsilon}_{ij}}{\frac{\hat{\sigma}}{\sqrt{np}} \sqrt{\left[p(n-1) - 12 \frac{[i-(n+1)/2]^2}{(n^2-1)}\right]}} \right| \leq t_{np-p-1;\alpha}.$$

Les résidus ainsi normalisés sont appelés *résidus studentisés* ou *résidus standardisés* et seront notés $\hat{\varepsilon}_{ij}^S$. En toute rigueur il faut estimer l'erreur ainsi que l'ensemble des paramètres en supprimant la variable Y_{ij} . Le nouveau résidu standardisé ainsi obtenu est appelé résidu par *validation croisée* et sera noté $\hat{\varepsilon}_{ij}^V$. On montre qu'il s'exprime simplement en fonction du précédent (*cf.* Antoniadis et al., p. 33),

$$\hat{\varepsilon}_{ij}^V = \hat{\varepsilon}_{ij}^S \sqrt{\frac{np - p - 2}{np - p - 1 - \hat{\varepsilon}_{ij}^{S^2}}},$$

et qu'il suit la loi de Student à $np - p - 2$ degrés de liberté.

Le premier test portant sur Y_{ij} est moins sévère que celui utilisant $\hat{\varepsilon}_{ij}^S$. Par contre, il n'y a pas d'ordre systématique entre ceux utilisant $\hat{\varepsilon}_{ij}^S$ ou $\hat{\varepsilon}_{ij}^V$, bien qu'il y ait une relation monotone entre ces deux variables, car le seuil $t_{np-p-2;\alpha}$ est plus faible pour $\hat{\varepsilon}_{ij}^V$ que $t_{np-p-1;\alpha}$ pour $\hat{\varepsilon}_{ij}^S$. Bien souvent la série est suffisamment longue pour que tous ces tests soient équivalents et reviennent à considérer que la variable $\hat{\varepsilon}_{ij}/\hat{\sigma}$ est $\mathcal{N}(0; 1)$. C'est le cas dans l'exemple du chiffre d'affaires de la presse parisienne. Avec $\alpha = 5\%$ on obtient $t_\alpha = 2,01$ pour les trois premiers tests et $t_\alpha = 1,96$ pour l'approximation normale. Les différents résidus sont donnés dans le Tableau 2.5 et la représentation de $\hat{\varepsilon}_{ij}^V$ fait l'objet de la Figure 2.15.

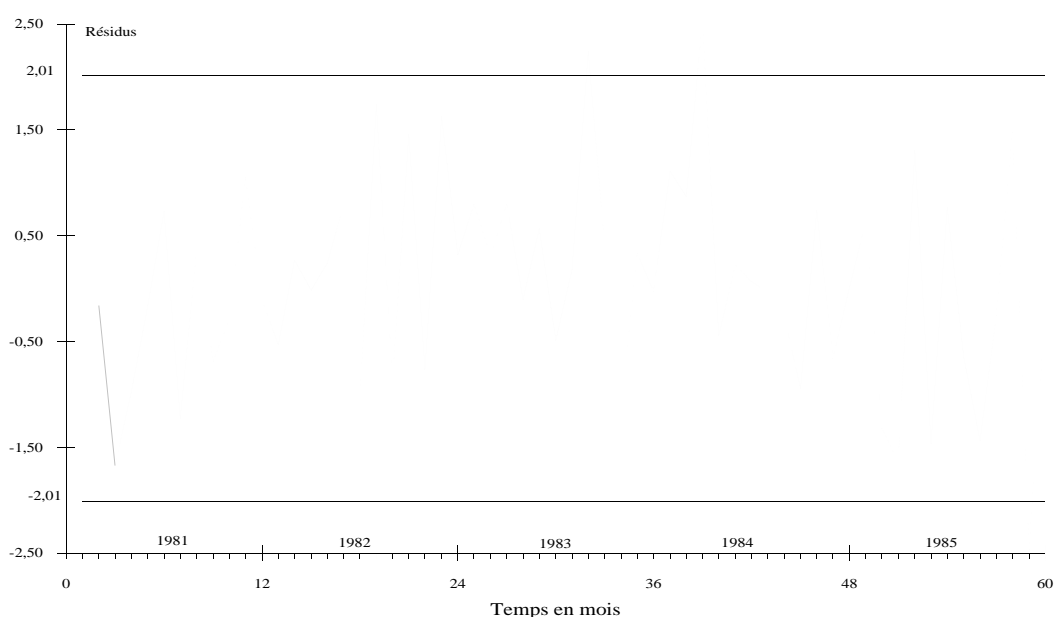


Figure 2.15: Chiffre d'affaires mensuel de la presse parisienne dans une petite ville de province : représentation des résidus obtenus par validation croisée

Dans le cadre du modèle de Buys-Ballot, il est naturel de vouloir apprécier l'importance de l'effet saisonnier. La Figure 2.16 donne la représentation graphique de cet effet pour notre illustration.

Pour comparer entre eux les coefficients saisonniers, il faut prendre en

année	mois	janv	fév	mars	avr	mai	juin	juil	août	sept	oct	nov	déc
	Résidus												
1981	Prévus	-1,61	-0,13	-1,29	-0,77	-0,13	0,58	-0,96	0,31	-0,54	-0,23	0,83	-0,09
	Studentisés	-2,04	-0,16	-1,64	-0,98	-0,16	0,74	-1,22	0,39	-0,69	-0,29	1,05	-0,11
	Validés	-2,11	-0,16	-1,67	-0,98	-0,16	0,73	-1,23	0,39	-0,69	-0,29	1,05	-0,11
	Normalisés	-1,78	-0,14	-1,44	-0,86	-0,14	0,64	-1,07	0,34	-0,61	-0,26	0,92	-0,10
1982	Prévus	-0,43	0,22	-0,01	0,20	0,64	-0,85	1,38	-0,71	1,17	-0,62	1,29	0,26
	Studentisés	-0,53	0,27	-0,02	0,24	0,79	-1,05	1,70	-0,87	1,44	-0,77	1,60	0,32
	Validés	-0,53	0,27	-0,01	0,24	0,79	-1,06	1,74	-0,87	1,46	-0,77	1,62	0,32
	Normalisés	-0,48	0,24	-0,01	0,22	0,70	-0,94	1,51	-0,78	1,28	-0,68	1,42	0,29
1983	Prévus	0,65	0,25	0,65	-0,08	0,46	-0,40	0,15	1,75	0,36	-0,91	0,27	-0,02
	Studentisés	0,80	0,31	0,80	-0,10	0,57	-0,49	0,18	2,15	0,44	-1,11	0,34	-0,03
	Validés	0,80	0,31	0,80	-0,10	0,56	-0,49	0,18	2,24	0,44	-1,11	0,33	-0,03
	Normalisés	0,72	0,28	0,72	-0,09	0,51	-0,44	0,16	1,92	0,39	-0,99	0,30	-0,02
1984	Prévus	0,90	0,71	1,95	-0,37	0,18	0,05	-0,03	-0,22	-0,77	0,60	-0,53	0,01
	Studentisés	1,11	0,87	2,41	-0,45	0,22	0,07	-0,04	-0,27	-0,95	0,74	-0,66	0,02
	Validés	1,11	0,87	2,54	-0,45	0,22	0,07	-0,04	-0,27	-0,95	0,74	-0,66	0,01
	Normalisés	0,98	0,78	2,14	-0,40	0,20	0,06	-0,03	-0,24	-0,84	0,66	-0,59	0,01
1985	Prévus	0,50	-1,04	-1,27	1,02	-1,14	0,61	-0,52	-1,12	-0,20	1,15	-1,85	-0,16
	Studentisés	0,64	-1,32	-1,61	1,30	-1,45	0,77	-0,66	-1,42	-0,26	1,46	-2,35	-0,21
	Validés	0,63	-1,33	-1,63	1,31	-1,47	0,77	-0,65	-1,44	-0,26	1,47	-2,47	-0,20
	Normalisés	0,56	-1,15	-1,41	1,14	-1,27	0,67	-0,57	-1,24	-0,23	1,28	-2,05	-0,18

Tableau 2.5: Chiffre d'affaires mensuel de la presse parisienne dans une petite ville de province : les différentes formes de résidus

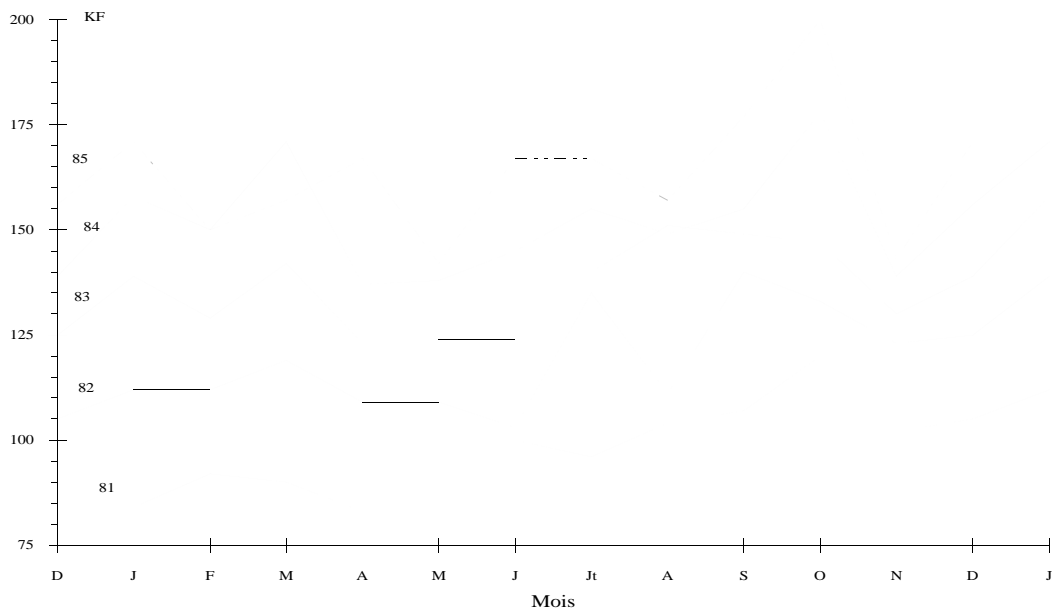


Figure 2.16: Mouvement saisonnier du chiffre d'affaires mensuel de la presse parisienne dans une petite ville de province

compte leurs variances. On définit ainsi les *coefficients standardisés* :

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

Le Tableau 2.6 donne ces résultats dans le cas de notre illustration.

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

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

Tableau 2.6: Coefficients saisonniers du chiffre d'affaires mensuel de la presse parisienne dans une petite ville de province

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

Seuil du test de Fisher-Snedecor de niveau 5% à 11 et 47 d.l. : 1,98

Tableau 2.7: Test de l'effet saisonnier du chiffre d'affaires mensuel de la presse parisienne dans une petite ville de province

Sous l'hypothèse $S_j = 0$, la variable suit la loi de Student à $np - p - 1$ degrés de liberté. Cela permet de considérer leurs valeurs sur une échelle de mesure standard. Rappelons que le niveau de signification du test de $S_j = 0$ contre $S_j \neq 0$ ne peut être contrôlé rigoureusement que si un tel test est envisagé pour un seul mois donné avant de consulter les observations. Notons que ce test de Student équivaut à comparer par un test de Fisher-Snedecor le modèle global au sous-modèle défini par l'hypothèse $S_j = 0$. Par contre il est possible de tester "systématiquement" la présence ou non de l'effet saisonnier de façon globale. Dans ce cas le sous-modèle est limité à la tendance linéaire. Celle-ci est estimée par la droite des moindres carrés ajustée sur l'ensemble des observations mensuelles :

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

où

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

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

L'estimateur sans biais de σ^2 , sous cette hypothèse, est

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

où

$$\text{var}(Y_{ij}) = \frac{1}{np} \sum_{t=1}^{np} [Y_t - \bar{y}_{..}]^2.$$

Sous l'hypothèse nulle, la variable

$$\frac{(np-2)\hat{\sigma}_0^2 - (np-p-1)\hat{\sigma}^2}{(p-1)\hat{\sigma}^2}$$

suit la loi de Fisher-Snedecor à $(p-1, np-p-1)$ degrés de liberté. Les résultats de ce test pour notre illustration sont reportés dans le Tableau 2.7.

2.6.4 Modèle linéaire général

Le modèle de Buys-Ballot est un cas particulier du modèle suivant :

$$Y_t = f(t) + S(t) + \varepsilon_t = \sum_{j=0}^q \alpha_j \phi_j(t) + \sum_{j=1}^p \beta_j \psi_j(t) + \varepsilon_t, \quad t = 1, \dots, T,$$

dans lequel $\phi_j(t)$ et $\psi_j(t)$ sont des fonctions connues du temps, utilisées pour modéliser respectivement la tendance $f(t)$ et la composante saisonnière $S(t)$ à l'aide de paramètres naturels α_j et β_j , l'erreur ε_t étant un bruit blanc gaussien. En effet la tendance linéaire est obtenue avec $\phi_0(t) = 1$ et $\phi_1(t) = t$. Pour la partie périodique de période p , nous introduisons les fonctions

$$\psi_j(t) = \delta_j(t) = 1 \text{ si } t = j \text{ mod}(p), 0 \text{ sinon, } \quad t = 1, \dots, T, \quad j = 1, \dots, p.$$

Les paramètres naturels sont $\alpha_0 = \beta$, $\alpha_1 = \alpha$ et $\beta_j = S_j$, $j = 1, \dots, p$. Cependant pour que le modèle soit identifiable, nous avons imposé la contrainte $S_1 + \dots + S_p = 0$. En général les fonctions $\phi_j(t)$, $j = 0, \dots, q$ constituent $q+1$ vecteurs linéairement indépendants de \mathbb{R}^T et il en est de même pour les p fonctions ψ_j , $j = 1, \dots, p$. Par contre l'intersection des sous-espaces de dimensions $q+1$ et p ainsi définis est rarement réduite au vecteur nul. En d'autres termes, les deux systèmes de fonctions engendrent un sous-espace de \mathbb{R}^T de dimension $r \leq p+q+1$. Il sera donc nécessaire d'introduire $p+q+1-r$

contraintes sur les paramètres naturels. Cela conduit à restructurer le modèle sous la forme,

$$Y_t = f(t) + S(t) + \varepsilon_t = \sum_{j=1}^r \theta_j x_j(t) + \varepsilon_t, \quad t = 1, \dots, T,$$

où les fonctions $x_j(t)$, $j = 1, \dots, r$, combinaisons linéaires des fonctions $\phi_j(t)$ et $\psi_j(t)$ initiales, définissent r vecteurs linéairement indépendants de \mathbb{R}^T .

Par exemple une généralisation immédiate du modèle de Buys-Ballot consiste à poser :

$$f(t) = \sum_{k=0}^q \alpha_k t^k, \quad S(t) = \sum_{j=1}^p \left[\sum_{k=0}^s \beta_{jk} t^k \right] \delta_j(t) = \sum_{j=1}^p S_j(t) \delta_j(t), \quad t = 1, \dots, T.$$

Dans ce cas la tendance est un polynôme de degré q et les “coefficients saisonniers” évoluent au cours du temps comme un polynôme de degré s . La relation

$$\sum_{j=1}^p \delta_j(t) = 1, \quad t = 1, \dots, T,$$

conduit aux contraintes

$$\sum_{j=1}^k \beta_{jk} = 0, \quad k = 0, \dots, \min(s, q),$$

qui seront suffisantes si T est assez grand. Notons qu’il n’est plus nécessaire de supposer que T soit un multiple de la période.

De façon générale le modèle initial, en fonction des paramètres naturels, et sa restructuration se décrivent sous forme vectorielle par :

$$Y = \tilde{X}\tilde{\theta} + \varepsilon = X\theta + \varepsilon, \quad \tilde{X} = XG, \quad \tilde{\theta} = G^{-1}\theta, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I_T).$$

Dans cette écriture $Y = {}^t[Y_1, \dots, Y_T]$ est le vecteur des observations, $\varepsilon = {}^t[\varepsilon_1, \dots, \varepsilon_T]$ est celui des erreurs, I_T étant la matrice identité d’ordre T , et $\tilde{\theta} = {}^t[\alpha_0, \dots, \alpha_q, \beta_1, \dots, \beta_p]$ est le vecteur des paramètres naturels. La matrice \tilde{X} , de dimension $T \times (p + q + 1)$, représente le système des fonctions initiales $\phi_j(t)$ et $\psi_j(t)$ et X , de dimensions $T \times r$, celui des nouvelles fonctions $x_j(t)$. La matrice G , de dimension $r \times (p + q + 1)$, est alors parfaitement définie. Le choix de l’inverse généralisé G^{-} traduit les contraintes sur $\tilde{\theta}$ assurant son identification à partir du paramètre $\theta = {}^t[\theta_1, \dots, \theta_r]$ du modèle. Le choix des fonctions $x_j(t)$ peut être guidé par l’interprétation du paramètre

θ qui en résulte. Par exemple, on pourra faire en sorte que l'absence d'une composante (saisonnière ou tendance) se traduise par la nullité des dernières composantes de θ .

Illustration

Les résultats concernant l'ajustement d'un modèle avec une tendance quadratique (noté modèle quadratique) sur le chiffre d'affaires de la presse parisienne sont indiqués dans les Tableaux 2.8 et 2.9 ainsi que sur la Figure 2.17. La représentation graphique des résidus normalisés par l'approximation gaussienne est donnée dans la Figure 2.18 et les tests concernant la présence de l'effet saisonnier ou la nécessité d'introduire le terme quadratique dans la tendance sont regroupés dans le Tableau 2.10.

année	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
1981	1	84	92	90	83	85	100	96	104	107	120	102	105
	ajustée	93	87	97	85	82	91	102	98	110	121	93	106
1982	2	112	112	119	109	109	103	135	111	140	133	123	125
	ajustée	116	110	120	108	104	113	124	120	131	142	114	126
1983	3	139	129	142	123	124	124	140	151	149	147	130	139
	ajustée	136	130	139	127	123	131	142	138	149	159	131	143
1984	4	158	150	171	137	138	145	155	149	155	178	139	156
	ajustée	153	146	155	143	139	146	157	152	163	173	145	156
1985	5	171	150	157	167	142	167	167	157	177	200	143	171
	ajustée	166	159	168	155	150	158	168	164	174	184	155	166
	coef. sais.	7	-1	7	-7	-12	-6	4	-2	8	17	-13	-2
1986	prévues	176	169	177	164	159	166	176	171	181	191	161	172

Tendance : "pente" = 2,13 ; "ordonnée à l'origine" = 84 ; terme quadratique = -0,0121

Tableau 2.8: Chiffre d'affaires mensuel de la presse parisienne dans une petite ville de province, modèle quadratique (unité : 1KF)

mois	janv	fév	mars	avr	mai	juin	juil	août	sept	oct	nov	déc
Coef. réels	6,7	-1,0	6,7	-6,7	-12,4	-5,6	3,8	-1,7	8,1	16,8	-12,7	-2,2
Coef. studentisés	2,0	-0,3	2,0	-2,0	-3,6	-1,6	1,1	-0,5	2,4	4,9	-3,7	-0,6

Seuil du test de Student de niveau 5% à 46 d.l. : 2,01

Tableau 2.9: Coefficients saisonniers du chiffre d'affaires mensuel de la presse parisienne dans une petite ville de province, modèle quadratique

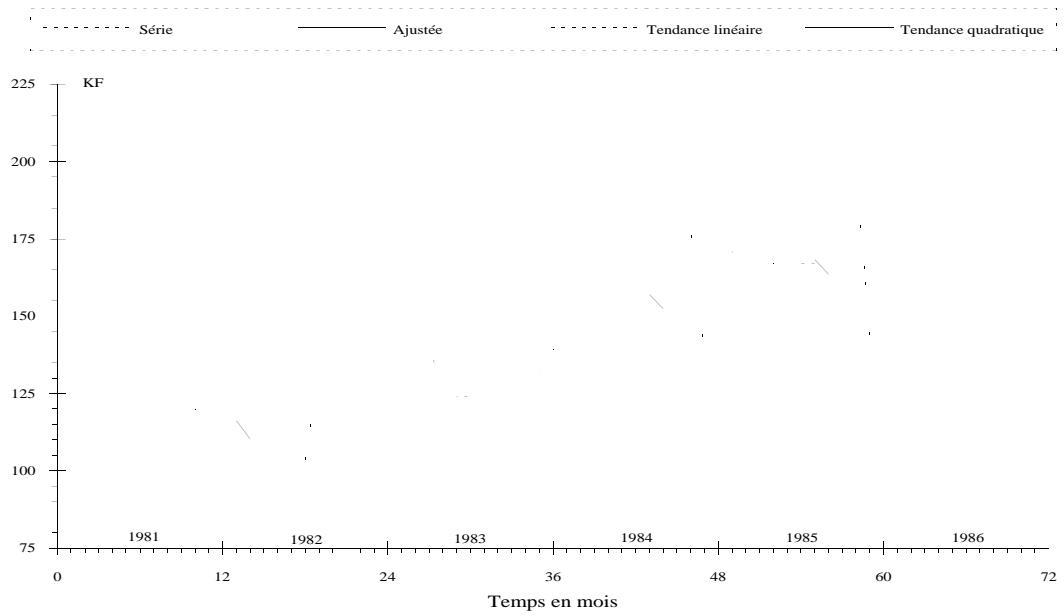


Figure 2.17: Chiffre d'affaires de la presse parisienne dans une petite ville de province, modèle quadratique

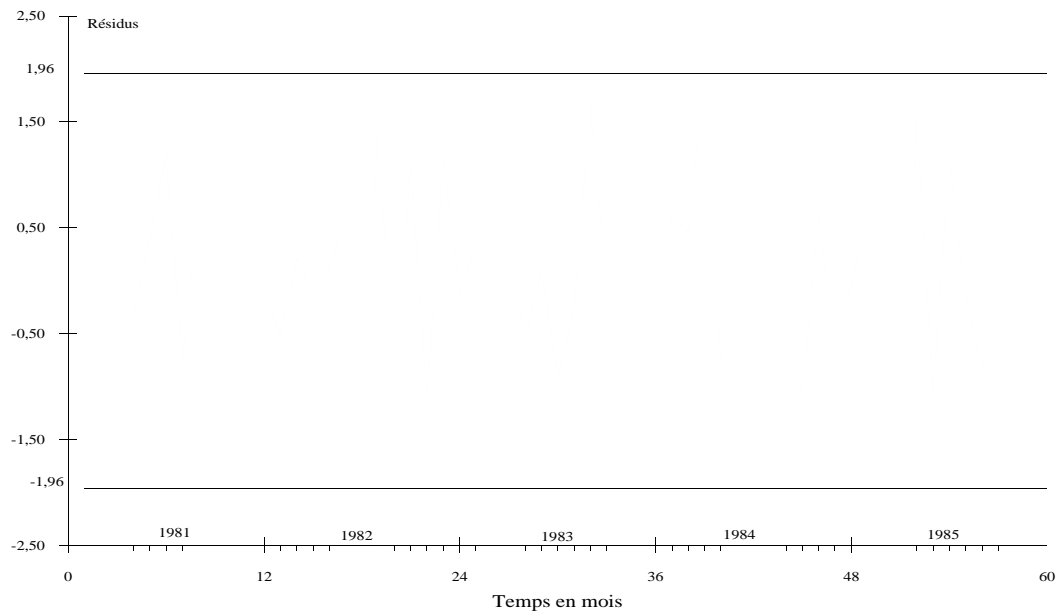


Figure 2.18: Résidus normalisés du chiffre d'affaires de la presse parisienne dans une petite ville de province, modèle quadratique

<i>Paramètres du modèle</i>	$\hat{\theta}_0$	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\sigma}$	$\hat{\sigma}^2$
Avec effet saisonnier	83,9	2,131	-0,0121	7,9	62,65
Sans effet saisonnier	84,0	2,120	-0,0120	11,2	125,35

Seuil du test de Fisher-Snedecor de niveau 5% à 11 et 46 d.l.	1,97
Statistique du test observée pour la présence du mouvement saisonnier	6,19
Seuil du test de Fisher-Snedecor de niveau 5% à 1 et 46 d.l.	4,06
Statistique du test observée pour la présence du terme quadratique	10,14

Tableau 2.10: Tests de l'effet saisonnier et de la présence du terme quadratique pour le chiffre d'affaires mensuel de la presse parisienne dans une petite ville de province, modèle quadratique

Liste des tableaux

1.1	Vitesse coronarienne y_i et poids x_i de 18 patients	11
1.2	Valeurs ajustées et résidus pour la vitesse coronarienne	17
1.3	Chiffre d'affaires et marge des entreprises	18
1.4	Résultats sur les entreprises	21
1.5	Exemple de non-sens	21
1.6	Table de contingence	24
1.7	Données et quelques résultats pour la décroissance D_A	37
1.8	Répartition des conscrits selon la taille et le poids	38
1.9	Caractéristiques conditionnelles du poids en fonction de la taille	38
1.10	Résultats concernant les conscrits	38
1.11	Répartition des ménages selon le nombre de personnes et de pièces habitables	39
1.12	Caractéristiques du nombre de pièces conditionnelles au nom- bre de personnes	39
1.13	Caractéristiques du nombre de personnes conditionnelles au nombre de pièces	40
1.14	Résultats concernant les ménages	40
2.1	Maintenance d'un réseau de distributeurs de boissons	55
2.2	Analyse de la variance du modèle de régression	65
2.3	Parabole bruitée	74
2.4	Chiffre d'affaires mensuel de la presse parisienne dans une pe- tite ville de province (unité : 1KF)	83
2.5	Chiffre d'affaires mensuel de la presse parisienne dans une pe- tite ville de province : les différentes formes de résidus	89
2.6	Coefficients saisonniers du chiffre d'affaires mensuel de la presse parisienne dans une petite ville de province	90
2.7	Test de l'effet saisonnier du chiffre d'affaires mensuel de la presse parisienne dans une petite ville de province	90
2.8	Chiffre d'affaires mensuel de la presse parisienne dans une pe- tite ville de province, modèle quadratique (unité : 1KF)	93

2.9	Coefficients saisonniers du chiffre d'affaires mensuel de la presse parisienne dans une petite ville de province, modèle quadratique	93
2.10	Tests de l'effet saisonnier et de la présence du terme quadratique pour le chiffre d'affaires mensuel de la presse parisienne dans une petite ville de province, modèle quadratique	95

Liste des Figures

1	Aspect géométrique du modèle linéaire	3
1.1	Régression du poids par rapport à la taille des étudiants de la promotion 97 d'IUP 3	4
1.2	Droites associées à un ensemble de points selon différents critères	10
1.3	Vitesse coronarienne en fonction du poids chez 18 patients . .	12
1.4	Régression linéaire simple dans l'espace des observations . . .	14
1.5	Exemples de données structurées	16
1.6	Résidus en fonction des valeurs ajustées pour la vitesse coronarienne	17
1.7	Marge en fonction du chiffre d'affaires de l'ensemble des entreprises, unité = 1MF	19
1.8	Zoom sur la marge en fonction du chiffre d'affaires de l'ensemble des entreprises, unité = 1MF	19
1.9	Marge en fonction du chiffre d'affaires des petites entreprises, unité = 1MF	20
1.10	Marge en fonction du chiffre d'affaires des moyennes ou grandes entreprises, unité = 1MF	20
1.11	Déficiences mentales en fonction des licences radio	22
1.12	Déficiences mentales en fonction du prénom du président	22
1.13	Courbe de régression	25
1.14	Courbe des moindres carrés	26
1.15	Décroissance D_A en fonction du temps	37
1.16	Courbe et droite de régression du poids en fonction de la taille d'un ensemble de conscrits	38
1.17	Courbes et droites de régression pour la répartition des ménages selon le nombre de personnes et le nombre de pièces habitables	40
2.1	Espace des variables du modèle linéaire	54
2.2	Espace des observations du modèle linéaire	54
2.3	Aspect géométrique du test de Fisher	64

2.4	Chronique des résidus par validation croisée pour le réseau de distributeurs	68
2.5	Résidus standardisés en fonction des valeurs ajustées pour le réseau de distributeurs de boissons	69
2.6	Temps en fonction du nombre de caisses pour le réseau de distributeurs de boissons	71
2.7	Temps en fonction de la distance pour le réseau de distributeurs de boissons	71
2.8	Nombre de caisses en fonction de la distance pour le réseau de distributeurs de boissons	72
2.9	Régression simple et ajustement d'une parabole sur une parabole bruitée	75
2.10	Régression simple sur le carré de x pour une parabole bruitée	75
2.11	Résidus par validation croisée pour une parabole bruitée	76
2.12	Résidus standardisés en fonction des valeurs ajustées pour une parabole bruitée	76
2.13	Régression de x en fonction de son carré x^2	77
2.14	Chiffre d'affaires mensuel de la presse parisienne dans une petite ville de province	83
2.15	Chiffre d'affaires mensuel de la presse parisienne dans une petite ville de province : représentation des résidus obtenus par validation croisée	88
2.16	Mouvement saisonnier du chiffre d'affaires mensuel de la presse parisienne dans une petite ville de province	89
2.17	Chiffre d'affaires de la presse parisienne dans une petite ville de province, modèle quadratique	94
2.18	Résidus normalisés du chiffre d'affaires de la presse parisienne dans une petite ville de province, modèle quadratique	94

Bibliographie

- [ABC92] A. ANTONIADIS, J. BERRUYER, and R. CARMONA. *Régression non linéaire et applications*. Economica, Paris, 1992.
- [BAI89a] G. BAILLARGEON. *Méthodes statistiques de l'ingénieur*, volume 1. SMG, 3^e édition, 1989.
- [BAI89b] G. BAILLARGEON. *Probabilités, statistique et techniques de régression*. SMG, 1989.
- [BAR71] J.R. BARRA. *Notions fondamentales de statistique mathématique*. Dunod, Paris, 1971.
- [BB69] J.R. BARRA and A. BAILLE. *Problèmes de statistique mathématique*. Dunod, Paris, 1969.
- [BJR94] G.E.P. BOX, G.M. JENKINS, and G.C. REINSEL. *Time Series Analysis, Forecasting and Control*. Prentice Hall, Englewood Cliffs, New Jersey, third edition, 1994.
- [BL87] D. BOSQ and J.-P. LECOUTRE. *Théorie de l'estimation fonctionnelle*. Economica, Paris, 1987.
- [CAL65] G. CALOT. *Cours de statistique descriptive*. Dunod, Paris, 3^e édition, 1965. Collection Statistique et Programmes Économiques.
- [DAG98a] P. DAGNELIE. *Statistique théorique et appliquée. Tome 1 : Statistique descriptive et bases de l'inférence statistique*. De Boeck Université, 1998.
- [DAG98b] P. DAGNELIE. *Statistique théorique et appliquée. Tome 2 : Inférence statistique à une et deux dimensions*. De Boeck Université, 1998.

- [DFT89] J.-J. DROESBECKE, B. FICHET, and P. TASSI. *Séries chronologiques : théorie et pratique des modèles ARIMA*. Economica, Paris, 1989.
- [DFT94] J.-J. DROESBECKE, B. FICHET, and P. TASSI. *Modélisation ARCH. Théorie statistique et applications dans le domaine de la finance*. Editions de l'Université de Bruxelles, Bruxelles et Editions Ellipses, Paris, 1994.
- [FER67] T. FERGUSON. *Mathematical statistic*. Academic Press, 1967.
- [FF67] C. FOURGEAUX and A. FUCHS. *Statistique*. Dunod, 1967.
- [GM89] C. GOURIEROUX and A. MONFORT. *Statistique et modèles économétriques*, volume 1 & 2. Economica, Paris, 1989.
- [GM90] C. GOURIEROUX and A. MONFORT. *Séries temporelles et modèles dynamiques*. Economica, Paris, 1990.
- [GOU92] C. GOURIEROUX. *Modèles ARCH et applications financières*. Economica, Paris, 1992.
- [JAF96] P. JAFFARD. *Initiation aux méthodes de la statistique et du calcul des probabilités*. Masson, 3^e édition, 1996.
- [KS71] M.G. KENDALL and A. STUART. *The advanced theory of statistics*. Charles Griffin, London, 3^e édition, 1971. 3 volumes.
- [LEH59] E.L. LEHMANN. *Testing statistical hypothesis*. John Wiley, 1959.
- [LMF79] L. LEBART, A. MORINEAU, and J.-P. FÉNELON. *Traitement des données statistiques, méthodes et programmes*. Dunod, 1979.
- [MCN89] P. Mc CULLAGH and J.A. NEDLER. *Generalized linear models*. Chapman and Hall, 2nd édition, 1989.
- [MEL90] G. MELARD. *Méthodes de prévision à court terme*. Editions de l'Université de Bruxelles, Bruxelles et Editions Ellipses, Paris, 1990.
- [MON82] A. MONFORT. *Cours de statistique mathématique*. Economica, Paris, 1982.
- [MP82] D.C. MONTGOMERY and E.A. PECK. *Introduction to linear regression analysis*. John Wiley, 1982.

- [NWK90] J. NETER, W. WASSERMAN, and M.H. KUTNER. *Applied linear statistical models. Regression, analysis of variance, and experimental designs*. Irwin, 3rd edition, 1990.
- [RAO73] C.R. RAO. *Linear Statistical Inference and its Applications*. Wiley, New-York, 1973.
- [SAP90] G. SAPORTA. *Probabilités, Analyse des données et Statistique*. Technip, 1990.
- [SCH59] H. SCHEFFE. *The analysis of variance*. John Wiley, 1959.
- [SR93] R.R. SOKAL and F.J. ROHLF. *Biometry: the principles and practice of statistics in biological research*. W.H. Freeman and company, 3rd edition, 1993.
- [TAS89] P. TASSI. *Méthodes statistiques*. Economica, Paris, 2^e edition, 1989.
- [VOL85] M. VOLLE. *Analyse des données*. Economica, Paris, 3rd edition, 1985.
- [WIL62] S. WILKS. *Mathematical statistics*. John Wiley, 1962.

Index

- équations normales, 58
- absence de corrélation, 28
- absence réciproque de corrélation, 28
- analyse de la covariance, 2, 3
- analyse de la déviance, 3
- analyse de la variance, 2, 3, 65
- analyse en composantes principales, 7
- borne de Cramer-Rao, 44, 62
- Buyss-Ballot, 80
- coefficient de corrélation linéaire, 5, 34
- coefficient de corrélation linéaire empirique, 15
- coefficient de détermination, 15, 58
- coefficient de détermination corrigé, 59
- coefficients standardisés, 90
- composante fondamentale, 80
- composante saisonnière, 80
- corrélation partielle, 70
- corrélée, 28
- courbe de régression, 25
- courbes de régression, 5
- distance de Cook, 73
- droite de régression, 5, 14
- droite de régression orthogonale, 7
- droite des moindres carrés, 5
- droite des moindres distances, 7
- droite des moindres rectangles, 8
- décomposition de Cholesky, 79
- efficace, 44
- ellipsoïde de confiance, 45, 63
- erreurs, 13
- espace des moyennes, 13, 53
- espace des observations, 13, 53
- espace des variables, 13, 53
- estimateur de Gauss-Markov, 42, 60
- estimateur de maximum de vraisemblance, 43
- estimateur des moindres carrés, 57
- estimateur efficace, 62
- estimateur sans biais, 41
- estimateur studentisé, 46
- estimateurs des moindres carrés, 14
- Fisher-Snedecor, 45, 64
- fonction de régression, 11
- indépendance, 27
- influyente, 73
- intervalle de prévision, 48, 66
- intervalles de confiance, 44, 63
- liaison fonctionnelle, 28
- liaison fonctionnelle réciproque, 28
- logit, 2
- loi de Student, 44
- matrice chapeau, 73
- matrice d'information de Fisher, 43, 62
- matrice de prédiction, 60

- matrice des régresseurs, 52
- matrice “chapeau”, 60
- maximum de vraisemblance, 62
- modèle linéaire, 2, 53
- modèle linéaire généralisé, 3
- moindres carrés généralisés, 57
- méthode des moindres carrés, 13

- non corrélée, 27
- non linéairement corrélées, 15

- orthogonalisation de Gram-Schmidt, 78

- paramètres, 12
- plan d’expérience, 2, 12
- plan d’expérience, 52
- probit, 2
- prévision, 48, 66, 84, 86
- prévision inverse, 49

- rapport de corrélation, 29
- rappports de corrélation, 5
- régresseur ajusté, 70
- régression linéaire, 2, 53
- régression linéaire multiple, 3
- régression linéaire simple, 3, 4
- régression non linéaire, 5
- régression poissonnienne, 3
- régression polynomiale, 73
- résidu, 86
- résidu par validation croisée, 48
- résidus, 16, 47, 67
- résidus par validation croisée, 67
- résidus standardisés, 47, 67, 87
- résidus studentisés, 47, 67, 87

- série ajustée, 84, 86

- Table de Buys-Ballot, 84
- table de contingence, 23
- tendance, 80
- test de Fisher, 46, 63, 66

- tests de Student, 46
- tests du chi-deux, 46

- valeurs ajustées, 16
- validation croisée, 87
- var inter, 25
- var intra, 25
- variable de liaison, 25
- variable de réponse ajustée, 69
- variable explicative, 11
- variable expliquée, 11
- vecteur gaussien, 33

Table des matières

INTRODUCTION	2
1 RÉGRESSION LINÉAIRE SIMPLE	4
1.1 DROITES ASSOCIÉES À UN ENSEMBLE DE POINTS DU PLAN	5
1.1.1 Droites des moindres carrés	5
1.1.2 Droite des moindres distances	7
1.1.3 Droite des moindres rectangles	8
1.1.4 Optimisation numérique	9
1.1.5 Exemple	9
1.2 DROITE DE RÉGRESSION LINÉAIRE	11
1.2.1 Illustration	11
1.2.2 Les hypothèses du modèle	12
1.2.3 Estimateur des moindres carrés	13
1.2.4 Coefficient de corrélation linéaire empirique	15
1.2.5 Analyse descriptive des résidus	16
1.2.6 Exemples	17
1.3 COURBES DE RÉGRESSION	23
1.3.1 Courbe des moindres carrés	23
1.3.2 Rapports de corrélation	28
1.3.3 Droites des moindres carrés pondérés	30
1.3.4 Vecteurs aléatoires gaussiens	32
1.3.5 Transformations sur les variables	35
1.3.6 Exemples	36
1.4 PROPRIÉTÉS DES ESTIMATEURS	39
1.4.1 Biais et variances	41
1.4.2 Propriété de Gauss-Markov	42
1.4.3 Lois de probabilités	42
1.4.4 Maximum de vraisemblance	43
1.5 INTERVALLES DE CONFIANCE ET TESTS	44

1.5.1	Intervalles de confiance	44
1.5.2	Tests	46
1.5.3	Étude des résidus	47
1.5.4	Prévision	48
1.5.5	Prévision inverse	49
2	RÉGRESSION LINÉAIRE MULTIPLE	51
2.1	LES HYPOTHÈSES DU MODÈLE	52
2.1.1	Modèle standard	52
2.1.2	Illustration	55
2.1.3	Modèle général	56
2.2	MÉTHODE DES MOINDRES CARRÉS	57
2.2.1	Estimateur des moindres carrés	57
2.2.2	Coefficient de détermination	58
2.3	PROPRIÉTÉS DES ESTIMATEURS	59
2.3.1	Biais et variance	59
2.3.2	Propriété de Gauss-Markov	60
2.3.3	Lois de probabilités	61
2.3.4	Maximum de vraisemblance	61
2.4	INTERVALLES DE CONFIANCE ET TESTS	62
2.4.1	Intervalles de confiance	63
2.4.2	Tests	63
2.4.3	Prévision	66
2.4.4	Étude des résidus	67
2.4.5	Étude des coefficients de régression	69
2.4.6	Étude du plan d'expérience	72
2.5	RÉGRESSION POLYNOMIALE	73
2.5.1	Illustration	73
2.5.2	Généralités	77
2.5.3	Régression orthogonale	78
2.6	APPLICATION AUX CHRONIQUES	80
2.6.1	Modèle de Buys-Ballot	80
2.6.2	Moyenne et variance des estimateurs	84
2.6.3	Inférence statistique	87
2.6.4	Modèle linéaire général	91
	LISTE DES TABLEAUX	93
	LISTE DES FIGURES	97

<i>TABLE DES MATIÈRES</i>	107
BIBLIOGRAPHIE	99
INDEX	102
TABLE DES MATIÈRES	105