Régression

Théorie et applications

Springer
Paris
Berlin
Heidelberg
New York
Hong Kong
Londres
Milan
Tokyo

Pierre-André Cornillon
Éric Matzner-Løber

Régression

Théorie et applications

Springer

Pierre-André Cornillon
Laboratoire de Statistique - UFR de Sciences sociales
Université Rennes 2
35043 Rennes Cedex

Éric Matzner-Løber
Laboratoire de Statistique - UFR de Sciences sociales
Université Rennes 2
35043 Rennes Cedex

ISBN-10 : 2-287-39692-6 Springer Paris Berlin Heidelberg New York

ISBN-13 : 978-2-287-39692-2 Springer Paris Berlin Heidelberg New York

SPIN : 11855965

Maquette de couverture : Jean-François Montmarché
Image de couverture : © Gaëtan de Séguin des Hons – « Il prend sa place » (détail).

Collection
Statistiques et probabilités appliquées
dirigée par Yadolah Dodge

Professeur Honoraire
Université de Neuchâtel
2002 Neuchâtel - Suisse

Dans la même collection :

— *Statistique. La théorie et ses applications,*
Michel Lejeune, avril 2004
— *Le choix Bayésien. Principes et pratique,*
Christian P. Robert, novembre 2005
— *Maîtriser l'aléatoire. Exercices résolus de probabilités et statistique,*
Eva Cantoni, Philippe Huber et Elvezio Ronchetti, octobre 2006

REMERCIEMENTS

Cet ouvrage, s'appuyant sur des exemples, n'existerait pas sans ceux-ci. A l'heure actuelle, s'il est très facile de traiter des données, il est beaucoup plus difficile de les proposer comme exemple pour une diffusion. Les données sont devenues confidentielles et les variables mesurées, jusqu'à leur intitulé même, représentent une avancée stratégique vis-à-vis des concurrents. Il est ainsi presque impensable de traiter des données issues du monde industriel ou du marketing, bien que les exemples y soient nombreux. Cependant, trois organismes, *via* leur directeur, ont pris l'initiative de permettre la diffusion de leurs données. Nous avons donc un très grand plaisir à remercier Magali Coron (Association Air Breizh), Bernard Mallet (CIRAD forêt) et Jean-Noël Marien (UR2PI). Nous souhaitons bien sûr associer tous les membres de l'unité de recherche pour la productivité des plantations industrielles (UR2PI) passés ou présents. Les membres de cet organisme de recherche congolais gèrent de nombreux essais tant génétiques que sylvicoles et nous renvoyons toutes les personnes intéressées auprès de cet organisme ou auprès du CIRAD, département forêt (`http:\\wwww.cirad.fr`), qui est un des membres fondateurs et un participant actif au sein de l'UR2PI.

Par ailleurs, la version actuelle de cet ouvrage résulte de l'action à des degrés divers de nombreuses personnes. Nous souhaitons donc remercier tous nos collègues de l'université Rennes 2, tous les étudiants de la filière MASS de Rennes 2 et ceux de l'ENSAI, qui ont permis l'élaboration de ce livre à partir de notes de cours.

Cependant, le livre ne serait pas ce qu'il est sans la patience et la minutie d'Arnaud Guyader. Entre deux énervements à peine contenus sur la qualité du manuscrit, il a débusqué d'innombrables erreurs tant sur la forme que sur le fond. Nous n'oublions pas les relecteurs exigeants que sont Christophe Abraham et Frank Rimek qui nous ont toujours poussé vers une plus grande clarté théorique ou pratique et dont les remarques ont toujours été pertinentes. Enfin, Nathalie Chèze, Julie Josse et Vincent Lefieux ont permis par leurs conseils avisés d'améliorer le document au moment même où l'on croyait arriver au but. Bien évidemment, après ces relectures successives, nous avons encore modifié quelques phrases et donc sûrement rajouté quelques fautes.

Nos remerciements vont également à Nathalie Huilleret de Springer-Verlag (Paris) et Yadolah Dodge, directeur de la collection, pour la confiance qu'ils nous ont accordée.

AVANT-PROPOS

L'objectif de cet ouvrage est de rendre accessible au plus grand nombre une des méthodes les plus utilisées de la statistique : la régression. Nous souhaitons aborder de manière simultanée les fondements théoriques et les questions inévitables que l'on se pose lorsque l'on modélise des phénomènes réels. En effet, comme pour toute méthode statistique, il est nécessaire de comprendre précisément la méthode et de savoir la mettre en œuvre. Si ces deux objectifs sont atteints, il sera alors aisé de transposer ces acquis à d'autres méthodes, moyennant un investissement modéré, tant théorique que pratique. Les grandes étapes - modélisation, estimation, choix de variables, examen de la validité du modèle choisi - restent les mêmes d'une méthode à l'autre. Cet aspect apparaît nettement dans le dernier chapitre consacré à certaines extensions de la régression linéaire. Ces extensions ont chacune un caractère spécifique, mais les différentes étapes vues en régression se retrouvent dans chaque méthode.

Cet ouvrage s'adresse aux étudiants des filières scientifiques, élèves ingénieurs, chercheurs dans les domaines appliqués (économie, biologie, sciences de la vie...) et plus généralement à tous les chercheurs souhaitant modéliser des relations de causalité. Il utilise aussi les notions d'intervalle de confiance, de test et les lois de probabilités classiques. Pour les lecteurs n'ayant aucune notion de ces concepts, le livre de Lejeune (2004) dans la même collection pourra constituer une aide précieuse pour certains paragraphes. Cet ouvrage nécessite la connaissance des bases du calcul matriciel : définition d'une matrice, somme, produit, inverse, ainsi que valeurs propres et vecteurs propres pour le dernier chapitre. Des résultats classiques sont toutefois rappelés en annexes afin d'éviter de consulter trop souvent d'autres ouvrages.

Cet ouvrage souhaite concilier les fondements théoriques nécessaires à la compréhension et à la pratique de la méthode. Nous avons donc souhaité un livre avec toute la rigueur scientifique possible mais dont le contenu et les idées ne soient pas noyés dans les démonstrations et les lignes de calculs. Pour cela, seules quelques démonstrations, que nous pensons importantes, sont conservées dans le corps du texte. Les autres résultats sont démontrés à titre d'exercice. Des exercices, de difficulté variable, sont proposés en fin de chapitre. La présence de † indique des exercices plus difficiles que la majorité des exercices proposés. Des questions de cours sous la forme de QCM sont aussi proposées afin d'aider aux révisions du chapitre. Les corrections de tous les exercices sont fournies en annexe A. Une partie « notes » présente en fin de chapitre des discussions ou extensions, cette partie pourra être ignorée lors d'une première lecture.

Afin que les connaissances acquises ne restent pas théoriques, nous avons intégré des exemples traités avec le logiciel libre GNU-R (`http://www.r-project.org`). Afin que les lecteurs puissent se familiariser avec le logiciel et retrouver les mêmes résultats que ceux donnés dans le livre, les commandes sont rapportées dans le livre. Nous encourageons donc les lecteurs à utiliser les données (qui se trouvent sur les pages web des auteurs) et les codes afin de s'approprier la théorie mais aussi la pratique.

Au niveau de l'étude des chapitres, le premier de ceux-ci, consacré à la régression simple, est traité afin de présenter de nombreux concepts et idées. Il est donc important de le lire afin de se familiariser avec les problèmes et les solutions envisagés ainsi qu'avec l'utilité des hypothèses de la régression.

Le second chapitre présente l'estimation et la géométrie de la méthode des moindres carrés. Il est donc fondamental.

Le troisième chapitre aborde la partie inférentielle. Il représente la partie la plus technique et la plus calculatoire de cet ouvrage. En première lecture, il pourra apparaître comme fastidieux, mais la lecture et la compréhension de la géométrie des tests entre modèles emboîtés semblent nécessaires. Le calcul des lois pour le praticien peut être omis.

Le quatrième chapitre présente très peu de calculs. Il permet de vérifier que le modèle, et donc les conclusions que l'on peut en tirer, sont justes. Cette partie est donc fondamentale pour le praticien. De plus, les idées sous-jacentes sont utilisées dans de très nombreuses méthodes statistiques. La lecture de ce chapitre est indispensable.

Le cinquième chapitre présente l'introduction de variables explicatives qualitatives dans le modèle de régression, soit en interaction avec une variable quantitative (analyse de la covariance), soit seules (analyse de la variance). La présentation oublie volontairement les formules classiques des estimateurs à base de somme et de moyenne par cellule. Nous nous focalisons sur les problèmes de paramètres et de contraintes, problèmes qui amènent souvent une question naturelle à la vue des listings d'un logiciel : « Tiens, il manque une estimation d'un paramètre ». Nous avons donc souhaité répondre simplement à cette question inhérente à la prise en compte de variables qualitatives.

Le sixième chapitre présente le choix de variables (ou de modèles). Nous présentons le problème *via* l'analyse d'un exemple à 3 variables. A partir des conclusions tirées de cet exemple, nous choisissons un critère de sélection (erreur quadratique moyenne ou EQM) et nous proposons des estimateurs cohérents. Ensuite, nous axons la présentation sur l'utilisation des critères classiques et des algorithmes de choix de modèles présents dans tous les logiciels et nous comparons ces critères. Enfin, nous discutons des problèmes engendrés par cette utilisation classique. Ce chapitre est primordial pour comprendre la sélection de modèle et ses problèmes.

Le septième chapitre propose les premières extensions de la régression. Il s'agit principalement d'une présentation succincte de certaines méthodes utilisées en moindres carrés généralisés. Elle présente aussi une approche de la régression par la méthode des noyaux.

Enfin, le huitième chapitre présente des extensions classiques (ridge, régression sur composantes principales) ou plus actuelles (lasso ou PLS) de la régression. D'un point de vue théorique, elles permettent d'approfondir les problèmes de contraintes sur le vecteur de coefficients. Chaque méthode est présentée d'un point de vue pratique de manière à permettre une prise en main rapide de la méthode. Elles sont illustrées sur le même exemple de spectroscopie, domaine d'application désormais très classique pour ces méthodes.

Table des matières

Chapitre 1

La régression linéaire simple

1.1 Introduction

L'origine du mot régression vient de Sir Francis Galton. En 1885, travaillant sur l'hérédité, il chercha à expliquer la taille des fils en fonction de celle des pères. Il constata que lorsque le père était plus grand que la moyenne, *taller than mediocrity*, son fils avait tendance à être plus petit que lui et, *a contrario*, que lorsque le père était plus petit que la moyenne, *shorter than mediocrity*, son fils avait tendance à être plus grand que lui. Ces résultats l'ont conduit à considérer sa théorie de *regression toward mediocrity*. Cependant l'analyse de causalité entre plusieurs variables est plus ancienne et remonte au milieu du XVIII^{e} siècle. En 1757, R. Boscovich, né à Ragussa, l'actuelle Dubrovnik, proposa une méthode minimisant la somme des valeurs absolues entre un modèle de causalité et les observations. Ensuite Legendre dans son célèbre article de 1805, « Nouvelles méthodes pour la détermination des orbites des comètes », introduit la méthode d'estimation par moindres carrés des coefficients d'un modèle de causalité et donna le nom à la méthode. Parallèlement, Gauss publia en 1809 un travail sur le mouvement des corps célestes qui contenait un développement de la méthode des moindres carrés, qu'il affirmait utiliser depuis 1795 (Birkes & Dodge, 1993).

Dans ce chapitre, nous allons analyser la régression linéaire simple : nous pouvons la voir comme une technique statistique permettant de modéliser la relation linéaire entre une variable explicative (notée X) et une variable à expliquer (notée Y). Cette présentation va nous permettre d'exposer la régression linéaire dans un cas simple afin de bien comprendre les enjeux de cette méthode, les problèmes posés et les réponses apportées.

1.1.1 Un exemple : la pollution de l'air

La pollution de l'air constitue actuellement une des préoccupations majeures de santé publique. De nombreuses études épidémiologiques ont permis de mettre

en évidence l'influence sur la santé de certains composés chimiques comme le dioxyde de souffre (SO_2), le dioxyde d'azote (NO_2), l'ozone (O_3) ou des particules sous forme de poussières contenues dans l'air. L'influence de cette pollution est notable sur les personnes sensibles (nouveau-nés, asthmatiques, personnes âgées). La prévision des pics de concentration de ces composés est donc importante.

Nous allons nous intéresser plus particulièrement à la concentration en ozone. Nous possédons quelques connaissances *a priori* sur la manière dont se forme l'ozone, grâce aux lois régissant les équilibres chimiques. La concentration de l'ozone sera fonction de la température ; plus la température sera élevée, plus la concentration en ozone va augmenter. Cette relation très vague doit être améliorée afin de pouvoir prédire les pics d'ozone.

Afin de mieux comprendre ce phénomène, l'association Air Breizh (surveillance de la qualité de l'air en Bretagne) mesure depuis 1994 la concentration en O_3 (en μg/ml) toute les 10 minutes et obtient donc le maximum journalier de la concentration en O_3, noté dorénavant `O3`. Air Breizh collecte également à certaines heures de la journée des données météorologiques comme la température, la nébulosité, le vent... Les données sont disponibles en ligne (cf. Avant-propos). Le tableau suivant donne les 10 premières mesures effectuées.

Tableau 1.1. 10 données de température à 12 h et teneur en ozone.

Individu	`O3`	`T12`
1	63.6	13.4
2	89.6	15
3	79	7.9
4	81.2	13.1
5	88	14.1
6	68.4	16.7
7	139	26.8
8	78.2	18.4
9	113.8	27.2
10	41.8	20.6

Nous allons donc chercher à expliquer le maximum de `O3` de la journée par la température à 12 h. D'un point de vue pratique *le but de cette régression est double* :

- ajuster un modèle pour expliquer la concentration en `O3` en fonction de T12 ;
- prédire les valeurs de concentration en `O3` pour de nouvelles valeurs de T12.

Avant toute analyse, il est intéressant de représenter les données. Voici donc une représentation graphique des données. Chaque point du graphique (fig.1.1) représente, pour un jour donné, une mesure de la température à 12 h et le pic d'ozone de la journée.

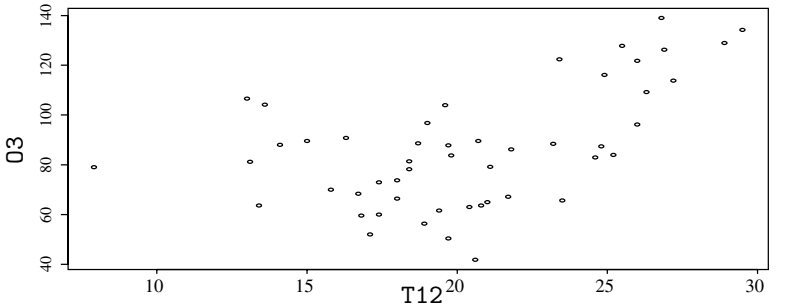

Fig. 1.1. 50 données journalières de température et O3.

Pour analyser la relation entre les x_i (température) et les y_i (ozone), nous allons chercher une fonction f telle que

$$y_i \approx f(x_i).$$

Pour définir $\approx$, il faut donner un critère quantifiant la qualité de l'ajustement de la fonction f aux données et une classe de fonctions $\mathcal{G}$ dans laquelle est supposée se trouver la vraie fonction inconnue. Le problème mathématique peut s'écrire de la façon suivante :

$$\underset{f \in \mathcal{G}}{\operatorname{argmin}} \sum_{i=1}^{n} l(y_i - f(x_i)), \tag{1.1}$$

où n représente le nombre de données à analyser et $l(.)$ est appelée *fonction de coût* ou encore *fonction de perte*.

1.1.2 Un deuxième exemple : la hauteur des arbres

Cet exemple utilise des données fournies par l'UR2PI et le CIRAD forêt (cf. Avant-propos). Lorsque le forestier évalue la vigueur d'une forêt, il considère souvent la hauteur des arbres qui la compose. Plus les arbres sont hauts, plus la forêt ou la plantation produit. Si l'on cherche à quantifier la production par le volume de bois, il est nécessaire d'avoir la hauteur de l'arbre pour calculer le volume de bois grâce à une formule du type « tronc de cône ». Cependant, mesurer la hauteur d'un arbre d'une vingtaine de mètres n'est pas aisé et demande un dendromètre. Ce type d'appareil mesure un angle entre le sol et le sommet de l'arbre. Il nécessite donc une vision claire de la cime de l'arbre et un recul assez grand afin d'avoir une mesure précise de l'angle et donc de la hauteur.

Dans certains cas, il est impossible de mesurer la hauteur, car ces deux conditions ne sont pas réunies, ou la mesure demande quelquefois trop de temps ou encore le forestier n'a pas de dendromètre. Il est alors nécessaire d'estimer la hauteur grâce à une mesure simple, la mesure de la circonférence à 1 mètre 30 du sol.

Nous possédons des mesures sur des eucalyptus dans une parcelle plantée et nous souhaitons à partir de ces mesures élaborer un modèle de prévision de la hauteur. Les eucalyptus étant plantés pour servir de matière première dans la pâte à papier, ils sont vendus au volume de bois. Il est donc important de connaître le volume et par là même la hauteur, afin d'évaluer la réserve en matière première dans la plantation (ou volume sur pied total). Les surfaces plantées sont énormes, il n'est pas question de prendre trop de temps pour la mesure et prévoir la hauteur par la circonférence est une méthode permettant la prévision du volume sur pied. La parcelle d'intérêt est constituée d'eucalyptus de 6 ans, âge de « maturité » des eucalyptus, c'est-à-dire l'âge en fin de rotation avant la coupe. Dans cette parcelle, nous avons alors mesuré $n = 1429$ couples circonférence-hauteur. Le tableau suivant donne les 10 premières mesures effectuées.

Tableau 1.2. Hauteur et circonférence (`ht` et `circ`) des 10 premiers eucalyptus.

Individu	`ht`	`circ`
1	18.25	36
2	19.75	42
3	16.50	33
4	18.25	39
5	19.50	43
6	16.25	34
7	17.25	37
8	19.00	41
9	16.25	27
10	17.50	30

Nous souhaitons donc expliquer la hauteur par la circonférence. Avant toute modélisation, nous représentons les données. Chaque point du graphique 1.2 représente une mesure du couple circonférence/hauteur sur un eucalyptus.

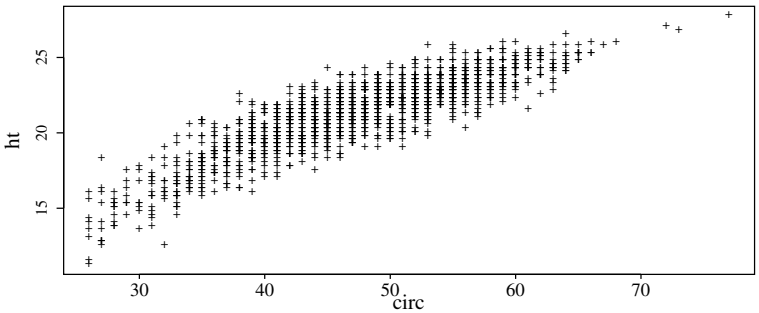

Fig. 1.2. Représentation des mesures pour les $n = 1429$ eucalyptus mesurés.

Pour prévoir la hauteur en fonction de la circonférence, nous allons donc chercher une fonction f telle que

$$y_i \approx f(x_i)$$

pour chaque mesure $i \in \{1, \ldots, 1429\}$.

Afin de quantifier précisément le symbole $\approx$, nous allons choisir une classe de fonctions $\mathcal{G}$. Cette classe représente tous les modèles de prévisions que l'on s'autorise afin de prévoir la hauteur en fonction de la circonférence. Ensuite, nous cherchons parmi ces modèles le meilleur, c'est-à-dire nous cherchons la fonction de $\mathcal{G}$ qui soit la plus proche possible des données selon une fonction de coût. Cela s'écrit

$$\arg\min_{f \in \mathcal{G}} \sum_{i=1}^{n} l(y_i - f(x_i)),$$

où n représente le nombre de données à analyser et $l(.)$ est appelée *fonction de coût* ou encore *fonction de perte*.

Remarque
Le calcul du volume proposé ici est donc fait en deux étapes : dans la première on estime la hauteur et dans la seconde on utilise une formule de type « tronc de cône » pour calculer le volume avec la hauteur estimée et la circonférence. Une autre méthode de calcul de volume consiste à ne pas utiliser de formule incluant la hauteur et prévoir directement le volume en une seule étape. Pour cela il faut calibrer le volume en fonction de la circonférence et il faut donc la mesure de nombreux volumes en fonction de circonférences, ce qui est très coûteux et difficile à réactualiser.

1.2 Modélisation mathématique

Nous venons de voir que le problème mathématique peut s'écrire de la façon suivante (cf. équation 1.1) :

$$\arg\min_{f \in \mathcal{G}} \sum_{i=1}^{n} l(y_i - f(x_i)),$$

où $l(.)$ est appelée *fonction de coût* et $\mathcal{G}$ un ensemble de fonctions données. Dans la suite de cette section, nous allons discuter du choix de la fonction de coût et de l'ensemble $\mathcal{G}$. Nous présenterons des graphiques illustratifs bâtis à partir de 10 données fictives de température et de concentration en ozone.

1.2.1 Choix du critère de qualité et distance à la droite

De nombreuses fonctions de coût $l(.)$ existent, mais les deux principales utilisées sont les suivantes :

– $l(u) = u^2$ coût quadratique ;
– $l(u) = |u|$ coût absolu.
Ces deux fonctions sont représentées sur le graphique 1.3 :

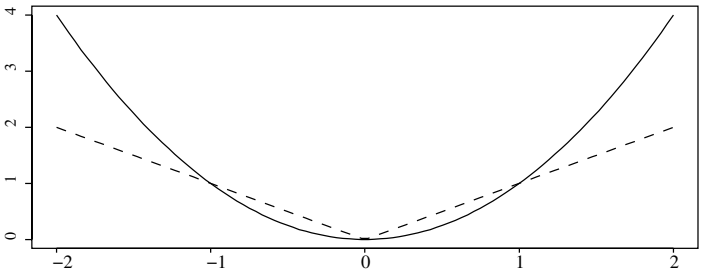

Fig. 1.3. Coût absolu (pointillés) et coût quadratique (trait plein).

Ces fonctions sont positives, symétriques, elles donnent donc la même valeur lorsque l'erreur est positive ou négative et s'annulent lorsque u vaut zéro.

La fonction l peut aussi être vue comme la distance entre une observation (x_i, y_i) et son point correspondant sur la droite $(x_i, f(x_i))$ (voir fig. 1.4).

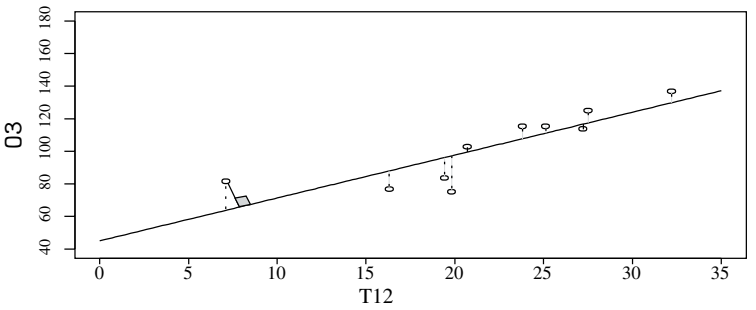

Fig. 1.4. Distances à la droite : coût absolu (pointillés) et distance d'un point à une droite.

Par point correspondant, nous entendons « évalué » à la même valeur x_i. Nous aurions pu prendre comme critère à minimiser la somme des distances des points (x_i, y_i) à la droite [1] (cf. fig. 1.4), mais ce type de distance n'entre pas dans le cadre des fonctions de coût puisqu'au point (x_i, y_i) correspond sur la droite un point $(x_i', f(x_i'))$ d'abscisse et d'ordonnée différentes.

Il est évident, que par rapport au coût absolu, le coût quadratique accorde une importance plus grande aux points qui restent éloignés de la droite ajustée, la distance étant élevée au carré (cf. fig. 1.3). Sur l'exemple fictif, dans la classe

[1] La distance d'un point à une droite est la longueur de la perpendiculaire à cette droite passant par ce point.

$\mathcal{G}$ des fonctions linéaires, nous allons minimiser $\sum_{i=1}^{n}(y_i - f(x_i))^2$ (coût quadratique) et $\sum_{i=1}^{n}|y_i - f(x_i)|$ (coût absolu). Les droites ajustées sont représentées sur le graphique ci-dessous :

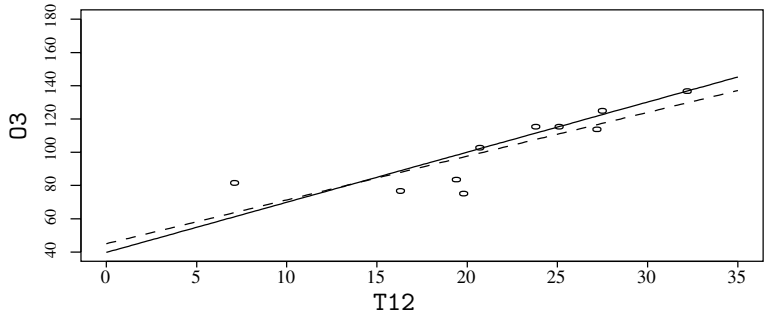

Fig. 1.5. 10 données fictives de température et O3, régressions avec un coût absolu (trait plein) et quadratique (pointillé).

La droite ajustée avec un coût quadratique propose un compromis où aucun point n'est très éloigné de la droite : le coût quadratique est sensible aux points aberrants qui sont éloignés de la droite. Ainsi (fig. 1.5) le premier point d'abscisse approximative 7^oC est assez éloigné des autres. La droite ajustée avec un coût quadratique lui accorde une plus grosse importance que l'autre droite et passe relativement donc plus près de lui. En enlevant ce point (de manière imaginaire), la droite ajustée risque d'être très différente : le point est dit influent et le coût quadratique peu robuste. Le coût absolu est plus robuste et la modification d'une observation modifie moins la droite ajustée. Les notions de points influents, points aberrants, seront approfondies au chapitre 4.

Malgré cette non-robustesse, le coût quadratique est le coût le plus souvent utilisé, ceci pour plusieurs raisons : historique, calculabilité, propriétés mathématiques. En 1800, il n'existait pas d'ordinateur et l'utilisation du coût quadratique permettait de calculer explicitement les estimateurs à partir des données. A propos de l'utilisation d'autres fonctions de coût, voici ce que disait Gauss (1809) : « Mais de tous ces principes, celui des moindres carrés est le plus simple : avec les autres, nous serions conduits aux calculs les plus complexes ». En conclusion, *seul le coût quadratique sera automatiquement utilisé dans la suite de ce livre, sauf mention contraire.* Les lecteurs intéressés par le coût absolu peuvent consulter le livre de Dodge & Rousson (2004).

1.2.2 Choix des fonctions à utiliser

Si la classe $\mathcal{G}$ est trop large, par exemple la classe des fonctions continues ($\mathcal{C}_0$), un grand nombre de fonctions de cette classe minimisent le critère (1.1). Ainsi toutes les fonctions de la classe qui passent par tous les points (interpolation), quand c'est possible, annulent la quantité $\sum_{i=1}^{n} l(y_i - f(x_i))$.

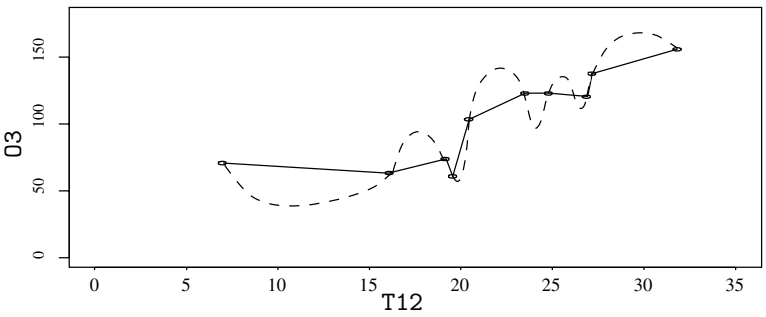

Fig. 1.6. Deux fonctions continues annulant le critère (1.1).

La fonction continue tracée en pointillés sur la figure (fig. 1.6) semble inappropriée bien qu'elle annule le critère (1.1). La fonction continue tracée en traits pleins annule aussi le critère (1.1). D'autres fonctions continues annulent ce critère, la classe des fonctions continues est trop vaste. Ces fonctions passent par tous les points et c'est là leur principal défaut. Nous souhaitons plutôt une courbe, ne passant pas par tous les points, mais possédant un trajet harmonieux, sans trop de détours. Bien sûr le trajet sans aucun détour est la ligne droite et la classe $\mathcal{G}$ la plus simple sera l'ensemble des fonctions affines. Par abus de langage, on emploie le terme de fonctions linéaires. D'autres classes de fonctions peuvent être choisies et ce choix est en général dicté par une connaissance *a priori* du phénomène et (ou) par l'observation des données.

Ainsi une étude de régression linéaire simple débute toujours par un tracé des observations (x, y). Cette première représentation permet de savoir si le modèle linéaire est pertinent. Le graphique suivant représente trois nuages de points différents.

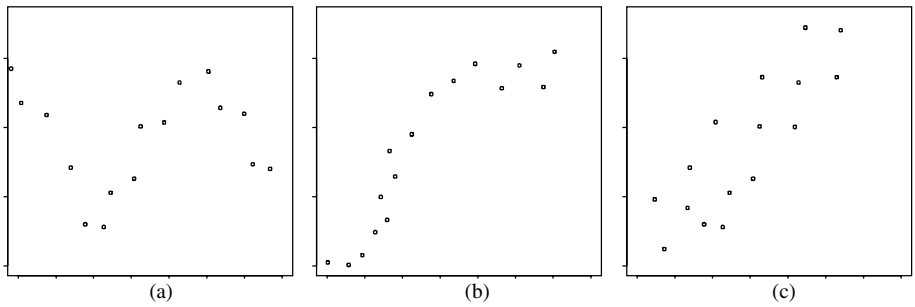

Fig. 1.7. Exemples fictifs de tracés : (a) fonction sinusoïdale, (b) fonction croissante sigmoïdale et (c) droite.

Au vu du graphique, il semble inadéquat de proposer une régression linéaire pour les deux premiers graphiques, le tracé présentant une forme sinusoïdale ou

sigmoïdale. Par contre, la modélisation par une droite de la relation entre X et Y pour le dernier graphique semble correspondre à la réalité de la liaison. Dans la suite de ce chapitre, nous prendrons $\mathcal{G} = \left\{f : f(x) = ax + b, \quad (a, b) \in \mathbb{R}^2\right\}$.

1.3 Modélisation statistique

Lorsque nous ajustons par une droite les données, nous supposons implicitement qu'elles étaient de la forme

$$Y = \beta_1 + \beta_2 X.$$

Dans l'exemple de l'ozone, nous supposons donc un modèle où la concentration d'ozone dépend linéairement de la température. Nous savons pertinemment que toutes les observations mesurées ne sont pas sur la droite. D'une part, il est irréaliste de croire que la concentration de l'ozone dépend linéairement de la température et de la température seulement. D'autre part, les mesures effectuées dépendent de la précision de l'appareil de mesure, de l'opérateur et il arrive souvent que, pour des valeurs identiques de la variable X, nous observions des valeurs différentes pour Y.

Nous supposons alors que la concentration d'ozone dépend linéairement de la température mais cette liaison est perturbée par un « bruit ». Nous supposons en fait que les données suivent le modèle suivant :

$$Y = \beta_1 + \beta_2 X + \varepsilon. \tag{1.2}$$

L'équation (1.2) est appelée **modèle de régression linéaire** et dans ce cas précis **modèle de régression linéaire simple**. Les β_j, appelés les paramètres du modèle (constante de régression et coefficient de régression), sont fixes mais inconnus, et nous voulons les estimer. La quantité notée ε est appelée bruit, ou erreur, et est aléatoire et inconnue.

Afin d'estimer les paramètres inconnus du modèle, nous mesurons dans le cadre de la régression simple une seule variable explicative ou variable exogène X et une variable à expliquer ou variable endogène Y. La variable X est souvent considérée comme non aléatoire au contraire de Y. Nous mesurons alors n observations de la variable X, notées x_i, où i varie de 1 à n et n valeurs de la variable à expliquer Y notées y_i.

Nous supposons que nous avons collecté n couples de données (x_i, y_i) où y_i est la réalisation de la variable aléatoire Y_i. Par abus de notation, nous confondrons la variable aléatoire Y_i et sa réalisation, l'observation y_i. Avec la notation ε_i, nous confondrons la variable aléatoire avec sa réalisation. Suivant le modèle (1.2), nous pouvons écrire

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i, \qquad i = 1, \cdots, n$$

où

– les x_i sont des valeurs connues non aléatoires ;

– les paramètres β_j, $j = 1, 2$ du modèle sont inconnus ;
– les ε_i sont les réalisations inconnues d'une variable aléatoire ;
– les y_i sont les observations d'une variable aléatoire.

1.4 Estimateurs des moindres carrés

Définition 1.1 (estimateurs des MC)
On appelle estimateurs des moindres carrés (MC) de β_1 et β_2, les estimateurs $\hat{\beta}_1$ et $\hat{\beta}_2$ obtenus par minimisation de la quantité

$$S(\beta_1, \beta_2) = \sum_{i=1}^{n} (y_i - \beta_1 - \beta_2 x_i)^2 = \|Y - \beta_1 \mathbb{1} - \beta_2 X\|^2,$$

où $\mathbb{1}$ est le vecteur de $\mathbb{R}^n$ dont tous les coefficients valent 1. Les estimateurs peuvent également s'écrire sous la forme suivante :

$$(\hat{\beta}_1, \hat{\beta}_2) = \underset{(\beta_1, \beta_2) \in \mathbb{R} \times \mathbb{R}}{\operatorname{argmin}} S(\beta_1, \beta_2).$$

1.4.1 Calcul des estimateurs de β_j, quelques propriétés

La fonction $S(\beta_1, \beta_2)$ est strictement convexe. Si elle admet un point singulier, celui-ci correspond à l'unique minimum. Annulons les dérivées partielles, nous obtenons un système d'équations appelées « équations normales » :

$$\begin{cases} \dfrac{\partial S(\hat{\beta}_1, \hat{\beta}_2)}{\partial \beta_1} &= -2 \displaystyle\sum_{i=1}^{n} (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) = 0, \\ \dfrac{\partial S(\hat{\beta}_1, \hat{\beta}_2)}{\partial \beta_2} &= -2 \displaystyle\sum_{i=1}^{n} x_i (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) = 0. \end{cases}$$

La première équation donne

$$\hat{\beta}_1 n + \hat{\beta}_2 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

et nous avons un estimateur de l'ordonnée à l'origine

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x}, \tag{1.3}$$

où $\bar{x} = \sum x_i / n$. La seconde équation donne

$$\hat{\beta}_1 \sum_{i=1}^{n} x_i + \hat{\beta}_2 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i.$$

En remplaçant $\hat{\beta}_1$ par son expression (1.3) nous avons une première écriture de

$$\hat{\beta}_2 = \frac{\sum x_i y_i - \sum x_i \bar{y}}{\sum x_i^2 - \sum x_i \bar{x}},$$

et en utilisant astucieusement la nullité de la somme $\sum(x_i - \bar{x})$, nous avons d'autres écritures pour l'estimateur de la pente de la droite

$$\hat{\beta}_2 = \frac{\sum x_i(y_i - \bar{y})}{\sum x_i(x_i - \bar{x})} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})(x_i - \bar{x})} = \frac{\sum(x_i - \bar{x})y_i}{\sum(x_i - \bar{x})^2}. \tag{1.4}$$

Pour obtenir ce résultat, nous supposons qu'il existe au moins deux points d'abscisses différentes. Cette hypothèse notée $\mathcal{H}_1$ s'écrit $x_i \neq x_j$ pour au moins deux individus. Elle permet d'obtenir l'unicité des coefficients estimés $\hat{\beta}_1, \hat{\beta}_2$.

Une fois déterminés les estimateurs $\hat{\beta}_1$ et $\hat{\beta}_2$, nous pouvons estimer la droite de régression par la formule

$$\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X.$$

Si nous évaluons la droite aux points x_i ayant servi à estimer les paramètres, nous obtenons des $\hat{y}_i$ et ces valeurs sont appelées les valeurs ajustées. Si nous évaluons la droite en d'autres points, les valeurs obtenues seront appelées les valeurs prévues ou prévisions. Représentons les points initiaux et la droite de régression estimée. La droite de régression passe par le centre de gravité du nuage de points $(\bar{x}, \bar{y})$ comme l'indique l'équation (1.3).

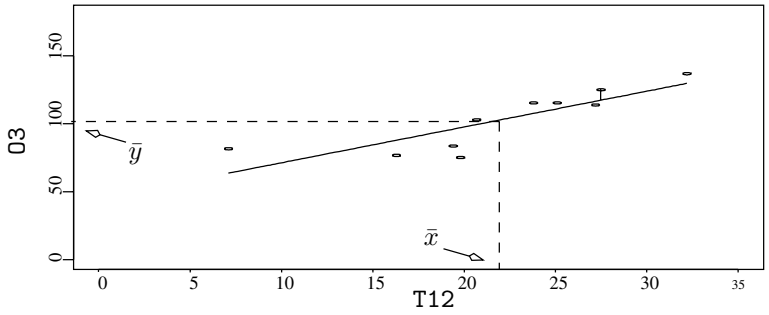

Fig. 1.8. Nuage de points, droite de régression et centre de gravité.

Nous avons réalisé une expérience et avons mesuré n valeurs (x_i, y_i). A partir de ces n valeurs, nous avons obtenu un estimateur de β_1 et de β_2. Si nous refaisions une expérience, nous mesurerions n nouveaux couples de données (x_i, y_i). A partir de ces données, nous aurions un nouvel estimateur de β_1 et de β_2. Les estimateurs sont fonction des données mesurées et changent donc avec les observations collectées (fig. 1.9). Les vraies valeurs de β_1 et β_2 sont inconnues et ne changent pas.

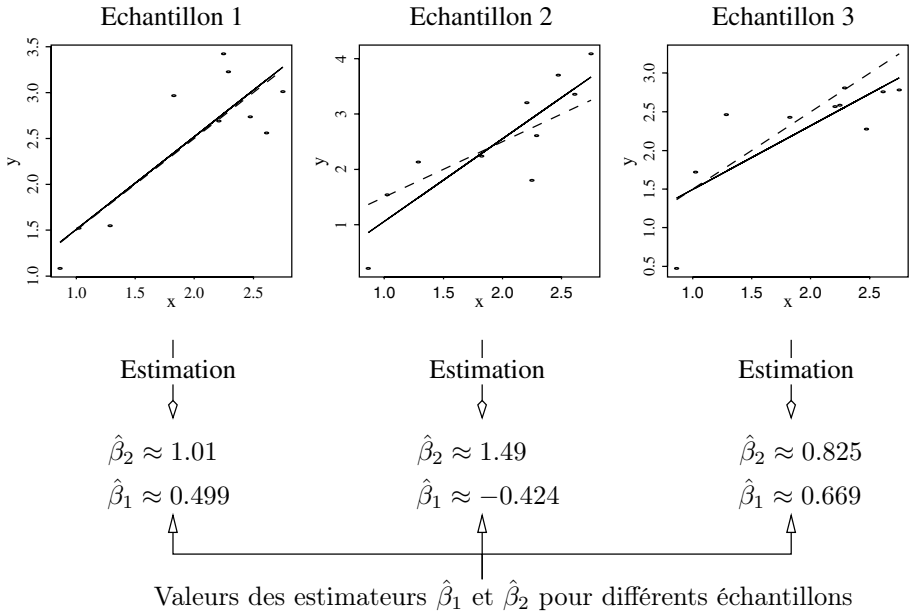

Fig. 1.9. Exemple de la variabilité des estimations. Le vrai modèle est $Y = X + 0.5 + \varepsilon$, où ε est choisi comme suivant une loi $\mathcal{N}(0, 0.25)$. Nous avons ici 3 répétitions de la mesure de 10 points (x_i, y_i), ou 3 échantillons de taille 10. Le trait en pointillé représente la vraie droite de régression et le trait plein son estimation.

Le statisticien cherche en général à vérifier que les estimateurs utilisés admettent certaines propriétés comme :

- un estimateur $\hat{\beta}$ est-il sans biais ? Par définition $\hat{\beta}$ est sans biais si $\mathbb{E}(\hat{\beta}) = \beta$. En moyenne sur toutes les expériences possibles de taille n, l'estimateur $\hat{\beta}$ moyen sera égal à la valeur inconnue du paramètre. En français, cela signifie qu'en moyenne $\hat{\beta}$ « tombe » sur β ;
- un estimateur $\hat{\beta}$ est-il de variance minimale parmi les estimateurs d'une classe définie ? En d'autres termes, parmi tous les estimateurs de la classe, l'estimateur utilisé admet-il parmi toutes les expériences la plus petite variabilité ?

Pour cela, nous supposons une seconde hypothèse notée $\mathcal{H}_2$ qui s'énonce aussi comme suit : les erreurs sont centrées, de même variance (homoscédasticité) et non corrélées entre elles. Elle permet de calculer les propriétés statistiques des estimateurs. $\mathcal{H}_2 : \mathbb{E}(\varepsilon_i) = 0$, pour $i = 1, \cdots, n$ et $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = \delta_{ij}\sigma^2$, où $\mathbb{E}(\varepsilon)$ est l'espérance de ε, $\mathrm{Cov}(\varepsilon_i, \varepsilon_j)$ est la covariance entre ε_i et ε_j et $\delta_{ij} = 1$ lorsque $i = j$ et $\delta_{ij} = 0$ lorsque $i \neq j$. Nous avons la première propriété de ces estimateurs (voir exercice 1.2)

Proposition 1.1 (Biais des estimateurs)
$\hat{\beta}_1$ et $\hat{\beta}_2$ sont des estimateurs sans biais de β_1 et β_2, c'est-à-dire que

$\mathbb{E}(\hat{\beta}_1) = \beta_1$ *et* $\mathbb{E}(\hat{\beta}_2) = \beta_2$.

Les estimateurs $\hat{\beta}_1$ et $\hat{\beta}_2$ sont sans biais, nous allons nous intéresser à leur variance. Afin de montrer que ces estimateurs sont de variances minimales dans leur classe, nous allons d'abord calculer leur variance (voir exercices 1.3, 1.4, 1.5). C'est l'objet de la prochaine proposition.

Proposition 1.2 (Variances de $\hat{\beta}_1$ et $\hat{\beta}_2$)
Les variances et covariance des estimateurs des paramètres valent :

$$\begin{aligned} \mathrm{V}(\hat{\beta}_2) &= \frac{\sigma^2}{\sum(x_i - \bar{x})^2} \\ \mathrm{V}(\hat{\beta}_1) &= \frac{\sigma^2 \sum x_i^2}{n\sum(x_i - \bar{x})^2} \\ \mathrm{Cov}(\hat{\beta}_1, \hat{\beta}_2) &= -\frac{\sigma^2 \bar{x}}{\sum(x_i - \bar{x})^2}. \end{aligned}$$

Cette proposition nous permet d'envisager la précision de l'estimation en utilisant la variance. Plus la variance est faible, plus l'estimateur sera précis. Pour avoir des variances petites, il faut avoir un numérateur petit et (ou) un dénominateur grand. Les estimateurs seront donc de faibles variances lorsque :

- la variance σ^2 est faible. Cela signifie que la variance de Y est faible et donc les mesures sont proches de la droite à estimer ;
- la quantité $\sum(x_i - \bar{x})^2$ est grande, les mesures x_i doivent être dispersées autour de leur moyenne ;
- la quantité $\sum x_i^2$ ne doit pas être trop grande, les points doivent avoir une faible moyenne en valeur absolue. En effet, nous avons

$$\frac{\sum x_i^2}{\sum(x_i - \bar{x})^2} = \frac{\sum x_i^2 - n\bar{x}^2 + n\bar{x}^2}{\sum(x_i - \bar{x})^2} = 1 + \frac{n\bar{x}^2}{\sum(x_i - \bar{x})^2}.$$

L'équation (1.3) indique que la droite des MC passe par le centre de gravité du nuage $(\bar{x}, \bar{y})$. Supposons $\bar{x}$ positif, alors si nous augmentons la pente, l'ordonnée à l'origine va diminuer et vice versa. Nous retrouvons donc le signe négatif pour la covariance entre $\hat{\beta}_1$ et $\hat{\beta}_2$.

Nous terminons cette partie concernant les propriétés par le théorème de Gauss-Markov qui indique que, parmi tous les estimateurs linéaires sans biais, les estimateurs des MC possèdent la plus petite variance (voir exercice 1.6).

Théorème 1.1 (Gauss-Markov)
Parmi les estimateurs sans biais linéaires en Y, les estimateurs $\hat{\beta}_j$ sont de variance minimale.

1.4.2 Résidus et variance résiduelle

Nous avons estimé β_1 et β_2. La variance σ^2 des ε_i est le dernier paramètre inconnu à estimer. Pour cela, nous allons utiliser les résidus : ce sont des estimateurs des erreurs inconnues ε_i.

Définition 1.2 (Résidus)
Les résidus sont définis par

$$\hat{\varepsilon}_i = y_i - \hat{y}_i$$

où $\hat{y}_i$ est la valeur ajustée de y_i par le modèle, c'est-à-dire $\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i$.

Nous avons la propriété suivante (voir exercice 1.7).

Proposition 1.3
Dans un modèle de régression linéaire simple, la somme des résidus est nulle.

Intéressons-nous maintenant à l'estimation de σ^2 et construisons un estimateur sans biais $\hat{\sigma}^2$ (cf. exercice 1.8) :

Proposition 1.4 (Estimateur de la variance du bruit)
La statistique $\hat{\sigma}^2 = \sum_{i=1}^{n} \hat{\varepsilon}_i^2/(n-2)$ est un estimateur sans biais de σ^2.

1.4.3 Prévision

Un des buts de la régression est de proposer des prévisions pour la variable à expliquer Y. Soit x_{n+1} une nouvelle valeur de la variable X, nous voulons prédire y_{n+1}. Le modèle indique que

$$y_{n+1} = \beta_1 + \beta_2 x_{n+1} + \varepsilon_{n+1}$$

avec $\mathbb{E}(\varepsilon_{n+1}) = 0$, $\mathrm{V}(\varepsilon_{n+1}) = \sigma^2$ et $\mathrm{Cov}(\varepsilon_{n+1}, \varepsilon_i) = 0$ pour $i = 1, \cdots, n$. Nous pouvons prédire la valeur correspondante grâce au modèle estimé

$$\hat{y}_{n+1}^p = \hat{\beta}_1 + \hat{\beta}_2 x_{n+1}.$$

En utilisant la notation $\hat{y}_{n+1}^p$ nous souhaitons insister sur la notion de prévision : la valeur pour laquelle nous effectuons la prévision, ici la $(n+1)^{\mathrm{e}}$, n'a pas servi dans le calcul des estimateurs. Remarquons que cette quantité sera différente de la valeur ajustée, notée $\hat{y}_i$, qui elle fait intervenir la i^{e} observation.

Deux types d'erreurs vont entacher notre prévision, l'une due à la non-connaissance de ε_{n+1} et l'autre due à l'estimation des paramètres.

Proposition 1.5 (Variance de la prévision $\hat{y}_{n+1}^p$)
La variance de la valeur prévue de $\hat{y}_{n+1}^p$ vaut

$$\mathrm{V}\left(\hat{y}_{n+1}^p\right) = \sigma^2 \left(\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right).$$

La variance de $\hat{y}_{n+1}^p$ (voir exercice 1.9) nous donne une idée de la stabilité de l'estimation. En prévision, on s'intéresse généralement à l'erreur que l'on commet entre la vraie valeur à prévoir y_{n+1} et celle que l'on prévoit $\hat{y}_{n+1}^p$. L'erreur peut être simplement résumée par la différence entre ces deux valeurs, c'est ce que nous appellerons l'erreur de prévision. Cette erreur de prévision permet de quantifier la capacité du modèle à prévoir. Nous avons sur ce thème la proposition suivante (voir exercice 1.10).

Proposition 1.6 (Erreur de prévision)
L'erreur de prévision, définie par $\hat{\varepsilon}^p_{n+1} = y_{n+1} - \hat{y}^p_{n+1}$ *satisfait les propriétés suivantes :*

$$\begin{aligned} \mathbb{E}(\hat{\varepsilon}^p_{n+1}) &= 0 \\ \mathrm{V}(\hat{\varepsilon}^p_{n+1}) &= \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right). \end{aligned}$$

Remarque
La variance augmente lorsque x_{n+1} s'éloigne du centre de gravité du nuage. Effectuer une prévision lorsque x_{n+1} est « loin » de $\bar{x}$ est donc périlleux, la variance de l'erreur de prévision peut alors être très grande !

1.5 Interprétations géométriques

1.5.1 Représentation des individus

Pour chaque individu, ou observation, nous mesurons une valeur x_i et une valeur y_i. Une observation peut donc être représentée dans le plan, nous dirons alors que $\mathbb{R}^2$ est l'espace des observations. $\hat{\beta}_1$ correspond à l'ordonnée à l'origine alors que $\hat{\beta}_2$ représente la pente de la droite ajustée. Cette droite minimise la somme des carrés des distances verticales des points du nuage à la droite ajustée.

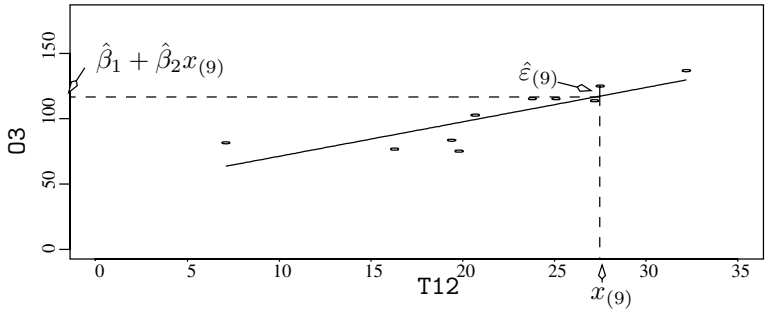

Fig. 1.10. Représentation des individus.

Les couples d'observations (x_i, y_i) avec $i = 1, \ldots, n$ ordonnées suivant les valeurs croissantes de x sont notés $(x_{(i)}, y_{(i)})$. Nous avons représenté la neuvième valeur de x et sa valeur ajustée $\hat{y}_{(9)} = \hat{\beta}_1 + \hat{\beta}_2 x_{(9)}$ sur le graphique, ainsi que le résidu correspondant $\hat{\varepsilon}_{(9)}$.

1.5.2 Représentation des variables

Nous pouvons voir le problème d'une autre façon. Nous mesurons n couples de points (x_i, y_i). La variable X et la variable Y peuvent être considérées

comme deux vecteurs possédant n coordonnées. Le vecteur X (respectivement Y) admet pour coordonnées : les observations $x_1, x_2, \ldots, x_n$ (respectivement $y_1, y_2, \ldots, y_n$). Ces deux vecteurs d'observations appartiennent au même espace $\mathbb{R}^n$: l'espace des variables. Nous pouvons donc représenter les données dans l'espace des variables. Le vecteur $\mathbb{1}$ est également un vecteur de $\mathbb{R}^n$ dont toutes les composantes valent 1. Les 2 vecteurs $\mathbb{1}$ et X engendrent un sous-espace de $\mathbb{R}^n$ de dimension 2. Nous avons supposé que $\mathbb{1}$ et X ne sont pas colinéaires grâce à $\mathcal{H}_1$ mais ces vecteurs ne sont pas obligatoirement orthogonaux. Ces vecteurs sont orthogonaux lorsque $\bar{x}$, la moyenne des observations $x_1, x_2, \ldots, x_n$ vaut zéro (voir la remarque ci-dessous).

La régression linéaire peut être vue comme la projection orthogonale du vecteur Y dans le sous-espace de $\mathbb{R}^n$ engendré par $\mathbb{1}$ et X, noté $\Im(X)$. Les coefficients $\hat{\beta}_1$ et $\hat{\beta}_2$ s'interprètent comme les composantes de la projection orthogonale notée $\hat{Y}$ de Y sur ce sous-espace. Voyons cela sur le graphique suivant :

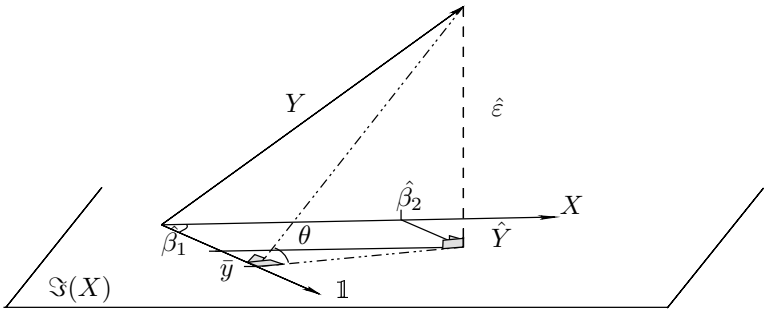

Fig. 1.11. Représentation de la projection dans l'espace des variables.

Remarque
Les vecteurs $\mathbb{1}$ et X de normes respectives $\sqrt{n}$ et $\sqrt{\sum_{i=1}^{n} x_i^2}$ ne forment pas une base orthogonale. Afin de savoir si ces vecteurs sont orthogonaux, calculons leur produit scalaire. Le produit scalaire est la somme du produit terme à terme des composantes des deux vecteurs et vaut ici $\sum_{i=1}^{n} x_i \times 1 = n\bar{x}$. Les vecteurs forment une base orthogonale lorsque la moyenne de X est nulle. En effet $\bar{x}$ vaut alors zéro et le produit scalaire est nul. Les vecteurs n'étant en général pas orthogonaux, cela veut dire que $\hat{\beta}_1 \mathbb{1}$ n'est pas la projection de Y sur la droite engendrée par $\mathbb{1}$ et que $\hat{\beta}_2 X$ n'est pas la projection de Y sur la droite engendrée par X. Nous reviendrons sur cette différence au chapitre suivant.

1.5.3 Le coefficient de détermination R^2

Un modèle, que l'on qualifiera de bon, possédera des estimations $\hat{y}_i$ proches des vraies valeurs y_i. Sur la représentation dans l'espace des variables (fig. 1.11) la qualité peut être évaluée par l'angle θ. Cet angle est compris entre -90 ° et

90 ° . Un angle proche de -90 ° ou de 90 ° indique un modèle de mauvaise qualité. Le cosinus carré de θ est donc une mesure possible de la qualité du modèle et cette mesure varie entre 0 et 1.

Le théorème de Pythagore nous donne directement que

$$\begin{aligned} \|Y - \bar{y}\mathbb{1}\|^2 &= \|\hat{Y} - \bar{y}\mathbb{1}\|^2 + \|\hat{\varepsilon}\|^2 \\ \sum_{i=1}^{n}(y_i - \bar{y})^2 &= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}\hat{\varepsilon}_i^2 \\ \mathrm{SCT} &= \mathrm{SCE} + \mathrm{SCR}, \end{aligned}$$

où SCT (respectivement SCE et SCR) représente la somme des carrés totale (respectivement expliquée par le modèle et résiduelle).

Le coefficient de détermination R^2 est défini par

$$\mathrm{R}^2 = \frac{\mathrm{SCE}}{\mathrm{SCT}} = \frac{\|\hat{Y} - \bar{y}\mathbb{1}\|^2}{\|Y - \bar{y}\mathbb{1}\|^2},$$

c'est-à-dire la part de la variabilité expliquée par le modèle sur la variabilité totale. De nombreux logiciels multiplient cette valeur par 100 afin de donner un pourcentage.

Remarques
Dans ce cas précis, R^2 est le carré du coefficient de corrélation empirique entre les x_i et les y_i et

- le R^2 correspond au cosinus carré de l'angle θ ;
- si $\mathrm{R}^2 = 1$, le modèle explique tout, l'angle θ vaut donc zéro, Y est dans $\Im(X)$ c'est-à-dire que $y_i = \beta_1 + \beta_2 x_i$;
- si $\mathrm{R}^2 = 0$, cela veut dire que $\sum(\hat{y}_i - \bar{y})^2 = 0$ et donc que $\hat{y}_i = \bar{y}$. Le modèle de régression linéaire est inadapté ;
- si R^2 est proche de zéro, cela veut dire que Y est quasiment dans l'orthogonal de $\Im(X)$, le modèle de régression linéaire est inadapté, la variable X utilisée n'explique pas la variable Y.

1.6 Inférence statistique

Jusqu'à présent, nous avons pu, en choisissant une fonction de coût quadratique, ajuster un modèle de régression, à savoir calculer $\widehat{\beta}_1$ et $\widehat{\beta}_2$. Grâce aux coefficients estimés, nous pouvons donc prédire, pour chaque nouvelle valeur x_{n+1} une valeur de la variable à expliquer $\widehat{y}^p_{n+1}$ qui est tout simplement le point sur la droite ajustée correspondant à l'abscisse x_{n+1}. En ajoutant l'hypothèse $\mathcal{H}_2$, nous avons pu calculer l'espérance et la variance des estimateurs. Ces propriétés permettent d'appréhender de manière grossière la qualité des estimateurs proposés. Le théorème de Gauss-Markov permet de juger de la qualité