Ifremer *Centre de Brest*

UTILISATION DU MODELE LINEAIRE.

Rappels de base – Méthodes de validation.

Philippe GROS

Avertissement.

L'essentiel de ce texte a été publié en juillet 2000 dans la collection «documents océanographiques» de l'Institut Océanographique de Paris [*Océanis* 23(3) : 359 - 515, 1997]. Cette version complétée a été éditée pour diffusion élargie, entre autres aux Laboratoires Côtiers de l'Ifremer.

Novembre 2000

Résumé.

La régression linéaire est l'un des modèles statistiques les plus employés : son champ d'application s'étend de la description et de l'analyse des données expérimentales jusqu'à la prévision, et il est aussi utilisé pour l'interpolation, ou pour l'aide à la mise en évidence de relations causales, par exemple. Il est par conséquent indispensable que le praticien possède une solide connaissance des prérequis, de la portée et des limites du modèle linéaire. C'est pourquoi les deux premiers chapitres sont consacrés au modèle "classique" simple (une seule variable contrôlée, ou régresseur) : le chapitre 1 expose, principalement à l'aide de représentations géométriques, les principes généraux de l'identification du modèle par les moindres carrés ordinaires. Le chapitre 2 aborde la modélisation statistique proprement dite ; il rappelle d'abord les concepts essentiels de la théorie de l'estimation, il présente ensuite l'estimation des paramètres du modèle par le maximum de vraisemblance, et il traite enfin l'inférence statistique dans le contexte gaussien, c'est-à-dire lorsque la loi des erreurs aléatoires indépendantes (les "résidus") est normale. Au chapitre 3 sont présentés d'une part la généralisation de cet ensemble de résultats au cas de la régression multiple (deux régresseurs ou plus), et d'autre part le diagnostic de l'impact de la "structure des régresseurs" (*i.e.*, des propriétés de la matrice du plan d'expérience) sur l'estimation des paramètres du modèle ; plusieurs palliatifs du problème de la "colinéarité" sont commentés à ce propos.
Les deux chapitres suivants s'adressent au coeur du sujet : l'utilisateur constate que dans la plupart des applications, les données "n'entrent pas exactement" dans le cadre défini par la théorie, qu'elles ne sont "pas tout à fait conformes" aux hypothèses formulées pour construire le modèle. La confrontation aux situations concrètes soulève ainsi trois grandes questions : comment reconnaître et caractériser la manière dont les données s'écartent du modèle postulé ? Quelles sont les conséquences de cette déviation sur la qualité des résultats obtenus ? Quelles mesures correctives peut-on appliquer ? Au chapitre 4 sont examinées les principales méthodes de mise en évidence et de gestion du non-respect des hypothèses de variance stable, d'absence d'autocorrélation, et éventuellement de normalité de la composante aléatoire du modèle linéaire. Au chapitre 5, l'attention est accordée à l'identification des points qui jouent dans l'ajustement un rôle prépondérant, et qui peuvent être parfois considérés comme douteux : plusieurs techniques complémentaires de détection des éléments influents et de reconnaissance des observations aberrantes sont présentées. Ce chapitre introduit en particulier l'application à la régression des outils de la statistique robuste (notions de fonction d'influence et de point de rupture d'un estimateur).
Le sixième chapitre mentionne enfin deux extensions du modèle linéaire : d'une part le modèle linéaire dit "généralisé" (au sens où la densité de probabilité du résidu peut être décrite par une loi de la famille exponentielle), et d'autre part la relation structurelle.

Abstract.

Linear regression modelling is one of the most widely used statistical tools: usual applications of this "core technique" encompass description and analysis of experimental data, interpolation, help in recognizing causal relationships, and forecasting. It is thus necessary for the practicioner to obtain an understanding of the basic principles necessary to apply regression methods in a variety of settings. Accordingly, the first two chapters provide the standard results for the simple linear regression (only one controlled variable, or regressor). Since the emphasis is on practical applications, theoretical results are stated without proof, and the major guidelines are building models, assessing fit and reliability, and drawing conclusions. Chapter 1 relies upon a geometrical approach to introduce the identification of the model by ordinary least squares. Chapter 2 focuses specifically on statistical modelling: the fundamental concepts of estimation theory are first recalled, and the maximum likelihood estimation of the model parameters is then presented; statistical inferences are treated here in the "classical" (gaussian) framework, *i.e.*, the errors are assumed to be independent and identical normal random variables. The generalization to the multiple linear regression model (two regressors at least) is described in Chapter 3 using matrix algebra; this chapter examines the design matrix properties generating multicollinearity problems: included are their sources, their harmful effects, and a review of available diagnostics and remedial measures.
The next two chapters form the nucleus of this practically oriented textbook on regression analysis, whose successful use requires a capacity both in checking the model adequacy, and in managing the practical difficulties that arise when the technique is employed with real-world data. Chapters 4 and 5 put therefore the emphasis on the art of exploratory data analysis rather than on statistical theory, and cover several procedures designed to detect various types of disagreement between observations and the assumed model. Chapter 4 introduces diagnostics for investigating departures from the usual assumptions on the random error component of the model (*e.g.*, heteroscedasticity, autocorrelation, or non-normality); remedial actions are also examined, for instance analytical methods for selecting transformations to stabilize residual variance. Chapter 5 goes beyond the residual analysis, by introducing methods for assessing the influence of individual observations, with the purpose of pinpointing outlying values both in the response variable and in the explanatory part of the model (the so-called "leverage points" in the latter case). This chapter also emphasizes a complementary line of inquiry, through the introduction of robust (or resistant) regression methods that require progressively fewer untenable assumptions, and whose results remain trustworthy even if a certain amount of observations are outliers. The concepts of breakdown point and influence function of an estimator are introduced; it is further stressed that robust methods provide powerful tools in identifying outliers, or, more generally, "troublesome" observations.
Despite its broad range of application, linear regression calls for generalizations; two of them are examined in Chapter 6: the first one is a brief introduction to logistic regression, which offers a didactic example of one special case in the class of generalized linear models; the second one deals with the structural relationship.

Novembre 2000

UTILISATION DU MODELE LINEAIRE.
RAPPELS DE BASE - METHODES DE VALIDATION.

Philippe GROS * — Ifremer — *Centre de Brest.*

* *E-mail* : phgros@ifremer.fr

Liminaire.

Ce document est destiné aux utilisateurs de l'outil statistique. Il constitue le support d'un enseignement dispensé aux étudiants qui abordent le troisième cycle d'océanographie biologique ; il a par ailleurs fait l'objet de plusieurs exposés dans le cadre de formations organisées au sein de l'IFREMER. Le niveau de connaissance nécessaire pour sa lecture correspond à celui acquis à l'issue du premier cycle universitaire d'une "filière" scientifique. Plus précisément, les bases de la théorie de l'estimation statistique, ainsi que celles des tests, sont supposées maîtrisées ; par précaution, les concepts essentiels de la théorie de l'estimation sont néanmoins rappelés au début de la deuxième partie.

Le premier objectif est de proposer à l'utilisateur un guide lui permettant d'exploiter au mieux les possibilités offertes par la régression linéaire : description de résultats expérimentaux, interpolation, prévision, aide à la recherche de liens causaux, ... Il ne s'agit nullement de dresser un "catalogue de recettes", mais au contraire d'amener le lecteur à s'interroger sur l'éventail des méthodes susceptibles d'être employées dans la pratique : c'est-à-dire, dans les situations (fréquentes !) où les données "n'entrent pas exactement" dans le cadre défini par la théorie.

Les deux premières parties sont donc logiquement consacrées à la présentation résumée du modèle classique ; elles sont simplement un aide-mémoire, qui privilégie l'exposé des résultats, sans recourir aux démonstrations formelles et détaillées que l'on trouvera dans les ouvrages de Statistique (*vide infra*, références citées).

· La première partie présente succinctement les principes généraux de l'estimation par les moindres carrés ordinaires (MCO) ; une large place y est accordée aux représentations géométriques, qui permettent d'appréhender directement plusieurs résultats établis dans le cadre de l'algèbre linéaire. Cette approche initiale, centrée sur les MCO, vise une simple description des données expérimentales, résumées à l'aide d'une droite, d'un plan, ..., par exemple. Il s'agit à ce niveau d'identifier un modèle.

· La seconde partie aborde le problème de la modélisation statistique proprement dite. En général, l'écriture d'un modèle qui résume les observations appelle des développements complémentaires : il est en particulier nécessaire de lui "donner un sens", *i.e.*, répondre à des questions telles que "peut on comparer les paramètre du modèle à des valeurs données *a priori* ?", ou encore " de quelle erreur sont entachées les prévisions réalisées à l'aide du modèle ?", par exemple. On aborde là le problème de la gestion des incertitudes. En ce sens, la seconde partie présente l'estimation par le maximum de vraisemblance, et traite des inférences statistiques usuelles dans le contexte du modèle probabiliste gaussien, *i.e.*, lorsque la loi des erreurs aléatoires indépendantes (les "résidus") est normale. Comme dans l'ensemble du document, c'est le point de vue "fréquentiste" qui est retenu.

Le lecteur averti pourra parcourir rapidement ces deux parties, à l'exception peut-être de la présentation des différences entre modèles linéaires et non linéaires (§ 2.17, exemple emprunté à Ratkowsky [1]).

· La troisième partie généralise à la régression multiple les résultats auparavant rappelés pour le modèle linéaire simple. Ainsi qu'il est désormais d'usage, c'est le formalisme matriciel qui est adopté : outre la présentation concise qu'il autorise, il convient de souligner que c'est aussi le formalisme employé dans les environnements logiciels évolués. L'exposé est délibérément limité : il n'inclut pas la régression sur variables qualitatives, qui établit le lien entre modèle linéaire et analyse de la variance appliquée dans le cadre des protocoles expérimentaux (l'élaboration de plans d'expérience est une spécialité à part entière, dont la théorie est présentée dans de nombreux ouvrages, tel celui de Scheffé [2], ou encore dans le récent manuel en langue française de Bergonzini & Duby [3]). Cependant, la question de la comparaison de droites de régression étant

régulièrement posée par les utilisateurs, la solution fondée sur l'emploi de variables indicatrices est présentée. Enfin, l'impact sur l'estimation des paramètres du modèle de la "structure des régresseurs" (*i.e.*, des propriétés de la matrice du plan d'expérience) est abordé sous l'angle du classique problème de la "colinéarité" ; les informations qui sont données sur ce point ont pour vocation essentielle de permettre la consultation des textes qui traitent le sujet en profondeur, par exemple le manuel de Belsley [4], ou plus simplement les chapitres *ad hoc* des ouvrages de Chatterjee & Price [5] et de Montgomery & Peck [6].

Les quatrième et cinquième parties, qui représentent *ca*. la moitié du document, concernent le coeur du sujet : l'utilisateur constate que dans la plupart des situations concrètes, les données expérimentales ne se conforment pas exactement aux contraintes requises par le modèle postulé. Cela pose trois grandes questions : comment reconnaître et caractériser la manière dont les données "s'écartent" du modèle théorique ? Quelles sont les conséquences de cette déviation sur la qualité des résultats ? Quelles mesures correctives peut-on appliquer ?

· La quatrième partie explore quelques unes des démarches qui visent à mettre en évidence le non-respect des hypothèses relatives à la composante aléatoire du modèle linéaire (variance stable, absence d'autocorrélation, et, éventuellement, loi normale), ainsi que les palliatifs envisageables. Parmi ces derniers figure la transformation des variables, famille de méthodes auxquelles est consacré le livre de Carroll & Ruppert [7].

· Dans la cinquième partie, l'attention est plutôt accordée à l'identification des points qui jouent dans l'ajustement un rôle prépondérant, et qui parfois peuvent être considérés comme douteux : plusieurs techniques complémentaires de détection de ces éléments influents sont présentées. Les références citées traitent pour la plupart cette rubrique, mais il convient de signaler que le sujet est très bien couvert, en *ca.* 80 pages, par l'opuscule de Fox [8]. En complément de ces méthodes de diagnostic, des informations sont enfin données sur la régression robuste. Il faut cependant souligner que la robustesse statistique (et les notions telles que la fonction d'influence, le point de rupture) est introduite d'une façon très qualitative. En effet, les bases mathématiques de la robustesse (comme l'analyse de fonctionnelles) sont d'un niveau bien plus avancé que celles nécessaires à la compréhension du reste du document (voir par exemple l'ouvrage de Lecoutre & Tassi [9]). Pour autant, il existe des présentations fort didactiques, conçues pour les praticiens : par exemple le manuel de Rousseeuw & Leroy [10], et surtout le chapitre 6 et l'annexe 2 de l'ouvrage de Hamilton [11].

· La dernière partie aborde quelques extensions du modèle linéaire. Comme dans la troisième partie, des choix limitatifs ont été opérés : les modèles non linéaires, qui posent des problèmes tout à fait spécifiques, n'y sont pas mentionnés (il faut toutefois observer que la régression robuste est non linéaire). Au demeurant, l'utilisateur intéressé dispose dans ce domaine d'excellents manuels : Gallant [12], Bates & Watts [13], Seber & Wild [14]. Ne figurent pas non plus dans cette sixième partie des techniques plus spécialisées, telles que la régression non paramétrique (*Cf.* par exemple Härdle [15]). En revanche, la relation structurelle y est présentée, surtout à cause des débats récurrents entre biométriciens que suscite son utilisation.

Le second objectif de ce document est d'inciter le lecteur à consulter les ouvrages qui traitent le sujet de façon plus complète, et/ou qui en approfondissent certains aspects. Outre celles qui ont déjà été citées, la littérature est riche d'intéressantes références : Draper & Smith ([16], réédité en 1981) peuvent être considérés comme les auteurs qui ont inauguré la présentation "moderne" du thème. Les ouvrages de Cook & Weisberg [17], et de Weisberg [18], sont focalisés sur le traitement des difficultés rencontrées dans les applications, le second se situant à un niveau de plus grande généralité que le premier. Pour les manuels de langue française, signalons le chapitre 3 du livre de Lebart *et al.* [19], exposé formel et concis de la théorie classique, et aussi l'ouvrage de

Tomassone *et al.* [20], de conception toute différente, articulé autour de l'analyse détaillée et commentée de situations expérimentales réelles.

Références citées :

[1] RATKOWSKY, D.A., 1983, *Nonlinear Regression Modeling. A unified practical approach*, M. Dekker ed., 276 p.

[2] SCHEFFE, H., 1959, *The Analysis of Variance*, J. Wiley & Sons ed., 477 p.

[3] BERGONZINI, J.-Cl., & C. DUBY, 1995, *Analyse et planification des expériences. Les dispositifs en blocs*, Masson éd., Paris, Milan, Barcelone, 353 p.

[4] BELSLEY, D.A., 1991, *Conditioning diagnostics. Collinearity and weak data in regression*, J. Wiley & Sons ed., 396 p.

[5] CHATTERJEE, S., & B. PRICE, 1991, *Regression analysis by example*, 2nd edition, J. Wiley & Sons ed., 278 p.

[6] MONTGOMERY, D.C., & E.A. PECK, 1992, *Introduction to linear regression analysis*, 2nd edition, J. Wiley & Sons ed., 527 p.

[7] CARROLL, R.J., & D. RUPPERT, 1988, *Transformation and weighting in regression*, Chapman & Hall ed., New York, London, 249 p.

[8] FOX, J., 1991, *Regression diagnostics*, Quantitative applications in the social sciences 79, SAGE Univ. papers, Newbury Park, California, 92 p.

[9] LECOUTRE, J.-P., & Ph. TASSI, 1987, *Statistique non paramétrique et robustesse*, éd. Economica, 455 p.

[10] ROUSSEEUW, P.J., & A.M. LEROY, 1987, *Robust regression and outlier detection*, J. Wiley & Sons ed., 329 p.

[11] HAMILTON, L.C., 1992, *Regression with graphics. A second course in applied statistics*, Duxbury Press, Wadsworth Publishing Co., Belmont, California, 363 p.

[12] GALLANT, A.R., 1987, *Nonlinear Statistical Models*, J. Wiley & Sons ed., 610 p.

[13] BATES, D.M., & D.G. WATTS, 1988, *Nonlinear Regression Analysis and its Applications*, J. Wiley & Sons ed., 365 p.

[14] SEBER, G.A.F., & C.J. WILD, 1989, *Nonlinear Regression*, J. Wiley & Sons ed., 768 p.

[15] HÄRDLE, W., 1990, *Applied nonparametric regression*, Cambridge University Press, New York, 333 p.

[16] DRAPER, N.R., & H. SMITH, 1966, *Applied Regression Analysis*, J. Wiley & Sons ed., 407 p.

[17] COOK, D., & S. WEISBERG, 1982, *Residuals and Influence in Regression*, Chapman & Hall ed., 230 p.

[18] WEISBERG, S., 1985, *Applied Linear Regression*, 2nd edition, J. Wiley & Sons ed., 324 p.

[19] LEBART, L., A. MORINEAU, & J.-P. FENELON, 1979, *Traitement des données statistiques. Méthodes et programmes*, Dunod éd., 510 p.

[20] TOMASSONE, R., S. AUDRAIN, E. LESQUOY-de TURCKHEIM, & C. MILLIER, 1992, *La Régression. Nouveaux regards sur une ancienne méthode statistique*, INRA, coll. actualités scientifiques & agronomiques 13, Masson éd., 2nde éd., 188 p.

Introduction

INTRODUCTION.

· ***Ecriture générale :***

Considérons les **variables mathématiques** y et x, et la fonction *f* : $y = f(x)$
Hormis le cas élémentaire y = x, la relation fait habituellement aussi intervenir un ou plusieurs **paramètres** θ_i , *e.g.* :

$$y = \theta x \ , \text{ ou bien encore : } \ y = \theta_1 + \theta_2 x^{\theta_3}$$

On notera enfin que la fonction *f* peut exprimer la relation entre y et plusieurs variables $x_1, x_2, ..., x_m$:

$$y = \theta_1 x_1 + x_2^{\theta_2} \ , \text{ par exemple.}$$

Sous sa forme la plus générale, le modèle qui exprime la dépendance de la variable mathématique y vis-à-vis des x_i peut donc s'écrire :

$$\boxed{y = f(x_1, \ldots, x_m \, ; \ \theta_1, \ldots, \theta_p)}$$

On se limitera au cas où *f* est une fonction réelle d'une ou plusieurs variables réelles. On remarquera aussi que l'on peut employer un formalisme plus condensé en considérant $(x_1, x_2, ..., x_m)$ et $(\theta_1, \theta_2, ..., \theta_p)$ comme des **vecteurs** :

$$\mathbf{x} \in \mathbb{R}^m \ , \ q \in \mathbb{R}^p \xrightarrow{\ f \ } y = f(\mathbf{x} \, ; \, q) \in \mathbb{R}$$

· ***Modèle linéaire :***

Par définition, le modèle $y = f(\mathbf{x} \, ; \, q)$ est dit linéaire si la fonction *f* est **une combinaison linéaire des paramètres** θ_i ; ainsi :

$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_1 x_2 \ , \ \text{ ou bien : } \ y = \theta_1 \sin x_1 + \theta_2 \sin x_2 \, ,$$

sont deux exemples de modèles linéaires (sous-entendu : suivant les paramètres). *A contrario*, un modèle tel que :

$$y = \theta_0 + \theta_1 \exp(\theta_2 x)$$

est non linéaire, car la fonction *f* n'est pas une combinaison linéaire des θ_i.

· ***Modèles non linéaires, mais linéarisables :***

Un exemple bien connu en biologie est celui de la relation d'allométrie, qui vise à décrire la croissance relative de deux organes de dimensions (*e.g.*, de longueurs) respectives x et y. Soient k_x et k_y les taux d'accroissement instantanés relatifs correspondants :

$$\frac{1}{x}\frac{dx}{dt} = k_x \quad , \quad \text{et :} \quad \frac{1}{y}\frac{dy}{dt} = k_y$$

$$\text{En posant :} \quad \beta = k_y / k_x \quad , \quad \text{il vient :} \quad \frac{dy}{y} = \beta \frac{dx}{x}$$

Sous l'hypothèse β = constante, l'intégration conduit à la relation : $y = \alpha x^{\beta}$

Ce modèle est non linéaire en (α, β). Il peut cependant être **linéarisé à l'aide d'une transformation** logarithmique : soient $Y = \ln(y)$, $X = \ln(x)$, et $\gamma = \ln(\alpha)$.

Avec ces nouvelles variables, le modèle devient :

$Y = \gamma + \beta X$, linéaire suivant les paramètres (γ, β).

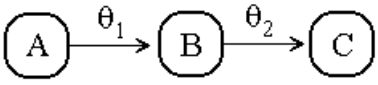
· ***Quelques autres cas simples :***

Il existe de très nombreux exemples de modèles définis par une combinaison non linéaire des paramètres, combinaison qui peut néanmoins être linéarisée. Quelques fonctions classiques sont rassemblées dans le tableau ci-dessous, qui ne donne qu'un très petit aperçu de la variété des situations rencontrées en pratique.

Fonction	Transformation	Modèle linéarisé
$y = \theta_0 \exp(\theta_1 x)$	$Y = \ln(y)$	$Y = \gamma + \theta_1 x$, $\gamma = \ln(\theta_0)$
$y = \theta_0 + \theta_1 \ln(x)$	$X = \ln(x)$	$y = \theta_0 + \theta_1 X$
$y = x/(-\theta_1 + \theta_0 x)$	$X = 1/x$, $Y = 1/y$	$Y = \theta_0 - \theta_1 X$
$y = \dfrac{\exp(\theta_0 + \theta_1 x)}{1 + \exp(\theta_0 + \theta_1 x)}$	$Y = \ln\left(y/(1-y)\right)$	$Y = \theta_0 + \theta_1 x$

· ***Modèles intrinsèquement non linéaires :***

Il s'agit de modèles qu'il est impossible de linéariser à l'aide de transformations. A titre d'illustration, nous emprunterons un exemple à la chimie : un composé A se transforme en un composé B, qui lui-même se transforme en un composé C, les réactions impliquées étant des cinétiques du premier ordre irréversibles. Le schéma du système, ainsi que les 3 équations différentielles ordinaires qui décrivent son évolution au cours du temps, se présentent comme suit :



$$\theta_1 \geq \theta_2 \geq 0$$

$$\frac{dA(t)}{dt} = -\theta_1 A(t)$$

$$\frac{dB(t)}{dt} = \theta_1 A(t) - \theta_2 B(t)$$

$$\frac{dC(t)}{dt} = \theta_2 B(t)$$

Conditions initiales (*i.e.*, à l'instant t = 0) **:** $A(0) = 1$, $B(0) = C(0) = 0$.

Solution :

$$\theta_1 > \theta_2$$

$$A(t) = \exp(-\theta_1 t)$$

$$B(t) = \frac{\theta_1}{\theta_1 - \theta_2}\left(e^{-\theta_2 t} - e^{-\theta_1 t}\right)$$

$$C(t) = 1 - B(t)$$

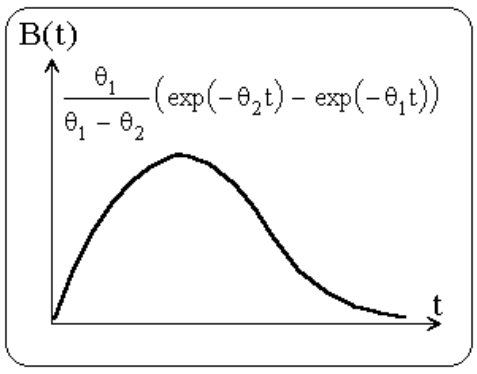
$$\theta_1 = \theta_2 = \theta$$

$$A(t) = \exp(-\theta t)$$

$$B(t) = \theta\, t\, e^{-\theta t}$$

$$C(t) = 1 - e^{-\theta t} - \theta\, t\, e^{-\theta t}$$

Représentation graphique : La variation au cours du temps de la quantité du composé B est représentée par une courbe dont l'allure est indiquée ci-dessous :



C'est là un exemple classique de modèle intrinsèquement non linéaire (*i.e.*, non linéarisable).

On retiendra que la séparation entre les modèles linéaires et non linéaires va bien au-delà des différences formelles qui ont été rappelées ; les seconds soulèvent des problèmes spécifiques difficiles, et ils constituent un champ d'étude à part entière.

· ***Régression linéaire :***

C'est une méthode statistique qui tient une position centrale, aussi bien dans la planification expérimentale que dans l'analyse des résultats concernant du matériel sujet à variabilité (d'où son importance dans les Sciences de la Vie, par exemple). Au plan conceptuel, le modèle régressif se compose d'**une partie** "**fixe**" (*i.e.*, qui exprime l'effet déterministe de la, ou des, variable(s) contrôlée(s) par l'expérimentateur), et d'**une partie aléatoire**, qui restitue la variabilité de l'objet étudié. Selon cette dichotomie, la régression linéaire est définie par deux contraintes fortes :

***** la partie fixe est un modèle linéaire déterministe,
***** la loi de la composante aléatoire est normale (dans le modèle "classique").

Ce que l'on peut résumer, en utilisant les notations déjà introduites :

variable aléatoire modélisée	=	$f(\mathbf{x}\,;\,\theta)$, modèle linéaire déterministe	+	erreur aléatoire

Au plan technique, la méthode soulève divers types de problèmes. Certains sont depuis longtemps résolus, et présentés comme des "classiques" (*e.g.*, ajustement du modèle et identification des paramètres, propriétés statistiques des estimateurs et précision des estimations, inférences), tandis que d'autres suscitent encore de nouveaux développements : ces derniers concernent pour l'essentiel l'étude du comportement du modèle lorsque les hypothèses de base, et spécialement celles formulées pour sa composante aléatoire, ne sont que "partiellement vérifiées". Cela recouvre d'une part la mise au point d'outils diagnostiques des points dits influents (*i.e.*, ceux qui jouent un rôle déterminant dans l'ajustement), et d'autre part la recherche de méthodes d'estimation "peu perturbées" par un écart aux hypothèses sur lesquelles est fondé le modèle (*e.g.*, régression robuste).

On retiendra par ailleurs qu'il existe deux voies pour s'affranchir des deux contraintes qui président à la définition de la régression linéaire : *(i)* la régression non linéaire, dont la partie fixe est un modèle déterministe non linéaire, et *(ii)* les modèles linéaires généralisés, dont la composante aléatoire suit un loi de la famille exponentielle. La façon dont on s'engage dans l'une ou l'autre de ces deux voies est directement conditionnée par la manière dont on cherche à sortir du cadre imposé par le modèle linéaire : il s'ensuit qu'une solide connaissance de ce dernier constitue un préalable indispensable.

Dans ce qui va suivre, et par souci de didactisme, les résultats de base sont donnés pour la régression linéaire simple (une seule variable contrôlée). Hormis quelques points de détail, tous se généralisent immédiatement à la régression linéaire multiple (deux variables contrôlées, ou plus). Cela n'exclut évidemment pas que ce second modèle pose des problèmes spécifiques, le plus connu étant la difficulté engendrée par la non-orthogonalité des régresseurs. La présentation adopte des notations quasi-conventionnelles, que l'on retrouvera à quelques nuances près dans tous les manuels spécialisés. Ainsi, pour se conformer à l'usage le plus répandu, les paramètres du modèle seront désormais notés β.

♦ **Définitions et hypothèses pour le modèle "classique" de régression linéaire :**

$$\underbrace{Y}_{\text{Variable aléatoire.}} = \underbrace{\beta_0 + \beta_1 x}_{\text{Composante déterministe.}} + \underbrace{\varepsilon}_{\text{Variable aléatoire : "Résidu".}}$$

Y : variable réponse, ou "expliquée", ou dépendante, ou endogène.

β_0, β_1 : paramètres ;

x : régresseur, ou variable "explicative", (ou indépendante, ou exogène).

♦ **Caractérisation de la nature des éléments du modèle :**

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ $i = 1, \ldots, n$	OBSERVEES	NON OBSERVABLES
QUANTITES ALEATOIRES	$y_1, y_2, \ldots, y_n$	$\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$
QUANTITES CERTAINES	$x_1, x_2, \ldots, x_n$	β_0, β_1

♦ **Hypothèses classiques (dites de "Gauss-Markov") sur les résidus :**

$$\begin{cases} E[\varepsilon_i] = 0 & \text{Espérance nulle,} \\ Var[\varepsilon_i] = \sigma^2 & \text{Homoscédasticité,} \\ Cov[\varepsilon_i, \varepsilon_j] = 0, \quad i \neq j, & \text{Résidus non corrélés .} \end{cases}$$

Problème : estimer (β_0, β_1).

Chapitre 1

Estimation des paramètres du modèle linéaire simple.

Présentation géométrique ; solution aux moindres carrés.

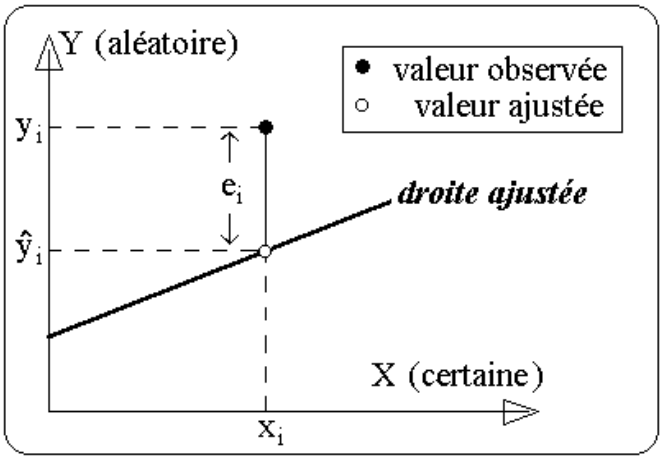
Sommaire du chapitre 1

1.1. IDENTIFICATION DES PARAMETRES.

L'identification des paramètres inconnus β_0 et β_1 procède de deux choix de base :

(i) celui de la mesure de l'écart entre une observation et la valeur prévue par le modèle,
(ii) et le choix du critère d'optimalité.

· ***Définition de l'écart à l'ajustement :***



L'écart e_i exprime, si le modèle est correct, l'incertitude engendrée par le caractère aléatoire de la variable réponse y. C'est donc la distance entre la valeur observée et la valeur calculée correspondante, mesurée parallèlement à l'axe des y.

· ***Critère d'optimalité :***

Les estimations $\hat{\beta}_0$ et $\hat{\beta}_1$ sont les valeurs pour lesquelles est atteint l'extrêmum d'une fonction-objectif (ou critère). Plusieurs critères sont envisageables : on pourrait ainsi rechercher $\hat{\beta}_0$ et $\hat{\beta}_1$ qui réalisent $\min\{\Sigma|e_i|\}$, ou bien encore $\min\{\max(e_i)\}$. Celui qui sera ici retenu est le classique **critère des moindres carrés ordinaires (MCO)** :

$$S(b_0, b_1) = \sum_i e_i^2 = \sum_i \left(y_i - (b_0 + b_1 x_i)\right)^2 \quad , \text{ et : } \quad S(\hat{\beta}_0, \hat{\beta}_1) = \min\left\{S(b_0, b_1)\right\}$$

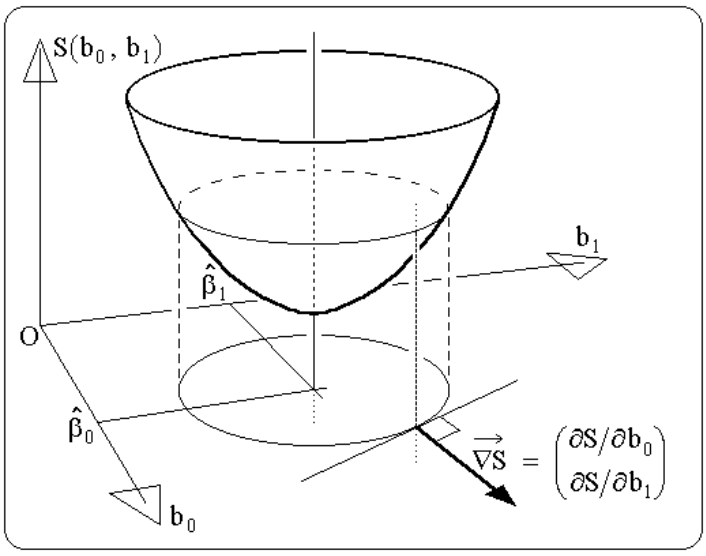
où b_0 et b_1 désignent les valeurs courantes des paramètres du modèle. L'intérêt du critère des MCO réside dans la simplicité des calculs qu'il nécessite, dans les possibilités de représentation géométrique qu'il autorise, et, sous certaines hypothèses, dans sa relation avec le maximum de vraisemblance (défini au deuxième chapitre). En contrepartie, on retiendra la faible robustesse des estimateurs des MCO.

Remarque : le critère des MCO est un cas particulier des **moindres carrés pondérés**, dont il sera question au chapitre 4 :

$$S(b_0, b_1) = \sum_i w_i e_i^2 \text{, où } w_i \text{ est la masse attribuée à l'observation } y_i.$$

· ***Représentation géométrique de la fonction critère :***

Le critère des MCO est, dans le cas du modèle linéaire simple, une fonction de deux variables seulement (b_1 : pente de la droite, b_0 : ordonnée à l'origine), les données expérimentales étant considérées comme des constantes, *i.e.*, fixées à leurs valeurs observées x_i et y_i. Il est par conséquent possible de représenter la fonction critère S par une surface (en l'occurrence, un paraboloïde) dans un espace à 3 dimensions :



Si l'on coupe le paraboloïde par un plan parallèle au plan $\{Ob_0, Ob_1\}$, l'intersection obtenue est une courbe fermée d'équation $S(b_0, b_1) =$ constante, représentative des isovaleurs de S. La figure ci-dessus montre la projection de cette courbe sur le plan $\{Ob_0, Ob_1\}$, parallèlement à l'axe OS : on obtient dans $\{Ob_0, Ob_1\}$ une "courbe de niveau", qui est ici une ellipse. En un point quelconque de cette courbe de niveau, ***le vecteur gradient ∇S indique la direction de la plus forte pente***, *i.e.*, il est orienté dans la direction suivant laquelle le taux d'accroissement de la fonction-critère S est maximal. Au point le plus bas du paraboloïde (le minimum dont on recherche les coordonnées), le vecteur gradient s'annule.

La condition nécessaire que doivent vérifier les estimations $\hat{\beta}_0$ et $\hat{\beta}_1$ s'écrit donc :

$$\left.\vec{\nabla} S\right|_{\hat{\beta}_0,\, \hat{\beta}_1} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

Pour le critère des MCO appliqué au modèle linéaire, la condition est aussi suffisante, ce qui n'est pas le cas en général.

· ***Résolution du système des "équations normales" :***

La recherche du minimum du critère des MCO est un simple problème d'optimisation, qui est résolu en écrivant que les estimations $\hat{\beta}_0$ et $\hat{\beta}_1$ vérifient la condition nécessaire d'extrêmum, *i.e.*, l'annulation du gradient de $S(b_0, b_1)$:

$$\frac{\partial}{\partial b_0}\sum_i e_i^2 = \frac{\partial}{\partial b_1}\sum_i e_i^2 = 0$$

avec :

$$\frac{\partial}{\partial b_0}\sum_i e_i^2 = 2\sum_{i=1}^{n}\left\{e_i \frac{\partial}{\partial b_0}\left[y_i - (b_0 + b_1 x_i)\right]\right\} = -2\sum_i e_i$$

et de même :

$$\frac{\partial}{\partial b_1}\sum_i e_i^2 = -2\sum_i e_i x_i$$

Les estimations $\hat{\beta}_0$ et $\hat{\beta}_1$ sont donc les solutions du système des équations normales :

$$\begin{cases} \sum_{i=1}^{n} e_i = \sum_{i=1}^{n}\left[y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)\right] = 0 & [1] \\ \sum_{i=1}^{n} e_i x_i = \sum_{i=1}^{n} x_i\left[y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)\right] = 0 & [2] \end{cases}$$

$[1] \Rightarrow \sum y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum x_i$, soit encore : $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$: la droite ajustée passe par le barycentre $(\bar{x}, \bar{y})$ du nuage des points expérimentaux. En outre, on remarque que la somme des écarts à l'ajustement e_i est nulle pour un modèle avec terme constant β_0.

$$[1], [2] \Rightarrow \sum x_i y_i - \left(\sum x_i\right)\left(\sum y_i\right)/n - \hat{\beta}_1\left[\sum x_i^2 - \left(\sum x_i\right)^2/n\right] = 0$$

$$\Rightarrow \hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \qquad i = 1, \ldots, n$$

Remarque : avec un critère des moindres carrés pondérés (*"weighted least squares"*, par opposition à *"ordinary LS"*), l'estimateur de la pente devient :

$$\hat{\beta}_1 = \sum_i w_i (x_i - \bar{x}_w)(y_i - \bar{y}_w) \Big/ \sum_i w_i (x_i - \bar{x}_w)^2$$

1.2. REPRESENTATION GEOMETRIQUE DES M.C.O.

$$E[Y_i] = \beta_0 + \beta_1 x_i \qquad y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i \qquad i = 1, \ldots, n$$

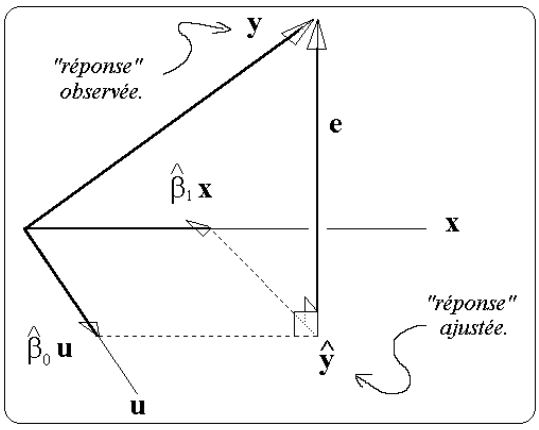
Soient les vecteurs de $\mathbb{R}^n$:

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad \mathbf{u} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \quad \mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \quad \mathbf{e} = \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix}$$

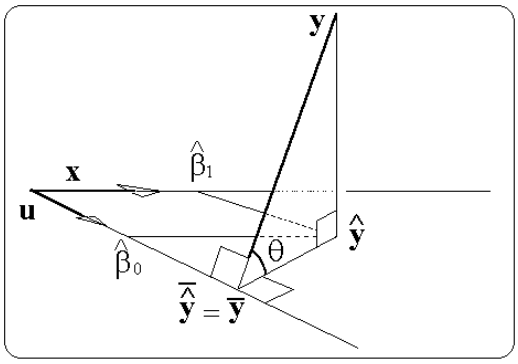
D'où l'écriture du système :

$$\mathbf{y} = \underbrace{\hat{\beta}_0 \mathbf{u} + \hat{\beta}_1 \mathbf{x}}_{\equiv \hat{\mathbf{y}}} + \mathbf{e}$$

et, ci-contre, la représentation géométrique :



· ***Résultats directement obtenus à l'aide de la représentation géométrique des MCO :***



(i) : $\mathbf{e} \perp \mathbf{u} \Rightarrow \mathbf{e}.\mathbf{u} = 0 \Rightarrow \sum_{i=1}^{n} e_i = 0$

C'est-à-dire : $\bar{y} = \bar{\hat{y}}$; par construction, la somme des écarts à l'ajustement est nulle pour un modèle incluant un terme constant β_0.

(ii) : dans le triangle rectangle $\{y, \hat{y}, \bar{y}\}$:

$$\Sigma(y_i - \bar{y})^2 = \Sigma e_i^2 + \Sigma(\hat{y}_i - \bar{\hat{y}})^2$$

En divisant par n, et en remarquant que $\bar{e} = 0$, on obtient la classique ***équation d'analyse de la variance*** :

$$\frac{1}{n}\Sigma(y_i - \bar{y})^2 = \frac{1}{n}\Sigma(\hat{y}_i - \bar{y})^2 + \frac{1}{n}\Sigma(e_i - \bar{e})^2$$

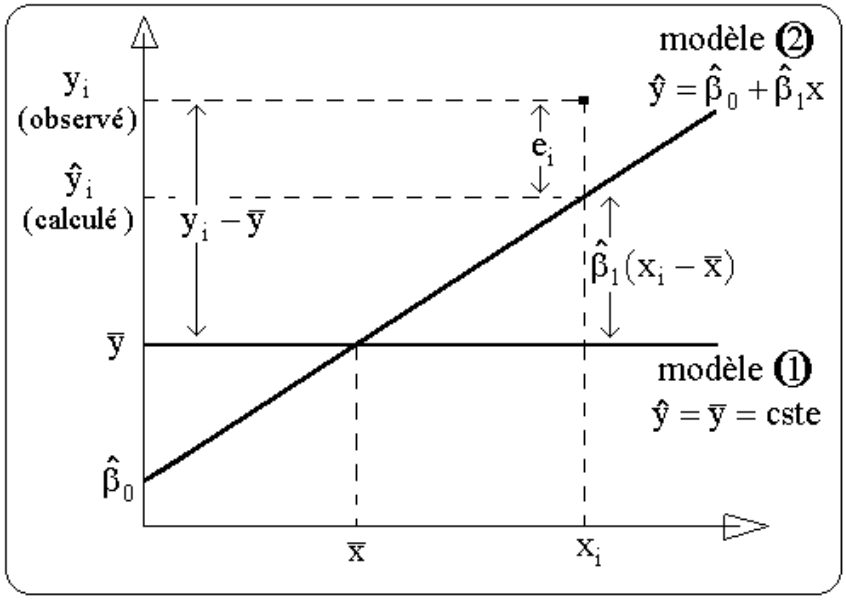
Variance totale ***V. " expliquée"*** ***Var. résiduell e***

(iii) Critère de "qualité de l'ajustement" : le ***coefficient de détermination*** R^2, $R^2 \in [0, 1]$, défini par,

$$R^2 = \cos^2\theta = \frac{\Sigma(\hat{y}_i - \bar{y})^2}{\Sigma(y_i - \bar{y})^2} = \frac{\text{var. expliquée}}{\text{var. totale}}$$

1.3. EQUATION D'ANALYSE DE LA VARIANCE.

· ***Représentation graphique des termes de l'équation:***



· ***Décomposition de la somme des carrés d'une régression :***

$$\sum (y_i - \bar{y})^2 \quad = \quad \sum (\hat{y}_i - \bar{y})^2 \quad + \quad \sum (y_i - \hat{y}_i)^2$$

SC totale (autour de la moyenne $\bar{y}$).	SC du nuage ajusté (due à la régression).	SC résiduelle (non "expliquée" par la régression).

Formulation équivalente : $\sum (y_i - \bar{y})^2 = \hat{\beta}_1^2 \sum (x_i - \bar{x})^2 + \sum e_i^2$

Ce résultat est habituellement représenté par le tableau suivant :

Source	***Somme des Carrés***	***d.d.l.***	***Carrés Moyens***
due à la régression	$\hat{\beta}_1^2 \sum (x_i - \bar{x})^2$	1	*idem*
résiduelle	$\sum e_i^2$	$n - 2$	$\hat{\sigma}^2 = \sum e_i^2 / (n-2)$
totale	$\sum (y_i - \bar{y})^2$	$n - 1$	

L'intérêt du tableau précédent est de permettre de décider si la régression est "significative" ou non, autrement dit ***choisir entre le modèle (2)*** :

$$E[Y] = \beta_0 + \beta_1 x,$$

et le modèle (1) qui exprime l'absence de relation linéaire entre Y et x :

$$E[Y] = \beta_0 = \text{constante}.$$

Ce choix est fondé sur le test des ***hypothèses statistiques*** suivantes:

H_0 : régression linéaire "non significative" [$\Rightarrow$ choisir le modèle (1)],
contre H_1 : régression linéaire "significative" [$\Rightarrow$ choisir le modèle (2)].

La ***statistique du test*** est le rapport : $\dfrac{\text{Carré moyen dû à la régression}}{\text{Carré moyen résiduel}}$

Bien évidemment, l'hypothèse nulle sera repoussée pour les "fortes" valeurs de ce rapport. Pour pouvoir juger qu'une valeur du rapport est "plutôt forte", il est cependant nécessaire de connaître la distribution de l'ensemble des valeurs qu'il peut prendre. C'est-à-dire qu'***il faut connaître la loi de la statistique du test quand*** H_0 ***est vraie. Pour cela, il est nécessaire de préciser la loi des résidus*** (*Cf.* chapitre 2).

Remarque : Le test peut aussi être effectué en utilisant pour statistique le coefficient de détermination R^2, ou bien encore en comparant à 0 l'estimation de β_1. Ces équivalences sont rassemblées au paragraphe 2.13.

Chapitre 2

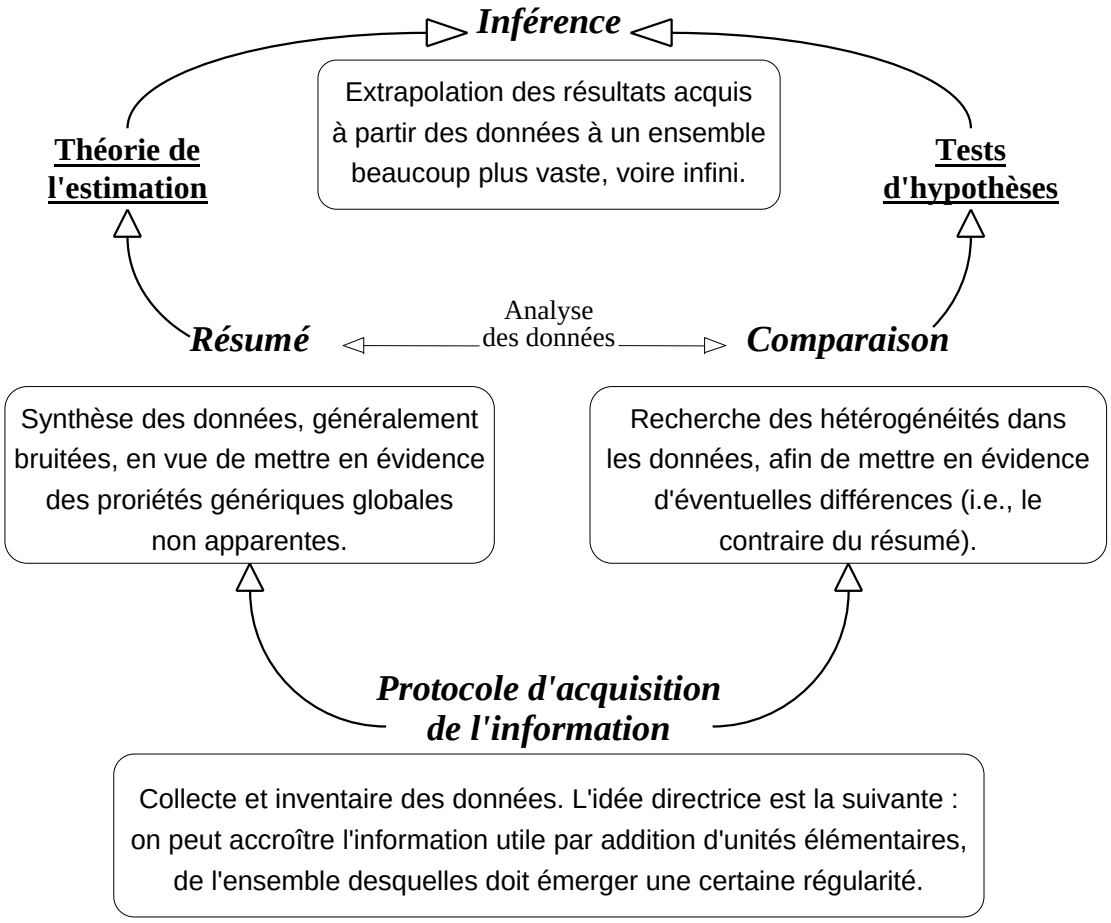
Estimation des paramètres du modèle linéaire simple par le maximum de vraisemblance.

Inférences dans le cadre gaussien.

Sommaire du chapitre 2

2.1. LA PLACE DE L'ESTIMATION DANS LA DEMARCHE STATISTIQUE.

L'objet de la Statistique est l'élaboration et l'application de méthodes optimales pour l'identification de phénomènes réels plus ou moins "masqués par du bruit", méthodes assorties de procédures de vérification du caractère non artéfactuel des résultats obtenus. Les principales étapes de la démarche sont schématiquement présentées ci-dessous :



Le point de départ est l'acquisition planifiée des données, de telle sorte que considérées dans leur ensemble, les observations révèlent un "signal", une "structure". Le protocole de collecte est l'objet de la théorie de l'échantillonnage (si l'on s'engage sur la voie du *résumé*), ou de la théorie des plans d'expérience (si l'on vise la *comparaison*).

Le *résumé* procède par agrégation de l'ensemble des données, en vue d'exprimer un ou plusieurs caractères communs (*e.g.*, une tendance globale). La *comparaison* adopte le point de vue symétrique, et vise à séparer des sous-ensembles d'observations, pour s'intéresser aux différences qui les distinguent. Enfin, l'*inférence* est la généralisation des conclusions, que l'on décidera d'extrapoler à un ensemble beaucoup plus vaste, *i.e.*, à la "population" dont sont issues les observations, ou encore au processus qui les a engendrées.

Dans ce qui va suivre, il sera surtout question de l'*estimation*, théorie qui traite la construction du résumé, ainsi que les inférences conduites à partir de celui ci.

2.2. LE MODELE D'ECHANTILLONNAGE.

· Supposons que l'on étudie une population finie (*e.g.*, les pieds de vigne d'une parcelle, les poissons d'un étang, ...), nommée *population parente*, ou *population-cible*. Elle est formée de *N* entités distinctes et identifiables, appelées *individus*. On s'intéresse à un caractère précis de chaque individu, caractère qui peut être qualitatif [*e.g.*, le cépage du pied considéré (syrah, mourvèdre,...)], quantitatif discret (*e.g.*, le nombre de grappes par pied), ou encore quantitatif continu (*e.g.*, le poids du poisson). Dans tous les cas, on peut se ramener à une *population de caractéristiques*, qui est une collection de valeurs numériques, toutes bien déterminées pour chacun des *N* individus. La population n'est en pratique jamais étudiée en totalité, mais seulement à partir d'un *sondage* : un sous-ensemble (un *échantillon*) de n individus ($n << N$) est extrait de la population. Plus exactement, on va créer une variable aléatoire en définissant un protocole d'*échantillonnage aléatoire*, qui attribue à chaque individu une probabilité connue d'être extrait de la population. C'est alors l'acte d'échantillonnage aléatoire qui introduit le hasard : par exemple, bien que précisément définie pour chaque poisson, la valeur du caractère "poids" est *a priori* incertaine pour l'un quelconque de ceux qui vont constituer l'échantillon observé : le poids individuel est une variable aléatoire.

· Dans de nombreuses situations, il est toutefois impossible de faire référence à un ensemble d'individus concrètement identifiables, et seule la variable aléatoire possède une signification. Ainsi, si l'on entreprend une étude statistique de l'intensité du vent en une station météorologique donnée, la notion d'individu n'est pas pertinente. On va s'intéresser directement à n réalisations de la variable aléatoire "intensité", observations issues d'une infinité de réalisations possibles (*i.e.*, d'une population abstraite infinie). On dira que l'on échantillonne un *processus aléatoire*. Ce processus sera souvent décrit par un *modèle paramétrique*, selon lequel la loi de la variable échantillonnée appartient à une famille de lois caractérisée par un ou plusieurs paramètres θ. Par exemple, pour décrire les variations du débit d'un fleuve, on pourra décider d'utiliser la loi lognormale de paramètres $q = (\mu, \sigma^2)$.

· Considérons donc un phénomène décrit par un modèle probabiliste. Pour simplifier, on se limitera à un modèle ne contenant qu'une seule variable aléatoire, notée Y ; cette variable aléatoire suit une loi inconnue, notée F_θ. Cette dernière notation souligne que la loi inconnue de Y dépend d'un ou plusieurs paramètres réels θ, *i.e.*, que la loi de Y appartient à une famille paramétrique, et donc que l'on postule une certaine forme mathématique pour sa fonction de répartition F_θ.

Le *modèle d'échantillonnage paramétrique* procède de l'observation répétée de Y, en vue d'affiner la connaissance de F_θ ; formellement :

• on recueille n observations indépendantes $(y_1, \ldots, y_n)$: c'est une réalisation du n-uple de variables aléatoires *indépendantes* $(Y_1, \ldots, Y_n)$, variables qui possèdent chacune *la même loi* F_θ ; ce que l'on note :

$$Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} F_\theta$$

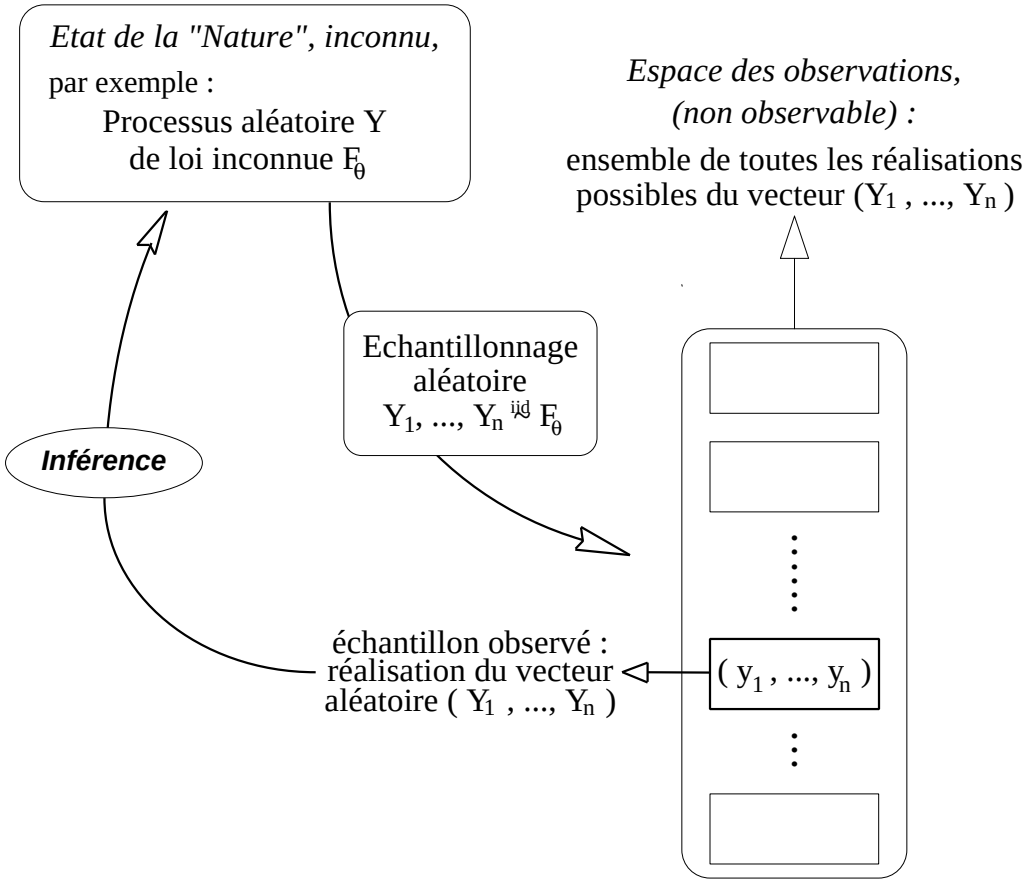
• le vecteur aléatoire $(Y_1, \ldots, Y_n)$ est appelé *échantillon* d'une population précisément définie (concrètement, ou bien au plan théorique), population dont Y est la *variable parente* (F_θ est parfois appelé "processus probabiliste parental").

• l'ensemble des observations $(y_1, \ldots, y_n)$ est interprété comme la réalisation de l'échantillon aléatoire $(Y_1, \ldots, Y_n)$; on notera que dans le langage courant, l'appellation "échantillon" confond en général l'*échantillon aléatoire* $(Y_1, \ldots, Y_n)$ avec sa réalisation, *i.e.*, avec *l'échantillon observé* $(y_1, \ldots, y_n)$.

· *N.B.* : le modèle d'échantillonnage n'est pas nécessairement paramétrique. Par exemple, on peut simplement spécifier que Y suit une loi continue, sans en préciser la forme, et adopter un modèle d'échantillonnage *non paramétrique* : $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} F$.

2.3. VARIABLE PARENTE - ECHANTILLON - INFERENCE.

Le schéma ci-dessous éclaire les relations entre les notions qui viennent d'être présentées. Il est essentiel de comprendre que **l'acte d'échantillonnage aléatoire, en même temps qu'il fournit les** n **observations** $(y_1, \ldots, y_n)$, **engendre un ensemble théorique, formé de toutes les réalisations possibles du vecteur aléatoire** $(Y_1, \ldots, Y_n)$. Cet ensemble est appelé *espace des observations*. Il rassemble tous les résultats expérimentaux que l'on aurait pu obtenir à la place de celui qui a été effectivement observé.



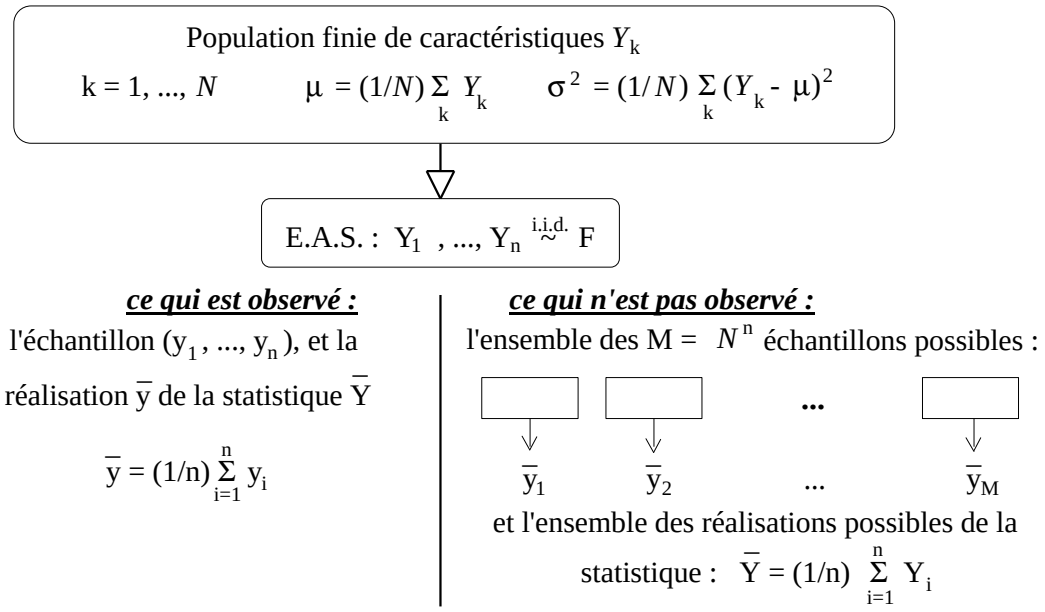
Si l'on connaissait l'*état de la Nature*, et en particulier la loi de Y, on pourrait (pour un protocole d'échantillonnage aléatoire donné) déterminer l'ensemble de tous les jeux de résultats $(y_1, \ldots, y_n)$ possibles, *i.e.*, l'espace des observations.

En pratique, on se trouve dans la situation opposée, celle où F_θ n'est pas connue : c'est précisement ce processus qui fait l'objet de l'étude. L'inférence statistique procède alors dans le sens suivant : à partir de l'observation d'un échantillon, le statisticien va tenter de reconstruire le modèle probabiliste le mieux à même d'avoir engendré le résultat expérimental obtenu ; il va aussi évaluer les incertitudes attachées au choix qu'il aura arrêté.

L'inférence statistique est *un problème de décision en environnement incertain* : *(i)* la Nature "choisit" la loi de Y dans un espace de lois de probabilité. *(ii)* Le statisticien, qui ignore le choix de la Nature, prend une décison ("fait un pari" sur la loi de Y) dans un ensemble de décisions possibles. *(iii)* De la différence entre ces deux choix (celui de la Nature et celui du statisticien) résulte une conséquence, qu'il faudra quantifier à l'aide d'un critère numérique (ce sera une *fonction d'utilité*, ou bien une *fonction de perte*).

2.4. QU'EST-CE QU'UNE STATISTIQUE ?

· Pour aborder la notion de statistique, le plus simple est de considérer une population finie de N individus. A chacun est associé un caractère quantitatif, noté Y, que l'on souhaite étudier. Dans l'ensemble de la population, les valeurs du caractère se distribuent autour d'une valeur centrale $\mu = (1/N)\sum Y_k$, k = 1, ..., N, avec une dispersion $\sigma^2 = (1/N)\sum(Y_k - \mu)^2$. Les grandeurs descriptives μ et σ^2 sont des quantités *certaines*, inconnues, que l'on souhaite évaluer ; on se limitera ici au cas de μ. A cette fin, on peut recourir à l'*échantillonnage aléatoire simple* (EAS), suivant lequel tout individu de la population se voit attribuer la même probabilité $1/N$ d'appartenir à l'échantillon de taille n. Pour garantir l'indépendance, les individus sont tirés avec remise. On observe alors $y_1, ..., y_n$:



· Dans cet exemple, on a défini une fonction $T : \mathbb{R}^n \rightarrow \mathbb{R}$, telle que $T(Y_1, \ldots, Y_n)$ soit une variable aléatoire. Une telle variable, construite sur l'échantillon, est appelée une *statistique*. On a considéré ici la *moyenne empirique* $\bar{Y} = (1/n)\sum Y_i$. Ayant observé $(y_1, \ldots, y_n)$, on calcule la fonction T sur les données expérimentales : $T(y_1, \ldots, y_n) = \bar{y} = (1/n)\sum y_i$. Si l'on pouvait accéder à toutes les réalisations possibles de $(Y_1, \ldots, Y_n)$, et calculer pour chacune la valeur correspondante de T, on déterminerait ainsi la loi appelée *distribution d'échantillonnage* de la statistique $\bar{Y} = T(Y_1, \ldots, Y_n)$.

· La question centrale est la suivante : ***que nous apprend*** $\bar{y}$ ***sur la valeur inconnue*** μ ***?*** Cela conduit à s'interroger sur l'*exactitude* et sur la *précision* avec lesquelles la quantité μ est approchée par $\bar{y}$. Ces deux qualités font référence à la distribution d'échantillonnage de $\bar{Y}$. On parlera d'exactitude si les réalisations possibles de la variable aléatoire $\bar{Y}$ sont en moyenne centrées sur μ, *i.e.*, si $E[\bar{Y}] = \mu$. Si leur dispersion autour de $E[\bar{Y}]$ est faible, *i.e.*, si $Var[\bar{Y}]$ est petite, on parlera alors de précision.

A cet égard, le cas de la moyenne empirique est particulièrement simple ; tout d'abord, le théorème de la limite centrale garantit que dans une grande variété de situations, la distribution d'échantillonnage de la statistique $\overline{Y}$ tend vers la normalité quand la taille n de l'échantillon croît. En outre : $E[\overline{Y}] = \mu$ (*absence de biais*), et $Var[\overline{Y}] = \sigma^2/n$.

2.5. QU'EST-CE QU'UN ESTIMATEUR ?

· Dans le problème présenté au paragraphe précédent, on a considéré l'évaluation de la vraie moyenne μ (inconnue) d'une population finie sondée par EAS. Dans ce but a été construite la statistique moyenne empirique $\overline{Y} = (1/n)\sum Y_i$: c'est un exemple d'*estimateur*. Il convient d'insister sur le fait qu'*un estimateur est une variable aléatoire*. C'est une statistique, dont la valeur calculée sur l'échantillon observé est appelée *estimation* : ici, $\bar{y} = (1/n)\sum y_i$, qui est une réalisation de $\overline{Y}$, constitue une estimation de μ.

· Ces définitions, introduites à partir d'un exemple particulier, s'étendent à un contexte beaucoup plus général : population infinie ou processus aléatoire, échantillonnage autre que l'EAS (*e.g.*, stratifié, par grappes, ...). Dans ce qui va suivre, on se situera dans le cadre paramétrique, et l'on s'intéressera directement à l'étude d'un processus décrit par une variable aléatoire Y, de loi F_θ. Pour simplifier, on va supposer $\theta \in \mathbb{R}$, et adopter le modèle d'échantillonnage $Y_1, \ldots, Y_n \overset{i.i.d.}{\sim} F_\theta$.

· Notons $\hat{\theta}_n$ la fonction du vecteur aléatoire $(Y_1, \ldots, Y_n)$, estimateur de la vraie valeur inconnue (notée $\theta^\bullet$) du paramètre θ. On distingue l'*estimation ponctuelle*, qui approche $\theta^\bullet$ par la valeur prise par l'estimateur $\hat{\theta}_n$, de l'*estimation par intervalle*, qui définit un domaine ayant une probabilité fixée *a priori* de contenir $\theta^\bullet$; pour un paramètre réel, on construira deux estimateurs $\hat{\theta}_n^L$ et $\hat{\theta}_n^U$ tels que :

$$\text{Proba}\{\ \hat{\theta}_n^L \le \theta^\bullet \le \hat{\theta}_n^U\ \} = 1 - \alpha \qquad \alpha \in\ ]\,0,\,1[$$

$[\hat{\theta}_n^L\ ,\ \hat{\theta}_n^U\ ]$ est appelé *intervalle de confiance* au niveau $1 - \alpha$ pour le paramètre θ. Cet intervalle est une grandeur aléatoire (alors que $\theta^\bullet$ est une quantité inconnue, mais certaine). La probabilité α représente le *risque* que l'intervalle ne renferme pas la vraie valeur $\theta^\bullet$.

· Il a été souligné précédemment que l'inférence statistique est un problème de décision en environnement incertain ; il en est de même de l'estimation. Par souci de simplicité, limitons nous à l'estimation d'un paramètre réel θ ($\theta \in \mathbb{R}$). C'est un problème décisionnel caractérisé par :

- un *espace des états* de la Nature, noté Θ, et formé de l'ensemble des éléments extérieurs à l'agent décideur (*i.e.*, au statisticien), éléments qui spécifient complètement l'environnement de ce dernier. Dans le cas présent, $\Theta = \mathbb{R}$. Dans l'espace des états, la Nature "choisit" un point $\theta^\bullet$.
- Un *espace* D *des décisions possibles* pour le statisticien qui, ne connaissant pas $\theta^\bullet$, va prendre une décision d dans D. En l'occurrence, il s'agira de l'estimation ponctuelle représentée par la valeur réelle $\hat{\theta}$ (D est ici confondu avec $\mathbb{R}$; notons que la décision d pourrait aussi être un intervalle de confiance).

- Un *ensemble* C *de conséquences*, résultant d'une décision d face à un état de la Nature, conséquences qu'il convient de hiérarchiser. A cette fin, à toute décision d'estimer $\theta^\bullet$ par d, et à toute valeur exogène θ (*i.e.*, appartenant à l'espace des états), est associée une *fonction de perte* définie sur C, et notée $\mathcal{L}$. C'est une application $\Theta \times D \rightarrow \mathbb{R}^+$, qui souvent en statistique est de la forme $\mathcal{L}(d, \theta^\bullet) = (d-\theta^\bullet)^2$ (perte quadratique), ou encore $\mathcal{L}(d, \theta^\bullet) = |d-\theta^\bullet|$.

La décision "rationnelle" visera à minimiser la perte, ou l'espérance de la perte. Le préalable nécessaire est la connaissance des propriétés statistiques de l'estimateur.

2.6. CARACTERISATION D'UN ESTIMATEUR.

L'estimateur étant défini comme une variable aléatoire, ses qualités (exactitude, précision) sont exprimées par les propriétés de sa distribution de probabilité. On s'intéressera en particulier à l'espérance et à la variance de cette loi, dite *distribution d'échantillonnage de l'estimateur*. En outre, on cherchera à caractériser son *comportement asymptotique*, *i.e.*, les propriétés qu'acquiert l'estimateur lorsque $n \rightarrow +\infty$.

· ***Biais d'un estimateur*** **:** notons $\hat{\theta}_n$ l'estimateur du paramètre $\theta \in \mathbb{R}$, dont la vraie valeur inconnue est $\theta^\bullet$, et soit $B(n, \theta^\bullet)$ le *biais* de $\hat{\theta}_n$; par définition :

$$B(n, \theta^\bullet) = E[\hat{\theta}_n] - \theta^\bullet \qquad \text{si } B(n, \theta^\bullet) = 0, \text{ l'estimateur } \hat{\theta}_n \text{ est dit } \textit{non biaisé}.$$

Par ailleurs, $\hat{\theta}_n$ est *asymptotiquement non biaisé* si $\lim_{n \rightarrow +\infty} B(n, \theta^\bullet) = 0$.

Par exemple, considérons $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} F$, F de variance finie σ^2. On peut estimer σ^2 à l'aide de l'estimateur *variance empirique* $\hat{\theta}_n = (1/n)\sum(Y_i - \overline{Y})^2$, qui est asymptotiquement sans biais. On montre en effet que dans ce cas : $B(n, \sigma^2) = -\sigma^2/n$.

· ***Comparaison d'estimateurs sans biais - Efficacité.***

- Relation de préférence : soient $\hat{\theta}_n^{(1)}$ et $\hat{\theta}_n^{(2)}$ deux estimateurs *non biaisés* de $\theta^\bullet$; on dira que $\hat{\theta}_n^{(1)}$ est *préférable* à $\hat{\theta}_n^{(2)}$ au sens de la variance si $\text{Var}[\hat{\theta}_n^{(1)}] \leq \text{Var}[\hat{\theta}_n^{(2)}]$. Par exemple, si $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} P(\theta)$, loi de Poisson de paramètre $\theta > 0$, on sait que $\theta = E[Y] = \text{Var}[Y]$. On peut donc estimer θ soit par $\hat{\theta}_n^{(1)} = \overline{Y}$, soit par $\hat{\theta}_n^{(2)} = (1/(n-1))\sum(Y_i - \overline{Y})^2$; on montre facilement que $\text{Var}[\hat{\theta}_n^{(1)}]/\text{Var}[\hat{\theta}_n^{(2)}] = (n-1)^2/(n^2(1+2\theta)) < 1$, et donc que la moyenne empirique est ici préférable à la variance empirique modifiée (*i.e.*, corrigée de son biais).

- Estimateur optimal : notons $\mathcal{B}_0$ *l'ensemble des estimateurs sans biais* de $\theta^\bullet$; on peut se demander s'il existe dans $\mathcal{B}_0$ un estimateur $\hat{\theta}_n^{(0)}$ préférable à tous les autres. Si tel est le cas, *i.e.*, si : $\text{Var}[\hat{\theta}_n^{(0)}] \leq \text{Var}[\hat{\theta}_n]$, $\forall\ \hat{\theta}_n \in \mathcal{B}_0$, l'estimateur $\hat{\theta}_n^{(0)}$ est dit *optimal*.

- Borne de Fréchet : la question qui se pose est alors celle de l'existence d'une limite inférieure à la variance d'un estimateur sans biais. La réponse est apportée par un résultat remarquable, l'inégalité de Fréchet, Darmois, Cramer & Rao ("*inégalité FDCR*"), qui énonce que l'incertitude attachée à l'estimation de θ ne peut pas descendre en dessous d'un seuil déterminé par l'information dont on dispose :

si $\hat{\theta}_n \in \mathcal{B}_0$, alors $\mathrm{Var}[\hat{\theta}_n] \geq \dfrac{1}{n.I_Y(\theta)}$ $\qquad \left[n.I_Y(\theta)\right]^{-1}$: *borne de Fréchet*

La quantité $n.I_Y(\theta)$, définie au paragraphe suivant, est l'information de Fisher qu'apporte l'échantillon aléatoire $(Y_1, \ldots, Y_n)$ sur le paramètre θ. Lorsque $\mathrm{Var}[\hat{\theta}_n]$ est égale à la borne de Fréchet, l'estimateur $\hat{\theta}_n$ est dit *efficace* (il est *asymptotiquement efficace* si cette borne est la limite de $\mathrm{Var}[\hat{\theta}_n]$ quand $n \to +\infty$).

· ***Remarques*** **:**

- un estimateur efficace est optimal, mais il n'y a aucune raison pour qu'un estimateur optimal soit efficace.

- Tout ce qui précède s'applique aussi à l'estimation d'une fonction dérivable $\varphi(\theta)$ du paramètre θ. La borne de Fréchet vaut alors $\left[\varphi'(\theta)\right]^2 / \left[n.I_Y(\theta)\right]$, avec $\varphi'(\theta) = d\varphi(\theta)/d\theta$.

2.7. L'INFORMATION AU SENS DE R. FISHER.

On considère les n variables aléatoires : $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} F_\theta$
et l'on note $\mathbf{Y}_n$ l'échantillon aléatoire $(Y_1, \ldots, Y_n)$; par souci de simplicité, on va aussi supposer que F_θ est une loi continue, qui ne dépend que d'un seul paramètre réel inconnu θ. La question abordée ici est la suivante : *comment mesurer l'information qu'apporte l'échantillon* $\mathbf{Y}_n$ *sur le paramètre inconnu* θ *?*

· ***Information de Fisher du vecteur aléatoire*** $\mathbf{Y}_n$ ***sur le paramètre*** θ

Soit $f(\mathbf{Y}_n; \theta)$ la fonction de densité de $\mathbf{Y}_n$; $f(\mathbf{Y}_n; \theta) > 0$; $\theta \in \mathbb{R}$
Moyennant des hypothèses peu restrictives sur la dérivabilité de f par rapport à θ, on définit le *score* sur θ de l'échantillon $\mathbf{Y}_n$:

$$s(\mathbf{Y}_n, \theta) = \frac{\partial}{\partial\theta} \ln\left[f(\mathbf{Y}_n; \theta)\right]$$

Le score est une variable aléatoire d'espérance nulle : $E[s(\mathbf{Y}_n, \theta)] = 0$. Pour une réalisation donnée $\mathbf{y}_n$ de $\mathbf{Y}_n$, le score peut s'interpréter comme la "sensibilité" de ln(f) à une variation infinitésimale de θ. La *quantité d'information* qu'apporte sur θ l'échantillon $\mathbf{Y}_n$ est définie par :

$$I_{\mathbf{Y}_n}(\theta) = E\left[s^2(\mathbf{Y}_n, \theta)\right] \quad ; \quad I_{\mathbf{Y}_n}(\theta) \geq 0$$

Il n'est guère facile d'interpréter intuitivement cette quantité (non aléatoire), qui n'est autre que *la variance du score*. On peut néanmoins remarquer que si l'on dispose de plusieurs réalisations de $\mathbf{Y}_n$, l'information est d'autant plus élevée que les valeurs correspondantes de la "sensibilité" de ln(f) à θ sont hétérogènes, et donc que les rélisations du score sont d'autant moins redondantes entre elles.

L'information se calcule à l'aide de l'équation : $I_{\mathbf{Y}_n}(\theta) = -E\left\{ \frac{\partial^2}{\partial\theta^2} \ln[f(\mathbf{Y}_n ; \theta)] \right\}$

Les propriétés de I(θ) sont conformes à ce que l'on attend d'une mesure de l'information :

- $I_{\mathbf{Y}_n}(\theta) = 0 \Leftrightarrow$ la loi du vecteur aléatoire $\mathbf{Y}_n$ ne dépend pas de θ
- Soient Y et Z deux variables aléatoires *indépendantes* de lois respectives F_θ et G_θ, dépendant chacune du même paramètre réel θ, et le vecteur aléatoire $\mathbf{w} = (Y, Z)$; alors :

$$I_{\mathbf{w}}(\theta) = I_Y(\theta) + I_Z(\theta)$$

En particulier, si $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} F_\theta$, et $\mathbf{Y}_n = (Y_1, \ldots, Y_n)$,

alors : $I_{\mathbf{Y}_n}(\theta) = n.I_Y(\theta)$

e.g., si les Y_i sont de densité exponentielle, $f(y) = \theta.e^{-\theta y}$, $\theta > 0$, on a : $I_{\mathbf{Y}_n}(\theta) = n/\theta^2$

· ***Présentation qualitative de la notion d'exhaustivité*** **:**

Considérons l'échantillon aléatoire $\mathbf{Y}_n$, et une statistique $T(\mathbf{Y}_n)$. La question est alors la suivante : *la réduction des données opérée par la statistique* T *fait elle perdre de l'information* ? Un résultat général établit que l'information ne peut que diminuer ou rester constante quand on passe de l'échantillon $\mathbf{Y}_n$ à une statistique $T(\mathbf{Y}_n)$, ce que l'on exprime par $I_T(\theta) \leq I_{\mathbf{Y}_n}(\theta)$, où $I_T(\theta)$ est l'information contenue sur θ dans la distribution d'échantillonnage de $T(\mathbf{Y}_n)$. Si la quantité d'information demeure constante, T est une *statistique exhaustive* :

$$I_T(\theta) = I_{\mathbf{Y}_n}(\theta) \Leftrightarrow T \text{ exhaustive.}$$

2.8. LE CHOIX D'UN ESTIMATEUR.

Pour choisir un estimateur, on considèrera ses propriétés dites "à distance finie" (*i.e.*, celles qu'il possède pour une taille d'échantillon n finie), et aussi ses propriétés asymptotiques (celles qui apparaissent lorsque $n \to +\infty$) ; pour les secondes, on tiendra également compte du taux de convergence, *i.e.*, de la "vitesse" avec laquelle les propriétés asymptotiques sont approchées lorsque n croît. En général, les critères de choix sont l'espérance (biais *vs.* non biais) et la variance de la loi de l'estimateur, et aussi cette loi elle-même.

· ***Erreur quadratique moyenne :***

Au paragraphe 2.6 a été présentée la comparaison entre estimateurs sans biais. Dans la pratique, on est aussi amené à comparer entre eux des estimateurs biaisés, ou encore des estimateurs biaisés à des estimateurs non biaisés. Le critère est alors l'*erreur quadratique moyenne* EQM (en anglais, *mean square error* ou MSE), grandeur qui intègre à la fois le carré du biais et la variance :

$$EQM[\hat{\theta}_n] = Var[\hat{\theta}_n] + [B(n, \theta^\bullet)]^2$$

L'introduction de ce critère nous permet de relativiser la notion d'estimateur efficace (*i.e.*, sans biais, et de variance égale à la borne de Fréchet). Supposons qu'il existe un tel estimateur $\hat{\theta}_n^{(e)}$ de $\theta^\bullet$; il n'est en aucun cas exclu qu'existe un autre estimateur $\hat{\theta}_n^{(b)}$ de $\theta^\bullet$, biaisé, mais de plus petite EQM : $\hat{\theta}_n^{(b)}$ serait alors préférable à $\hat{\theta}_n^{(e)}$.

· ***Propriétés asymptotiques*.**

- **Estimateur convergent :** $\hat{\theta}_n$ est dit convergent s'il *converge en probabilité* vers $\theta^\bullet$:

$$\forall\, \varepsilon \in \mathbb{R}^+ , \quad \lim_{n \to +\infty} \text{Proba } \{\ |\hat{\theta}_n - \theta^\bullet| \leq \varepsilon\ \} = 1$$

Selon cette propriété, quand la taille de l'échantillon devient infinie, alors $E[\hat{\theta}_n] \to \theta^\bullet$, et $\text{Var}[\hat{\theta}_n] \to 0$. On notera que si un estimateur est efficace, alors il est aussi convergent. La moyenne et la variance empiriques sont des exemples d'estimateurs convergents.

- **Estimateur asymptotiquement gaussien :** considérons la suite d'estimateurs $\{\hat{\theta}_n\}_{n \geq 1}$ d'un même paramètre θ, de valeur inconnue $\theta^\bullet$. C'est une suite de variables aléatoires ; elle est dite *convergente asymptotiquement normale* si la loi de $\hat{\theta}_n$ converge vers la loi normale. Désignons par Φ la fonction de répartition de la loi normale centrée réduite N(0, 1), alors :

$$\lim_{n \to +\infty} \text{Proba } \{\ \sqrt{n}(\hat{\theta}_n - \theta^\bullet)\big/\sqrt{V_{AS}}\ \leq u\ \} = \Phi(u)$$

$$\text{ce que l'on note : } \sqrt{n}\,(\,\hat{\theta}_n - \theta^\bullet\,) \xrightarrow{\text{loi}} N(0, V_{AS}),$$

où V_{AS} est la variance asymptotique de la suite $\{\hat{\theta}_n\}_{n \geq 1}$. Si $\hat{\theta}_n$ est de plus asymptotiquement efficace, sa variance tend alors vers la borne de Fréchet (*Cf.* § 2.6).

· ***Méthodes d'estimation ponctuelle :***

La méthode des moindres carrés ordinaires (MCO) a déjà été présentée. En effet, c'est dans le contexte particulier du modèle linéaire, quand les hypothèses de Gauss-Markov sont vérifiées (*Cf.* Introduction), que les estimateurs MCO possèdent des propriétés d'optimalité, y compris à distance finie (paragraphe suivant).

Il existe plusieurs autres méthodes d'estimation. Parmi les plus classiques, citons la méthode des moments (dont le principe consiste à identifier les moments empiriques aux moments théoriques), qui fournit des estimateurs pourvus d'intéressantes propriétés asymptotiques (convergence, normalité asymptotique). Dans la suite, on retiendra la méthode du maximum de vraisemblance (présentée aux paragraphes 2.10 et 2.11), l'une des plus utilisées.

2.9. MODELE LINEAIRE SIMPLE : PROPRIETES STATISTIQUES DES ESTIMATEURS DES MOINDRES CARRES ORDINAIRES.

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad , \quad E(\varepsilon_i) = 0 \ , \ Cov(\varepsilon_i, \varepsilon_j) = \sigma^2 \delta_{ij} \ , \ \delta_{ij} = \begin{cases} 1 \text{ si } i = j \\ 0 \text{ si } i \neq j \end{cases}$$

· ***Les estimateurs*** $\hat{\beta}_0$ ***et*** $\hat{\beta}_1$ ***sont sans biais :***

On remplace, dans l'expression de $\hat{\beta}_1$, la différence $Y_i - \overline{Y}$ par $\beta_1(x_i - \overline{x}) + \varepsilon_i$:

$$\hat{\beta}_1 = \beta_1 + \frac{\sum(x_i - \overline{x})\varepsilon_i}{\sum(x_i - \overline{x})^2} \quad \Rightarrow \quad E(\hat{\beta}_1) = \beta_1$$

Par ailleurs :

$$\left.\begin{array}{l} \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{x} \\ \beta_0 = E(Y) - \beta_1 \overline{x} \end{array}\right\} \Rightarrow \hat{\beta}_0 = \beta_0 + \overline{Y} - E(Y) + (\beta_1 - \hat{\beta}_1)\overline{x} \Rightarrow E(\hat{\beta}_0) = \beta_0$$

· ***Variances et covariances :***

- Variance de la pente :

$$\left.\begin{array}{l} Var(\hat{\beta}_1) = E(\hat{\beta}_1 - \beta_1)^2 \\ \hat{\beta}_1 - \beta_1 = \dfrac{\sum(x_i - \overline{x})\varepsilon_i}{\sum(x_i - \overline{x})^2} \end{array}\right\} \Rightarrow \quad Var(\hat{\beta}_1) = \frac{\sum\limits_{i,j}\sum (x_i - \overline{x})(x_j - \overline{x})E(\varepsilon_i \varepsilon_j)}{\left(\sum(x_i - \overline{x})^2\right)^2} = \frac{\sigma^2}{\sum(x_i - \overline{x})^2}$$

- Covariance de la pente avec $\overline{Y}$: Soient $\gamma = 1/n$ et $\lambda_i = (x_i - \overline{x})/\sum(x_i - \overline{x})^2$

$$\left.\begin{array}{l} \overline{Y} = \gamma Y_1 + \ \ldots \ + \gamma Y_n \\ \hat{\beta}_1 = \lambda_1 Y_1 + \ \ldots \ + \lambda_n Y_n \\ Cov(Y_i, Y_j) = \sigma^2 \delta_{ij} \end{array}\right\} \Rightarrow Cov(\overline{Y}, \hat{\beta}_1) = (\gamma\lambda_1 + \ \ldots \ + \gamma\lambda_n)\sigma^2$$

$$= \frac{\sigma^2}{n\sum(x_i - \overline{x})^2} \sum(x_i - \overline{x}) = 0$$

- Variance de l'ordonnée à l'origine :

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{x} \quad \Rightarrow \quad Var(\hat{\beta}_0) = Var(\overline{Y}) + \overline{x}^2 Var(\hat{\beta}_1) - 2\overline{x} Cov(\overline{Y}, \hat{\beta}_1)$$

$$\Rightarrow \quad Var(\hat{\beta}_0) = \sigma^2 \left[(1/n) + \overline{x}^2 / \sum(x_i - \overline{x})^2 \right]$$

- Covariance entre les estimateurs des paramètres :

$$Cov(\hat{\beta}_0, \hat{\beta}_1) = Cov\left((\overline{Y} - \hat{\beta}_1 \overline{x}), \hat{\beta}_1 \right) = - \overline{x} \, Var(\hat{\beta}_1) = - \frac{\overline{x}\sigma^2}{\sum(x_i - \overline{x})^2}$$

Les résultats qui précèdent nécessitent simplement le respect des hypothèses de Gauss-Markov (*Cf.* Introduction ; on rappelle qu'elles ne spécifient aucunement quelle est la loi des résidus). Dans ce cadre, les estimateurs MCO sont *optimaux* au sens suivant : ils sont sans biais, et de variance plus faible que tout autre estimateur qui serait à la fois non biaisé et fonction linéaire des variables aléatoires Y_i.

2.10. LE MAXIMUM DE VRAISEMBLANCE.

Considérons les n observations $(y_1, \ldots, y_n)$, réalisations de : $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} F_\theta$
Disposant des n observations y_i, et *ayant choisi un modèle paramétrique* de la loi commune F_θ des Y_i, on se propose d'estimer le paramètre inconnu θ. Pour simplifier, on considère $\theta \in \mathbb{R}$, mais tout ce qui va être présenté se généralise immédiatement au cas d'un vecteur de p paramètres, *i.e.*, $q \in \mathbb{R}^p$, p entier > 1.

· Pour l'une quelconque des n variables aléatoires Y_i, F_θ désigne la fonction de répartition, *i.e.*, Proba$\{ Y_i \leq y \} = F_\theta(y)$. Si F_θ est différentiable, on peut aussi caractériser la loi de Y_i par sa fonction de densité : $f_\theta(y) = dF_\theta(y)/dy$. Deux cas sont à distinguer : ***(i)*** $f_\theta(y)$ est une probabilité lorsque Y_i est une variable discrète ; *e.g.*, $Y_i \in$ N, $Y_i \sim P(\theta)$, loi de Poisson de paramètre $\theta > 0$, alors : $f_\theta(y) = e^{-\theta}\theta^y/y! =$ Proba$\{ Y_i = y \}$. ***(ii)*** Si Y_i est continue, par exemple si Y_i suit une loi normale, $f_\theta(y)$ n'est pas une probabilité (sa valeur peut excéder 1). Pour $Y_i \in \mathbb{R}$ et une petite valeur $\Delta > 0$: Proba$\{ Y_i \in [y, y + \Delta] \} \approx f_\theta(y).\Delta$; le calcul exact requiert l'intégration de la fonction de densité :

$$[a, b] \text{ intervalle de } \mathbb{R} \; ; \quad \text{Proba}\{ Y_i \in [a, b] \} = \int_a^b f_\theta(y)dy = F_\theta(b) - F_\theta(a).$$

· Désignons par $\mathbf{Y}_n$ l'échantillon aléatoire $(Y_1, \ldots, Y_n)$, et notons $\mathbf{y}_n$ sa réalisation, *i.e.*, les n observations $(y_1, \ldots, y_n)$. Soit $f(\mathbf{Y}_n ; \theta)$ la fonction de densité du vecteur $\mathbf{Y}_n$:

$$f(\mathbf{Y}_n ; \theta) = f_\theta(y_1).f_\theta(y_2). \ldots .f_\theta(y_n), \text{ classiquement notée : } f(\mathbf{Y}_n ; \theta) = \prod_{i=1}^{n} f_\theta(y_i)$$

$f(\mathbf{Y}_n ; \theta)$ est une densité de probabilité : c'est une fonction des n variables y_i, le paramètre θ étant fixé. Le maximum de vraisemblance est fondé sur *la fonction de vraisemblance* ("*likelihood function*") :

$$L(\theta ; \mathbf{y}_n) = \prod_{i=1}^{n} f_\theta(y_i)$$

dont la définition apparaît à première vue identique à celle de la densité $f(\mathbf{Y}_n ; \theta)$. De fait, la vraisemblance $L(\theta ; \mathbf{y}_n)$ est la valeur de la densité de $\mathbf{Y}_n$ qui correspond à l'échantillon observé $\mathbf{y}_n$ (c'est exactement la probabilité d'obtenir cet échantillon si les Y_i sont discrètes, *vide supra*). Mais à la différence de $f(\mathbf{Y}_n ; \theta)$, la vraisemblance $L(\theta ; \mathbf{y}_n)$ n'est pas une fonction des variables y_i, et n'est donc pas une densité : $L(\theta ; \mathbf{y}_n)$ est une fonction de θ, dans laquelle les données observées $y_1, \ldots, y_n$ sont considérées comme des quantités fixées.

Estimer le paramètre θ par le maximum de vraisemblance consiste à identifier la valeur $\hat{\theta}$ qui maximise la probabilité d'obtenir l'échantillon $\mathbf{y}_n$ que l'on a effectivement observé. Pour cela, on résout les *équations de la vraisemblance, i.e.,* on recherche la solution $\hat{\theta}$ du système :

$$\partial L/\partial\theta\,\big|_{\theta=\hat{\theta}} = 0 \ , \text{ et : } \ \partial^2 L/\partial\theta^2\,\big|_{\theta=\hat{\theta}} < 0$$

La première équation peut posséder plusieurs solutions ; seules celles qui vérifient la seconde correspondent à un maximum de $L(\theta\,;\,\mathbf{y}_n)$.

· En pratique, compte tenu de la forme des lois de probabilité usuelles, il est plus aisé d'employer le logarithme de L, que l'on appelle la *log-vraisemblance*. On résoudra donc :

$$\partial \ln(L)/\partial\theta\,\big|_{\theta=\hat{\theta}} = 0 \ , \text{ avec : } \ \partial \ln(L)/\partial\theta = \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \ln\big[f_\theta(y_i)\big] \ ,$$

en respectant la condition nécessaire de maximum.

Les estimateurs du maximum de vraisemblance (EMV) sont parmi les plus habituellement utilisés, et pourtant leurs premières propriétés sont plutôt de type "négatif". En effet, *il n'existe aucune raison*
- qu'un EMV soit non biaisé (et donc, *a fortiori,* qu'il soit efficace ou simplement optimal),
- qu'un EMV soit unique.

Mais les EMV possèdent par ailleurs d'intéressantes propriétés, en particulier asymptotiques.

· ***Propriétés à distance finie*** **:**

Considérons le modèle d'échantillonnage paramétrique $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} F_\theta$, et notons $\hat{\theta}_n$ un EMV du paramètre $\theta \in \mathbb{R}$, de vraie valeur $\theta^\bullet$, inconnue. Les propriétés de $\hat{\theta}_n$ seront seulement énoncées, sans détailler leurs conditions d'existence (en particulier, la densité de $\mathbf{Y}_n$ doit être deux fois dérivable par rapport à θ). Le lecteur intéressé par les démonstrations est invité à consulter les ouvrages généraux de Statistique.

[1] S'il existe une statistique exhaustive $T(\mathbf{Y}_n)$ pour l'échantillon aléatoire $\mathbf{Y}_n$ (*Cf.* § 2.7), alors l'EMV $\hat{\theta}_n$ est fonction de $T(\mathbf{Y}_n)$.
N.B. : cela n'implique pas que l'EMV soit lui-même exhaustif.

[2] S'il existe un estimateur efficace du paramètre θ, alors cet estimateur est un EMV de θ.

[3] Invariance : soit une fonction $\varphi(\theta)$, suffisamment régulière. Si $\hat{\theta}_n$ est un EMV de θ, alors $\varphi(\hat{\theta}_n)$ est un EMV de $\varphi(\theta)$.

· ***Propriétés asymptotiques*** **:**

[4] Tout EMV est convergent.

[5] Sous certaines condition de régularité de F_θ, tout EMV $\hat{\theta}_n$ du paramètre θ est tel que :

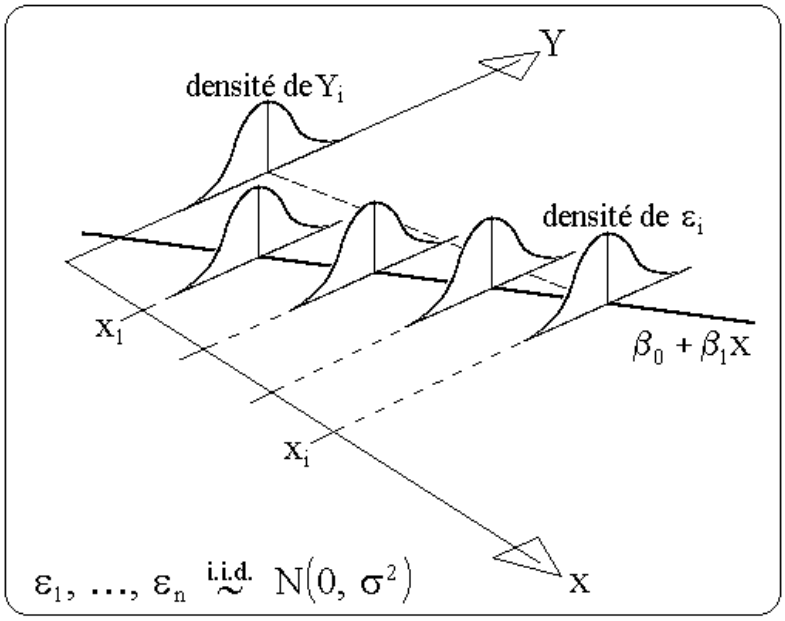
$$\sqrt{n}(\hat{\theta}_n - \theta^\bullet) \xrightarrow{\text{loi}} N[\,0,\, 1/\sqrt{I_Y(\theta^\bullet)}\,]$$

On retiendra plus particulièrement que les propriétés [2] et [5] permettent d'affirmer que, si $\hat{\theta}_n$ est un EMV de θ :

- ou bien il est efficace, et c'est alors le meilleur estimateur non biaisé de θ,
- ou bien il n'existe pas d'estimateur efficace du paramètre θ, mais l'EMV $\hat{\theta}_n$ tend à se comporter comme un estimateur efficace quand n augmente, tandis que sa distribution se rapproche de la normalité.

2.11. MODELE LINEAIRE SIMPLE : ESTIMATION DES PARAMETRES LORSQUE LES RESIDUS SONT NORMAUX.

Le modèle paramétrique adopté pour décrire la densité de probabilité des résidus peut être représenté par la figure suivante :



· ***Estimation des paramètres par le maximum de vraisemblance :***

$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \Rightarrow Y_1, \ldots, Y_i, \ldots, Y_n$ variables aléatoires normales, avec :

$$E[Y_i] = \beta_0 + \beta_1 x_i \;,\quad Var[Y_i] = \sigma^2$$

Densité de Y_i : $f(y\,;\,\beta_0,\beta_1,\sigma) = (\,1/(\sigma\sqrt{2\pi})\,)\exp\left\{ -(y-\beta_0-\beta_1 x_i)^2/(2\sigma^2) \right\}$

Echantillon observé : $\{(x_1, y_1), \ldots, (x_i, y_i), \ldots, (x_n, y_n)\}$

Notons b_0, b_1, et s^2 les valeurs courantes des paramètres β_0, β_1, et σ^2 du modèle, et L la fonction de vraisemblance :

$$L(b_0, b_1, s\,;\, \mathbf{x}, \mathbf{y}) = \prod_{i=1}^{n} f(b_0, b_1, s\,;\, x_i, y_i) = s^{-n}(2\pi)^{-n/2} \exp\left\{ -\frac{1}{2s^2} \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 \right\}$$

On voit immédiatement que pour toute valeur fixée de $s > 0$, on obtient l'équivalence :

Déterminer $\hat{\beta}_0$ et $\hat{\beta}_1$ qui **maximisent** L	$\Leftrightarrow$	Déterminer $\hat{\beta}_0$ et $\hat{\beta}_1$ qui **minimisent** $\sum e_i^2$

C'est à dire que **dans le contexte gaussien, et seulement dans ce cas, le maximum de vraisemblance fournit des estimations identiques à celles qui seraient obtenues avec les moindres carrés**. Il convient par ailleurs de souligner que l'équivalence ne vaut que pour les estimations des paramètres du modèle, et qu'elle ne s'applique pas à la gestion des incertitudes.

· ***Estimation de la variance résiduelle :***

Dans le premier chapitre, où aucune hypothèse probabiliste n'a été formulée sur la loi des ε_i, la minimisation du critère des MCO n'a pas donné d'estimation de la variance résiduelle σ^2 [il demeurait néanmoins possible de l'estimer par la variance empirique des $(Y_i - \hat{y}_i)$]. En revanche, le maximum de vraisemblance permet de résoudre ce problème. Soit donc le modèle paramétrique :

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad , \text{ avec } \varepsilon_1, \ldots, \varepsilon_n \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$$

et le n-échantillon observé : $\{(x_1, y_1), \ldots, (x_i, y_i), \ldots, (x_n, y_n)\}$

Comme précédemment, désignons par b_0, b_1, et s^2 les valeurs courantes des paramètres du modèle. La log-vraisemblance s'écrit :

$$\ln(L(b_0, b_1, s^2\,;\, \mathbf{x}, \mathbf{y})) = -\frac{n}{2}\ln(2\pi s^2) - \frac{1}{2s^2} \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

D'où la condition nécessaire d'extrêmum :

$$\partial \ln(L)/\partial (s^2)\Big|_{s^2 = \hat{\sigma}^2} = -\frac{n}{2\hat{\sigma}^2} + \frac{1}{2(\hat{\sigma}^2)^2} \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 = 0$$

Et si l'on note $\hat{\sigma}^2_{MV}$ (au lieu de $\hat{\sigma}^2$) l'estimateur du maximum de vraisemblance de la variance résiduelle :

$$\hat{\sigma}^2_{MV} = \frac{1}{n} \sum_{i=1}^{n} e_i^2$$

· ***Correction du biais de l'EMV*** $\hat{\sigma}^2_{MV}$**.** D'après l'équation d'analyse de la variance :

$$E\left(\sum e_i^2\right) = E\left(\sum\left(Y_i - \overline{Y}\right)^2\right) - \left(\sum\left(x_i - \overline{x}\right)^2\right)E\left(\hat{\beta}_1^2\right), \text{ avec :}$$

$$\left.\begin{aligned} E\left(\sum\left(Y_i - \overline{Y}\right)^2\right) &= \beta_1^2 \sum\left(x_i - \overline{x}\right)^2 + (n-1)\sigma^2 \\ E\left(\hat{\beta}_1^2\right) &= E^2\left(\hat{\beta}_1\right) + Var\left(\hat{\beta}_1\right) = \beta_1^2 + \sigma^2 / \sum\left(x_i - \overline{x}\right)^2 \end{aligned}\right\} \Rightarrow E\left(\sum e_i^2\right) = (n-2)\sigma^2$$

D'où **l'estimation non biaisée** de σ^2 : $\hat{\sigma}^2 = \left(\sum e_i^2\right)/(n-2)$

On montre d'autre part que $(n-2)\hat{\sigma}^2/\sigma^2$ suit une loi de χ^2 à n–2 d.d.l. ; par conséquent :

$$Var\left(\hat{\sigma}^2\right) = 2\sigma^4/(n-2)$$

· ***Information de Fisher sur les paramètres du modèle linéaire dans le cadre gaussien.***

La définition de l'information pour un paramètre réel (*Cf.* § 2.7) s'étend au cas d'un paramètre vectoriel ; soit donc $q = (\beta_0, \beta_1, \sigma^2)$. L'information qu'apporte le n-échantillon (**x**, **Y**) sur q est exprimée par la matrice d'information de Fisher :

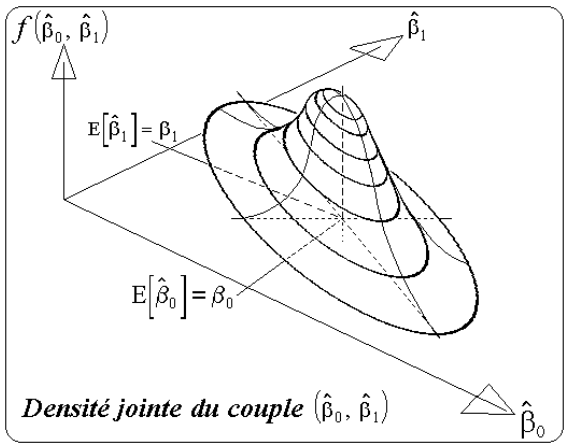
$$I_{(\mathbf{x},\,\mathbf{Y})}(\beta_0, \beta_1, \sigma^2) = \frac{1}{\sigma^2}\begin{pmatrix} n & \sum x_i & 0 \\ \sum x_i & \sum x_i^2 & 0 \\ 0 & 0 & n/(2\sigma^2) \end{pmatrix}$$

L'inverse de cette matrice définit la borne de Fréchet (généralisation de la notion présentée au paragraphe 2.6), qui permet de conlure quant à l'efficacité des EMV de β_0 et de β_1. La matrice d'information est séparable par blocs : on vérifie facilement que l'inverse du bloc nord-ouest (2, 2) est identique à la matrice de variances-covariances des EMV de β_0 et de β_1 (*Cf.* ci-dessous) : on en conclut que ces estimateurs sont **efficaces**. En revanche, l'inverse du bloc sud-est (formé d'un seul élément) de la matrice d'information est égal à $2\sigma^4/n$, *i.e.*, il est inférieur à la variance de l'estimateur non biaisé $\hat{\sigma}^2$ précedemment défini : ce dernier n'est donc pas efficace pour σ^2.

· ***Loi jointe des estimateurs*** $\left(\hat{\beta}_0, \hat{\beta}_1\right)$ **:**

$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix}$ de loi :

$$\mathbf{N_2}\left\{\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix},\ \frac{\sigma^2}{\sum (x_i - \overline{x})^2}\begin{pmatrix} \frac{\sum x_i^2}{n} & -\overline{x} \\ -\overline{x} & 1 \end{pmatrix}\right\}$$



Densité jointe du couple $(\hat{\beta}_0, \hat{\beta}_1)$

2.12. TEST SUR LES PARAMETRES DU MODELE LINEAIRE SIMPLE DANS LE CADRE GAUSSIEN.

L'hypothèse de normalité des résidus permet de connaître la loi des estimateurs (paragraphe précédent) ; il est par conséquent possible de tester des valeurs *a priori* de ceux-ci. On présente ici le test de Student sur une valeur donnée de la pente (en l'occurrence, 0), mais le principe demeure identique pour toute autre valeur, et vaut aussi pour le paramètre "position". On définit d'abord les hypothèses nulle et alternative, puis la statistique du test : c'est une variable aléatoire, dont la loi sous H_0 permet de décider si le résultat expérimental observé est compatible avec H_0, ou si au contraire sa réalisation est plus conforme à l'alternative H_1.

· ***Test sur l'estimation*** $\hat{\beta}_1$ ***de la pente*** β_1 ***:***

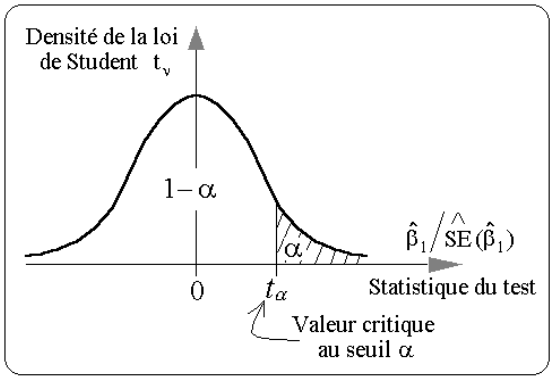
$$①\ :\ \hat{\beta}_1 \sim N\left\{\beta_1,\ \frac{\sigma^2}{\sum(x_i-\bar{x})^2}\right\} \Rightarrow \frac{\hat{\beta}_1-\beta_1}{\sigma\Big/\sqrt{\sum(x_i-\bar{x})^2}} \sim N(0,\ 1)$$

$$②\ :\ \frac{\hat{\sigma}^2}{\sigma^2} \sim \frac{\chi^2_{n-2}}{n-2} \qquad ③\ :\ \text{on montre que } \hat{\beta}_1 \text{ et } \hat{\sigma} \text{ sont indépendants.}$$

$$①\ ②\ ③ \Rightarrow \frac{\dfrac{\hat{\beta}_1-\beta_1}{\sigma\Big/\sqrt{\sum(x_i-\bar{x})^2}}}{\hat{\sigma}/\sigma} = \frac{\hat{\beta}_1-\beta_1}{\hat{\sigma}\Big/\sqrt{\sum(x_i-\bar{x})^2}} \overset{H_0}{\sim} t_{n-2}$$

· ***Application directe :*** tester $H_0 : \beta_1 = 0$, contre $H_1 : \beta_1 > 0$, par exemple.

Statistique du test : $S = \left(\hat{\beta}_1 - 0\right)\Big/\hat{SE}\left(\hat{\beta}_1\right)$; si H_0 est vraie, la densité de probabilité de S se présente comme suit :



Soit t_{obs} la valeur observée de la statistique S du test. La décision repose sur la comparaison au risque α de p = Proba{ $S > t_{obs}$ }, calculée en supposant que H_0 est vraie.

$$\text{En effet : Proba}_{H_0}\left\{\hat{\beta}_1\Big/\hat{SE}\left(\hat{\beta}_1\right) > t_a\right\} = \alpha$$

2.13. RECAPITULATION DES TESTS DE "SIGNIFICATIVITE" DE LA REGRESSION LINEAIRE SIMPLE.

Il existe plusieurs tests équivalents pour décider si la régression est "significative" ou non, autrement dit ***choisir entre le modèle (2)*** :

$$E[Y] = \beta_0 + \beta_1 x,$$

et le modèle (1) qui exprime l'absence de dépendance de Y vis-à-vis de x :

$$E[Y] = \beta_0 = \text{constante}.$$

Ce choix est fondé sur le test des ***hypothèses statistiques*** suivantes:

H_0 : régression linéaire "non significative" [$\Rightarrow$ choisir le modèle (1)],
contre H_1 : régression linéaire "significative" [$\Rightarrow$ choisir le modèle (2)].

Il revient au même d'exprimer comme suit ces hypothèses :

$H_0 : \beta_1 = 0$, contre $H_1 : \beta_1 \neq 0$; ou bien encore $H_0 : \rho = 0$, contre $H_1 : \rho \neq 0$,

où ρ désigne le coefficient de corrélation linéaire, estimé par r. Les statistiques utilisées, ainsi que les lois qu'elles suivent sous H_0 , sont rappelées dans le tableau ci-dessous :

S : statistique du test.	Loi de S sous H_0 .	Rejet de H_0 au seuil α si :
$S = \hat{\beta}_1 \big/ \hat{SE}(\hat{\beta}_1)$	Loi de Student à $\nu = n-2$ d.d.l.	$\lvert S \rvert > t_{n-2\ ;\ \alpha/2}$
$S = r\sqrt{n-2} \big/ \sqrt{1-r^2}$	Loi de Student à $\nu = n-2$ d.d.l.	$\lvert S \rvert > t_{n-2\ ;\ \alpha/2}$
$S = \hat{\beta}_1^2 \sum (x_i - \bar{x})^2 \big/ \hat{\sigma}^2$	Loi de Fisher à $\nu_1 = 1$ et $\nu_2 = n-2$ d.d.l.	$S > F_{1,\ n-2\ ;\ \alpha}$

<u>Remarques</u> :

(i) Le coefficient de détermination R^2 représente la proportion de la variabilité de Y "expliquée" par la régression sur x ; on montre facilement qu'il est égal au carré du coefficient de corrélation empirique r, *i.e.*, $R^2 = r^2$.

(ii) Pour n grand, l'approximation suivante est couramment utilisée : $\frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) \overset{\rho=0}{\sim} N\left(0,\ \frac{1}{\sqrt{n-3}}\right)$

(iii) Si $n \to +\infty$, alors r converge en loi vers $N\left(0,\ 1/\sqrt{n}\right)$ quand $\rho = 0$.

2.14. MODELE LINEAIRE SIMPLE : INTERVALLES DE CONFIANCE DANS LE CADRE GAUSSIEN.

· ***Intervalle de confiance attaché à l'estimation*** $\hat{\beta}_1$ ***de la pente*** β_1 ***:***

Cet intervalle à $100(1-\alpha)\%$ sera noté $IC_\alpha(\hat{\beta}_1)$; par définition, c'est l'ensemble des valeurs conjecturées du paramètre inconnu β_1 que l'on ne repoussera pas dans un test (ici bilatéral) effectué au seuil α :

$$IC_\alpha\left(\hat{\beta}_1\right) = \left[\hat{\beta}_1 - t_{n-2\,;\,\alpha/2}\,\hat{SE}\left(\hat{\beta}_1\right)\ ,\ \hat{\beta}_1 + t_{n-2\,;\,\alpha/2}\,\hat{SE}\left(\hat{\beta}_1\right)\right]$$

avec : $\hat{SE}(\hat{\beta}_1) = \hat{\sigma}\Big/\sqrt{\sum(x_i-\overline{x})^2}$, écart-type estimé de l'estimateur $\hat{\beta}_1$ de la pente β_1.

On rappelle que la vraie valeur β_1 possède la probabilité α de ne pas appartenir à l'intervalle "de confiance" IC_α... En effet, le seuil α, habituellement fixé *a priori* à la valeur 0.05, représente la probabilité de commettre l'erreur de première espèce, *i.e.*, repousser à tort l'hypothèse nulle. IC_α s'interprète de la façon suivante : si l'on pouvait créer un très grand nombre de n-échantillons (x_i, y_i), et calculer pour chacun l'intervalle de confiance attaché à l'estimation de la pente, alors $100(1-\alpha)\%$ d'entre eux contiendraient la vraie valeur β_1.

· ***Remarque :*** Supposons que les inférences concernent à la fois la pente β_1 et la position β_0. On pourrait attacher un intervalle de confiance à 95%, par exemple, à l'estimation de chaque paramètre. Si ces intervalles étaient construits indépendamment l'un de l'autre, la probabilité qu'ils soient simultanément corrects serait $(.95)^2 = .9025$; le risque d'erreur effectif diffère donc de la valeur nominale 5%. Au surplus, les deux intervalles, obtenus à partir du même ensemble de couples (x_i, y_i), ne sont pas indépendants, et cela introduit une complication supplémentaire.

Afin de conclure simultanément sur les deux paramètres avec un risque d'erreur α connu, on construit la région de confiance jointe attachée à l'estimation du couple (pente, position). Il s'agit d'une ellipse dont la surface dépend de f_α, borne du fractile de taille α de la loi de Fisher $F_{2\,;\,n-2}$. L'équation de l'ellipse qui limite la région de confiance est établie en remarquant que les deux variables aléatoires *indépendantes* :

$$\zeta = n(\overline{Y} - E[Y])^2/\sigma^2 + (\hat{\beta}_1 - \beta_1)^2 \sum(x_i-\overline{x})^2/\sigma^2\ ,\ \text{et}\ \xi = (n-2)\hat{\sigma}^2/\sigma^2$$

suivent respectivement une loi de χ^2_2 et de χ^2_{n-2} ; le rapport $(\zeta/2)\big/\big(\xi/(n-2)\big)$ suit donc la loi $F_{2\,;\,n-2}$, *i.e.*, $\text{Proba}\left\{(\zeta/2)\big/\big(\xi/(n-2)\big) \le f_\alpha\right\} = 1-\alpha$; l'estimateur de la position apparaît au numérateur de ce rapport, car il est fonction de $\overline{Y}$ (*Cf.* chapitre 1). Cela permet de tester au seuil α la compatibilité avec l'estimation obtenue $(\hat{\beta}_0, \hat{\beta}_1)$ de valeurs hypothétiques (b_0, b_1). La décision est fondée sur la relation probabiliste :

$$\text{Proba}\left\{\varphi(b_0-\hat{\beta}_0,\ b_1-\hat{\beta}_1) \le f_\alpha\right\} = 1-\alpha,$$

dont l'expression analytique est présentée plus loin.

· ***Domaine de confiance joint du couple d'estimations*** $(\hat{\beta}_0,\ \hat{\beta}_1)$ ***:***

Définir ce domaine équivaut à déterminer **l'ensemble des droites** que l'on ne rejettera pas dans un test au seuil α. Pour cela, on a vu précédemment qu'il n'est pas possible d'utiliser simplement $IC_\alpha(\hat{\beta}_0)$ et $IC_\alpha(\hat{\beta}_1)$. Soient donc :

$$d_{\hat{b}} \;=\; \begin{pmatrix} \hat{\beta}_0 - \beta_0 \\ \hat{\beta}_1 - \beta_1 \end{pmatrix}, \quad \mathbf{V}_{\hat{b}} \;=\; \begin{pmatrix} \mathrm{Var}\left(\hat{\beta}_0\right) & \mathrm{Cov}\left(\hat{\beta}_0,\ \hat{\beta}_1\right) \\ \mathrm{Cov}\left(\hat{\beta}_0,\ \hat{\beta}_1\right) & \mathrm{Var}\left(\hat{\beta}_1\right) \end{pmatrix}, \quad Q \;=\; d'_{\hat{b}}\left(\mathbf{V}_{\hat{b}}\right)^{-1} d_{\hat{b}}$$

La forme quadratique Q s'exprime aussi (pour la définition de la matrice **X**, *cf*. chapitre 3) :

$$Q = \frac{1}{\sigma^2}\left(d'_{\hat{b}}\mathbf{X}'\mathbf{X}d_{\hat{b}}\right) \;=\; \frac{1}{\sigma^2}\left[n\left(\hat{\beta}_0-\beta_0\right)^2 - 2n\bar{x}\left(\hat{\beta}_0-\beta_0\right)\left(\hat{\beta}_1-\beta_1\right) + \left(\hat{\beta}_1-\beta_1\right)^2 \sum x_i^2\right]$$

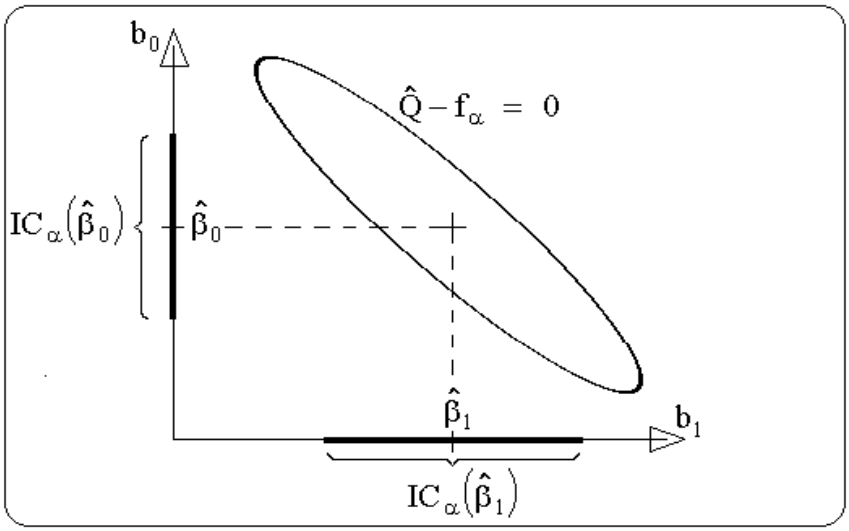
D'après les propriétés des vecteurs aléatoires multinormaux, Q suit un loi de χ^2 à 2 d.d.l., et par définition de la loi de Fisher-Snedecor :

$$(Q/2)/(\hat{\sigma}^2/\sigma^2) \;=\; (\chi^2_2/2)/\left(\chi^2_{n-2}/(n-2)\right) \;\sim\; F_{2\,;\,n-2}$$

Notons b_0 et b_1 les valeurs courantes des estimations $\hat{\beta}_0$ et $\hat{\beta}_1$; en pratique, le contour du domaine de confiance joint du couple d'estimations $(\hat{\beta}_0,\ \hat{\beta}_1)$ est représenté par l'ellipse d'équation :

$$\frac{1}{2\hat{\sigma}^2}\left\{n\left(b_0-\hat{\beta}_0\right)^2 - 2n\bar{x}\left(b_0-\hat{\beta}_0\right)\left(b_1-\hat{\beta}_1\right) + \left(b_1-\hat{\beta}_1\right)^2 \sum x_i^2\right\} - f_\alpha \;=\; 0$$

où la valeur f_α est définie par : $\mathrm{Proba}\left\{F_{2\,;\,n-2} > f_\alpha\right\} = \alpha$



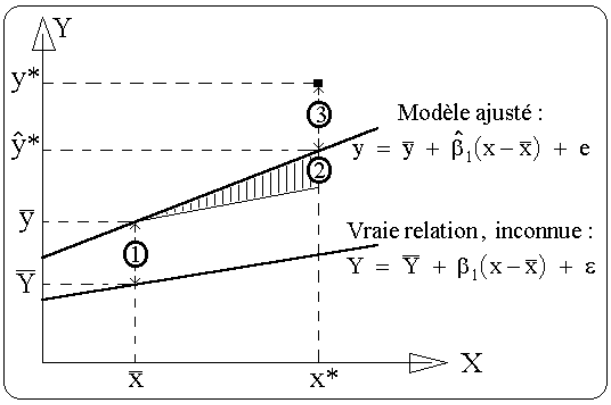
2.15. PREVISION A L'AIDE DU MODELE LINEAIRE SIMPLE.

· ***Deux problèmes :***

(i) Calculer la précision de $\hat{y}^*$, valeur <u>moyenne</u> estimée de la variable réponse, prévue pour la valeur x^* du régresseur.

(ii) Calculer la précision avec laquelle est estimée y^*, réponse <u>individuelle</u> prévue pour une valeur donnée x^* du régresseur.

· ***Trois sources d'incertitude :***



① erreur commise sur l'estimation de $\overline{Y}$ } Problème ***(i)***

② erreur commise sur l'estimation de β_1 }

③ il existe un résidu "autour" de la droite } Problème ***(ii)***

· ***Premier problème :*** variance de la valeur moyenne de la réponse en $x = x^*$.

La vraie "réponse moyenne" (inconnue) vaut : $E\left[Y^*\right] = \overline{Y}^* = E[Y] + \beta_1\left(x^* - \overline{x}\right)$

Dans l'expression de la réponse moyenne calculée $\hat{y}^*$, il est possible de faire apparaître les composantes de l'erreur commise sur la prévision :

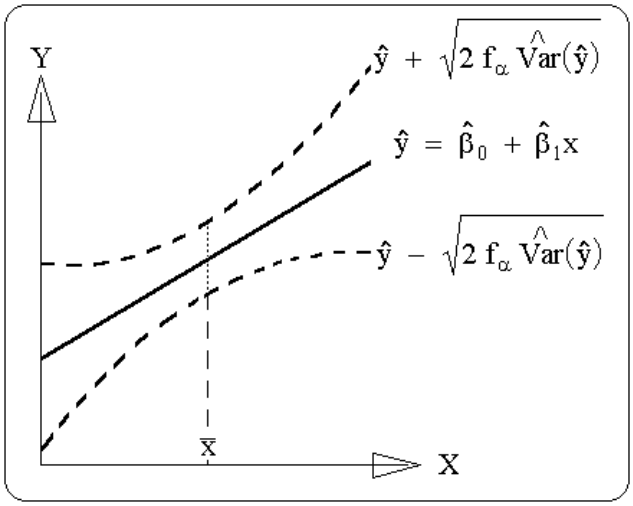
$$\hat{y}^* - \overline{Y}^* \;=\; \overline{Y} \;-\; E[Y] \;+\; \left(x^* - \overline{x}\right)\left(\hat{\beta}_1 - \beta_1\right)$$

On montre que (*Cf.* § 2.9.) : $\mathrm{Cov}\left(\overline{Y},\, \hat{\beta}_1\right) \;=\; 0$; donc :

$$E\left[\hat{y}^* - \overline{Y}^*\right]^2 \;=\; \mathrm{Var}\left(\hat{y}^*\right) \;=\; \mathrm{Var}\left(\overline{Y}\right) \;+\; \left(x^* - \overline{x}\right)^2 \mathrm{Var}\left(\hat{\beta}_1\right)$$

Finalement : $$\mathrm{Var}\left(\hat{y}^*\right) \;=\; \sigma^2 \left[\; \frac{1}{n} \;+\; (x^* - \overline{x})^2 \Big/ \sum_i \left(x_i - \overline{x}\right)^2 \;\right]$$

· ***Application :*** Construction de la **région de confiance** à 100(1-α)% autour de la droite ajustée (schéma ci-après). On se situe ici encore dans le contexte paramétrique, défini par : $\varepsilon_1, \ldots, \varepsilon_n \overset{\text{i.i.d.}}{\sim} N\left(0,\, \sigma^2\right)$; f_α désigne la limite du fractile de taille α de la loi $F_{\nu_1=2,\, \nu_2=n-2}$, *i.e.*, Proba $\left\{F_{2\,;\,n-2} \;>\; f_\alpha\right\} \;=\; \alpha$.



· ***Second problème :*** variance d'une "réponse individuelle" en $x = x^*$.
i.e., calculer la précision d'une valeur prévue y^* de la variable aléatoire Y^* :

$$Y^* = E[Y] + \beta_1\left(x^* - \bar{x}\right) + \varepsilon \text{ , avec : } \varepsilon_1, \ldots, \varepsilon_n \overset{\text{i.i.d.}}{\sim} N\left(0, \sigma^2\right)$$

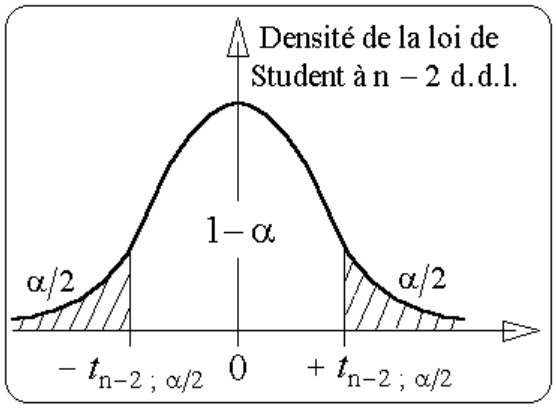
Sachant que la meilleure prévision pour ε est $E[\varepsilon] = 0$, la valeur prévue y^* de Y^* s'exprime :

$$y^* = \bar{Y} + \hat{\beta}_1\left(x^* - \bar{x}\right)$$

Composantes de l'erreur de prévision : comme dans le premier problème interviennent les termes $\bar{Y} - E[Y]$, et $\left(x^* - \bar{x}\right)\left(\hat{\beta}_1 - \beta_1\right)$. Il s'y ajoute une troisième source d'erreur, décorrélée des deux précédentes, la variance résiduelle σ^2. Donc :

$$Var\left(y^*\right) = \sigma^2 \left[1 + \frac{1}{n} + \left(x^* - \bar{x}\right)^2 \Big/ \sum_i \left(x_i - \bar{x}\right)^2 \right]$$

· ***Application :*** attacher un **intervalle de confiance** au niveau $100(1-\alpha)\%$ à l'estimation y^* d'une réponse individuelle prévue en $x = x^*$



$$y^* \pm t_{n-2\,;\,\alpha/2} \sqrt{\hat{Var}\left(y^*\right)}$$

où $\hat{Var}\left(y^*\right)$ s'obtient en remplaçant σ^2 par $\hat{\sigma}^2$ dans l'expression de $Var\left(y^*\right)$, et où la quantité $t_{n-2\,;\,\alpha/2}$ désigne la limite du fractile de taille $\alpha/2$ de la loi de Student à n–2 d.d.l., comme l'indique le schéma ci-contre.

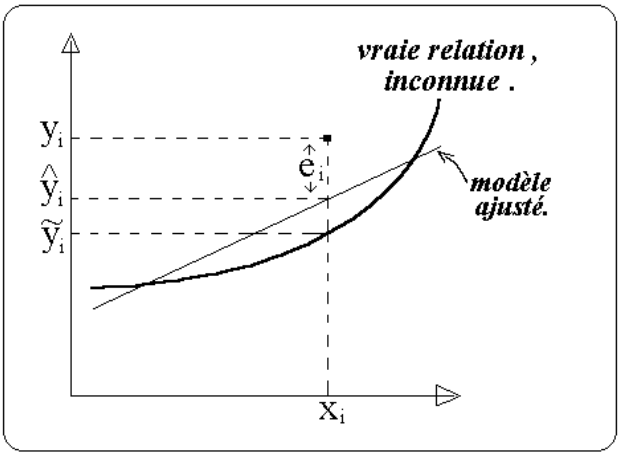
2.16. MODELE LINEAIRE SIMPLE : DEFAUT D'AJUSTEMENT *vs.* ERREUR PURE.

· ***Formulation du problème :***

Il s'agit ici de préciser à quel objectif répond le modèle. A cet égard, il convient de distinguer deux grands types d'applications : ***(i)*** le *modèle qui résume* le plus fidèlement les données observées, et qu'à cette fin on compliquera éventuellement, *vs*. ***(ii)*** le *modèle d'inférence*, qui vise la généralisation des conclusions obtenues à partir des résultats expérimentaux, et qui requiert par conséquent une connaissance précise du degré d'incertitude attaché aux estimations. En pratique, ces deux stratégies peuvent conduire à des choix aux effets antagonistes : le prix de la "description fine" est l'augmentation du nombre de paramètres, option génératrice d'instabilités. La qualité de l'inférence est quant à elle tributaire de la précision, et l'objectif ***(ii)*** incitera donc à ne pas ambitionner un modèle complexe.

Sur le schéma ci-après figure en gras la "vraie" relation, inconnue, que l'on cherche à modéliser. Notons la $\tilde{y} = \phi(x)$, avec : $E[Y_i - \phi(x_i)] = 0$, pour i = 1, ..., n. On a aussi représenté le modèle ajusté (une droite), pour lequel l'hypothèse $E[Y_i - (\beta_0 + \beta_1 x_i)] = 0$ n'est pas vérifiée si $\beta_0 + \beta_1 x \neq \phi(x)$. On va donc tester :

$$H_0 : \phi(x) = \beta_0 + \beta_1 x \text{ , } vs. \text{ } H_1 : \phi(x) \neq \beta_0 + \beta_1 x.$$



La réponse moyenne vraie en x_i est notée : $\tilde{y}_i = E[Y_i]$

De façon purement formelle, on peut écrire :

$$y_i - \hat{y}_i = y_i - \hat{y}_i - E[y_i - \hat{y}_i] + E[y_i - \hat{y}_i]$$

$$= \underbrace{\left((y_i - \hat{y}_i) - (\tilde{y}_i - E[\hat{y}_i])\right)}_{\substack{\textit{Variable aléatoire d 'espérance} \\ \textit{nulle, que le modèle soit} \\ \textit{correct ou non.}}} + \underbrace{(\tilde{y}_i - E[\hat{y}_i])}_{\substack{\textit{Défaut d' ajus-} \\ \textit{tement :} \ B_i}}$$

On montre que : $$E\left[\frac{\sum(y_i-\hat{y}_i)^2}{n-2}\right] = \begin{cases} \sigma^2 & \text{lorsque le modèle est correct} \\ \sigma^2+\dfrac{\sum B_i^2}{n-2} & \text{si le modèle n'est pas correct} \end{cases}$$

Donc, s'il y a défaut d'ajustement (*"lack of fit"*), l'estimateur $\hat{\sigma}^2$ estime la variance résiduelle (*"pure error"*) plus la moyenne des carrés des termes de biais.

· ***Méthode de détection du défaut d'ajustement :***

Il est nécessaire de disposer, de préférence **pour chaque** x_i, de **plusieurs répétitions indépendantes de la variable réponse** y_i ; les données se présentent alors comme suit :

$$\left\{x_i \ ; \ y_{i1}, y_{i2}, \ldots, y_{ij}, \ldots, y_{ip_i}\right\}, \qquad p_i \geq 2, \qquad i = 1, \ldots, n$$

Notons $\bar{y}_i$ la moyenne empirique de la réponse en x_i : $\bar{y}_i = \dfrac{1}{p_i}\sum_{j=1}^{p_i} y_{ij}$

Contribution de l'erreur pure à la dispersion des réponses en $x = x_i$	$\sum_{j=1}^{p_i}(y_{ij}-\bar{y}_i)^2$	*Degrés de liberté :* $p_i - 1$
Somme des carrés totale due à l'erreur pure	$\sum_{i=1}^{n}\sum_{j=1}^{p_i}(y_{ij}-\bar{y}_i)^2$	$n_e = \sum_{i=1}^{n}(p_i-1)$
Variance estimée de l'erreur pure	$\hat{\sigma}_e^2 = \dfrac{1}{n_e}\sum_{i=1}^{n}\sum_{j=1}^{p_i}(y_{ij}-\bar{y}_i)^2$	

· ***Test :*** H_0 = pas de défaut d'ajustement,
contre : H_1 = il existe un défaut d'ajustement.

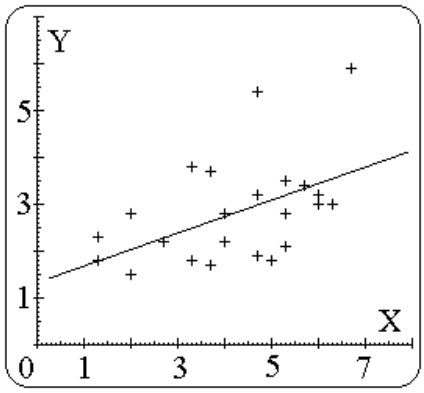
Notons n_r le nombre de d.d.l. de la variance résiduelle estimée $\hat{\sigma}^2$, et S la statistique du test. Cette dernière est simplement le carré moyen dû au défaut d'ajustement, divisé par la variance estimée de l'erreur pure. A l'évidence, H_0 sera repoussée pour les "trop grandes" valeurs de S. Plus précisément, pour décider qu'une valeur est trop grande, on utilise le fait que S suit une loi de Fisher-Snedecor si H_0 est vraie :

$$S = \frac{\left(\hat{\sigma}^2-\hat{\sigma}_e^2\right)/\left(n_r-n_e\right)}{\hat{\sigma}_e^2/n_e} \overset{H_0}{\sim} F_{\nu_1,\nu_2}, \quad \nu_1 = n_r - n_e, \quad \nu_2 = n_e$$

On observe que la mise en évidence d'un éventuel défaut d'ajustement procède d'une classique analyse de la variance : il s'agit d'identifier un signal "masqué" par du bruit. En l'occurrence, on cherche à détecter, *dans la dispersion des réponses autour du modèle ajusté*, un "effet inter-groupes" (une différence entre les moyennes des écarts en chacun des x_i), qui émergerait de la variabilité "intra-groupe" (*i.e.*, de la dispersion des réponses autour de leurs valeurs ajustées). La dispersion attribuable à cet effet inter-groupes est évaluée par la quantité $\hat{\sigma}^2 - \hat{\sigma}_e^2$ qui apparaît au numérateur de S.

Exemple : cet exemple est extrait de l'ouvrage de N. R. DRAPER & H. SMITH (1966) : *Applied Regression Analysis*, J. Wiley & Sons ed.

Données :			
x_i	y_{ij}		
1.3	1.8, 2.3	.125	1
2.0	1.5, 2.8	.845	1
2.7	2.2		
3.3	1.8, 3.8	2.000	1
3.7	1.7, 3.7	2.000	1
4.0	2.2, 2.8, 2.8	.240	2
4.7	1.9, 3.2, 5.4	6.260	2
5.0	1.8		
5.3	2.1, 2.8, 3.5	.980	2
5.7	3.4		
6.0	3.0, 3.2	.020	1
6.3	3.0		
6.7	5.9		
SC totale due à l'erreur pure :		12.470	
et d.d.l. :			11



$$\hat{\sigma}_e^2 = \frac{1}{n_e} \sum_{i,j} \left(y_{ij} - \bar{y}_i \right)^2 = 1.134$$

D'où le tableau d'analyse de la variance:

Source	***SC***	***ddl***	***CM***	***Statistiques des tests***
Totale	27.518	23		
Régression	6.326	1	6.326	6.959 , valeur significative au seuil $\alpha = 0.05$
Résiduelle, dont :	21.192	22	0.963	
défaut d'ajustement	*8.722*	*11*	*0.793*	*0.699 , valeur non significative au au seuil $a = 0.05$*
et erreur pure	*12.470*	*11*	*1.134*	

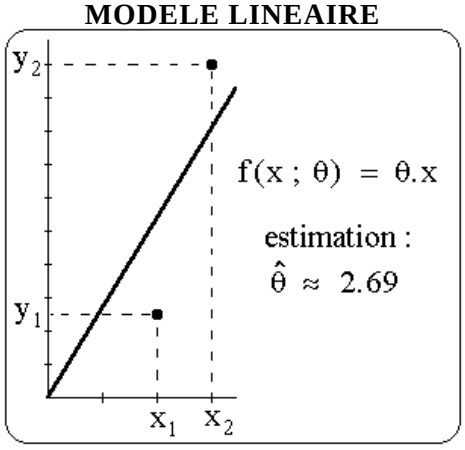
2.17. REGRESSION LINEAIRE *vs.* REGRESSION NON LINEAIRE.

Modèle : $Y_i = f(x_i ; \theta) + \varepsilon_i \qquad i = 1, \ldots, n$

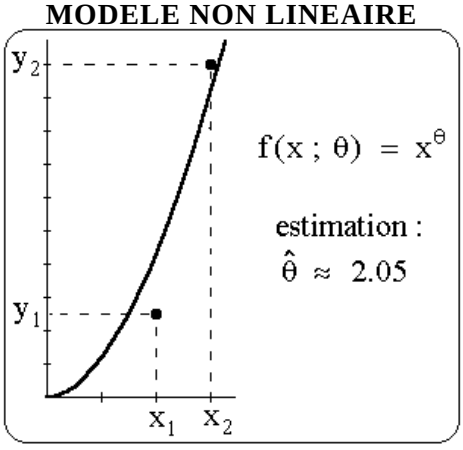
Hypothèses : $\left\{ \begin{array}{l} \left.\begin{array}{l} \text{les } x_i \text{ connus sans erreur,} \\ \theta \text{ paramètre inconnu, fixé.} \end{array}\right\} \text{Quantités certaines.} \\ \varepsilon_1, \ldots, \varepsilon_n \ \textit{i.i.d.}\ N(0, \sigma^2), \text{ composante aléatoire.} \end{array} \right.$

Exemple : pour n=2, ajustement du modèle à $\left\{ \begin{array}{l} (x_1, y_1) = (2.0, 2.5) \\ (x_2, y_2) = (3.0, 10.0) \end{array} \right.$

MODELE LINEAIRE



MODELE NON LINEAIRE



Critère d'optimalité et estimation du paramètre inconnu θ :

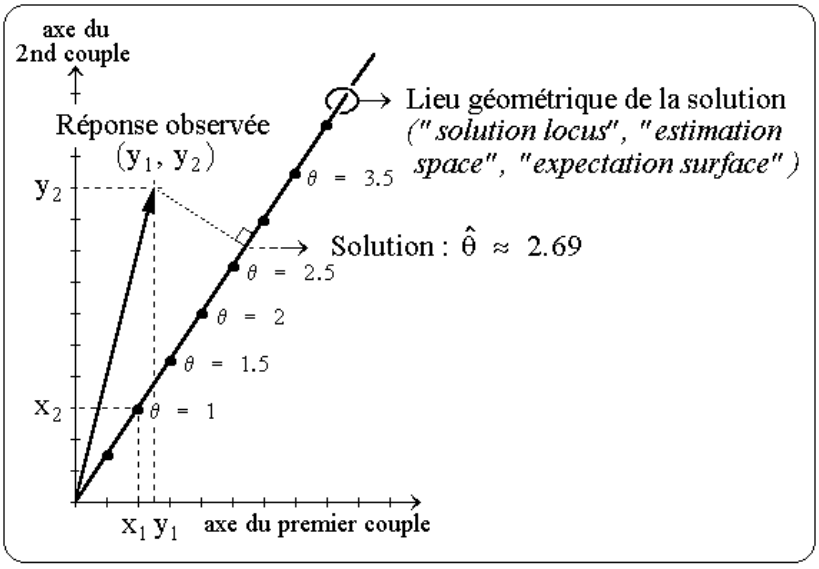
$S(\hat{\theta}) = \min \sum (y_i - \theta x_i)^2$ $\Rightarrow \hat{\theta} = \sum x_i y_i / \sum x_i^2$	$S(\hat{\theta}) = \min \sum (y_i - x_i^{\theta})^2$ $\Rightarrow \sum y_i (\ln x_i) x_i^{\hat{\theta}} = \sum (\ln x_i) x_i^{2\hat{\theta}}$

Propriétés statistiques de l'estimateur du paramètre θ :

$\hat{\theta}$: combinaison linéaire des variables aléatoires Y_i *i.i.d.* $N(\theta x_i, \sigma^2)$ $\Rightarrow \hat{\theta} \sim N(\theta, \sigma^2 / \sum x_i^2)$ L'estimateur $\hat{\theta}$ est non biaisé, et de variance mininale.	$\hat{\theta}$ n'est pas une combinaison linéaire des Y_i, et sa loi n'est en général pas normale. De plus, l'estimateur $\hat{\theta}$ n'est ni sans biais, ni de variance minimale (sauf asymptotiquement).

· ***Lieu géométrique de la solution, cas linéaire :***

Représentation dans l'espace de l'échantillon (*"sample space", "response space"*). Pour un échantillon de n couples, c'est l'espace engendré par n axes de coordonnées, chaque axe étant associé à l'une des observations (x_i, y_i). Dans cet exemple élémentaire emprunté à D.A. RATKOWSKY (*Nonlinear Regression Modeling*, 1983, M. Dekker ed.), on a : n = 2 ; d'où la possibilité d'une simple représentation plane.



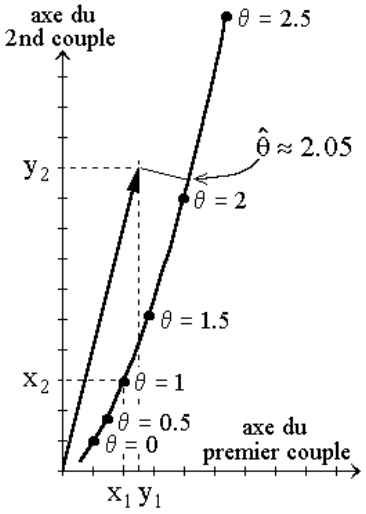
Le lieu géométrique de la solution est un sous-espace de l'espace de l'échantillon, et sa dimension est égale au nombre de paramètres du modèle (1 dans cet exemple). Ce lieu géométrique est défini par la combinaison du modèle et des données ; c'est-à-dire, ici, par $E[y] = \theta x$, et par les deux valeurs $x_1 = 2$ et $x_2 = 3.0$.

Caractérisation du modèle linéaire :

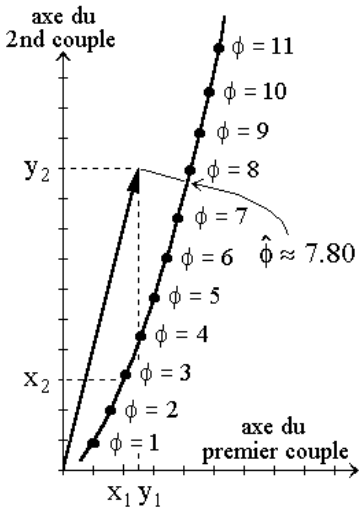
(i) Le lieu géométrique de la solution est linéaire (une droite pour un modèle à 1 paramètre, un plan pour 2 paramètres, ...).
(ii) A des incréments égaux des valeurs du paramètre θ correspondent, sur ce lieu géométrique, des distances égales entre les points (sur une droite), ou entre les droites (dans un plan) , ..., représentatifs des différentes valeurs "équi-espacées" de θ.

· ***Lieu géométrique de la solution, cas non linéaire :***

Modèle : $E[Y_i] = x_i^{\theta}$



Modèle : $E[Y_i] = x_i^{\ln \phi}$



Caractérisation du modèle non linéaire :

Deux mesures du degré de non-linéarité attaché à la combinaison {modèle - protocole expérimental - paramétrisation} ont été proposées par D.M. BATES & D.G. WATTS [1980, *J. R. Statist. Soc.* B **42**(1) : 1-25]. On trouvera une présentation plus récente de ce problème au chapitre 7 de leur ouvrage de 1988, ou au chapitre 4 du livre de G.A.F. SEBER & C.J. WILD (1989) ; *Cf.* références [13] et [14], citées à la page iii du liminaire.

(i) "*intrinsic nonlinearity*" : le lieu géométrique de la solution est courbe.

(ii) "*parameter-effects nonlinearity*" : illustré ci-dessus ; à une discrétisation régulière des valeurs de θ correspondent des intervalles inégaux sur le lieu géométrique (figure de gauche). Cela peut entraîner une dissymétrie de la distribution d'échantillonnage de l'estimateur de θ, effet qu'il est cependant possible de corriger par reparamétrisation du modèle (figure de droite, où $\ln\phi = \theta$).

Chapitre 3

Présentation sommaire de la régression linéaire multiple.

Sommaire du chapitre 3

3.1. MODELE LINEAIRE SIMPLE : PRESENTATION ADOPTANT LE FORMALISME MATRICIEL.

Modèle théorique : $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

Modèle ajusté : $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \ , \ y_i - \hat{y}_i = e_i$

avec i = 1, 2 , ..., n.
On définit :

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \qquad \mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \qquad b = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \qquad e = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Avec ces notations, le modèle théorique s'écrit : $\mathbf{y} = \mathbf{X}b + e$, et le vecteur $\hat{b}$ des estimateurs des moindres carrés ordinaires est celui qui minimise $\| \mathbf{e} \| = \sqrt{\mathbf{e}'\mathbf{e}}$, norme du vecteur **e** des écarts à l'ajustement : $\mathbf{e} = \mathbf{y} - \mathbf{X}\hat{b}$.

$\hat{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, où le symbole $'$ désigne la transposition.

Notons $\mathbf{I}_n$ la matrice identité d'ordre n.
Sous les hypothèses de Gauss-Markov : $E(e) = \mathbf{0}$, $Cov(e) = \sigma^2 \mathbf{I}_n$, il vient :

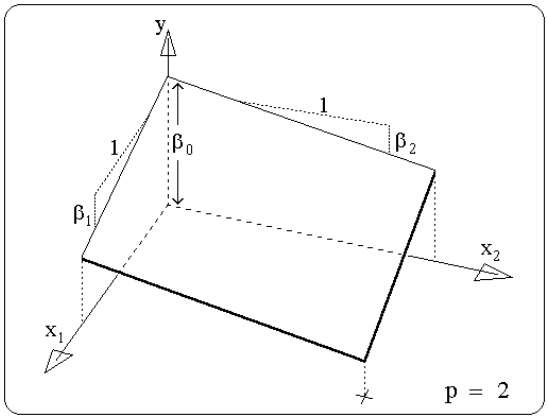
$$E\left(\hat{b}\right) = b \ , \qquad Cov\left(\hat{b}\right) = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$$

et la quantité inconnue σ^2 est estimée par : $\hat{\sigma}^2 = \dfrac{\mathbf{e}'\mathbf{e}}{n-2}$

· ***Intérêt du formalisme matriciel :*** Les résultats présentés pour le modèle linéaire simple (*i.e.*, ne faisant intervenir que p = 1 régresseur, plus un terme constant) se généralisent immédiatement au cas du modèle linéaire multiple (p>1), le dénominateur de la variance résiduelle estimée valant alors n–p–1.

On remarquera aussi que le calcul des estimateurs MCO nécessite la non singularité de la matrice $\mathbf{X}'\mathbf{X}$, ou, ce qui revient au même, que la matrice **X** soit de plein rang p+1. C'est toujours le cas en pratique, ce qui n'exclut cependant pas que la matrice $\mathbf{X}'\mathbf{X}$ puisse parfois être "mal conditionnée" : cette dernière situation, engendrée par le phénomène de *quasi-dépendance linéaire ("colinéarité") entre régresseurs*, a pour effet de déstabiliser les estimateurs MCO des paramètres du modèle (*Cf.* paragraphes 3.6 et suivants).

3.2. MODELE LINEAIRE MULTIPLE : PRESENTATION.



Il est possible de visualiser le modèle à p = 2 régresseurs, plus un terme constant :

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

qui définit le plan de régression partiellement figuré ci-contre.

Ce schéma indique aussi comment s'interprète l'un quelconque des coefficients de régression β_i : c'est la variation de y due à une variation d'une unité de x_i , tous les autres régresseurs étant maintenus constants.
Nous considérerons désormais le cas général : $p \geq 2$; les valeurs des p régresseurs sont rangées dans une matrice **X** formée de n vecteurs-lignes $\mathbf{x}_i$:

$$\mathbf{x}_i = \left(1,\ x_{i1},\ \ldots,\ x_{ij},\ \ldots,\ x_{ip}\right) , \quad i = 1,\ \ldots,\ n , \quad n > p+1$$

La matrice **X** est appelée ***matrice du plan d'expérience*** (*"design matrix"*). Elle est de dimensions $n \times p$ s'il n'y a pas de terme constant, ou bien $n \times (p+1)$ pour un modèle qui inclut terme constant ; seul ce second cas sera envisagé par la suite, et **X** sera supposée être de plein rang p+1. Soit b le vecteur-colonne des paramètres. Avec ces notations, le modèle s'écrit simplement :

$$b = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix} , \quad E(y_i) = \mathbf{x}_i b , \quad i = 1,\ \ldots,\ n$$

3.3. MODELE LINEAIRE MULTIPLE : RESULTATS GENERAUX.

Les résultats qui généralisent ceux établis pour le modèle linéaire simple (p = 1) seront ici rappelés en utilisant le formalisme matriciel. On traitera du modèle avec terme constant β_0 : la première des p+1 colonnes (avec $p \geq 2$) de la matrice **X** ne contient donc que des 1. Le modèle théorique s'énonce :

$$\mathbf{y} = \mathbf{X}b + e , \quad E(e) = \mathbf{0} , \quad Var(e) = \sigma^2 \mathbf{I}$$

Les dimensions des vecteurs-colonnes **y**, e, et b valent respectivement $n \times 1$, $n \times 1$, et $(p+1) \times 1$; **I** est la matrice identité d'ordre n. La $n \times (p+1)$ matrice **X** est non singulière. Les paramètres sont estimés par minimisation du critère des MCO, *i.e.*,

$$(\mathbf{y} - \mathbf{X}\hat{\mathrm{b}})'(\mathbf{y} - \mathbf{X}\hat{\mathrm{b}}) \;=\; \mathbf{e}'\mathbf{e} \;=\; \|\,\mathbf{e}\,\|^2 \;=\; \min\,!$$

où $'$ désigne la transposition ; **e** est le $n \times 1$ vecteur des écarts à l'ajustement. Les estimateurs non biaisés des paramètres et de la variance résiduelle valent respectivement :

$$\hat{\mathrm{b}} \;=\; (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \quad , \quad \hat{\sigma}^2 \;=\; \mathbf{e}'\mathbf{e} \big/ \big(n - (p+1)\big)$$

La matrice de covariance des estimateurs vaut : $\mathrm{Var}\big(\hat{\mathrm{b}}\big) \;=\; \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$

L'équation d'analyse de la variance s'écrit désormais : $\mathbf{y}'\mathbf{y} \;=\; \hat{\mathrm{b}}'\mathbf{X}'\mathbf{y} \;+\; \mathbf{e}'\mathbf{e}$, d'où :

Source de variation	Somme des carrés	d.d.l.
Régression, dont :		
terme constant	$(\Sigma y_i)^2 / n$	1
effet de $x_1, x_2, \ldots, x_p$	$\hat{\mathrm{b}}'\mathbf{X}'\mathbf{y} - (\Sigma y_i)^2 / n$	p
Résiduelle	$\mathbf{e}'\mathbf{e}$	n – (p+1)
Totale	$\mathbf{y}'\mathbf{y}$	n

La qualité de l'ajustement est appréciée à l'aide du coefficient de détermination R^2 :

$$R^2 \;=\; 1 \;-\; \frac{\mathbf{e}'\mathbf{e}}{\mathbf{y}'\mathbf{y} - (\Sigma y_i)^2 / n} \;=\; 1 \;-\; \frac{\text{SC résiduelle}}{\text{SC totale corrigée}}$$

La valeur de R^2 dépend de p : elle diminue quand on augmente le nombre de régresseurs, toutes choses égales par ailleurs. Aussi définit on le ***coefficient de détermination ajusté***, qui tient compte des nombres de d.d.l. :

$$R^2_{\text{ajusté}} \;=\; 1 \;-\; \frac{\mathbf{e}'\mathbf{e} \big/ (n-p)}{\big[\mathbf{y}'\mathbf{y} - (\Sigma y_i)^2 / n\big] \big/ (n-1)} \;=\; 1 \;-\; \frac{n-1}{n-p}\big(1 - R^2\big)$$

· ***Valeurs ajustées ; écarts à l'ajustement :***

Le vecteur des réponses estimées s'obtient par : $\hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{b}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$

que l'on note classiquement : $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$, où : $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$

Nous reviendrons au cinquième chapitre sur la n × n matrice **H**, lorsqu'il sera question de la détection des points influents. Cette matrice est symétrique ; le i-ème élément diagonal h_{ii} a pour expression :

$$h_{ii} = \mathbf{x}_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i'$$

Pour un modèle avec terme constant, et c des n lignes de **X** égales à $\mathbf{x}_i$: $1/n \le h_{ii} \le 1/c$

On a évidemment : $\mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{y}$, et l'on montre que : $\mathrm{Var}(\mathbf{e}) = \sigma^2(\mathbf{I} - \mathbf{H})$

Donc, contrairement aux résidus non observables ε_i , les écarts à l'ajustement e_i ne sont pas de variance constante : $\mathrm{Var}(e_i) = \sigma^2(1-h_{ii})$. Ils sont de plus corrélés entre eux, avec : $\mathrm{Cov}(e_i, e_k) = -\sigma^2 h_{ik}$. Quand on examine les écarts à l'ajustement, on les standardise préalablement afin d'éliminer l'effet dû aux différences entre leurs variances. L'une des standardisations habituelles consiste à définir :

$$i\text{-ème écart standardisé} = \frac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}} \text{ , noté : } e_i^*$$

Bien entendu, ces résultats valent aussi pour le modèle linéaire simple (p = 1) ; h_{ii} vaut alors :

$$h_{ii} = 1/n + (x_i - \bar{x})^2 \Big/ \sum(x_i - \bar{x})^2$$

· ***Prévisions à l'aide du modèle ajusté :***

A l'instar du modèle linéaire simple, la régression multiple peut servir à réaliser une prévision, par exemple en $\mathbf{x} = \mathbf{x}^*$; elle est obtenue par la relation : $\hat{y}^* = \mathbf{x}^*\hat{\mathbf{b}}$

La variance estimée d'une réponse moyenne prévue par le modèle ajusté vaut :

$$\hat{\mathrm{Var}}(\hat{y}^*) = \hat{\sigma}^2 h^* \quad , \quad \text{où : } \quad h^* = \mathbf{x}^*(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}^{*\prime}$$

et dans le cas d'une réponse individuelle prévue en $\mathbf{x} = \mathbf{x}^*$: $\hat{\mathrm{Var}}(y^*) = \hat{\sigma}^2(1 + h^*)$.

Ces résultats permettent d'obtenir des limites de confiance, calculées à l'aide des fractiles d'ordre $\alpha/2$ de la variable *t* de Student à $\nu = n - (p+1)$ d.d.l.

3.4. MODELE LINEAIRE MULTIPLE : DIAGRAMME DE LA VARIABLE AJOUTEE, *"ADDED VARIABLE PLOT"*.

· ***Formulation du problème :***

Considérons la variable y, dépendant des deux variables contrôlées x_1 et x_2, et les trois modèles :

Régression :	de y sur x_1	de y sur x_2	de y sur x_1 et x_2
Proportion de la variabilité "expliquée" de y :	R_1^2	R_2^2	$R_{1,2}^2$

Que peut on dire *a priori* de $R_{1,2}^2$ connaissant R_1^2 et R_2^2 ? La réponse est : pas grand chose, sinon que :

$$R_{1,2}^2 \geq \max\left(R_1^2, R_2^2\right)$$

- Cas particulier : $\mathrm{Cov}(x_1, x_2) = 0$, alors : $R_{1,2}^2 = R_1^2 + R_2^2$

Les deux variables contrôlées x_1 et x_2 sont non corrélées, on dit qu'il y a ***orthogonalité des régresseurs*** ; leurs effets peuvent être caractérisés séparément sans ambiguïté.

- Cas général : $\mathrm{Cov}(x_1, x_2) \neq 0$

Si x_1 et x_2 sont des variables redondantes, qui "expliquent" à peu près de la même façon une partie de la variabilité de y, alors :

$$\max\left(R_1^2, R_2^2\right) \leq R_{1,2}^2 < R_1^2 + R_2^2$$

Si au contraire l'interaction des variables x_1 et x_2 est telle que connaître les deux ensemble apporte plus d'information que de les considérer séparément, alors :

$$R_{1,2}^2 > R_1^2 + R_2^2$$

Par exemple, pour un végétal donné : x_1 = longueur de la feuille, x_2 = largeur de la feuille, y = surface de la feuille.

La difficulté d'emploi de la régression multiple réside précisément dans cette incapacité à connaître *a priori* le lien entre la variable réponse et chacun des régresseurs introduits dans le modèle (sauf s'ils sont orthogonaux). D'où la nécessité de disposer d'outils permettant de caractériser l'effet de l'inclusion d'un régresseur supplémentaire dans un modèle. L'une de ces techniques est la construction du diagramme de la variable ajoutée (*added variable plot,* ou *partial regression leverage plot,* encore appelé *partial residual plot,* la dernière dénomination recouvrant en fait deux méthodes légèrement différentes).

· ***"Added variable plot", modèle :*** $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_j x_j + \ldots + \beta_p x_p + \varepsilon$ [1]

· ***Objectif :*** Visualiser l'effet de la variable contrôlée x_j ($j \in [1, p]$) sur la variable réponse y, après élimination de l'effet des p–1 autres régresseurs. Soit donc :

$$Reg(-x_j) \equiv \text{ensemble des p–1 régresseurs } \{x_1, ..., x_{j-1}, x_{j+1}, ..., x_p\}$$

On s'intéresse à la relation entre y et x_j, toutes deux ajustées par $Reg\ (-x_j)$. Autrement dit, on recherche le lien éventuel entre la variabilité de y "non expliquée" par $Reg\ (-x_j)$, et la variabilité de x_j "non expliquée" par $Reg\ (-x_j)$.

· ***Démarche : (i)*** Ajuster le modèle :

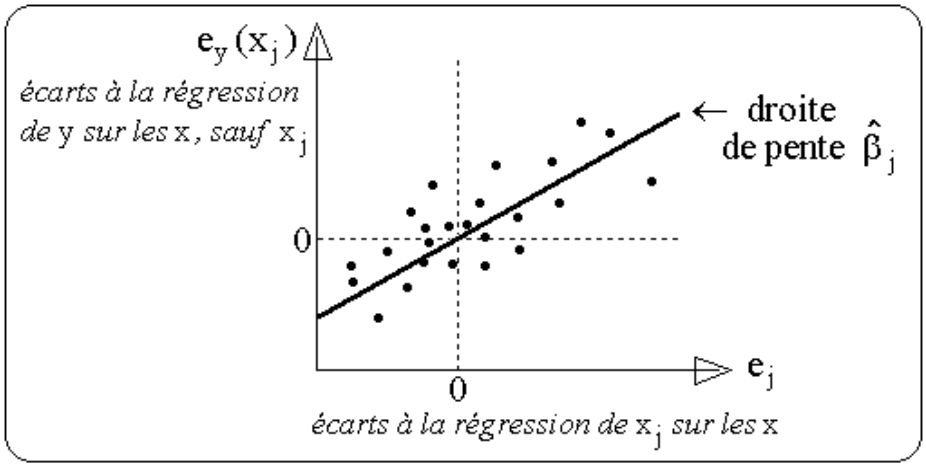
$$\hat{y} = \hat{\alpha}_0 + \hat{\alpha}_1 x_1 + \ldots + \hat{\alpha}_{j-1} x_{j-1} + \hat{\alpha}_{j+1} x_{j+1} + \ldots + \hat{\alpha}_p x_p + e$$

Les écarts à l'ajustement e seront notés : $e_y(x_j)$. Ils représentent la variabilité de y "non expliquée" par les p–1 régresseurs $x_1, ..., x_{j-1}, x_{j+1}, ..., x_p$.

(ii) Ajuster : $\hat{x}_j = \hat{\gamma}_0 + \hat{\gamma}_1 x_1 + \ldots + \hat{\gamma}_{j-1} x_{j-1} + \hat{\gamma}_{j+1} x_{j+1} + \ldots + \hat{\gamma}_p x_p + e$

Les écarts à l'ajustement e seront cette fois notés e_j : ils correspondent aux variations de x_j non expliquées par les p–1 régresseurs $x_1, ..., x_{j-1}, x_{j+1}, ..., x_p$.

(iii) Diagramme de la variable ajoutée :



Ce graphique permet d'évaluer la pertinence de la prise en compte de la variable x_j dans le modèle [1] : une forte tendance linéaire révèle une dépendance marquée de y (ajusté par les autres x) vis-à-vis de x_j. Au surplus, la pente de la régression des $e_y(x_j)$ sur les e_j obtenue par les MCO est identique au coefficient de régression de x_j dans l'ajustement du modèle [1]. La technique apporte enfin une aide précieuse pour l'identification des points douteux, ou encore pour la détection d'éventuelles non-linéarités.

3.5. COMPARAISON DE DROITES DE REGRESSION.

· ***Comparaison de deux droites :*** *i.e.*, comparer deux droites ajustées séparément à deux "populations" de points (x_{hi} , y_{hi}), où h indice la population (ici : h = 1, 2), et i indice l'élément dans la population (i = 1, ..., n_h). L'objectif est de décider si les deux populations peuvent être décrites par une seule droite. On remarque que ce problème n'a de sens qu'à la condition que les variances résiduelles soient identiques, *i.e.*, $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

Modèle "une seule droite" :	***vs.***	**Modèle "2 droites distinctes" :**
$y_{hi} = \beta_0 + \beta_1 x_{hi} + \varepsilon_{hi}$		$\begin{cases} y_{1i} = \beta_0^{(1)} + \beta_1^{(1)} x_{1i} + \varepsilon_{1i} \\ y_{2i} = \beta_0^{(2)} + \beta_1^{(2)} x_{2i} + \varepsilon_{2i} \end{cases}$

Hypothèses statistiques : $H_0 : \begin{cases} \beta_0^{(1)} = \beta_0^{(2)} = \beta_0 \\ \beta_1^{(1)} = \beta_1^{(2)} = \beta_1 \end{cases}$ contre $H_1 : \begin{cases} \beta_0^{(1)} \neq \beta_0^{(2)} \\ \beta_1^{(1)} \neq \beta_1^{(2)} \end{cases}$

Méthode de résolution : construire un modèle "augmenté" à l'aide de la variable indicatrice z_{hi} : $z_{hi} = 0$ si h = 1, et $z_{hi} = 1$ si h = 2.

$$\begin{aligned} y_{hi} &= \theta_0 + \theta_1 x_{hi} + \theta_2 z_{hi} + \theta_3 z_{hi} x_{hi} + \varepsilon_{hi} \\ &= (\theta_0 + \theta_2 z_{hi}) + (\theta_1 + \theta_3 z_{hi}) x_{hi} + \varepsilon_{hi} \end{aligned}$$

donc : $\begin{cases} H_0 : \theta_2 = \theta_3 = 0 \\ H_1 : \theta_2 \neq 0, \ \theta_3 \neq 0 \end{cases}$

On voit immédiatement que si l'on retient H_0 : (β_0 , β_1) = (θ_0 , θ_1). Au contraire, si H_1 est retenue, alors : ($\beta_0^{(1)}$, $\beta_1^{(1)}$) = (θ_0 , θ_1), et : ($\beta_0^{(2)}$, $\beta_1^{(2)}$) = ($\theta_0 + \theta_2$, $\theta_1 + \theta_3$). Le choix entre H_0 et H_1 repose sur la significativité de la diminution relative de la variance résiduelle attachée au modèle qui inclut z et zx.

	Source de variation	Somme des carrés	d.d.l.
Modèle "augmenté" :	*Régression sur x, z, zx*	SCrég*	p+2
	Résiduelle (x, z, zx)	SCrés*	N – 2(p+1)
Une seule droite :	*Régression sur x*	SCrég	p
	Résiduelle (x)	SCrés	N – (p+1)
	Totale	SCtot	N – 1

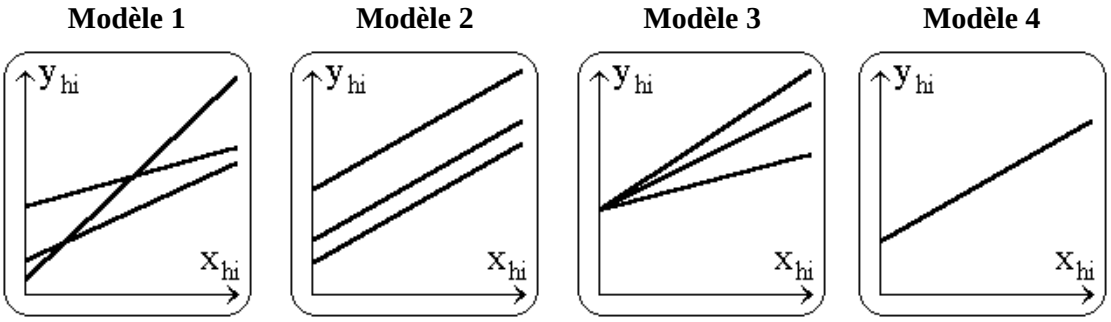
Avec, dans le cas présent, p = 1 (régressions linéaires simples), et $N = n_1 + n_2$.

Base de la décision : $\dfrac{\left(SCrés - SCrés^*\right)/2}{SCrés^*/\left(N - 2(p+1)\right)} \overset{H_0}{\sim} F_{\nu_1 = 2 \ ; \ \nu_2 = N-2(p+1)}$

· ***Cas général, plus de deux droites :***

comparer entre elles H droites distinctes ajustées séparément à H "populations" de points (x_{hi} , y_{hi}), où h indice la population (ici : h = 1, 2, ..., H), et i indice l'élément dans la population (i = 1, ..., n_h).

L'objectif est de décider si les H populations peuvent être décrites par un ensemble de droites défini par un nombre réduit de paramètres, à la limite par une seule droite. Quatre choix traditionnels sont schématisés ci-après, où trois populations seulement ont été considérées (H = 3), afin de ne pas alourdir la présentation :



Méthode de résolution : elle consiste à tester la significativité de l'augmentation relative de la variance résiduelle quand on passe du modèle le plus riche en paramètres (ici, le modèle 1) à un modèle simplifié (*e.g.*, le modèle 3). Le passage d'un modèle "simple" à un modèle plus complexe ("augmenté") est encore une fois réalisé à l'aide de ***variables indicatrices*** z : $z_h = 1$ si l'élément $(x_i , y_i) \in$ population h, et $z_h = 0$ sinon. Elles sont introduites dans les différents modèles comme indiqué ci-après :

Termes de la combinaison linéaire	**Modèle 1**	**Modèle 2**	**Modèle 3**	**Modèle 4**
Terme constant β_0			•	•
x		•		•
z_1	•	•		
z_2	•	•		
z_3	•	•		
z_1x_1	•		•	
z_2x_2	•		•	
z_3x_3	•		•	
d.d.l. de la SC résiduelle :	$dl_1 = N–H(p+1)$	$dl_2 = N–H–p$	$dl_3 = N–Hp–1$	$dl_4 = N–(p+1)$

Seul le cas de régressions linéaires simples a été présenté ici, donc p = 1. La méthode peut cependant être appliquée à des groupes de p>1 régresseurs. Par ailleurs, N désigne ici la somme des n_h , pour h = 1, 2, ..., H.

Hypothèses statistiques : H_0 : modèle m (m = 2, 3, ou bien 4),
contre H_1 : modèle 1.

Base de la décision :

$$\frac{\left(SCrés_{\ m} - SCrés_{\ 1}\right)/\left(dl_m - dl_1\right)}{SCrés_{\ 1}/dl_1} \overset{H_0}{\sim} F_{\nu_1 = dl_m - dl_1\ ;\ \nu_2 = dl_1}$$

Remarques :

(i) Dans son principe, la méthode consiste à tester si la simplification d'un modèle (*i.e.*, la réduction du nombre de ses paramètres) n'altère pas significativement son "pouvoir explicatif" (au sens statistique) : on le compare donc à un modèle plus général. C'est pour cela que le modèle 1 (toutes les droites sont distinctes) joue le rôle de référence dans la statistique ci-dessus, mais d'autres combinaisons peuvent être envisagées.
Ainsi en est il du cas particulier classique de ***"l'analyse de la covariance"***. Elle consiste à tester H_0 : modèle 4, contre H_1 : modèle 2. Dans cette analyse, le régresseur ("covariable") est supposé posséder le même effet pour toutes les populations ("groupes"). Les différences inter-groupes, représentées par les différences entre ordonnées à l'origine, sont attribuées à l'effet additif des "traitements".

(ii) Un problème communément rencontré est celui du "test d'égalité des pentes" de deux droites (*i.e.*, H_0 : modèle 4, H_1 : modèle 2, pour h = 1,2). Il est possible de s'affranchir de l'analyse de la variance, car le test F avec $\nu_1 = 1$ est équivalent au test t. Si les variances résiduelles sont homogènes, on estime σ^2 par :

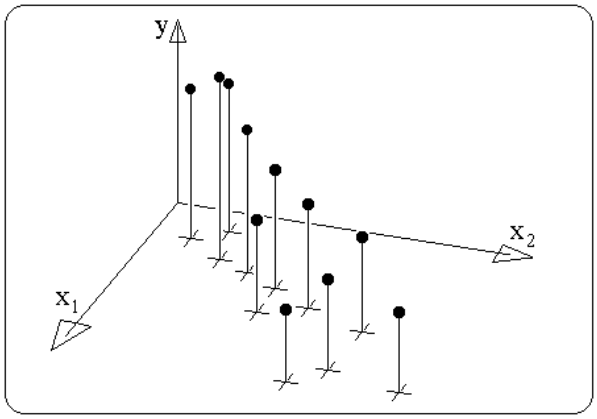
$$\hat{\sigma}^2 = \frac{\left(n_1 - 2\right)\hat{\sigma}_1^2 + \left(n_2 - 2\right)\hat{\sigma}_2^2}{n_1 + n_2 - 4}$$

et, toujours avec des notations évidentes :

$$\frac{\hat{\beta}_1^{(1)} - \hat{\beta}_1^{(2)}}{\hat{\sigma}\sqrt{1/\Sigma\left(x_{1i} - \bar{x}_1\right)^2 + 1/\Sigma\left(x_{2i} - \bar{x}_2\right)^2}} \overset{H_0}{\sim} t_{\nu\ =\ n_1 + n_2 - 4}$$

(iii) Le modèle 3, où les droites sont concourantes en $(0, \beta_0)$, peut aisément être transformé en un modèle où elles sont concourantes en un point d'abscisse *c*.

3.6. MODELE LINEAIRE MULTIPLE : CONSEQUENCES DE LA NON-ORTHOGONALITE DES REGRESSEURS.



Présentation du problème : supposons, pour simplifier, que l'on souhaite ajuster un modèle à p = 2 régresseurs aux données figurées ci-contre. De manière imagée, on peut dire qu'un plan ajusté à ce nuage de points sera "posé en équilibre instable" (*the "wobbly table" metaphor of collinearity…*), et qu'en particulier il "pourra basculer facilement" de part et d'autre de la direction d'allongement maximal du nuage.

Ce ne serait pas le cas si les couples de valeurs des régresseurs étaient régulièrement réparties dans le plan (Ox_1, Ox_2), plutôt que rassemblées au voisinage d'une droite de ce plan.

Du point de vue statistique, la conséquence de cette quasi-colinéarité des régresseurs x_1 et x_2 (habituellement, on dit simplement ***colinéarité***) est la forte variance des estimateurs des paramètres du modèle ; en effet, si l'on note r_{12} le coefficient de corrélation entre x_1 et x_2, on montre que :

$$\mathrm{Var}\left(\hat{\beta}_j\right) = \frac{1}{1-r_{12}} \times \frac{\sigma^2}{\sum\left(x_{ij}-\bar{x}_j\right)^2} \quad , \quad j = 1,\ 2$$

Ce résultat se généralise à p > 2 ; soit R_j^2 le coefficient de détermination de la régression de x_j sur les p−1 autres régresseurs. Alors :

$$\mathrm{Var}\left(\hat{\beta}_j\right) = \frac{1}{1-R_j^2} \times \frac{\sigma^2}{\sum\left(x_{ij}-\bar{x}_j\right)^2} \quad , \quad j = 1,\ \ldots,\ p$$

Le terme $1/(1-R_j^2)$ est appelé j-ème ***facteur d'inflation de la variance***.

3.7. "COLINEARITE" DES REGRESSEURS DU MODELE LINAIRE MULTIPLE : EXEMPLE DE SYMPTÔMES.

Par souci de didactisme, on a choisi plus haut une illustration excessivement simple de la colinéarité. Dans la pratique, le phénomène n'est jamais aussi trivial. Ainsi, plusieurs groupes de régresseurs pourront-ils être engagés dans différentes relations "presque" linéaires, sans qu'apparaissent pour autant de fortes corrélations entre les colonnes de **X**. L'exemple qui suit est emprunté à D. A. BELSLEY (1991) : deux expérimentateurs,

que nous désignerons par A et B, disposent d'un enregistrement commun des variations de la réponse **y** ; en revanche, les expérimentateurs A et B collectent séparément les valeurs correspondantes des régresseurs, d'où une légère différence de lecture des éléments de **X** :

y	$\mathbf{x}_1$		$\mathbf{x}_2$		$\mathbf{x}_3$	
	A	B	A	B	A	B
3.3979	–3.138	–3.136	1.286	1.288	0.169	0.170
1.6094	–0.297	–0.296	0.250	0.251	0.044	0.043
3.7131	–4.582	–4.581	1.247	1.246	0.109	0.108
1.6767	0.301	0.300	0.498	0.498	0.117	0.118
0.0419	2.279	2.730	–0.280	–0.281	0.035	0.036
3.3768	–4.836	–4.834	0.350	0.349	–0.094	–0.093
1.1661	0.065	0.064	0.208	0.206	0.047	0.048
0.4701	4.102	4.103	1.069	1.069	0.375	0.376

Corrélations entre régresseurs :

	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$
$\mathbf{x}_1$	1.000		
$\mathbf{x}_2$	–0.346	1.000	
$\mathbf{x}_3$	0.533	0.610	1.000

L'impact sur les résultats des ajustements réalisés par les deux expérimentateurs est sans commune mesure avec les différences entre les valeurs des régresseurs :

	Coefficients de régression estimés (écart-type entre parenthèses)					
	du terme cst.	de $\mathbf{x}_1$	de $\mathbf{x}_2$	de $\mathbf{x}_3$	$\hat{\sigma}$	R^2
Exp. A	**1.255** (0.091)	**0.974** (3.818)	**9.022** (23.602)	–**38.440** (108.97)	.162	.992
Exp. B	**1.275** (0.093)	**0.247** (2.307)	**4.511** (14.207)	–**17.644** (65.709)	.163	.992

• Considérés séparément, les résultats obtenus par chaque expérimentateur sont tout à fait décevants : même si la valeur de R^2 est élevée, seul l'estimateur $\hat{\beta}_0$ apparaît significativement non nul. C'est là une indication d'une éventuelle colinéarité ; on notera par ailleurs qu'aucune des corrélations entre deux quelconques des régresseurs n'est supérieure à *ca.* 0.6 en valeur absolue.

• Le résultat le plus surprenant est l'ampleur des différences entre les estimations obtenues par A et B. Cette instabilité exprime la forte sensibilité des estimateurs à de petites perturbations de **X**, ici les différences de lecture entre A et B. La sensibilité des estimateurs constitue aussi un indice de colinéarité.

• Un autre symptôme doit être mentionné : les estimations obtenues ne sont pas conformes à la connaissance *a priori* que l'on pourrait posséder sur les valeurs des paramètres. Dans l'exemple traité, la "vraie" relation (inconnue, que l'on cherche à estimer) est en réalité :

$$\mathbf{y} = 1.2 - 0.4\mathbf{x}_1 + 0.6\mathbf{x}_2 + 0.9\mathbf{x}_3 + \mathrm{e}, \text{ avec } \varepsilon_1, ..., \varepsilon_n \text{ i.i.d. } N(0, \sigma = 0.1)$$

3.8. "COLINEARITE" DES REGRESSEURS : POSITION DU PROBLEME.

L'interprétation des résultats de l'ajustement du modèle linéaire multiple repose implicitement sur l'hypothèse que les régresseurs ne sont pas fortement interdépendants (n'oublions pas que les régresseurs sont les "variables indépendantes" du modèle...). Si tel n'est pas le cas, il n'est guère possible d'évaluer l'effet "marginal" de l'un quelconque des régresseurs, *i.e.*, son influence sur la variable réponse lorsque les autres sont maintenus constants.

S'il n'existe aucune relation linéaire entre les régresseurs, ils sont dits orthogonaux. Il leur correspond une matrice $\mathbf{X'X}$ diagonale. Dans la majorité des applications, cette situation de parfaite orthogonalité ne se présente jamais.

D'un point de vue théorique, on lui oppose celle d'une exacte relation linéaire entre régresseurs ("parfaite colinéarité") : il existe au moins un $(p+1) \times 1$ vecteur-colonne $\mathbf{z}$ non nul, tel que $\mathbf{Xz} = \mathbf{0}$. La matrice $\mathbf{X'X}$ est alors singulière (*i.e.*, de rang $< p+1$, non inversible). Ce cas n'est pas non plus rencontré en pratique (le calcul des estimateurs des moindres carrés serait alors impossible), mais en revanche se présentent de nombreuses situations de "quasi-colinéarité" (on dit simplement, suivant un abus de langage communément admis, colinéarité). Il existe alors un ou plusieurs vecteurs $\mathbf{z}$ tels que les produits $\mathbf{Xz}_1$, $\mathbf{Xz}_2$, ..., sont "petits", "voisins du vecteur nul" : un ou plusieurs groupes de régresseurs sont "presque" linéairement liés (*near-linear dependency*, synonyme de *collinearity*, ou encore parfois *multicollinearity*). La matrice $\mathbf{X}$ est alors dite "mal conditionnée" (*ill conditioning*).

La colinéarité est donc *un problème spécifique aux données, plus précisément à la structure de la matrice* $\mathbf{X}$. De façon informelle, on peut dire que l'information contenue dans $\mathbf{X}$ est déficiente, au sens où l'on peut approcher certaines colonnes par des combinaisons linéaires des autres. **La conséquence statistique de ces redondances entre régresseurs est la mauvaise identification de certains paramètres du modèle (fortes variances des estimateurs, sensibilité à de petites petites perturbations des données)**.

Compte tenu des conséquences lourdes du phénomène, il importe donc :
- de déterminer le nombre et la "force" des relations de quasi-dépendance linéaire entre colonnes de $\mathbf{X}$,
- d'identifier les régresseurs engagés dans chacune de ces relations,
- et de quantifier la dégradation de la précision, due à la colinéarité, des estimateurs des MCO du modèle linéaire multiple.

De nombreux diagnostics de colinéarité ont été proposés, fondés par exemple sur l'examen de la matrice $\mathbf{R}$ des corrélations entre régresseurs, le calcul des facteurs d'inflation de la variance (*VIF : variance inflation factor*), la valeur du déterminant de $\mathbf{X'X}$ (ou de $\mathbf{R}$), la recherche des fortes corrélations partielles, ..., pour ne citer que les plus connus. Ces techniques permettent de mettre le phénomène en évidence, mais pour atteindre les trois objectifs précédemment énoncés, il est nécessaire de procéder à une analyse approfondie de

la structure de **X**. En ce sens, le diagnostic précis de la colinéarité repose sur la décomposition de $\mathbf{X'X}$ en éléments propres, ou mieux encore, sur la décomposition en valeurs singulières de la matrice **X** elle même.

3.9. DETECTION DE LA COLINEARITE.

Dans cette partie traitant de la caractérisation de la colinéarité, on considèrera indifféremment le modèle avec ou bien sans terme constant : **X** *possède alors respectivement* p+1 *ou bien* p *colonnes. Afin d'alléger l'écriture, on désignera par* s *le nombre de colonnes de* **X** *dans l'un et l'autre cas. Mais on aura toujours* s < n, *et* **X** *de plein rang* s. $\mathbf{X'}$, $\mathbf{X'X}$ *et* $\mathbf{XX'}$ *seront donc aussi de rang* s.

· ***Diagnostic fondé sur la décomposition en éléments propres de*** $\mathbf{X'X}$.

Les vecteurs propres de la s × s matrice $\mathbf{X'X}$ sont les s vecteurs $\mathbf{v}_1$, ..., $\mathbf{v}_s$ non nuls tels que : $\mathbf{X'Xv} = \lambda\mathbf{v}$, *i.e.*, les vecteurs qui sont simplement "étirés" ($\lambda > 1$) ou bien "rétrécis" ($\lambda < 1$) quand ils sont prémultipliés par $\mathbf{X'X}$. Le "facteur d'élongation" λ_j est la valeur propre associée au vecteur propre $\mathbf{v}_j$. Ici, la matrice $\mathbf{X'X}$ est définie positive et symétrique : ses valeurs propres sont donc réelles et toutes positives, et les vecteurs propres associés à des valeurs propres distinctes sont mutuellement orthogonaux.

La relation avec le diagnostic de colinéarité est évidente : si un vecteur propre $\mathbf{v}_k$ était associé à une valeur propre nulle, alors $\mathbf{X'Xv}_k = \mathbf{0}$, soit encore $\mathbf{Xv}_k = \mathbf{0}$. La matrice $\mathbf{X'X}$ serait dans ce cas singulière, et les composantes de $\mathbf{v}_k$ décriraient la relation d'exacte dépendance linéaire entre les colonnes de **X**.

D'où l'idée (suggérée par M. G. Kendall il y a une trentaine d'années environ) de ***considérer les "petites" valeurs propres de*** $\mathbf{X'X}$ ***comme révélatrices de relations proches de la dépendance linéaire entre régresseurs***. L'intérêt de cette démarche réside dans sa capacité à "débrouiller" les situations complexes où coexistent plusieurs relations : à chacune correspond un "petit λ".

Au plan des applications, la question centrale est alors celle de décider à partir de quel seuil une valeur propre doit être considérée comme petite. Dans le cas présent, la mauvaise référence est zéro ; il faut en effet envisager l'étendue des valeurs propres, *i.e.*, comparer les plus petites à la plus grande. Cela conduit à la notion de nombre de condition de la matrice $\mathbf{X'X}$:

$$\kappa(\mathbf{X'X}) = \sqrt{\lambda_{max}/\lambda_{min}}$$

Ce nombre est toujours supérieur à 1 ; selon S. CHATTERJEE & B. PRICE (1991), qui le calculent pour la matrice **R** des corrélations entre régresseurs, et qui se fondent sur leur propre expérience, il convient de s'alarmer lorsqu'il dépasse 15, et de prendre des mesures correctives à partir de la valeur 30.

Une interrogation de même nature apparaît pour un vecteur propre associé à une petite valeur propre ; on considère habituellement comme "colinéaires" les régresseurs associés aux plus grandes composantes de ce vecteur propre. L'insuffisance de cette procédure est la suivante : l'une des composantes peut être aussi petite que l'on veut, et néanmoins correspondre à l'une des colonnes de **X** qui participe à la relation de quasi-dépendance linéaire.

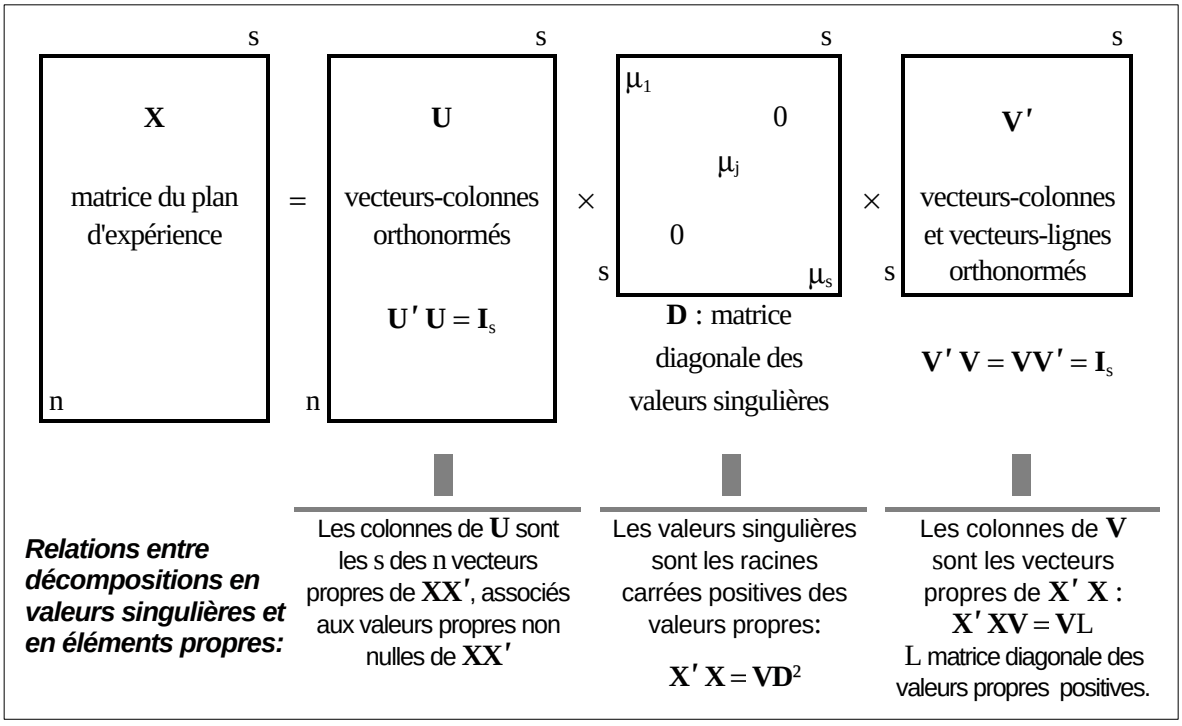
Ces limitations nécessitent de préciser la signification du nombre de condition, et conduisent de plus à préconiser la décomposition de **X** en valeurs singulières (*SVD : singular-value decomposition*), plutôt que celle de $\mathbf{X'X}$ en éléments propres. Outre que la première généralise la seconde, elle est aussi obtenue par des algorithmes numériquement plus stables lorsque **X** est mal conditionnée.

· *Décomposition en valeurs singulières de la matrice* X.

Cette décomposition s'applique à toute matrice, pas nécessairement carrée. On considère ici la $n \times s$ matrice **X** du plan d'expérience, avec $n > s$; sa ***décomposition en valeurs singulières*** s'écrit :

$$\mathbf{X} = \mathbf{UDV'}, \text{ avec : } \mathbf{U'U} = \mathbf{V'V} = \mathbf{I}_s \text{ , et } \mathbf{D} \text{ diagonale}$$

La matrice **X** étant de rang s, aucun des éléments diagonaux de la $s \times s$ matrice **D** n'est nul (le nombre de ces éléments non nuls est en effet égal au rang de **X**). Ils sont tous positifs, et appelés valeurs singulières de **X**. On les note souvent $\mu_1, ..., \mu_s$:



Inverse généralisée de Moore-Penrose (ou ***pseudo-inverse***) : toute matrice possède une inverse généralisée. Celle de la n × s matrice **X** est la s × n matrice notée $\mathbf{X}^+$, calculée comme suit :

$$\mathbf{X}^+ = \mathbf{V}\mathbf{D}^+\mathbf{U}' \ , \ \mathbf{D}^+ \text{ diagonale, } \mathbf{D}^+ = \left(d_{jj}^+\right)_{j=1,\ldots,s} \ , \ d_{jj}^+ = \begin{cases} 1/\mu_j & \text{si } \mu_j > 0 \\ 0 & \text{si } \mu_j = 0 \end{cases}$$

$\mathbf{X}^+$ est unique, et si **X** (n, s) est de plein rang s en colonne (s < n) : $\mathbf{X}^+ = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$

C'est-à-dire : $\hat{b} = \mathbf{X}^+\mathbf{y}$;
on retiendra qu'il est préférable de calculer l'estimateur des MCO à l'aide de l'inverse généralisée, plutôt qu'à partir de la formule classique issue des équations normales. En effet, le nombre de condition de $\mathbf{X}^+$ est égal à $\kappa(\mathbf{X})$, alors que celui de $\mathbf{X}'\mathbf{X}$ est égal à $\kappa^2(\mathbf{X})$. La décomposition en valeurs singulières permet à cet égard d'éviter le calcul de $\mathbf{X}'\mathbf{X}$ et de son inverse.

· ***Outils de diagnostic : nombre de condition, indices de condition.***

- ***Norme spectrale de la matrice*** **X** (n, s) **:** soit le s × 1 vecteur unitaire **z** (*i.e.*, $\|\mathbf{z}\| = \sqrt{\mathbf{z}'\mathbf{z}} = 1$) ; la norme spectrale de **X**, que l'on note ||**X**||, est par définition :

$$\|\mathbf{X}\| = \sup_{\|\mathbf{z}\|=1} \|\mathbf{X}\mathbf{z}\| = \mu_{max}$$

i.e., le maximum de la norme euclidienne du vecteur **Xz**, qui représente la norme spectrale de **X**, est égal à sa plus grande valeur singulière (donc $\|\mathbf{X}^+\| = 1/\mu_{min}$). Comme la norme euclidienne qui l'induit, la norme spectrale est une vraie norme : c'est une fonction à valeurs réelles positives (nulle seulement si **X** = **0**) qui vérifie, pour tout scalaire c, ||c**X**|| = |c|.||**X**||, et qui vérifie aussi "l'inégalité du triangle".

- ***Sensibilité de l'estimateur*** $\hat{b}$ ***des paramètres du modèle linéaire multiple.***

Soient $\delta\mathbf{X}$ et $\delta\mathbf{y}$ de petites perturbations de la matrice des régresseurs **X** et de la variable réponse **y**. On montre que la sensibilité des estimations à ces perturbations est bornée supérieurement par une expression qui dépend du nombre de condition $\kappa(\mathbf{X})$:

$$\frac{\left\|\delta\hat{b}\right\|}{\left\|\hat{b}\right\|} \leq \gamma \frac{\kappa(\mathbf{X})}{\hat{R}} \left(2 + \kappa(\mathbf{X})\sqrt{1-\hat{R}^2}\right) \ , \quad \text{avec : } \begin{array}{c} \gamma = \max\left(\|\delta\mathbf{y}\|/\|\mathbf{y}\| \ , \ \|\delta\mathbf{X}\|/\|\mathbf{X}\|\right) \\ \kappa(\mathbf{X}) = \mu_{max}/\mu_{min} \ , \ \hat{R}^2 = 1 - \mathbf{e}'\mathbf{e}/\mathbf{y}'\mathbf{y} \end{array}$$

La sensibilité de $\hat{b}$ à de petites perturbations des données est donc d'autant plus grande que **X** est mal conditionnée, *i.e.*, que $\kappa(\mathbf{X})$ est lui-même grand. Une mauvaise qualité de l'ajustement, exprimée par une faible valeur du coefficient de détermination non centré $\hat{R}^2$, accroît la sensibilité : celle-ci se rapproche alors de $\kappa^2(\mathbf{X})$.

- Indices de condition.

A la n × s matrice **X** est associé un ensemble de s valeurs réelles ≥ 1, rangées en ordre croissant :

$$\eta_j = \mu_{max}/\mu_j \ , \ j = 1, \ldots, s \ ; \ N.B. : \ \eta_s = \kappa(\mathbf{X})$$

Le diagnostic de la colinéarité est fondé sur ces indices de condition η_j ; à chaque valeur singulière μ_j petite devant μ_{max} correspond une relation de quasi-dépendance linéaire entre les régresseurs. Le nombre de ces relations est égal au nombre des indices de condition η_j qui possèdent une valeur élevée.

- Contributions à la variance des estimateurs.

La variance de chaque estimateur $\hat{\beta}_k$ peut s'écrire comme une somme de s termes, dépendant chacun d'une valeur singulière μ_j de $\mathbf{X} = \mathbf{UDV'}$. Ainsi peuvent être évaluées séparément les influences sur $\mathrm{Var}(\hat{\beta}_k)$ des éventuelles relations de quasi-dépendance entre régresseurs :

$$\mathbf{Var}\left(\hat{b}\right) = \sigma^2(\mathbf{X'X})^{-1} = \sigma^2\mathbf{VD}^{-2}\mathbf{V'} \ , \ i.e. : \ \mathrm{Var}\left(\hat{\beta}_k\right) = \sigma^2 \sum_{j=1}^{s} \frac{v_{kj}^2}{\mu_j^2}$$

Les *proportions* anormalement fortes des variances de deux (ou plus de deux) estimateurs $\hat{\beta}_k$, proportions correspondant à la même petite valeur singulière, révèlent une colinéarité des régresseurs attachés à ces mêmes paramètres β_k.

· Diagnostic de D. A. BELSLEY, 1991.

Ce diagnostic est fondé sur les éléments précédemment présentés : indices de condition, et proportions correspondantes des variances des estimateurs. On rappelle que s = p+1 ou bien p, selon que le modèle inclut ou non un terme constant. Les étapes de la procédure sont les suivantes :

(i) "équilibrage" des colonnes de la matrice **X** : chaque vecteur-colonne $\mathbf{x}_k$ est divisé par sa norme euclidienne $\|\mathbf{x}_k\|$; les colonnes de **X** sont donc préalablement transformées en vecteurs unitaires. Par ailleurs, elles ne sont pas centrées.

(ii) décomposition de **X** en valeurs singulières, et calcul :

- des s indices de condition $\eta_j = \mu_{max}/\mu_j$,
- des contributions à la variance de chacun des estimateurs des s paramètres du modèle (s proportions π_{jk} pour le k-ème estimateur) ;

$$\mathrm{Var}\left(\hat{\beta}_k\right) = \sigma^2\phi_k \ , \ \phi_k = \sum_{j=1}^{s}\phi_{kj} \ , \ \phi_{kj} = v_{kj}^2/\mu_j^2$$

Pour chaque $\hat{\beta}_k$, les proportions de sa variance associées aux valeurs singulières sont simplement :

$$\pi_{jk} = \phi_{kj}/\phi_k \quad , \quad j, k = 1, \ldots, s \quad \text{et} \quad \sum_j \pi_{jk} = 1$$

(iii) rassembler dans un même tableau les éléments du diagnostic : en colonnes, les estimateurs des paramètres du modèle, et en lignes, les indices de condition classés par ordre croissant (cela correspond à un classement en ordre décroissant des valeurs singulières). Les cellules du tableau contiennent les proportions de variance :

Indices de condition η_j	**Proportions π_{jk} de la variance de l'estimateur :**				
	$\hat{\beta}_1$	...	$\hat{\beta}_k$	...	$\hat{\beta}_s$
$\eta_1 = 1$	π_{11}	...	π_{1k}	...	π_{1s}
⋮	⋮				
$\eta_j = \mu_{max}/\mu_j$	π_{j1}	...	$\pi_{jk} = \phi_{kj}/\phi_k$	...	π_{js}
⋮	⋮				
$\eta_s = \kappa(\mathbf{X})$	π_{s1}	...	π_{sk}	...	π_{ss}

(iv) à l'aide des indices de condition η_j, identifier la (ou les) éventuelle(s) relation(s) de quasi-dépendance linéaire entre régresseurs. Selon BELSLEY, les valeurs comprises entre 5 et 10 traduisent de faibles dépendances, et celles allant de 30 à 100 ou plus révèlent des relations modérées à fortes. On retiendra que la valeur de 30 n'est qu'indicative : elle sera sans grande signification si elle coexiste avec une valeur de 3000, par exemple.

(v) pour chaque indice > 30, rechercher les deux (ou plus de deux) proportions de variance "anormalement élevées" (disons, > .5 ; ces proportions appartiennent à une même ligne du tableau ci-dessus). Elles correspondent aux régresseurs qui sont dans une situation proche de la dépendance linéaire, et quantifient la dégradation de la précision des coefficients de régression qui leur sont attachés.

Pour de plus amples informations sur cette méthode de diagnostic, il est conseillé de consulter l'ouvrage de D. A. BELSLEY (1991 ; *cf.* référence [4] au début du présent document), spécialement les chapitres 5 à 7.

3.10. COMMENT PALLIER LA COLINEARITE DES REGRESSEURS ?

Remarque : L'interprétation des coefficients de régression peut s'avérer illusoire lorsque les variables explicatives sont fortement intercorrélées. La construction de la matrice **X** du plan d'expérience ne pouvant pas toujours être maîtrisée, différentes méthodes ont été proposées en vue d'obtenir des estimateurs plus stables (*vide infra*). En tout premier lieu, il est important de retenir qu'il n'existe aucune panacée pour extraire des données une information qu'elles ne contiennent pas. Idéalement, il faudrait donc pouvoir *(i)* collecter des données supplémentaires, et/ou *(ii)* utiliser une information *a priori* sur les paramètres. La première option, outre qu'elle n'est pas toujours réalisable, demeure inopérante si la colinéarité est une propriété intrinsèque des régresseurs. Quant à la seconde, elle est souvent difficile à mettre en pratique.

Concernant le choix des régresseurs, aucune démarche ne saurait remplacer la phase de réflexion sur la nature des données, au cours de laquelle l'expérimentateur peut être conduit à l'élaboration d'un modèle fondé sur une sélection de variables autre que celle qui serait issue d'une procédure automatique (du type de celles décrites dans les manuels spécialisés, *e.g.*, sélection *"forward"*, ou *"backward"*, ou *"stepwise"*) ; au demeurant, ces techniques ne donnent des résultats satisfaisants qu'en l'absence de colinéarité. Diagnostic de colinéarité et sélection des régresseurs sont par nature étroitement liés.

· ***Régression sur composantes principales, ou "orthogonalisée" :*** dans cette méthode, les régresseurs sont orthogonaux : ce sont des combinaisons linéaires des variables contrôlées initiales, obtenues par décomposition en vecteurs et valeurs propres de la matrice $\mathbf{X}'\mathbf{X}$, ou bien d'une matrice qui en est directement déduite (*e.g.*, matrice des covariances, ou des corrélations entre régresseurs). Le profond intérêt de la démarche est très bien illustré par l'exemple traité au chapitre 6 de l'ouvrage de R. TOMASSONE *et al.* (2-ème édition, 1992), auquel nous renvoyons le lecteur. Sur ce même sujet, nous recommandons aussi la lecture du chapitre 7 de l'ouvrage de S. CHATTERJEE & B. PRICE (2nd ed., 1991). Les références précises figurent dans le liminaire du présent document.
Même si l'utilisateur ne souhaite pas employer la régression orthogonalisée, les techniques classiques d'analyse factorielle lui apporteront une aide précieuse pour comprendre la structure des régresseurs (voir par exemple le chapitre 8 du livre de L. C. HAMILTON, 1992) : elles peuvent donc être systématiquement appliquées dans ce but. L'ouvrage de B. ESCOFIER & J. PAGES (*Analyse factorielles simples et multiples. Objectifs, méthodes et interprétation*, Dunod éd., 2nde édition 1990) offre un exposé synthétique de cette famille de méthodes.

· ***Régression pseudo-orthogonalisée ("Ridge regression") :*** mentionnée pour mémoire, elle se rattache à un groupe de méthodes qui visent à minimiser la somme des erreurs quadratiques moyennes (SEQM) des estimateurs des paramètres, auxquels n'est plus imposée la condition de non biais. Elles sont donc fondées sur un critère d'optimalité de la forme :

$$\text{SEQM} \;=\; \sum_{j=1}^{p} \left\{ \text{Var}(\hat{\beta}_j) + \left[\text{Biais}(\hat{\beta}_j) \right]^2 \right\}$$

i.e., l'intérêt porte surtout sur l'estimation des paramètres (le même poids est accordé à chacun), plutôt que sur d'autres aspects de la régression. On notera que les covariances entre paramètres sont ignorées, et aussi que SEQM n'est pas invariant par changement d'échelle. Dans la *"ridge regression"*, l'estimateur est de la forme :

$$\hat{b}_{RR} = (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{y} \quad , \quad k \geq 0, \text{ paramètre de contrôle ou "}\textit{biasing parameter}\text{"}$$

Le paramètre k est choisi de façon à diminuer le nombre de condition de $\mathbf{X}'\mathbf{X}$, en vue d'établir un compromis entre l'augmentation du biais et la diminution de la variance des estimateurs. Même si certains auteurs (*e.g.*, D. C. MONTGOMERY & E. A. PECK, 1992) présentent la *ridge regression* comme une méthode alternative, il est assez généralement admis qu'elle offre plutôt un moyen graphique de détection de la colinéarité, et d'aide à la sélection des régresseurs (en utilisant la *ridge trace*, voir par exemple S. CHATTERJEE & B. PRICE, *op. cit.*, ainsi que ci-après).

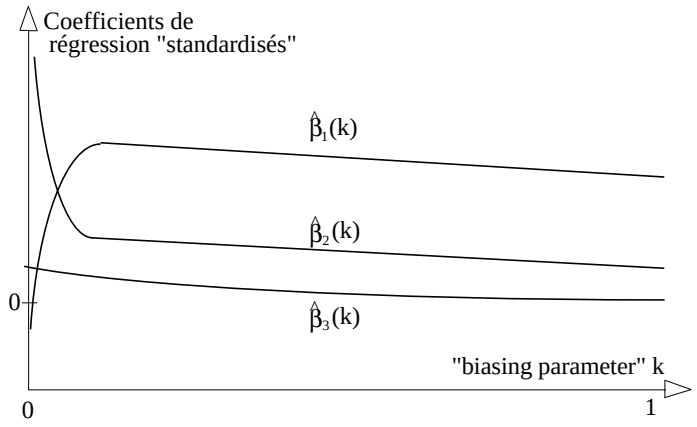
· ***Utilisation de la "ridge trace" pour la détection de la colinéarité, et pour l'aide à la sélection des régresseurs :***

Le paramètre k, dont le rôle est d'améliorer le conditionnement de $\mathbf{X}'\mathbf{X}$, fonde la différence entre les moindres carrés (k = 0) et la régression pseudo-orthogonale (k > 0). Le but est d'obtenir avec la seconde méthode des estimateurs plus stables qu'avec la première, mais biaisés. Dans ce qui suit, les colonnes de $\mathbf{X}$ (ainsi que $\mathbf{y}$) sont centrées et réduites (le terme constant disparaît) ; la matrice $\mathbf{X}'\mathbf{X}$ est donc, à une constante près, la p × p matrice des corrélations entre régresseurs. Les p coefficients de régression, qui sont alors dits "standardisés", sont directement intercomparables.

$\hat{b}_{RR}(k) = (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{y}$; la somme des erreurs quadratiques moyennes des composantes du vecteur aléatoire $\hat{b}_{RR}(k)$ est formée de deux termes :

$$\underbrace{E\left[(\hat{b}_{RR}(k)-b)'(\hat{b}_{RR}(k)-b)\right]}_{\text{Total des erreurs quadratiques moyennes des coefficients de la régression "}\textit{ridge}\text{"}} = \underbrace{\sigma^2 \sum_j \frac{\lambda_j}{(\lambda_j + k)^2}}_{\substack{\text{Somme des variances des composantes de } \hat{b}_{RR}(k), \\ \text{fonction décroissante de k}}} + \underbrace{k^2 b'(\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-2} b}_{\substack{\text{Carré du biais de } \hat{b}_{RR}(k), \\ \text{fonction croissante de k}}}$$

où les $\lambda_j > 0$ sont les p valeurs propres de $\mathbf{X}'\mathbf{X}$. On montre qu'il existe une valeur de k pour laquelle les estimateurs *"ridge"* sont peu sensibles à de petites perturbations des données. En pratique, cette valeur est approchée en représentant graphiquement les variations des estimateurs des paramètres du modèle, en fonction des valeurs de k, pour $k \in [0, 1]$: ce graphe est appelé ***ridge trace***.



Comme le montre l'exemple simple ci-contre, la *ridge trace* permet de repérer visuellement les estimateurs instables, et donc d'établir un premier diagnostic de colinéarité : on remarque ainsi les fortes variations de $\hat{\beta}_1$ et de $\hat{\beta}_2$ (le premier change de signe) pour les petites valeurs de k. Le troisième coefficient standardisé apparaît stable, mais son "pouvoir explicatif" est sans doute faible, car ses valeurs restent proches de zéro.

La *ridge trace* peut aussi être employée pour éliminer du "modèle complet" un ou plusieurs régresseurs. Les règles pratiques sont les suivantes :

(i) éliminer les régresseurs associés aux estimateurs standardisés stables mais petits (*i.e.*, les régresseurs à faible "capacité explicative") ;

(ii) supprimer aussi les régresseurs associés aux estimateurs qui sont instables, et qui de plus tendent vers 0 lorsque k s'approche de 1 ;

(iii) enlever enfin un (ou plusieurs) régresseur(s) affecté(s) d'un coefficient instable.
Pour vérifier que les régresseurs conservés dans le modèle "final" sont dans une configuration proche de l'orthogonalité, on compare les graphes des variations de :

$$\sum_j \hat{\beta}_j^2(k) \quad vs. \quad k, \text{ et de : } \quad \left(\sum_j \hat{\beta}_j^2(0)\right) \Big/ (1+k)^2 \quad vs. \quad k \;;\; k \in [0, 1]$$

(*N.B.* : $\hat{\beta}_j(0)$ est l'estimateur des MCO) ;

si les régresseurs sélectionnés sont presque orthogonaux, les deux courbes obtenues sont quasiment confondues. Un exemple d'application est présenté au chapitre 9 du manuel de S. CHATTERJEE & B. PRICE (1991).

Chapitre 4

La pratique de la régression linéaire : les techniques classiques de validation du modèle.

Sommaire du chapitre 4

4.1. DIAGNOSTIC DES DEFAUTS D'UTILISATION DU MODELE LINEAIRE.

Rappelons que le modèle de régression linéaire est formé de deux composantes (*Cf.* Introduction) : une *partie fixe*, qui est une combinaison linéaire déterministe des paramètres que l'on souhaite estimer, et un *résidu aléatoire*. Les propriétés que l'on postule pour cette seconde composante peuvent être exprimées par les hypothèses de Gauss-Markov, ou bien décrites par un modèle probabiliste paramétrique (la loi normale en Statistique "classique", deuxième chapitre).

La validité des inférences fondées sur le résultat de l'ajustement est tributaire du respect des contraintes qui précèdent. On s'intéressera ici à *la vérification des hypothèses formulées sur les résidus* ; que l'on soit ou non dans le contexte paramétrique, les résidus sont en général l'objet d'un "corpus minimal" de spécifications, qui correspondent aux hypothèses de Gauss-Markov. Il conviendra donc de s'assurer :

- que les ε_i sont d'espérance nulle ; cela est réalisé en vérifiant l'absence de défaut d'ajustement, *i.e.*, d'une éventuelle "dérive" de l'ensemble des observations par rapport à la composante déterministe du modèle (*Cf.* § 2.16.). La question se pose aussi pour des observations prises isolément ("points aberrants"), elle sera considérée dans le cinquième chapitre.
- Que la variance résiduelle est stable ; lorsque ce n'est pas le cas, le palliatif le plus habituel est la transformation des variables. Cette méthode est abordée aux paragraphes qui suivent.
- Que les ε_i ne sont pas corrélés. C'est sans doute le point à la fois le plus important et le plus délicat. Deux situations sont à envisager : (i) la formulation du modèle postule l'absence de corrélation, et il faut donc vérifier que la matrice de covariance des ε_i est diagonale. Le cas élémentaire où cette hypothèse est opposée à l'alternative d'une matrice tridiagonale (autocorrélation d'ordre 1) est présenté au paragraphe 4.7. (ii) Si la matrice de covariance des ε_i est connue *a priori*, et qu'elle est non diagonale, les paramètres seront estimés par les moindres carrés généralisés (§ 4.9).

Dans le contexte paramétrique classique, la loi des erreurs est décrite par le modèle gaussien ; deux procédures de validation (examen graphique + test) sont exposées au paragraphe 4.8.

Les principales qualités requises pour les techniques considérées sont résumées ci-après.

(i) Le comportement de la statistique utilisée doit être connu (au moins approximativement), aussi bien :
- lorsque le modèle est correct,
- que lorsqu'une ou plusieurs des hypothèses de base ne sont pas vérifiées.

Cela suggère l'emploi de *méthodes relativement spécifiques*, plutôt que le recours à des techniques "multi-usages". De ce point de vue, l'efficacité d'une approche globale telle que l'examen du nuage des écarts à l'ajustement est assez rapidement limitée.

(ii) La (ou les) statistique(s) qui fonde(nt) le diagnostic ne doivent pas nécessiter de trop pénibles calculs. Cette condition est généralement remplie dans le cas du modèle linéaire.

(iii) Il est souhaitable que le diagnostic s'appuie sur des représentations graphiques.

(iv) Enfin, si la technique conduit à diagnostiquer un défaut d'utilisation du modèle, elle doit aussi fournir des indications sur les corrections à apporter.

La majorité des méthodes employées procèdent par "augmentation" du modèle initial : ses performances sont comparées à celles d'un modèle plus complexe, dans lequel des modifications sont introduites pour tester spécifiquement les diverses hypothèses mentionnées plus haut.

4.2. HETEROSCEDASTICITE : PALLIATIFS.

Cette situation est celle du ***non respect de l'hypothèse :*** $\mathrm{Var}(\varepsilon_i) = \text{cste} = \sigma^2$

Il est important de mettre en évidence et de corriger l'hétéroscédasticité. En effet, lorsque la variance résiduelle n'est pas stable, les estimateurs des MCO sont toujours sans biais, mais ils sont moins efficaces (ils ne sont pas de variance minimale). Cette perte d'efficacité se conçoit aisément : il existe dans les données une "structure de la variance" que le modèle statistique néglige.

Deux palliatifs sont alors envisageables :

(i) Utiliser un critère de *moindres carrés pondérés* (*Cf.* § 4.10).

(ii) Appliquer à la variable réponse une *transformation stabilisante* de la variance résiduelle. Le tableau ci-dessous (inspiré de S. WEISBERG, *in :* "*Applied Linear Regression*", 2nd ed., 1985, J. Wiley & Sons) recense les transformations les plus classiques :

Situation	***Transformation***	***Remarques***
$\mathrm{Var}(\varepsilon_i) \propto \mathrm{E}(y_i)$	$\sqrt{y}$	Base théorique : transformation conçue pour des dénombrements issus d'une loi de Poisson.
	$\sqrt{y} + \sqrt{y+1}$	Transformation dite de Freeman-Tukey, utilisée lorsque certains des y sont nuls, ou très petits.
$\mathrm{Var}(\varepsilon_i) \propto [\mathrm{E}(y_i)]^2$	$\ln(y)$	Très communément employée, surtout quand l'étendue des valeurs des y (>0) couvre un à plusieurs ordres de grandeur.
	$\ln(y+1)$	*idem*, quand certains y sont nuls.
$\mathrm{Var}(\varepsilon_i) \propto [\mathrm{E}(y_i)]^4$	$1/y$	Recommandée lorsque les valeurs de y sont regroupées au voisinage de 0, et que de fortes valeurs sont observées en bien moins grand nombre. Condition : y>0.
	$1/(y+1)$	*idem*, quand certains y sont nuls.
$\mathrm{Var}(\varepsilon_i) \propto \mathrm{E}(y_i)[1 - \mathrm{E}(y_i)]$	$\arcsin(\sqrt{y})$	Pour les proportions binômiales ; $y \in [0, 1]$

4.3. DETECTION DE L'HETEROSCEDASTICITE.

· ***Pincipe :*** "augmenter" le modèle (concrètement, lui ajouter un paramètre), afin de convertir l'hypothèse $\mathrm{Var}(\varepsilon_i) = \text{cste} = \sigma^2$ en une hypothèse paramétrique testable.

· ***Exemple :*** on considère l'ajustement d' *un modèle linéaire simple*. On soupçonne que la variance résiduelle σ^2 n'est pas stable, et que ses variations dépendent des valeurs de la variable réponse. Cela peut s'écrire :

$$\mathrm{Var}(\varepsilon_i) = \sigma^2\left[\exp(\lambda\, y_i)\right] \qquad \begin{array}{c} i = 1, \ldots, n \\ \lambda : \text{paramètre inconnu.} \end{array}$$

Le test consiste alors à *détruire l'une des deux hypothèses suivantes* :

$$H_0 : \lambda = 0 \qquad H_1 : \lambda \neq 0$$

Etapes de la procédure :

1. Calculer les n valeurs $u_i = e_i^2 / \hat{\sigma}^2_{MV}$, où : $\hat{\sigma}^2_{MV} = \frac{1}{n}\sum_{j=1}^{n} e_j^2$.

2. Calculer la régression (avec terme constant) des u_i sur les valeurs ajustées $\hat{y}_i$. Soit SSreg la somme des carrés due à cette régression : elle possède 1 d.d.l. (la variance σ^2 n'est fonction, dans cet exemple, que de la variable réponse).

$$\mathrm{SSreg} = \left[\sum(u_i - \bar{u})(\hat{y}_i - \bar{y})\right]^2 \Big/ \sum(\hat{y}_i - \bar{y})^2$$

3. La statistique du test est S = SSreg/2. La décision repose sur le fait que la loi asymptotique de la statisitique S est un χ^2 à 1 d.d.l. sous H_0 (*i.e.*, si $\lambda = 0$).

4. Appui graphique : examen du nuage des n couples $\{ (1-h_{ii})\hat{y}_i , (e_i^*)^2 \}$. Ce nuage de points présente une "forme en coin" sous H_1.

· ***Remarque :*** En posant le modèle $\mathrm{Var}(\varepsilon_i) = \sigma^2\left[\exp(\mathbf{l}\, \mathbf{z}_i')\right]$, où $\mathbf{l}$ est un 1×q vecteur-ligne inconnu, et où $\mathbf{z}_i$ est un vecteur-ligne formé de q ≤ p composantes de $\mathbf{x}_i$, on voit que la démarche se généralise aisément, et qu'elle permet de tester une dépendance monotone de σ^2 vis-à-vis d'un (ou d'un groupe de q) régresseur(s). SSreg et S possèdent alors q d.d.l..

4.4. TRANSFORMATION DE LA VARIABLE REPONSE.

Objectifs : *(i)* Stabiliser la variance σ^2.

(ii) Linéariser la relation entre la réponse et les régresseurs.

(iii) Modifier la loi des résidus de telle sorte qu'elle soit plus proche de la normalité, en particulier, plus symétrique.

(iv) Permettre l'emploi d'un modèle plus simple.

Les objectifs *(i)* à *(iv)* ne sont généralement pas atteints à l'aide d'une même transformation. En pratique, il faudra établir des compromis.

· ***Choix d'une transformation : méthode de* BOX & COX.**

Principe : "augmenter" le modèle, de façon à traiter le problème du choix de la transformation comme celui de l'estimation d'un paramètre supplémentaire λ, *i.e.* :

$$\mathbf{z}^{\lambda} = \mathbf{X}b + e \ , \quad \text{avec } \mathrm{Cov}(e) = \sigma^2\mathbf{I} \ , \quad \text{et } \mathbf{z}^{\lambda} = (z_1^{\lambda},\ldots,z_i^{\lambda},\ldots,z_n^{\lambda})'$$

$$\text{où :} \quad z_i^{\lambda} = \begin{cases} \dfrac{y_i^{\lambda}-1}{\lambda\left[\mathrm{GM}(y)\right]^{\lambda-1}} & \text{si } \lambda \neq 0 \\ \mathrm{GM}(y).\ln(y_i) & \text{si } \lambda = 0 \end{cases} \qquad i = 1, \ldots, n$$

GM(y) désigne la moyenne géométrique des y_i : $\mathrm{GM}(y) = \left[\prod y_i\right]^{1/n}$

On observera qu'après application de cette transformation, qui correspond à une forme un peu compliquée de la ***famille des transformations puissance***, le modèle est alors non linéaire.

Problème pratique : estimer simultanément (b,σ^2,λ). Dans la méthode de Box & Cox, le *critère d'optimalité* est construit de telle manière que la transformation rapproche le plus possible la distribution des résidus de la normalité. Il s'agit donc essentiellement d'une ***transformation "normalisante"***.

Règle de choix :		
	$\hat{\lambda} \approx 1 \Rightarrow$	pas de transformation.
	$\hat{\lambda} \approx 0 \Rightarrow$	transformation logarithmique.
	Autrement :	transformation de type puissance.

· ***Mise en pratique :*** supposons que λ soit connu ; on peut alors calculer :

$$\hat{b}_{\lambda} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{z}^{\lambda} \qquad \mathbf{e}_{\lambda} = \mathbf{z}^{\lambda} - \mathbf{X}\hat{b}_{\lambda} \qquad \mathrm{RSS}(\lambda) = \mathbf{e}_{\lambda}'\mathbf{e}_{\lambda}$$

En fait, comme λ est inconnu, on calcule les quantités ci-dessus dans une gamme de valeurs plausibles de λ, *i.e.*, entre –2 et +2. Pour des valeurs de λ extérieures à cet intervalle, l'application de la méthode devient douteuse.

L'intérêt d'avoir choisi pour $\mathbf{z}^{\lambda}$ une formulation un peu compliquée réside dans le fait que, quel que soit λ, la somme des carrés résiduelle RSS(λ) est toujours exprimée dans les mêmes unités : pour λ variant entre –2 et +2, les différentes valeurs de RSS(λ) sont donc directement comparables entre elles, et la plus petite est celle qui correspond à la valeur cherchée $\hat{\lambda}$: RSS($\hat{\lambda}$) = min{RSS(λ)}.

L'estimation $\hat{\lambda}$ est aussi celle qui maximise la log-vraisemblance L(λ) :

$$L(\hat{\lambda}) = \max\{L(\lambda)\} \qquad L(\lambda) = -\frac{n}{2}\ln[RSS(\lambda)]$$

En pratique, une résolution graphique est d'une précision suffisante pour identifier $\hat{\lambda}$: on représente les variations de la courbe en dôme L(λ) *vs.* λ, et l'on retient l'abscisse du maximum. Cette abscisse est en général arrondie à une valeur proche telle que –1, –1/2, 0, ..., qui détermine une transformation simple (voir l'exemple ci-dessous).

Remarques :

(i) La transformation de BOX & COX étant de type puissance, **elle n'est applicable que si la variable réponse est strictement positive.**

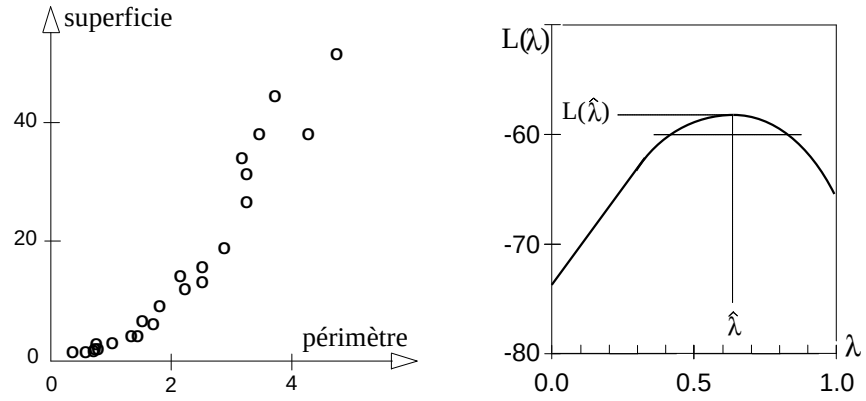
(ii) Pour guider le choix de la transformation, il est possible d'attacher à $\hat{\lambda}$ un intervalle de confiance approché à 100(1-α)%, défini par l'ensemble des valeurs de λ telles que :

$$L(\lambda) > L(\hat{\lambda}) - \frac{1}{2}\chi^2_{\alpha\,;\,1}$$

· ***Exemple d'application :*** cet exemple est extrait de l'ouvrage de Sanford WEISBERG (1985, référence [18] du liminaire du présent document). Il traite de la relation entre le périmètre et la superficie de 25 églises romanes, d'après des données publiées pas S. J. GOULD, et présentées dans le tableau suivant :

église	**périmètre** (× 100 m)	**superficie** (× 100 m²)	**église**	**périmètre** (× 100 m)	**superficie** (× 100 m²)
St. Albans	3.48	38.83	*Byland*	3.14	34.27
Durham	3.69	43.92	*Roche*	2.04	17.61
Blyth	1.43	9.14	*Carmel*	1.77	13.37
Binham	2.05	16.66	*Bengeo*	0.59	2.04
Gloucester	3.05	36.16	*Copford*	0.69	2.22
Norwich	4.19	38.66	*Kempley*	0.50	1.46
Leominster	2.43	17.74	*Birkin*	0.69	1.92
Southwell	2.40	19.46	*Hales*	0.63	1.86
Chertsey	2.72	23.00	*Moccas*	0.58	1.69
Hereford	2.99	29.75	*Peterchurch*	0.86	3.31
Canterbury	4.78	51.19	*Little Tey*	0.41	1.13
Lindesfarne	1.33	6.60	*Melbourne*	1.23	6.74
Tintern	1.47	9.04			

Le nuage des points (ci-dessous à gauche) révèle que la superficie des bâtiments ne varie pas linéairement en fonction de leur périmètre (comme on s'y attend *a priori*...). Le graphe de droite montre la variation de la log-vraisemlance L(λ) en fonction de λ, exposant de la transformation de type puissance que l'on appliquera à la variable "superficie".



Dans cet exemple, la log-vraisemblance est maximale pour $\hat{\lambda} \approx 0.63$; l'intervalle de confiance approché attaché à cette estimation inclut la valeur 1/2. Cela permet, par souci de simplicité, de choisir la transformation racine carrée.

4.5. TRANSFORMATION DE LA VARIABLE CONTRÔLEE.

Deux cas : *(i)* La variable réponse passe par un extrêmum dans le domaine de la (ou des) variable(s) contrôlée(s) → *Régression polynômiale*.

(ii) La variable réponse croît ou bien décroît de façon monotone en fonction de la variable contrôlée, mais à un taux non constant.

→ *Situation examinée ci-après pour le modèle linéaire simple. Ce qui sera présenté se généralise immédiatement au modèle multiple.*

· ***Méthode :*** "augmenter" le modèle, afin :
- de tester la nécessité d'une transformation,
- de choisir éventuellement la transformation adéquate,
- et d'obtenir un moyen d'évaluation graphique.

Modèle "augmenté" : $y_i = \beta_0 + \beta_1 x_i^{\alpha} + \varepsilon_i \qquad i = 1, ..., n$

Plutôt que d'estimer directement les paramètres de ce modèle, on commence par le linéariser en approchant exp(α.ln(x)) par son développement à l'ordre 1 au voisinage de α = 1 :

$$x^{\alpha} \cong x + (\alpha - 1)\frac{\partial}{\partial \alpha} x^{\alpha}\Big|_{\alpha=1} \quad \Rightarrow \quad x^{\alpha} \cong x + (\alpha - 1)\, x \ln(x)$$

D'où l'approximation linéaire du modèle "augmenté" :

$$y_i = \beta_0 + \beta_1 x_i + \eta\, x_i \ln(x_i) + \varepsilon_i \qquad \text{avec :} \quad \eta = \beta_1(\alpha - 1)$$

La décision de transformer ou non la variable contrôlée x repose sur la valeur estimée du coefficient de régression η.

· ***Etapes de la procédure de BOX & TIDWELL :***

1. Obtenir $\hat{\eta}$ en calculant le plan de régression de y sur x et x.ln(x) ;

Remarque : si le rapport max(x)/min(x) est voisin de 10, ou < 10, alors les deux régresseurs x et x.ln(x) sont dans une situation proche de la colinéarité, et le paramètre η est mal identifié (*Cf.* chapitre 3).

2. Décider de la nécessité d'une transformation à l'aide du test :

$$H_0 : \eta = 0 \qquad \text{contre :} \quad H_1 : \eta \neq 0$$

Base de la décision : $\hat{\eta} / \hat{SE}(\hat{\eta}) \sim t_{\nu = n-3}$ sous l'hypothèse nulle.

3. Si H_0 est repoussée, estimer α : $\hat{\alpha} = (\hat{\eta} / \hat{\beta}_1) + 1$
où $\hat{\beta}_1$ est l'estimation obtenue dans la régression de y sur x.

4. Diagnostic graphique : *"**added variable plot**"*, technique classiquement employée en régression multiple (*Cf.* § 3.4.), et permettant ici d'apprécier la relation entre y et x.ln(x), pour y ajusté par x et x.ln(x) . Dans le cas présent, la question pertinente est celle de la caractérisation de l'effet de la variable supplémentaire x.ln(x), ajoutée au modèle qui n'inclut que le régresseur x. Pour répondre à cette question, *on modélise la variabilité de la réponse y que "n'explique pas" x par la variabilité de x.ln(x) "non expliquée" par x*. On représente pour cela le nuage des n points de coordonnées (ζ_i, ξ_i) :

en abscisse, les valeurs ζ_i, écarts à la régression de x.ln(x) sur x,
en ordonnée, les valeurs ξ_i, écarts à la régression de y sur x.

Si le diagramme fait apparaître une relation linéaire forte entre les ξ_i et les ζ_i, il est alors judicieux d'inclure la variable additionnelle au modèle. La valeur du coefficient de régression qui lui est attaché (ici : η) est donnée par l'estimateur MCO de la pente de la droite de régression des ξ_i sur les ζ_i.

· ***Exemple d'application :*** considérons à nouveau la relation superficie *vs*. périmètre, étudiée sur un échantillon de 25 églises romanes de Grande Bretagne (§ 4.4). L'application de la méthode de BOX & COX suggère le modèle :

$$\sqrt{superficie} = \beta_0 + \beta_1(périmètre) + \varepsilon$$

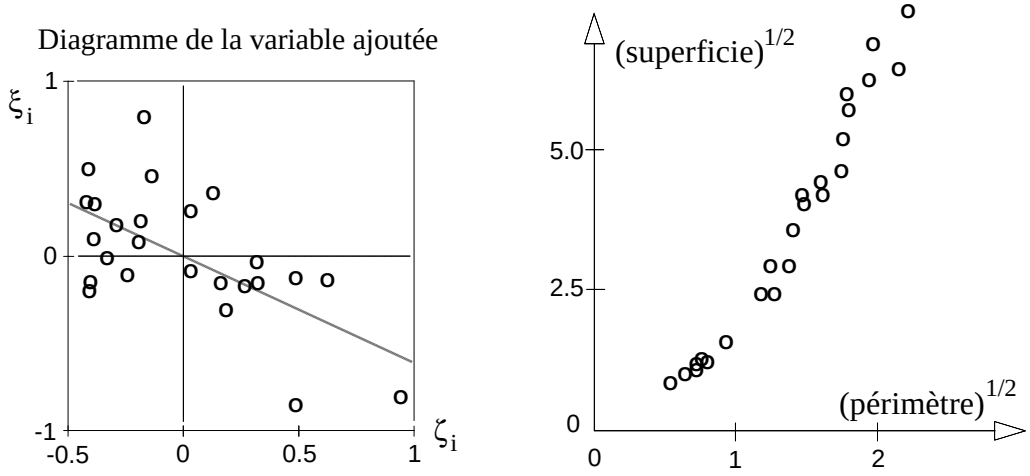
Paramètre	Estimation	Erreur std.	Statistique t	d.d.l.	
β_0	0.5998	0.1374	4.36	23	$\hat{\sigma}^2 = 0.1344$
β_1	1.5437	0.0589	26.20		$R^2 = 0.9676$

L'hypothèse $\beta_1 = 0$ étant rejetée sans ambiguïté, l'intérêt d'appliquer une transformation puissance à la variable "périmètre" est évalué à l'aide du modèle :

$$\sqrt{superficie} = \beta_0 + \beta_1(périmètre) + \eta(périmètre)[\ln(périmètre)] + \varepsilon$$

Paramètre	Estimation	Erreur std.	Statistique t	d.d.l.	
β_0	–0.5036	0.2679	–1.88	22	
β_1	2.6966	0.2625	10.27		$\hat{\sigma}^2 = 0.0739$
η	–0.6727	0.1510	–4.45		$R^2 = 0.9830$

L'exposant α de la transformation est estimé par : $\hat{\alpha} = -.6727 / 1.544 + 1 = .56$



On observera que les petites églises s'écartent de la tendance linéaire globale entre $\sqrt{\text{superficie}}$ et $\sqrt{\text{périmètre}}$. Incidemment, cette déviation conduit à rappeler que ***les méthodes "automatiques" du choix des transformations peuvent aboutir à des résultats qui ne sont pas totalement satisfaisants.***

4.6. TRANSFORMATION DES VARIABLES DU MODELE LINEAIRE : RECOMMANDATIONS PRATIQUES.

1. Il existe une étroite dépendance entre la transformation de la variable réponse et celle(s) de la (ou des) variable(s) contrôlée(s). Des techniques de transformation simultanée ont été proposées (*Cf.* COOK, D.R., & S. WEISBERG, 1982, *Residuals and Influence in Regression*, Chapman & Hall, ed., pp. 83-86), mais il n'est généralement pas utile d'y recourir pour résoudre la majorité des problèmes rencontrés.

2. S. WEISBERG (1985) recommande la démarche suivante, sous réserve que les données soient toutes strictement positives :

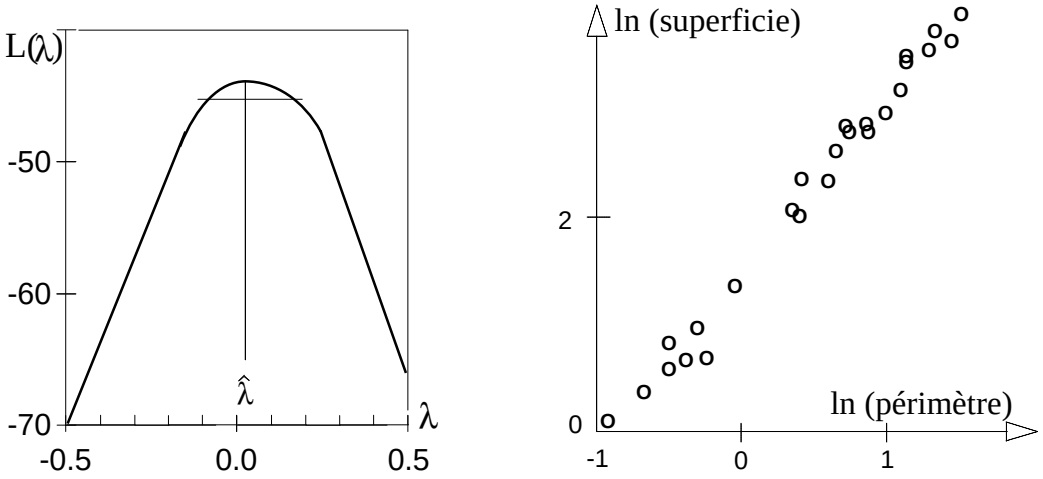
(i)	Appliquer la transformation logarithmique à toute variable contrôlée pour laquelle le rapport (valeur maximale)/(valeur minimale) vaut *ca.* 10 ou plus.
(ii)	Utiliser ensuite la méthode de BOX & COX pour traiter le cas de la variable réponse, et la transformer éventuellement.
(iii)	Revenir enfin sur la (ou les) variable(s) contrôlée(s) dont le coefficient de régression est associé à une forte valeur du t de Student, et appliquer la méthode de BOX & TIDWELL.

3. ***Exemple :*** relation superficie *vs*. périmètre des églises romanes :

(i) étendue des valeurs du régresseur proche de 10 $\rightarrow$ *ln(périmètre)*,

(ii) méthode de BOX & COX : $\hat{\lambda} \approx .02 \rightarrow$ *ln(superficie)*.

Le profil de vraisemblance (à gauche) et le nuage de points après transformation logarithmique des variables périmètre et superficie (à droite) apparaissent ci-dessous :



4.7. L'HYPOTHESE D'ABSENCE DE CORRELATION ENTRE LES RESIDUS.

C'est une hypothèse fondamentale, selon laquelle ***la valeur d'un résidu attaché à une observation n'est aucunement influencée par les valeurs que prennent les autres résidus***. La meilleure vérification du respect de cette hypothèse procède d'une réflexion approfondie sur la nature du processus qui engendre les données. En effet, l'une des sources d'autocorrélation des résidus est souvent l'absence de prise en compte d'un (au moins) régresseur qui influence la variable réponse étudiée. Par ailleurs, il conviendra aussi d'être vigilant lorsque les observations sont ordonnées dans le temps et/ou l'espace, et que les éléments voisins sont susceptibles de se contaminer mutuellement (dans les manuels de langue anglaise, on parle alors de "*serial correlation*").

L'autocorrélation des résidus entraîne de graves conséquences :
- les estimateurs des MCO demeurent non biaisés, mais ils ne sont pas de variance minimale (ils ne sont pas efficaces) ;
- σ^2 est en général sévèrement sous-estimé, donnant ainsi une fausse impression de bonne précision des estimations ;
- les inférences fondées sur les lois t et F ne sont plus valides.

· ***Modèle autorégressif d'ordre 1 (AR-1) des résidus :***

Les techniques de détection de la corrélation entre les résidus sont en général très difficiles à mettre en pratique, sauf dans quelques circonstances. En particulier, lorsque les données sont collectées suivant un pas de temps (ou d'espace) régulier, le test de DURBIN-WATSON peut alors être appliqué. Cette procédure suppose que les résidus sont engendrés par un processus autorégressif d'ordre 1 (elle ne décèlera pas nécessairement une structure d'autocorrélation plus complexe). Soit $\mathbf{y} = \mathbf{X}b + e$. La corrélation la plus simple entre les ε_i est décrite par un modèle AR-1 :

$$\text{AR-1 :}\quad \varepsilon_i = \rho\,\varepsilon_{i-1} + \omega_i \quad , \quad |\rho| < 1 \quad , \quad \omega_1, \ldots, \omega_n \overset{\text{i.i.d.}}{\sim} N\left(0,\ (1-\rho^2)\sigma_\varepsilon^2\right)$$

modèle dans lequel ce sont les résidus ω_i qui sont non corrélés. Le coefficient d'autocorrélation d'ordre 1 (parfois noté ρ_1) est estimé par :

$$\hat{\rho} = \sum_{i=2}^{n} e_i\, e_{i-1} \Big/ \sum_{i=1}^{n} e_i^2$$

où les e_i sont les écarts à l'ajustement du modèle $\mathbf{y} = \mathbf{X}b + e$, *qui inclut un terme constant*. La décision de détruire H_0 ($\rho = 0$) ou bien H_1 ($\rho \neq 0$) repose sur la statistique de DURBIN-WATSON :

$$d = \sum_{i=2}^{n}(e_i - e_{i-1})^2 \Big/ \sum_{i=1}^{n} e_i^2 \quad , \ \textit{i.e.} : \quad d \approx 2(1-\hat{\rho}) \qquad 0 \le d \le 4$$

Les tables fournissent deux seuils d_L et d_U fonctions de n, de p, et du risque α.

	rejeter H_0	*pas de conclusion*	*conserver* H_0
$\hat{\rho} > 0$	$0 < d < d_L$	$d_L < d < d_U$	$d_U < d < 2$
$\hat{\rho} < 0$	$4 - d_L < d < 4$	$4 - d_U < d < 4 - d_L$	$2 < d < 4 - d_U$

· ***Exemple :***

L'exemple qui va être traité est présenté dans le manuel de S. CHATTERJEE & B. PRICE (1991, référence [5] citée dans le liminaire du présent document). Il s'agit de la relation entre les dépenses de consommation trimestrielles, y, et les réserves monétaires, x, toutes deux exprimées en milliards de dollars. Ces données concernent l'économie des USA au cours des années 1952-56 :

Année	trimestre	y	x	Année	trimestre	y	x
1952	1	214.6	159.3	1954	3	238.7	173.9
	2	217.7	161.2		4	243.2	176.1
	3	219.6	162.8	1955	1	249.4	178.0
	4	227.2	164.6		2	254.3	179.1
1953	1	230.9	165.9		3	260.9	180.2
	2	233.3	167.9		4	263.3	181.2
	3	234.1	168.3	1956	1	265.6	181.6
	4	232.3	169.7		2	268.2	182.5
1954	1	233.7	170.5		3	270.4	183.3
	2	236.5	171.6		4	275.6	184.3

Résultat de l'ajustement du modèle : $y_t = \beta_0 + \beta_1 x_t + \varepsilon_t \quad ; \quad t = 1, ..., 20$

Paramètre	Estimation	Erreur std.	Statistique t	d.d.l.	
β_0	–154.700	19.850	–7.790	18	$\hat{\sigma} = 3.983$
β_1	2.300	0.115	20.080		$R^2 = 0.955$

Si l'on néglige l'indispensable étape de l'examen critique des résultats, l'ajustement conduit à la conclusion suivante : la variation d'une unité des réserves monétaires entraîne, au seuil $\alpha = .05$, une multiplication par 2.06 - 2.54 des dépenses. Et la valeur de R^2 indique que *ca*. 95% de la variabilité de y est "expliquée" par celle de x.

Ces conclusions ne sont valides que si les conditions nécessaires aux inférences sont respectées ; ayant ici affaire à une série chronologique, il est judicieux de rechercher une éventuelle autocorrélation des résidus. L'outil standard est dans ce cas **le graphe des écarts "standardisés"** (*Cf.* § 3.3) **en fonction du temps**, présenté à la page suivante.

A ces écarts correspond l'estimation $\hat{\rho} = 0.874$ du coefficient d'autocorrélation d'ordre 1 (il s'agit d'une autocorrélation positive, comme c'est le plus souvent le cas). La statistique d de Durbin-Watson est égale à 0.328 : elle est inférieure à la valeur critique ($d_L = 1.18$, $\alpha = .05$), et l'hypothèse alternative d'une relation AR-1 entre les résidus est acceptée.

· ***Elimination de l'autocorrélation d'ordre 1 :***

Dans le modèle [1] : $y_t = \beta_0 + \beta_1 x_t + \varepsilon_t$, lorsque les résidus sont engendrés par un processus AR-1, les variables peuvent être transformées comme suit :

$$u_t = x_t - \rho x_{t-1} \ , \quad v_t = y_t - \rho y_{t-1} \ , \quad \omega_t = \varepsilon_t - \rho \varepsilon_{t-1}$$

et le modèle [2] dont les résidus ω_t sont des variables aléatoires i.i.d. $N\left(0,\ (1-\rho^2)\sigma_\varepsilon^2\right)$ s'écrit :

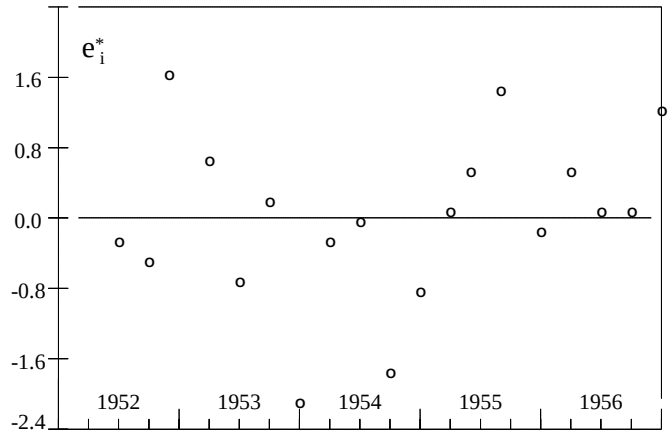
$$v_t = \beta_0(1-\rho) + \beta_1 u_t + \omega_t \quad , \quad t = 1, \ldots, n \qquad [2]$$

Les transformations de x en u et de y en v nécessitent de connaître ρ, qu'il faut par conséquent estimer. Plusieurs démarches sont envisageables ;

♦ **Méthode de Cochrane-Orcutt :**

(i) estimer β_0 et β_1 par les MCO pour le modèle [1] ;
(ii) estimer ρ à partir des écarts à l'ajustement effectué en *(i)* ;
(iii) ajuster le modèle [2] ($y_t - \hat{\rho} y_{t-1}$ *vs.* $x_t - \hat{\rho} x_{t-1}$), et estimer à nouveau β_0 et β_1 ;
(iv) examiner les écarts à l'ajustement effectué à l'étape *(iii)* ;

Appliquée à l'exemple considéré plus haut, cette méthode donne les résultats suivants :



La nouvelle valeur de la statistique d est égale à 1.607 ; on ne peut pas utiliser une seconde fois le test de Durbin-Watson, car une estimation de ρ intervient dans cet ajustement (étape *iii* ci-dessus) ; d constitue néanmoins un indice pour apprécier la réduction du phénomène d'autocorrélation, et considérer le résultat obtenu comme satisfaisant. Si tel n'était pas le cas, on remarque que la procédure de Cochrane-Orcutt pourrait être itérée en répétant alternativement les étapes *ii* (estimation de ρ) et *iii* (estimation de β_0 et β_1).

♦ **Estimation simultanée de** ρ, β_0 et β_1 **:**

$$\left.\begin{aligned} y_t &= \beta_0 + \beta_1 x_t + \varepsilon_t \\ \varepsilon_t &= \rho\,\varepsilon_{t-1} + \omega_t \end{aligned}\right\} \Rightarrow\; y_t - \rho y_{t-1} = \beta_0(1-\rho) + \beta_1(x_t - \rho x_{t-1}) + \omega_t$$

A la différence de la méthode de Cochrane-Orcutt, les trois paramètres sont directement identifiés par minimisation d'un critère des moindres carrés :

$$\sum_{t=2}^{n} \left(y_t - \hat{\rho} y_{t-1} - \hat{\beta}_0(1-\hat{\rho}) - \hat{\beta}_1(x_t - \hat{\rho} x_{t-1}) \right)^2 = S(\hat{\beta}_0,\, \hat{\beta}_1,\, \hat{\rho}) = \text{min!}$$

On remarque que le modèle est non linéaire, à cause des produits de paramètres $\rho\beta_0$ et $\rho\beta_1$. La recherche du minimum du critère s'effectue donc par itérations. L'écart-type de $\hat{\beta}_1$ peut être approché par :

$$\hat{SE}(\hat{\beta}_1) \approx \left(S(\hat{\beta}_0,\, \hat{\beta}_1,\, \hat{\rho}) / (n-2) \right) \Big/ \sqrt{\Sigma \left(x_t - \hat{\rho} x_{t-1} - \bar{x}(1-\hat{\rho}) \right)^2}$$

♦ **Comparaison des résultats :** on vérifie aisément qu'une autocorrélation positive entraîne une surestimation de la précision des estimateurs des MCO. Dans l'exemple traité, ignorer l'autocorrélation aboutit à diviser par 4 l'écart-type estimé de $\hat{\beta}_1$!

	Méthode d'estimation		
Paramètre	Moindres carrés ordinaires	Cochrane-Orcutt	Moindres carrés non linéaires
ρ		0.874	0.824
β_0	–154.700	–324.44	–235.509
β_1	2.300	2.758	2.753
$SE(\hat{\beta}_1)$	0.115	0.444	0.436

D'autres méthodes ont été proposées ; en règle générale, elles pallient les déficiences des MCO en présence d'un processus AR-1 des résidus ; au plan de l'efficacité, elles sont comparables lorsque $|\rho|$ est inférieur à *ca*. .8 ; si l'autocorrélation est plus forte, les méthodes d'estimation en deux étapes ou bien itératives doivent être préférées, car le défaut qui consiste à sous-estimer ρ semble moins accentué chez celles-ci.

· ***Correction d'une spécification erronée du modèle :***

Bien que la statistique de Durbin-Watson ait été spécifiquement conçue pour mettre en évidence un processus AR-1 des résidus, **il conviendra d'interpréter toute valeur "significative" de d comme révélatrice de l'existence éventuelle d'un problème d'une autre nature**, et en particulier une mauvaise définition du modèle due à un "régresseur manquant". Il peut typiquement s'agir de l'omission d'une variable explicative dont les valeurs changent en fonction du temps. Une illustration simple est proposée au chapitre 6 de l'ouvrage de S. CHATTERJEE & B. PRICE (1991, référence [5] citée dans le liminaire du présent document).

4.8. L'HYPOTHESE DE NORMALITE DES RESIDUS.

C'est sur l'hypothèse $\mathbf{e} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$ que sont fondées les inférences qui utilisent les tests t et F, ou encore la construction de régions de confiance : d'où l'importance accordée à la vérification de la normalité des résidus. Or, dans le cas des petits échantillons, cette vérification est très délicate.

En effet, les résidus sont des quantités non observables, et les tests de normalité ne peuvent porter que sur les écarts à l'ajustement. Sachant que $\mathbf{e} = (\mathbf{I}-\mathbf{H})\mathbf{e}$, il existe entre un écart donné et l'ensemble des résidus la relation :

$$e_i = \varepsilon_i - \sum_{j=1}^{n} h_{ij}\varepsilon_j$$

♦ ***Petits échantillons :***

Si le nombre de d.d.l. n–(p+1) est faible, et si certains h_{ij} possèdent des valeurs élevées, alors, dans l'égalité ci-dessus, la somme peut l'emporter sur ε_i pour déterminer la loi de e_i. Au surplus, l'effet "limite centrale" tendra à conférer à cette somme une distribution proche de la normalité, même si la loi des ε_i (de variance finie) n'est pas normale.

On désigne ce phénomène en parlant de *supernormalité des écarts à l'ajustement*. Il s'ensuit qu'appliqués aux petits échantillons, les tests de normalité donnent des résultats qui dépendent de n, p, et **H**, et dont l'interprétation est souvent douteuse.

♦ ***Grands échantillons :***

Quand n s'accroît à p fixé, alors $h_{ii} \to 0$, et ε_i domine $\sum_j h_{ij}\varepsilon_j$

Dans ce cas, les tests appliqués aux écarts fournissent une information équivalente à celle qu'ils donneraient s'ils étaient appliqués aux résidus eux-mêmes.

Techniques classiquement employées :

(i) outils graphiques ("probability plots", "rankit plots"),

(ii) tests, *e.g.*, celui de SHAPIRO & WILK.

· ***Diagnostic fondé sur l'examen du graphe des scores normaux ("rankits") :***

Soit la variable aléatoire Z de loi N(0, 1), dont la fonction de répartition est classiquement notée Φ. Et soit le modèle d'échantillonnage $Z_1, \ldots, Z_n \overset{\text{i.i.d.}}{\sim} N(0, 1)$; les valeurs de ce n-échantillon sont rangées en ordre croissant :

$$(Z_1, \ldots, Z_n) \xrightarrow[\text{d'ordre}]{\text{Statistique}} (Z_{(1)}, \ldots, Z_{(n)}), \text{ avec : } Z_{(1)} < Z_{(2)} < \ldots < Z_{(n)}$$

Par définition, on appelle **scores normaux** les n quantités $u_n(i)$ telles que :

$$\boxed{u_n(i) = E(Z_{(i)})}$$

Pour alléger l'écriture, ces n scores normaux seront notés $u_{(i)}$; leur calcul est assez compliqué, mais il existe des tables qui donnent les valeurs des $u_{(i)}$ en fonction de n. Il existe en outre des algorithmes qui fournissent d'excellentes approximations.

♦ ***Principe du diagnostic :*** pour apprécier visuellement la normalité d'une variable aléatoire X dont on possède n réalisations indépendantes x_i, on commence par classer par ordre croissant les n valeurs observées. En notant $(x_{(1)}, \ldots, x_{(n)})$ l'ensemble des observations ordonnées :

$$\text{Si } X \sim N(\mu_x, \sigma_x^2) \Rightarrow \begin{array}{c}\text{les n points de coordonnées } (u_{(i)}, x_{(i)}) \text{ sont} \\ \text{approximativement alignés : } x_{(i)} \approx \mu_x + \sigma_x u_{(i)}\end{array}$$

Habituellement, on *standardise* les x_i : $x_i \xrightarrow[\text{réduction}]{\text{centrage et}} x_i^* = (x_i - \hat{\mu}_x)/\hat{\sigma}_x$

Après standardisation puis rangement en ordre croissant, et si la loi de X est normale, alors les n points $(u_{(i)}, x^*_{(i)})$ sont situés au voisinage de la première bissectrice.

♦ ***Application aux écarts à l'ajustement d'une régression :*** les écarts e_i sont d'abord standardisés, *i.e.*,

$$e_i \rightarrow e_i^* = \frac{(e_i - \bar{e})}{\hat{SE}(e_i)}, \text{ soit : } e_i^* = \frac{e_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}$$

avec, dans le cas du modèle linéaire simple : $h_{ii} = 1/n + (x_i - \bar{x})^2 / \Sigma(x_i - \bar{x})^2$

Pour vérifier l'hypothèse de normalité des ε_i , on construit le *"rankit plot"* : les écarts standardisés ordonnés *vs.* les scores normaux associés (*vide supra*), et l'on examine l'alignement des points ainsi obtenus.

On souhaite tester H_0 : $\varepsilon_1, \ldots, \varepsilon_i, \ldots, \varepsilon_n \overset{i.i.d.}{\sim} N(0, \sigma^2)$

contre l'alternative d'une distribution non gaussienne des résidus. On commence par définir la statistique d'ordre des n écarts standardisés :

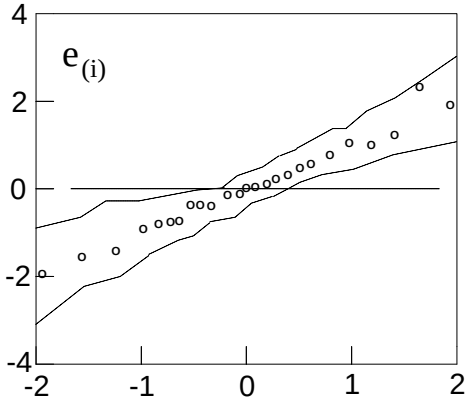
$$(e_1^*, \ldots, e_n^*) \rightarrow (e_{(1)}, \ldots, e_{(n)}) \quad \text{avec :} \quad e_{(1)} < e_{(2)} < \ldots < e_{(n)}$$

i.e., les n valeurs des écarts standardisés rangées en ordre croissant. On définit en outre les n scores normaux ("*rankits*) :

$$u_{(1)} < u_{(2)} < \ldots < u_{(n)}$$

où $u_{(i)}$ désigne l'espérance de la i-ème composante du vecteur associé à la statistique d'ordre d'un n-échantillon de N(0,1). Une approximation simple des rankits est donnée par :

$$u_{(i)} \cong \Phi^{-1}\{(i-3/8)/(n-1/4)\} \text{ , avec : } \Phi(t) = (1/\sqrt{2\pi}) \int_{-\infty}^{t} \exp(-s^2/2)\,ds$$



Scores normaux $u_{(i)}$

♦ ***Rankit plot :*** technique graphique permettant d'évaluer visuellement la plausibilité de l'hypothèse de normalité des résidus. En effet, si cette hypothèse est vraie, alors *les points de coordonnées* $(u_{(i)}, e_{(i)})$ *sont alignés sur une droite*. On retiendra que l'emploi de cette technique nécessite de l'expérience et de l'entraînement. A titre d'illustration, le graphe correspondant à l'exemple étudié aux paragraphes 4.4 à 4.6 [modèle ln(superficie) *vs.* ln(périmètre)] est représenté ci-dessus. L'enveloppe des points est un intervalle de confiance obtenu par la méthode d'ATKINSON (*Biometrika*, **68** : 13-20, 1981).

♦ ***Note sur le calcul des scores normaux :***

- Définition de la statistique d'ordre : Soit X une variable aléatoire réelle, dont la fonction de répartition est notée F, et la densité f. Et soit le modèle d'échantillonnage $X_1, \ldots, X_n \overset{i.i.d.}{\sim} F$. On appelle statistique d'ordre l'application :

$$(X_1, \ldots, X_n) \rightarrow (X_{(1)}, \ldots, X_{(n)})$$

dans laquelle les $X_{(i)}$ sont rangés en ordre croissant : $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$, les inégalités étant strictes si X est absolument continue (seul ce cas, qui exclut la possibilité d'*ex-aequo*, sera considéré ici).
N.B. : *la notation* $X_{(k)}$ *n'a de sens que pour* n *donné* (k = 1, ...,n).

- Loi de la k-ème coordonnée $X_{(k)}$ **:** On établit les résultats généraux suivants :

$$\text{Densité :} \quad f_k(x) = \frac{n!}{(n-k)!\ (k-1)!}\ F^{k-1}(x)\ [1-F(x)]^{n-k}\ f(x)$$

$$\text{Fonction de répartition :} \quad F_k(x) = \sum_{i=k}^{n} C_n^i\ F^i(x)\ [1-F(x)]^{n-i}$$

N.B. : la fonction de répartition est calculée à l'aide des tables de la loi bêta incomplète :
$F_k(x) = I_{F(x)}(k,\ n-k+1)$, *i.e.*, l'intégrale β incomplète tronquée en F(x).

- Calcul des scores normaux : D'après la définition des scores normaux, et compte tenu des résultats précédents,

$$u_n(k) = E\left(X_{(k)}\right) = \frac{n!}{(n-k)!\ (k-1)!} \int_{-\infty}^{+\infty} t\ \{\Phi(t)\}^{k-1}\ \{1-\Phi(t)\}^{n-k}\ \phi(t)\ dt$$

$$\text{avec :}\ \ \phi(t) = (1/\sqrt{2\pi})\ \exp(-t^2/2)\ \text{, et}\ \ \Phi(x) = \int_{-\infty}^{x} \phi(t)\ dt$$

Le problème est celui du calcul de l'intégrale, qui ne peut être effectué que numériquement. Deux algorithmes (*NSCOR*1 et *NSCOR*2) ont été développées par J.P. ROYSTON [1982, Algorithm AS 177, Expected Normal Order Statistics (exact and approximate), *Appl. Statist.* **31** : 161-165], qui fournissent les scores normaux avec cinq (*NSCOR*2) ou au moins sept (*NSCOR*1) décimales exactes.

N.B. : il est possible de réaliser un calcul approché des moments de la statistique d'ordre à l'aide d'un développement limité. En s'en tenant à l'espérance et au cas où F = Φ, on obtient, à l'ordre 2 :

$$E\left(X_{(k)}\right) \approx \Phi^{-1}\left(\frac{k}{n+1}\right) + \frac{k(n-k+1)}{2(n+1)^2(n+2)}\ \frac{d^2}{dt^2}\Phi^{-1}(t)\Big|_{t\ =\ k/(n+1)}$$

d'où la possibilité d'une approximation des scores normaux n'utilisant que la fonction inverse de la fonction de répartition de la loi N(0,1).

· ***Diagnostic fondé sur le test de Shapiro-Wilk :***

♦ ***Remarques préliminaires :***

(i) La motivation principale est ici la suivante : vérifier que le modèle normal décrit les données de façon satisfaisante, et constitue de ce fait une approximation raisonnable. L'ambition n'est pas d'établir que le processus probabiliste qui engendre les résidus est exactement gaussien.

(ii) Les techniques graphiques (*e.g.* "rankit plots") conduisent à des décisions qui sont entachées par un manque d'objectivité, spécialement dans les cas douteux. Il est donc recommandé de corroborer les conclusions en s'aidant d'une procédure à laquelle est attaché un risque de première espèce (*i.e.*, une probabilité de rejeter à tort l'hypotèse nulle) connu.

(iii) Une statistique de test ne livre jamais une quantité d'information comparable à celle contenue dans un graphique : les deux techniques doivent être appliquées conjointement.

♦ ***Statistique*** W ***de Shapiro-Wilk (1965) :***

La statistique W peut être présentée comme analogue au carré du coefficient de corrélation entre les points du "rankit plot", et l'hypothèse de normalité est donc repoussée pour les trop faibles valeurs de W. Il s'agit plus précisément du rapport de deux statistiques qui estiment chacune la variance de la distribution dans le cadre gaussien, et qui estiment des quantités différentes hors de ce contexte (*vide infra*). Ce test est généralement conseillé du fait de sa bonne puissance.

Le détail des calculs pour des échantillons de taille 3 à 50 figure dans :

SHAPIRO, S.S., & M.B. WILK (1965), An analysis of variance test for normality (complete samples), *Biometrika* **52**(3-4) : 591-611.

La généralisation aux grands échantillons (n de l'ordre de 2000) est due à :

ROYSTON, J.P. (1982), An extension of Shapiro and Wilk's W test for normality to large samples, *Appl. Statist.* **31**(2) : 115-124.

La traduction informatisée de la procédure peut être réalisée à partir des codes FORTRAN *AS 66, AS 111, AS 177,* et *AS 181* publiés dans *Applied Statistics.*

♦ ***Note sur le test de Shapiro-Wilk :***

Il s'agit d'un ***test de détection d'un écart à la normalité***, généralement considéré comme le meilleur "test omnibus" au sens où il possède une bonne puissance face à une vaste classe de distributions non gaussiennes. Il est en particulier optimal lorsque celles ci sont dissymétriques, ou bien symétriques mais platykurtiques. On considèrera un échantillon de n réalisations indépendantes (x_1, ..., x_n) de la variable aléatoire réelle X.

- Statistique du test : sa construction découle de l'idée suivante : sous l'hypothèse nulle $X \sim N(\mu_X, \sigma^2_X)$, la relation entre les valeurs ordonnées $x_{(i)}$ de l'échantillon, et les scores normaux $u_{(i)}$ correspondants, est linéaire (la pente de la droite vaut σ_X). Dans son principe, le test consiste à comparer :

- le carré de la pente estimée de la régression des valeurs ordonnées $x_{(i)}$ sur les scores normaux $u_{(i)}$, qui sous H_0 estime, à une constante multiplicative près, la variance σ^2_X de la population,
- avec la somme des carrés totale, qui estime aussi, à une constante multiplicative près, la variance de la population.

La statistique du test est donc définie par : $W = b^2 / S^2$, avec : $S^2 = \sum (x_i - \bar{x})^2$

Les composantes $x_{(i)}$ de la statistique d'ordre ne sont pas indépendantes. Par conséquent, la pente de la régression des $x_{(i)}$ sur les $u_{(i)}$ est estimée par les moindres carrés généralisés (paragraphe suivant), employés lorsque la matrice de covariance des

résidus est symétrique, définie positive, mais non diagonale. En introduisant quelques simplifications, il vient :

$$b^2 = \left(\sum_{i=1}^{[n/2]} u_{(n-i+1)} \left(x_{(n-i+1)} - x_{(i)} \right) \right)^2$$

où les $u_{(i)}$ désignent les scores normaux, et [n/2] la partie entière de n/2.

Si H_0 est vraie, la loi de la statistique W ne dépend que de n (elle ne dépend ni de μ_X , ni de σ_X) ; le numérateur et le dénominateur de W estiment la même quantité.

Sous l'alternative (X non gaussienne), b^2 et S^2 estiment au contraire des quantités différentes, et des simulations ont montré que E(W) est alors plus faible que sous H_0 , et que Var(W) est plus grande.

- Base de la décision : $\text{Proba} \left\{ W < w(n, \alpha) \mid H_0 \right\} = \alpha$

On repoussera donc H_0 , avec un risque de première espèce fixé α, dès que la valeur calculée de W sera inférieure à la valeur critique w(n, α). Ces valeurs critiques ont été tabulées pour n = 3, 4, ..., 50 et α = .01, .02, .05, .10, .50, .90, .95, .98, et .99 ; la généralisation à $3 \leq n \leq 2000$ nécessite le recours à l'algorithme AS181 [*Appl. Statist.* **31** : 115-124, & 176-180 (1982)].

4.9. MODELE LINEAIRE : MOINDRES CARRES GENERALISES.

L'hypothèse de non-corrélation (ou même d'indépendance) des résidus est souvent adoptée par nécessité, car il est exceptionnel que l'on dispose d'une connaissance *a priori* de leurs variances. Cependant, si l'on sait quelle est la structure de covariance des résidus, les moindres carrés généralisés permettent d'intégrer cette information supplémentaire. Soit donc la matrice symétrique et définie positive S, telle que :

$$\text{Var}(e) = \sigma^2 S, \quad S \text{ connue}, \quad \sigma^2 > 0 \text{ pas nécessairement connue},$$

et le modèle : $\mathbf{y} = \mathbf{X}b + e$, où la matrice $\mathbf{X}$ est de dimensions $n \times (p+1)$ et de rang $p + 1$.

L'***estimateur des moindres carrés généralisés*** a pour expression :

$$\hat{b} = \left(\mathbf{X}' S^{-1} \mathbf{X} \right)^{-1} \mathbf{X}' S^{-1} \mathbf{y}$$

Avec les moindres carrés généralisés, les écarts à l'ajustement ne jouent pas tous le même rôle ; et en particulier, une importance moindre est attribuée à ceux auxquels correspondent des résidus de forte variance. L'estimateur ci-dessus peut être calculé directement, mais le problème peut aussi être transformé de façon à être résolu par les MCO, et bénéficier ainsi des résultats établis pour ces derniers.

Soit donc la $n \times n$ matrice symétrique $\mathbf{C}$, telle que : $\mathbf{C}'\mathbf{C} = \mathbf{C}\mathbf{C}' = S^{-1}$

On vérifie aisément que : $\mathrm{Var}(\mathbf{C}e) = \sigma^2 \mathbf{I}_n$,

et en prémultipliant par $\mathbf{C}$ chacun des membres de l'égalité qui définit le modèle théorique, il vient :

$$\mathbf{C}y = \mathbf{C}\mathbf{X}b + \mathbf{C}e$$

Si l'on définit : $\mathbf{z} = \mathbf{C}y$, $\mathbf{M} = \mathbf{C}\mathbf{X}$, $f = \mathbf{C}e$,

alors le modèle devient : $\mathbf{z} = \mathbf{M}b + f$, avec : $\mathrm{Var}(f) = \sigma^2 \mathbf{I}_n$,

et l'estimateur des moindres carrés généralisés se calcule comme celui des MCO, *i.e.*,

$$\hat{b} = (\mathbf{M}'\mathbf{M})^{-1}\mathbf{M}'\mathbf{z}$$

Dans le cas général, la seule difficulté numérique est celle du calcul de la matrice "racine carrée" $\mathbf{C}$. Il existe un cas particulier où ce calcul est tout spécialement simple, celui où la matrice S est diagonale : les résidus n'ont pas tous la même variance, et ils sont non corrélés. Ce problème est résolu par les moindres carrés pondérés.

4.10. MODELE LINEAIRE : MOINDRES CARRES PONDERES.

Ce critère est employé lorsque les résidus sont non corrélés, mais que leurs variances sont inégales (hétéroscédasticité). La matrice des covariances est alors diagonale, et les variances des résidus s'écrivent :

$$\mathrm{Var}(\varepsilon_i) = \sigma^2 / w_i \ , \quad w_i > 0 \, , \quad i = 1, \ldots, n$$

Dans ce cas, la matrice S est habituellement notée $\mathbf{W}^{-1}$, avec $\mathbf{W} = \mathrm{diag}(w_i)$, et l'on a :

$$S = \mathbf{W}^{-1} = \begin{pmatrix} 1/w_1 & & & \\ & \ddots & 0 & \\ & & 1/w_i & \\ & 0 & & \ddots \\ & & & 1/w_n \end{pmatrix} \Rightarrow \mathbf{C} = \begin{pmatrix} \sqrt{w_1} & & & \\ & \ddots & 0 & \\ & & \sqrt{w_i} & \\ & 0 & & \ddots \\ & & & \sqrt{w_n} \end{pmatrix}$$

Concrètement, la pondération signifie qu'*aux observations les moins "incertaines"* (i.e., *aux éléments de plus faible variance) est accordée une plus grande influence dans l'ajustement*. Par exemple :

- si les variances des résidus sont égales, et que la mesure de la réponse y_i est la moyenne de n_i réplicats,

$$\mathrm{Var}(y_i) = \sigma^2/n_i \quad \Rightarrow \quad w_i = n_i$$

- si la variance des résidus est proportionnelle aux valeurs du régresseur,

$$\mathrm{Var}(y_i) = x_i\sigma^2 \quad \Rightarrow \quad w_i = 1/x_i$$

... etc. Dans ces conditions, pour un modèle à p régresseurs, et incluant un terme constant, la $n \times (p+1)$ matrice $\mathbf{M}$ et le $n \times 1$ vecteur $\mathbf{z}$ valent respectivement :

$$\mathbf{M} = \begin{pmatrix} \sqrt{w_1} & \sqrt{w_1}x_{11} & \cdots & \sqrt{w_1}x_{1p} \\ \sqrt{w_2} & \sqrt{w_2}x_{21} & \cdots & \sqrt{w_2}x_{2p} \\ \vdots & \vdots & & \vdots \\ \sqrt{w_n} & \sqrt{w_n}x_{n1} & \cdots & \sqrt{w_n}x_{np} \end{pmatrix} \quad \mathbf{z} = \begin{pmatrix} \sqrt{w_1}y_1 \\ \sqrt{w_2}y_2 \\ \vdots \\ \sqrt{w_n}y_n \end{pmatrix}$$

Dans le cas du modèle linéaire simple (un seul régresseur, p = 1), l'estimateur des MC pondérés s'exprime, en notant $\bar{x}_w$ et $\bar{y}_w$ les moyennes empiriques pondérées :

$$\hat{b} = (\mathbf{M}'\mathbf{M})^{-1}\mathbf{M}'\mathbf{z} \overset{p=1}{\Rightarrow} \hat{\beta}_1 = \frac{\sum w_i(x_i - \bar{x}_w)(y_i - \bar{y}_w)}{\sum w_i(x_i - \bar{x}_w)^2}$$

Chapitre 5

La pratique de la régression linéaire : comment identifier et traiter les points "suspects" ou anormalement influents ?

Sommaire du chapitre 5

5.1. MODELE LINEAIRE CLASSIQUE (SIMPLE OU MULTIPLE) ; DIAGNOSTIC DES ELEMENTS INFLUENTS.

Soit le modèle $y_i = \mathbf{x}_i b + \varepsilon_i$, où le vecteur-ligne $\mathbf{x}_i$ désigne la i-ème ligne de la matrice **X** des p variables explicatives :

$$\mathbf{x}_i = (1, x_{i1}, x_{i2}, \ldots, x_{ip}) , \quad \text{où} \quad i = 1, 2, \ldots, n$$

Le doublet ($\mathbf{x}_i$, y_i) sera désormais appelé "i-ème donnée ponctuelle", ou plus simplement "i-ème élément". On notera que la matrice **X** possède p+1 colonnes, et que la première, qui correspond au paramètre β_0, ne contient que des 1.

· **Objectifs des techniques de diagnostic :**

Elles procèdent essentiellement de deux préoccupations interdépendantes :

(i) vérifier que la distribution des données est conforme aux hypothèses sur lesquelles est fondé le modèle (*i.e.*, hypothèses de Gauss-Markov, et éventuellement normalité des résidus), cadre en dehors duquel les propriétés d'optimalité des estimateurs ne sont plus garanties.

(ii) identifier en ce sens les éléments "suspects", ou tout simplement ceux qui possèdent une influence anormalement forte sur les estimations ou sur les valeurs ajustées.

· **Trois techniques de diagnostic seront examinées :**

(i) La quantification de "*l'effet de levier*" d'un élément sur l'ensemble des valeurs ajustées, technique entièrement fondée sur les proriétés de la matrice **H** (*"**hat matrix**"*), et donc de **X**, puisque $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

(ii) Les "*écarts Studentisés*", encore appelés "*t-écarts*", qui font intervenir à la fois le plan expérimental suivant lequel sont arrangées les valeurs des régresseurs (matrice **X**, *"**design matrix**"*), et les valeurs de la variable réponse.

(iii) L'évaluation des influences combinées de l'effet de levier et des écarts à l'ajustement sur les valeurs calculées de la réponse : la statistique DFITS.

5.2. DIAGNOSTIC DE "L'EFFET DE LEVIER".

Modèle théorique : $\mathbf{y} = \mathbf{X}b + e$, $E(e) = \mathbf{0}$, $Cov(e) = \sigma^2\mathbf{I}$

Modèle ajusté : $$\hat{\mathbf{y}} = \mathbf{X}\hat{b} = \mathbf{X}\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}\right] = \mathbf{H}\mathbf{y}$$

$\hat{\mathbf{y}}$ est la projection orthogonale de la réponse observée **y** sur le sous-espace de dimension p+1 engendré par les colonnes de la matrice **X**. Cette projection est caractérisée par la matrice $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, symétrique (*i.e.* **H** = **H**′) et idempotente (*i.e.* $\mathbf{H} = \mathbf{H}^2$). La matrice **H**, qui transforme le vecteur **y** en $\hat{\mathbf{y}}$, est pour cette raison appelée "matrice chapeau" (***hat matrix***).

La relation entre $\hat{y}_i$, valeur calculée en $\mathbf{x}_i$ de la variable réponse, et l'ensemble de toutes les valeurs observées y_j s'exprime donc :

$$\hat{y}_i = \sum_{j=1}^{n} h_{ij}\, y_j \quad , \text{ avec : } \quad \mathbf{H} = \left(h_{ij}\right)_{i\,=\,1,\,\ldots,\,n\;;\;j\,=\,1,\,\ldots,\,n}$$

Cela signifie qu'***une perturbation de la réponse observée*** y_j ***modifie la valeur ajustée*** $\hat{y}_i$ ***au taux*** h_{ij} . Soit $\mathbf{h}_j$ le j-ème vecteur colonne de la matrice **H** :

$$\mathbf{h}_j = (\, h_{1j},\, \ldots,\, h_{nj}\,)' \quad ; \quad \text{on montre que : } \left\| \mathbf{h}_j \right\|^2 = h_{jj}$$

Les composantes du vecteur $\mathbf{h}_j$ renseignent donc sur la façon dont les n valeurs calculées dépendent de la j-ème valeur observée (d'où l'appellation anglo-saxonne *leverage vector*). Le carré de la norme de ce vecteur, qui est précisément égal au terme diagonal h_{jj} (propriété de la matrice **H**), permet ainsi d'apprécier l'effet global de l'observation y_j. Pour cette raison, on dit de façon imagée (et quelque peu inexacte, *Cf.* plus loin) que *"**l'effet de levier" de l'observation*** y_j ***vaut*** h_{jj} **.**

Remarques : ***(i)*** Cet effet ne dépend **que** de la matrice **X** .

(ii) La conséquence pratique importante est q'une erreur affectant une observation à fort effet de levier est susceptible de contaminer les valeurs calculées en de nombreux autres points.

· ***Propriétés de la matrice*** **H** ***:***

rappelons préalablement quelles sont les composantes du modèle ajusté,

$$\begin{pmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_i \\ \vdots \\ \hat{y}_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} & \cdots & x_{1p} \\ \vdots & \vdots & & \vdots & & \vdots \\ 1 & x_{i1} & \cdots & x_{ik} & \cdots & x_{ip} \\ \vdots & \vdots & & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} & \cdots & x_{np} \end{pmatrix} \times \begin{pmatrix} \hat{\beta}_0 \\ \vdots \\ \hat{\beta}_p \end{pmatrix} = \begin{pmatrix} h_{11} & \cdots & h_{1j} & \cdots & h_{1n} \\ \vdots & & \vdots & & \vdots \\ h_{i1} & \cdots & h_{ij} & \cdots & h_{in} \\ \vdots & & \vdots & & \vdots \\ h_{n1} & \cdots & h_{nj} & \cdots & h_{nn} \end{pmatrix} \times \begin{pmatrix} y_1 \\ \vdots \\ y_j \\ \vdots \\ y_n \end{pmatrix}$$

· ***Propriétés des éléments diagonaux*** h_{ii} ***:***

	élément $\mathbf{x}_i$ *non répliqué*	*élément* $\mathbf{x}_i$ *répliqué* c *fois*
*Modèle **avec** terme constant*	$\frac{1}{n} \le h_{ii} \le 1$	$\frac{1}{n} \le h_{ii} \le \frac{1}{c}$
*Modèle **sans** terme constant*	$0 \le h_{ii} \le 1$	$0 \le h_{ii} \le \frac{1}{c}$

En outre : $\sum_{i=1}^{n} h_{ii} = \text{trace}(\mathbf{H}) = \text{rang}(\mathbf{H}) = \text{rang}(\mathbf{X}) = p+1$

c'est-à-dire que ***la valeur moyenne des éléments diagonaux est*** (p+1)/n . Si l'on note $\mathbf{X}$ la matrice formée des n vecteurs-lignes $\mathbf{x}_i$ des régresseurs centrés :

$$\mathbf{x}_i = \left(x_{i1} - \bar{x}_1, \ldots, x_{ij} - \bar{x}_j, \ldots, x_{ip} - \bar{x}_p\right)$$

Avec ces notations : $h_{ii} = (1/n) + \mathbf{x}_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i'$

en particulier, pour *le modèle linéaire simple* : $h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_j - \bar{x})^2}$

Cette dernière relation montre que, dans le cas du modèle simple (*i.e.* p=1), ***la valeur de*** h_{ii} ***croît avec le carré de l'écart entre*** x_i ***et la moyenne des valeurs du régresseur***. Dans le cas du modèle multilinéaire (p>1), les isosurfaces définies par h_{ii} = *constante* sont des (hyper)ellipsoïdes centrés sur le point $\bar{\mathbf{x}}$; h_{ii} est alors d'autant plus grand qu'il caractérise un élément $\mathbf{x}_i$ faisant un petit angle avec un vecteur propre de $\mathbf{X}'\mathbf{X}$ associé à une petite valeur propre, sauf si le produit scalaire $\mathbf{x}_i\mathbf{x}_i'$ est lui-même petit.

· ***Diagnostics utilisant la matrice*** **H** ***:***

On considère le modèle à p régresseurs, plus un terme constant β_0 ; la matrice $\mathbf{X}$ possède donc p+1 colonnes. Les vecteurs e et $\mathbf{e}$ désignent respectivement les résidus (non observables, définis sur le modèle théorique) et les écarts à l'ajustement (observables). Ces variables aléatoires sont liées par la matrice $\mathbf{H}$:

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \left[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]\mathbf{y} = (\mathbf{I} - \mathbf{H})\mathbf{y}$$

La matrice $\mathbf{I}$–$\mathbf{H}$, symétrique et idempotente, caractérise la projection sur le sous-espace orthogonal à celui engendré par les colonnes de $\mathbf{X}$, donc :

$$(\mathbf{I} - \mathbf{H})\mathbf{y} = (\mathbf{I} - \mathbf{H})(\mathbf{X}b + e) = (\mathbf{I} - \mathbf{H})e \quad \Rightarrow \quad \boxed{\mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{y} = (\mathbf{I} - \mathbf{H})e}$$

Conséquence : $E(\mathbf{e}) = \mathbf{0}$, $Cov(\mathbf{e}) = \sigma^2(\mathbf{I}-\mathbf{H})$. Et si l'on considère le i-ème écart :

$$e_i = (1-h_{ii})\,\varepsilon_i - \sum_{j\neq i}^{n} h_{ij}\,\varepsilon_j$$

Les éléments diagonaux de **H** sont tous compris entre 0 et 1 : la relation ci-dessus montre donc qu'une forte erreur attachée à y_i (*i.e.* un fort résidu ε_i) peut ne pas engendrer un grand écart e_i quand $h_{ii} \approx 1$. La conséquence pratique importante est la suivante : ***un fort écart*** e_i ***entre une valeur calculée et une valeur observée n'implique pas nécessairement que cette dernière soit suspecte ; l'amplitude de l'écart peut être due à une autre observation "aberrante"*** y_j ***, s'il lui est associée une valeur*** h_{ij} ***élevée.***

Cette remarque conduit à nuancer l'appellation imagée de "levier" ; en effet :

$$\hat{y}_i = h_{ii}y_i + \sum_{j\neq i}^{n} h_{ij}\,y_j \;,\quad h_{ii} \in [0,\,1] \;,\text{ et } \sum_j h_{ij} = \sum_i h_{ij} = 1$$

Lorsque $h_{ii} \approx 1$, alors $\hat{y}_i$ tend à s'approcher de y_i (d'où le nom de levier...). Mais cette interprétation néglige le rôle éventuel des quantités aléatoires y_j , qui elles aussi contribuent à la valeur $\hat{y}_i$: pour cette raison, on préfèrera parler de ***potentiel*** plutôt que de levier.

Par ailleurs : $Var(\hat{y}_i) = \sigma^2 h_{ii}$, $Var(e_i) = \sigma^2(1-h_{ii})$, relations qui montrent qu'à un élément de potentiel élevé correspond une valeur ajustée instable (de relativement forte variance), mais un écart de faible variance (et *vice versa*).

5.3. ECARTS STANDARDISES ; ECARTS "STUDENTISES".

· ***Idée de base :*** Pour pouvoir comparer entre eux les écarts e_i, et décider que l'un quelconque d'entre eux est "aberrant", il est nécessaire de recourir à une standardisation préalable. En effet, les plus grands écarts tendent à apparaître pour les $\mathbf{x}_i$ proches de $\bar{\mathbf{x}}$, car $Var(e_i)$ croît lorsque h_{ii} diminue. La standardisation la plus naturelle est suggérée par la relation $\mathbf{e} = (\mathbf{I}-\mathbf{H})\mathbf{e}$, qui entraîne $Cov(\mathbf{e}) = (\mathbf{I}-\mathbf{H})\sigma^2$; d'où :

$$e_i^* = \frac{y_i - \hat{y}_i}{\hat{\sigma}\sqrt{1-h_{ii}}} \quad , \quad \text{où} \quad \hat{\sigma}^2 = \mathbf{e}'\,\mathbf{e} \,/\, (n-p-1)$$

Si le modèle est correct, ces ***écarts standardisés*** (*"standardized residuals", "internally Studentized residuals"*) sont d'espérance nulle et possèdent ***une même variance égale à*** 1, indépendante de σ^2 et des h_{ii}. De plus :

$$\mathbf{e} \sim N(\mathbf{0},\,\sigma^2\mathbf{I}) \;\Rightarrow\; (e_i^*)^2 \,/\, (n-p-1) \;\sim\; \text{loi bêta de paramètres 1/2 et } (n-p-2)/2$$

La diffusion confidentielle des tables de la loi β a cependant incité à rechercher une autre forme de standardisation des écarts à l'ajustement. Notons :

$(\mathbf{X}_{(-i)}, \mathbf{y}_{(-i)})$ l'ensemble des données ***moins*** l'élément $(\mathbf{x}_i, y_i)$, et,

$\hat{b}_{(-i)} = (\mathbf{X}'_{(-i)}\mathbf{X}_{(-i)})^{-1}\mathbf{X}'_{(-i)}\mathbf{y}_{(-i)}$ l'estimateur MCO correspondant

On calcule alors l'écart : $e_{(-i)} = y_i - \tilde{y}_i$, où : $\tilde{y}_i = \mathbf{x}_i \hat{b}_{(-i)}$

On notera que y_i et $\tilde{y}_i$ sont indépendants, le premier n'étant pas utilisé pour calculer le second ; par ailleurs :

$$\mathrm{Var}(y_i - \tilde{y}_i) = \sigma^2\left[1 + \mathbf{x}_i(\mathbf{X}'_{(-i)}\mathbf{X}_{(-i)})^{-1}\mathbf{x}'_i\right]$$

que l'on estime en remplaçant σ^2 par $\hat{\sigma}^2_{(-i)}$, calculée en utilisant la relation :

$$(n-p-2)\hat{\sigma}^2_{(-i)} = (n-p-1)\hat{\sigma}^2 - e_i^2/(1-h_{ii})$$

D'où la définition des "***écarts Studentisés***" (*"externally Studentized residuals", "t-residuals", "jackknifed residuals", "cross-validatory residuals"*) :

$$e^*_{(-i)} = \frac{y_i - \hat{y}_i}{\hat{\sigma}_{(-i)}\sqrt{1-h_{ii}}}$$

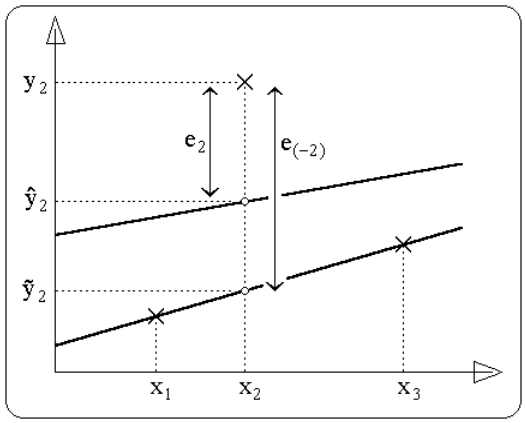
· ***Ecart attaché au retrait d'une observation ("predicted residual") :***

La figure ci-contre représente la droite ajustée à : $\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \end{pmatrix}$, $\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}$,

et aussi celle obtenue pour :

$\mathbf{X}_{(-2)} = \begin{pmatrix} 1 & x_1 \\ 1 & x_3 \end{pmatrix}$, $\mathbf{y}_{(-2)} = \begin{pmatrix} y_1 \\ y_3 \end{pmatrix}$.

Les deux écarts $y_2 - \hat{y}_2$ et $y_2 - \tilde{y}_2 = e_{(-2)}$ ("*predicted residual*") sont aussi représentés.



Remarque : les estimations des paramètres de la régression ajustée au données dont on a enlevé l'élément i se déduisent aisément des estimations obtenues pour l'échantillon complet ; en effet, si l'on note $\mathbf{x}_i$ le i-ème vecteur-ligne de la matrice $\mathbf{X}$, et $\mathbf{X}_{(-i)}$ la matrice $\mathbf{X}$ *moins* la ligne $\mathbf{x}_i$:

$$\left(\mathbf{X}'_{(-i)}\mathbf{X}_{(-i)}\right)^{-1} = \left(\mathbf{X}'\mathbf{X}\right)^{-1} + \frac{\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{x}'_i\,\mathbf{x}_i\left(\mathbf{X}'\mathbf{X}\right)^{-1}}{1-h_{ii}}$$

Cette remarquable équation fut employée dès 1821 par Gauss. Dans le cadre de la régression linéaire, elle est à la base de toutes les relations entre résultats obtenus soit en incluant, soit en excluant le i-ème élément.

· ***Diagnostics utilisant les écarts "Studentisés" :***

L'écart Studentisé $e^*_{(-i)}$ est une statistique sensible à une réponse y_i aberrante au sens où elle n'est pas conforme au modèle postulé, *e.g.* : $y_i = \mathbf{x}_i b + \delta + \varepsilon_i$

C'est-à-dire que pour le i-ème élément on a : $E(y_i) = \mathbf{x}_i b + \delta$.

L'écart $e^*_{(-i)}$ peut être utilisé pour tester $H_0 : \delta = 0$, contre $H_1 : \delta \neq 0$. En effet, ***les écarts Studentisés suivent une loi de "Student" sous*** H_0 ***:***

$$e^*_{(-i)} = \frac{e_i \big/ \left(\sigma\sqrt{1-h_{ii}}\right)}{\hat{\sigma}_{(-i)} \big/ \sigma} \sim \frac{N(0,1)}{\sqrt{\chi^2_{n-p-2} \big/ (n-p-2)}} \sim t_{n-p-2}$$

à condition que $e \sim N(\mathbf{0}, \sigma^2\mathbf{I})$.

Les écarts Studentisés fournissent donc *un moyen de détection des réponses aberrantes ("outliers").* Le test ne doit cependant pas être employé sans précaution.

♦ **<u>Premier cas</u> : On suspecte *a priori* que le i-ème élément est aberrant** ; dans ce cas, la décision s'appuie sur la comparaison de :

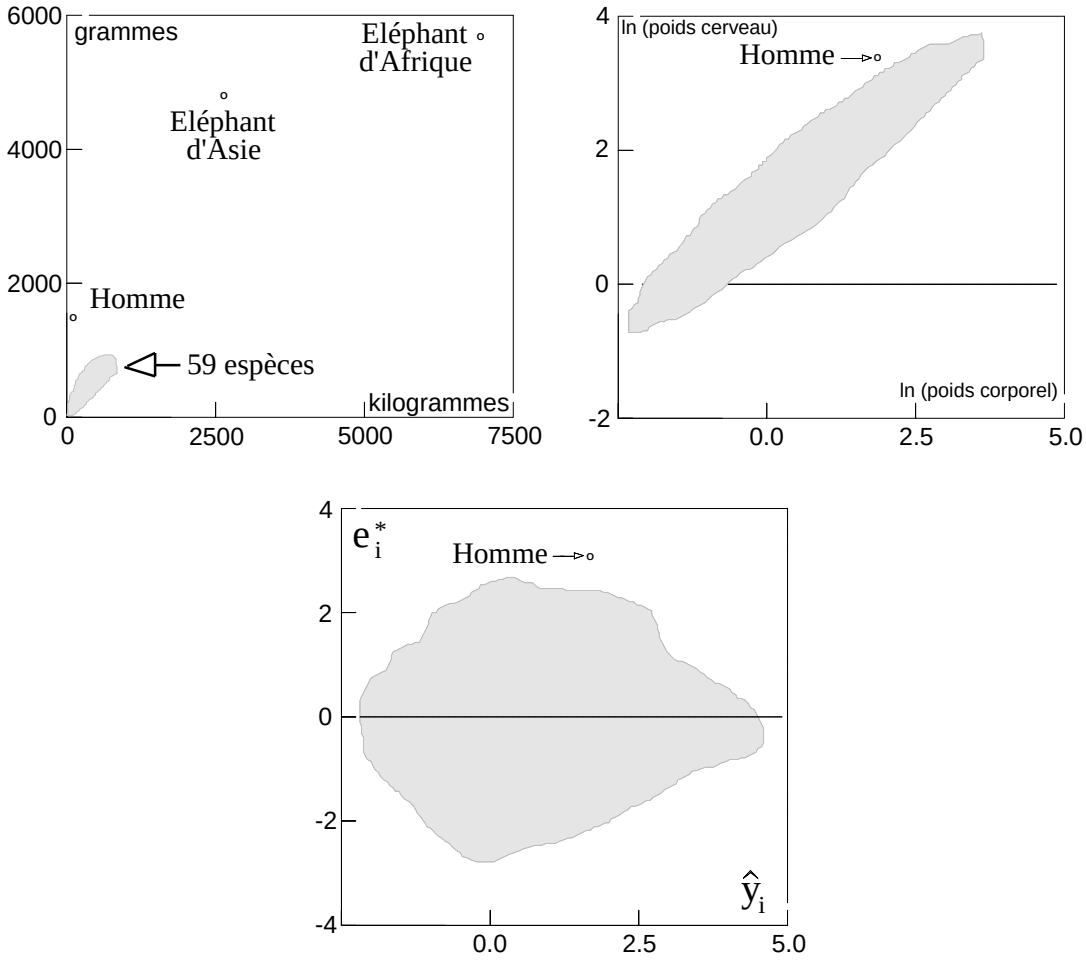
$$\text{Proba}\left\{ \left| e^*_{(-i)} \right| > t_{\nu;\, \alpha/2} \right\}$$

au risque de première espèce α.

<u>Exemple de test de conformité au modèle d'un élément *a priori* suspect</u> :

L'exemple qui suit est dû à S. WEISBERG (1985, référence [18] dans le liminaire du présent document). Il s'agit d'appliquer le modèle d'allométrie (*Cf.* Introduction) à la description de la relation entre le poids du cerveau (en g) et le poids corporel (en kg) chez 62 espèces de Mammifères. On suppose *a priori* que le cas de l'espèce humaine n'est pas correctement décrit par la relation globale, et par conséquent que le point représentant l'Homme, "décalé" par rapport au modèle qui décrit la relation chez les autres espèces, est un "*outlier*".

Le nuage des points est représenté avant (ci-après, à gauche) et après (à droite) transformation logarithmique ; en bas apparaît le graphe des écarts standardisés *vs.* la réponse calculée.



$$e_i^* = 2.848 \ , \quad e_{(-i)}^* = 3.04 \ ; \quad \text{Proba}_{H_0} \left(e_{(-i)}^* > t_{\nu=60} \right) \approx .005$$

On souhaite conclure au seuil $\alpha = .05$: probabilité critique $< \alpha \quad \Rightarrow$ rejet de H_0 .

♦ **<u>Second cas</u> : on ne connaît pas *a priori* la (ou les) réponse(s) suspecte(s).**

Dans cette situation, on peut être tenté de répéter pour différents éléments la procédure appliquée dans le cas précédent ; l'obstacle à une telle démarche est le suivant : les $e_{(-i)}^*$ ne sont pas indépendants, et *le seuil* α *du test n'est pas conservé dans cet ensemble de plusieurs comparaisons.*

Le palliatif habituel consiste à ne comparer à une valeur-seuil que le plus grand des écarts. Mais alors *la valeur critique ne peut pas être celle employée dans le test t classique, qui suppose que l'écart à tester est choisi* a priori, *sans que l'on soit influencé par l'observation des autres écarts.*

Illustration : considérons l'ajustement d'un modèle avec n = 65 et p = 3, et supposons qu'aucun point ne soit aberrant. Si l'on teste l'un des 65 écarts "studentisés", choisi *a priori*, la probabilité d'obtenir $|t_{\nu=60}| > 2$ est égale à .05 ; en revanche, avec 65 tests indépendants, la probabilité que la plus grande valeur de la statistique $|t_{\nu=60}|$ excède 2 devient $1-(.95)^{65} = .964$; cet exemple illustre ***la nécessité d'une redéfinition du seuil*** (sachant qu'en outre, dans le problème traité ici, les tests ne sont pas indépendants).

Palliatif. On choisit τ tel que : $\text{Proba}\left\{\begin{matrix}\text{une valeur } \left|e^*_{(-i)}\right| \\ \text{choisie au hasard parmi n}\end{matrix} > \tau\right\} \approx \alpha$,

α fixé *a priori*. Utilisant l'**inégalité de Bonferroni** (*vide infra*), on pose :

$$\tau = t_{\nu\,;\,\alpha/(2n)}\ ,\ \nu = n-p-2$$

La probabilité que les n tests bilatéraux soient simultanément corrects est égale à $1-\alpha$ *au moins* : le risque effectif est légèrement inférieur au risque nominal α. Le raisonnement sous-jacent peut être illustré à l'aide de l'exemple simple suivant. Considérons les 2 événements E_i, i = 1, 2 : "décider *à tort* que la i-ème réponse y_i est aberrante". Alors : $\text{Proba}(E_1) = \text{Proba}(E_2) = \alpha/2$; et notons $\overline{E}_i$ l'événement : "décider *avec raison* que la i-ème réponse est aberrante". La probabilité pour que les deux décisions soient simultanément correctes s'exprime alors :

$$\text{Proba}\,(\overline{E}_1 \cap \overline{E}_2) = 1 - \text{Proba}\,(E_1 \cup E_2)$$

avec : $\text{Proba}\,(E_1 \cup E_2) = \text{Proba}(E_1) + \text{Proba}(E_2) - \text{Proba}\,(E_1 \cap E_2)$, il vient :

$$\text{Proba}\,(\overline{E}_1 \cap \overline{E}_2) = 1 - \text{Proba}(E_1) - \text{Proba}(E_2) + \text{Proba}\,(E_1 \cap E_2)$$

$$\Rightarrow \text{Proba (les 2 décisions sont correctes)} \geq 1 - \alpha/2 - \alpha/2$$

$$\geq 1 - \alpha$$

Si l'on reprend l'exemple cité plus haut (n = 65, p = 3), et que l'on cherche la valeur τ que la statistique $|t_{\nu=60}|$ dépasse avec une probabilité voisine de .05 dans 65 tests bilatéraux simultanés, on trouve : $\tau = t_{60\,;\,0.05/(2\times65)} \approx 3.53$ (au lieu de 2). En pratique, on procèdera de même pour déterminer la valeur-seuil au delà de laquelle les écarts "studentisés" seront considérés comme révélateurs de points aberrants.

5.4. EVALUATION DES EFFETS COMBINES DU PLAN D'EXPERIENCE ET DES ECARTS A L'AJUSTEMENT.

Dans la matrice $\mathbf{X}$, on individualise le i-ème vecteur-ligne $\mathbf{x}_i$:

$$\mathbf{X} = \begin{bmatrix}\mathbf{X}_{(-i)} \\ \mathbf{x}_i\end{bmatrix} \qquad \text{N.B. : } \mathbf{X}'_{(-i)}\,\mathbf{X}_{(-i)} = \mathbf{X}'\mathbf{X} - \mathbf{x}'_i\,\mathbf{x}_i$$

De la sorte, on peut évaluer la ***contribution à l'ajustement*** de l'élément ($\mathbf{x}_i$, y_i) :

$$\hat{b} - \hat{b}_{(-i)} = \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}'_i\, e_i}{1-h_{ii}} \qquad \hat{y}_i - \mathbf{x}_i \hat{b}_{(-i)} = \hat{y}_i - \tilde{y}_i = \frac{h_{ii}}{1-h_{ii}} e_i$$

La j-ème composante du vecteur défini par l'équation ci-dessus (à gauche) est souvent désignée par $DFBETA_{ij}$. Quant à la contribution à l'ajustement exprimée par l'équation de droite, on la standardise par l'écart-type estimé de la i-ème réponse :

$$\hat{SE}(\hat{y}_i) = \hat{\sigma}_{(-i)}\sqrt{h_{ii}}$$

définissant ainsi la statistique DFITS (*"Stantardized Difference of Fitted values"*) :

$$DFITS_i = \frac{\hat{y}_i - \tilde{y}_i}{\hat{\sigma}_{(-i)}\sqrt{h_{ii}}} = e^*_{(-i)}\sqrt{\frac{h_{ii}}{1-h_{ii}}}$$

· ***Règle pratique d'utilisation :***

La moyenne des h_{ii} valant (p+1)/n, et les $\left|e^*_{(-i)}\right| > 2$ étant considérés comme "significatifs" au seuil α = .05, il est classiquement recommandé d'examiner attentivement les éléments ($\mathbf{x}_i$, y_i) pour lesquels :

$$\left|DFITS_i\right| > 2\sqrt{\frac{(p+1)/n}{1-(p+1)/n}} = 2\sqrt{\frac{p+1}{n-p+1}}$$

Remarque : selon certains auteurs, le coefficient 2 engendre des bornes "trop optimistes", au sens où des points douteux sont souvent encadrés par ± DFITS. On recommande donc parfois d'utiliser plutôt une valeur de 1.5 dans les applications.

5.5. OUTILS DE DIAGNOSTIC ET ROBUSTESSE.

· ***Introduction :***

Un modèle statistique est dit *paramétrique* quand il inclut une loi appartenant à une famille caractérisée par un petit nombre de paramètres réels inconnus. Cette loi de probabilité est choisie pour décrire le processus stochastique qui engendre les données observées. En statistique "classique", il s'agit de la loi normale.

Pour tout modèle, qui n'est qu'une représentation simplifiée de phénomènes réels, se pose la question de la validité des résultats qu'il produit, dès l'instant que les hypothèses sur lesquelles il est fondé ne sont que "partiellement compatibles" avec les données expérimentales. Une

idée qui a longtemps prévalu fut celle d'une modeste dégradation des qualités des procédures statistiques standard quand le modèle postulé n'est qu'approximativement exact. Mais en fait, la plupart des méthodes se sont révélées sensibles à ce type de déviation : c'est en particulier le cas des estimateurs des moindres carrés des paramètres de la régression linéaire, qui peuvent être déstabilisés par une seule observation discordante [par exemple, une observation qui serait incompatible avec le modèle $\varepsilon_1, ..., \varepsilon_n$ *iid* $N(0, \sigma^2)$].

La relative fragilité de la régression classique (*i.e.*, résidus normaux, moindres carrés) a motivé deux orientations de recherche. L'une d'elles est l'élaboration de méthodes visant à déceler la présence des points douteux, ou plus généralement à attirer l'attention de l'analyste sur les observations qui possèdent une forte influence sur l'estimation. Plusieurs de ces outils de diagnostic ont été présentés précédemment. L'autre voie est celle de la régression "robuste" : **en statistique, une méthode est dite robuste si ses performances ne sont que peu altérées par un non-respect modéré du modèle probabiliste qu'elle inclut dans sa définition**. Le concept de robustesse renvoie donc à deux questions : *(i)* celle de la "violation tolérable" du modèle théorique adopté pour décrire la variabilité des données, et *(ii)*, celle du choix des performances qui sont le moins affectées par cet écart à la loi prise comme hypothèse du modèle.

Les techniques de diagnostic des points influents et la régression robuste visent le même objectif (se prémunir contre les dégâts provoqués par les éléments aberrants presque toujours présents dans les bases de données), mais elles procèdent différemment. Les premières servent à détecter les anomalies qui apparaissent lorsque les données sont ajustées à l'aide des méthodes classiques. Elles sont utilisées pour l'examen critique des résultats, préalablement, s'il en est besoin, à un nouvel ajustement à des données "propres". En revanche, **la régression robuste est construite pour résister à l'influence d'éventuels points suspects, et elle vise à "n'ajuster qu'une majorité des observations" (celles qui sont en accord avec le modèle paramétrique postulé), tout en réduisant, voire en annulant, l'effet sur l'ajustement des observations non conformes à ce même modèle**. Ces dernières seront alors en général révélées par leur éloignement vis-à-vis de la droite (du plan, ...) robuste.

D'un point de vue pratique, l'utilisateur dispose donc de deux familles de méthodes complémentaires, appelées à être employées conjointement. Dans ce chapitre, il sera désormais question des méthodes robustes appliquées à la régression linéaire.

· ***Notions utiles pour la régression robuste :***

Une présentation très générale du problème de l'estimation est la suivante : estimer un paramètre inconnu q, $q \in \mathbb{R}^p$, à partir d'un échantillon de n réalisations indépendantes d'une variable aléatoire réelle X, dont la fonction de répartition (f. r.) est notée F. Dans de nombreux cas, q est associé à la fonction de répartition inconnue F par une fonctionnelle, notée T, et définie sur un ensemble de lois de probabilité ; q peut alors s'écrire : $q = T(F)$. Si l'on note $\hat{F}_n$ la f. r. empirique obtenue à partir du n-échantillon de X, l'***estimateur naturel*** de q est simplement $\hat{q} = T(\hat{F}_n)$. Ces définitions seront précisées plus loin.

Comme en analyse des fonctions, l'étude du comportement de la fonctionnelle T au voisinage de la loi F repose sur un développement de Taylor de $T(\hat{F}_n) - T(F)$. En d'autres termes, F est considérée comme "un point" d'une famille de lois de probabilité, et l'on est conduit à étendre à cet ensemble les notions habituelles de l'analyse, *i.e.*, continuité, différentiabilité, mesures de distance entre lois de probabilité.

♦ Une définition communément admise de la robustesse fait appel à la continuité de la fonctionnelle T au voisinage de F : si la loi G est "proche de" F, on dira que T est robuste si les lois images de G et F par T sont elles aussi voisines. Selon ce point de vue, on peut établir ***un indicateur qualitatif de la robustesse : le point de rupture*** (*breakdown point*). C'est le seuil qui définit le plus grand voisinage possible de la loi F, voisinage à l'intérieur duquel tout éloignement vis-à-vis de F sera sans effets désastreux sur les propriétés statistiques de l'estimateur $\hat{q} = T(\hat{F}_n)$.

Pour estimer les paramètres du modèle linéaire, une méthode robuste possédant un point de rupture élevé consiste à minimiser la médiane des écarts quadratiques (*Least Median of Squares Regression*). Elle sera exposée à la fin de ce chapitre.

♦ Si la fonctionnelle T est non seulement continue mais aussi dérivable, il existe un autre moyen d'en apprécier la robustesse : ***la fonction d'influence***, notée IF,

$$T(\hat{F}_n) - T(F) = \frac{1}{n} \sum_{i=1}^{n} IF(x_i ; T, F) + o\left(dist(\hat{F}_n, F)\right)$$

où dist(.) est une distance entre lois de probabilités, et où o(.) désigne un terme d'ordre inférieur. L'équation ci-dessus montre que $IF(x_i ; T, F)$ représente la contribution de l'observation x_i à $T(\hat{F}_n) - T(F)$, *i.e.*, à la valeur approchée de l'erreur que l'on commet en estimant q par $\hat{q}$.

De cette définition découle l'idée suivante : pour construire un estimateur robuste, on interdira que sa fonction d'influence dépasse une valeur maximale. Cela sera illustré à l'aide du cas simple de quelques M-estimateurs de position bien connus, dont certains possèdent une fonction d'influence bornée. La fonction d'influence du modèle linéaire classique sera aussi présentée, ainsi que l'emploi des M-estimateurs dans ce contexte.

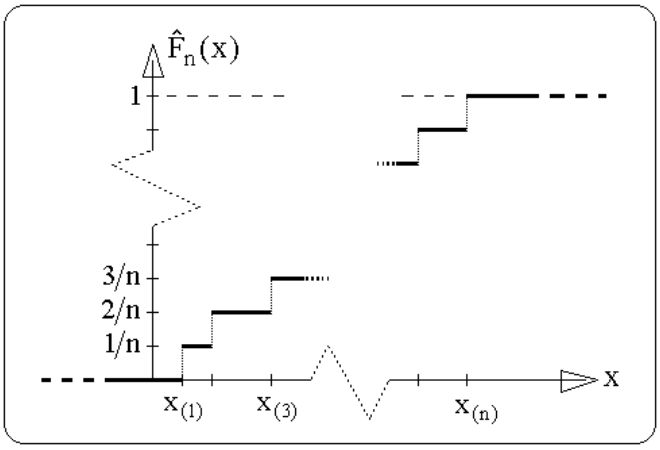
5.6. ESTIMATEUR NATUREL ; FONCTION D'INFLUENCE.

· ***Fonction de répartition empirique ("sample c.d.f.") :***

Soit : $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} F$, où F désigne la vraie fonction de répartition des X_i,

et le n-échantillon observé : $\{X_1 = x_1, \ldots, X_n = x_n\}$

Statistique d'ordre du n-échantillon : $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$; avec ces notations :



$$\hat{F}_n(x) = \begin{cases} 0 \text{ si } x < x_{(1)} \\ i/n \text{ si } x_{(i)} \leq x < x_{(i+1)} \\ 1 \text{ si } x_{(n)} \leq x \end{cases}$$

♦ **Convergences (résultats asymptotiques) :**

(i) La loi faible des grands nombres garantit la convergence des fréquences relatives vers les probabilités ; donc, en tout point x fixé :

$$\forall \varepsilon > 0, \lim_{n \to \infty} \text{Proba} \left\{ \left| \hat{F}_n(x) - F(x) \right| < \varepsilon \right\} = 1$$

(ii) Il existe un théorème plus fort (théorème de Glivenko-Cantelli), qui garantit la convergence globale en tout x :

$$\boxed{\text{Proba} \left\{ \lim_{n \to \infty} \left(\sup_{-\infty < x < +\infty} \left| \hat{F}_n(x) - F(x) \right| \right) = 0 \right\} = 1}$$

· ***Fonctionnelle statistique ; estimateur naturel :***

F : fonction de répartition dépendant d'un paramètre inconnu θ, $\theta \in \Theta$, $\Theta \subset \mathbb{R}^p$

Dans de nombreux cas, le paramètre θ peut être considéré comme une fonction associée à F par une ***fonctionnelle*** (*i.e.*, une fonction d'une fonction) que l'on notera T :

$$\{\text{ensemble des lois de probabilité sur } \boldsymbol{X}\} \xrightarrow{T} \Theta \text{, c'est-à-dire : } \theta = T(F)$$

Limitons nous au cas où $\mathcal{X} = \mathbb{R}$, *i.e.*, où X est variable aléatoire réelle continue, de loi F ; on peut alors citer quelques exemples classiques :

$$\theta \equiv \text{espérance} \ : \ T(F) \ = \ \int_{\mathbb{R}} x \, dF(x) \ = \ E_F(X)$$

$$\theta \equiv \text{variance} \ : \ T(F) \ = \ \int_{\mathbb{R}} x^2 \, dF(x) - E_F^2(X) \ = \ Var_F(X)$$

$$\theta \equiv \text{moyenne } \alpha\text{-tronquée} \ : \ T(F) \ = \ \frac{1}{1-2\alpha} \int_{\alpha}^{1-\alpha} F^{-1}(x) \, dx$$

... etc. Si $\mathcal{X} = \mathbb{R}^2$, on a aussi $\theta \equiv \rho$, coefficient de corrélation linéaire, par exemple.

♦ ***Estimateur naturel :*** Soit $\hat{\theta}$ l'estimateur naturel de θ,
calculé à partir d'un n-échantillon de $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} F$,

$$\boxed{\hat{\theta} = T(\hat{F}_n) \ , \quad \text{où} \quad \hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{]-\infty, \, x[}(x_i)}$$

avec : $\mathbf{1}_{\Omega}(x_i) = \begin{cases} 1 \text{ si } x_i \in \Omega \\ 0 \text{ si } x_i \notin \Omega \end{cases}$; on note généralement : $\hat{\theta} = \theta(\hat{F})$

Et parmi les exemples cités plus haut :

$$T(F) = E_F(X) \ ; \ T(\hat{F}_n) = \frac{1}{n} \sum_i x_i = \bar{x}$$

$$T(F) = Var_F(X) \ ; \ T(\hat{F}_n) = \frac{1}{n} \sum_i x_i^2 - \bar{x}^2 = S^2$$

· ***Fonction d'influence. Définition :***

♦ **Contaminée** $F_\tau(G)$ **:**

F, G : fonctions de répartition de deux lois de probabilité P et Q ; $\tau \in [0, 1]$;
on appelle "contaminée de F par G au taux τ" la fonction : $F_\tau(G) = (1-\tau)F + \tau G$
La loi Q est appelée "loi de probabilité contaminante".

♦ **Dérivée au sens de Gâteaux d'une fonctionnelle statistique :**

On rappelle qu'une *fonctionnelle statistique* est une statistique pouvant s'écrire comme une fonction d'une loi de probabilité. Soient donc $\theta = T(F)$, et $\hat{\theta} = T(\hat{F}_n)$ l'estimateur naturel de θ. On considèrera $T(\hat{F}_n)$ comme un "bon estimateur" du paramètre $\theta = T(F)$ si la fonctionnelle T "se comporte bien" au voisinage de F. On sait en effet que $\hat{F}_n$ converge vers F quand $n \to \infty$.

Pour caractériser plus spécialement la robustesse de $T(\hat{F}_n)$, il faut se doter d'outils permettant d'apprécier le comportement de T quand le modèle postulé F n'est pas tout à fait exact, par exemple lorsque F est contaminée par une loi G. On définit ainsi $T'_G(F)$, dérivée (au sens de Gâteaux) en F de la fonctionnelle T dans la direction de la loi contaminante G :

$$T'_G(F) = \lim_{\tau \to 0} \frac{T\big((1-\tau)F + \tau G\big) - T(F)}{\tau} = \lim_{\tau \to 0} \frac{T\big(F_\tau(G)\big) - T(F)}{\tau}$$

♦ **Fonction d'influence :**

On se place maintenant dans le cas particulier où G est la f. r. de la loi $Q = \delta_x$, *i.e.*, **la probabilité contaminante est celle qui attribue la masse** 1 **au point** x :

$$IF(x\,;\,T,\,F) = T'_{\delta_x}(F) = \lim_{\tau \to 0} \frac{T\big((1-\tau)F + \tau\,\delta_x\big) - T(F)}{\tau}$$

La fonction d'influence quantifie l'effet produit sur la fonctionnelle statistique T par l'addition d'une observation en x, quand $n \to \infty$; la fonction IF mesure donc ***la sensibilité de T à une contamination infinitésimale en*** x. Dans la pratique, on utilise la fonction d'influence empirique, approximation de IF pour des échantillons de taille finie.

5.7. FONCTION D'INFLUENCE DU MODELE LINEAIRE.

La fonction d'influence répondra à un double objectif : permettre d'apprécier l'effet des éléments (**x**, y) auxquels est associé un résidu dont les valeurs sont "anormalement fortes", mais aussi l'effet des **x** de potentiel élevé. Pour pouvoir tenir compte de ce second aspect, on calcule la fonction d'influence IF *en considérant* (**x**, y) *comme un vecteur aléatoire*, même si la théorie de la régression traite les x comme des quantités fixées. Soit donc F la fonction de répartition jointe du $1 \times ((p+1)+1)$ vecteur (**x**, y). Avec ces notations :

$$E_F\left[\begin{pmatrix}\mathbf{x}'\\ y\end{pmatrix}(\mathbf{x}\ y)\right] = \begin{pmatrix} S(F) & g(F)\\ g'(F) & \tau(F)\end{pmatrix} \qquad T(F) = (S(F))^{-1} g(F)$$

où S est la $(p+1)\times(p+1)$ matrice $E(\mathbf{x}'\mathbf{x})$, g le $(p+1)\times 1$ vecteur $E(\mathbf{x}'y)$, et T la fonctionnelle statistique qui correspond à l'estimateur des MCO du vecteur b des paramètres. La fonction d'influence de T en (**x**, y) s'exprime :

$$IF\big((\mathbf{x},\,y)\,;\,T,\,F\big) = (S(F))^{-1}\mathbf{x}'\big(y - \mathbf{x}T(F)\big)\ ,\quad \text{avec :}\ T(F) = b$$

Cette expression montre qu'au point (**x**, y), la fonction d'influence inclut le produit de ***deux termes non bornés lorsque*** b ***est estimé par les moindres carrés ordinaires*** :

- le terme y–**x**b, relatif au résidu ε attaché à (**x**, y),
- le vecteur-ligne **x**, relatif aux valeurs des régresseurs.

Le modèle linéaire ajusté par les moindres carrés n'est donc pas robuste face aux éventuelles valeurs extrêmes de l'un ou l'autre de ces deux termes.

· ***Utilisation pratique de la fonction d'influence :***

L'évaluation de l'influence consiste à introduire de petites perturbations dans le modèle, puis à en examiner les conséquences sur les résultats. L'intérêt réside dans l'information apportée sur la dépendance de ces derniers vis-à-vis de certaines spécifications du modèle postulé. Cela nécessite d'effectuer trois choix : ***(i)*** la nature de la perturbation, ***(ii)*** le type de résultat auquel on va s'intéresser, et ***(iii)*** la mesure de l'effet produit. On utilisera à cette fin la fonction d'influence empirique, c'est-à-dire celle calculée en remplaçant dans l'équation précédente :

$$F \text{ par : } \hat{F}_n \text{ , et } T(F) \text{ par : } \hat{b} = T(\hat{F}_n)$$

d'où :

$$\boxed{\hat{IF}\left((\mathbf{x},\ y)\ ;\ \hat{b},\ \hat{F}_n\right) = n(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}'\ (y - \mathbf{x}\hat{b})}$$

♦ **Cas d'une perturbation infinitésimale :**

On considère un échantillon infiniment grand d'éléments (**x**, y), que l'on perturbe en lui ajoutant un nouvel élément ($\mathbf{x}_i$, y_i). Les composantes du vecteur EIC_i (*"**Empirical Influence Curve**"*) sont les taux de variation correspondants des estimateurs des paramètres du modèle :

$$EIC_i = EIC(\mathbf{x}_i,\ y_i) = n(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i'\ (y_i - \mathbf{x}_i\hat{b})$$

De manière équivalente, on peut définir $EIC_{(-i)}$ lorsque la perturbation est le retrait du i-ème élément :

$$EIC_{(-i)} = (n-1)(\mathbf{X}'_{(-i)}\ \mathbf{X}_{(-i)})^{-1}\mathbf{x}_i'\ \left(e_i/(1-h_{ii})^2\right)$$

♦ **Cas d'une perturbation de données expérimentales :**

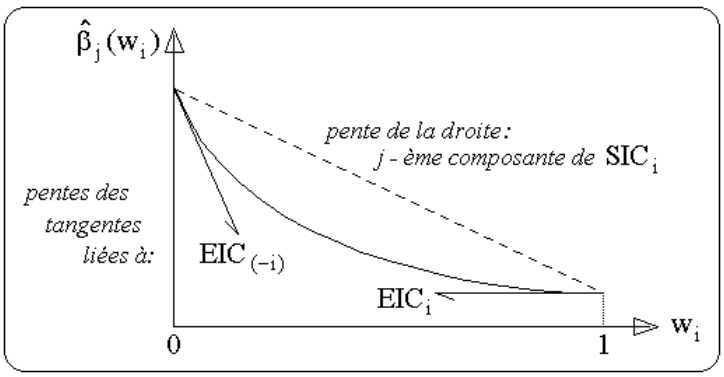
On considère cette fois un échantillon de taille finie, et l'on quantifie l'effet du retrait de l'élément i par le vecteur SIC_i (*"**Sample Influence Curve**"*), dont les composantes sont proportionnelles aux taux de variations des estimations des paramètres du modèle.

$$SIC_i = (n-1)(\hat{b} - \hat{b}_{(-i)}) = (n-1)(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i' \frac{e_i}{1-h_{ii}}$$

♦ **Interprétation :** afin de comparer entre elles ces trois formes de la fonction d'influence empirique, supposons qu'à tous les éléments, sauf au i-ème, est attaché un résidu de variance σ^2. L'élément $(\mathbf{x}_i, y_i)$ possède un résidu de variance σ^2/w_i , $w_i > 0$. L'estimateur des MC pondérés vaut :

$$\hat{b}(w_i) = \hat{b} - \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i'(1-w_i)e_i}{1-(1-w_i)h_{ii}} \quad , \quad \text{et :} \quad \frac{\partial}{\partial w_i}\hat{b}(w_i) = \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i' e_i}{\left(1-(1-w_i)h_{ii}\right)^2}$$

Notons $D\hat{b}(w_i)$ la dérivée par rapport à w_i ; avec cette notation :



$$EIC_i = n\, D\hat{b}(1)$$

$$EIC_{(-i)} = (n-1)\lim_{w_i \to 0} D\hat{b}(w_i)$$

$$SIC_i = (n-1)\int_0^1 D\hat{b}(w_i)\, dw_i$$

· ***Influence normée :***

On considèrera l'influence empirique SIC_i : $SIC_i = (n-1)(\hat{b} - \hat{b}_{(-i)})$

Le problème est de définir une norme, *i.e.*, une application : $SIC_i \in \mathbb{R}^{p+1} \to \mathbb{R}$, qui permettra de ***hiérarchiser les éléments*** $(\mathbf{x}_i, y_i)$ ***en fonction de leur influence***.

On définit donc $D_i(\mathbf{M}, s) \in \mathbb{R}$, où $\mathbf{M}$ est une (p+1)×(p+1) matrice symétrique (semi)-définie positive, et où la constante réelle *s* désigne le facteur d'échelle :

$$D_i(\mathbf{M}, s) = \frac{(n-1)^2}{s}(SIC_i)'\mathbf{M}(SIC_i) = \frac{1}{s}(\hat{b} - \hat{b}_{(-i)})'\mathbf{M}(\hat{b} - \hat{b}_{(-i)})$$

L'ensemble des isovaleurs $D_i(\mathbf{M}, s)$ = *constante* définit un ellipsoïde de dimension égale au rang de **M**. Par ailleurs, si l'on fait apparaître les écarts standardisés :

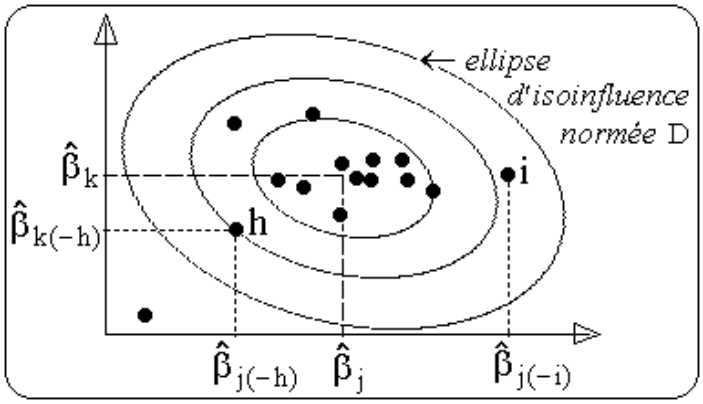
$$D_i(\mathbf{M}, s) = \frac{\hat{\sigma}^2}{s} \frac{\mathbf{x}_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{M}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i'}{1-h_{ii}} (e_i^*)^2 \quad , \quad \text{où :} \quad e_i^* = \frac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$$

Il existe plusieurs variantes de cette norme, qui correspondent à différents choix de **M** et de *s*. Mentionnons deux distances classiques, pour lesquelles $\mathbf{M} = \mathbf{X}'\mathbf{X}$:

$D_i(\mathbf{X}'\mathbf{X}, \hat{\sigma}^2_{(-i)}) = (DFITS_i)^2$, carré de la statistique déjà rencontrée (paragraphe 5.4.),

$D_i(\mathbf{X}'\mathbf{X}, (p+1)\hat{\sigma}^2)$ = "distance de COOK", notée D_i.

La figure ci-contre montre des ellipses d'isovaleurs (de la distance de Cook, par exemple) dans l'espace de deux estimateurs. Chaque point a pour coordonnées les estimations obtenues quand l'élément correspondant $(\mathbf{x}_i, y_i)$ est enlevé. Les éléments les plus influents se situent à la périphérie.



♦ **Remarques :**

(i) Calcul de la distance de COOK : $D_i = \frac{1}{p+1}(e_i^*)^2 \frac{h_{ii}}{1-h_{ii}}$

(ii) Le point i est considéré comme "influent" si $D_i > 1$.

(iii) La distance de COOK est induite par la métrique $\mathbf{X}'\mathbf{X}$ dans l'espace $\mathbb{R}^{p+1}$ des paramètres, et par la métrique $\mathbf{I}_n$ dans l'espace $\mathbb{R}^n$ des réponses prévues :

Distance dans l'espace paramétrique :	$\mathbf{X'IX=M}$	*dans l'espace des réponses :*
$\left((p+1)\hat{\sigma}^2\right) D_i = (\hat{b} - \hat{b}_{(-i)})'(\mathbf{X}'\mathbf{X})(\hat{b} - \hat{b}_{(-i)})$	$\leftrightarrow$	$= (\hat{\mathbf{y}} - \hat{\mathbf{y}}_{(-i)})'(\hat{\mathbf{y}} - \hat{\mathbf{y}}_{(-i)})$

5.8. NOTION DE M-ESTIMATEUR.

· ***Définition :*** on appelle M-estimateur (*"Maximum likelihood type estimator"*) $\hat{\theta}$ de θ l'estimateur solution de :

$$\sum_{i=1}^{n} \rho\left(x_i ; \hat{\theta}\right) = \min !$$

où ρ est une fonction arbitraire non constante. L'estimateur classique du maximum de vraisemblance correspond à : $\rho(x ; \theta) = -\ln f(x ; \theta)$; on note que $\hat{\theta}$ est aussi solution de :

$$\sum_{i=1}^{n} \psi\left(x_i ; \hat{\theta}\right) = 0 \ , \quad \text{avec :} \quad \psi(u) = \frac{\partial}{\partial u} \rho(u) \ , \quad u \in \mathbb{R}$$

N.B. : $\hat{\theta} = \theta(\hat{F}_n)$, la fonctionnelle θ étant définie par : $\int \psi(x ; \theta(F))\, dF(x) = 0$

♦ **Cas des estimateurs de position** (*e.g.*, moyenne, médiane, ...) **:**

$\hat{\theta}$ est alors solution de : $\sum \rho\left(x_i - \hat{\theta}\right) = \min !$, *i.e.*, $\sum \psi\left(x_i - \hat{\theta}\right) = 0$

· ***Fonction d'influence d'un M-estimateur :*** de façon imagée et abusive, on peut présenter la fonction d'influence IF comme l'outil qui permet d'apprécier l'impact, sur l'estimation, d'une observation x supplémentaire. La définition rigoureuse est beaucoup plus abstraite, et fait appel à la notion de contamination infinitésimale en x (voir plus haut, § 5.6.). Pour les M-estimateurs, on établit le résultat important suivant :

$$IF(x ; \theta, F) = \text{cste} \times \psi(x ; \theta(F))$$

où la constante vaut : $\text{cste} = -\left[\int (\partial/\partial\theta)\, \psi(x ; \theta(F))\, F(dx)\right]^{-1}$

Si $\hat{\theta}$ est un estimateur de position : $IF(x ; \theta, F) = \text{cste} \times \psi(x - \theta(F))$

♦ M-estimateurs de position : exemples.

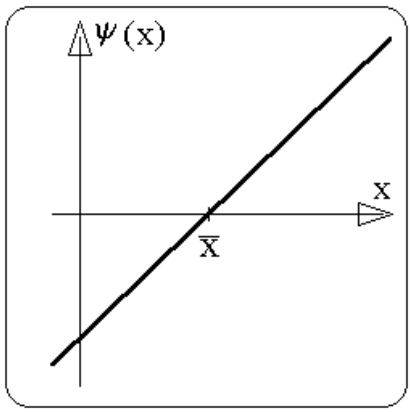
Les trois exemples qui suivent illustrent comment la fonction d'influence IF (ou, de manière équivalente, la fonction ψ) révèle la plus ou moins forte "résistance" d'un estimateur vis-à-vis des valeurs extrêmes.

Moyenne arithmétique : c'est l'estimateur du maximum de vraisemblance pour F normale, donc :

$$\rho(u) = u^2/2 + \text{cste}, \text{ et } \psi(u) = u$$

$$\Sigma\left(x_i - \hat{\theta}\right) = 0 \Rightarrow \hat{\theta} = \bar{x}$$

La fonction ψ n'est pas bornée, et donc IF ne l'est pas non plus : l'influence d'une observation "extrême" n'est donc elle-même pas bornée.

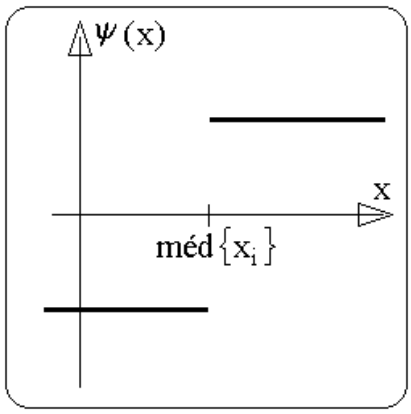


Médiane : c'est l'estimateur du maximum de vraisemblance si F est exponentielle double.

$$\rho(u) = |u| + \text{cste}, \text{ et } \psi(u) = \text{sign}(u)$$

$$\Sigma\left|x_i - \hat{\theta}\right| = \min ! \Rightarrow \hat{\theta} = \text{méd}\left\{x_1, \ldots, x_n\right\}$$

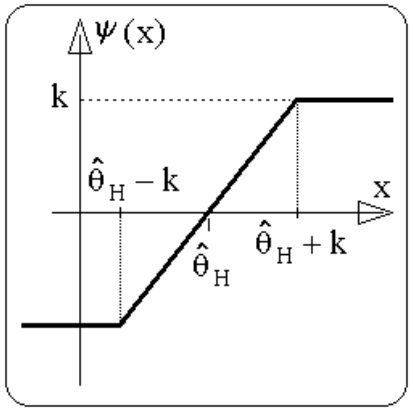
La fonction ψ est bornée : l'influence d'une observation est constante en valeur absolue, quelle que soit sa distance à la médiane.



M-estimateur de Huber : construit, du point de vue théorique, pour obtenir une variance asymptotique minimale lorsque F est "de type gaussien" en son milieu, et "alourdie" (*e.g.*, de type exponentiel) aux extrémités. **La fonction ψ est bornée** :

$$\psi(u) = \max\{-k, \min(u, k)\}$$

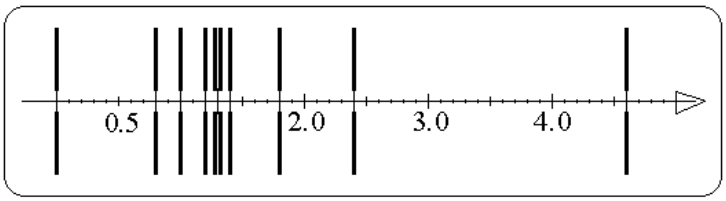
La solution $\hat{\theta}_H$ de $\Sigma\,\psi\left(x_i - \hat{\theta}_H\right) = 0$ est atteinte à l'aide d'une méthode itérative.



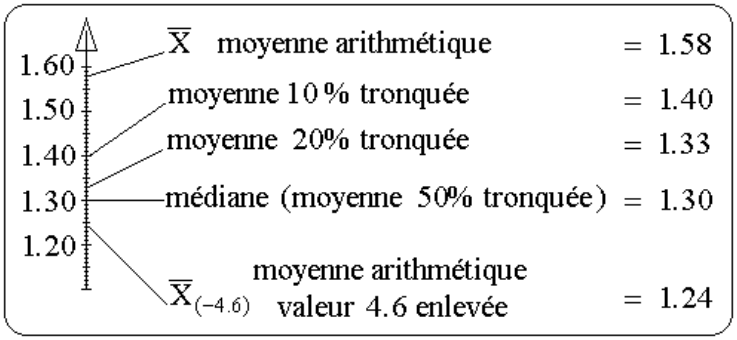
♦ Fonctions d'influence empiriques de trois M-estimateurs de position :

On a représenté ci-après les estimations obtenues avec trois estimateurs (moyenne arithmétique, moyenne 10% ou 20% tronquée, médiane) de la valeur centrale d'un échantillon de 10 observations x_i (parmi lesquelles la dernière pourrait être un "*outlier*"). La figure du bas montre les fonctions d'influence empiriques, qui révèlent comment sont modifiées les estimations lorsque l'observation x_{10} varie, les neuf autres étant fixées.

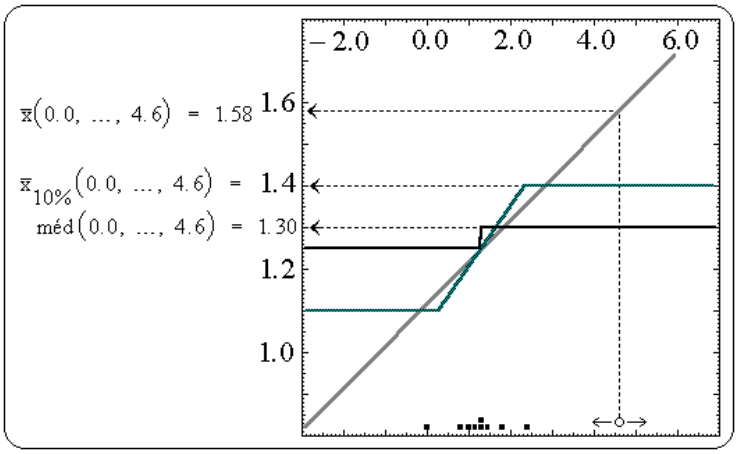
Observations : { 0.0, 0.8, 1.0, 1.2, 1.3, 1.3, 1.4, 1.8, 2.4, 4.6 }



Résultats obtenus avec différents estimateurs de position :



Fonctions d'influence empiriques :



5.9. REGRESSION ROBUSTE FONDEE SUR LES M-ESTIMATEURS.

N.B. : dans ce chapitre a été choisie une présentation très simplifiée ; on n'y évoquera pas les difficiles questions théoriques soulevées par la régression robuste (*Cf.* par exemple le chapitre 6 de l'ouvrage de F.R. HAMPEL *et al.* : *Robust Statistics. The approach based on influence functions*, J. Wiley & Sons, 1986).

· ***Modèle linéaire simple*** (p = 1) ***:***

Les équations normales s'expriment désormais :

$$\sum_{i=1}^{n} \psi\left(e_i/\hat{\sigma}\right) = 0 \ , \quad \sum_{i=1}^{n} x_i \, \psi\left(e_i/\hat{\sigma}\right) = 0 \ , \quad e_i = y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)$$

Si l'on choisit pour ψ la fonction : $\psi(u) = u.\min\{1,\ k/|\,u\,|\}$, avec : $u = e_i/\hat{\sigma}$, il est alors équivalent d'écrire le système des équations normales en faisant apparaître les pondérateurs $w_i = w(e_i)$:

$$\begin{cases} \sum_{i=1}^{n} \left\{ w(e_i) \, \dfrac{y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)}{\hat{\sigma}} \right\} = 0 \\ \sum_{i=1}^{n} \left\{ x_i \, w(e_i) \, \dfrac{y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)}{\hat{\sigma}} \right\} = 0 \end{cases} \quad , \quad w(e_i) = \min\left\{ 1,\ k\hat{\sigma}/|\,e_i\,| \right\}$$

D'autres choix sont possibles pour la fonction ψ, mais toujours prévaut l'idée directrice de borner l'influence des valeurs "extrêmes". Ici, les pondérateurs $w(e_i)$ ont pour effet de réduire l'impact des points (x_i, y_i) auxquels est associé un fort écart e_i : il leur est attribué un poids $w_i < 1$ dans l'ajustement. Les points "peu éloignés" de la droite possèdent au contraire un poids w_i égal à 1. Formellement :

$$\psi(e_i/\hat{\sigma}) = w(e_i) \, \frac{e_i}{\hat{\sigma}} = \begin{cases} e_i/\hat{\sigma} & \text{si } |e_i| \le k\hat{\sigma} \\ k \, \text{sign}(e_i) & \text{si } |e_i| \ge k\hat{\sigma} \end{cases}$$

Le système des équations normales ne possède pas de solution analytique, et il doit être résolu par itérations (le principe de l'algorithme de repondération itérative des moindres carrés est donné plus loin). L'estimateur robuste du paramètre d'échelle σ est souvent la médiane des écarts absolus à la médiane empirique (notée MAD : *"Median Absolute Deviation"*),

$$\hat{\sigma} = 1.483 \, \text{MAD} = 1.483 \, \underset{j}{\text{méd}} \left| e_j - \underset{i}{\text{méd}} \, (e_i) \right| \qquad 1.483 \approx 1 / \Phi^{-1}(.75)$$

♦ **Remarque :**

$$\psi(e_i/\hat{\sigma}) \;=\; \frac{1}{\hat{\sigma}}\left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\right) \;\times\; \left(\min\left\{1,\ k\hat{\sigma}/|e_i|\right\}\right)$$

La fonction d'influence du M-estimateur de Huber est le produit de deux termes :	***terme non borné : pas de robustesse face aux valeurs extrêmes du régresseur*** x.	***terme borné : limitation de l'effet des grandes valeurs des écarts.***

· ***Modèle linéaire multiple*** (p > 1) ***:***

♦ **Critère d'optimalité :** $S(\mathbf{b},\ s) \;=\; \sum_{i=1}^{n} s\,\rho\left(\left(y_i - \mathbf{x}_i\mathbf{b}\right)/s\right)$

ρ désigne une fonction convexe donnée, et s la valeur courante du paramètre d'échelle σ.

♦ **Estimation des paramètres :** $e_i \;=\; y_i - \mathbf{x}_i\mathbf{b}$, $\dfrac{\partial}{\partial b_j} S(\mathbf{b},\ s) \;=\; \sum_{i=1}^{n} x_{ij}\,\psi(e_i/s)$

$i = 1, ..., n$
$j = 0, 1, ..., p$

$$\frac{\partial}{\partial s} S(\mathbf{b},\ s) \;=\; \sum_{i=1}^{n} \chi(e_i/s)$$

avec : $\psi(u) = d\rho(u)/du$, et : $\chi(u) = \rho(u) - u.\psi(u)$, $u \in \mathbb{R}$, $u = e_i/s$

Les estimateurs des paramètres β_j et σ, qui réalisent le minimum de $S(\mathbf{b}, s)$, sont les solutions du système de p+2 équations :

$$\sum_{i=1}^{n} x_{ij}\,\psi\left(\left(y_i - \mathbf{x}_i\hat{b}\right)/\hat{\sigma}\right) \;=\; 0 \ , \quad j \;=\; 0,\ 1,\ \ldots,\ p$$

$$\text{et} \quad \sum_{i=1}^{n} \chi\left(\left(y_i - \mathbf{x}_i\hat{b}\right)/\hat{\sigma}\right) \;=\; 0$$

♦ **Remarque sur le choix de la fonction ψ :** un choix classique est la fonction ψ_H de Huber. Il lui correspond :

$$\chi_H(u) \;=\; \left(\psi_H^2(u) - \text{cste}\right)/2 \ , \quad \text{avec :}\ \text{cste} \;=\; E_\Phi\left(\psi_H^2\right) \ \text{si}\ F_\varepsilon \ \text{normale.}$$

Il a été vu précédemment qu'adopter la fonction ψ_H équivaut à utiliser les MC pondérés, avec les poids w_i fonction des écarts e_i :

$$w_i \;=\; w(e_i) \;=\; \min\left\{1,\ k\hat{\sigma}/|e_i|\right\}$$

On pourrait choisir *w* fonction des écarts standardisés : $e_i^* \;=\; e_i/(\hat{\sigma}\sqrt{1-h_{ii}})$

Mais si l'on souhaite réduire l'importance des points qui "s'ajustent mal", et qui de plus ont un "effet de levier" prononcé, on utilisera plutôt :

$$w_i = w(e_i, \mathbf{x}_i) = \min \left\{ 1, \frac{k\hat{\sigma}}{|e_i|} \sqrt{\frac{1-h_{ii}}{h_{ii}}} \right\}$$

· ***Généralisation des M-estimateurs appliqués à la régression***

Supposons, pour simplifier l'exposé, que $\sigma^2 = 1$. On a vu que les M-estimateurs solutions du système :

$$\sum_{i=1}^{n} \mathbf{x}_i \, \psi(y_i - \mathbf{x}_i \hat{b}) = \mathbf{0} \quad , \quad \text{où : } \mathbf{x}_i = (1, x_{i1}, \ldots, x_{ip})$$

sont robustes vis-à-vis des fortes valeurs des écarts, mais qu'ils ne le sont pas face à d'éventuelles valeurs aberrantes des régresseurs. D'où l'idée d'élargir le domaine de définition de la fonction ψ pour y inclure $\mathbf{x}$, *i.e.* :

$$\text{remplacer } \psi(e_i), \text{ fonction de } \mathbb{R} \to \mathbb{R}, \text{ par : } \eta(\mathbf{x}_i, e_i), \text{ de } \mathbb{R}^{p+1} \times \mathbb{R} \to \mathbb{R}$$

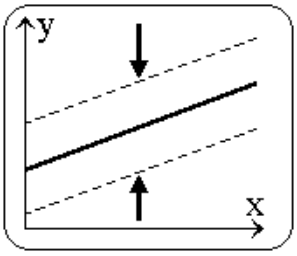
Les estimateurs sont alors solutions de : $\sum_{i=1}^{n} \mathbf{x}_i \, \eta(\mathbf{x}_i, y_i - \mathbf{x}_i \hat{b}) = \mathbf{0}$

On parle alors quelquefois de ***GM-estimateurs***. La fonction η est continue et impaire, et les différentes options qui ont été proposées sont de la forme :

$$\eta(\mathbf{x}, e) = w(\mathbf{x}).\psi(e.v(\mathbf{x})) \quad , \quad w(\mathbf{x}), v(\mathbf{x}) \text{ fonctions de } \mathbb{R}^{p+1} \to \mathbb{R}^+$$

D'où l'écriture du système d'équations à résoudre : $\sum_{i=1}^{n} \mathbf{x}_i . w(\mathbf{x}_i).\psi[\,(y_i - \mathbf{x}_i \hat{b}).v(\mathbf{x}_i)\,] = \mathbf{0}$

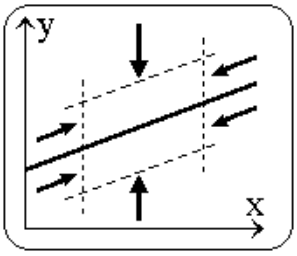
Les différents estimateurs se distinguent par la manière dont sont définis les pondérateurs w et v, et donc par les modalités de limitation de l'influence des écarts ou des régresseurs. Trois exemples sont schématisés ci-dessous :



Huber : $w(x) = 1$, $v(x) = 1$

$\sum x_i \psi(e_i) = 0$

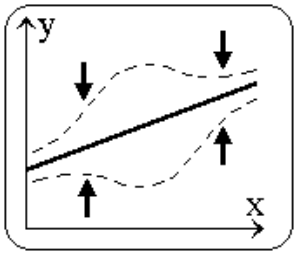
influence des x_i non bornée.



Mallows : $v(x) = 1$

$\sum x_i w_i \psi(e_i) = 0$

influence des x_i bornée, quelle que soit l'amplitude de l'écart correspondant.

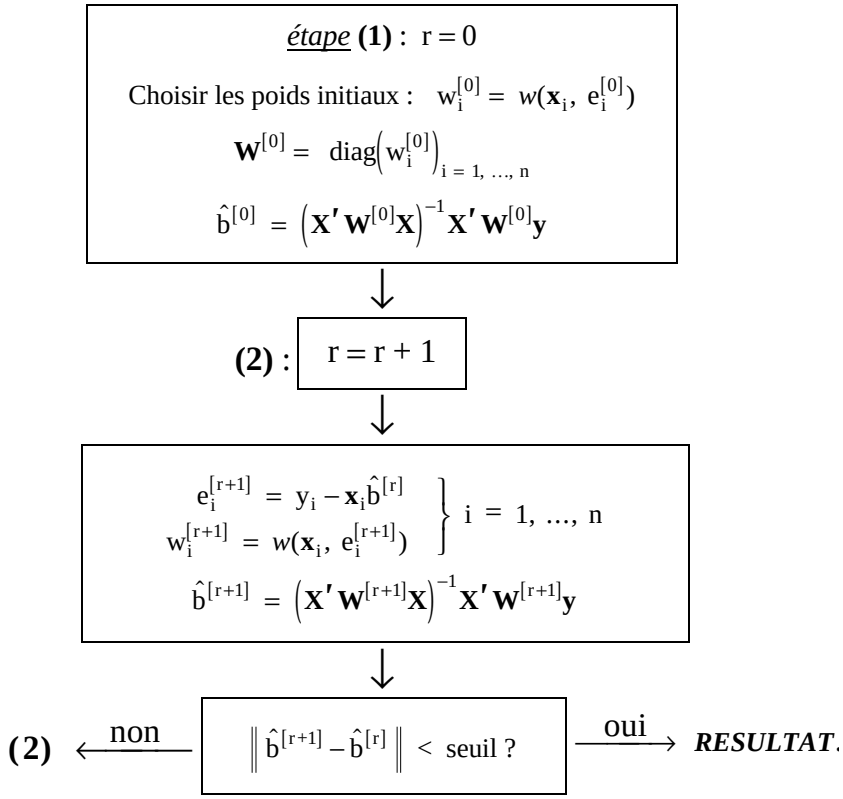


Schweppe : $v(x) = 1/w(x)$

$\sum x_i w_i \psi(e_i/w_i) = 0$

rôle de "l'effet de levier" diminué seulement si la valeur de e_i est grande.

· ***Algorithme de repondération itérative des moindres carrés ("iteratively reweighted least squares, IRLS") :***



Les MC pondérés sont traditionnellement employés lorsque les résidus sont non corrélés et de variances inégales. Plus récemment, cette méthode d'estimation a été utilisée en vue de réduire l'influence des éléments "suspects". Dans ce second cas, les poids w_i sont des fonctions prédéfinies des des données : $w_i = w(\mathbf{x}_i, e_i)$. La fonction *w* est construite de telle sorte qu'elle exprime la dépendance des poids w_i vis-à-vis du (ou des) régresseur(s) d'une part (elle est donc **sensible aux** "**effets de levier**"), et bien sûr vis-à-vis des écarts e_i d'autre part (**sensibilité éventuelle aux résidus de forte variance**).

5.10. LE POINT DE RUPTURE COMME CRITERE DE ROBUSTESSE.

· ***Définition du point de rupture pour un échantillon de taille finie :***

Considérons le n-échantillon : $\mathcal{Z}_n = \{ (\mathbf{x}_1, y_1), ..., (\mathbf{x}_i, y_i), ..., (\mathbf{x}_n, y_n) \}$,
et un estimateur T des paramètres du modèle de régression, *i.e.*, : $T(\mathcal{Z}_n) = \hat{b}$
Soit $\mathcal{A}$ l'ensemble de tous les n-échantillons "contaminés" $\tilde{\mathcal{Z}}_n$, que l'on obtient en remplaçant m quelconques des n points de $\mathcal{Z}_n$ par des points de coordonnées arbitraires (cela autorise l'apparition de valeurs totalement aberrantes).

Et soit enfin B_{max} la valeur maximale du biais de T engendré par cette contamination :

$$B_{max}(m\,;T,\boldsymbol{\mathcal{L}}_n) = \sup_{\tilde{\boldsymbol{\mathcal{L}}}_n \in \boldsymbol{\mathcal{M}}} \left\| T(\tilde{\boldsymbol{\mathcal{L}}}_n) - T(\boldsymbol{\mathcal{L}}_n) \right\| \qquad m \le n/2$$

Lorsque B_{max} devient infini, cela signifie qu'une contamination au taux m/n peut "éloigner" de $T(\boldsymbol{\mathcal{L}}_n)$ l'estimateur T d'une distance arbitrairement grande : *"the estimator breaks down"*. Pour un n-échantillon fini, le point de rupture ε_n^* de l'estimateur T est la plus petite proportion m/n de contamination tolérable :

$$\varepsilon_n^*(T,\boldsymbol{\mathcal{L}}_n) = \min\{\, m/n \text{ tel que : } B_{max}(m\,;T,\boldsymbol{\mathcal{L}}_n) \text{ infini} \,\}$$

N.B. : Il existe plusieurs manières de définir le point de rupture ; c'est la plus aisément accessible aux praticiens qui a été adoptée ici.

· ***Exemples :***

Critère d'optimalité définissant l'estimateur T des paramètres		Point de rupture n fini : ε_n^*	Point de rupture asymptotique : ε^*	T résistant vis-à-vis de valeurs aberrantes (*outliers*) des y_i	T résistant vis-à-vis de valeurs aberrantes (*outliers*) des $\mathbf{x}_i$
$\min \sum e_i^2$	norme L_2	1/n	0 %	non	non
$\min \sum \lvert e_i \rvert$	norme L_1	1/n	0 %	oui	non
$\min \sum \rho(e_i)$	M-estimateurs	1/n	0 %	oui	non

Le tableau indique la valeur du point de rupture pour trois classes d'estimateurs ; on notera la vulnérabilité des M-estimateurs aux *"leverage points"*, *i.e.*, ils ne résistent pas aux valeurs extrêmes des régresseurs. C'est pour pallier cette faiblesse qu'ont été introduits les GM-estimateurs (paragraphe précédent), parfois appelés "estimateurs à influence bornée". Mais le point de rupture de ces derniers n'atteint que *ca.* 30% au mieux, et sa valeur est une fonction décroissante du nombre p des variables explicatives.

· ***Estimateurs des paramètres du modèle linéaire robustes au sens du point de rupture :***

Il convient de rappeler que le point de rupture (dont la valeur ne peut pas dépasser 50%) est un indicateur qualitatif assez sommaire de la robustesse. S'il est souhaitable (voire nécessaire, selon certains auteurs) d'obtenir une valeur élevée de ε^*, cela ne constitue nullement une condition suffisante pour définir un "bon" estimateur. De nombreuses méthodes ont été proposées depuis près d'un demi-siècle, pour la plupart desquelles ε^* n'atteint pas 30%. Trois exemples seulement sont présentés ci-après.

♦ **Exemples pour le modèle linéaire simple** ($y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, $i = 1, \ldots, n$).

Estimateur.	ε^*	Remarques.
$\hat{\beta}_1 = \underset{1 \le i < j \le n}{\text{méd}} \left\{ \frac{y_j - y_i}{x_j - x_i} \right\}$	29%	***Auteur : H. Theil*** (1950). Médiane des n(n−1)/2 pentes possibles entre couples. Efficacité asymptotique élevée. Propriétés d'invariance : 1 et 2.
$\hat{\beta}_1 = \underset{i}{\text{méd}} \left\{ \underset{j \ne i}{\text{méd}} \left(\frac{y_j - y_i}{x_j - x_i} \right) \right\}$	50%	***Auteur : A.F. Siegel*** (1982). *"Repeated Median estimator"*, ou *"RM estimator"*. Les propriétés asymptotiques ne sont pas connues. Propriétés d'invariance : 1 et 2.
$\underset{i}{\text{méd}} \left\{ (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \right\} = \min !$	50%	***Auteur : P.J. Rousseeuw*** (1984). *"Least Median of Squares"*, ou *"LMS"*. Faible convergence (en $n^{-1/3}$). Propriétés d'invariance : 1, 2 et 3.

Propriétés d'invariance de l'estimateur T des paramètres du modèle de régression multiple : Soient **v** un vecteur colonne, *c* une constante réelle, et **A** une matrice carrée non singulière, arbitraires,

(1) *"regression equivariant"* : $T(\mathbf{x}_i, y_i + \mathbf{x}_i\mathbf{v}) = T(\mathbf{x}_i, y_i) + \mathbf{v}$

(2) *"scale equivariant"* : $T(\mathbf{x}_i, cy_i) = cT(\mathbf{x}_i, y_i)$ $\qquad i = 1, ..., n$

(3) *"affine equivariant"* : $T(\mathbf{x}_i\mathbf{A}, y_i) = \mathbf{A}^{-1}T(\mathbf{x}_i, y_i)$

N.B. : la propriété (3) autorise un changement de coordonnées par combinaison linéaire des régresseurs, sans modifier les valeurs calculées de la variable réponse.

Il est important de souligner que les estimateurs *"RM"* et *"LMS"* se généralisent au modèle linéaire multiple (p > 1 régresseurs). Cette généralisation est immédiate pour l'estimateur de Rousseeuw. L'estimateur de Siegel nécessite de former les C_n^p sous-ensembles de p points

$$(\mathbf{x}_{i_1}, y_{i_1}), \ldots, (\mathbf{x}_{i_p}, y_{i_p}),$$

à chacun desquels correspond l'unique solution :

$$\mathbf{t}, \text{ vecteur dont la j-ème composante est notée } t_j(i_1, \ldots, i_p),$$

qui définit l'hyperplan passant exactement par ces p points. La j-ème composante du *"RM estimator"* est égale à :

$$\hat{\beta}_j = \underset{i_1}{\text{méd}} \left(\underset{i_2}{\text{méd}} \left(\ldots \left(\underset{i_p}{\text{méd}} \; t_j(i_1, i_2, \ldots, i_p) \right) \ldots \right) \right)$$

5.11. REGRESSION ROBUSTE AU SENS DE LA MOINDRE MEDIANE DES CARRES DES ECARTS (*LMS Regression*).

Pour simplifier, cette méthode est ici présentée pour le modèle linéaire simple. L'extension au cas du modèle multiple est immédiate. Soit donc les données :

$$\{ (x_i, y_i) \}_{i = 1, \ldots, n} \text{ , et le modèle : } y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Le résidu aléatoire ε est supposé vérifier les propriétés classiques ; on souhaite néanmoins protéger l'ajustement des dégâts que pourraient causer d'éventuelles valeurs aberrantes, et se donner un moyen de mettre en évidence ces dernières. Dans ce but, les paramètres du modèle sont identifiés par minimisation d'une mesure robuste de la dispersion des écarts :

$$\underset{i = 1, \ldots, n}{\text{méd}} \left\{ (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \right\} = \min !$$

· ***Interprétation géométrique :***

Identifier dans le plan {Ox, Oy} **la bande rectiligne la plus étroite qui contient la moitié des observations : la "droite *LMS*" en est l'axe médian.** Cette définition géométrique appelle trois remarques : *(i)* "moitié" signifie [n/2]+1, où [.] désigne la partie entière ; *(ii)* la largeur de la bande est mesurée verticalement, *i.e.*, parallèlement à l'axe des y ; *(iii)* au modèle multiple (p > 1), correspond un volume minimal entre hyperplans parallèles.

Cette interprétation géométrique permet de comprendre pourquoi, malgré son point de rupture élevé, la droite *LMS* peut être déstabilisée par de petites perturbations dans la "partie centrale" des données : si les observations se partagent équitablement entre deux bandes de largeurs voisines, alors une petite modification de l'une des observations peut "faire sauter" la droite *LMS* d'une bande à l'autre.

· ***Point de rupture :*** pour le modèle avec p > 1 régresseurs, et **X** de plein rang p,

$$\varepsilon^* = \left([n/2] - p + 2 \right) / n \text{ , où } [n/2] \text{ est le plus grand entier } \leq n/2 .$$

· ***"Exact fit property" :*** c'est une conséquence de la valeur élevée du point de rupture. Cette propriété peut être présentée comme suit, pour le modèle à p > 1 régresseurs : supposons qu'il existe un vecteur q, tel que au moins n – [n/2] + p – 1 des observations vérifient exactement $y_i = \mathbf{x}_i q$; alors : $\hat{b}_{LMS} = q$, et cela quelles que soient les valeurs des autres observations.

· ***Moindres carrés tronqués*** *(Least trimmed squares, LTS regression)*. Il est souvent recommandé d'employer cette méthode, plutôt que la régression *LMS* : la convergence est meilleure (en $n^{-1/2}$ au lieu de $n^{-1/3}$), et la fonction-objectif est aussi plus régulière. Le point de départ du calcul des estimateurs est la statistique d'ordre des écarts quadratiques :

$$(e^2)_{(1)} \leq (e^2)_{(2)} \leq \ldots \leq (e^2)_{(n)} \text{ , puis on minimise le critère } LTS : \sum_{i=1}^{h} (e^2)_{(i)}$$

Pour $h = [n(1-\alpha)] + [\alpha(p+1)]$, avec $\alpha \in]0, 1/2[$, la valeur du point de rupture ε_n^* est proche de la proportion de troncature α ; elle est maximale et vaut $([(n-p)/2]+1)/n$ lorsque $\alpha = 1/2$.

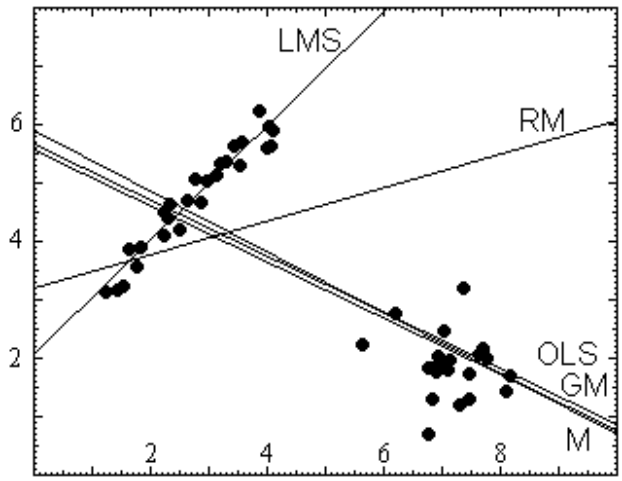
Pour la présentation de problèmes concrets traités à l'aide de ces méthodes robustes, le lecteur est invité à consulter l'ouvrage de P. J. ROUSSEEUW & A. M. LEROY (1987, *cf*. références au début du présent document).

· ***Expérience numérique illustrative : "breakdown plot".***

L'exemple qui suit est emprunté à l'ouvrage de P.J. ROUSSEEUW (1987). Il vise à faire apparaître concrètement, par génération d'échantillons pseudo-aléatoires, les différences de comportement entre quelques uns des estimateurs examinés jusqu'ici.

(i) **Effets d'une forte contamination (taux : 40%).** Le nuage des "bonnes observations" est construit suivant une direction d'allongement. Ce n'est pas le cas des points contaminants, qui sont dispersés de façon isotrope autour de leur barycentre :

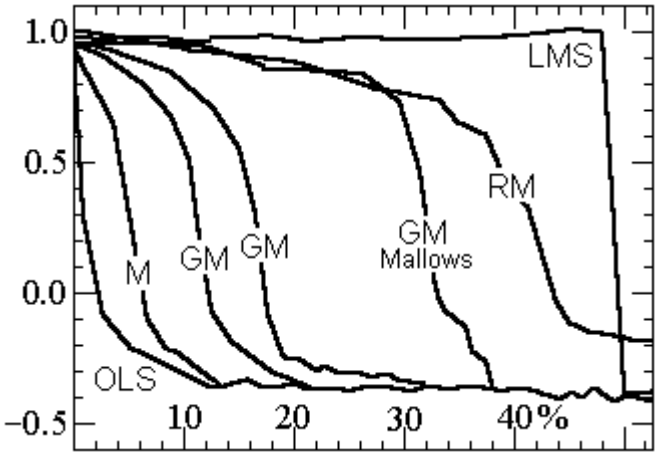
nombre total de points (x_i, y_i) : $n = 50$	"bonnes" observations : $y_i = x_i + 2 + \varepsilon_i$, $x_i \sim U[1, 4]$, $\varepsilon_i \sim N(0, \sigma)$, $\sigma = .2$	loi des $m = 20$ points aberrants : $(x_i, y_i) \overset{\text{i.i.d.}}{\sim} N_2\left\{\begin{pmatrix}7\\2\end{pmatrix} \begin{pmatrix}1/2 & 0\\0 & 1/2\end{pmatrix}\right\}$



Si le nuage des points aberrants était translaté "vers le bas", la droite ajustée selon le critère *RM* ne subirait qu'un déplacement fini (pas de "rupture"). La droite ajustée au sens de la moindre médiane des carrés des écarts ne serait quant à elle pas affectée.

LMS : *Least Median of Squares.*
RM : *Repeated Median.*
OLS : *Ordinary Least Squares.*
GM : *Mallows' and Scheppe's Generalized M-estimators.*
M : *Huber's M-estimator.*

***(ii)* Comportement des estimateurs vis-à-vis d'une perturbation croissante :** les mêmes caractéristiques que précédemment sont utilisées pour créer les données. L'augmentation progressive de la contamination m/n s'obtient en remplaçant un "bon" point par un *"outlier"* à chaque étape. Pour n = 100, m prend donc successivement les valeurs 0, 1, 2, ..., 50. Résultat :



*"**Breakdown plot**"* des différents estimateurs. Le graphe montre les variations des valeurs estimées de la pente (ordonnées), en fonction de l'accroissement de la contamination (abscisses : m/n exprimé en %).

5.12. MODELE LINEAIRE : ESTIMATION ROBUSTE EN DEUX ETAPES.

· ***Etape 1 : Minimiser la médiane des carrés des écarts ("LMS regression").***

Identifier l'estimateur $\hat{b}_{LMS}$: $\underset{i}{\text{méd}} \left\{ (y_i - \mathbf{x}_i \hat{b}_{LMS})^2 \right\} = \min !$

où b est le (p+1)×1 vecteur des paramètres du modèle, et calculer une valeur initiale robuste $s^{[0]}$ du paramètre d'échelle :

$$s^{[0]} = C \sqrt{\underset{i}{\text{méd}} \, (e_i^2)} \qquad C = 1.4826(1 + 5/(n-p-1))$$

Dans la constante *C*, le facteur $1.4826 \approx \Phi^{-1}(.75)$ vise à garantir une certaine consistance lorsque les résidus sont normaux. Le rôle du facteur (empirique) en n et p est d'éviter que la valeur de $s^{[0]}$ ne devienne trop faible, spécialement lorsque n est lui-même petit.

Déterminer les poids initiaux : $w_i^{[0]} = \begin{cases} 1 \text{ si } \left| e_i / s^{[0]} \right| \le 2.5 \\ 0 \text{ sinon} \end{cases}$

pour l'estimation du paramètre d'échelle : $\hat{\sigma}_{LMS} = \sqrt{\sum w_i^{[0]} e_i^2 \Big/ \left(\sum w_i^{[0]} - p - 1 \right)}$

L'estimateur $\hat{\sigma}_{LMS}$ (dont le point de rupture vaut 50%) sert de référence pour apprécier l'amplitude des écarts au modèle ajusté par le critère *"LMS"*. A l'étape suivante, $\hat{\sigma}_{LMS}$ permet d'attribuer une masse nulle aux points aberrants vis-à-vis de l'ajustement robuste.

· ***Etape 2 : Minimiser la somme des carrés repondérés ("RLS regression").***

Recalculer les pondérateurs : $w_i = \begin{cases} 1 \text{ si } \left| e_i / \hat{\sigma}_{LMS} \right| \leq 2.5 \\ 0 \text{ sinon} \end{cases}$

et identifier $\hat{b}_{RLS}$ qui minimise le critère des moindres carrés repondérés (*Reweighted Least Squares*) :

$$\sum_{i=1}^{n} w_i (y_i - \mathbf{x}_i \hat{b}_{RLS})^2 = \min !$$

Intérêt au plan des applications : $\hat{b}_{RLS}$ est l'estimateur des moindres carrés ordinaires calculé sur l'ensemble des données, *moins* les valeurs douteuses (celles auxquelles est attribué un poids $w_i = 0$). D'où l'intérêt pratique évident d'ajuster conjointement, pour les comparer entre elles, la *"RLS regression"* et la régression classique (non robuste) : lorsque les résultats de l'ajustement diffèrent sensiblement, les écarts à la première offrent un moyen de détecter les points qui déstabilisent la seconde.

5.13. EVALUATION DES INCERTITUDES HORS DU CADRE PARAMETRIQUE CLASSIQUE.

· ***Exposé du problème* :** l'identification des paramètres d'un modèle est fondée sur la recherche de l'extrêmum d'une certaine fonction S (*Cf.* §§ 1.1, 2.10, 2.17, 3.10, 4.4, 4.7, ainsi que le présent chapitre). Ce critère S quantifie l'écart entre le "morceau de réalité" auquel on accède par sondage et/ou expérimentation, et la représentation qu'en restitue le modèle (*e.g.*, un résumé synthétique, une prévision, …). Par conséquent, la première étape de la démarche consiste à définir une mesure Q de l'écart à l'ajustement e :

$$Q[e] = e^2, \quad \text{ou encore :} \quad Q[e] = |e|, \quad \text{par exemple.}$$

Avec, dans le cas du modèle linéaire, et en employant les mêmes notations que jusqu'à présent :

$$e_i(\mathbf{b}) = y_i - \mathbf{x}_i \mathbf{b} \; ; \quad i = 1, 2, \ldots, n$$

La solution recherchée est le $(p+1) \times 1$ vecteur $\hat{b}$ qui réalise le minimum du critère S(**b**) :

$$S(\mathbf{b}) = \sum_{i=1}^{n} Q\left[e_i(\mathbf{b}) \right]$$

La condition nécessaire pour que $S(\hat{b})$ soit un extrêmum de S(**b**) s'écrit :

$$\frac{\partial}{\partial \beta_j} \sum_{i=1}^{n} Q\left[e_i(\mathbf{b}) \right] \Big|_{\mathbf{b}=\hat{b}} = 0 \quad ; \quad j = 0, 1, 2, \ldots, p$$

Et si l'on note q[e] la dérivée de Q[e] (*i.e.*, q[e] = dQ[e]/de), avec $e_i(\mathbf{b}) = y_i - \mathbf{x}_i\mathbf{b}$:

$$\mathbf{X}' \mathbf{q}(\mathbf{b}) \big|_{\mathbf{b}=\hat{b}} = \mathbf{0} \ , \quad \text{où : } \mathbf{q}(\mathbf{b}) = \left(q[e_1(\mathbf{b})], \ldots, q[e_n(\mathbf{b})] \right)'$$

On remarque immédiatement qu'à $Q[e] = e^2$, soit encore q[e] = 2e, il correspond le système des p+1 "équations normales" dont la solution est l'estimateur des MCO :

$$2\,\mathbf{X}'\,\mathbf{e}(\mathbf{b}) \big|_{\mathbf{b}=\hat{b}} = 2\,\mathbf{X}'(\mathbf{y} - \mathbf{x}\hat{b}) = \mathbf{0}$$

Soulignons que dans ce contexte, on obtient non seulement une solution analytique pour le calcul de l'estimateur $\hat{b}$, mais que la théorie permet de plus d'établir l'équation des erreurs standard des composantes $\hat{\beta}_j$ de $\hat{b}$. Autrement dit, on accède par la voie analytique à un indicateur de la "stabilité" de $\hat{b}$ (*Cf.* par exemple §§ 2.9, 3.3). Rappelons que la notion de stabilité fait ici référence à la plus ou moins grande similitude entre les estimations que l'on obtiendrait si l'on avait la possibilité de répéter un grand nombre de fois l'acquisition des observations et leur ajustement au modèle.

On a vu qu'il est essentiel de disposer du moyen d'évaluer la stabilité des estimations. Mais en dehors du cadre "classique" (modèle linéaire, moindres carrés), pour lequel on dispose des résultats de l'algèbre linéaire, il n'est pas garanti *a priori* que l'on saura calculer analytiquement l'erreur standard, sinon au prix d'approximations. Ainsi, la théorie n'offre pas de solution si l'on choisit Q[e] = |e|, mesure à laquelle correspond le critère S "moindres valeurs absolues des écarts" (*Least Absolute Deviation*, ou *LAD*). Il en est de même, entre autres exemples, pour les estimateurs $\hat{b}_{LMS}$, $\hat{b}_{LTS}$, et $\hat{b}_{RLS}$, présentés au paragraphes 5.11 et 5.12 (rappelons que les modèles de régression robuste sont non linéaires).

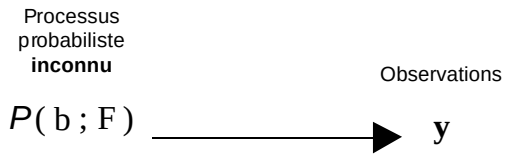
Pour résoudre ce type de problème réfractaire à la statistique mathématique, l'alternative suivante a été proposée : utiliser la puissance de calcul des ordinateurs. Plus précisément, on fait appel à une procédure de ***rééchantillonnage*** des données ; la plus connue est aujourd'hui le ***bootstrap***, dont le principe général est exposé en annexe (pp. 143-146). On limitera ici l'exposé au cas de l'application du bootstrap à la régression.

· ***Application du rééchantillonnage à la régression* :**

Considérons le vecteur **y** des n observations, avec,

$$y_i = f(b ; \mathbf{x}_i) + \varepsilon_i \qquad \varepsilon_1, ..., \varepsilon_n \overset{\text{i.i.d.}}{\sim} F \qquad F \text{ inconnue}$$

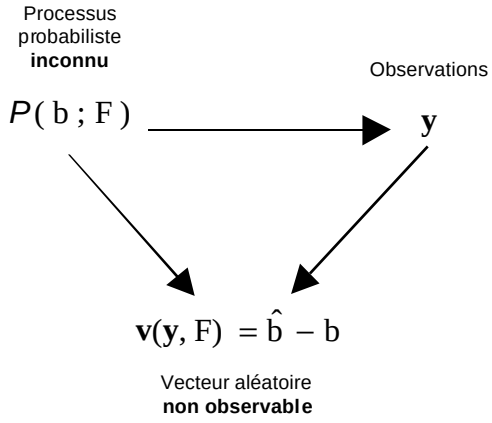
Les résidus indépendants suivent une loi de probabilité inconnue F, centrée sur 0 au sens où $E_F[\varepsilon] = 0$, ou encore $\text{Proba}_F\{\varepsilon < 0\} = 1/2$, par exemple. On peut donc noter $P(b ; F)$ le processus probabiliste inconnu qui engendre les observations **y** :



Suivant la démarche rappelée plus haut, on obtient une estimation $\hat{b}$ du vecteur inconnu b, estimation optimale au sens d'un critère S(**b**) :

$$\hat{y}_i = f(\hat{b} ; \mathbf{x}_i) + e_i \qquad i = 1, \ldots, n$$

La question traitée ici s'énonce : comment évaluer la qualité de l'estimateur $\hat{b}$? En d'autres termes, que peut-on dire des propriétés du vecteur aléatoire non observable $\hat{b} - b$? Posons $\mathbf{v}(\mathbf{y} ; F) = \hat{b} - b$:



La stabilité de l'estimateur $\hat{b}$ est classiquement mesurée par la matrice de variances-covariances de $\hat{b} - b$, que l'on va chercher à estimer :

$$S(F) = E_F[(\hat{b} - b)(\hat{b} - b)'] = E_F[\mathbf{v}\mathbf{v}']$$

A ce niveau, deux voies sont envisageables pour statuer sur la qualité de l'estimation :

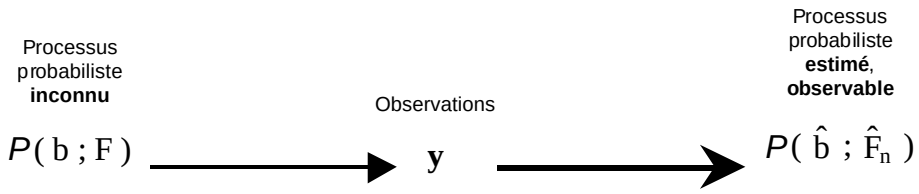
- celle qui est ici supposée inopérante, *i.e.*, utiliser une expression analytique pour le calcul de la matrice S(F) ; on s'est en effet placé dans la situation où l'on ne sait pas établir une telle équation.
- Répéter plusieurs fois l'expérience qui a permis d'obtenir les données **y**, et ajuster le modèle à chacun des ensembles d'observations ainsi répliquées. En pratique, une telle démarche n'est quasiment jamais applicable, car l'analyste ne dispose en général que d'un unique échantillon **y**. C'est-à-dire qu'il est presque toujours impossible de réaliser des réplications indépendantes de **y** à partir du processus générateur $P(b ; F)$.

La démarche alternative est celle du bootstrap : dans son principe (*Cf.* Annexe), la méthode consiste simplement à **répliquer les données non pas à partir du processus générateur inconnu** $P(b ; F)$, **mais à partir d'un processus estimé**. Cela peut être éclairé à l'aide du schéma des relations entre observations et entités inconnues, que l'on a présenté plus haut.

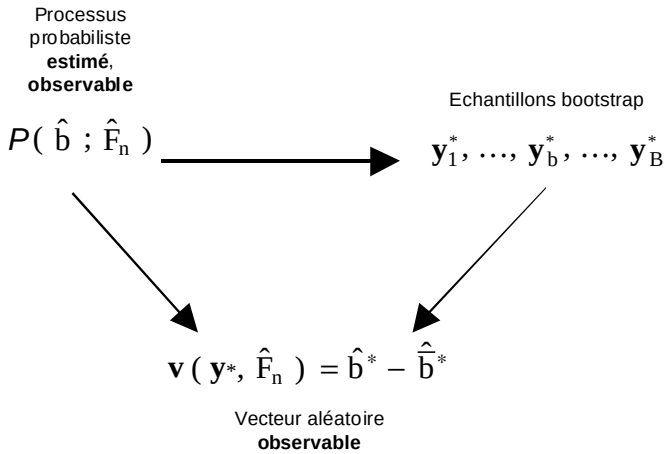
Soit $\hat{F}_n$ la fonction de répartition empirique (*Cf.* § 5.6.) de la loi inconnue F :

$\hat{F}_n$: attribution de la probabilité 1/n à chaque écart e_i

Connaissant $\hat{b}$, désignons alors par $P(\hat{b} ; \hat{F}_n)$ le processus estimé :

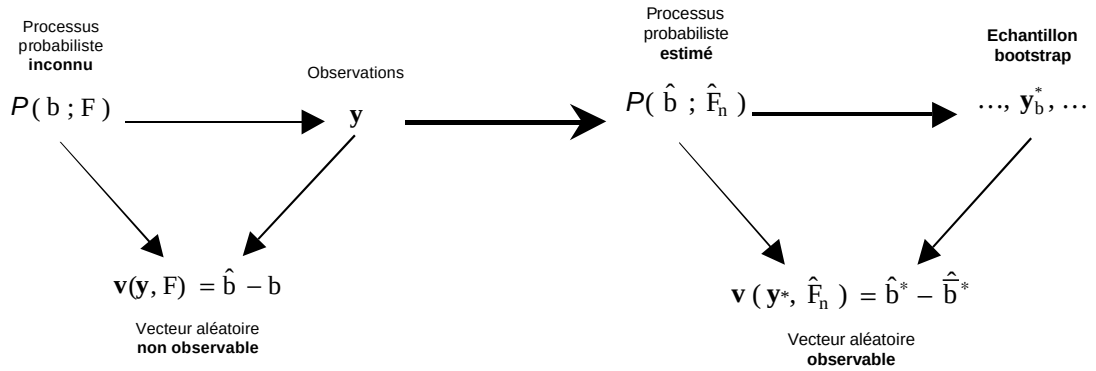


L'idée de base du bootstrap est de recréer par simulation un grand nombre (noté B) ***d'échantillons*** (notés $\mathbf{y}^*$)***, selon un algorithme de tirage pseudo-aléatoire dans le processus probabiliste estimé***. A chaque "échantillon bootstrap" $\mathbf{y}^*$ est associé un "réplicat bootstrap" $\hat{b}^*$, ce que l'on peut résumer par le schéma suivant :



Le schéma qui précède illustre comment le rééchantillonnage plonge la représentation de "l'état (non observable) de la Nature" dans un "monde bootstrap" accessible à l'analyste. Ainsi obtient-on par simulation les "réplicats bootstrap" $\mathbf{v}(\mathbf{y}_1^*, \hat{F}_n)$, …, $\mathbf{v}(\mathbf{y}_B^*, \hat{F}_n)$, dont la distribution offre une image du "comportement statistique" de la quantité aléatoire non observable $\mathbf{v}(\mathbf{y}, F)$.

Le résumé ci-après souligne que l'étape cruciale est l'estimation du processus générateur des données, étape qui permet le passage du monde réel au "monde bootstrap". Dans le modèle de régression examiné ici, la statistique-clef est la fonction de répartition empirique $\hat{F}_n$, qui estime la loi F du résidu aléatoire attaché au modèle $E_F[y_i] = f(b, \mathbf{x}_i)$.



· ***Algorithme du bootstrap appliqué à la régression*** **:**

En pratique, il s'agit de calculer B estimations indépendantes du paramètre inconnu b. Dans ce qui suit, on forme des échantillons bootstrap d'écarts ; étant donné que l'on connaît le modèle ajusté, la présentation est équivalente à celle qui précède.

♦ **Etape 1.** Estimer F par $\hat{F}_n$ en attribuant la masse 1/n à chaque écart centré $e_i - \bar{e}$,

$$\bar{e} = \frac{1}{n}\sum_{i=1}^{n} e_i \quad ; \quad \hat{F}_n(e) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{]-\infty,\, e[}\left(e - \bar{e}\right)$$

♦ **Etape 2.** Réaliser dans $\hat{F}_n$ n tirages aléatoires indépendants, avec remise. Concrètement, on forme un n-échantillon bootstrap noté $\{e_1^*, \ldots, e_n^*\}$ en effectuant n tirages aléatoires avec remise dans l'ensemble $\{ e_1 - \bar{e}, \ldots, e_n - \bar{e}, \}$; au rééchantillon $\{e_1^*, \ldots, e_n^*\}$ est associé le critère $S^*(\mathbf{b})$:

$$S^*(\mathbf{b}) = \sum_{i=1}^{n} Q\,[\, e_i^*(\mathbf{b})\,]$$

L'optimisation de ce critère fournit le "réplicat bootstrap" $\hat{b}^*$.

♦ **Etape 3.** Effectuer B répétitions indépendantes de l'**étape 2**, pour obtenir B "réplicats bootstrap" $\hat{b}_1^*, \ldots, \hat{b}_b^*, \ldots, \hat{b}_B^*$.

A l'issue de l'étape 3, on peut alors estimer, par exemple, la variance de l'estimateur par :

$$\hat{Var}_{*}(\hat{\beta}_j) = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{\beta}^{*}_{b,j} - \bar{\beta}^{*}_{j})^2 \quad \text{avec :} \quad \bar{\beta}^{*}_{j} = \frac{1}{B}\sum_{b=1}^{B}\hat{\beta}^{*}_{b,j}$$

Autrement dit, on estime la matrice S(F) par : $\hat{\sigma}^{*}(\hat{F}_n) = E_{*}[\,(\hat{b}^{*} - \bar{b}^{*})\,(\hat{b}^{*} - \bar{b}^{*})'\,]$, où la notation $E_{*}[\ ]$ rappelle qu'il s'agit de l'espérance de réplicats bootstrap.

Il convient d'attirer l'attention sur le point suivant : l'algorithme qui précède opère un rééchantillonnage des écarts à l'ajustement ; ce choix est fondé sur l'hypothèse selon laquelle la loi F du résidu e ne dépend pas des valeurs du (ou des) régresseurs. C'est là une hypothèse forte ; si on ne lui accorde que peu de confiance, il est alors préférable de rééchantillonner les éléments $(\mathbf{x}_i, y_i)$, en suivant les étapes qui viennent d'être exposées : on formera B rééchantillons $(\mathbf{x}_i, y_i)^*$. Cette remarque permet de rappeler que **le bootstrap ne peut pas être employé sans précautions**, sous peine d'aboutir à des résultats inconsistants. Nous conseillons donc vivement au praticien confronté à un problème concret de s'appuyer sur les ouvrages mentionnés ci-après.

Une abondante littérature a été consacrée au bootstrap. Le lecteur intéressé par la mise en pratique de cette méthode consultera avec profit les chapitres traitant des applications au modèle de régression dans les deux ouvrages suivants : celui de B. Efron & R.J. Tibshirani (1993), et celui de A.C. Davison & D.V. Hinkley (1997), dont les références sont données plus loin en Annexe. Et aussi, pour sa présentation très didactique et soutenue par une riche iconographie, l'Annexe 2 (pp. 303-331) de l'ouvrage de L.C. Hamilton (*Cf.* Liminaire, référence [11] p. iii).

Chapitre 6

Quelques extensions du modèle linéaire classique.

Sommaire du chapitre 6

6.1. NOTION DE MODELE LINEAIRE GENERALISE.

· ***Exemple : la régression logistique.***

Le modèle linéaire classique est caractérisé par deux contraintes fortes : la linéarité de sa composante déterministe, et la normalité de sa composante aléatoire. Le modèle linéaire généralisé est conçu pour traiter les problèmes où la loi de la variable réponse est décrite par un modèle paramétrique autre que le modèle gaussien (*e.g.*, loi de Poisson). L'exemple qui suit, emprunté à l'ouvrage de Sanford WEISBERG (1985), est celui où la loi de la variable réponse est binômiale, problème traité dans le cadre de la régression logistique.

Variable contrôlée $x_1, \ldots, x_i, \ldots, x_N$	Nb. d'épreuves de Bernouilli : n_i	Nb. de "succès" $y_1, \ldots, y_i, \ldots, y_N$	Proportion de succès $y_1/n_1, \ldots, y_N/n_N$
0	70	0	.000
1	70	9	.129
2	70	21	.300
3	70	47	.671
4	70	60	.857
5	70	63	.900

♦ **Modèle :** $y_i \sim B(n_i, \theta_i)$, $\theta_i \in [0, 1]$, $i = 1, \ldots, N$

♦ **Objectif de la régression logistique :** exprimer la dépendance de la probabilité de succès inconnue θ_i en fonction du régresseur x_i . Le modèle $E(y_i/n_i) = \beta_0 + \beta_1 x_i$ étant à l'évidence inapproprié, on applique classiquement ***la transformation "logit"*** :

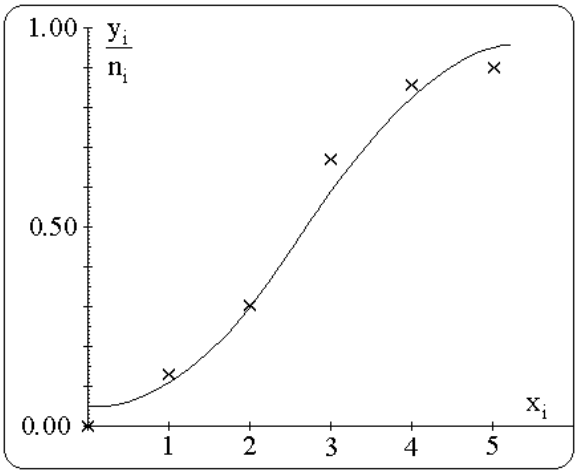
$$\theta_i \in [0, 1] \xrightarrow{\text{logit}} \text{logit}(\theta_i) = \ln\left(\theta_i/(1-\theta_i) \right) \in \mathbb{R}$$

i.e., le logarithme népérien du rapport de la probabilité de succès à la probabilité d'échec. Cette transformation aboutit à deux formulations équivalentes du modèle de régression :

$$E\left[\text{logit}(\theta_i)\right] = \beta_0 + \beta_1 x_i \quad , \quad E(y_i/n_i) = \theta_i = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}$$

♦ **Estimation des paramètres :** les données consistent en N triplets (y_i , n_i , x_i). La variance de la variable réponse n'est pas stable : $\text{Var}(y_i/n_i) = \theta_i (1-\theta_i)/n_i$. La régression pondérée des $\text{logit}(y_i/n_i)$ sur les x_i, avec les poids $w_i = n_i / [\theta_i (1-\theta_i)]$, semblerait *a priori* la technique adéquate. Malheureusement, les θ_i sont inconnus, et cela nécessite de faire appel à une procédure itérative. Habituellement, les utilisateurs des modèles linéaires généralisés emploient le logiciel GLIM (*Generalized Linear Interactive Modeling*, BAKER, R.J., & J.A. NELDER, 1978).

♦ **Résultat de l'ajustement :**



$\hat{\beta}_0 = -3.301$

$\hat{SE}(\hat{\beta}_0) = .3238$

$\hat{\beta}_1 = 1.246$

$\hat{SE}(\hat{\beta}_1) = .1119$

deviance = 9.35

(4 d.d.l.)

♦ **Test sur les paramètres du modèle :** comme dans le modèle linéaire classique, la question est celle de la dépendance de la réponse vis-à-vis du régresseur (*i.e.*, β_1 différent de, ou bien égal à zéro ?). Dans le cas présent, une réponse approchée est obtenue en comparant la valeur du rapport (estimation de β_1) / (écart-type estimé de l'estimateur) avec le fractile d'ordre $\alpha/2$ de la loi N(0, 1) : 1.246/.1119 = 11.13, valeur qui conduit à repousser H_0 ($\beta_1 = 0$) pour conserver H_1 ($\beta_1 \neq 0$).

Il existe un test plus fiable, fondé sur la statistique *"**deviance**"* ; elle joue ici le rôle de la somme des carrés résiduelle du modèle linéaire classique. Dans le cas de la régression logistique, elle s'exprime :

$$\textit{deviance} \;=\; 2 \sum_{i=1}^{N} \left\{ y_i \; \ln\left(\frac{y_i}{n_i \hat{\theta}_i} \right) + \left(n_i - y_i \right) \ln\left(\frac{n_i - y_i}{n_i - \hat{\theta}_i} \right) \right\}$$

Le nombre de d.d.l. vaut N moins le nombre de paramètres. La statistique du test est la différence entre la déviance du modèle "complet" (*i.e.*, qui inclut β_1) et la déviance du modèle $E[y_i] = \exp(\beta_0) / [1+\exp(\beta_0)]$, qui exprime l'absence de relation entre la variable réponse et le régresseur (*"**change in deviance method**"*). Sous H_0 , la différence suit une loi de χ^2, avec ν = différence entre les nombres de d.d.l. ; ici :

$$\textit{change in deviance} = 250.5 - 9.4 = 241.1 \;, \quad \text{avec } 5 - 4 = 1 \text{ d.d.l.}$$

· ***Présentation résumée de quelques autres cas élémentaires :***

♦ **Architecture du modèle linéaire généralisé :** l'exemple de la régression logistique, qui est un cas particulier des modèles linéaires généralisés, permet d'illustrer la structure de ces modèles. Elle inclut fondamentalement :

(i) une fonction linéaire des paramètres (*"**linear predictor**"*), qui exprime l'espérance de la i-ème réponse. Dans l'exemple traité : $\beta_0 + \beta_1 x_i$;

(ii) une fonction de lien (*"**link function**"*), qui permet de transformer l'espérance de la variable réponse en une combinaison linéaire des paramètres du modèle. Dans l'exemple traité, il s'agit de la transformation logit.

(iii) une loi des résidus (*"**error function**"*), qui représente la composante aléatoire du modèle. L'hypothèse d'indépendance des résidus est conservée ; la généralisation provient de l'abandon du modèle normal, et de la possibilité de choisir une loi de la famille exponentielle (la loi binômiale dans le cas de la régression logistique).

♦ **Exemples :**

En général, la partie déterministe du modèle inclut p régresseur et p paramètres (p+1 s'il y a un terme constant). Soit donc $\mathbf{x}_i$ le i-ème vecteur-ligne de la matrice **X**, et b le vecteur-colonne des paramètres. Avec ces notations, trois modèles communément utilisés sont présentés, regroupés avec le modèle classique dans le tableau suivant :

Fonction de lien	Modèle de régression	Composante aléatoire
$\mathbf{x}_i b = E(y_i)$	$E(y_i) = \mathbf{x}_i b$	Normale
$\mathbf{x}_i b = \ln[E(y_i)]$	$E(y_i) = \exp(\mathbf{x}_i b)$	Poisson
$\mathbf{x}_i b = \text{logit}[E(y_i)]$	$E(y_i) = \dfrac{\exp(\mathbf{x}_i b)}{1+\exp(\mathbf{x}_i b)}$	Binômiale
$\mathbf{x}_i b = 1/E(y_i)$	$E(y_i) = 1/\mathbf{x}_i b$	Gamma

On retiendra que cette présentation ne donne qu'une idée très succincte de la richesse des possibilités offertes par le modèle linéaire généralisé ; ainsi, des associations entre fonction de lien et loi des résidus différentes de celles exposées ci-dessus peuvent elles être envisagées. On remarquera aussi que l'analyse des variables discrètes et continues est englobée dans le même cadre. L'ouvrage de P. McCULLAGH & J.A. NELDER (1983, *Generalized Linear Models*, Chapman & Hall ed., London) expose ces méthodes de façon synthétique, tant du point de vue théorique que pratique.

6.2. RELATION FONCTIONNELLE, RELATION STRUCTURELLE.

Jusqu'à présent, seul a été envisagé le contexte de la régression, où la variable x est certaine. L'objectif est alors de formuler la dépendance de l'espérance de la variable aléatoire Y vis-à-vis de x, dans le but de **prévoir E(Y) pour un x donné**. Un problème distinct est assez communément rencontré : l'objectif est **l'estimation des paramètres de la relation, sans ambition de prévision, mais à des fins de description ou de comparaison à la théorie**. Cette seconde question amène à considérer des relations dont le modèle de régression "classique" n'est qu'un cas particulier.

· ***Relation fonctionnelle :***

C'est le modèle mathématique postulé (ou fondé sur une loi établie), qui décrit la relation linéaire entre deux variables "mathématiques" X et Y :

$$Y = \beta_0 + \beta_1 X$$

Problème : estimer les paramètres β_0 et β_1 de ce modèle théorique. Si X et Y étaient connues sans erreur (ce qui n'est jamais le cas de variables expérimentalement mesurées), la solution serait obtenue par simple résolution d'un système d'équations.

· ***Relation structurelle :***

♦ **Premier cas :** X et Y représentent des quantités certaines, mais dont la mesure est entachée d'une erreur. On ne peut donc observer que les réalisations de deux variables aléatoires ξ et η :

$$\xi_i = X_i + \delta_i \quad , \quad \eta_i = Y_i + \varepsilon_i \quad , \quad i = 1, \ldots, n$$

notations qui expriment qu'à chaque observation d'une "vraie" valeur est attachée une erreur aléatoire. Les hypothèses du modèle structurel sont les suivantes :

$$E[e] = \mathbf{0} \ , \ E[ee'] = \sigma_\varepsilon^2 \mathbf{I}_n \ , \ E[d] = \mathbf{0} \ , \ E[dd'] = \sigma_\delta^2 \mathbf{I}_n \ , \ E[ed'] = \mathbf{0}$$

et la relation entre les variables aléatoires observables ξ et η s'exprime :

$$\boxed{\eta = \beta_0 + \beta_1 \xi + (\varepsilon - \beta_1 \delta)}$$

L'estimation des paramètres de ce modèle pose un problème distinct de celui de la régression. Tout d'abord, ξ est une variable aléatoire. De plus, elle est corrélée au terme d'erreur :

$$\mathrm{Cov}(\xi, \ \varepsilon - \beta_1 \delta) = -\beta_1 \sigma_\delta^2$$

Remarque : ***La régression "classique" est le cas particulier correspondant à*** $\sigma_\delta^2 = 0$.

♦ **Second cas :** généralise le premier, X et Y étant cette fois des variables aléatoires. On retrouve les mêmes résultats que précédemment, à l'exception de la covariance, qui ne doit plus être calculée en considérant X comme une constante :

$$\mathrm{Cov}(\xi,\ \varepsilon - \beta_1\delta) \ = \ \mathrm{E}[X\varepsilon] \ - \ \beta_1\mathrm{E}[X\delta] \ - \ \beta_1\sigma_\delta^2$$

On se ramène au modèle précédent en formulant les hypothèses supplémentaires :

$$\mathrm{Cov}(X,\ \delta) \ = \ \mathrm{Cov}(X,\ \varepsilon) \ = \ \mathrm{Cov}(Y,\ \delta) \ = \ \mathrm{Cov}(\ Y,\ \varepsilon) \ = \ 0$$

6.3. MODELE LINEAIRE SIMPLE : RELATION STRUCTURELLE.

· ***Définition des sources de variabilité à l'aide d'un exemple :***

On va considérer la relation entre les deux variables aléatoires X et Y :

$$X \ = \ \mathrm{Log}(L) \quad , \quad Y \ = \ \mathrm{Log}(W)$$

où L représente la longueur d'un organisme (*e.g.*, un représentant d'une espèce animale) et W son poids. On souhaite estimer le "coefficient d'allométrie", *i.e.*, le paramètre β_1 de la relation linéaire :

$$Y \ = \ \beta_0 \ + \ \beta_1 X$$

(i) Mesures réalisées sur un seul individu :

Sur un unique individu, caractérisé par $X = x^*$ et $Y = y^*$, on réalise au même instant une série de k couples de mesures indépendantes (taille, poids). On obtient ainsi k réalisations du couple aléatoire :

$$x^* \ + \ \delta \quad , \quad y^* \ + \ \varepsilon$$

où l'aléa δ (respectivement ε) est l'erreur attachée à la mesure de la quantité x^* (resp. y^*). On supposera :

$$\mathrm{E}(d) \ = \ \mathrm{E}(e) \ = \ \mathrm{E}(ed') \ = \ \mathbf{0} \quad , \quad \mathrm{E}(dd') \ = \ \sigma_\delta^2\mathbf{I} \quad , \quad \mathrm{E}(ee') \ = \ \sigma_\varepsilon^2\mathbf{I}$$

hypothèses qui généralisent celles formulées dans le cadre de la régression.

(ii) Mesures réalisées sur un échantillon aléatoire d'individus :

On considère désormais la situation habituellement rencontrée en pratique, où l'on dispose d'un échantillon de n individus, et où l'on mesure la longueur et le poids de chacun. D'où les n réalisations des variables aléatoires ξ et η :

$$\xi_i \ = \ X_i + \delta_i \quad , \quad \eta_i \ = \ Y_i + \varepsilon_i \quad , \quad i \ = \ 1,\ ...,\ n$$

Dans ce cas, le caractère aléatoire des variables observées ξ et η provient :
- des erreurs commises lors de la mesure de X et de Y (*Cf.* ci-dessus),
- et aussi du fait que la longueur et le poids d'individus échantillonnés dans une population sont des variables aléatoires.

On retiendra donc qu'il existe ***une variabilité intrinsèque aux quantités*** X ***et*** Y que l'on étudie, variabilité ***à laquelle se superpose un bruit inhérent à l'acte de mesure***. Ces deux sources de variation sont supposées non corrélées, hypothèse qui s'exprime :

$$\mathrm{Cov}(X,\,\delta) \;=\; \mathrm{Cov}(X,\,\varepsilon) \;=\; \mathrm{Cov}(Y,\,\delta) \;=\; \mathrm{Cov}(Y,\,\varepsilon) \;=\; 0$$

L'estimation du paramètre β_1 pose un problème différent de celui résolu pour la régression "classique", car les variances des erreurs δ et ε jouent désormais un rôle essentiel.

· ***Estimation des paramètres par le maximum de vraisemblance :***

♦ **Modèle :**

$$\xi_i \;=\; X_i + \delta_i \;\;,\;\; \eta_i \;=\; Y_i + \varepsilon_i \;\;,\;\; Y_i \;=\; \beta_0 + \beta_1 X_i \quad ;\quad i \;=\; 1,\,\ldots,\,n$$

Ce modèle structurel postule une relation linéaire entre les deux variables aléatoires X et Y, non observables. Au corps d'hypothèses déjà constitué (*Cf.* ci-dessus), on ajoutera la suivante : les couples aléatoires observables $(\xi_i,\ \eta_i)$ constituent n réalisations indépendantes d'une même loi normale bivariée, *i.e.*,

$$\begin{pmatrix}\xi\\ \eta\end{pmatrix} \;\sim\; N_2\left\{\begin{pmatrix}\mu\\ \beta_0+\beta_1\mu\end{pmatrix},\ \begin{pmatrix}\sigma_X^2+\sigma_\delta^2 & \beta_1\sigma_X^2\\ \beta_1\sigma_X^2 & \beta_1^2\sigma_X^2+\sigma_\varepsilon^2\end{pmatrix}\right\}$$

N.B. : Selon cette hypothèse, toutes les variables structurelles X_i possèdent la même espérance, et il en est bien entendu de même pour les Y_i . Les problèmes rencontrés si l'on relaxe cette restriction sont traités dans KENDALL, M. & A. STUART (*The Advanced Theory of Statistics*, Vol. 2, *Inference and relationship*, Ch. Griffin & Co., 4th ed., 1979).

♦ **Estimation des paramètres :**

$$\begin{cases}\hat\mu \;=\; \bar\xi\\ \hat\beta_0+\hat\beta_1\hat\mu \;=\; \bar\eta\\ \hat\sigma_X^2+\hat\sigma_\delta^2 \;=\; \hat\sigma_\xi^2\\ \hat\beta_1^2\hat\sigma_X^2+\hat\sigma_\varepsilon^2 \;=\; \hat\sigma_\eta^2\\ \hat\beta_1\hat\sigma_X^2 \;=\; \hat\sigma_{\xi,\,\eta}\end{cases}$$

L'ensemble {moyennes, variances, et covariance empiriques} sont des statistiques exhaustives pour la loi normale bivariée du couple (ξ, η), et sont aussi les EMV de ces paramètres. Les estimations MV des 6 paramètres :

$$\mu,\ \beta_0,\ \beta_1,\ \sigma_X^2,\ \sigma_\delta^2,\ \sigma_\varepsilon^2,$$

sont donc calculées en résolvant le système de 5 équations ci-contre, sous la contrainte que les estimations des variances soient positives.

On remarque qu'il s'agit d'un système de 5 équations à 6 inconnues, et seul le paramètre μ est identifiable.

En l'absence d'information (ou bien d'un hypothèse) complémentaire, deux résultats seulement sont obtenus :

(i) $\hat{\beta}_0 + \hat{\beta}_1 \bar{\xi} = \bar{\eta}$, *i.e.*, la droite ajustée passe par le barycentre $(\bar{\xi}, \bar{\eta})$ des n couples observés (ξ_i, η_i) ;

(ii) $\left|\hat{\sigma}_{\xi,\eta}\right| / \hat{\sigma}_\xi^2 \;\leq\; \hat{\beta}_1 \;\leq\; \hat{\sigma}_\eta^2 / \left|\hat{\sigma}_{\xi,\eta}\right|$, si $\hat{\beta}_1, \hat{\sigma}_{\xi,\eta} \neq 0$;

relations d'ordre qui signifient que la pente de la "droite structurelle" est bornée, en valeur absolue, inférieurement par le coefficient de régression de η sur ξ, et supérieurement par l'inverse du coefficient de régression de ξ sur η.

L'interprétation géométrique de ces résultats est évidente, et conforme à l'intuition : le modèle structurel est "encadré" par les deux droites de régression (η sur ξ, ξ sur η), et les trois droites passent par le barycentre du nuage des points observés.

♦ Comment lever l'indétermination du système d'équations ?

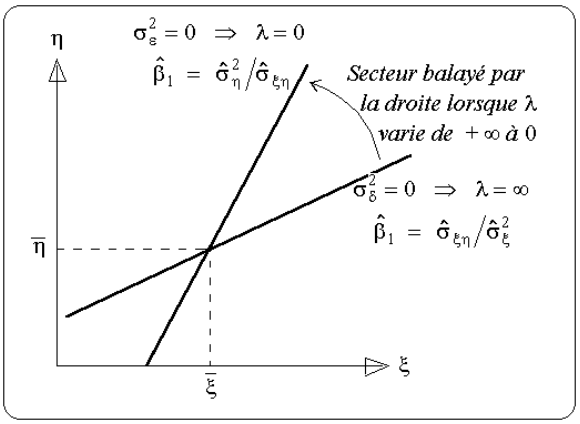
Il n'est pas possible d'utiliser une connaissance de β_0 ou β_1 , qui sont précisément les paramètres que l'on souhaite estimer ; il en est de même de la variance de X, cette variable n'étant pas observable. Dans cette situation, on est conduit à formuler une hypothèse sur les variances des erreurs δ et ε. Plusieurs solutions peuvent être envisagées, mais la démarche la plus classique consiste à poser :

$$\lambda = \sigma_\varepsilon^2 / \sigma_\delta^2 \quad , \quad \lambda \text{ fixé, connu ou estimable.}$$

Avec cette condition supplémentaire, on obtient :

$$\hat{\beta}_1 = \left(Q + \sqrt{Q^2 + 4\lambda\hat{\sigma}_{\xi,\eta}^2} \right) \Big/ 2\hat{\sigma}_{\xi,\eta} \quad , \quad \text{avec : } Q = \hat{\sigma}_\eta^2 - \lambda\hat{\sigma}_\xi^2$$

♦ Représentation graphique :

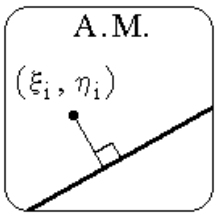


La figure ci-contre montre les deux positions extrêmes de la relation structurelle : les droites de régression de η sur ξ et de ξ sur η, correspondant à $\text{Var}(\delta) = 0$ et $\text{Var}(\varepsilon) = 0$ respectivement. **L'angle que forment entre elles ces deux droites est d'autant plus petit que la valeur absolue du coefficient de corrélation linéaire ρ est proche de 1 (les deux régressions sont confondues pour $\rho = \pm 1$).** En pratique, une valeur absolue élevée de l'estimation r de ρ (disons de l'ordre de 0.9 ou plus) incitera à considérer tous les modèles comme équivalents, et le choix de la régression "classique" de η sur ξ constituera alors le compromis le plus judicieux.

· ***Deux modèles classiques :***

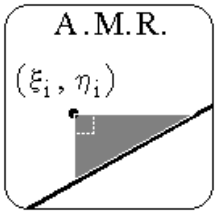
Lorsque λ n'est pas connu *a priori*, ou bien lorsque les données ne permettent pas de l'estimer, le modèle retenu est souvent soit l'axe majeur (*"major axis"*), soit l'axe majeur réduit (ou "droite des moindres rectangles", *"reduced major axis"*, *"geometric mean regression"*). A chacun correspond une hypothèse précise sur la valeur de λ.

L'axe majeur :



Les variances des erreurs sont égales (ou supposées telles) : $\lambda = 1$. L'axe majeur minimise la somme des carrés des écarts mesurés perpendiculairement à la droite ajustée (*schéma ci-contre*) ; c'est donc le premier axe de l'ACP de la matrice de variances-covariances des variables ξ et η, d'où son nom.

L'axe majeur réduit :

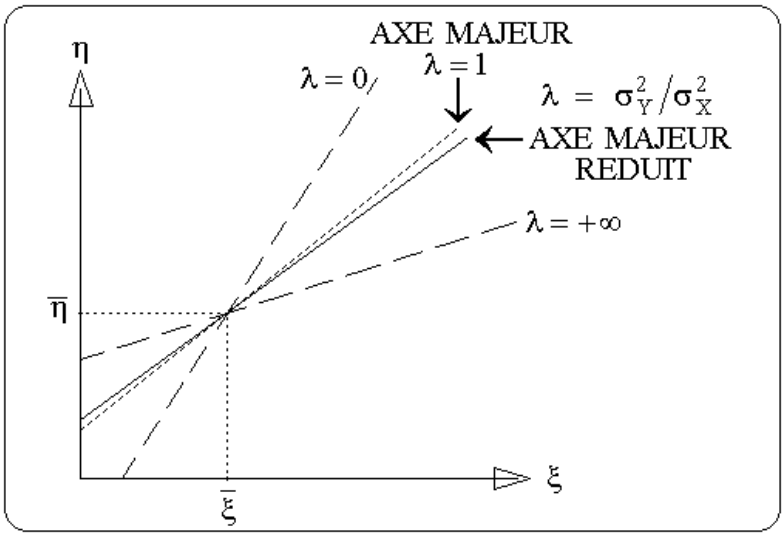


Encore appelé "droite des moindres rectangles", car il minimise la somme des surfaces figurées en gris sur le schéma ci-contre. Les variances des erreurs sont proportionnelles à celles de leurs variables respectives, *i.e.*, λ = Var(Y)/Var(X). L'estimateur de la pente vaut :

$$\hat{\beta}_1 = \hat{\sigma}_\eta / \hat{\sigma}_\xi \text{ , avec le signe de } \hat{\text{cov}}(\xi, \eta)$$

♦ **Représentation dans le plan des variables ξ et η :**

Quand $|\rho| \to 1$, les deux droites de régression (de ξ sur η, de η sur ξ) convergent vers une droite limite commune, qui est précisément l'axe majeur réduit.



· ***Régression prédictrice ou bien relation structurelle ?***

Il s'agit là d'une question assez fréquemment soulevée, motivée par celle du choix entre la droite des moindres carrés ordinaires, et l'axe majeur réduit, par exemple. Deux catégories d'éléments de réponse sont à considérer :

(i) Les objectifs de l'étude : si le but est la prévision des variations de la variable "expliquée" en fonction de celles de la variable "explicative", on a affaire à un problème de régression. Les deux variables ne jouent pas le même rôle. En particulier, les valeurs du régresseur sont en principe contrôlées par l'expérimentateur ; un exemple typiquement rencontré en physiologie est celui des relations dose-effet.

Si au contraire les deux variables jouent des rôles "symétriques", et que le but n'est pas de prévoir les valeurs prises par l'une d'elles en fonction de celles de l'autre, mais de décrire le plus précisément possible leur relation, on choisira d'ajuster un modèle structurel ; l'exemple classique en biologie est celui de l'estimation du coefficient d'allométrie. Quant au choix du "meilleur" modèle structurel (qui peut d'ailleurs être la droite des MCO ...), il est déterminé par la nature des variables étudiées (*Cf.* ci-après).

(ii) La nature des données : on supposera que le processus qui engendre les observations est correctement décrit par une relation linéaire. Si la dispersion des points expérimentaux autour de la tendance linéaire est faible ($|\rho|$ voisin de 1), tous les modèles possibles sont alors si proches les uns des autres que la droite des MCO constitue le choix le plus simple.

D'un point de vue pratique, le problème ne se pose réellement que lorsque $|\rho|$ est faible, et que l'emploi de la droite des MCO aboutit, si elle n'est pas le bon modèle, à sous-estimer β_1 . Il convient alors de déterminer λ, le rapport des variances des erreurs, et de s'interroger sur la sensibilité de l'estimation au choix de la valeur de λ. On montre que, asymptotiquement,

$$\frac{\partial \hat{\beta}_1}{\partial \lambda} = -\tau \beta_1 \Big/ \left(\beta_1^2 + \lambda\right) \quad , \quad \text{où} \quad \tau = \frac{\sigma_\delta^2}{\sigma_X^2} \quad \text{désigne le rapport bruit / signal.}$$

L'estimateur de la pente est donc relativement insensible à λ lorsque la vraie valeur de λ est elle-même grande, et que le rapport bruit/signal de la variable X est voisin de zéro. En dehors de ce contexte, des simulations ont établi qu'un choix erroné de la valeur de λ compromet le bénéfice attendu de l'utilisation du modèle structurel (plutôt que de la droite des MCO).

Deux cas peuvent se présenter :

- On connaît λ (*e.g.*, il a été estimé expérimentalement). L'estimation de β_1 peut alors être calculée directement.

- On ne connaît pas λ, et l'on ne dispose même pas des moyens de formuler une hypothèse réaliste sur sa valeur approchée. La solution optimale ne sera donc pas obtenue, mais l'on peut néanmoins être tenté de supposer que l'on "réduira les dégâts" en adoptant l'AMR (ou l'AM) à la place de la droite des MCO. Une recommandation pragmatique simple pourrait être la suivante : *si l'on sait que l'on commet une erreur appréciable sur la mesure de* X, calculer la pente de l'AMR et celle de la droite des MCO. Si la différence est sensible, choisir l'AMR. Sinon, choisir la droite des MCO.

ANNEXE.

Echantillonnage, rééchantillonnage : le *bootstrap.*

Le bootstrap *est une méthode relativement récente de calcul intensif sur ordinateur : l'article "fondateur" a été publié il y a une vingtaine d'années [B. Efron (1979). The 1977 Rietz lecture. Bootstrap methods: another look at the jackknife,* Ann. Statist **7**(1): 1-26*], et il a ensuite fallu attendre plus d'une décennie pour qu'apparaissent les premiers ouvrages de synthèse destinés aux praticiens [*e.g.*, B. Efron & R. Tibshirani (1993),* An introduction to the bootstrap*, Chapman & Hall, 436 p., et aussi : A.C. Davison & D.V. Hinkley (1997),* Bootstrap methods and their application*, Cambridge University Press, 582 p.]. Le principal objectif du* bootstrap *est de quantifier la précision des estimateurs ; la mesure habituellement employée à cette fin est l'erreur-standard attachée à l'estimation,* i.e.*, la racine carrée de la variabilité de l'estimateur autour de son espérance (autrement dit, on appelle erreur-standard l'écart-type de la loi de l'estimateur). Le principe général du bootstrap consiste à substituer aux calculs analytiques la puissance des simulations, spécialement dans les cas, fréquents en pratique, où établir l'équation de l'erreur-standard (par exemple) constitue un problème "réfractaire" aux outils de la statistique mathématique. C'est pourquoi le domaine d'application privilégié du bootstrap est celui de l'estimation à partir de données issues de protocoles complexes* (e.g.*, plans d'échantillonnage combinant plusieurs stratégies, modèles régressifs paramétriques ou non, séries temporelles, ...). Par souci de clarté, le principe général du bootstrap sera néanmoins exposé à partir d'un exemple simple.*

• Supposons que l'on échantillonne un processus aléatoire de loi inconnue F, caractérisée par un paramètre θ que l'on souhaite estimer ($\theta \in \Theta, \Theta \subset \mathbb{R}^p$) ; dans de nombreux cas, ce paramètre θ peut être considéré comme une fonction de la fonction de répartition F, c'est-à-dire comme une fonctionnelle, ce que l'on notera : $\theta = T(F)$. Prenons un exemple élémentaire : le processus étudié est une variable aléatoire réelle continue X, et $\theta = E[X]$; on définit T comme suit :

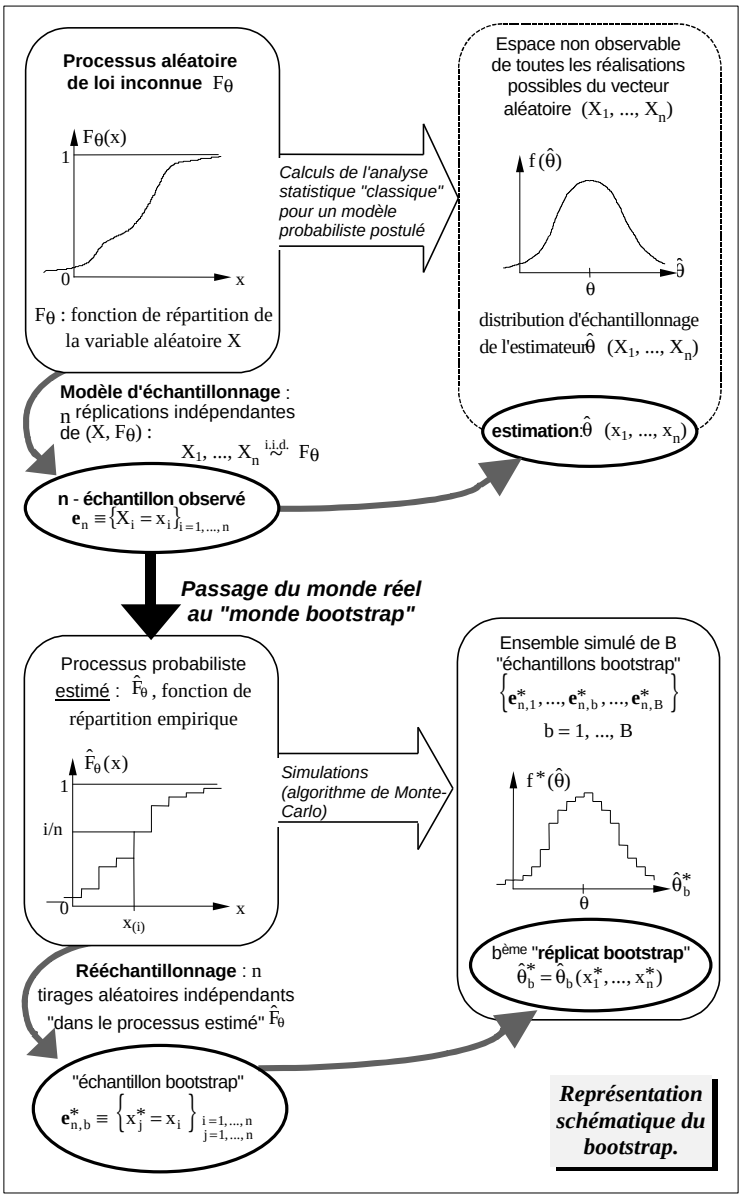
$$\{\text{ensemble des lois de probabilité sur } \mathbb{R}\} \xrightarrow{T} \Theta \text{ ; ici : } \Theta = \mathbb{R} \text{ et } \theta = T(F) = \int_{\mathbb{R}} x\, dF(x)$$

Soit le modèle d'échantillonnage $X_1, ..., X_n \overset{\text{i.i.d.}}{\sim} F$. L'*estimateur naturel* du paramètre θ est défini par : $\hat{\theta} = T(\hat{F}_n)$, où $\hat{F}_n$ est la fonction de répartition empirique associée au n-échantillon $\mathbf{e}_n = \{X_1 = x_1, ..., X_n = x_n\}$, *i.e.* :

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{\{x < x_i\}} \text{ ; } I_{\{\,\}} \text{ désigne la fonction indicatrice.}$$

On note que $\hat{F}_n$ est l'estimation non paramétrique de F au sens du maximum de vraisemblance. Si θ est l'espérance de F, on a évidemment : $T(\hat{F}_n) = (1/n) \sum x_i$. La question centrale est alors la suivante : que nous apprend $\hat{\theta}$ sur la valeur inconnue θ ? cela conduit à s'interroger sur l'exactitude et sur la précision avec lesquelles la quantité inconnue θ est approchée par $\hat{\theta}$. Ces deux critères peuvent être quantifiés si l'on connaît la distribution d'échantillonnage de l'estimateur $\hat{\theta}$: on parlera d'exactitude (ou d'absence de biais) si $E_F[\hat{\theta}] = \theta$, et de précision si $Var_F[\hat{\theta}]$ est "petite". L'exemple choisi ($\hat{\theta}$ moyenne empirique) est à cet égard particulièrement simple : le théorème de la limite centrale garantit que dans une grande variété de situations, la loi de $\hat{\theta}$ tend vers la normalité lorsque n augmente ; de plus, on sait que la moyenne empirique est non biaisée, et aussi que : $Var_F[\hat{\theta}] = Var[X]/n$. Il s'agit ici d'un problème d'estimation ponctuelle ; mentionnons aussi celui de l'estimation par intervalle, qui consiste à définir un domaine ayant une probabilité $1 - \alpha$ fixée *a priori* de contenir la valeur inconnue θ ; si $\Theta = \mathbb{R}$, on construira deux estimateurs $\hat{\theta}_L$ et $\hat{\theta}_U$ tels que : $\text{Proba}\{\hat{\theta}_L \le \theta \le \hat{\theta}_U\} = 1 - \alpha$, où α représente le risque que l'intervalle ne renferme pas la vraie valeur θ.

• Le principal intérêt du bootstrap est qu'il permet d'estimer l'erreur-standard, le biais, ou de calculer des intervalles de confiance *quel que soit le degré de complexité de la fonctionnelle* T. La démarche repose sur le "plug-in principle", qui consiste à estimer le paramètre $\theta = T(F)$ par la statistique $\hat{\theta} = T(\hat{F}_n)$. La méthode est justifiée par le fait que $\hat{F}_n$ est une statistique exhaustive pour la loi F, *i.e.*, que $\hat{F}_n$ "contient autant d'information" sur F que l'échantillon observé $\{x_1, ..., x_n\}$; si l'on dispose d'une source d'information autre que l'échantillon observé, il conviendra de l'utiliser : par exemple, si l'on sait *a priori* que la loi F appartient à une famille connue de modèles paramétriques, on recourra au bootstrap paramétrique.



Représentation schématique du bootstrap.

La figure ci-contre illustre le concept de rééchantillonnage appliqué au bootstrap. Dans la partie supérieure, on reconnaît les trois éléments de base qui fondent l'inférence statistique dans le cadre de l'estimation : *l'état de la Nature*, inconnu, ici le processus probabiliste de loi F_θ, *l'échantillon aléatoire* de taille n, dont la réalisation engendre *l'univers de tous les échantillons que l'on aurait pu observer*. De ces trois entités, seul le n-échantillon est observable. Ce que l'on peut résumer par :

F_θ (processus probabiliste inconnu) ⟶ $\{x_1, ..., x_n\}$ (échantillon observé)

L'objectif du bootstrap est de recréér de manière approchée les entités non observables, de façon à pouvoir simuler la répétition d'autant "d'échantillons virtuels" que l'on veut, *i.e.* :

$\hat F_\theta$ (processus probabiliste estimé) ⟶ $\{x^*_1, ..., x^*_n\}$ (échantillon bootstrap)

Bien évidemment, l'étape cruciale est le passage du monde réel aux entités virtuelles accessibles :

$\{x_1, ..., x_n\}$ (échantillon observé) ⟶ $\hat F_\theta$ (processus probabiliste estimé)

D'après ce qui précède, l'échantillon bootstrap est issu de n tirages *indépendants*, et opérés *avec remise*, des n valeurs x_i de l'échantillon observé ; si par exemple n = 7, on pourra obtenir $\{x_4, x_2, x_7, x_3, x_7, x_4, x_6\}$, que l'on notera : $\mathbf{e}^*_7 = \{x^*_1, x^*_2, \ldots, x^*_7\}$.

• L'algorithme est alors le suivant :

(1) on commence par créer B échantillons bootstrap indépendants, notés $\mathbf{e}^*_{n,1}$, ..., $\mathbf{e}^*_{n,b}$, ..., $\mathbf{e}^*_{n,B}$; pour estimer une erreur-standard, on choisira pour B une valeur généralement comprise entre 50 et 200 ; le calcul d'un intervalle de confiance nécessitera une valeur de B dix fois plus grande. Remarquons que le nombre maximum d'échantillons bootstrap distincts que l'on peut engendrer à partir de n observations est $\binom{2n-1}{n}$, nombre qui croît rapidement avec n (il est égal à 35 quand n = 4, et à 77 558 710 quand n = 15) ;

(2) à chaque échantillon bootstrap correspond un "réplicat bootstrap" $\hat\theta^*_b$, *i.e.*, l'estimation $\hat\theta$ calculée pour $\mathbf{e}^*_{n,b}$;

(3) l'estimateur bootstrap de $SE_F[\hat\theta]$, erreur-standard de $\hat\theta$, est $\hat{SE}_{\hat F}[\hat\theta^*]$. On l'approche par :

$$\hat{SE}_B[\hat\theta^*] = \left\{ \sum_{b=1}^{B} [\hat\theta^*_b - \hat\theta^*_{(.)}]^2 \Big/ (B-1) \right\}^{1/2}, \quad \text{avec :} \quad \hat\theta^*_{(.)} = \sum_{b=1}^{B} \hat\theta^*_b \Big/ B$$

Comme toute statistique, les estimateurs bootstrap sont entachés d'une erreur, dont la source est ici double : d'abord la variabilité d'échantillonnage, due à ce que l'on n'observe qu'une seule réalisation du n-uple de variables aléatoires i.i.d. $(X_1, ..., X_n)$, et ensuite la variabilité du rééchantillonnage lui-même, due à ce que l'on ne forme que B échantillons bootstrap. Si l'on s'intéresse à la "stabilité" de l'estimateur bootstrap de l'erreur-standard présenté ci-dessus, on montre que sa variance est voisine de $c_1/n^2 + c_2/(n.B)$, où les constantes c_1 et c_2 dépendent de la loi inconnue F, mais pas de la taille d'échantillon n ni du nombre B de rééchantillons.

• Le bootstrap n'est pas la seule méthode de rééchantillonnage, et il existe en particulier une technique plus ancienne et bien connue, celle du *jackknife* ; pour éclairer la relation entre bootstrap et jackknife, il est utile d'introduire la notion de vecteur multinomial de rééchantillonnage. En effet, la création d'un échantillon bootstrap consiste en n répétitions indépendantes d'une épreuve qui possède n événements élémentaires (en l'occurrence, les n valeurs fixées $x_1, x_2, ..., x_n$ de l'échantillon observé) dont les probabilités de réalisation sont $p_i^0 = \text{Proba}\{x_i\} = 1/n$, $\forall\, i \in [1, 2, ..., n]$. Soit P_i^* (où $i = 1, 2, ..., n$) la variable aléatoire discrète à valeurs dans {0, 1/n, 2/n, ..., 1} ; P_i^* représente la proportion de valeurs x^* égales à x_i dans un n-échantillon bootstrap, elle est donc définie par $P_i^* = \#\{ x^* = x_i\}/n$; (le symbole # signifie "nombre de fois où"). Et soit $\mathbf{P}^*$ le vecteur de rééchantillonnage défini par : $\mathbf{P}^* = (P_1^*, ..., P_n^*)$, et : $\sum_{i=1}^n P_i^* = 1$. Il est clair que le vecteur aléatoire $n\mathbf{P}^*$ suit une loi multinômiale de paramètres n et $\mathbf{p}^0 = (1/n, 1/n, ..., 1/n)$, et que créer B échantillons bootstrap équivaut à engendrer B réalisations indépendantes $\mathbf{P}_b^*$ du vecteur aléatoire $\mathbf{P}^* \sim \mathit{Mult}(n, \mathbf{p}^0)/n$. En effet, à chaque réalisation de $\mathbf{P}^*$ correspond une fonction de répartition empirique repondérée $\hat{F}^*$, qui attribue les masses $P_1^*, ..., P_n^*$ respectivement aux observations $x_1, ..., x_n$ (on rappelle que $\hat{F}_n$ attribue la masse 1/n à chaque x_i) ; par conséquent, à la b-ème réalisation $\mathbf{P}_b^*$ de $\mathbf{P}^*$ correspond aussi le réplicat bootstrap $\hat{\theta}_b^* = T(\hat{F}_b^*)$, que l'on notera $T(\mathbf{P}_b^*)$. Pour résumer, l'estimateur bootstrap est obtenu par un algorithme de Monte-Carlo, qui exécute B tirages indépendants dans la loi $\mathit{Mult}(n, \mathbf{p}^0)/n$. On estime le paramètre $\theta = T(F)$ par la statistique $\hat{\theta} = T(\hat{F}_n) = T(\mathbf{p}^0)$, et l'écart-type bootstrap de l'estimateur par :

$$\hat{SE}_B[\hat{\theta}^*] = \sqrt{\hat{Var}_*[\, T(\mathbf{P}^*)\,]}\ ,\ \text{où}\ \hat{Var}_*\ \text{est la variance estimée à l'aide de 1, 2, ..., b, ..., B réplicats simulés}\ T(\mathbf{P}_b^*).$$

• L'estimateur jackknife ne fait quant à lui intervenir que les n vecteurs :

$$\mathbf{p}_{(-i)} = (\tfrac{1}{n-1},\ \tfrac{1}{n-1},\ \ldots,\ 0,\ \ldots,\ \tfrac{1}{n-1}),\ \text{où la valeur 0 occupe la i-ème place,}$$

à chaque $\mathbf{p}_{(-i)}$ sont associés un "échantillon jackknife" $\{x_1, ..., x_{i-1}, x_{i+1}, ..., x_n\}$ de taille $n - 1$, et une fonction de répartition empirique notée $\hat{F}_{n-1}^{(-i)}$. Le i-ème des n "réplicats jackknife" est évidemment : $\hat{\theta}_{(-i)} = T(\hat{F}_{n-1}^{(-i)}) = T(\mathbf{p}_{(-i)})$. Le jackknife a été à l'origine conçu comme une technique de correction du biais, estimé par : $\hat{b}_J = (n-1)(\hat{\theta}_{(\cdot)} - \hat{\theta})$, où $\hat{\theta}_{(\cdot)} = \sum_{i=1}^n \hat{\theta}_{(-i)}/n$, et $\hat{\theta} = T(\hat{F}_n) = T(\mathbf{p}^0)$; quant à l'estimateur jackknife de l'erreur-standard de $\hat{\theta}$, il s'exprime :

$$\hat{SE}_J(\hat{\theta}) = \left\{ \tfrac{n-1}{n} \sum_{i=1}^n [\, \hat{\theta}_{(-i)} - \hat{\theta}_{(\cdot)}]^2 \right\}^{1/2}$$

Considérons maintenant une statistique linéaire $T^{LIN}(\mathbf{P}^*) = c_0 + (\mathbf{P}^* - \mathbf{p}^0)\mathbf{u}$, où c_0 est un constante, et $\mathbf{u}$ un vecteur $n \times 1$ dont les composantes u_i vérifient $\sum_{i=1}^n u_i = 0$; T^{LIN} associe un hyperplan au n-dimensionnel simplexe des vecteurs $\mathbf{P}^*$ qui vérifient les contraintes $0 \le P_i^* \le 1$ et $\sum_{i=1}^n P_i^* = 1$. Notons T_J^{LIN} la statistique qui définit l'unique hyperplan contenant les n "points jackknife" $[\mathbf{p}_{(-i)}, T(\mathbf{p}_{(-i)})]_{i = 1, ..., n}$: c'est l'approximation linéaire en $\mathbf{P}^*$ de $T(\mathbf{P}^*)$ qui coïncide avec $T(\mathbf{p}_{(-i)})$ en chacun des n échantillons jackknife $\{x_1, ..., x_{i-1}, x_{i+1}, ..., x_n\}$. Analytiquement, T_J^{LIN} est définie par : $c_0 = \hat{\theta} = T(\mathbf{p}^0)$, et par : $u_i = (n-1)[\, \hat{\theta}_{(-i)} - \hat{\theta}_{(\cdot)}]$; on montre que :

$$\hat{SE}_J(\hat{\theta}) = \left\{ \tfrac{n}{n-1} \hat{Var}_*[\, T_J^{LIN}(\mathbf{P}^*)] \right\}^{1/2}$$

Autrement dit, *l'estimateur jaccknife de la variance de* $\hat{\theta} = T(\hat{F}_n)$ *est égal à l'estimateur bootstrap de la variance de la statistique linéaire* T_J^{LIN}, au facteur multiplicatif n/(n–1) près (ce facteur garantit l'absence de biais quand $\hat{\theta} = \overline{X}$). Par conséquent, la qualité de l'approximation par le jackknife (qui ne nécessite que n réplicats) d'une "statistique bootstrap" (qui requiert le calcul de B réplicats) dépend de la qualité de l'approximation de $T(\mathbf{P}^*)$ par T_J^{LIN}. L'efficacité du jackknife se détériore avec l'augmentation de la non-linéarité de T : un exemple classique est l'incapacité du jackknife à estimer l'erreur-standard de la médiane.

• ***Relation entre le bootstrap et la delta-méthode non paramétrique*** : la delta-méthode est une technique d'estimation des variances applicable dans un cas particulier, celui où l'on a affaire à *des statistiques qui sont des fonctions de moyennes empiriques*. Contrairement au bootstrap, qu'elle a précédé d'environ 150 ans, elle ne recourt pas à un algorithme de Monte-Carlo, mais repose sur un développement de Taylor au voisinage des espérances (d'où ses autres appellations : *method of statistical differentials*, *propagation of errors formula*, *Taylor series method*). La delta-méthode est réputée posséder une tendance à sous-estimer les variances.

Soient m variables aléatoires $X_1,\ldots,X_i,\ldots,X_m$, avec $E[X_i] = \mu_i$ et $E[(X_i - \mu_i)(X_j - \mu_j)] = Cov(X_i, X_j)$ pour i, j = 1, ..., m ; et soit une fonction dérivable $\varphi(X_1,\ldots,X_m)$, notée $\varphi(\mathbf{X})$. On définit :

$$\Delta_i(m) = \partial_{X_i}\varphi(\mathbf{X})\big|_{X_1=\mu_1,\ ...,\ X_m=\mu_m} \qquad \Delta_{ij}(m) = \partial^2_{X_iX_j}\varphi(\mathbf{X})\big|_{X_1=\mu_1,\ ...,\ X_m=\mu_m}$$

Avec ces notations, le développement de Taylor de $\varphi(\mathbf{X})$ au voisinage du vecteur m des espérances, et jusqu'à l'ordre 2, s'écrit : $\varphi(\mathbf{X}) \approx \varphi(m) + \sum_{i=1}^{m}(X_i-\mu_i).\Delta_i(m) + \sum_i\sum_j(X_i-\mu_i)(X_j-\mu_j)\Delta_{ij}(m)/2$

Application 1 : *Approximation à l'ordre 2 du biais de* $\varphi(\mathbf{X})$

$$E[\varphi(\mathbf{X})] - \varphi(m) \approx \sum_{i=1}^{m}\Delta_{ii}(m)Var(X_i)/2 + \sum\sum_{i<j}\Delta_{ij}(m)Cov(X_i, X_j)$$

Application 2 : *Approximation à l'ordre 1 de la variance de* $\varphi(\mathbf{X})$

Première simplification: on néglige le biais éventuel, *i.e.*, $Var[\varphi(\mathbf{X})] = E[\varphi(\mathbf{X}) - E(\varphi(\mathbf{X}))]^2 \approx E[\varphi(\mathbf{X}) - \varphi(m)]^2$

Seconde simplification: on ne retient que les termes d'ordre 1 du développement de Taylor ;

$$Var[\varphi(\mathbf{X})] \approx E\left[\sum_{i=1}^{m}(X_i-\mu_i).\Delta_i(m)\right]^2 \approx E\left[\sum_i\sum_j(X_i-\mu_i)(X_j-\mu_j)\Delta_i(m)\Delta_j(m)\right]$$

D'où la valeur approchée de la variance vraie de $\varphi(\mathbf{X})$:

$$Var[\varphi(\mathbf{X})] \approx \sum_{i=1}^{m}\Delta_i^2(m)Var(X_i) + 2\sum\sum_{i<j}\Delta_i(m)\Delta_j(m)\,Cov(X_i, X_j)$$

Lorsque φ est une fonction linéaire des X_i, *i.e.*, de la forme $\varphi(\mathbf{X}) = \sum_{i=1}^{m}\lambda_iX_i$, où les λ_i sont des constantes réelles, la formule ci-dessus est alors exacte (avec bien évidemment $\Delta_i = \lambda_i$). Dans le cas général, un exemple d'application classiquement cité est celui du coefficient de corrélation. Pour faire apparaître qu'il s'agit bien d'une statistique fonction de moyennes empiriques, on va l'écrire : $\varphi(\mathbf{X}) = \varphi(X_1, X_2) = r(\overline{Q}_1, \overline{Q}_2, ..., \overline{Q}_5)$, où $Q_1(\mathbf{X}) = X_1$, $Q_2(\mathbf{X}) = X_2$, $Q_3(\mathbf{X}) = X_1^2$, $Q_4(\mathbf{X}) = X_1X_2$, et $Q_5(\mathbf{X}) = X_2^2$. Soit n le nombre de réalisations du vecteur aléatoire $\mathbf{X}$; avec ces notations : $\varphi(\mathbf{X}) = r(\overline{Q}_1,\ldots,\overline{Q}_5) = (\overline{Q}_4 - \overline{Q}_1\overline{Q}_2)/[(\overline{Q}_3 - \overline{Q}_1^2)^{1/2}(\overline{Q}_5 - \overline{Q}_2^2)^{1/2}]$. On montre que :

$$Var[\varphi(\mathbf{X})] \approx \frac{r^2}{4n}\left\{\frac{\mu_{40}}{\mu_{20}^2} + \frac{\mu_{04}}{\mu_{02}^2} + \frac{2\mu_{22}}{\mu_{20}\mu_{02}} + \frac{4\mu_{22}}{\mu_{11}^2} - \frac{4\mu_{31}}{\mu_{11}\mu_{20}} - \frac{4\mu_{13}}{\mu_{11}\mu_{02}}\right\}$$

où $\mu_{\alpha\beta} = E[(X_1 - E(X_1))^\alpha(X_2 - E(X_2))^\beta]$. La *delta-méthode non paramétrique* estime la variance en remplaçant r par $\hat{r}$, et les moments $\mu_{\alpha\beta}$ par les moments empiriques $\hat{\mu}_{\alpha\beta} = \sum_{i=1}^{n}(x_{1i} - \bar{x}_1)^\alpha(x_{2i} - \bar{x}_2)^\beta$; il existe entre le bootstrap et la delta-méthode non paramétrique un terme de comparaison non trivial : celle-ci fournit des résultats identiques au *jackknife infinitésimal* appliqué à des fonctions de moyennes, l'estimation de variance fournie par ce dernier étant, avec les notations déjà employées, définie par :

$$\hat{Var}_{IJ}(\hat{\theta}) = \hat{Var}_*[T^{TAN}(\mathbf{P}^*)], \text{ où : } T^{TAN}(\mathbf{P}^*) = T(\mathbf{p}^0) + (\mathbf{P}^* - \mathbf{p}^0)\mathbf{U},$$

La statistique linéaire T^{TAN} désigne cette fois l'hyperplan tangent à $T(\mathbf{P}^*)$ en $T(\mathbf{p}^0)$. Les n composantes du vecteur colonne $\mathbf{U}$ sont les *influences empiriques* $U_i = \lim_{\varepsilon\to 0}[T(\mathbf{p}^0 + \varepsilon(\mathbf{e}_i - \mathbf{p}^0)) - T(\mathbf{p}^0)]/\varepsilon$, avec : $\mathbf{e}_i$ = (0, ..., 0, 1, 0, ..., 0), la valeur 1 occupant la i-ème place. L'influence empirique peut aussi s'écrire $U_i = \partial_\varepsilon(\hat{F}_i^{(\varepsilon)})\big|_{\varepsilon=0}$, où $\hat{F}_i^{(\varepsilon)}$ est la "contaminée de $\hat{F}$ en x_i au taux ε", *i.e.*, $\hat{F}_i^{(\varepsilon)}$ attribue la probabilité $(1-\varepsilon)/n + \varepsilon$ à x_i, et les probabilités $(1-\varepsilon)/n$ aux n–1 observations $x_j \neq x_i$. B. Efron a montré que la variance estimée par la delta-méthode non paramétrique est identique à celle obtenue par le jackknife infinitésimal, *i.e.*, par la formule suggérée par L. Jaeckel : $\hat{Var}_{IJ}(\hat{\theta}) = \sum_{i=1}^{n}U_i^2/n^2$.

Outre les travaux cités en préambule, Bradley Efron a consacré à la théorie et aux applications du bootstrap de nombreux articles ; mentionnons entre autres *SIAM Review* **21**(4) : 460-480 (1979), *Biometrika* **68** (3) : 589-599 (1981), *Can. J. Statist.* **9** (2) : 139-172 (1981), *J. Amer. Statist. Assoc.* **78**(382) :316-331 (1983), *SIAM CBMS-NSF* **38**, 92 p. (1982), *Stat. Sci.* **1**(1) : 54-77 (1986), *J. Amer. Statist. Assoc.* **82**(397) :171-185 (1987), *Stat. Sci.* **11**(2) : 189-228 (1996). Citons enfin l'article de vulgarisation paru dans la revue *Scientific American* **248**(5) : 96-108 (1983), et aussi celui de présentation générale publié dans *Science, Wash.* **253** : 390-395 (1991).