Cours d'analyse de données

L3 MIAGE (Informatique) S1, Université Paris Dauphine – PSL

Pierre Wolinski[1]

2025–2026

1. Maître de conférences, Université Paris Dauphine – PSL.

Table des matières

Chapitre 1

Introduction

L'analyse de données est une discipline qui a pour but de traiter des ensembles de données afin d'en extraire des informations utiles. Ces informations peuvent prendre la forme de graphiques, de statistiques, de tableaux de données, ou encore de modèles, qui synthétisent l'information contenue dans l'ensemble de données. Celles-ci peuvent ensuite être utilisées pour divers usages, tels que :

1. l'analyse exploratoire des données : les données sont traitées de façon à trouver des liens entre les différentes composantes d'un problème. Le but est d'identifier les aspects d'un problème qui nécessitent une étude plus approfondie ;
2. l'inférence : le but de l'inférence est d'utiliser l'ensemble des données disponibles pour prédire des données qui n'ont pas encore été observées. Cela est utile pour l'aide à la décision et pour la prédiction ;
3. la compression : la compression de l'ensemble de données permet d'en diminuer la taille et de diminuer le temps de calcul nécessaire à leur traitement ;
4. la modélisation : en représentant l'ensemble de données avec un modèle mathématique simple, on en facilite tout traitement mathématique ultérieur (simulation de nouvelles données, évaluation de risque, etc.).

1.1 Buts du cours

Ce cours a pour but de fournir les outils de base en analyse de données, notamment :

1. outils statistiques et graphiques pour l'analyse univariée : moyenne, médiane, écart-type, quartiles, histogrammes, boîtes à moustaches...
2. outils statistiques pour l'analyse de données bivariée : covariance, corrélation de Pearson, corrélation de Spearman, tableau des corrélations ;
3. analyse multivariée pour l'inférence : régression linéaire simple, régression linéaire multiple, analyse statistique de la régression linéaire ;
4. analyse multivariée pour l'analyse de données exploratoire, la visualisation et la réduction de dimension : analyse en composantes principales ;
5. tests statistiques pour l'analyse de données.

1.2 Les données

Définition 1 (Tableau de données)**.** *Un* tableau de données *est la représentation d'un ensemble de données sous la forme d'un tableau à double entrée où :*

1. *chaque colonne représente une* variable *(ou* attribut*, ou* champ*) ;*
2. *chaque ligne représente une* observation *;*
3. *la case située ligne i, colonne j contient la* valeur *de la variable j pour l'observation i.*

Définition 2 (Observation)**.** *Une* observation *est un objet singulier qui a été observé (individu, événement, cas d'étude, etc.).*

Définition 3 (Variable)**.** *Les* variables *sont les caractéristiques mesurables des observations.*

Définition 4 (Population)**.** *La* population *est l'ensemble des observations possibles muni de la probabilité d'occurrence de chacune d'elles. En général, nous n'y avons pas accès.*

Définition 5 (Échantillon)**.** *Un* échantillon *est un ensemble (fini) d'observations tiré dans la population.*

Définition 6 (Série de données)**.** *Une* série de données *est la liste des valeurs prises par une variable dans un échantillon. Par exemple, la colonne d'un tableau de données est une série de données.*

Un tableau de données est donc une représentation limitée de la réalité. D'un côté, il ne contient qu'un échantillon de la population totale, ce qui ne nous donne pas accès aux observations rares, et peut nous fournir une représentation biaisée de la réalité. De l'autre, nous n'avons accès aux observations que via les variables, c'est-à-dire les mesures qui ont été effectuées. Certaines caractéristiques des objets observés sont donc manquantes.

1.3 Les variables

Opérations sur les variables. On commence par distinguer deux types de variables : les variables catégorielles et les variables quantitatives.

Définition 7 (Variables quantitatives)**.** *Une variable* quantitative *est une valeur numérique sur laquelle il peut être intéressant d'effectuer des opérations élémentaires (addition, soustraction, multiplication, division). Par exemple : prix d'un article, taille d'un individu, nombre d'employés dans une entreprise...*

On pourra traiter les variables quantitatives avec les outils standards d'algèbre et d'analyse.

Définition 8 (Variables catégorielles, ordinales et nominales)**.** *Une variable* catégorielle *indique la catégorie à laquelle appartient chaque observation. Une variable catégorielle peut être :*

1. ordinale *: il existe une façon judicieuse de les ordonner (classement à un concours, note attribuée à un service...),*
2. nominale *: autres variables (couleur d'un article en vente, département de résidence d'un client, genre...).*

On ne peut pas effectuer les opérations d'arithmétique élémentaire sur les variables catégorielles. Mais, pour les variables ordinales, il est possible d'effectuer des opérations de comparaison.

Remarque 1. *Des valeurs entières peuvent correspondre à une variable catégorielle ou quantitative selon le contexte (nombre d'enfants dans une famille, année de sortie d'un film...).*

Plage des valeurs prises par une variable. Dans certaines circonstances (par exemple, pour la représentation graphique d'une séries de données), on doit distinguer les variables en fonction de la plage de valeurs qu'elles peuvent prendre : variables discrètes, variables continues.

Définition 9 (Variables discrètes, modalités)**.** *Une* variable discrète *est une variable pouvant prendre un nombre fini (ou infini dénombrable) de valeurs.*

La modalité *d'une variable discrète est une valeur que peut prendre cette variable.*

*L'*ensemble des modalités *est l'ensemble des valeurs que peut prendre une variable discrète.*

Définition 10 (Variables continues)**.** *Une* variable continue *est une variable pouvant prendre n'importe quelle valeur dans un intervalle donné (ou une union d'intervalles).*

1.4 Éléments d'histoire et bibliographie

Les outils d'analyse de données présentés dans ce cours ont été élaborés dans des contextes très différents depuis le début du XIXe siècle.

La méthode des moindres carrés. La première publication connue présentant une régression linéaire est celle de [Legendre, 1805], qui a été baptisée "méthode des moindres carrés". Legendre était confronté à un système d'équations linéaires *surdéterminé* : le nombre d'équations était supérieur au nombre d'inconnues. Cette situation survient lorsqu'une loi physique lie p paramètres inconnus, et qu'on dispose de n observations, avec $p < n$: on a alors p inconnues liées par n équations.

Dans cette situation, il est généralement impossible de trouver une solution qui satisfasse simultanément l'ensemble des équations. L'idée de Legendre est alors de chercher une solution *approchée*, et non exacte, au système d'équations. Il a proposé de trouver les paramètres qui minimisent la somme des *carrés des erreurs* qu'on commet à chaque équation, d'où l'appellation "méthode des moindres carrés", qui a donné la régression linéaire telle qu'elle est présentée dans ce cours.

Cette méthode, mise en œuvre par Legendre pour calculer des paramètres de l'équation de la surface de la Terre,[1] fut à nouveau utilisée en astronomie par [Hubble, 1929] dans un article célèbre pour déterminer la valeur de la "constante de Hubble".

Dans un autre registre, les recherches sur l'hérédité, dans la lignée des travaux de Charles Darwin, ont conduit à construire des méthodes d'analyse de données. Par exemple, les expériences de Galton sur des graines de *Lathyrus Odoratus* l'ont conduit à effectuer une régression linéaire et à tracer la première droite de régression en 1877 [Pearson, 1930]. L'objectif de Galton était de visualiser et quantifier l'influence de l'hérédité dans la taille des graines.

La corrélation de Pearson. À la fin du XIXe siècle, les travaux sur l'hérédité conduisent Pearson à proposer une quantité permettant de mesurer le degré de dépendance entre deux variables aléatoires gaussiennes : la *corrélation de Pearson* [Pearson, 1896].

L'analyse en composantes principales. Durant la première moitié du XXe siècle, l'analyse en composantes principales a d'abord été formulée par Pearson [Pearson, 1901], puis redécouverte et utilisée en psychologie [Hotelling, 1933]. Le but de ces travaux était de découvrir si un petit nombre d'aptitudes intellectuelles, fondamentales et indépendantes, permettait d'expliquer les scores obtenus par des élèves dans un ensemble de tâches donné (en arithmétique et en lecture).

Les tests statistiques. Les tests statistiques sont nés de la nécessité d'estimer la moyenne d'une quantité sur une population entière à partir d'un échantillon de taille limitée. En particulier, quelle taille d'échantillon faut-il choisir pour avoir une bonne estimation ? L'un des premiers tests statistiques d'usage courant est le *test de Student*. Celui-ci a été conçu par l'employé de brasserie William Sealy Gosset [Dodge, 2008], qui était chargé d'évaluer la qualité de la bière produite à partir d'un échantillon de taille limitée. Il publia ses résultats sous le nom de "Student", d'où le nom du premier test statistique : le test de Student [Student, 1908].

Bibliographie. Ce cours a été conçu à l'aide des références suivantes : [James et al., 2023], [Albright and Winston, 2020], [Heumann and Shalabh, 2016].

1. La Terre n'étant pas une sphère parfaite, plusieurs paramètres sont nécessaires pour caractériser l'excentricité de sa surface.

Chapitre 2

Analyse univariée

L'analyse univariée d'une série de données consiste à synthétiser l'information qu'elle contient, via des outils comme la moyenne, l'écart-type, ou l'histogramme. Dans le cas de l'analyse univariée de plusieurs séries de données ou d'un tableau de données, chaque série (ou colonne) est étudiée indépendamment des autres.

L'analyse univariée est utile pour comparer facilement plusieurs séries de données : leur centre (moyenne ou médiane) est-il proche ? Leur dispersion (écart-type) est-elle similaire ? Ont-elles toutes la même proportion de valeurs extrêmes ? Etc.

Ces outils permettent typiquement de comparer des séries de données obtenues de manière similaire. Par exemple, une entreprise faisant du commerce de détail peut comparer la valeur du panier de chacun de ses clients d'une année à l'autre, ou d'un point de vente à l'autre, dans le but de détecter une tendance ou de comparer le comportement des clients en fonction du point de vente.

Ce sont également des outils d'analyse exploratoire des données, que l'on peut appliquer à chaque colonne d'un tableau de données avant d'utiliser des outils plus sophistiqués.

2.1 Rappels sur les probabilités

Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace probabilisé. L'ensemble Ω est l'ensemble de toutes les issues possibles $\omega \in \Omega$ d'une expérience aléatoire, $\mathcal{F}$ est l'ensemble des parties $F \subset \Omega$ mesurables de Ω (qu'on appelle aussi *événements*), et $\mathbb{P}$ est une mesure de probabilité.

Exemple 1 (Lancer de pièce équilibrée)**.** *On modélise le lancer d'une pièce équilibrée, pouvant tomber sur pile (P) ou face (F) :*

- $\Omega = \{P, F\}$;
- $\mathcal{F} = \mathcal{P}(\Omega)$*, l'ensemble des parties de* Ω ;
  on a donc : $\mathcal{F} = \{\emptyset, \{P\}, \{F\}, \{P, F\}\}$;
- $\mathbb{P}(\emptyset) = 0$ *(la probabilité que l'expérience n'ait pas d'issue est nulle),*
  $\mathbb{P}(\{P\}) = \mathbb{P}(\{F\}) = \frac{1}{2}$ *(les issues P et F sont équiprobables),*
  $\mathbb{P}(\{P, F\}) = 1$ *(la probabilité que l'une des issues se réalise vaut 1).*

À partir de maintenant, pour décrire un espace probabilisé discret [1], on considérera que $\mathcal{F} = \mathcal{P}(\Omega)$, où $\mathcal{P}(\Omega)$ est l'ensemble des parties de Ω, et, pour caractériser $\mathbb{P}$, on se

1. C'est-à-dire Ω est un ensemble fini ou infini dénombrable (i.e. de même cardinal que $\mathbb{N}$).

contentera de fournir $\mathbb{P}(\{\omega\})$ pour tout $\omega \in \Omega$. Pour simplifier les notations, on pourra écrire $\mathbb{P}(\omega)$ à la place de $\mathbb{P}(\{\omega\})$.

Exemple 2 (Lancer de dé)**.** *On modélise le lancer d'un dé équilibré à six faces :*
- $\Omega = \{1, 2, 3, 4, 5, 6\}$ *;*
- $\mathbb{P}(1) = \mathbb{P}(2) = \cdots = \mathbb{P}(6) = \frac{1}{6}$.

Exemple 3 (Lancers de deux dés)**.** *On modélise les lancers indépendants de deux dés équilibrés à six faces :*
- $\Omega = \{1, 2, 3, 4, 5, 6\}^2 = \{(1, 1), (1, 2), \cdots, (6, 5), (6, 6)\}$ *;*
- *pour toute issue* $(a, b) \in \Omega$, $\mathbb{P}((a, b)) = \frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$ *; par exemple,* $\mathbb{P}((2, 5)) = \frac{1}{36}$.

Exemple 4 (n lancers d'une pièce équilibrée)**.** *On modélise n lancers indépendants d'une pièce équilibrée :*
- $\Omega = \{P, F\}^n$ *; pour* $n = 5$, Ω *contient par exemple* (P, F, P, F, F) *;*
- *pour toute issue* $\omega \in \Omega$, $\mathbb{P}(\omega) = \frac{1}{2^n}$.

Exemple 5 (Durée d'attente)**.** *On modélise la durée d'attente à un arrêt de bus :*
- $\Omega = \mathbb{R}$ *;*
- $\mathcal{F} = \mathcal{B}(\mathbb{R})$*, l'ensemble des boréliens de* $\mathbb{R}$ *;*
- *on mesure la durée d'attente en minutes et on choisit la mesure de probabilité uniforme sur* $[5, 15]$ *pour* $\mathbb{P}$*. On a ainsi, pour tout intervalle* $[a, b]$ *:* $\mathbb{P}([a, b]) = \int_a^b \frac{1}{10} \mathbb{1}_{[5,15]}(t)\, \mathrm{d}t$*, où* $\mathbb{1}_{[5,15]}$ *est la fonction indicatrice de l'intervalle* $[5, 15]$ *:* $\mathbb{1}_{[5,15]}(t) = 1$ *si* $t \in [5, 15]$ *et* 0 *sinon.*

Définition 11. *Une* variable aléatoire réelle X *est une fonction mesurable de* $(\Omega, \mathcal{F}, \mathbb{P})$ *dans* $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$*, où* $\mathcal{B}(\mathbb{R})$ *est la tribu borélienne sur* $\mathbb{R}$.

Exemple 6. *On considère le lancer indépendant de deux dés et on souhaite étudier la somme obtenue.*

Pour modéliser cette expérience, on définit Ω *comme l'ensemble des issues possibles du lancer :* $\Omega = \{(1, 1), (1, 2), \cdots (6, 5), (6, 6)\}$*. Pour tout* $\omega \in \Omega$*, la probabilité de l'événement* $\{\omega\}$ *est :* $\mathbb{P}(\{\omega\}) = \frac{1}{36}$.[2] *On définit une variable aléatoire* X *de la façon suivante :* $X((a, b)) = a + b$*.* X *représente la somme des résultats des deux dés.*

Définition 12. *La* loi d'une variable aléatoire réelle X *est la mesure de probabilité* P *telle que, pour tout* $B \in \mathcal{B}(\mathbb{R})$, $\mathrm{P}(B) = \mathbb{P}(X^{-1}(B))$*, où* $X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\}$.

Exemple 7. *On reprend l'exemple du lancer indépendant de deux dés, où* X *est une variable aléatoire représentant la somme des deux dés. On souhaite décrire la loi* P *de* X*.*

Par exemple, calculons la probabilité que X *vaille* 2 *ou* 4*. Par définition,* $\mathrm{P}(\{2, 4\}) = \mathbb{P}(X^{-1}(\{2, 4\}))$*. Or,* $X^{-1}(\{2, 4\}) = \{(1, 1)\} \cup \{(1, 3), (2, 2), (3, 1)\} = \{(1, 1), (1, 3), (2, 2), (3, 1)\}$*. Par conséquent,* $\mathbb{P}(X^{-1}(\{2, 4\})) = \mathbb{P}(\{(1, 1), (1, 3), (2, 2), (3, 1)\}) = \frac{4}{36} = \frac{1}{9}$.

Notation 1. *Dans la suite, on adoptera les notations suivantes, pour toute variable aléatoire* X *de loi* P *:*
- *pour tout réel* a, $\mathbb{P}(X = a) := \mathrm{P}(X^{-1}(\{a\}))$ *;*
- *pour tout ensemble mesurable* I, $\mathbb{P}(X \in I) := \mathrm{P}(X^{-1}(I))$.

2. Comme les résultats des deux dés sont indépendants et que les résultats du lancer d'un seul dé sont équiprobables, alors $\mathbb{P}(\{(a, b)\}) = \frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$.

2.2 Variables aléatoires discrètes et à densité

Tout au long du cours, on ne considérera que des variables aléatoires discrètes ou à densité.

Caractérisation 1 (Variable aléatoire discrète)**.** *Une variable aléatoire discrète X est une variable aléatoire à valeurs dans $\mathbb{N}$. Sa loi* P *est caractérisée par la suite $(p_n)_{n\in\mathbb{N}}$ définie par :*

$$\forall n \in \mathbb{N}, \quad \mathbb{P}(X = n) = p_n, \tag{2.1}$$

où pour tout entier n, $0 \leq p_n \leq 1$, et $\sum_{n=0}^{\infty} p_n = 1$.

Exemple 8 (Lancer d'une pièce équilibrée)**.** *Soit X une variable aléatoire modélisant le résultat du lancer d'une pièce équilibrée. Il y a deux issues possibles : pile et face, que l'on représentera respectivement par les entiers 0 et 1. On définit donc X de la façon suivante :*

$$p_0 := \mathbb{P}(X = 0) = \frac{1}{2}, \quad p_1 := \mathbb{P}(X = 1) = \frac{1}{2}. \tag{2.2}$$

Exemple 9 (Lancer d'un dé)**.** *Soit X une variable aléatoire modélisant le résultat du lancer d'un dé. Il y a six issues possibles : $\{1, 2, 3, 4, 5, 6\}$. On définit donc X à l'aide de la suite $(p_n)_n$:*

$$p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = \frac{1}{6}. \tag{2.3}$$

Ainsi, pour toute issue $i \in \{1, \cdots, 6\}$, $\mathbb{P}(X = i) = p_i = \frac{1}{6}$.

Exemple 10 (Somme des lancers de deux dés)**.** *Soit X une variable aléatoire modélisant la somme des lancers indépendants de deux dés. Les issues possibles sont : $\{2, 3, \cdots, 11, 12\}$. On définit donc X à l'aide de la suite $(p_n)_n$:*

$$p_2 = p_{12} := \frac{1}{36}, \qquad p_3 = p_{11} := \frac{2}{36}, \qquad \cdots \qquad p_6 = p_8 := \frac{5}{36}, \qquad p_7 := \frac{6}{36}. \tag{2.4}$$

Caractérisation 2 (Variable aléatoire à densité)**.** *Une variable aléatoire à densité X est une variable aléatoire à valeurs dans $\mathbb{R}$. Sa loi* P *est caractérisée par une fonction $p_X : \mathbb{R} \to \mathbb{R}^+$ vérifiant la propriété suivante :*

$$\textit{pour tout intervalle } [a, b], \quad \mathbb{P}(X \in [a, b]) = \int_a^b p_X(x)\,\mathrm{d}x, \tag{2.5}$$

où pour tout réel x, $0 \leq p_X(x)$, et $\int_{-\infty}^{\infty} p_X(x)\,\mathrm{d}x = 1$.

Exemple 11 (Variable aléatoire uniforme)**.** *Soit X une variable aléatoire réelle uniforme sur l'intervalle $[a, b]$, ce qu'on note $X \sim \mathcal{U}(a, b)$. Sa densité est définie par :*

$$p_X(x) := \frac{1}{b-a}\mathbb{1}_{[a,b]}(x), \tag{2.6}$$

où $\mathbb{1}_{[a,b]}$ est la fonction indicatrice du segment $[a, b]$:

$$\mathbb{1}_{[a,b]}(x) = \begin{cases} 1 & \textit{si } x \in [a, b], \\ 0 & \textit{si } x \notin [a, b]. \end{cases} \tag{2.7}$$

Exemple 12 (Variable aléatoire gaussienne)**.** *Soit X une variable aléatoire gaussienne de paramètres μ et σ^2, ce qu'on note $X \sim \mathcal{N}(\mu, \sigma^2)$. Sa densité est définie par :*

$$p_X(x) := \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{2.8}$$

2.2.1 Espérance d'une variable aléatoire

Définition 13 (Espérance dans le cas discret)**.** *L'espérance d'une variable aléatoire discrète X de loi* P *peut s'écrire :*

$$\mathbb{E}_{X\sim \mathrm{P}}[X] = \sum_{n=0}^{\infty} n\, p_n, \tag{2.9}$$

où $p_n = \mathbb{P}(X = n)$.

Étant donnée une fonction $f : \mathbb{N} \to \mathbb{R}$, l'espérance de la variable aléatoire $Y = f(X)$ est donnée par :

$$\mathbb{E}_{X\sim \mathrm{P}}[Y] = \mathbb{E}_{X\sim \mathrm{P}}[f(X)] = \sum_{n=0}^{\infty} f(n)\, p_n. \tag{2.10}$$

Définition 14 (Espérance dans le cas d'une variable aléatoire à densité)**.** *L'espérance d'une variable aléatoire réelle X de loi* P *dont la densité est p_X peut s'écrire :*

$$\mathbb{E}_{X\sim \mathrm{P}}[X] = \int_{-\infty}^{\infty} x\, p_X(x)\mathrm{d}x. \tag{2.11}$$

Étant donnée une fonction mesurable $f : \mathbb{R} \to \mathbb{R}$, l'espérance de la variable aléatoire $Y = f(X)$ est donnée par :

$$\mathbb{E}_{X\sim \mathrm{P}}[Y] = \mathbb{E}_{X\sim \mathrm{P}}[f(X)] = \int_{-\infty}^{\infty} f(x)\, p_X(x)\mathrm{d}x. \tag{2.12}$$

Dans la suite, on utilisera indistinctement les notations $\mathbb{E}_{X\sim \mathrm{P}}[Y]$, $\mathbb{E}_X[Y]$ et $\mathbb{E}[Y]$ en l'absence d'ambiguïté sur la loi de X ou sur la variable aléatoire intégrée dans l'espérance.

Exemple 13. *L'espérance d'une variable aléatoire n'est pas toujours définie.*

Soit X la variable aléatoire réelle telle que, pour tout entier $n \geq 1$, $\mathbb{P}(X = 2^n) = 2^{-n}$. Pour tout x ne s'écrivant pas sous la forme 2^n avec n entier strictement positif, on pose $\mathbb{P}(X = x) = 0$. La loi de X est bien définie, puisque : $\sum_{n=1}^{\infty} \mathbb{P}(X = 2^n) = \sum_{n=1}^{\infty} 2^{-n} = 1$.

Calcul de l'espérance de X : $\mathbb{E}[X] = \sum_{n=1}^{\infty} 2^{-n} \times 2^n = \sum_{n=1}^{\infty} 1 = \infty$.

2.2.2 Variance d'une variable aléatoire

Soit X une variable aléatoire discrète ou réelle à densité.

Définition 15 (Variance)**.** *La variance de X est définie par :*

$$\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]. \tag{2.13}$$

Remarque 2. *On a également :* $\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$.

Remarque 3. *Soit X une variable aléatoire gaussienne $X \sim \mathcal{N}(\mu, \sigma^2)$. Par définition, sa densité est :* $p_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$. *Un calcul montre que μ et σ^2 sont respectivement la moyenne et la variance de X :*

$$\mathbb{E}X = \int_{-\infty}^{\infty} x\, p_X(x)\mathrm{d}x = \mu, \tag{2.14}$$

$$\mathrm{Var}(X) = \int_{-\infty}^{\infty} x^2\, p_X(x)\mathrm{d}x - \left(\int_{-\infty}^{\infty} x\, p_X(x)\mathrm{d}x\right)^2 = \sigma^2. \tag{2.15}$$

(Penser à utiliser l'intégration par parties)

La variance est un indicateur de l'étalement de la distribution de X : plus la variance de X est grande, plus les valeurs prises par X sont susceptibles d'être éloignées de son espérance $\mathbb{E}[X]$.

2.2.3 Fonction de répartition

Définition 16 (Fonction de répartition)**.** *Soit X une variable aléatoire réelle. La* fonction de répartition *F_X de X est définie par :*

$$\forall x \in \mathbb{R}, \quad F_X(x) = \mathbb{P}(X \leq x). \tag{2.16}$$

Remarque 4. *Soit $F : \mathbb{R} \to [0,1]$ une fonction telle que :*

(i) F est croissante sur $\mathbb{R}$;

(ii) F est continue à droite ;

(iii) $\lim_{-\infty} F = 0$ *et* $\lim_{\infty} F = 1$.

Alors F est la fonction de répartition d'une variable aléatoire X et la loi P *de X est intégralement déterminée par F.*

Réciproquement, toute fonction de répartition vérifie les propriétés (i), (ii), (iii).

Cas particulier des variables aléatoires réelles à densité. Soit X une variable aléatoire réelle de densité $p_X : \mathbb{R} \to \mathbb{R}$ et de fonction de répartition F_X.

Remarque 5. *On a, pour tout $x \in \mathbb{R}$:*

$$p_X(x) = F_X'(x), \qquad F_X(x) = \int_{-\infty}^{x} p_X(t)\,\mathrm{d}t. \tag{2.17}$$

De plus, F_X est une fonction continue.

2.2.4 Médiane et quantiles

Soit X une variable aléatoire réelle à densité. Pour simplifier les définitions, nous supposons que F_X est strictement croissante (ou, ce qui est équivalent, que la densité p_X soit strictement positive).

Définition 17 (Médiane)**.** *La* médiane *de X est le réel $q_{1/2}$ tel que $F_X(q_{1/2}) = \frac{1}{2}$.*

Comme F_X est continue (X étant à densité) avec $\lim_{-\infty} F = 0$ et $\lim_{\infty} F = 1$, alors $q_{1/2}$ existe. Comme F_X est strictement croissante, alors $q_{1/2}$ est unique.

Dans ce cadre, la médiane $q_{1/2}$ de X est un réel divisant la distribution de X en deux parties égales : $\mathbb{P}(X \leq q_{1/2}) = \mathbb{P}(X > q_{1/2}) = \frac{1}{2}$.

Définition 18 (Quantiles)**.** *Soit $p \in]0, 1[$ une probabilité. Le* quantile *d'ordre p de X est le réel q_p tel que $F_X(q_p) = p$.*

Comme F_X est continue (X étant à densité) avec $\lim_{-\infty} F = 0$ et $\lim_{\infty} F = 1$, alors q_p existe. Comme F_X est strictement croissante, alors q_p est unique.

En d'autres termes, le quantile q_p est défini de sorte que : $\mathbb{P}(X \leq q_p) = p$.

Remarque 6. *Le quantile $q_{1/2}$ d'ordre $\frac{1}{2}$ est la médiane.*

2.3 Application à des échantillons

Dans la partie précédente, nous avons défini formellement l'espérance, la variance, la médiane et les quantiles d'une variable aléatoire X de loi P. Nous avons vu comment calculer toutes ces quantités quand nous disposons d'une caractérisation de la loi de X, comme la suite $(\mathbb{P}(X = n))_n$ dans le cas où X est une variable aléatoire discrète, ou la densité de probabilité de X dans le cas où X est réelle à densité.

En analyse de données, on cherche à effectuer l'opération inverse. Étant donné un échantillon constitué de n variables aléatoires $(X_1, \cdots, X_n)$ réelles indépendantes et identiquement distribuées, on souhaite caractériser la loi P des $(X_i)_i$. Par exemple, on peut calculer la moyenne empirique et la variance empirique de l'échantillon, ce qui nous renseigne respectivement sur l'espérance et la variance de la loi P.

Dans cette partie, nous définissons les outils de base permettant de décrire la distribution de variables quantitatives. Ces outils, tels que la moyenne empirique et la variance empirique, sont adaptés à l'étude d'une variable, indépendamment de toutes les autres. Ils nous informent sur la position du centre de la distribution d'une variable et sa dispersion, ce qui permet déjà de connaître l'ordre de grandeur de chaque variable et leur variabilité sur l'ensemble des observations. De plus, ces quantités peuvent être utilisés pour *standardiser* l'ensemble de données, c'est-à-dire transformer les données pour que chaque variable évolue sur une plage de données similaire, afin de pouvoir les combiner et les comparer plus facilement.

Pour simplifier les notations, on suppose dans cette partie que les observations $(X_1, \cdots, X_n)$ de l'ensemble de données sont des scalaires (et non des vecteurs) : $\forall i \in \{1, \cdots, n\}, X_i \in \mathbb{R}$.

Cependant, les quantités définies ici ne sont pas conçues pour nous renseigner sur les relations de dépendance entre les différentes variables. Par conséquent, il sera nécessaire d'introduire des objets mathématiques et des méthodes plus sophistiquées.

2.3.1 Moyenne

Définition 19 (Moyenne)**.** *La* moyenne *de l'échantillon* $(X_1, \cdots, X_n)$ *est définie par :*

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{2.18}$$

Théorème 1 (Loi forte des grands nombres)**.** *Soit* $(X_1, \cdots, X_n, \cdots)$ *une suite (infinie) de variables aléatoires réelles indépendantes et identiquement distribuées. Supposons que* $\mathbb{E}[|X_1|] < \infty$.[3] *On a donc :*

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow[n\to\infty]{\text{p.s.}} \mathbb{E}[X_1]. \tag{2.19}$$

Par conséquent, sous réserve d'intégrabilité de $|X_1|$, la moyenne empirique tend vers l'espérance de X_1 lorsque le nombre n d'observations tend vers l'infini. Il est donc raisonnable de vouloir estimer $\mathbb{E}[X_1]$ avec $\bar{X}$.

2.3.2 Variance empirique

Définition 20 (Variance empirique)**.** *La* variance empirique *de l'échantillon* $(X_1, \cdots, X_n)$ *est définie par :*

$$\tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2. \tag{2.20}$$

Définition 21 (Écart-type empirique)**.** *L'*écart-type *de l'échantillon* $(X_1, \cdots, X_n)$ *de variance empirique* $\tilde{\sigma}^2$ *est* $\tilde{\sigma} \geq 0$.

Remarque 7 (Correction de Bessel)**.** *La variance empirique* $\tilde{\sigma}^2$ *n'est pas le seul moyen d'estimer la variance* σ^2 *de la loi des* $(X_i)_i$*. On peut aussi utiliser la formule suivante :*

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2, \tag{2.21}$$

appelée variance empirique avec correction de Bessel.

La seule différence avec la variance empirique est la présence de $n-1$ *au dénominateur au lieu de* n*. L'estimateur* $\hat{\sigma}^2$ *de la variance* σ^2 *a la propriété d'être centré en* σ^2 *(on dit alors qu'il est* non biaisé*) :*

$$\mathbb{E}\hat{\sigma}^2 = \sigma^2. \tag{2.22}$$

Ce n'est pas le cas de la variance empirique $\tilde{\sigma}^2$*, qui a tendance à sous-estimer la variance :*

$$\mathbb{E}\tilde{\sigma}^2 = \frac{n-1}{n}\sigma^2. \tag{2.23}$$

3. Comme les X_i sont i.i.d., alors $\mathbb{E}[|X_1|] = \cdots = \mathbb{E}[|X_n|] = \cdots$. On aurait pu tout autant écrire cette condition pour X_2, X_3, etc.

2.3.3 Médiane et quantiles

Définition 22 (Médiane d'un échantillon)**.** *Soit un échantillon* $\mathsf{X} = (X_1, \cdots, X_n)$*. On note* $\mathsf{X}' = (X'_1, \cdots, X'_n)$ *l'échantillon* X *réordonné dans l'ordre croissant :* $X'_1 \leq X'_2 \leq \cdots \leq X'_n$*. La* médiane m *de l'échantillon* X *est :*

- *si* n *est impair,* $m = X'_{(n+1)/2}$*, c'est-à-dire la valeur "centrale" de l'échantillon ;*
- *si* n *est pair,* $m = \frac{1}{2}(X'_{n/2} + X'_{n/2+1})$*, c'est-à-dire la moyenne des deux valeurs "centrales" de l'échantillon.*

La médiane m *de l'échantillon partage celui-ci en deux ensembles de taille égale, l'un contenant les valeurs inférieures à* m*, l'autre contenant les valeurs supérieures à* m*.*

Exemple 14. *Soit un échantillon* $(-7, -3, 13, 12, 0, 3, -6, -5)$*. Un tri des valeurs donne :* $(-7, -6, -5, -3, 0, 3, 12, 13)$*. L'échantillon étant de taille 8, la médiane se calcule en moyennant les* 4^e *et* 5^e *valeurs :* $m = \frac{-3+0}{2} = -1.5$*.*

Remarque 8. *Contrairement à la moyenne, la médiane n'est pas sensible aux valeurs extrêmes.*

Exemple 15. *On reprend l'échantillon* $\mathsf{X} = (-7, -3, 13, 12, 0, 3, -6, -5)$*, de médiane* -1.5 *et de moyenne* 0.875*. Si l'on ajoute la valeur* 1000*, on obtient un nouvel échantillon* $\mathsf{X}' = (-7, -3, 13, 12, 0, 3, -6, -5, 1000)$*, de moyenne* 111.9 *et de médiane* 0*.*

Dans ce cas, la médiane a peu changé, mais la moyenne est devenue considérablement plus grande.

Définition 23 (Quantiles d'un échantillon)**.** *Soit un échantillon* $\mathsf{X} = (X_1, \cdots, X_n)$*. On note* $\mathsf{X}' = (X'_1, \cdots, X'_n)$ *l'échantillon* X *réordonné dans l'ordre croissant :* $X'_1 \leq X'_2 \leq \cdots \leq X'_n$*. Le* quantile d'ordre p *de l'échantillon* X*, noté* $\hat{q}_p$ *est :*

$$\begin{aligned} \hat{q}_p &:= X'_r(r - p(n-1)) + X'_{r+1}(p(n-1) - r + 1) \\ \textit{avec :} \quad r &:= \lfloor p(n-1) \rfloor + 1, \end{aligned} \tag{2.24}$$

où $\lfloor x \rfloor$ *est la partie entière de* x*. Voir [Hyndman and Fan, 1996, méthode 7].*

Remarque 9. *Lorsque* $p(n-1)$ *est un nombre entier, alors* $r = \lfloor p(n-1) \rfloor + 1 = p(n-1) + 1$*. On a donc :*

$$\hat{q}_p := X'_r. \tag{2.25}$$

Définition 24 (Quartiles et déciles)**.** *On définit des quantiles particuliers, très couramment utilisés en analyse de données :*

- *le premier quartile* $\hat{Q}_1$ *et le troisième (ou dernier) quartile* $\hat{Q}_3$ *sont définis de la façon suivante :*

$$\hat{Q}_1 := \hat{q}_{1/4} \qquad\qquad \hat{Q}_3 := \hat{q}_{3/4}; \tag{2.26}$$

- *le premier décile* $\hat{D}_1$ *et le neuvième (ou dernier) décile* $\hat{D}_9$ *sont définis de la façon suivante :*

$$\hat{D}_1 := \hat{q}_{1/10} \qquad\qquad \hat{D}_9 := \hat{q}_{9/10}. \tag{2.27}$$

Remarque 10. *Le premier quartile est la valeur séparant les* 25% *de données les plus basses des autres. Le troisième quartile est la valeur séparant les* 25% *de données les plus élevées des autres. Naturellement,* $\hat{Q}_3 \geq \hat{Q}_1$.

Définition 25 (Écart inter-quartile)**.** *L'*écart inter-quartile *est la quantité* $\mathrm{EI} := \hat{Q}_3 - \hat{Q}_1$.

Remarque 11. *La moitié des données sont situées entre le premier quartile et le troisième quartile. Il s'agit des données "les moins extrêmes".*

2.3.4 Terminologie

Définition 26 (Mode)**.** *Le mode d'une série de données discrète est sa modalité la plus fréquente.*

Exemple 16. *Soit une série de données* $(3, 1, 2, 3, 5, 4, 5, 8, 5, 10)$. *La modalité la plus fréquente est* 5*, donc son mode est* 5.

Un *indicateur de tendance centrale* est un outil permettant de mesurer la position du centre d'une série de données, comme la moyenne, la médiane, le mode...

Un *indicateur de dispersion* est un outil permettant de mesurer l'étalement typique des valeurs d'une série de données, comme l'écart-type, l'écart inter-quartile...

2.4 Outils graphiques et tableaux

Dans cette partie, nous nous intéressons à la représentation de caractéristiques d'une série de données, sous forme de tableau ou de graphique.

Au préalable, il est nécessaire de distinguer deux types de variables : les variables discrètes (voir Déf. 9) et les variables continues (voir Déf. 10).

2.4.1 Tableau d'effectifs, tableau de fréquences

Définition 27 (Modalité)**.** *Pour une variable donnée, une* modalité *est une valeur que peut prendre cette variable.*

Dans le cas d'une variable discrète, on peut définir un tableau d'effectifs et un tableau de fréquences.

Définition 28 (Tableau d'effectifs, tableau de fréquences)**.** *Le* tableau d'effectifs *d'une série de données discrètes* $\mathsf{X} = (X_1, \cdots, X_n)$ *est un tableau indiquant le nombre de fois que chaque modalité apparaît. En notant* $(a_1, \cdots, a_p)$ *les modalités de* X*, l'effectif de la modalité* a_j *s'écrit :* $n_j = \sum_{i=1}^{n} \mathbb{1}_{X_i = a_j}$.

Le tableau de fréquences *de* X *est un tableau indiquant la fréquence de chaque modalité. La fréquence de la modalité* a_j *s'écrit :* $f_j = \frac{n_j}{n}$.

Exemple 17. *Soit une série de données* $\mathsf{X} = (3, 2, 3, 1, 4, 2, 3, 1, 2, 1, 1, 1)$. *Les modalités de la variable sont :* $\{1, 2, 3, 4\}$. *Le tableau combinant les effectifs et les fréquences est :*

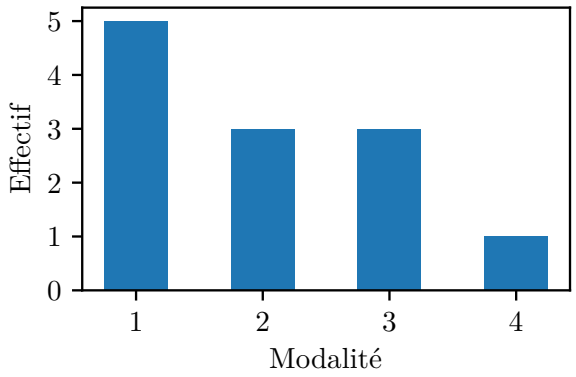
modalité	effectif	fréquence
1	5	42%
2	3	25%
3	3	25%
4	1	8%

2.4.2 Histogramme

L'histogramme est un outil permettant de visualiser le nombre ou la fréquence des différentes valeurs prises par une variable dans une série de données.

Variable discrète. Dans le cas d'une variable discrète, l'histogramme est simplement la représentation graphique du tableau d'effectifs, avec la modalité en abscisse et l'effectif en ordonnée.

Exemple 18 (Variable discrète)**.** *Soit une série de données* $(3, 2, 3, 1, 4, 2, 3, 1, 2, 1, 1, 1)$.



Variable continue. Dans le cas d'une variable continue, il est nécessaire de classer les valeurs en différents groupes avec de tracer l'histogramme.

Exemple 19 (Variable continue)**.** *On considère une variable représentant la taille d'un individu (en cm) :* $(161, 165, 165, 165, 171, 172, 173, 173, 174, 176, 176, 182)$. *Pour représenter un histogramme de cette variable, on décide de regrouper les valeurs par intervalle de taille 5 :* $[160, 165[$, $[165, 170[$, *etc. Ainsi, puisqu'on a 1 individu ayant une taille située dans* $[160, 165[$, *on trace une barre de taille 1 entre les valeurs 160 et 165, etc.*

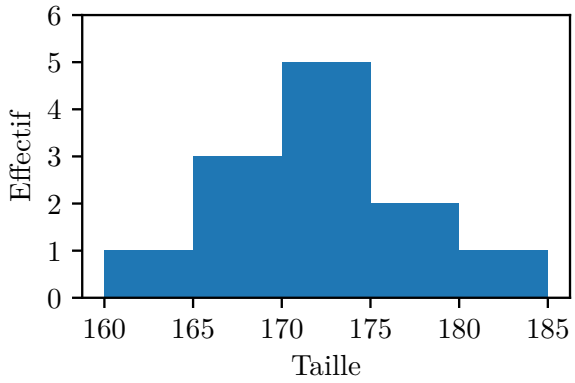
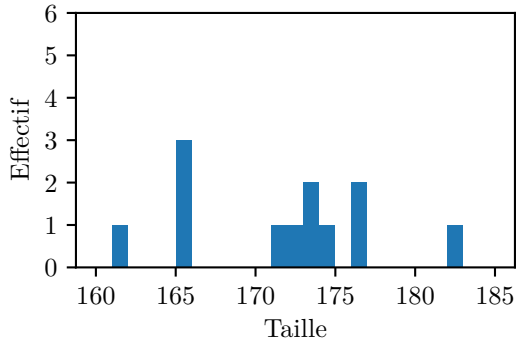
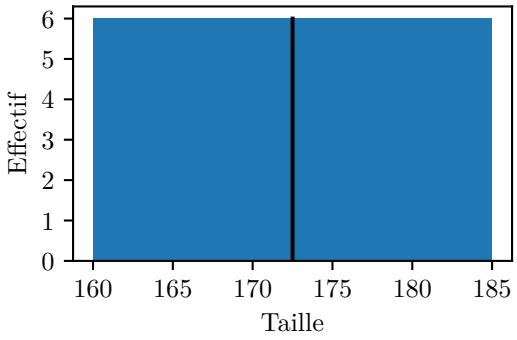


FIGURE 2.1 – Taille des intervalles : 5.

Remarque 12. *Le choix de regroupement peut changer complètement l'allure de l'histogramme. Avec les données de l'exemple précédent, on obtient les deux histogrammes suivants, lorsqu'on fixe respectivement la taille des intervalles à 1 et à 12.5.*



(a) Taille des intervalles : 1 (fin).



(b) Taille des intervalles : 12.5 (grossier).

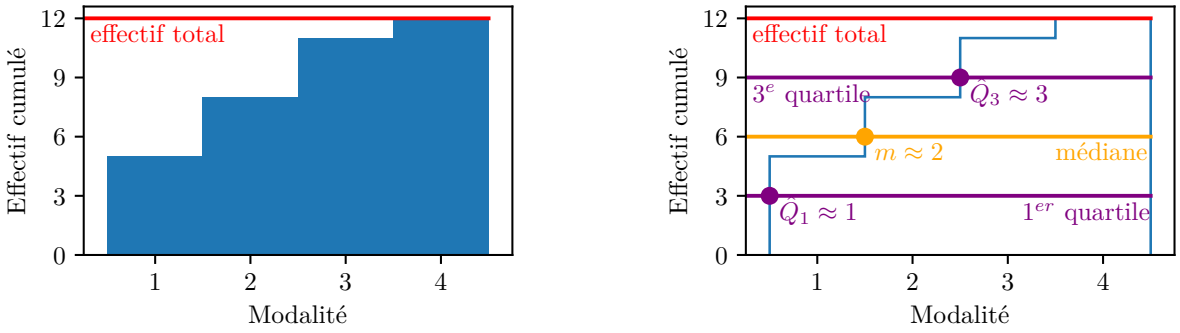
FIGURE 2.2 – Deux histogramme tracés avec la même série de données, mais différentes tailles d'intervalles.

Le choix de regroupement choisi dans la figure 2.2a est trop fin : les barres sont de hauteur très variable et il est difficile d'identifier facilement la distribution des données. Dans la figure 2.2b, le choix de regroupement est trop grossier : en choisissant des groupes trop larges, l'information contenue dans l'ensemble de données a été en grande partie oubliée. Le choix effectué dans la figure 2.1 semble le plus adapté : on arrive à distinguer la forme générale de la distribution.

2.4.3 Histogramme cumulé

Dans le cas d'une variable quantitative discrète, l'histogramme cumulé représente, pour chaque modalité a_i, la somme des effectifs de a_i et de tous les $a_j < a_i$.

Exemple 20. *On reprend la série de données* $3, 2, 3, 1, 4, 2, 3, 1, 2, 1, 1, 1$*). L'histogramme cumulé est :*



(a) Histogramme cumulé.

(b) Calcul approximatif des quantiles.

FIGURE 2.3 – Histogrammes cumulés.

Remarque 13 (Lien avec les quantiles)**.** *Comme indiqué dans l'exemple ci-dessus, il est possible d'utiliser un histogramme cumulé pour obtenir une approximation de la médiane et de différents quantiles. Par exemple, pour approximer le premier quartile, il suffit de tracer la droite horizontale ayant pour ordonnée* $1/4$ *de l'effectif total* n*, et de prendre la modalité de la barre qu'elle intersecte.*

On peut effectuer l'opération pour n'importe quel quantile d'ordre p*, en traçant la droite horizontale ayant pour ordonnée* $p \times n$*.*

Attention : *cette opération n'est pas toujours exacte.*

2.4.4 Boîte à moustaches

La *boîte à moustache* (ou *boxplot*) est une représentation graphique d'une série de données permettant de visualiser ses principales caractéristiques de concentration et de dispersion, comme la médiane, les quartiles, les valeurs minimale et maximale, ainsi que les valeurs atypiques.

En représentation horizontale, la boîte à moustaches typique a la forme suivante :

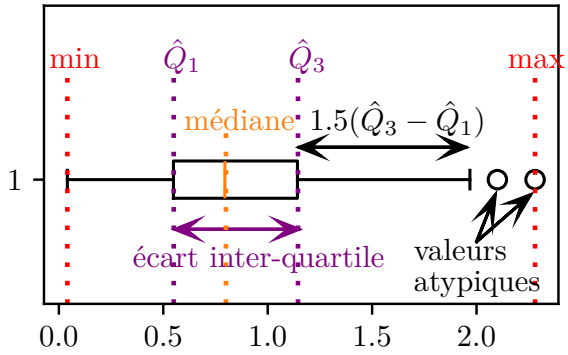


FIGURE 2.4 – Visualisation des différentes quantités résumant la forme d'une série de données sur un diagramme à moustaches.

Plus précisément :

— l'écart inter-quartile est représenté par une boîte ;

— la médiane est représentée par une barre située sur la boîte ;
— deux segments partent de la boîte de chaque côté et s'arrêtent :
  — à la valeur extrémale (minimum ou maximum) si celle-ci est située à une distance inférieure à $1.5(\hat{Q}_3 - \hat{Q}_1)$ de la boîte,
  — sinon, à une distance de $1.5(\hat{Q}_3 - \hat{Q}_1)$ de la boîte ; dans ce cas, les valeurs situées au-delà sont considérées comme atypiques et sont représentées par des points.

Exemple 21. *On peut utiliser des boîtes à moustaches pour comparer la distribution de plusieurs variables.*

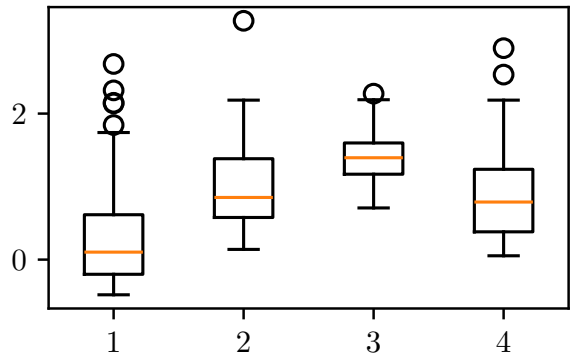
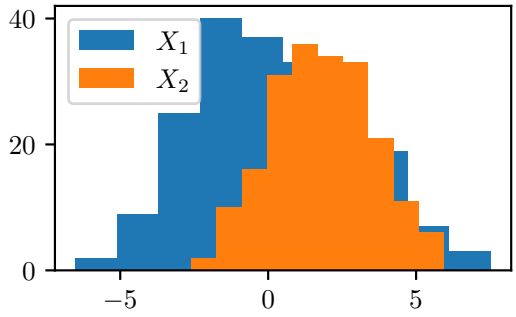


FIGURE 2.5 – Ce graphique permet d'identifier rapidement la tendance centrale et la dispersion des séries de données considérées, et de repérer celles qui contiennent des valeurs atypiques.

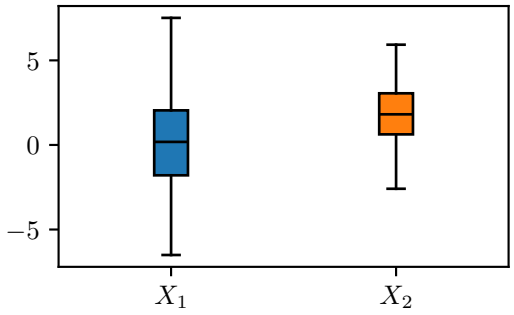
Remarque 14. *Une boîte à moustaches permet de visualiser très facilement la forme de la distribution des données :*
— *la boîte indique la plage de valeurs "centrale" ;*
— *la médiane indique si, dans cette plage de valeurs centrale, les valeurs sont distribuées de manière symétrique ou non ;*
— *la taille des segments indique si les valeurs relativement éloignées du centre sont probables ou non ;*
— *la présence de points au-delà des extrémités des segments signale l'existence le valeurs très éloignées du centre (valeurs atypiques).*

2.5 Limites de l'analyse univariée

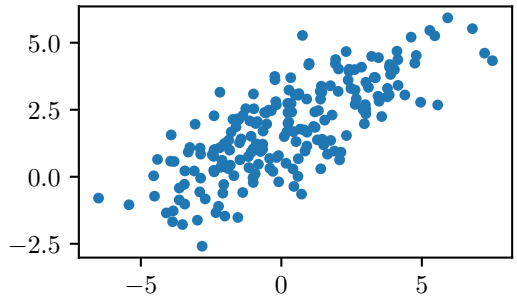
Quand on étudie un tableau de données, on peut commencer par effectuer l'analyse univariée de chaque variables. Mais cette analyse ne permet pas de capturer les relations de dépendance entre les différentes variables. Cela est illustré sur la figure 2.6.



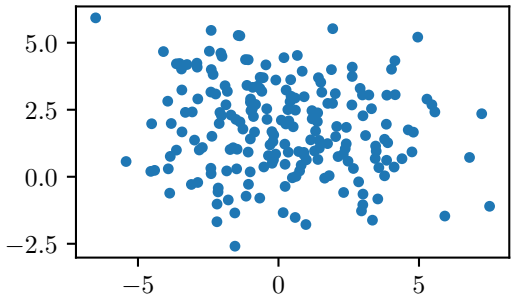
(a) Histogrammes.



(b) Boîtes à moustaches.



(c) Cas 1 : dépendance entre X_1 et X_2.



(d) Cas 2 : indépendance entre X_1 et X_2.

FIGURE 2.6 – Analyse univariée de deux tableaux de données différents, dont les variables sont notées X_1 et X_2 : dans les deux cas, les histogrammes et les boîtes à moustaches sont identiques (ligne du haut), mais les nuages de points sont différents (ligne du bas). Cela indique que l'analyse univariée est aveugle aux interactions entre les variables.

Chapitre 3

Analyse de corrélation

À partir de ce chapitre, nous ne cherchons plus à faire l'analyse de plusieurs séries de données de manière indépendante. Nous considérons un tableau de données à plusieurs variables : chaque observation est représentée par une ligne du tableau, et chaque variable correspond à une colonne. Dans ce contexte, les variables sont dépendantes, puisque les valeurs prises par les différentes variables *dans une même ligne* correspondent à *une même observation*.

Par exemple, si l'on observe plusieurs individus et que l'on reporte, pour chacun d'eux, son âge, sa taille et son poids, on s'attend à trouver une certaine structure dans le tableau de données : il serait invraisemblable d'avoir, pour une même observation, un âge de 1 an et une taille de 2,00 m, ou une taille de 2,00 m et un poids de 20 kg. Pourtant, prises indépendamment, chacune de ces valeurs est parfaitement vraisemblable : il existe des individus ayant 1 an, des individus mesurant 2,00 m, et des individus (jeunes) pesant 20 kg.

Le but d'une analyse multivariée est de tirer parti de la relation de dépendance entre les variables du tableau afin de le synthétiser efficacement. Dans le cas où l'on analyse la dépendance des variables *deux à deux*, on parle d'*analyse bivariée*. Enfin, lorsqu'on utilise comme outil la *corrélation* entre deux variables, on parle d'*analyse de corrélation*.

3.1 Nuage de points

Lorsqu'on étudie la relation entre deux variables, l'outil le plus simple à notre disposition est le *nuage de points*. C'est un outil graphique permettant de visualiser les valeurs prises par un couple de variables.

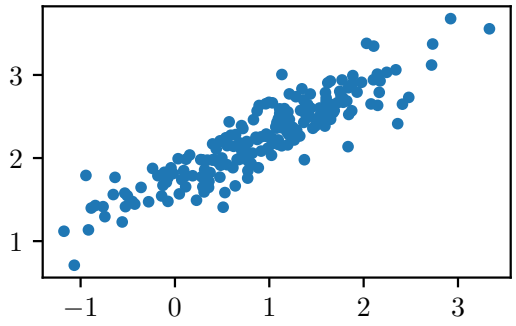
3.1.1 Définition et propriétés

Soient X et Y deux variables d'un tableau de données. On note $(X_1, \cdots, X_n)$ et $(Y_1, \cdots, Y_n)$ les séries de données associées.

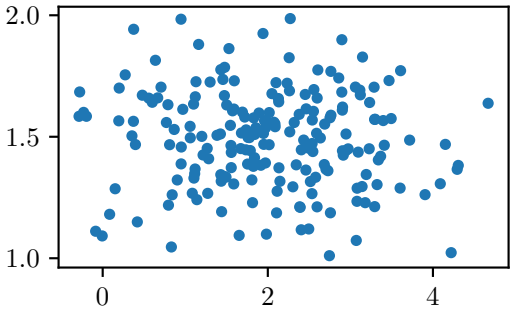
Définition 29 (Nuage de points)**.** *Le nuage de points (X, Y) est le graphique composé des points de coordonnées (X_i, Y_i), pour $i \in \{1, \cdots n\}$.*

Le nuage de points permet de mettre en évidence une relation de dépendance entre les variables X et Y, la nature de la dépendance et l'influence du bruit :

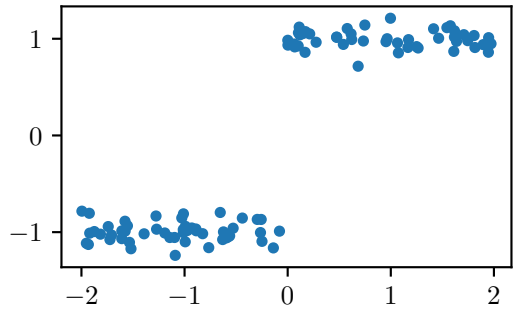
1. si le fait de connaître X nous donne de l'information sur Y (ou inversement), alors les variables sont dépendantes ;
2. lorsque deux variables sont dépendantes, on peut caractériser la nature de la dépendance. Par exemple : plus X est grand, plus Y est grand ; plus X est grand, plus Y est proche de 0 ; X et Y sont toujours de même signe...
3. lorsque deux variables sont dépendantes, on peut voir si la relation de dépendance est fortement bruitée ou non. En l'absence de bruit, le fait de connaître X_i permet de prédire Y_i avec précision. Lorsque le bruit est important, le fait de connaître X_i nous renseigne assez peu sur Y_i.



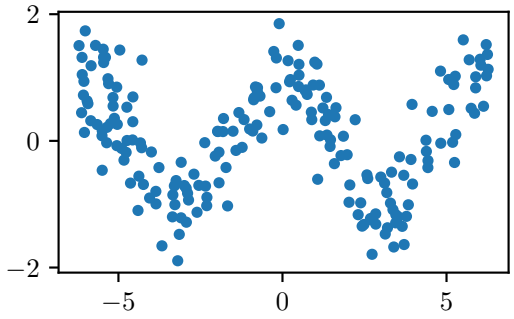
(a) Dépendance linéaire.



(b) Pas de dépendance.



(c) Dépendance non linéaire.



(d) Dépendance non linéaire, très bruitée.

FIGURE 3.1 – Nuages de points tracés pour différentes configurations de dépendance.

On parle de relation linéaire entre les variables X et Y lorsqu'on peut modéliser fidèlement le nuage de points (X, Y) par une équation affine (avec du bruit) :

$$Y_i = aX_i + b + \epsilon_i, \tag{3.1}$$

où ϵ_i est une variable aléatoire représentant le bruit, et a et b sont des nombres réels.

3.1.2 Tableau de nuages de points

Un seul nuage de points ne permet de visualiser que la relation entre deux variables. Il est néanmoins possible d'étendre le procédé à un nombre quelconque de variables : on trace, pour chaque couple de variables $(X^{(i)}, X^{(j)})$, le nuage de points correspondant, puis on place la figure obtenue dans la case de coordonnées (i, j) d'un tableau.

La figure 3.2 représente un tableau de nuages de points dans le cas où l'on étudie 3 variables (le tableau a alors 3 lignes et 3 colonnes). Un tel tableau permet d'identifier facilement les couples de variables fortement dépendantes, pour lesquels il peut être intéressant de faire une analyse plus approfondie.

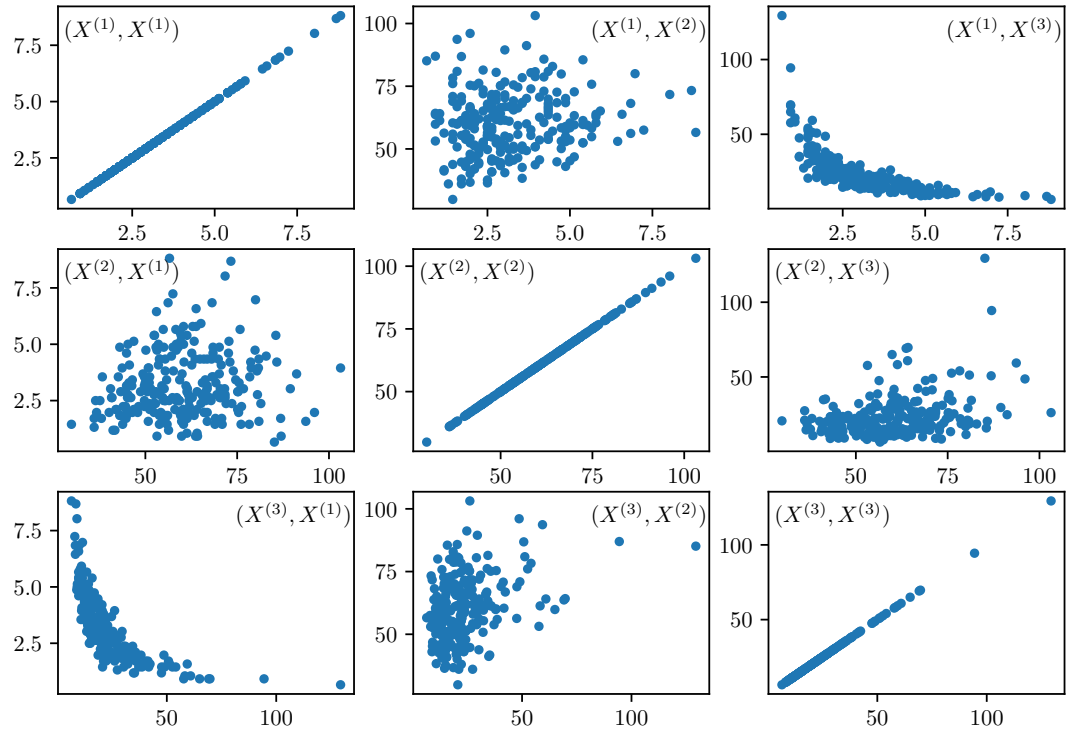


FIGURE 3.2 – Tableau de nuages de points avec 3 variables. Les 3 figures de la diagonale montrent une corrélation parfaite (points parfaitement alignés), ce qui n'est pas surprenant, puisque chaque variable est parfaitement corrélée avec elle-même. La case de coordonnées $(3, 1)$, représentant le nuage de points $(X^{(3)}, X^{(1)})$, est plus intéressante : on observe une relation de dépendance non linéaire (points concentrés autour d'une courbe, mais non alignés).

3.2 Covariance

Soient X et Y deux variables d'un tableau de données. On note $(X_1, \cdots, X_n)$ et $(Y_1, \cdots, Y_n)$ les séries de données associées. On définit ici la covariance de (X, Y).

Définition 30 (Covariance)**.** *On définit la covariance (empirique) du couple* (X, Y) *comme suit :*

$$\overline{\mathrm{Cov}}(X, Y) = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}), \tag{3.2}$$

où $\bar{X}$ *et* $\bar{Y}$ *sont respectivement la moyenne de* X *et la moyenne de* Y*.*

La covariance est l'outil fondamental qui va nous permettre d'évaluer la dépendance entre deux variables. En particulier, elle nous informe sur la partie *linéaire* de la dépendance entre celles-ci :

1. si $\overline{\mathrm{Cov}}(X, Y) = 0$, il n'y a pas de composante linéaire de la dépendance entre X et Y ;
2. si $\overline{\mathrm{Cov}}(X, Y) > 0$, alors il y a une dépendance linéaire entre X et Y, et les grandes valeurs de X sont associées à de grandes valeurs de Y ;
3. si $\overline{\mathrm{Cov}}(X, Y) < 0$, alors il y a une dépendance linéaire entre X et Y, et les grandes valeurs de X sont associées à de petites valeurs de Y.

Propriétés 1. *La covariance a les propriétés suivantes :*

1. *lien avec la variance :*

$$\overline{\mathrm{Cov}}(X, X) = \bar{\sigma}_X^2; \tag{3.3}$$

2. *symétrie :*

$$\overline{\mathrm{Cov}}(X, Y) = \overline{\mathrm{Cov}}(Y, X); \tag{3.4}$$

3. *annulation des constantes : soit $\lambda \in \mathbb{R}$:*

$$\overline{\mathrm{Cov}}(X + \lambda, Y) = \overline{\mathrm{Cov}}(X, Y). \tag{3.5}$$

4. *linéarité : soit Z une variable quelconque et soit $\lambda \in \mathbb{R}$:*

$$\overline{\mathrm{Cov}}(X + \lambda Y, Z) = \overline{\mathrm{Cov}}(X, Z) + \lambda \overline{\mathrm{Cov}}(Y, Z). \tag{3.6}$$

Remarque 15. *En règle générale, on ne peut pas tirer de conclusion précise de la valeur de $\overline{\mathrm{Cov}}(X, Y)$* seule.

Par exemple, prenons une variable $Y := \lambda X$ avec $\lambda > 0$. Il existe une relation linéaire parfaite (i.e., sans bruit) entre X et Y. Pourtant, la covariance entre X et Y vaut :

$$\overline{\mathrm{Cov}}(X, Y) = \overline{\mathrm{Cov}}(X, \lambda X) = \lambda \overline{\mathrm{Cov}}(X, X) = \lambda \bar{\sigma}_X^2, \tag{3.7}$$

qui dépend à la fois de λ et de $\bar{\sigma}_X^2$. Par conséquent, $\overline{\mathrm{Cov}}(X, Y)$ peut être aussi proche de 0 que l'on veut en prenant $\lambda \to 0$, et aussi grande que l'on veut en prenant $\lambda \to \infty$. $\overline{\mathrm{Cov}}(X, Y)$ dépend de la même façon de la valeur de $\bar{\sigma}_X^2$.

Une relation linéaire parfaite entre X et Y n'entraîne pas forcément une covariance élevée. Cet outil doit donc être affiné, notamment en tenant compte de $\bar{\sigma}_X^2$ et $\bar{\sigma}_Y^2$. Cela nous amène naturellement à la *corrélation de Pearson.*

3.3 Corrélation de Pearson

3.3.1 Définition et propriétés

La corrélation de Pearson permet de mesurer la partie linéaire de la dépendance entre deux variables.

Définition 31 (Corrélation de Pearson)**.** *On définit la corrélation de Pearson comme suit :*

$$\overline{\text{corr}}(X,Y) := \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i-\bar{Y})^2}} = \frac{\overline{\text{Cov}}(X,Y)}{\bar{\sigma}_X\bar{\sigma}_Y}, \tag{3.8}$$

où $\bar{\sigma}_X^2$ et $\bar{\sigma}_Y^2$ sont respectivement la variance de X et la variance de Y

Propriétés 2. *La corrélation de Pearson a les propriétés suivantes :*

1. $\overline{\text{corr}}(X,Y) \in [-1,1]$ *;*
2. $\overline{\text{corr}}(X,X) = 1$ *(si X n'est pas constante) ;*
3. *symétrie :*

$$\overline{\text{corr}}(X,Y) = \overline{\text{corr}}(Y,X); \tag{3.9}$$

4. *soient $a > 0$ et $b \in \mathbb{R}$:*

$$\overline{\text{corr}}(aX+b,Y) = \overline{\text{corr}}(X,Y). \tag{3.10}$$

La propriété 1 nous indique que la corrélation de Pearson, contrairement à la covariance, est bornée (et même restreinte à $[-1,1]$). De plus, la propriété 2 nous montre que la corrélation $\overline{\text{corr}}(X,Y)$ atteint son maximum lorsque $Y = X$. Enfin, la propriété 4 nous permet d'écrire, pour $Y = \lambda X$ (avec $\lambda > 0$) :

$$\overline{\text{corr}}(X,Y) = \overline{\text{corr}}(X,\lambda X) = \overline{\text{corr}}(X,X) = 1. \tag{3.11}$$

Par conséquent, la corrélation entre X et Y est de 1 lorsqu'il existe une relation linéaire parfaite entre X et Y, peu importe le coefficient de proportionnalité λ et les variances $\bar{\sigma}_X^2$ et $\bar{\sigma}_Y^2$. La corrélation de Pearson permet donc de résoudre l'inconvénient de la covariance illustré dans la remarque 15.

Définition 32. *Soient X et Y deux variables :*

1. *on dit que X et Y sont* non corrélées *si $\overline{\text{corr}}(X,Y) = 0$;*
2. *on dit que X et Y sont* corrélées positivement *si $\overline{\text{corr}}(X,Y) > 0$;*
3. *on dit que X et Y sont* corrélées négativement *si $\overline{\text{corr}}(X,Y) < 0$.*

Remarque 16 (Corrélation $\neq$ dépendance)**.** *Deux variables X et Y corrélées sont nécessairement dépendantes : si elles sont corrélées positivement, alors les grandes valeurs de X sont associées aux grandes valeurs de Y, donc connaître X_i nous donne de l'information sur la valeur de Y_i. De même si elles sont corrélées négativement.*

En revanche, deux variables X et Y dépendantes peuvent être non corrélées. Prenons par exemple le tableau de données suivant :

X	6	2	1	3	3	2	6	1
Y	6	-2	-1	-3	3	2	-6	1

On a $\overline{\text{corr}}(X,Y) = 0$. Pourtant, il existe une relation de dépendance très simple entre X et Y : si $X_i = x$, alors $Y_i = \pm x$.

Une corrélation nulle entre deux variables ne signifie pas qu'elles sont indépendantes.

3.3.2 Analyse de corrélation (Pearson)

De la même manière qu'avec les nuages de points, il est possible d'utiliser la corrélation dans le cadre de l'analyse d'un tableau avec un nombre quelconque de variables. L'*analyse de corrélation* consiste à étudier le *tableau des corrélations*, définie de la manière suivante :

Définition 33 (Tableau des corrélations)**.** *On dispose d'un tableau de données à p variables (quantitatives), notées $(X^{(1)}, \cdots, X^{(p)})$. Son tableau des corrélations est une matrice C de taille $p \times p$, dont chaque coefficient est défini par :*

$$C_{ij} = \overline{\text{corr}}(X^{(i)}, X^{(j)}). \tag{3.12}$$

Comme $\overline{\text{corr}}(X^{(i)}, X^{(i)}) = 1$, alors la diagonale principale de la matrice des corrélations ne contient que des 1. Par symétrie de $\overline{\text{corr}}$, la matrice des corrélations est toujours symétrique.

Exemple 22. *On considère le tableau de données suivant :*

$X^{(1)}$	0.28	3.10	2.37	5.62	1.99	1.36	0.52
$X^{(2)}$	4.51	6.05	4.28	8.94	5.40	4.06	3.87
$X^{(3)}$	1.30	3.36	1.42	6.17	1.25	1.54	0.77
$X^{(4)}$	3.32	3.79	3.68	3.89	3.76	3.80	3.48

Le tableau de corrélations associé est le suivant :

	$X^{(1)}$	$X^{(2)}$	$X^{(3)}$	$X^{(4)}$
$X^{(1)}$	1.00	0.92	0.93	0.78
$X^{(2)}$	0.92	1.00	0.96	0.59
$X^{(3)}$	0.93	0.96	1.00	0.61
$X^{(4)}$	0.78	0.59	0.61	1.00

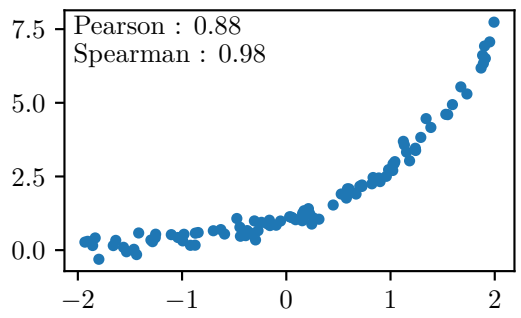
L'analyse de corrélation permet d'évaluer la partie linéaire de la dépendance entre chaque couple de variables. Elle permet d'identifier les variables "redondantes", c'est-à-dire celles qui sont très fortement corrélées (positivement ou négativement), et permet également de mesurer grossièrement l'influence de plusieurs variables sur une variable d'intérêt. Par exemple, on peut effectuer une analyse de corrélation dans le cas de l'analyse des données provenant d'une chaîne d'hôtels : on mesure l'effet de différents paramètres de l'hôtel (prix de la chambre, nombre d'étoiles, note donnée sur internet, densité de la population aux alentours, etc.) sur sa rentabilité, dans le but de proposer une stratégie pour l'entreprise (placement de nouveaux hôtels, etc.).

3.3.3 Limites de la corrélation de Pearson

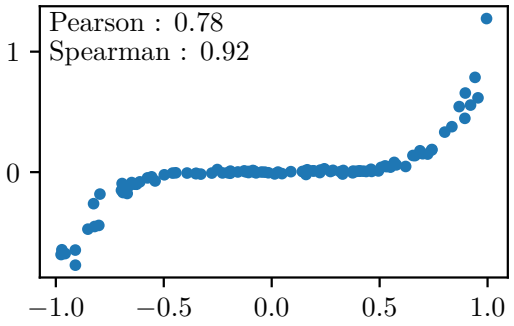
La corrélation de Pearson est facile à calculer et à interpréter, mais souffre de plusieurs inconvénients.

Dépendance non linéaire. La corrélation de Pearson est limitée à la mesure de la partie *linéaire* de la dépendance. Lorsque la structure de dépendance entre deux variables a une composante non linéaire, celle-ci est ignorée par la corrélation de Pearson.

La figure 3.3 illustre ce phénomène avec deux exemples. Dans les deux cas, la corrélation de Pearson capture partiellement la dépendance entre les deux variables, puisque les grandes valeurs de Y sont associées aux grandes valeurs de X. Cette relation n'est pas pour autant linéaire ou affine, comme les formules qui lient Y à X l'indiquent.



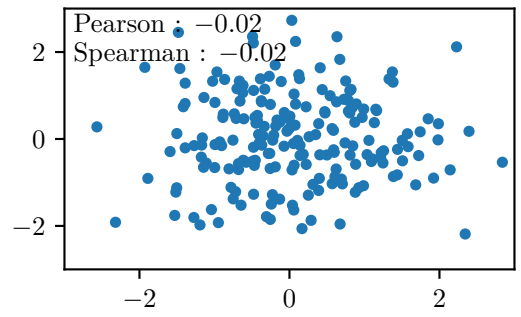
(a) Variables (X, Y) fortement dépendantes : $Y_i \approx \exp(X_i)$.



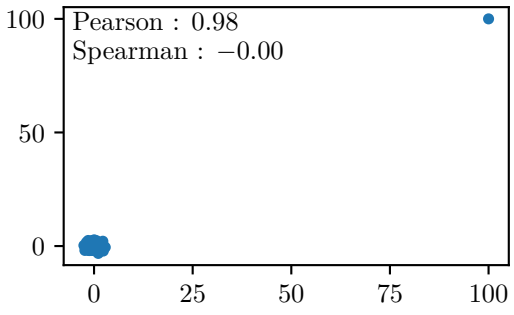
(b) Variables (X, Y) fortement dépendantes : $Y_i \approx X_i^3$.

FIGURE 3.3 – Cas où la relation de dépendance entre les variables a une composante non linéaire.

Valeurs aberrantes. La corrélation de Pearson est très sensible aux valeurs extrêmes ou aberrantes. La figure 3.4 illustre ce phénomène : la corrélation entre deux variables peut passer de 0 à 1 avec l'ajout d'une seule observation. En règle générale, on ne peut pas se fier à la valeur de la corrélation si l'on n'a pas vérifié au préalable qu'il n'y avait pas de valeurs extrêmes dans le tableau de données.



(a) Nuage de points pour un couple de variables indépendantes : corrélations de Pearson et de Spearman proches de 0.



(b) Même nuage de points, avec une valeur aberrante : corrélation de Pearson proche de 1, corrélation de Spearman restée proche de 0.

FIGURE 3.4 – Cas où la valeur de la corrélation de Pearson change radicalement avec l'ajout d'une seule observation, alors que la corrélation de Spearman change peu.

3.4 Corrélation de Spearman

La corrélation de Spearman est un outil fondé sur la corrélation de Pearson, qui quantifie non pas la dépendance linéaire entre deux variables, mais leur dépendance en termes de *monotonie*. Autrement dit, si les grandes valeurs de X sont systématiquement associées aux grandes valeurs de Y, alors la corrélation de Spearman sera très élevée, peu importe la nature exacte de la dépendance entre X et Y (linéaire, fonction $x \mapsto \exp(x)$, fonction $x \mapsto x^3$, etc.).

Définition 34 (Rang d'une valeur dans une série de données)**.** *Soit $(X_1, \cdots, X_n)$ une série de données. On note $\mathrm{rg}(X_i)$ le rang de la valeur X_i dans la série de données. On définit ce rang de la façon suivante.*

Si X_i est l'unique plus petite valeur dans la série de X, alors $\mathrm{rg}(X_i) = 1$, si elle est l'unique deuxième plus petite, alors $\mathrm{rg}(X_i) = 2$, etc.

Si la série de données contient des valeurs identiques, on attribue le même rang à celles-ci, valant la moyenne des rangs qu'elles auraient eus si on les avait triées. Par exemple, prenons le cas où il y a 3 valeurs identiques dans la série de données, qui occupent toutes le rang 5 à égalité. On trie arbitrairement ces valeurs : l'une devrait prendre le rang 5, une autre le rang 6 et la dernière le rang 7. Puis on leur attribue toutes la moyenne de ces rangs, c'est-à-dire $\frac{5+6+7}{3} = 6$.

Définition 35 (Corrélation de Spearman)**.** *Soient X et Y deux variables d'un tableau de données. On note $(X_1, \cdots, X_n)$ et $(Y_1, \cdots, Y_n)$ les séries de données associées.*

On construit un autre tableau de données, dont les variables rg_X et rg_Y sont définies par :

$$(\mathrm{rg}_X)_i = \mathrm{rg}(X_i) \qquad\qquad (\mathrm{rg}_Y)_i = \mathrm{rg}(Y_i). \tag{3.13}$$

La corrélation de Spearman est définie par :

$$r_s(X, Y) = \overline{\mathrm{corr}}(\mathrm{rg}_X, \mathrm{rg}_Y). \tag{3.14}$$

Exemple 23. *On considère le tableau de données suivant :*

X	6	5	0	2	2	5	4	5
Y	10	7	-4	-3	0	8	1	8

Le tableau des rangs est donc le suivant :

rg_X	8	6	1	2.5	2.5	6	4	6
rg_Y	8	5	1	2	3	6.5	4	6.5

La corrélation de Spearman de (X, Y) est donc : 0.98.

La corrélation de Spearman ne tient compte que de l'*ordre* des valeurs prises par les X_i et les Y_i, et non des valeurs elles-mêmes. Cette propriété a deux conséquences, qui répondent aux inconvénients de la corrélation de Pearson.

Premièrement, la corrélation de Spearman permet de capturer les relations de dépendance plus complexes que les dépendances linéaires. Par exemple, si Y est une fonction

monotone (croissante ou décroissante) de X, la corrélation de Spearman de (X, Y) sera élevée, peu importe la forme exacte de la fonction. En comparaison, la corrélation de Pearson ne sera élevée que si la fonction est à peu près affine. La figure 3.3 illustre cette propriété.

Deuxièmement, la corrélation de Spearman est peu sensible à l'ajout d'une observation (X_i, Y_i) pour laquelle X_i ou Y_i est démesurément plus grande que pour les autres observations. En effet, dans un ensemble à n observations, une valeur extrême aura une influence limitée sur le calcul de la corrélation de Spearman, alors que la corrélation de Pearson peut changer radicalement (passer de 0 à 1, ou même de -1 à 1). La figure 3.4 illustre cette propriété.

La corrélation de Spearman est donc adaptée aux situations où l'on veut démontrer une relation de dépendance plus générale qu'une simple dépendance linéaire, et aux cas où des valeurs extrêmes peuvent apparaître. La corrélation de Pearson reste pertinente pour mesurer la composante linéaire de la dépendance.

3.5 Corrélations fallacieuses et causalité

La corrélation (de Pearson et de Spearman) est un outil permettant de mettre en évidence une relation de dépendance entre deux variables d'un tableau de données. Néanmoins, la corrélation *seule* ne permet pas de conclure sur l'existence d'un lien de causalité entre deux variables, et ne permet pas de faire des extrapolations fiables.

Corrélations fallacieuses. Dans la grande masse de données disponibles, il est facile de trouver des couples de variables n'ayant aucun lien entre elles, mais dont la corrélation est proche de 1. On qualifie cette corrélation de *fallacieuse* dans la mesure où l'on *sélectionne* un couple de variables de façon à obtenir une très forte corrélation, au lieu de calculer les corrélations sur un ensemble de variables *déjà défini*.

L'existence de corrélations fallacieuses doit inciter à de bonnes pratiques :

1. en-dehors de tout contexte, on ne peut rien conclure d'un calcul de corrélation entre deux variables ; celui-ci doit toujours s'accompagner d'une *expertise* démontrant qu'il est pertinent d'étudier la relation entre celles-ci ;
2. la boucle consistant à traiter un tableau de données, puis ajouter des variables et recommencer doit être effectuée avec discernement : on court le risque d'ajouter une variable qui est, *par hasard*, corrélée à une variable déjà présente, ce qui va fausser l'interprétation de l'analyse de corrélation.

Corrélation et causalité. Si deux variables X et Y ont une corrélation proche de 1 (ou de -1), on ne peut pas nécessairement conclure à une relation de causalité entre l'une et l'autre. Il est possible que ce soit une troisième variable (encore non prise en compte) qui cause à la fois X et Y, induisant effectivement une corrélation entre X et Y. Mais il est également possible que la corrélation entre X et Y soit due au hasard.

Au mieux, une analyse de corrélation peut permettre de découvrir les couples de variables dépendantes, et ainsi servir de base pour une analyse plus poussée sur les liens de causalité. Mais, sans information supplémentaire ni connaissance d'expert, il est difficile d'interpréter une corrélation entre deux variables.

Chapitre 4

Régression linéaire

4.1 Enjeux

Dans de nombreuses situations, on dispose d'un tableau de données contenant au moins deux variables, et on souhaite exprimer l'une d'entre elles en fonction des autres.

Prise de décision. On souhaite prendre une décision qui va avoir l'effet désiré sur une variable donnée. On modélise la prise de décision comme le choix de la valeur de certaines variables, et on détermine quel choix aura l'effet désiré.

Prédiction. On est confronté à une nouvelle observation partielle, pour laquelle des variables sont connues, et d'autres sont encore inconnues. On cherche à prédire les variables inconnues en se servant d'un tableau de données contenant les observations passées (complètes) et de l'observation partielle. C'est une tâche de *prédiction* : on utilise la connaissance que l'on a déjà acquise pour faire une prédiction sur des données dont on ne dispose pas encore.

Modélisation. On dispose d'un tableau de données, et on souhaite traiter une variable (ou plusieurs) comme une fonction des autres variables. Cela survient lors de certaines analyses mathématiques, où l'on préfère traiter une fonction plutôt qu'un tableau de données. C'est aussi utile lorsqu'on souhaite quantifier l'influence de plusieurs variables sur une autre.

4.2 Régression linéaire simple

4.2.1 Définition du problème

Nous disposons d'un tableau de données constitué de deux variables quantitatives, notées X et Y, et de n observations, notées $((X_1, Y_1), \cdots (X_n, Y_n))$.

Dans ce contexte, nous voulons construire un modèle f, prédisant Y_i à partir de X_i :

$$Y_i \approx f(X_i), \tag{4.1}$$

où f est une fonction.

Pour ce faire, nous avons besoin de choisir deux éléments : le modèle f et l'objectif à atteindre.

Le modèle linéaire simple. Le *modèle linéaire simple* est le modèle proposant une relation affine entre une variable expliquée Y et une variable explicative X. La forme générale du modèle linéaire simple est la suivante :

$$f(X) = \beta_0 + \beta_1 X, \tag{4.2}$$

où β_0 et β_1 sont les *paramètres* du modèle à régler de façon à avoir la meilleure approximation $Y \approx f(X) = \beta_0 + \beta_1 X$.

La somme des carrés des résidus. Mais comment mesure-t-on la qualité de l'approximation $Y \approx f(X)$? Afin de quantifier celle-ci, on commence par définir les *résidus* du modèle par rapport au jeu de données $((X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n))$. Pour chaque observation $i \in \{1, \cdots, n\}$, le résidu e_i est donné par :

$$e_i = Y_i - (\beta_0 + \beta_1 X_i), \tag{4.3}$$

ce qui correspond à l'écart entre chaque valeur Y_i à approximer et l'estimation $f(X_i) = \beta_0 + \beta_1 X_i$ de cette valeur donnée par le modèle f. Ainsi, plus les résidus $(e_1, \cdots, e_n)$ sont proches de zéro, plus on considérera que l'approximation $Y \approx f(X) = \beta_0 + \beta_1 X$ est bonne.

Une fois les résidus définis, on peut construire un critère permettant de quantifier la qualité du modèle. Le critère le plus couramment utilisé en régression linéaire est la *Somme des Carrés des Résidus* (SCR) :

$$\mathrm{SCR}(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - f(X_i))^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 = \sum_{i=1}^{n} e_i^2. \tag{4.4}$$

En minimisant la SCR, on minimise la somme des carrés des écarts entre les données Y_i et les résultat du modèle $f(X_i)$. Si la SCR vaut zéro, cela signifie que tous les résidus e_i sont nuls, donc que le modèle f est parfait : pour tout i, $e_i = Y_i - f(X_i) = 0$, donc $Y_i = f(X_i)$.

Remarque 17. *Habituellement, on utilise indistinctement les termes "régression linéaire" et "méthode des moindres carrés". On peut toutefois introduire la nuance suivante :*

- *une régression linéaire consiste à utiliser un modèle linéaire sur un ensemble de données ; il est tout à fait possible d'effectuer une régression linéaire en utilisant un autre critère que la SCR pour quantifier sa qualité, par exemple :*

$$L(\beta_0, \beta_1) = \sum_{i=1}^{n} |e_i|; \tag{4.5}$$

- *la méthode des moindres carrés consiste à utiliser un modèle linéaire conjointement avec la SCR.*

Remarque 18. *On peut utiliser des modèles autres que linéaires. Par exemple, on peut vouloir modéliser la relation entre X et Y avec un polynôme de degré 2 (ou plus) :*

$$f(X) = \beta_0 + \beta_1 X + \beta_2 X^2 \ (+ \cdots). \tag{4.6}$$

Il s'agit toujours d'une régression, mais non linéaire.

4.2.2 Solution du problème

La solution du problème de minimisation de la SCR définie en (4.4) est donnée par la proposition suivante.

Proposition 1. *On suppose qu'on n'a pas $X_1 = \cdots = X_n$ (les $(X_i)_i$ ne sont pas tous simultanément égaux). Le point $(\hat{\beta}_0, \hat{\beta}_1)$ minimisant la SCR a pour coordonnées :*

$$\begin{cases} \hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1 \bar{X}, \\ \hat{\beta}_1 &= \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}, \end{cases} \tag{4.7}$$

où $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ et $\bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$.

Démonstration. Démonstration en fin de chapitre. □

Remarque 19 (Relation avec la corrélation de Pearson)**.** *Le coefficient directeur $\hat{\beta}_1$ de la régression linéaire est directement proportionnel à la corrélation de Pearson :*

$$\hat{\beta}_1 = \frac{\bar{\sigma}_Y}{\bar{\sigma}_X} \overline{\mathrm{corr}}(X, Y) = \frac{\overline{\mathrm{Cov}}(X, Y)}{\bar{\sigma}_X^2}. \tag{4.8}$$

Cela illustre le fait que la corrélation de Pearson entre X et Y quantifie la composante linéaire de la dépendance entre X et Y.

Exemple 24. *On a le tableau de données suivant :*

X	8	4	10	3	11	6
Y	-6	1	-12	0	-12	-2

On veut effectuer la régression de Y en fonction de X. On a :

$$\bar{X} = 7 \qquad \bar{Y} = -5.17 \qquad \mathrm{Cov}(X, Y) = -15.17 \qquad \mathrm{Var}(X) = 8,67. \tag{4.9}$$

Donc :

$$\hat{\beta}_1 = -1.75 \qquad \hat{\beta}_0 = 7.08. \tag{4.10}$$

Remarque 20 (Cas d'instabilité de la régression linéaire)**.** *Lorsque le nombre d'observations est faible et que la corrélation entre les deux variables est petite, les résultats de la régression linéaire peuvent changer considérablement selon l'échantillon choisi. Cela est illustré sur la figure 4.1.*

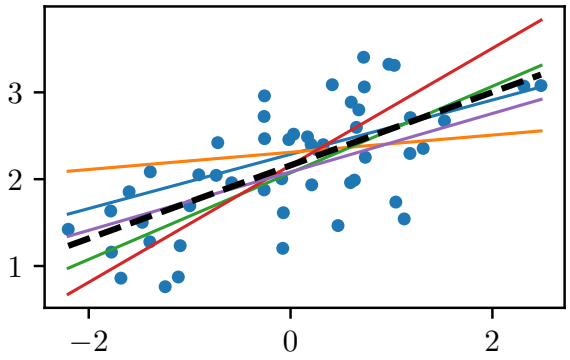


FIGURE 4.1 – On a effectué ici 5 régressions linéaires sur 5 échantillons du tableau de données différents (droites colorées), et la régression linéaire sur le tableau entier (droite noire en pointillés). On observe une variabilité importante entre les droites de régression.

Il faut donc proposer un moyen systématique d'évaluer la fiabilité d'une régression linéaire. Même s'il est toujours possible d'effectuer la régression linéaire d'une variable en fonction d'une autre, la fiabilité de cette régression n'est pas toujours assurée : il est possible que, sur un autre échantillon, les résultats de la régression soient très différents.

4.2.3 Évaluation de l'incertitude sur les paramètres

Jusqu'à maintenant, on a considéré que l'ensemble de données $((X_1, Y_1), \cdots (X_n, Y_n))$ était déterministe et fixé une bonne fois pour toutes. Mais, si nous décidons d'échantillonner un nouvel ensemble de données pour réaliser une autre régression linéaire, obtiendrons-nous un résultat similaire ? Autrement dit : est-on certain d'avoir acquis des connaissances sur la relation entre les variables X et Y, ou bien a-t-on construit une relation entre X et Y propre à l'ensemble de données qu'on a échantillonné au départ, et qui s'avérerait fausse sur un autre échantillon ?

On souhaite donc évaluer l'influence du bruit provenant de l'échantillonnage des données sur la valeur des paramètres optimaux $\hat{\beta}_0$ et $\hat{\beta}_1$. Pour y parvenir, nous devons nécessairement modéliser ce bruit, et donc introduire une composante aléatoire dans les données.

En pratique, on considère que les variables explicatives $(X_i)_i$ restent fixées et déterministes, et on modélise chaque variable expliquée Y_i comme une variable aléatoire d'espérance $\mathbb{E}[Y_i] = \beta_0 + \beta_1 X_i$, à laquelle on ajoute du bruit (additif) ε_i :

$$\forall i \in \{1, \cdots, n\}, \quad \begin{cases} X_i \text{ est fixé,} \\ Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \end{cases} \tag{4.11}$$

où les variables aléatoires $(\varepsilon_1, \cdots, \varepsilon_n)$ sont centrées, de même variance σ^2 et de covariance nulle ($\forall i \neq j, \mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$). Les paramètres β_0 et β_1 sont fixés, et sont *a priori* inconnus. Comme on part du principe que le modèle défini en (4.11) a généré les données, on souhaite en estimer les paramètres (β_0, β_1). Pour cela, on va se servir des estimateurs $(\hat{\beta}_0, \hat{\beta}_1)$ définis dans la proposition 1, et on cherchera à évaluer leur qualité.

Rappels sur les estimateurs. Avant de discuter de la qualité des estimateurs $\hat{\beta}_0$ et $\hat{\beta}_1$, nous effectuons un bref rappel sur les estimateurs en statistiques.

Définition 36 (Estimateur)**.** *Soient des variables aléatoires observées* $(X_1, \cdots, X_n)$*, i.i.d. et de loi* P*. Soit* $\theta \in \mathbb{R}$ *un réel quelconque, généralement un paramètre lié à la loi* P*, comme son espérance ou sa variance.*

(i) un estimateur *de* θ*, noté* $\hat{\theta}$*, est une fonction* $\mathbb{R}^n \to \mathbb{R}$ *mesurable par rapport à la tribu engendrée par* $(X_1, \cdots, X_n)$*. On souhaite que* $\hat{\theta}(X_1, \cdots, X_n) \approx \theta$ *avec une grande probabilité. Pour simplifier les notations, on utilisera indistinctement* $\hat{\theta}$ *et* $\hat{\theta}(X_1, \cdots, X_n)$ *;*

(ii) on dit que l'estimateur $\hat{\theta}$ *de* θ *est* sans biais *si :*

$$\mathbb{E}[\hat{\theta}] = \theta; \tag{4.12}$$

(iii) on appelle erreur quadratique moyenne *de l'estimateur* $\hat{\theta}$ *de* θ *la quantité suivante, notée MSE (*Mean Squared Error*) :*

$$\mathrm{MSE}(\hat{\theta}) = \mathbb{E}[(\theta - \hat{\theta})^2]. \tag{4.13}$$

Lorsque l'estimateur $\hat{\theta}$ *est sans biais, on a :* $\theta = \mathbb{E}\hat{\theta}$*, donc :*

$$\mathrm{MSE}(\hat{\theta}) = \mathbb{E}[(\mathbb{E}\hat{\theta} - \hat{\theta})^2] = \mathrm{Var}(\hat{\theta}). \tag{4.14}$$

Exemple 25 (Estimateur de l'espérance)**.** *Soient des variables aléatoires observées* $(X_1, \cdots, X_n)$*, i.i.d. et de loi* P*. On suppose que* $\mathbb{E}|X_1| < \infty$ *et* $\mathrm{Var}(X_1) < \infty$*, et on note* $\mu = \mathbb{E}X_1$ *et* $\sigma^2 = \mathrm{Var}(X_1)$*. Soit* $\hat{\mu}$ *un estimateur de* μ *défini par :*

$$\hat{\mu} = \frac{1}{n}\sum_{i=1} X_i. \tag{4.15}$$

Cet estimateur a les propriétés suivantes :

— $\hat{\mu}$ *est sans biais :*

$$\mathbb{E}\hat{\mu} = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}X_i = \mu; \tag{4.16}$$

— *l'erreur quadratique moyenne de* $\hat{\mu}$ *est :*

$$\mathrm{MSE}(\hat{\mu}) = \mathbb{E}[(\mu - \hat{\mu})^2] = \mathbb{E}[(\mathbb{E}\hat{\mu} - \hat{\mu})^2] = \mathrm{Var}(\hat{\mu}) \tag{4.17}$$

$$= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{\sigma^2}{n}. \tag{4.18}$$

On remarque que, dans ce cas, la MSE tend vers zéro lorsque la taille n *de l'échantillon tend vers l'infini. Cela illustre le fait que l'estimateur* $\hat{\mu}$ *est plus fiable lorsqu'on dispose de plus d'observations.*

Estimateurs des paramètres (β_0, β_1)**.** On s'intéresse à la qualité des estimateurs $(\hat{\beta}_0, \hat{\beta}_1)$ de (β_0, β_1). On va donc vérifier qu'ils sont sans biais et on va calculer leur MSE.

Proposition 2. *Les estimateurs $\hat{\beta}_0$ et $\hat{\beta}_1$ ont pour espérance et MSE :*

$$\mathbb{E}\hat{\beta}_0 = \beta_0, \qquad \mathrm{MSE}(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right), \tag{4.19}$$

$$\mathbb{E}\hat{\beta}_1 = \beta_1, \qquad \mathrm{MSE}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2}. \tag{4.20}$$

Ils sont donc sans biais.

Démonstration. Démonstration en fin de chapitre. □

Estimation de σ^2. On a vu que, pour calculer $\mathrm{MSE}(\hat{\beta}_0)$ et $\mathrm{MSE}(\hat{\beta}_1)$, on a besoin de connaître σ^2. Or, en règle générale, σ^2 est un paramètre inconnu. Il faut donc l'estimer.

Notation 2. *On note $\hat{Y}_i := \hat{\beta}_0 + \hat{\beta}_1 X_i$ l'estimation de Y_i fournie par le modèle.*

Proposition 3. *On a :*

$$\mathbb{E}[\mathrm{SCR}(\hat{\beta}_0, \hat{\beta}_1)] = (n-2)\sigma^2 \tag{4.21}$$

Par conséquent, un estimateur non biaisé de σ^2 est :

$$\hat{\sigma}^2 := \frac{\mathrm{SCR}(\hat{\beta}_0, \hat{\beta}_1)}{n-2} = \frac{1}{n-2} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2, \tag{4.22}$$

c'est-à-dire, à un facteur $\frac{n-2}{n}$ près, la moyenne des carrés des résidus.

Démonstration. Démonstration en fin de chapitre. □

La valeur de $\hat{\sigma}$, appelée *erreur résiduelle standard* (ERS), indique le manque d'adéquation du modèle linéaire aux données. Une ERS $\hat{\sigma}$ élevée peut signifier deux choses qui ne s'excluent pas mutuellement : soit le modèle linéaire est inadapté à l'ensemble de données, soit le bruit est très important.

Le coefficient de détermination R^2.

Définition 37. *Le* coefficient de détermination R^2 *est défini par :*

$$R^2 := \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = \frac{\mathrm{SCE}(\hat{\beta}_0, \hat{\beta}_1)}{\mathrm{SCT}}, \tag{4.23}$$

où $\mathrm{SCT} := \sum_{i=1}^n (Y_i - \bar{Y})^2$ *est la appelée la* somme des carrés des écarts totale *et* $\mathrm{SCE}(\hat{\beta}_0, \hat{\beta}_1) := \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$ *est appelée la* somme des carrés des écarts expliqués.

Remarque 21. *Le coefficient R^2 s'exprime aussi de la façon suivante :*

$$R^2 = 1 - \frac{\mathrm{SCR}(\hat{\beta}_0, \hat{\beta}_1)}{\mathrm{SCT}}. \tag{4.24}$$

Démonstration. Démonstration en fin de chapitre. □

La SCT sert de point de référence pour évaluer la qualité d'une régression. En effet, la SCT est égale à la somme des carrés des écarts qu'on peut attendre si l'on décide d'estimer les valeurs Y_i avec la meilleure constante possible, c'est-à-dire leur moyenne $\bar{Y}$. On s'attend donc à ce que toute technique de régression fournisse une SCR inférieure ou égale à la SCT. C'est notamment le cas de la régression linéaire Comme par ailleurs la SCR est positive ou nulle, $\frac{\mathrm{SCR}(\hat{\beta}_0,\hat{\beta}_1)}{\mathrm{SCT}} \in [0,1]$:

- si $\frac{\mathrm{SCR}(\hat{\beta}_0,\hat{\beta}_1)}{\mathrm{SCT}} \approx 0$, alors la technique de régression est significativement plus performante que l'estimation des Y_i par leur moyenne Y : la variable X est très informative ;
- si $\frac{\mathrm{SCR}(\hat{\beta}_0,\hat{\beta}_1)}{\mathrm{SCT}} \approx 1$, alors la technique de régression est à peine meilleure que l'estimation des Y_i par leur moyenne : la variable X est peu informative.

Remarque 22. *Le coefficient R^2 représente la proportion de variabilité de Y qui peut être expliquée en utilisant X :*

- $0 \leq R^2 \leq 1$ *;*
- $R^2 \approx 0$ *signifie que la variabilité de Y n'est pas explicable par X, ce qui peut signifier que : 1) la variance σ^2 du bruit ϵ est élevée, 2) le modèle linéaire est inadapté, 3) la variable X est indépendante de Y (X est inutile pour prédire Y) ;*
- $R^2 \approx 1$ *signifie que le modèle linéaire est quasiment exact.*

4.3 Régression linéaire multiple

On souhaite maintenant effectuer une régression linéaire avec n observations, p variables explicatives $(X^{(1)}, \cdots, X^{(p)})$, et une variable expliquée Y. Dans le cas de la régression linéaire simple, X_i est un scalaire et le modèle est une fonction affine à une seule variable :

$$f(X_i) = \beta_1 X_i + \beta_0.$$

Pour généraliser cette démarche au cas où X_i est un vecteur (ligne) à p coordonnées notées $(X_i^{(1)}, \cdots, X_i^{(p)})$, on choisit comme modèle une fonction affine à plusieurs variables (ou, ce qui est équivalent, une fonction dont la variable est un vecteur) :

$$f(X_i) = \beta_1 X_i^{(1)} + \cdots + \beta_p X_i^{(p)} + \beta_{p+1} = X_i \boldsymbol{\beta},$$

où $X_i\boldsymbol{\beta}$ est le produit scalaire entre le vecteur (ligne) $X_i \in \mathbb{R}^{1\times(p+1)}$ et le vecteur (colonne) $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$, et, par convention, $X_i^{(p+1)} = 1$.

La perte à optimiser est donc :

$$\mathrm{SCR}(\boldsymbol{\beta}) = \sum_{i=1}^{n} (Y_i - X_i\boldsymbol{\beta})^2.$$

On peut exprimer la SCR avec une formule matricielle : soit $\mathbf{X} \in \mathbb{R}^{n\times(p+1)}$ le tableau de données qui contient les valeurs des variables explicatives. La valeur x_{ij} située en ligne i, colonne j vaut :

— si $j < p+1$, $x_{ij} = X_i^{(j)}$, c'est-à-dire la variable j de l'observation i ;
— si $j = p+1$, $x_{ij} = 1$.

On procède de manière similaire avec la variable expliquée Y. On note $\mathbf{Y} \in \mathbb{R}^n$ le vecteur dont la composante d'index i vaut Y_i (la variable Y associée à l'observation i). La SCR est donc égale à :

$$\mathrm{SCR}(\boldsymbol{\beta}) = \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2, \tag{4.25}$$

où $\|\mathbf{x}\|^2 = x_1^2 + x_2^2 + \cdots x_n^2$ est la norme au carré du vecteur $\mathbf{x} \in \mathbb{R}^n$.

Proposition 4. *Si $\mathbf{X}^T\mathbf{X}$ est inversible, alors la solution $\hat{\boldsymbol{\beta}}$ du problème de minimisation existe et est unique :*

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}. \tag{4.26}$$

Démonstration. Démonstration en fin de chapitre. □

Remarque 23. *La formule pour $\hat{\boldsymbol{\beta}}$ contient l'inversion de la matrice $\mathbf{X}^T\mathbf{X}$:*

1. *pour que $\mathbf{X}^T\mathbf{X}$ soit inversible, il est nécessaire que $n \geq p+1$ (i.e., le nombre de lignes de $\mathbf{X}$ est supérieur à son nombre de colonnes).*
   En effet, $\mathrm{rg}(\mathbf{X}^T\mathbf{X}) \leq \mathrm{rg}(\mathbf{X}) \leq \min(n, p+1)$. Pour que $\mathbf{X}^T\mathbf{X}$ soit inversible, il faut que $\mathrm{rg}(\mathbf{X}^T\mathbf{X}) = p+1$, donc que $p+1 \leq n$.
2. *Dans le cas où $\mathbf{X}^T\mathbf{X}$ n'est pas inversible, il est parfois possible d'utiliser le pseudo-inverse de Moore-Penrose de $\mathbf{X}$, noté $\mathbf{X}^+$. Il a les propriétés suivantes :*

   $$\mathbf{X}^+\mathbf{X}\mathbf{X}^+ = \mathbf{X}^+ \qquad\qquad \mathbf{X}\mathbf{X}^+\mathbf{X} = \mathbf{X}. \tag{4.27}$$

   Si $\mathbf{X}^T\mathbf{X}$ est inversible, alors $\mathbf{X}^+ = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$.
3. *Dans le cas où deux variables sont fortement corrélées, $\mathbf{X}^T\mathbf{X}$ tend à ne pas être inversible, le calcul numérique de $(\mathbf{X}^T\mathbf{X})^{-1}$ devient instable, et le calcul de $\hat{\boldsymbol{\beta}}$ aussi.*

Définition 38. *Le coefficient R^2 s'exprime de la façon suivante :*

$$R^2 = 1 - \frac{\mathrm{SCR}(\hat{\boldsymbol{\beta}})}{\mathrm{SCT}}. \tag{4.28}$$

4.4 Pré-traitement des variables

Au-delà du modèle linéaire. La régression linéaire multiple permet d'expliquer une variable avec une combinaison linéaire de variables explicatives. Cependant, il arrive fréquemment que la relation entre la variable expliquée Y et les variables explicatives $(X^{(j)})_j$ soit non linéaire. Pour modéliser ces relations plus complexes, il est possible d'étendre le modèle linéaire de façon à proposer des relations polynomiales entre Y et les $(X^j)_{(j)}$. Par exemple, dans le cas d'une seule variable explicative X, on modélise Y par :

$$Y \approx f(X) := \alpha_0 + \alpha_1 X + \alpha_2 X^2,$$

où f est un polynôme de degré 2 et les $(\alpha_i)_i$ sont des paramètres à régler. On constate que ce problème de *régression polynomiale* de Y en fonction de X a la même forme que le problème de *régression linéaire multiple* de Y en fonction de (X, X^2). Par conséquent, les paramètres $(\alpha_0, \alpha_1, \alpha_2)$ minimisant la SCR peuvent être calculés analytiquement en utilisant la proposition 4.

Passage au logarithme. Il arrive souvent qu'une variable prenne des valeurs positives s'étalant sur plusieurs ordres de grandeur. C'est notamment le cas de nombreuses variables issues de jeux de données économiques. On peut alors pré-traiter ces variables en leur appliquant la fonction log avant de procéder à la régression linéaire multiple. De cette façon, on s'assure que chaque variable prenne des valeurs raisonnablement étalées sur une échelle linéaire, et on peut s'attendre à un meilleur ajustement du modèle linéaire sur ces données pré-traitées. Par exemple, si on utilise $\log(Y)$ au lieu de Y et $\log(X)$ au lieu de X, on a le modèle :

$$\log(Y) \approx f(X) := \beta_0 + \beta_1 \log(X) \quad \text{soit} \quad Y \approx \exp(\beta_0) X^{\beta_1},$$

ce qui revient à découvrir une relation de puissance entre Y et X.

4.5 Régression ridge

Il est possible de modifier légèrement de problème de régression initial pour éviter les instabilités dans le calcul de l'inverse de $\mathbf{X}^T\mathbf{X}$, et donc dans le calcul de $\hat{\boldsymbol{\beta}}$. Le problème de régression modifié consiste à minimiser la fonction SCR_λ, définie par :

$$\mathrm{SCR}_\lambda(\boldsymbol{\beta}) := \mathrm{SCR}(\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|^2, \tag{4.29}$$

où $\lambda \geq 0$ est un paramètre à régler. Ce problème de minimisation s'appelle la *régression ridge* et le terme $\lambda\|\boldsymbol{\beta}\|^2$ s'appelle *régularisation de Tikhonov.*

Proposition 5. *Si $\lambda > 0$, alors ce problème admet une unique solution $\hat{\boldsymbol{\beta}}_\lambda$:*

$$\hat{\boldsymbol{\beta}}_\lambda = (\mathbf{X}^T\mathbf{X} + \lambda \mathbf{I})^{-1}\mathbf{X}^T\mathbf{Y}, \tag{4.30}$$

où $\mathbf{I}$ est la matrice identité de taille $p+1$.

Remarque 24. *Cette régularisation a deux effets :*

1. *en ajoutant $\lambda\mathbf{I}$ à $\mathbf{X}^T\mathbf{X}$, elle rend l'inversion de $\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}$ plus stable ;*
2. *elle pousse la solution $\hat{\boldsymbol{\beta}}_\lambda$ du problème à avoir une norme $\|\hat{\boldsymbol{\beta}}_\lambda\|$ plus faible. En effet, le terme $\lambda\|\boldsymbol{\beta}\|^2$ de $\mathrm{SCR}_\lambda(\boldsymbol{\beta})$ pousse $\boldsymbol{\beta}$ vers 0.*
   Notamment, lorsque $\lambda \to 0$, $\hat{\boldsymbol{\beta}}_\lambda \to \hat{\boldsymbol{\beta}}$, et lorsque $\lambda \to \infty$, $\hat{\boldsymbol{\beta}}_\lambda \to 0$. L'ajout de $\lambda\|\boldsymbol{\beta}\|^2$ dans la SCR introduit donc un biais. Il faut trouver un compromis entre stabilité et exactitude.

Validation simple. Pour trouver une bonne valeur pour λ, on peut effectuer une *validation simple*. Cela consiste à partitionner le tableau de données $(\mathbf{X}, \mathbf{Y})$ en deux parties, notées $(\mathbf{X}_E, \mathbf{Y}_E)$ (E pour *entraînement*) et $(\mathbf{X}_V, \mathbf{Y}_V)$ (V pour *validation*). On s'intéresse à la SCR calculée sur $(\mathbf{X}_E, \mathbf{Y}_E)$ et la SCR calculée sur $(\mathbf{X}_V, \mathbf{Y}_V)$, que l'on note respectivement $\mathrm{SCR}_\lambda(\boldsymbol{\beta}, \mathbf{X}_E, \mathbf{Y}_E)$ et $\mathrm{SCR}_\lambda(\boldsymbol{\beta}, \mathbf{X}_V, \mathbf{Y}_V)$

La validation simple permet de trouver une valeur $\hat{\lambda}$ pour λ :

— pour $\lambda \geq 0$ fixé, on utilise $(\mathbf{X}_E, \mathbf{Y}_E)$ pour calculer $\hat{\boldsymbol{\beta}}_{E,\lambda}$:

$$\hat{\boldsymbol{\beta}}_{E,\lambda} = \arg\min_{\boldsymbol{\beta}} \left[\mathrm{SCR}_\lambda(\boldsymbol{\beta}, \mathbf{X}_E, \mathbf{Y}_E)\right]. \tag{4.31}$$

— calcul de $\hat{\lambda}$:

$$\hat{\lambda} = \arg\min_{\lambda \geq 0} \left[\mathrm{SCR}(\hat{\boldsymbol{\beta}}_{E,\lambda}, \mathbf{X}_V, \mathbf{Y}_V) \right]. \tag{4.32}$$

Informellement, on peut dire que :
— on utilise $(\mathbf{X}_E, \mathbf{Y}_E)$ pour calculer $\hat{\boldsymbol{\beta}}_{E,\lambda}$;
— on utilise $(\mathbf{X}_V, \mathbf{Y}_V)$ pour calculer $\hat{\lambda}$.

4.6 Preuves

Preuve de la proposition 1.

Démonstration. On cherche le couple de réels $(\hat{\beta}_0, \hat{\beta}_1)$ minimisant la SCR. Soit $\boldsymbol{\beta} = (\beta_0, \beta_1) \in \mathbb{R}^2$.

Idée de la preuve :

1. on réécrit $\mathrm{SCR}(\beta_0, \beta_1)$ de façon a avoir une expression simple en fonction de β_0 et β_1 (il faut faire sortir ces variables de la somme $\sum_{i=1}^n$) ;
2. on effectue une translation bien choisie des données (centrage), de façon à simplifier les calculs qui suivent ;
3. on démontre que la SCR est strictement convexe par rapport à $\boldsymbol{\beta}$. Cela permet d'affirmer que, si la SCR a un point critique, alors il est unique et est un minimum global. Pour démontrer la stricte convexité de la SCR, il suffit de prouver que les valeurs propres de sa hessienne $\mathbf{H} \in \mathbb{R}^{2\times 2}$ sont strictement positives ;
4. on détermine le point critique de la SCR, c'est-à-dire $\hat{\boldsymbol{\beta}}$ tel que : $\nabla_{\boldsymbol{\beta}} \mathrm{SCR}(\hat{\boldsymbol{\beta}}) = 0$. Comme la SCR est strictement convexe, ce point critique est aussi un minimum global de la SCR ;
5. on effectue l'opération inverse à la translation faite au point 2.

Réécriture de la SCR :

$$\begin{aligned}
\mathrm{SCR}(\beta_0, \beta_1) &= \sum_{i=1}^n (Y_i - (\beta_0 + \beta_1 X_i))^2 \\
&= \sum_{i=1}^n \left[Y_i^2 - 2Y_i(\beta_0 + \beta_1 X_i) + (\beta_0 + \beta_1 X_i)^2 \right] \\
&= \sum_{i=1}^n \left[Y_i^2 - 2Y_i\beta_0 - 2X_iY_i\beta_1 + \beta_0^2 + 2\beta_0\beta_1 X_i + \beta_1^2 X_i^2 \right] \\
&= n\beta_0^2 + a\beta_1^2 + 2b\beta_0\beta_1 - 2c\beta_0 - 2d\beta_1 + f,
\end{aligned}$$

où $a = \sum_{i=1}^n X_i^2$, $b = \sum_{i=1}^n X_i$, $c = \sum_{i=1}^n Y_i$, $d = \sum_{i=1}^n X_iY_i$, $f = \sum_{i=1}^n Y_i^2$.

Translation des données. On aimerait simplifier cette expression. On remarque que, si les données $(X_i)_i$ et $(Y_i)_i$ étaient centrées, c'est-à-dire $\bar{X} = 0$ et $\bar{Y} = 0$, alors on aurait $b = n\bar{X} = 0$ et $c = n\bar{Y} = 0$, ce qui simplifierait grandement l'expression de SCR. On propose donc d'effectuer les calculs sur les données centrées $(\tilde{X}_i)_i$ et $(\tilde{Y}_i)_i$, définies par :

$\tilde{X}_i = X_i - \lambda_X$ et $\tilde{Y}_i = Y_i - \lambda_Y$, avec $\lambda_X = \bar{X}$ et $\lambda_Y = \bar{Y}$. Dans ce cadre, la fonction à minimiser devient :

$$\begin{aligned}\widetilde{\mathrm{SCR}}(\beta_0, \beta_1) &= \sum_{i=1}^{n} (\tilde{Y}_i - (\beta_0 + \beta_1 \tilde{X}_i))^2 \\ &= \sum_{i=1}^{n} (Y_i - \lambda_Y - (\beta_0 + \beta_1 (X_i - \lambda_X)))^2 \\ &= \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_i + \lambda_Y - \beta_1 \lambda_X))^2\end{aligned}$$

Il suffit maintenant de prouver que, pour tout choix de translation λ_X des variables $(X_i)_i$ et λ_Y des variables $(Y_i)_i$, la valeur minimale que peut atteindre SCR est également atteignable par $\widetilde{\mathrm{SCR}}$. Si $(\hat{\beta}_0, \hat{\beta}_1)$ minimise SCR, alors en posant $\tilde{\beta}_0 = \hat{\beta}_0 - \lambda_Y + \hat{\beta}_1 \lambda_X$ et $\tilde{\beta}_1 = \hat{\beta}_1$, on a :

$$\begin{aligned}\widetilde{\mathrm{SCR}}(\tilde{\beta}_0, \tilde{\beta}_1) &= \sum_{i=1}^{n} (Y_i - (\hat{\beta}_0 - \lambda_Y + \hat{\beta}_1 \lambda_X + \hat{\beta}_1 X_i + \lambda_Y - \hat{\beta}_1 \lambda_X))^2 \\ &= \sum_{i=1}^{n} (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i))^2 \\ &= \mathrm{SCR}(\hat{\beta}_0, \hat{\beta}_1),\end{aligned}$$

qui est par hypothèse le minimum de SCR.

À partir de maintenant, on considérera les données $\tilde{X}_i = X_i - \bar{X}$ et $\tilde{Y}_i = Y_i - \bar{Y}$ et la minimisation de $\widetilde{\mathrm{SCR}}$. On peut notamment écrire :

$$\widetilde{\mathrm{SCR}}(\beta_0, \beta_1) = n\beta_0^2 + \tilde{a}\beta_1^2 - 2\tilde{d}\beta_1 + \tilde{f},$$

où $\tilde{a} = \sum_{i=1}^{n} \tilde{X}_i^2$, $\tilde{d} = \sum_{i=1}^{n} \tilde{X}_i \tilde{Y}_i$, $f = \sum_{i=1}^{n} \tilde{Y}_i^2$. Les constantes $\tilde{b}$ et $\tilde{c}$ sont nulles, grâce à la translation effectuée.

Stricte convexité de $\widetilde{\mathrm{SCR}}$. On rappelle que la hessienne d'une fonction $f : \mathbb{R}^2 \to \mathbb{R}$ de deux variables (X_1, X_2) de classe $\mathcal{C}^2$ s'écrit :

$$\begin{pmatrix} \frac{\partial^2 f}{\partial X_1^2} & \frac{\partial^2 f}{\partial X_1 \partial X_2} \\ \frac{\partial^2 f}{\partial X_1 \partial X_2} & \frac{\partial^2 f}{\partial X_2^2} \end{pmatrix}.$$

La hessienne $\mathbf{H}$ de $\widetilde{\mathrm{SCR}}$ s'écrit donc :

$$\mathbf{H} = \begin{pmatrix} \frac{\partial^2 \widetilde{\mathrm{SCR}}}{\partial \beta_0^2} & 0 \\ 0 & \frac{\partial^2 \widetilde{\mathrm{SCR}}}{\partial \beta_1^2} \end{pmatrix} = 2 \begin{pmatrix} n & 0 \\ 0 & \tilde{a} \end{pmatrix}.$$

Pour démontrer que $\widetilde{\mathrm{SCR}}$ est strictement convexe, il suffit de prouver que les valeurs propres de $\mathbf{H}$ sont strictement positives, c'est-à-dire $n > 0$ et $\tilde{a} > 0$. On rappelle que $\tilde{a} = \sum_{i=1}^{n} \tilde{X}_i^2 = \sum_{i=1}^{n} (X_i - \bar{X})^2$. Par conséquent, $\tilde{a} = 0$ si, et seulement si, $X_1 = \cdots = X_n = \bar{X}$, ce qui n'est jamais vérifié (par hypothèse de la proposition).

Calcul du point critique. On sait maintenant que la fonction $\widetilde{\text{SCR}}$ est strictement convexe. Pour déterminer son minimum (s'il existe), il suffit de chercher un point critique. S'il existe, alors il est aussi l'unique minimum global de $\widetilde{\text{SCR}}$. On commence par le calcul du gradient de $\widetilde{\text{SCR}}$:

$$\begin{pmatrix} \frac{\partial \widetilde{\text{SCR}}}{\partial \beta_0} \\ \frac{\partial \widetilde{\text{SCR}}}{\partial \beta_1} \end{pmatrix} = 2 \begin{pmatrix} n\beta_0 \\ \tilde{a}\beta_1 - \tilde{d} \end{pmatrix}.$$

Par définition, (β_0, β_1) est un point critique si, et seulement si, le gradient de $\widetilde{\text{SCR}}$ s'annule en ce point, c'est-à-dire :

$$\begin{cases} n\beta_0 = 0, \\ \tilde{a}\beta_1 = \tilde{d}. \end{cases}$$

Puisque n et $\tilde{a}$ sont non nuls, alors ce système d'équations admet une unique solution, qu'on note $(\tilde{\beta}_0, \tilde{\beta}_1)$:

$$\begin{cases} \tilde{\beta}_0 = 0, \\ \tilde{\beta}_1 = \frac{\tilde{d}}{\tilde{a}}. \end{cases}$$

Retour aux données initiales. Pour trouver le point $(\hat{\beta}_0, \hat{\beta}_1)$ minimisant SCR, on rappelle que $\tilde{\beta}_0 = \hat{\beta}_0 - \lambda_Y + \hat{\beta}_1 \lambda_X$ et $\tilde{\beta}_1 = \hat{\beta}_1$. Comme on a choisi les translations $\lambda_X = \bar{X}$ et $\lambda_Y = \bar{Y}$, alors, en utilisant les définitions de $\tilde{a}$ et $\tilde{d}$:

$$\begin{cases} \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}, \\ \hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}. \end{cases}$$

□

Preuve de la proposition 2.

Démonstration. Calcul de $\mathbb{E}\hat{\beta}_1$ *:*

$$\begin{aligned} \mathbb{E}\hat{\beta}_1 &= \frac{\sum_{i=1}^n (X_i - \bar{X})\mathbb{E}(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{\sum_{i=1}^n (X_i - \bar{X})(\mathbb{E}Y_i - \frac{1}{n}\sum_{j=1}^n \mathbb{E}Y_j)}{\sum_{i=1}^n (X_i - \bar{X})^2} \\ &= \frac{\sum_{i=1}^n (X_i - \bar{X})(\beta_0 + \beta_1 X_i - \frac{1}{n}\sum_{j=1}^n (\beta_0 + \beta_1 X_j))}{\sum_{i=1}^n (X_i - \bar{X})^2} \\ &= \frac{\sum_{i=1}^n (X_i - \bar{X})(\beta_0 + \beta_1 X_i - \beta_0 - \beta_1 \frac{1}{n}\sum_{j=1}^n X_j)}{\sum_{i=1}^n (X_i - \bar{X})^2} \\ &= \beta_1 \frac{\sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \beta_1. \end{aligned}$$

Calcul de $\mathbb{E}\hat{\beta}_0$ *:*

$$\mathbb{E}\hat{\beta}_0 = \mathbb{E}\bar{Y} - \mathbb{E}[\hat{\beta}_1]\bar{X} = \frac{1}{n}\sum_{i=1}^n (\beta_0 + \beta_1 X_i) - \beta_1 \bar{X} = \beta_0 + \beta_1 \bar{X} - \beta_1 \bar{X} = \beta_0.$$

Calcul de $\mathrm{MSE}(\hat{\beta}_1)$ *:*

$$\mathrm{MSE}(\hat{\beta}_1) = \mathrm{Var}(\hat{\beta}_1) = \frac{\mathrm{Var}\left(\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})\right)}{\left(\sum_{i=1}^n (X_i - \bar{X})^2\right)^2}.$$

En notant $\alpha_i = X_i - \bar{X}$, on a :

$$\begin{aligned}
\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) &= \sum_{i=1}^n \alpha_i (Y_i - \bar{Y}) = \sum_{i=1}^n \alpha_i Y_i - \sum_{i=1}^n \alpha_i \bar{Y} \\
&= \sum_{i=1}^n \alpha_i Y_i - \sum_{i=1}^n \left[\alpha_i \frac{1}{n} \sum_{j=1}^n Y_j\right] = \sum_{i=1}^n \alpha_i Y_i - \sum_{j=1}^n \left[Y_j \sum_{i=1}^n \frac{\alpha_i}{n}\right] \\
&= \sum_{j=1}^n Y_j \left[\alpha_j - \sum_{i=1}^n \frac{\alpha_i}{n}\right] = \sum_{j=1}^n Y_j \alpha_j,
\end{aligned}$$

puisque $\sum_{i=1}^n \alpha_i = \sum_{i=1}^n (X_i - \bar{X}) = 0$. Donc, par indépendance des $(Y_i)_i$:

$$\mathrm{Var}\left(\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})\right) = \sum_{j=1}^n \mathrm{Var}(Y_j)\alpha_j^2 = \sigma^2 \sum_{j=1}^n \alpha_j^2 = \sigma^2 \sum_{i=1}^n (X_i - \bar{X})^2.$$

Finalement :

$$\mathrm{MSE}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2}.$$

Calcul de $\mathrm{MSE}(\hat{\beta}_0)$ *:*

$$\begin{aligned}
\mathrm{MSE}(\hat{\beta}_0) = \mathrm{Var}(\hat{\beta}_0) &= \mathrm{Var}\left(\bar{Y} - \hat{\beta}_1 \bar{X}\right) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^n Y_i - \bar{X}\frac{\sum_{i=1}^n Y_i \alpha_i}{\sum_{i=1}^n \alpha_i^2}\right) \\
&= \mathrm{Var}\left(\sum_{i=1}^n Y_i \left[\frac{1}{n} - \bar{X}\frac{\alpha_i}{\sum_{j=1}^n \alpha_j^2}\right]\right) = \sigma^2 \sum_{i=1}^n \left(\frac{1}{n} - \bar{X}\frac{\alpha_i}{\sum_{j=1}^n \alpha_j^2}\right)^2 \\
&= \sigma^2 \left(\frac{1}{n} + \bar{X}^2 \frac{\sum_{i=1}^n \alpha_i^2}{(\sum_{j=1}^n \alpha_j^2)^2}\right) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right).
\end{aligned}$$

□

Preuve de la proposition 3.

Démonstration. On a :

$$\begin{aligned}
\mathbb{E}\left[\mathrm{SCR}(\hat{\beta}_0, \hat{\beta}_1)\right] &= \mathbb{E}\left[\sum_{i=1}^n \left(Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)\right)^2\right] \\
&= \mathbb{E}\left[\sum_{i=1}^n \left(Y_i - \bar{Y} - \hat{\beta}_1 (X_i - \bar{X})\right)^2\right] \\
&= \mathbb{E}\left[\sum_{i=1}^n \left((Y_i - \bar{Y})^2 + \hat{\beta}_1^2 (X_i - \bar{X})^2 - 2(Y_i - \bar{Y})\hat{\beta}_1 (X_i - \bar{X})\right)\right]
\end{aligned}$$

Or :

$$\sum_{i=1}^{n} \hat{\beta}_1^2 (X_i - \bar{X})^2 = \sum_{i=1}^{n} \left(\frac{\sum_{j=1}^{n} (X_j - \bar{X})(Y_j - \bar{Y})}{\sum_{j=1}^{n} (X_j - \bar{X})^2} \right)^2 (X_i - \bar{X})^2$$
$$= \frac{\left(\sum_{j=1}^{n} (X_j - \bar{X})(Y_j - \bar{Y}) \right)^2}{\sum_{j=1}^{n} (X_j - \bar{X})^2},$$

et :

$$\sum_{i=1}^{n} (Y_i - \bar{Y}) \hat{\beta}_1 (X_i - \bar{X}) = \sum_{i=1}^{n} (Y_i - \bar{Y}) \frac{\sum_{j=1}^{n} (X_j - \bar{X})(Y_j - \bar{Y})}{\sum_{j=1}^{n} (X_j - \bar{X})^2} (X_i - \bar{X})$$
$$= \frac{\left(\sum_{j=1}^{n} (X_j - \bar{X})(Y_j - \bar{Y}) \right)^2}{\sum_{j=1}^{n} (X_j - \bar{X})^2}.$$

Donc :

$$\mathbb{E}\left[\mathrm{SCR}(\hat{\beta}_0, \hat{\beta}_1) \right] = \mathbb{E}\left[\sum_{i=1}^{n} (Y_i - \bar{Y})^2 \right] - \mathbb{E}\left[\frac{\left(\sum_{j=1}^{n} (X_j - \bar{X})(Y_j - \bar{Y}) \right)^2}{\sum_{j=1}^{n} (X_j - \bar{X})^2} \right].$$

D'une part :

$$\mathbb{E}\left[\sum_{i=1}^{n} (Y_i - \bar{Y})^2 \right] = \mathbb{E}\left[\sum_{i=1}^{n} Y_i^2 \right] - n\mathbb{E}\left[\bar{Y}^2 \right]$$
$$= \mathbb{E}\left[\sum_{i=1}^{n} (\beta_0 + \beta_1 X_i + \epsilon_i)^2 \right] - n\mathbb{E}\left[\left(\beta_0 + \beta_1 \bar{X} + \frac{1}{n} \sum_{i=1} \epsilon_i \right)^2 \right]$$
$$= n\beta_0^2 + \beta_1^2 \sum_{i=1}^{n} X_i^2 + n\sigma^2 + 2\beta_0\beta_1 \sum_{i=1}^{n} X_i$$
$$- n\left[\beta_0^2 + \beta_1^2 \bar{X}^2 + \frac{1}{n}\sigma^2 + 2\beta_0\beta_1 \bar{X} \right]$$
$$= (n-1)\sigma^2 + \beta_1^2 \left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2 \right).$$

D'autre part, on veut calculer :

$$\mathbb{E}\left[\frac{\left(\sum_{j=1}^{n} (X_j - \bar{X})(Y_j - \bar{Y}) \right)^2}{\sum_{j=1}^{n} (X_j - \bar{X})^2} \right].$$

On s'intéresse au numérateur, num $:= \mathbb{E}[(\sum_{j=1}^n (X_j - \bar{X})(Y_j - \bar{Y}))^2]$:

$$\begin{aligned}
\text{num} &= \mathbb{E}\left[\sum_{j,k=1}^n (X_j - \bar{X})(X_k - \bar{X})(Y_j - \bar{Y})(Y_k - \bar{Y})\right] \\
&= \mathbb{E}\left[\sum_{j,k=1}^n (X_j - \bar{X})(X_k - \bar{X})(\beta_0 + \beta_1 X_j + \epsilon_j - \bar{Y})(\beta_0 + \beta_1 X_k + \epsilon_k - \bar{Y})\right] \\
&= \mathbb{E}\left[\sum_{j,k=1}^n (X_j - \bar{X})(X_k - \bar{X})(\beta_1 X_j + \epsilon_j)(\beta_1 X_k + \epsilon_k)\right] \\
&= \sum_{j,k=1}^n (X_j - \bar{X})(X_k - \bar{X})(\beta_1^2 X_j X_k + \beta_1 X_k \mathbb{E}[\epsilon_j] + \beta_1 X_j \mathbb{E}[\epsilon_k] + \mathbb{E}[\epsilon_j \epsilon_k]),
\end{aligned}$$

donc :

$$\begin{aligned}
\text{num} &= \beta_1^2 \sum_{j,k=1}^n (X_j - \bar{X})(X_k - \bar{X}) X_j X_k + \sigma^2 \sum_{j,k=1}^n (X_j - \bar{X})(X_k - \bar{X}) \\
&= \beta_1^2 \sum_{j,k=1}^n (X_j - \bar{X})(X_k - \bar{X}) X_j X_k + \sigma^2 \sum_{j=1}^n (X_j - \bar{X})^2 \\
&= \beta_1^2 \left(\sum_{j=1}^n (X_j - \bar{X})^2\right)^2 + \sigma^2 \sum_{j=1}^n (X_j - \bar{X})^2.
\end{aligned}$$

Par conséquent :

$$\begin{aligned}
\mathbb{E}\left[\frac{\left(\sum_{j=1}^n (X_j - \bar{X})(Y_j - \bar{Y})\right)^2}{\sum_{j=1}^n (X_j - \bar{X})^2}\right] &= \beta_1^2 \sum_{j=1}^n (X_j - \bar{X})^2 + \sigma^2 \\
&= \beta_1^2 \left(\sum_{i=1}^n X_i^2 - n\bar{X}^2\right) + \sigma^2.
\end{aligned}$$

Finalement, en regroupant les deux termes :

$$\mathbb{E}\left[\text{SCR}(\hat{\beta}_0, \hat{\beta}_1)\right] = (n-2)\sigma^2.$$

□

Preuve de la remarque 21.

Démonstration. On veut montrer que :

$$\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 = \sum_{i=1}^n (Y_i - \bar{Y})^2 - \sum_{i=1}^n (Y_i - \hat{Y}_i)^2,$$

c'est-à-dire :

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2.$$

On calcule :

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + 2\sum_{i=1}^{n}(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}).$$

Il suffit donc de montrer que :

$$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) = 0.$$

On rappelle que $(\hat{\beta}_0, \hat{\beta}_1)$ minimise SCR, donc :

$$\frac{\partial \text{SCR}}{\partial \beta_0}(\hat{\beta}_0, \hat{\beta}_1) = 2\sum_{i=1}^{n}(\hat{Y}_i - Y_i) = 0, \tag{4.33}$$

$$\frac{\partial \text{SCR}}{\partial \beta_1}(\hat{\beta}_0, \hat{\beta}_1) = 2\sum_{i=1}^{n} X_i(\hat{Y}_i - Y_i) = 0. \tag{4.34}$$

On a donc :

$$\begin{aligned}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) &= \sum_{i=1}^{n}(Y_i - \hat{Y}_i)(\hat{\beta}_0 + \hat{\beta}_1 X_i - \bar{Y}) \\ &= (\hat{\beta}_0 - \bar{Y})\sum_{i=1}^{n}(Y_i - \hat{Y}_i) + \hat{\beta}_1 \sum_{i=1}^{n} X_i(Y_i - \hat{Y}_i) \\ &= 0,\end{aligned}$$

en utilisant les équations (4.33) et (4.34). □

Preuve de la proposition 4.

Démonstration. Nous ne faisons qu'une partie de la démonstration : la recherche des points critiques de la SCR, c'est-à-dire les $\boldsymbol{\beta}$ tels que $\frac{\partial \text{SCR}}{\partial \boldsymbol{\beta}} = 0$. Nous supposons que la SCR est convexe, ce qui implique que le point critique que nous trouverons est le minimum global de la SCR.

En rappelant que, pour tout vecteur $\mathbf{x}$, on a : $\|\mathbf{x}\|^2 = \mathbf{x}^T\mathbf{x}$, on a :

$$\begin{aligned}\text{SCR}(\boldsymbol{\beta}) &= (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \mathbf{Y}^T\mathbf{Y} - \mathbf{Y}^T\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{Y} + (\mathbf{X}\boldsymbol{\beta})^T\mathbf{X}\boldsymbol{\beta} \\ &= \mathbf{Y}^T\mathbf{Y} - 2\mathbf{Y}^T\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta},\end{aligned}$$

puisque, pour tous vecteurs $\mathbf{u}$ et $\mathbf{v}$ de $\mathbb{R}^n$, $\mathbf{u}^T\mathbf{v} = \mathbf{v}^T\mathbf{u}$, et pour toutes matrices $\mathbf{A} \in \mathbb{R}^{n\times p}$ et $\mathbf{B} \in \mathbb{R}^{p\times n}$, $(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$.

On veut calculer le gradient de la SCR.

Rappel : pour une fonction $f : \mathbb{R}^n \to \mathbb{R}$, le gradient de f en $\mathbf{x}_0 \in \mathbb{R}^n$ est le vecteur ∇f tel que :

$$f(\mathbf{x}_0 + \mathbf{h}) - f(\mathbf{x}_0) = \mathbf{h}^T(\nabla f) + o(\|\mathbf{h}\|),$$

où $\mathbf{h} \in \mathbb{R}^n$. Or on a :

$$\begin{aligned}
\mathrm{SCR}(\boldsymbol{\beta} + \mathbf{h}) - \mathrm{SCR}(\boldsymbol{\beta}) &= -2\mathbf{Y}^T\mathbf{X}\mathbf{h} + \mathbf{h}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} \\
&= -2\mathbf{h}^T\mathbf{X}^T\mathbf{Y} + 2\mathbf{h}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} + o(\|\mathbf{h}\|) \\
&= \mathbf{h}^T\left(-2\mathbf{X}^T\mathbf{Y} + 2\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}\right) + o(\|\mathbf{h}\|).
\end{aligned}$$

Donc le gradient de la SCR vaut :

$$\frac{\partial \mathrm{SCR}}{\partial \boldsymbol{\beta}} = -2\mathbf{X}^T\mathbf{Y} + 2\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}.$$

Celui-ci s'annule lorsque :

$$-2\mathbf{X}^T\mathbf{Y} + 2\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = 0,$$

c'est-à-dire :

$$\boldsymbol{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}.$$

□

Chapitre 5

Analyse en composantes principales

5.1 Description de la méthode

Soit $\mathbf{X} \in \mathcal{M}_{n,p}$ un tableau de données avec n observations et p variables. Chaque observation $X_i \in \mathbb{R}^p$ est un vecteur colonne de dimension p : $\mathbf{X}^T = (X_1 \cdots X_n) \in \mathcal{M}_{p,n}$. On souhaite construire une représentation des observations dans un espace de dimension $d \leq p$.

Formalisation du problème. On cherche à approximer chaque X_i avec $\hat{X}_i = \mathbf{UW}X_i$, où $\mathbf{W} \in \mathcal{M}_{d,p}$ est la *matrice de compression* et $\mathbf{U} \in \mathcal{M}_{p,d}$ est la *matrice de reconstruction*. On transporte donc les $X_i \in \mathbb{R}^p$ sur un espace de dimension $d < p$ avec $\mathbf{W}$ et on les reconstruit en transportant le vecteur $\mathbf{W}X_i$ de dimension d dans l'espace de départ (de dimension p).

Le problème auquel nous nous intéressons est la minimisation de l'erreur de reconstruction, définie par :

$$L(\mathbf{U}, \mathbf{W}) := \sum_{i=1}^{n} \|X_i - \mathbf{UW}X_i\|^2, \tag{5.1}$$

par rapport aux paramètres $(\mathbf{U}, \mathbf{W})$.

Remarque 25. *Pour tout $\mathbf{W} \in \mathcal{M}_{d,p}$, $\mathbf{U} \in \mathcal{M}_{p,d}$, les vecteurs reconstruits $\hat{X}_i = \mathbf{UW}X_i \in \mathbb{R}^p$ appartiennent tous à $\mathrm{Im}(\mathbf{UW})$, qui est un sous-espace vectoriel de $\mathbb{R}^p$ de dimension au plus d.*

Démonstration. On a :

$$\dim(\mathrm{Im}(\mathbf{UW})) = \mathrm{rg}(\mathbf{UW}) \leq \min(\mathrm{rg}(\mathbf{U}), \mathrm{rg}(\mathbf{W})).$$

Or, $\mathrm{rg}(\mathbf{U}) \leq d$ et $\mathrm{rg}(\mathbf{W}) \leq d$, donc $\mathrm{rg}(\mathbf{UW}) \leq d$. □

Résolution du problème. On résout le problème (5.1) en deux étapes :

1. réduction de l'espace de recherche de $(\mathbf{U}, \mathbf{W})$: on se ramène à la recherche d'une matrice $\mathbf{V} \in \mathcal{M}_{p,d}$ dont les colonnes sont orthonormales ;

2. construction d'une matrice $\mathbf{V}$ telle que définie ci-dessus de sorte que $(\mathbf{U}, \mathbf{W}) = (\mathbf{V}, \mathbf{V}^T)$ est solution du problème de minimisation (5.1).

Lemme 1 (Partie 1)**.** *Soit* $(\mathbf{U}, \mathbf{W})$ *une solution du problème* (5.1)*. Alors il existe* $\mathbf{V} \in \mathcal{M}_{p,d}$ *dont les colonnes sont orthonormales tel que* $(\mathbf{V}, \mathbf{V}^T)$ *est aussi une solution.*

Le problème (5.1) prend maintenant une nouvelle forme : on cherche $\mathbf{V} \in \mathcal{M}_{p,d}$ dont les colonnes sont orthonormales minimisant :

$$\sum_{i=1}^{n} \|X_i - \mathbf{V}\mathbf{V}^T X_i\|^2. \tag{5.2}$$

Proposition 6 (Partie 2)**.** *Soit* $\mathbf{A} = \sum_{i=1}^{n} X_i X_i^T = \mathbf{X}^T\mathbf{X}$*. Définie ainsi, la matrice* $\mathbf{A} \in \mathcal{M}_p$ *est diagonalisable en base orthonormale et ses valeurs propres sont positives ou nulles. On note* $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ *ses valeurs propres et* $(v_1, v_2, \cdots, v_p)$ *des vecteurs propres associés à celles-ci, formant une base orthonormale de* $\mathbb{R}^p$*. Alors, en définissant la matrice* $\mathbf{V} \in \mathcal{M}_{p,d}$ *définie par* $\mathbf{V} = (v_1, \cdots, v_d)$*,* $(\mathbf{V}, \mathbf{V}^T)$ *est solution du problème* (5.1).

Remarque 26. *Avec* $d = p$*,* $\mathbf{V}$ *est une matrice orthogonale, donc* $\mathbf{V}\mathbf{V}^T = I_p$*, et on a :* $\hat{X}_i = \mathbf{V}\mathbf{V}^T X_i = X_i$*. Dans ce cas, la reconstruction est parfaite.*

Lorsque d décroît, la reconstruction $\hat{X}_i$ *de* X_i *est de moins en moins bonne. Tout se passe comme si l'on négligeait ou on « coupait » les plus petites valeurs propres de* $\mathbf{A}$*.*

5.2 Pré-traitement des données

En règle générale, il est nécessaire de pré-traiter les données avant de procéder à une Analyse en Composantes Principales (ACP). On applique généralement un centrage et une réduction des données

Centrage. Le centrage des données correspond à l'opération suivante, pour toute variable $X^{(j)}$:

$$X^{(j)} \leftarrow X^{(j)} - \bar{X}^{(j)}. \tag{5.3}$$

De cette façon, chaque variable a une moyenne nulle.

On effectue ce centrage, parce que les opérations de compression/reconstruction effectuées par l'ACP sont des opérations linéaires, et non affines. À cause de cela, si les données sont concentrées et loin de l'origine, l'ACP risque de supprimer une partie importante de l'information présente dans les données

Réduction. Après centrage des données, la réduction correspond à l'opération suivante, pour toute variable $X^{(j)}$:

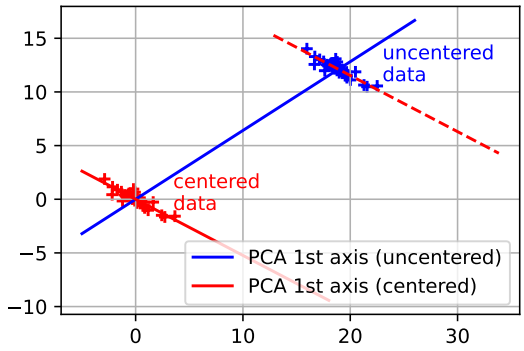
$$X^{(j)} \leftarrow \frac{X^{(j)} - \bar{X}^{(j)}}{\sigma_{X^{(j)}}}. \tag{5.4}$$

De cette façon, chaque variable a une variance de 1.

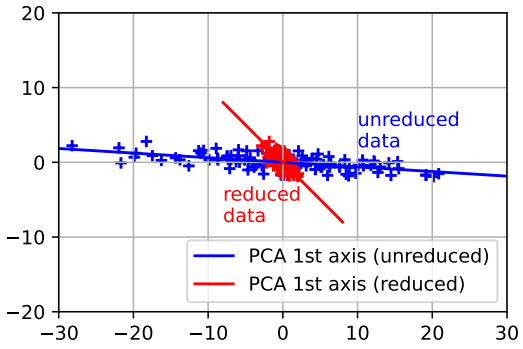
La réduction permet de rendre l'ACP invariante par changement d'échelle : le résultat de l'ACP ne change pas si l'on décide de multiplier par une constante λ toutes les valeurs correspondant à une variable.

Résumé. Les opérations de centrage et de réduction permettent d'éliminer les distorsions dues au positionnement et à l'échelle du nuage de points.

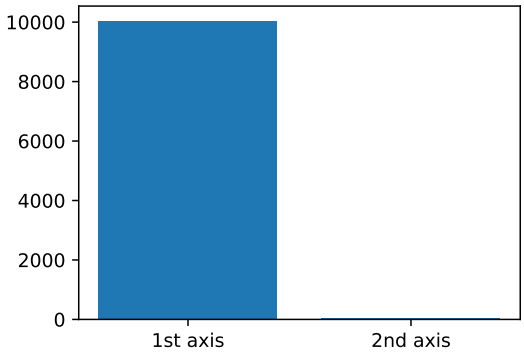
Remarque 27. *Après centrage et réduction, la matrice* $\frac{1}{n}\mathbf{X}^T\mathbf{X}$ *est la matrice de corrélation entre les* $X^{(j)}$.



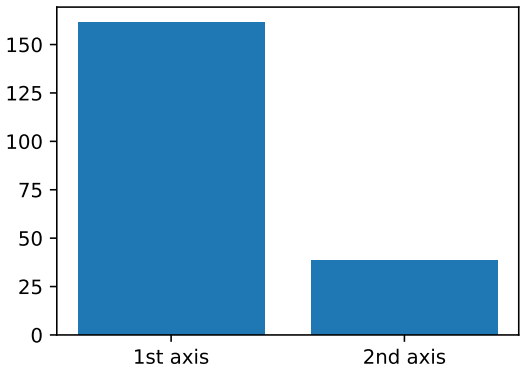
(a) Influence du centrage des données sur le premier axe de l'ACP : les directions obtenues sont quasiment orthogonales, seule celle obtenue après centrage est alignée avec l'axe de dispersion maximale des données.



(b) Influence de la réduction des données sur l'ACP : le poids de la coordonnée X sur le premier axe principal diminue après réduction des données, donc celui-ci provenait principalement de la grande variance de la variable X.



(c) Variance expliquée par les axes principaux sans réduction des données : la variance expliquée par le 2e axe principal semble négligeable.



(d) Variance expliquée par les axes principaux après réduction des données : la variance expliquée par le 2e axe principal n'est pas négligeable.

FIGURE 5.1 – Influence du centrage et de la réduction des données avant ACP.

5.3 Interprétation d'une ACP

Choix de d**.** Le choix de la dimension d est cruciale : en prenant d proche de p, on obtient une reconstruction presque exacte, mais sans réduction importante de la dimension ; en prenant d proche de 1, la reconstruction est moins précise, mais on réduit considérablement la dimension. Il a donc un compromis à trouver entre qualité de la reconstruction et réduction de la dimension.

Axes principaux. On appelle *axes principaux* les directions associées aux vecteurs $(v_1, \cdots, v_d)$, où chaque v_i est un vecteur propre associé à λ_i, avec $\lambda_1 \geq \cdots \geq \lambda_d$.

Pour un axe principal v_i donné, les composantes $(v_i)_j$ contiennent la contribution de chaque variable j à l'axe v_i :

— $(v_i)_j = 0$: pas de contribution de $X^{(j)}$ à l'axe v_i ;
— $(v_i)_j > 0$: contribution positive de $X^{(j)}$ à l'axe v_i ;
— $(v_i)_j < 0$: contribution négative de $X^{(j)}$ à l'axe v_i.

Valeurs propres. Les valeurs propres $\lambda_1 \geq \cdots \geq \lambda_d$ correspondent à l'importance de chaque axe $(v_1, \cdots, v_d)$ pour la reconstruction des données.

Pour mesurer l'importance de chaque axe, on construit les quantités suivantes :

— variance expliquée par l'axe v_i : $\frac{\lambda_i}{\lambda_1+\cdots+\lambda_p}$;
— variance expliquée par l'ACP de dimension d : $\frac{\lambda_1+\cdots+\lambda_d}{\lambda_1+\cdots+\lambda_p}$.

Plus la variance expliquée par l'ACP est proche de 1, plus l'ACP permet de compresser et reconstruire les données avec une bonne qualité.

Visualisation des données dans la base déterminée par les axes principaux. Dans la base déterminée par les axes principaux, chaque observation $X_i \in \mathbb{R}^p$ a pour coordonnées $\mathbf{V}^T X_i \in \mathbb{R}^d$. Pour ce qui concerne le tableau de données $\mathbf{X} \in \mathcal{M}_{n,p}$, celui-ci est transformé en $\mathbf{XV} \in \mathcal{M}_{n,d}$. Lorsque cet espace est de dimension $d = 2$, il est possible de tracer le nuage de points correspondant au tableau de données projeté $\mathbf{XV}$.

Cercle des corrélations. Le cercle des corrélations est un outil qui permet de visualiser la contribution de deux axes principaux pour reconstruire les variables du tableau de données initial.

On trace dans le cercle unité du plan p vecteurs ancrés à l'origine du repère, de coordonnées (x_i, y_i) $(i \in \{1, \cdots, p\})$. Les coordonnées (x_i, y_i) sont données par la procédure suivante :

1. on sélectionne les vecteurs directeurs de deux axes principaux, que l'on notera $v_1, v_2 \in \mathbb{R}^p$. Il s'agit en général des vecteurs correspondant aux deux premiers axes principaux, mais on peut en choisir d'autres ;

2. on définit les vecteurs $w_1, w_2 \in \mathbb{R}^n$ tels que :

$$w_1 = (\mathbf{X}v_1)^T \qquad\qquad w_2 = (\mathbf{X}v_2)^T, \tag{5.5}$$

   qui contiennent la contribution de chacun des deux axes à chacune des n observations. Si $(w_2)_i$ est grand, alors l'axe 2 (représenté par v_2) a une forte contribution dans l'observation i ;

3. pour chaque variable $i \in \{1, \cdots, p\}$, on définit :

$$x_i := \mathrm{corr}(X^{(i)}, w_1) \qquad\qquad y_i := \mathrm{corr}(X^{(i)}, w_2). \tag{5.6}$$

5.4 Régression en composantes principales

Enjeux. Lorsque qu'on veut effectuer une régression linéaire de $\mathbf{Y}\mathbb{R}^n$ en fonction de $\mathbf{X} \in \mathcal{M}_{n,p}$ avec un grand nombre p de variables, il peut arriver que certaines d'entre elles soient redondantes et aboutissent à une matrice $\mathbf{X}^T\mathbf{X}$ difficile ou impossible à inverser.

Le but de la régression en composantes principales (RCP) est de réduire le nombre de variables sur lesquelles effectuer la régression, en ne conservant que les combinaisons de variables les plus pertinentes. Elle permet de surmonter plusieurs difficultés :

— la régression est difficile à visualiser (trop de variables explicatives) ;
— la régression est difficile à interpréter ;
— les variables explicatives sont très corrélées.

Principe. Avant de procéder à la régression, on effectue une ACP sur $\mathbf{X}$ (après centrage et réduction) avec d axes principaux. On note $\mathbf{Z} = \mathbf{XV} \in \mathcal{M}_{n,d}$ le tableau de données projeté dans la base des axes principaux.

Puis on effectue la régression linéaire de $\mathbf{Y}$ en fonction de $\mathbf{Z}$. Tout d'abord, on ajoute une colonne de 1 dans le tableau de données $\mathbf{Z}$, et on obtient le nouveau tableau $\tilde{\mathbf{Z}} \in \mathcal{M}_{n,d+1}$.

La régression linéaire fournit donc un vecteur $\boldsymbol{\beta} \in \mathbb{R}^{d+1}$ permettant de modéliser $\mathbf{Y}$ en fonction de $\mathbf{Z}$:

$$\hat{\mathbf{Y}} = \tilde{\mathbf{Z}}\boldsymbol{\beta}. \tag{5.7}$$

Puis il suffit d'exprimer $\hat{\mathbf{Y}}$ en fonction de $\mathbf{X}$:

$$\begin{aligned} \hat{\mathbf{Y}} &= \mathbf{Z}\boldsymbol{\beta}_{1:d} + \beta_{d+1}\mathbb{1}_n & (5.8)\\ &= \mathbf{XV}\boldsymbol{\beta}_{1:d} + \beta_{d+1}\mathbb{1}_n, & (5.9) \end{aligned}$$

où $\boldsymbol{\beta}_{1:d} \in \mathbb{R}^d$ est le vecteur contenant les d premières coordonnées de $\boldsymbol{\beta}$, $\beta_{d+1} \in \mathbb{R}$ est la coordonnées $d+1$ de $\boldsymbol{\beta}$, et $\mathbb{1}_n \in \mathbb{R}^n$ est le vecteur de $\mathbb{R}^n$ ne contenant que des 1.

5.5 Preuves

Preuve du lemme 1.

Démonstration. Soient $\mathbf{U} \in \mathbb{R}^{p\times d}$ et $\mathbf{W} \in \mathbb{R}^{d\times p}$ deux matrices. On va montrer qu'il existe une matrice $\mathbf{V} \in \mathbb{R}^{p\times d}$ dont les colonnes sont orthonormales telle que :

$$\mathcal{L}(\mathbf{V}, \mathbf{V}^T) \leq \mathcal{L}(\mathbf{U}, \mathbf{W}).$$

On construit $\mathbf{V} \in \mathbb{R}^{p\times d}$ une matrice aux colonnes orthonormales telle que $\mathrm{Im}(\mathbf{U}) \subset \mathrm{Im}(\mathbf{V})$. Notons $d' = \mathrm{rg}(\mathbf{U})$. On construit une base orthonormale $(\mathbf{v}_1, \cdots, \mathbf{v}_{d'})$ de $\mathrm{Im}(\mathbf{U})$. Si $d' = d$, alors on pose $\mathbf{V} = (\mathbf{v}_1, \cdots, \mathbf{v}_d)$. Si $d' < d$, alors on la complète avec des vecteurs quelconques pour obtenir une famille orthonormale avec d vecteurs $(\mathbf{v}_1, \cdots, \mathbf{v}_{d'}, \mathbf{v}_{d'+1}, \cdots, \mathbf{v}_d)$ et on pose $\mathbf{V} = (\mathbf{v}_1, \cdots, \mathbf{v}_d)$. Ainsi, on a :

$$\mathrm{Im}(\mathbf{U}) \subset \mathrm{Im}(\mathbf{V}).$$

On rappelle que l'on veut minimiser :

$$\mathcal{L}(\mathbf{W}, \mathbf{U}) = \sum_{i=1}^{n} \|X_i - \mathbf{U}\mathbf{W}X_i\|^2.$$

Considérons un point de donnée quelconque $\mathbf{x} \in \mathbb{R}^p$. On va comparer son erreur de reconstruction $\|\mathbf{x} - \mathbf{U}\mathbf{W}\mathbf{x}\|^2$ à l'erreur de reconstruction minimale qu'on peut obtenir en restreignant sa reconstruction à $\mathrm{Im}(\mathbf{V})$. On veut donc calculer :

$$\min_{\tilde{\mathbf{x}} \in \mathrm{Im}(\mathbf{V})} \|\mathbf{x} - \tilde{\mathbf{x}}\|^2 = \min_{\mathbf{y} \in \mathbb{R}^d} \|\mathbf{x} - \mathbf{V}\mathbf{y}\|^2.$$

On veut minimiser la fonction $f : \mathbf{y} \mapsto \|\mathbf{x} - \mathbf{V}\mathbf{y}\|^2$. On a :

$$\begin{aligned} f(\mathbf{y}) &= \|\mathbf{x}\|^2 + \|\mathbf{V}\mathbf{y}\|^2 - 2\mathbf{x}^T\mathbf{V}\mathbf{y} \\ &= \|\mathbf{x}\|^2 + \mathbf{y}^T\mathbf{V}^T\mathbf{V}\mathbf{y} - 2\mathbf{x}^T\mathbf{V}\mathbf{y} \\ &= \|\mathbf{x}\|^2 + \mathbf{y}^T\mathbf{y} - 2\mathbf{x}^T\mathbf{V}\mathbf{y}, \end{aligned}$$

car $\mathbf{V}^T\mathbf{V} = \mathbf{I}_d$, les colonnes de $\mathbf{V}$ étant orthonormales. Pour trouver un minimum de f, il suffit de trouver un point critique (point où le gradient de f s'annule), et de montrer que f est convexe. On calcule le gradient de f :

$$\nabla f(\mathbf{y}) = 2\mathbf{y} - 2\mathbf{V}^T\mathbf{x},$$

qui s'annule en un unique point :

$$\mathbf{y} = \mathbf{V}^T\mathbf{x}.$$

Pour montrer que f est convexe, on peut calculer sa hessienne et montrer que ses valeurs propres sont toutes positives :

$$\mathbf{H}_f(\mathbf{y}) = 2\mathbf{I}_d.$$

Donc $\mathbf{H}_f(\mathbf{y})$ a pour unique valeur propre 2, qui est strictement positive. Donc f est convexe et atteint son minimum en son unique point critique $\mathbf{y} = \mathbf{V}^T\mathbf{x}$.

On a donc :

$$\min_{\tilde{\mathbf{x}} \in \mathrm{Im}(\mathbf{V})} \|\mathbf{x} - \tilde{\mathbf{x}}\|^2 = \|\mathbf{x} - \mathbf{V}\mathbf{V}^T\mathbf{x}\|^2.$$

Or $\mathrm{Im}(\mathbf{U}) \subset \mathrm{Im}(\mathbf{V})$, donc :

$$\|\mathbf{x} - \mathbf{U}\mathbf{W}\mathbf{x}\|^2 \geq \min_{\tilde{\mathbf{x}} \in \mathrm{Im}(\mathbf{U})} \|\mathbf{x} - \tilde{\mathbf{x}}\|^2 \geq \min_{\tilde{\mathbf{x}} \in \mathrm{Im}(\mathbf{V})} \|\mathbf{x} - \tilde{\mathbf{x}}\|^2 = \|\mathbf{x} - \mathbf{V}\mathbf{V}^T\mathbf{x}\|^2.$$

En appliquant ce résultat à $(X_1, \cdots, X_n)$, on a :

$$\forall i \in \{1, \cdots, n\}, \quad \|X_i - \mathbf{U}\mathbf{W}X_i\|^2 \geq \|X_i - \mathbf{V}\mathbf{V}^T X_i\|^2.$$

En sommant sur i, on a :

$$\sum_{i=1}^{n} \|X_i - \mathbf{U}\mathbf{W}X_i\|^2 \geq \sum_{i=1}^{n} \|X_i - \mathbf{V}\mathbf{V}^T X_i\|^2,$$

c'est-à-dire :

$$\mathcal{L}(\mathbf{U}, \mathbf{W}) \geq \mathcal{L}(\mathbf{V}, \mathbf{V}^T).$$

En particulier, prenons $(\mathbf{U}, \mathbf{W})$ qui minimise $\mathcal{L}$. On a montré qu'il existe une matrice $\mathbf{V} \in \mathbb{R}^{p \times d}$ dont les colonnes sont orthonormales telle que $\mathcal{L}(\mathbf{U}, \mathbf{W}) \geq \mathcal{L}(\mathbf{V}, \mathbf{V}^T)$. Comme par ailleurs $(\mathbf{U}, \mathbf{W})$ minimise $\mathcal{L}$, alors $\mathcal{L}(\mathbf{U}, \mathbf{W}) \leq \mathcal{L}(\mathbf{V}, \mathbf{V}^T)$. Donc $\mathcal{L}(\mathbf{U}, \mathbf{W}) = \mathcal{L}(\mathbf{V}, \mathbf{V}^T)$. □

Chapitre 6

Tests statistiques appliqués à l'analyse de données

Dans les méthodes présentées précédemment, on calcule des corrélations et des coefficients de régression linéaire. Il est souvent utile de savoir si de telles valeurs sont nulles ou non. En pratique, il est très rare qu'une corrélation ou qu'un coefficient de régression soit rigoureusement nul. Cela est dû au fait que les données sur lesquelles nous travaillons sont *bruitées* : elles contiennent un aléa qui est essentiellement inexplicable (il est impossible d'avoir une connaissance suffisante du monde pour prédire une valeur avec une précision infinie).

Par exemple, considérons deux variables aléatoires (X, Y). Supposons que X et Y soient indépendantes, et qu'on tire un échantillon $((X_1, Y_1), \cdots, (X_n, Y_n))$ de n observations i.i.d. de même loi que (X, Y). Comme X et Y sont indépendantes, alors $\mathrm{corr}(X, Y) = 0$. Pourtant, si l'on calcule la corrélation empirique $\overline{\mathrm{corr}}(X, Y)$ du tableau de données contenant $((X_1, Y_1), \cdots, (X_n, Y_n))$, il est probable que $\overline{\mathrm{corr}}(X, Y)$ soit non nul.

On rappelle que :

— lorsque X et Y sont des variables aléatoires :

$$\mathrm{corr}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{\mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)]}{\sqrt{\mathbb{E}[(X - \mathbb{E}X)^2]\mathbb{E}[(Y - \mathbb{E}Y)^2]}}; \tag{6.1}$$

— lorsque X et Y sont des variables d'un tableau de données :

$$\overline{\mathrm{corr}}(X, Y) = \frac{\overline{\mathrm{Cov}}(X, Y)}{\bar{\sigma}_X \bar{\sigma}_Y} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}}. \tag{6.2}$$

6.1 Introduction

Nous allons nous intéresser dans un premier temps à un cas d'étude simple des tests d'hypothèse : la pièce de monnaie truquée.

Présentation du problème. On dispose d'une pièce de monnaie qui tombe sur pile avec probabilité $\theta \in]0, 1[$ et sur face avec probabilité $1 - \theta$. On a effectué n lancers

indépendants, que l'on note $(X_1, \cdots, X_n)$. Par convention, $X_i = 1$ si le résultat du i-ième lancer est pile, et $X_i = 0$ si le résultat du i-ième lancer est face.

À partir de la séquence $(X_1, \cdots, X_n)$ obtenue, on souhaite déterminer si la pièce est truquée. Autrement dit, on cherche à vérifier si l'on a $\theta = 1/2$ (pièce non truquée), ou $\theta \neq 1/2$ (pièce truquée).

Intuition. Pour nous faire une idée de la difficulté de la tâche, considérons trois situations différentes :

1. on suppose que, sur 100 lancers effectués, on a obtenu 100 fois pile. Intuitivement, il est quasiment certain que la pièce est truquée ;
2. on suppose que, sur 100 lancers effectués, on a obtenu 50 fois pile et 50 fois face. Intuitivement, il est très difficile d'affirmer que la pièce est truquée ;
3. on suppose que, sur 100 lancers effectués, on a obtenu 63 fois pile et 37 fois face. Dans cette situation intermédiaire, il est difficile de se servir de son intuition pour conclure. Certes, la pièce semble "biaisée" en faveur de *pile*, mais on ne sait pas si ce "biais" provient de l'aléa de l'échantillonnage ou est dû au fait que la pièce est réellement truquée.

   Ce cas intermédiaire est analogue au cas où $\overline{\text{corr}}(X, Y) = 0.1$: les variables X et Y sont-elles fondamentalement décorrélées, et a-t-on obtenu 0.1 à cause de l'aléa d'échantillonnage ? ou cette corrélation positive provient-elle d'une corrélation positive fondamentale entre X et Y ?

Idée de base. Pour déterminer si la pièce de monnaie est truquée ou non, on procède par réfutation d'hypothèse. D'abord, on fait l'hypothèse que la pièce n'est pas truquée. Dans ce cadre, la proportion r de *pile* dans notre échantillon est censée être relativement proche de $1/2$. En particulier, les valeurs de r éloignées de $1/2$ sont moins probables que celles qui en sont proches. On peut donc se servir de r pour conserver ou réfuter l'hypothèse de départ.

Si la valeur trouvée pour r est très improbable (i.e., éloignée de $1/2$), alors on va considérer que l'hypothèse de départ, à savoir "la pièce n'est pas truquée", est réfutée, et on va adopter l'hypothèse alternative : la pièce est truquée. Si, par contre, la valeur trouvée pour r n'est pas improbable, alors on va considérer que nous ne disposons pas de données suffisantes pour réfuter l'hypothèse de départ, on va donc la conserver.

Les étapes du test sont donc les suivantes :

1. définition des hypothèses $\mathbb{H}_0$ et $\mathbb{H}_1$ ($\mathbb{H}_0$: pièce non truquée, $\mathbb{H}_1$: pièce truquée) ;
2. définition d'un *seuil de rejet* $\alpha \in]0, 1[$;
3. construction de la *statistique de test* (ici, il s'agira de la proportion r de *pile* dans l'échantillon) ;
4. sous l'hypothèse que la pièce n'est pas truquée, calcul de la probabilité p d'obtenir une proportion de *pile* plus extrême que r (i.e., plus éloignée de $1/2$ que r) ;
5. comparaison de p (la *p-valeur*) avec le seuil α : si $p \geq \alpha$, alors on conserve $\mathbb{H}_0$; si $p < \alpha$, alors on rejette $\mathbb{H}_0$ et on adopte $\mathbb{H}_1$.

Définition des hypothèses. La première étape du test statistique consiste à définir deux hypothèses : l'*hypothèse nulle* (ou *hypothèse conservatrice*), notée $\mathbb{H}_0$, et l'*hypothèse alternative*, notée $\mathbb{H}_1$. L'hypothèse $\mathbb{H}_0$ est celle que nous allons initialement considérer comme vraie, et que nous allons chercher à réfuter. L'hypothèse $\mathbb{H}_1$ est l'hypothèse que nous adopterons en cas de réfutation de $\mathbb{H}_0$.

Dans l'étude du cas de la pièce de monnaie truquée, on choisit :

— hypothèse $\mathbb{H}_0 : \theta = 1/2$ (pièce non truquée) ;

— hypothèse $\mathbb{H}_1 : \theta \neq 1/2$ (pièce truquée).

Nous discuterons ce choix à la fin de cette section.

Construction de la statistique de test. Pour déterminer si l'on doit conserver $\mathbb{H}_0$ ou adopter $\mathbb{H}_1$, on va construire une *statistique de test*. Une statistique de test est une variable aléatoire que l'on peut calculer à partir de l'échantillon $(X_1, \cdots, X_n)$. De plus, celle-ci doit être utilisable pour vérifier si $\mathbb{H}_0$ peut être réfutée ou non.

Dans le cas du test d'une pièce de monnaie, il peut être intéressant d'étudier la statistique R définie par :

$$R := \frac{1}{n} S_n \qquad\qquad S_n := \sum_{i=1}^{n} X_i. \tag{6.3}$$

En effet, R est égal à la proportion de 1 (représentant *pile*) dans l'échantillon $(X_1, \cdots, X_n)$. Intuitivement, R devrait être proche de $1/2$ si la pièce est non truquée, et devrait s'éloigner de $1/2$ si la pièce est truquée.

Formellement, les X_i sont des variables aléatoires de Bernoulli de même paramètre θ : $X_i = 1$ avec probabilité θ et $X_i = 0$ avec probabilité $1 - \theta$. On a notamment $\mathbb{E}[X_i] = \theta$. Par conséquent, quelle que soit la valeur de θ (peu importe que la pièce soit truquée ou non), on a :

$$\mathbb{E}[R] = \frac{1}{n}\mathbb{E}[S_n] = \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \frac{1}{n}\sum_{i=1}^{n}\theta = \frac{1}{n}\cdot n\theta = \theta. \tag{6.4}$$

Donc, que la pièce soit truquée ou non, la statistique R sera centrée autour du paramètre θ, qui représente la probabilité que la pièce testée donne *pile*. Par conséquent, si $\theta = 1/2$, R sera centrée autour de $1/2$, et si $\theta \neq 1/2$, R ne sera pas centrée autour de $1/2$.

On peut donc espérer pouvoir utiliser R pour déterminer si la pièce est truquée : une valeur de R très éloignée de $1/2$ est très improbable si l'on suppose que $\theta = 1/2$. Dans ce cas, il serait raisonnable de rejeter cette hypothèse et d'admettre que $\theta \neq 1/2$.

Loi de la statistique de test sous $\mathbb{H}_0$. La statistique de test que nous avons retenue est R, définie en (6.4). Pour effectuer le test, il faut connaître la loi de R sous $\mathbb{H}_0$, c'est-à-dire en supposant que $\mathbb{H}_0$ est vraie (i.e., $\theta = 1/2$). De façon générale, la loi de R n'est pas très difficile à obtenir, même si l'on ne suppose pas que $\theta = 1/2$. Il suffit de remarquer que S_n est une somme de n variables aléatoires de Bernoulli X_i de paramètre θ, indépendantes. Par conséquent, S_n est une variable aléatoire qui suit une loi binomiale de paramètres (n, θ) : $S_n \sim \mathcal{B}(n, \theta)$. La variable aléatoire $R = \frac{1}{n} S_n$ suit donc, à un facteur près, une loi proche d'une binomiale.

Cependant, pour des raisons pédagogiques, il est plus intéressant de montrer comment on procède avec une statistique de test de loi *continue*, et non discrète comme une binomiale.

On va donc approximer la loi de R par une loi continue. Cela est possible lorsque $n \gtrsim 30$, $n\theta \gtrsim 5$ et $n(1-\theta) \gtrsim 5$: dans ce cas, on peut approximer une loi binomiale $\mathcal{B}(n, \theta)$ par une loi gaussienne $\mathcal{N}(n\theta, n\theta(1-\theta))$. Sous ces conditions, S_n suit approximativement une loi $\mathcal{N}(n\theta, n\theta(1-\theta))$, et R suit approximativement une loi $\mathcal{N}(\theta, \frac{1}{n}\theta(1-\theta))$.

Dans la suite, on considérera donc que, sous $\mathbb{H}_0$ (i.e., en supposant $\theta = 1/2$) :

$$R \sim \mathcal{N}\left(\frac{1}{2}, \frac{1}{4n}\right). \tag{6.5}$$

Calcul de la p-valeur. Étant donné l'échantillon $(X_1, \cdots, X_n)$, on calcule la statistique de test sur celui-ci, que l'on note r. La quantité r est un nombre réel fixé, et non une variable aléatoire, c'est pourquoi on la note avec un r minuscule.

Sous $\mathbb{H}_0$, on s'attend à ce que r soit proche de $1/2$. C'est pourquoi on cherche maintenant à calculer la probabilité que, sous $\mathbb{H}_0$, R soit plus éloigné de $1/2$ que r. Cette probabilité nous permet de quantifier à quel point la statistique r obtenue sur notre échantillon est probable sous l'hypothèse $\mathbb{H}_0$. S'il est trop peu probable d'obtenir une statistique R "au moins aussi extrême" que r sous $\mathbb{H}_0$, alors il faudra réfuter $\mathbb{H}_0$.

Dans le cas du test de la pièce truquée, on définit donc la p-valeur comme suit :

$$p := \mathbb{P}\left(\left|R - \frac{1}{2}\right| \geq \left|r - \frac{1}{2}\right|\right) = 2\,\min(\mathbb{P}(R \geq r), \mathbb{P}(R \leq r)). \tag{6.6}$$

La probabilité p est la probabilité que, sous l'hypothèse $\mathbb{H}_0$, l'échantillon produise une statistique R plus éloignée de $1/2$ que r (i.e., plus extrême que r).

Étant donné un *seuil de rejet* $\alpha \in]0, 1[$, il ne reste plus qu'à comparer p et α pour conclure :

- si $p > \alpha$, alors on conserve $\mathbb{H}_0$: il n'est pas improbable que, si $\mathbb{H}_0$ est vraie, on tire un échantillon qui donne une statistique R au moins aussi extrême que r ;
- si $p \leq \alpha$, alors on rejette $\mathbb{H}_0$ et on adopte $\mathbb{H}_1$: il est improbable que, en admettant que $\mathbb{H}_0$ soit vraie, on tire un échantillon qui donne une statistique R au moins aussi extrême que r.

Les valeurs courantes pour α sont 0.05 ou 0.01. Plus le seuil α est bas, plus on exige d'avoir une statistique r éloignée de $1/2$ pour rejeter $\mathbb{H}_0$. On est alors plus "conservateur" qu'avec un seuil α élevé.

6.2 Résumé de la méthode

Procédure complète :

1. selon la situation, déterminer quel test utiliser et noter :
   - les hypothèses $\mathbb{H}_0$ et $\mathbb{H}_1$,
   - si le test est unilatéral ou bilatéral,
   - la formule de la statistique de test X à utiliser,
   - la loi de X sous $\mathbb{H}_0$;

2. fixer le seuil de rejet $\alpha \in]0,1[$;
3. calculer la statistique de test sur l'échantillon, qu'on notera x ;
4. selon les moyens à disposition, calculer la p-valeur ou calculer le seuil x_α de niveau α associé à la statistique de test :
   (a) dans le cas où l'on peut calculer numériquement $\mathbb{P}(X > x)$ ou $\mathbb{P}(|X| > x)$ (en utilisant la loi de X sous $\mathbb{H}_0$), on calcule la p-valeur, notée p :
      — si le test est unilatéral, $p = \mathbb{P}(X > x)$, ou plus rarement $p = \mathbb{P}(X < x)$ (selon les spécifications du test),
      — si le test est bilatéral, $p = \mathbb{P}(|X| > x)$,
   (b) dans le cas où l'on dispose d'une table donnant, pour la loi de X sous $\mathbb{H}_0$ et pour le seuil α choisi, un seuil x_α défini de sorte que :
      — si $x' \leq x_\alpha$, alors la p-valeur p associée à x' est plus grande que α,
      — si $x' > x_\alpha$, alors la p-valeur p associée à x' est plus petite que α.
5. on conclut en fonction de la méthode utilisée au point 4 :
   (a) on conserve $\mathbb{H}_0$ si $p > \alpha$, et on adopte $\mathbb{H}_1$ si $p \leq \alpha$,
   (b) on conserve $\mathbb{H}_0$ si $x < x_\alpha$, et on adopte $\mathbb{H}_1$ si $x \geq x_\alpha$.

Remarque 28 (Que conclure d'une conservation de $\mathbb{H}_0$?)**.** *Ce n'est pas parce qu'on a conservé $\mathbb{H}_0$ que $\mathbb{H}_0$ est vraie :*

— *il est possible qu'on ait manqué de données pour rejeter $\mathbb{H}_0$; il faudrait alors refaire le test avec un échantillon plus grand (n plus grand) ;*

— *il est possible que la statistique de test n'ait pas été judicieusement choisie : la* puissance du test *est alors faible, et il est difficile de réfuter $\mathbb{H}_0$, même avec beaucoup de données ; il faudrait construire un test plus puissant, si cela est possible.*

Remarque 29 (**Interprétation erronée de la p-valeur**)**.** *Certes la p-valeur, notée p, est une probabilité, et on rejette $\mathbb{H}_0$ lorsque p est petite, mais on ne peut pas interpréter p comme la probabilité que $\mathbb{H}_0$ soit vraie.*

Interprétation pratique du test. Les tests statistiques sont particulièrement intéressants dans les situations où il est raisonnable d'avoir une "hypothèse par défaut" $\mathbb{H}_0$, qu'on souhaite conserver tant qu'elle n'est pas réfutée.

Par exemple, dans le cadre du test d'un nouveau médicament, il est plus raisonnable de considérer que c'est à celui-ci de faire ses preuves, plutôt que de partir du principe qu'il fonctionne. De cette façon, on évite de mettre sur le marché un médicament dont l'efficacité n'est pas prouvée. Dans ce cas, l'hypothèse $\mathbb{H}_0$ est donc "le médicament n'a pas d'effet spécifique", et l'hypothèse $\mathbb{H}_1$ est "le médicament a un effet spécifique". Le test statistique doit donc être suffisamment bien conçu, avec suffisamment de données, pour que l'effet du médicament puisse être démontré avec une bonne p-valeur.

Cependant, il peut exister des situations où les tests tels que décrits ici ne sont pas adaptés. En effet, on ne peut pas choisir arbitrairement $\mathbb{H}_0$ et $\mathbb{H}_1$ en fonction de nos besoins. Reprenons par exemple le cas de la pièce de monnaie truquée, mais en intervertissant $\mathbb{H}_0$ et $\mathbb{H}_1$. Ce cadre d'étude peut être intéressant si l'on part du principe que la pièce est truquée, et que l'on cherche à réfuter cette hypothèse. Mais, avec de telles hypothèses $\mathbb{H}_0$ et $\mathbb{H}_1$, il sera impossible de suivre la procédure indiquée plus haut : sous $\mathbb{H}_0$, on a $\theta \neq 1/2$, ce qui est une hypothèse insuffisante pour déterminer la loi de la statistique de test. Or, sans la loi de la statistique de test, il est impossible de calculer la p-valeur.

6.3 Application à la corrélation

Soient $\tilde{X}$ et $\tilde{Y}$ deux variables aléatoires. On dispose d'un échantillon $((X_1, Y_1), \cdots (X_n, Y_n))$, où chaque observation (X_i, Y_i) est tirée de manière indépendante et selon la même loi que $(\tilde{X}, \tilde{Y})$. On note $r := \overline{\text{corr}}(X, Y)$ la corrélation empirique entre les variables X et Y sur l'échantillon, et on note $\rho = \text{corr}(\tilde{X}, \tilde{Y})$ la corrélation entre les variables aléatoires $\tilde{X}$ et $\tilde{Y}$.

En utilisant l'échantillon $((X_1, Y_1), \cdots (X_n, Y_n))$ et la corrélation empirique r, on souhaite déterminer si la corrélation ρ est nulle ou non.

Caractéristiques du test :

- hypothèses :
  - $\mathbb{H}_0$: $\rho = 0$: la corrélation entre les variables est nulle,
  - $\mathbb{H}_1$: $\rho \neq 0$: la corrélation entre les variables est non nulle ;
- statistique de test :

$$T := r\sqrt{\frac{n-2}{1-r^2}}; \tag{6.7}$$

- loi de T sous l'hypothèse $\mathbb{H}_0$: T suit une loi de Student à $n-2$ degrés de liberté, que l'on peut noter $\mathcal{T}(n-2)$;
- test bilatéral.

Calcul de la p-valeur : tableau à double entrée : colonnes : seuil α ; lignes : nombre de degrés de liberté.

Remarque 30. *Ce test est valable dans le cadre où $(\tilde{X}, \tilde{Y})$ est un vecteur gaussien. Si ce n'est pas le cas, on peut discuter la pertinence de cette méthode.*

Applications.

Exemple 26 (Absence de corrélation entre $\tilde{X}$ et $\tilde{Y}$)**.** *Cas où $\tilde{X}$ et $\tilde{Y}$ sont des gaussiennes indépendantes.*

n	10	20	50	100	200	500	1000	2000	5000	10000
T	-0.09	0.59	-1.10	-0.162	-0.882	-0.722	2.707	1.327	0.808	-0.75
p-valeur	0.931	0.562	0.277	0.872	0.379	0.471	0.007	0.185	0.419	0.453

Si le seuil de rejet α est de 0.05, alors on ne rejette $\mathbb{H}_0$ que pour l'expérience menée avec $n = 1000$: la p-valeur vaut $p = 0.007 < 0.05 = \alpha$.

Exemple 27 (Présence de corrélation entre $\tilde{X}$ et $\tilde{Y}$)**.** *Cas où $\tilde{X}$ et $\tilde{Y}$ sont des gaussiennes corrélées.*

n	10	20	50	100	200	500	1000	2000	5000	10000
T	2.45	3.00	6.25	11.8	14.1	20.2	33.1	44.8	69.4	100
p-valeur	0.04	0.0077	10^{-7}	0	0	0	0	0	0	0

Si le seuil de rejet α est de 0.01, alors on ne conserve $\mathbb{H}_0$ que pour l'expérience menée avec $n = 10$: la p-valeur vaut $p = 0.04 \geq 0.01 = \alpha$. On rejette $\mathbb{H}_0$ dans tous les autres cas.

6.4 Application à la régression linéaire

On effectue une régression linéaire avec le modèle suivant : $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\beta}} \in \mathbb{R}^{p+1}$, où $\hat{\boldsymbol{\beta}}$ est le paramètre minimisant la SCR (voir chapitre sur la régression linéaire).

On suppose que les Y_i ont été obtenus de la façon suivante :

$$Y_i = X_i^T \boldsymbol{\beta} + \epsilon_i, \tag{6.8}$$

où $X_i \in \mathbb{R}^{p+1}$ est le vecteur représentant l'observation i, $\boldsymbol{\beta}$ est le vecteur des paramètres sous-jacents $(\beta_1, \cdots, \beta_{p+1})$, et les ϵ_i sont des variables aléatoires gaussiennes indépendantes et identiquement distribuées.

On souhaite déterminer si les paramètres sous-jacents $\beta_1, \cdots, \beta_p$ sont tous nuls ou bien si au moins l'un d'entre eux est significativement éloigné de 0.

Caractéristiques du test :

- hypothèses :
  - $\mathbb{H}_0 : \beta_1 = \cdots = \beta_p = 0$: tous les paramètres sous-jacents sont nuls,
  - $\mathbb{H}_1 : \exists j : \beta_j \neq 0$: un des paramètres au moins est non nul ;
- statistique de test :

$$F := \frac{\frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{p}}{\frac{\sum_{i=1}^n (\hat{Y}_i - Y_i)^2}{n-p-1}}; \tag{6.9}$$

- loi de F sous l'hypothèse $\mathbb{H}_0$: F suit une loi de Fisher $\mathcal{F}(p, n-p-1)$ à p degrés de liberté pour le numérateur et $n-p-1$ degrés de liberté pour le dénominateur ;
- test unilatéral.

Calcul de la p-valeur : pour chaque seuil α, il existe un tableau à double entrée : on sélectionne la colonne correspondant au nombre de degrés de liberté au numérateur, et la ligne correspondant au nombre de degrés de liberté au dénominateur. Par exemple : $(p, n-p-1)$, ou $(q, n-p-1)$.

Applications.

Exemple 28 ($\beta_1 = 0$)**.** *Cas où $Y_i = \beta_0 + \epsilon_i$, c'est-à-dire $\beta_1 = 0$.*

n	10	20	50	100	200	500	1000	2000	5000	10000
F	5.881	1.147	0.873	3.224	0.001	0.975	2.115	0.209	1.933	1.350
p-valeur	0.042	0.298	0.355	0.076	0.981	0.324	0.146	0.648	0.165	0.245

Note : pour $n = 10$, la p-valeur est inférieure au seuil de 5 %, $\alpha = 0.05$. On doit donc rejeter l'hypothèse $\mathbb{H}_0$.

Exemple 29 (Présence de corrélation entre X et Y)**.** *Cas où $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$, avec $\beta_1 \neq 0$.*

n	10	20	50	100	200	500	1000	2000	5000	10000
F	3.33	14.5	55.3	91.6	237	382	1029	2122	5057	10296
p-valeur	0.11	0.001	$2 \cdot 10^{-9}$	10^{-15}	10^{-16}	10^{-16}	10^{-16}	10^{-16}	10^{-16}	10^{-16}

Note : pour $n = 10$, la p-valeur est supérieure au seuil de 5 %, $\alpha = 0.05$. On doit donc conserver l'hypothèse $\mathbb{H}_0$. Mais ce n'est pas anormal, l'échantillon étant très petit.

Extension du test. On peut étendre ce test au cadre où l'on veut vérifier si seulement une partie des paramètres est nulle :

- hypothèses :
  - $\mathbb{H}_0 : \beta_{p-q+1} = \cdots = \beta_p = 0$: les q derniers paramètres sous-jacents sont tous nuls,
  - $\mathbb{H}_1 : \exists j \in [p-q+1, p] : \beta_j \neq 0$: l'un au moins des paramètres sous-jacents est non nul ;
- statistique de test :

$$F := \frac{\frac{\sum_{i=1}^n (\hat{Y}_i^{1:p-q} - Y_i)^2 - \sum_{i=1}^n (\hat{Y}_i - Y_i)^2}{q}}{\frac{\sum_{i=1}^n (\hat{Y}_i - Y_i)^2}{n-p-1}}. \tag{6.10}$$

  où $\hat{Y}_i^{1:p-q}$ est l'estimation de Y_i dans le cas où l'on effectuerait la régression linéaire uniquement avec les $p-q$ premières variables. Dans le cas où $q = p$, $\hat{Y}_i^{1:p-q} = \bar{Y}$ et on retombe sur la formule donnée précédemment ;
- loi de F sous l'hypothèse $\mathbb{H}_0$: F suit une loi de Fisher $\mathcal{F}(q, n-p-1)$ à q degrés de liberté pour le numérateur et $n-p-1$ degrés de liberté pour le dénominateur ;
- test unilatéral.

Remarque 31. *N'importe quelle combinaison de paramètres peut être testée selon cette procédure. Il suffit au préalable de réordonner les paramètres pour placer à la fin les q paramètres à tester.*

Bibliographie

[Albright and Winston, 2020] Albright, S. C. and Winston, W. L. (2020). *Business analytics : Data analysis and decision making*. Cengage Learning, Inc.

[Dodge, 2008] Dodge, Y. (2008). *The concise encyclopedia of statistics*. Springer New York.

[Heumann and Shalabh, 2016] Heumann, C. and Shalabh, M. S. (2016). *Introduction to statistics and data analysis*. Springer.

[Hotelling, 1933] Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6) :417.

[Hubble, 1929] Hubble, E. (1929). A relation between distance and radial velocity among extra-galactic nebulae. *Proceedings of the national academy of sciences*, 15(3) :168–173.

[Hyndman and Fan, 1996] Hyndman, R. J. and Fan, Y. (1996). Sample quantiles in statistical packages. *The American Statistician*, 50(4) :361–365.

[James et al., 2023] James, G., Witten, D., Hastie, T., Tibshirani, R., and Taylor, J. (2023). *An introduction to statistical learning : With applications in python*. Springer Nature.

[Legendre, 1805] Legendre, A. M. (1805). *Nouvelles méthodes pour la détermination des orbites des comètes : avec un supplément contenant divers perfectionnemens de ces méthodes et leur application aux deux comètes de 1805*. Éditions stéréotypes.

[Pearson, 1896] Pearson, K. (1896). Mathematical contributions to the theory of evolution. — iii. regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character*, 187 :253–318.

[Pearson, 1901] Pearson, K. (1901). Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11) :559–572.

[Pearson, 1930] Pearson, K. (1930). *The Life, Letters and Labours of Francis Galton*, volume III. University Press, Cambridge.

[Student, 1908] Student (1908). The probable error of a mean. *Biometrika*, 6(1) :1–25.