

Chapitre 1

Le modèle de régression linéaire simple

1.1 Introduction

On dispose au point de départ des observations $(x_1, y_1), \dots, (x_n, y_n)$ de n couples $(X_1, Y_1), \dots, (X_n, Y_n)$ de variables aléatoires réelles.

Théorème de la variance totale : $\mathbb{E}[\text{var}(Y_i|X_i)] \leq \text{var}(Y_i)$. Interprétation : le phénomène aléatoire représenté par les X_i peut servir à *expliquer*, ou plutôt à *décrire*, celui représenté par les Y_i , puis éventuellement à le *prédire*.

On va donc chercher une fonction f telle que pour tout i , $f(X_i)$ "approche au mieux" Y_i .

Deux questions :

- Quel sens donner à "approcher au mieux" ?
- Quelle forme de fonction f choisir ?

Des réponses :

- Se donnant une *fonction de perte* (ou fonction de coût) l , comme par exemple la fonction de perte absolue définie par $l(y, y') = |y - y'|$ ou la fonction de perte quadratique définie par $l(y, y') = (y - y')^2$, on vise f minimisant $\mathbb{E} \left[\sum_{i=1}^n l(Y_i, f(X_i)) \right]$.
- Beaucoup de possibilités pour la forme de f , mais la plus simple et naturelle (valable dans de très nombreuses situations pratiques néanmoins) est la forme affine (ou linéaire par abus de langage).

1.2 De très nombreux exemples

1.2.1 Exemples historiques

Gauss et Legendre (1795/1809, 1805) : mécanique céleste et méthode des moindres carrés ordinaires.

Article de Francis Galton, *Regression towards mediocrity in hereditary stature*, Journal of the Anthropological Institute 15 : 246-63 (1886), à l'origine de l'anglicisme *régression*. Travaux antérieurs sur les diamètres de graines de pois de senteur et de leur descendance (1885).

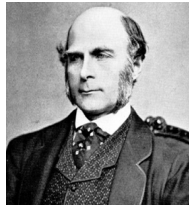
À l'origine de la régression



Adrien Marie Legendre
(1752-1833)



Carl Friedrich Gauss
(1777-1855)



Francis Galton
(1822-1911)

1.2.2 Prix et consommation de tabac

Données INSEE : prix relatif du tabac (indice 100 en 1970) et consommation de tabac (en grammes par adulte de 15 ans ou plus et par jour) en France de 1951 à 2009.

1.2.3 Consommation d'alcool et espérance de vie par pays

Données issues du rapport de l'OMS datant de février 2011 sur la consommation d'alcool (en L d'alcool pur par adulte de 15 ans ou plus) en projection pour l'année 2008, et l'espérance de vie en 2009 par pays.

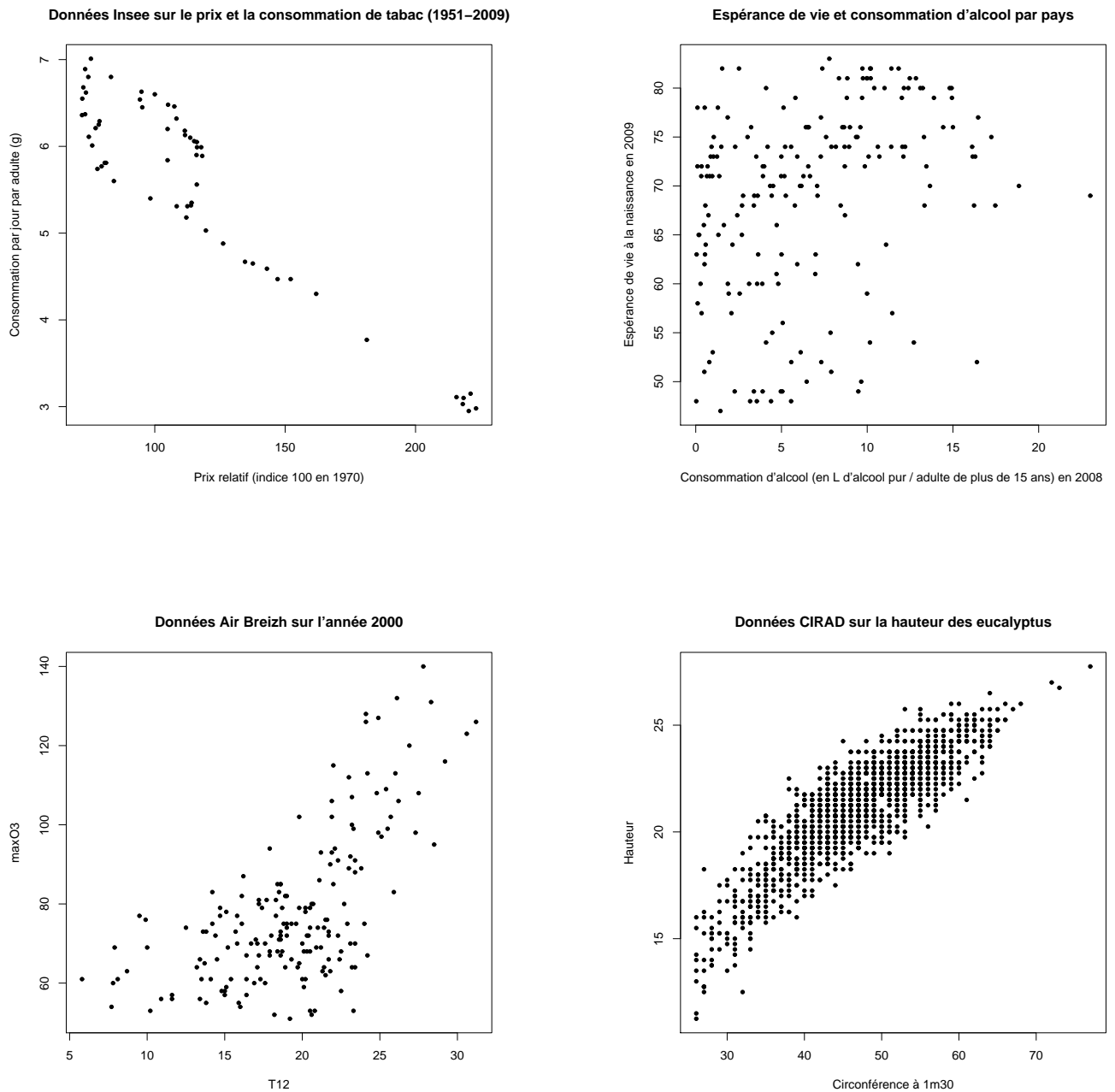
1.2.4 Qualité de l'air en Bretagne

Données fournies par Air Breizh : mesures du maximum journalier de la concentration en O_3 (en $\mu g/ml$) et de la température à 12h de 1994 à 2001.

1.2.5 Hauteur des eucalyptus

Données fournies par le Cirad (Centre de coopération internationale en recherche agronomique pour le développement) : mesures de la circonférence à 1 mètre 30 du sol et de la longueur du tronc d'eucalyptus d'une parcelle plantée.

FIGURE 1.1 – Représentations graphiques des nuages de points



1.3 Modèle de régression linéaire simple

1.3.1 Formulation analytique

Les Y_i et les X_i n'étant pas, dans l'immense majorité des cas, exactement liées de façon affine, on suppose qu'elles le sont "en moyenne" c'est à dire que $\mathbb{E}[Y_i] = \beta_0 + \beta_1 \mathbb{E}[X_i]$ pour tout $i = 1 \dots n$.

On introduit alors le modèle statistique suivant :

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad \text{pour } i = 1 \dots n,$$

où

- X_i est une variable aléatoire observée appelée *régresseur* ou *variable explicative*,
- Y_i est une variable aléatoire observée, appelée *variable à expliquer*,
- β_0 et β_1 sont des paramètres réels inconnus appelés *paramètres de régression* ou *coefficients de régression*,
- les ε_i sont des variables aléatoires indépendantes des X_i , non observées, appelées *erreurs* ou *bruits*, auxquelles on impose certaines conditions complémentaires.

Les conditions standards imposées aux ε_i sont les suivantes :

- (C₁) : $\mathbb{E}[\varepsilon_i] = 0$ pour tout $i = 1 \dots n$ (centrage),
- (C₂) : $\text{cov}(\varepsilon_i, \varepsilon_j) = 0$ pour tout $i \neq j$ (non corrélation),
- (C₃) : $\text{var}(\varepsilon_i) = \sigma^2$ (inconnue) pour tout $i = 1 \dots n$ (homoscédasticité).

Ce modèle est appelé *modèle de régression linéaire simple*.

Conditionnement sachant $X_i = x_i \Rightarrow$ on considère dans toute la suite du chapitre le modèle :

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \text{pour } i = 1 \dots n, \tag{1.1}$$

où

- x_i est déterministe, et il existe au moins un couple (i, j) tel que $x_i \neq x_j$,
- Y_i est une variable aléatoire observée,
- β_0 et β_1 sont des paramètres réels inconnus,
- les ε_i sont des variables aléatoires non observées vérifiant les conditions (C₁) à (C₃),

On a ainsi :

- $\mathbb{E}[Y_i] = \beta_0 + \beta_1 x_i$ pour tout $i = 1 \dots n$,
- $\text{cov}(Y_i, Y_j) = 0$ pour tout $i \neq j$ et $\text{var}(Y_i) = \sigma^2$ pour tout $i = 1 \dots n$.

Objectifs de statistique inférentielle :

- Estimation ponctuelle de (β_0, β_1) sur la base des observations y_1, \dots, y_n de Y_1, \dots, Y_n de façon à expliquer "au mieux" les variables Y_i en fonction des x_i , puis à prédire "au mieux" une valeur de Y_{n+1} à partir d'une nouvelle valeur x_{n+1} .
- Estimation ponctuelle de la variance σ^2 .
- Construction d'intervalles de confiance, de tests d'hypothèses et de critères permettant de juger de la qualité de l'explication ou de la prédiction : condition sur la loi des ε_i nécessaire.

1.3.2 Formulation vectorielle

$$Y = \beta_0 \mathbb{1} + \beta_1 x + \varepsilon = \mathbb{X}\beta + \varepsilon,$$

avec

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbb{1} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, \quad x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \quad \mathbb{X} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \quad \text{et} \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Les conditions (C_1) à (C_3) se traduisent par :

- $\mathbb{E}[\varepsilon] = 0$ et $\mathbb{E}[Y] = \beta_0 \mathbb{1} + \beta_1 x = \mathbb{X}\beta$,
- $\text{Var}(\varepsilon) = \text{Var}(Y) = \sigma^2 I_n$.

Représentation géométrique :

$\mathcal{E}(\mathbb{X})$ désigne le sous-espace vectoriel de \mathbb{R}^n engendré par les vecteurs $\mathbb{1}$ et x . On remarque que la projection orthogonale de Y sur le sous-espace vectoriel engendré par $\mathbb{1}$ est $\bar{Y}\mathbb{1}$.

1.4 Estimation (ponctuelle) et prédiction dans le cas général

1.4.1 Estimation ponctuelle des coefficients de régression

Rappel : on vise f affine minimisant $\mathbb{E} \left[\sum_{i=1}^n l(Y_i, f(x_i)) \right]$, c'est-à-dire un couple (β_0, β_1) minimisant $\mathbb{E} \left[\sum_{i=1}^n l(Y_i, \beta_0 + \beta_1 x_i) \right]$. La loi des Y_i étant inconnue, on applique le même principe que celui de la méthode des moments. Les paramètres (β_0, β_1) peuvent alors être estimés par $\hat{\beta}_0$ et $\hat{\beta}_1$ tels que $(\hat{\beta}_0, \hat{\beta}_1) \in \argmin_{(\beta_0, \beta_1) \in \mathbb{R}^2} \sum_{i=1}^n l(Y_i, \beta_0 + \beta_1 x_i)$.

Définition 1. On appelle droite de régression l'ensemble $\mathcal{D} = \{(x, y), y = \hat{\beta}_0 + \hat{\beta}_1 x\}$.

Représentation pour la perte absolue et pour la perte quadratique (moins robuste).

Choix usuel de la perte quadratique \Rightarrow moindres carrés ordinaires.

Moindres carrés ordinaires

Définition 2. On appelle estimateurs des moindres carrés ordinaires (MCO) de β_0 et β_1 les estimateurs $\hat{\beta}_0$ et $\hat{\beta}_1$ tels que $(\hat{\beta}_0, \hat{\beta}_1) \in \argmin_{(\beta_0, \beta_1) \in \mathbb{R}^2} \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 x_i)^2$.

Calcul des estimateurs des MCO :

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

Preuve. On note $L(\beta_0, \beta_1) = \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 x_i)^2$. La fonction L est une fonction de deux variables réelles. Ses points critiques sont obtenus par la résolution du système :

$$\begin{cases} \frac{\partial L}{\partial \beta_0}(\beta_0, \beta_1) = 0 \\ \frac{\partial L}{\partial \beta_1}(\beta_0, \beta_1) = 0. \end{cases}$$

On obtient le point critique $(\hat{\beta}_0, \hat{\beta}_1) = \left(\bar{Y} - \hat{\beta}_1 \bar{x}, \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \right)$, et on vérifie que ce point critique correspond à un minimum local à l'aide des notations de Monge. $p = \frac{\partial^2 L}{\partial \beta_0^2}(\hat{\beta}_0, \hat{\beta}_1) = 2n$, $q = \frac{\partial^2 L}{\partial \beta_0 \partial \beta_1}(\hat{\beta}_0, \hat{\beta}_1) = 2 \sum_{i=1}^n x_i$, $r = \frac{\partial^2 L}{\partial \beta_1^2}(\hat{\beta}_0, \hat{\beta}_1) = 2 \sum_{i=1}^n x_i^2$, donc $pr - q^2 = 4(n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2)$. L'inégalité de Cauchy-Schwarz donne $(\sum_{i=1}^n x_i)^2 \leq \sum_{i=1}^n 1^2 \cdot \sum_{i=1}^n x_i^2 \leq n \sum_{i=1}^n x_i^2$, avec égalité lorsque (x_1, \dots, x_n) est colinéaire à $(1, \dots, 1)$, c'est-à-dire lorsque tous les x_i sont égaux (ce qui n'est pas possible par hypothèse). On a donc $pr - q^2 > 0$ et $(\hat{\beta}_0, \hat{\beta}_1)$ est bien un minimum local.

Remarque : la droite de régression des MCO calculée passe par le centre de gravité (\bar{x}, \bar{y}) du nuage de points.

Retour sur les exemples (Figure 1.2) : tracé des droites de régression des MCO calculées sur les observations.

Propriétés des estimateurs des MCO

Théorème 1. Les estimateurs des MCO vérifient les propriétés suivantes.

1. $\hat{\beta}_0$ et $\hat{\beta}_1$ sont des estimateurs linéaires en $(Y_i)_{i=1..n}$, sans biais de β_0 et β_1 respectivement ;
2. $var(\hat{\beta}_0) = \frac{\sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2} \sigma^2$;
3. $var(\hat{\beta}_1) = \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \sigma^2$;
4. $cov(\hat{\beta}_0, \hat{\beta}_1) = \frac{-\bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \sigma^2$.

Preuve.

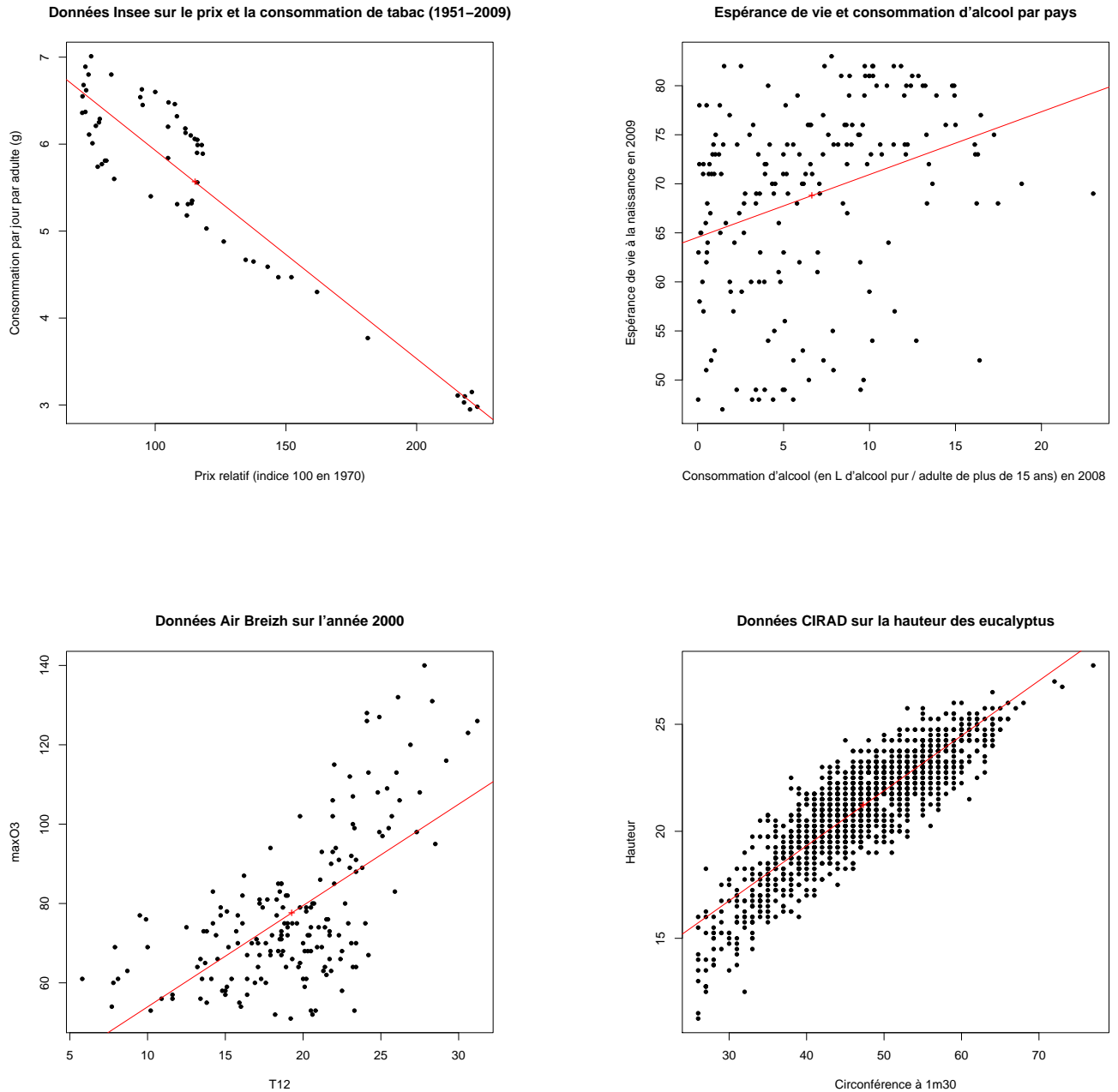
$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x}) Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = \sum_{i=1}^n w_i Y_i, \text{ avec } w_i = \frac{x_i - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

De même, $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x} = \sum_{i=1}^n \left(\frac{1}{n} - \bar{x} w_i \right) Y_i$, donc les estimateurs des MCO sont bien des estimateurs linéaires.

Pour la suite, on note au préalable que :

1. $\sum_{i=1}^n w_i = \sum_{i=1}^n \frac{x_i - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} = 0$,

FIGURE 1.2 – Représentations graphiques des droites de régression



$$2. \sum_{i=1}^n w_i x_i = \sum_{i=1}^n \frac{(x_i - \bar{x})x_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = \sum_{i=1}^n \frac{(x_i - \bar{x})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = 1,$$

$$3. \sum_{i=1}^n w_i^2 = \sum_{i=1}^n \frac{(x_i - \bar{x})^2}{(\sum_{i=1}^n (x_i - \bar{x})^2)^2} = \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

Alors :

$$\mathbb{E}[\hat{\beta}_1] = \mathbb{E}[\sum_{i=1}^n w_i Y_i] = \sum_{i=1}^n w_i (\beta_0 + \beta_1 x_i) = \beta_1 \sum_{i=1}^n w_i x_i = \beta_1, \text{ et}$$

$$\mathbb{E}[\hat{\beta}_0] = \mathbb{E}[\bar{Y} - \hat{\beta}_1 \bar{x}] = \mathbb{E}[\beta_0 + \beta_1 \bar{x} + \bar{\varepsilon}] - \beta_1 \bar{x} = \beta_0.$$

En outre,

$$\text{var}(\hat{\beta}_0) = \sum_{i=1}^n \left(\frac{1}{n^2} - 2 \frac{\bar{x} w_i}{n} + \bar{x}^2 w_i^2 \right) \sigma^2 = \left(\frac{1}{n} - 0 + \bar{x}^2 \sum_{i=1}^n w_i^2 \right) \sigma^2 = \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) \sigma^2 = \frac{\sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2} \sigma^2,$$

$$\text{cov}(\hat{\beta}_0, \hat{\beta}_1) = \text{cov}(\sum_{i=1}^n (\frac{1}{n} - \bar{x} w_i) Y_i, \sum_{i=1}^n w_i Y_i) = \sum_{i=1}^n \sum_{j=1}^n (\frac{1}{n} - \bar{x} w_i) w_j \text{cov}(Y_i, Y_j) = \sum_{i=1}^n (\frac{1}{n} - \bar{x} w_i) w_i \sigma^2 = - \frac{\bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \sigma^2,$$

$$\text{et } \text{var}(\hat{\beta}_1) = \sum_{i=1}^n w_i^2 \sigma^2 = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

Interprétation.

Théorème 2 (Gauss Markov). Parmi les estimateurs linéaires sans biais de β_0 et β_1 respectivement linéaires en $(Y_i)_{i=1, \dots, n}$, $\hat{\beta}_0$ et $\hat{\beta}_1$ sont de variance minimale.

La preuve sera vue dans le chapitre suivant.

1.4.2 Estimation ponctuelle de la variance, valeurs ajustées et résidus

Pourquoi estimer la variance σ^2 ?

$$\text{Var}(\hat{\beta}) = \sigma^2 \begin{pmatrix} \frac{\sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2} & \frac{-\bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ \frac{-\bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} & \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \end{pmatrix}$$

Problème : les ε_i ne sont pas observés donc pour pouvoir estimer la variance σ^2 , on introduit les résidus.

Définition 3. On appelle résidus les quantités $\hat{\varepsilon}_i = Y_i - \hat{Y}_i$, pour $i = 1 \dots n$, où $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. Les variables \hat{Y}_i sont appelées valeurs ajustées.

Les valeurs ajustées donnent une estimation de $\mathbb{E}[Y_i]$.

Propriétés des résidus : les $\hat{\varepsilon}_i$ sont des variables aléatoires observées, centrées, de somme nulle i.e. $\sum_{i=1}^n \hat{\varepsilon}_i = 0$ (donc non indépendantes), corrélées négativement et hétéroscédastiques.

Une idée naturelle : estimer la variance σ^2 par $\frac{1}{n} \sum_{i=1}^n \hat{\varepsilon}_i^2$, mais c'est un estimateur biaisé, d'espérance égale à $\frac{n-2}{n} \sigma^2$. On choisit donc plutôt :

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n \hat{\varepsilon}_i^2,$$

qui est un estimateur sans biais de σ^2 .

Preuve. On a

$$\hat{\varepsilon}_i = \beta_0 + \beta_1 x_i + \varepsilon_i - \hat{\beta}_0 - \hat{\beta}_1 x_i = \bar{Y} - \beta_1 \bar{x} - \bar{\varepsilon} + \beta_1 x_i + \varepsilon_i - \bar{Y} + \hat{\beta}_1 \bar{x} - \hat{\beta}_1 x_i = (\varepsilon_i - \bar{\varepsilon}) + (\beta_1 - \hat{\beta}_1)(x_i - \bar{x}). \text{ D'où}$$

$$\hat{\varepsilon}_i^2 = (\varepsilon_i - \bar{\varepsilon})^2 + (\beta_1 - \hat{\beta}_1)^2 (x_i - \bar{x})^2 + 2(\beta_1 - \hat{\beta}_1)(x_i - \bar{x})(\varepsilon_i - \bar{\varepsilon}).$$

$$\text{Or } \hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x}) Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\beta_0 \sum_{i=1}^n (x_i - \bar{x}) + \beta_1 \sum_{i=1}^n (x_i - \bar{x}) x_i + \sum_{i=1}^n (x_i - \bar{x}) \varepsilon_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = \beta_1 + \frac{\sum_{i=1}^n (x_i - \bar{x}) (\varepsilon_i - \bar{\varepsilon})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \text{ d'où}$$

$$\begin{aligned} \sum_{i=1}^n \hat{\varepsilon}_i^2 &= \sum_{i=1}^n (\varepsilon_i - \bar{\varepsilon})^2 + \sum_{i=1}^n (\beta_1 - \hat{\beta}_1)^2 (x_i - \bar{x})^2 - 2 \sum_{i=1}^n (\beta_1 - \hat{\beta}_1)^2 (x_i - \bar{x})^2 \\ &= \sum_{i=1}^n (\varepsilon_i - \bar{\varepsilon})^2 - \sum_{i=1}^n (\beta_1 - \hat{\beta}_1)^2 (x_i - \bar{x})^2, \text{ et } \mathbb{E} \left[\sum_{i=1}^n \hat{\varepsilon}_i^2 \right] = (n-1)\sigma^2 - \sigma^2 = (n-2)\sigma^2. \end{aligned}$$

1.4.3 Le coefficient de détermination R^2

Les \hat{Y}_i étant tels que $\sum_{i=1}^n \hat{\varepsilon}_i^2$ soit minimale, puisque $\sum_{i=1}^n \hat{Y}_i = \sum_{i=1}^n Y_i$, on a le théorème suivant :

Théorème 3. La somme des carrés totale (SCT) est égale à la somme des carrés expliquée (SCE) plus la somme des carrés résiduelle (SCR) :

$$\underbrace{\sum_{i=1}^n (Y_i - \bar{Y})^2}_{\text{SCT}} = \underbrace{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}_{\text{SCE}} + \underbrace{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}_{\text{SCR}}.$$

La preuve sera vue dans le paragraphe suivant.

La décomposition ci-dessus s'appelle l'équation d'analyse de la variance.

Interprétations.

Définition 4. Le coefficient de détermination R^2 est la fraction de la variabilité totale expliquée par la régression. Plus précisément,

$$R^2 = \frac{\text{SCE}}{\text{SCT}} = \left(\text{corr}(Y_i, \hat{Y}_i) \right)^2.$$

On a $R^2 \in [0, 1]$. Interprétation des cas limites.

1.4.4 Prédiction

A partir d'une nouvelle valeur explicative x_{n+1} , on souhaite prédire une nouvelle observation d'une variable $Y_{n+1} = \beta_0 + \beta_1 x_{n+1} + \varepsilon_{n+1}$, avec $\mathbb{E}[\varepsilon_{n+1}] = 0$, $\text{var}(\varepsilon_{n+1}) = \sigma^2$ et $\text{cov}(\varepsilon_{n+1}, \varepsilon_i) = 0$ pour tout $i = 1 \dots n$ i.e. Y_{n+1} non corrélée avec les $(Y_i)_{i=1 \dots n}$ utilisées pour estimer les coefficients de régression.

Pour cela, on introduit $\hat{Y}_{n+1}^p = \hat{\beta}_0 + \hat{\beta}_1 x_{n+1}$.

L'erreur de prédiction est définie par $\hat{\varepsilon}_{n+1}^p = Y_{n+1} - \hat{Y}_{n+1}^p$ (inconnue), dont la variance est égale à

$$\text{var}(\hat{\varepsilon}_{n+1}^p) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right).$$

Preuve laissée en exercice.

Remarque : attention à la prédiction lorsque la valeur x_{n+1} est éloignée de \bar{x} ...

1.4.5 Ecritures matricielles et interprétations géométriques

On rappelle ici la formulation vectorielle du modèle de régression linéaire simple :

$$Y = \beta_0 \mathbb{1} + \beta_1 x + \varepsilon = \mathbb{X}\beta + \varepsilon,$$

avec

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbb{1} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, \quad x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \quad \mathbb{X} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \quad \text{et} \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Estimateurs des moindres carrés ordinaires, valeurs ajustées, résidus et projection orthogonale

On commence par remarquer que $\sum_{i=1}^n (Y_i - \beta_0 - \beta_1 x_i)^2 = (Y - \mathbb{X}\beta)'(Y - \mathbb{X}\beta) = \|Y - \mathbb{X}\beta\|^2$, où $\|\cdot\|$ est la norme euclidienne de \mathbb{R}^n .

Si l'on note

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix},$$

alors par définition, $\hat{\beta} \in \text{argmin}_{\beta \in \mathbb{R}^2} (Y - \mathbb{X}\beta)'(Y - \mathbb{X}\beta)$.

On cherche donc le point critique de la fonction $L : \beta \mapsto (Y - \mathbb{X}\beta)'(Y - \mathbb{X}\beta) = Y'Y - Y'\mathbb{X}\beta - \beta'\mathbb{X}'Y + \beta'\mathbb{X}'\mathbb{X}\beta = Y'Y - 2\beta'\mathbb{X}'Y + \beta'\mathbb{X}'\mathbb{X}\beta$.

$\nabla L(\beta) = 0$ si et seulement si $\mathbb{X}'\mathbb{X}\beta = \mathbb{X}'Y$ et puisque $\mathbb{X}'\mathbb{X}$ est inversible (hypothèse sur les x_i),

$$\beta = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'Y.$$

Ce point critique correspond bien à un minimum puisque la matrice hessienne de L en $(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'Y$ vaut $2\mathbb{X}'\mathbb{X}$ qui est définie positive.

On a alors

$$\hat{\beta} = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'Y,$$

et

$$\hat{Y} = \mathbb{X}\hat{\beta} = \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'Y,$$

où la matrice $\Pi_{\mathbb{X}} = \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'$ est appelée la matrice "chapeau" (hat matrix) (puisqu'elle transforme Y en \hat{Y} !).

En fait, $\Pi_{\mathbb{X}}$ est la matrice de projection orthogonale sur le sous-espace vectoriel de \mathbb{R}^n $\mathcal{E}(\mathbb{X})$ engendré par les vecteurs colonnes de la matrice \mathbb{X} à savoir $\mathbb{1}$ et x .

On a en effet les propriétés suivantes :

- $\Pi_X' = \Pi_X$ (symétrie),
- $\Pi_X \Pi_X = X(X'X)^{-1}X'X(X'X)^{-1}X' = \Pi_X$ (idempotence),
- $\Pi_X X\beta = X\beta$ pour tout $\beta \in \mathbb{R}^2$.

On retrouve ainsi le résultat :

$\hat{\beta} \in \operatorname{argmin}_{\beta \in \mathbb{R}^2} \|Y - X\beta\|^2$ si et seulement si $X\hat{\beta}$ est la projection orthogonale de Y sur $\mathcal{E}(X)$.

En utilisant ces notations, le vecteur des résidus vérifie $\hat{\varepsilon} = (\hat{\varepsilon}_1, \dots, \hat{\varepsilon}_n)' = Y - \hat{Y} = (I_n - \Pi_X)Y$, c'est-à-dire que $\hat{\varepsilon}$ est le projeté orthogonal de Y sur $\mathcal{E}(X)^\perp$.

On peut ainsi montrer facilement que

- $X'\hat{\varepsilon} = (0, 0)'$ i.e. $\sum_{i=1}^n \hat{\varepsilon}_i = 0$ et $\sum_{i=1}^n x_i \hat{\varepsilon}_i = 0$,
- $\mathbb{E}[\hat{\varepsilon}] = 0$,
- $\operatorname{Var}(\hat{\varepsilon}) = \sigma^2(I_n - \Pi_X)$ (corrélation négative et hétéroscédasticité).

Représentation géométrique.

Coefficient de détermination R^2 et théorème de Pythagore

On rappelle que la somme des carrés totale, la somme des carrés estimée et la somme des carrés résiduelle désignent respectivement $\sum_{i=1}^n (Y_i - \bar{Y})^2$, $\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$ et $\sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n \hat{\varepsilon}_i^2$.

On a, en notant $\Pi_{\mathbb{1}}$ la matrice de projection orthogonale sur le sous-espace vectoriel de \mathbb{R}^n engendré par $\mathbb{1}$,

$$\begin{aligned}
 \sum_{i=1}^n (Y_i - \bar{Y})^2 &= (Y - \Pi_{\mathbb{1}}Y)'(Y - \Pi_{\mathbb{1}}Y) \\
 &= Y'(I_n - \Pi_{\mathbb{1}})'(I_n - \Pi_{\mathbb{1}})Y \\
 &= Y'(I_n - \Pi_X + \Pi_X - \Pi_{\mathbb{1}})'(I_n - \Pi_{\mathbb{1}})Y \\
 &= Y'(I_n - \Pi_X)(I_n - \Pi_{\mathbb{1}})Y + Y'(\Pi_X - \Pi_{\mathbb{1}})'(I_n - \Pi_{\mathbb{1}})Y \\
 &= Y'(I_n - \Pi_X)Y + Y'(\Pi_X - \Pi_{\mathbb{1}})(I_n - \Pi_{\mathbb{1}})Y \\
 &= Y'(I_n - \Pi_X)Y + Y'(\Pi_X - \Pi_{\mathbb{1}} - \Pi_{\mathbb{1}} + \Pi_{\mathbb{1}})Y \\
 &= Y'(I_n - \Pi_X)Y + Y'(\Pi_X - \Pi_{\mathbb{1}} - \Pi_{\mathbb{1}} + \Pi_{\mathbb{1}})Y \\
 &= Y'(I_n - \Pi_X)Y + Y'(\Pi_X - \Pi_{\mathbb{1}})Y \\
 &= \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2.
 \end{aligned}$$

On a donc bien le résultat suivant : SCT=SCR+SCE, et on a en fait redémontré ici le théorème de Pythagore dans ce cas c'est-à-dire :

$$\|Y - \Pi_{\mathbb{1}}Y\|^2 = \|(I_n - \Pi_X)(Y - \Pi_{\mathbb{1}}Y)\|^2 + \|\Pi_X(Y - \Pi_{\mathbb{1}}Y)\|^2 = \|(I_n - \Pi_X)Y\|^2 + \|\Pi_X Y - \Pi_{\mathbb{1}}Y\|^2.$$

Représentation graphique.

On a alors

$$R^2 = \frac{\|\Pi_X Y - \Pi_{\mathbb{1}}Y\|^2}{\|Y - \Pi_{\mathbb{1}}Y\|^2} = \cos^2 \theta,$$

où θ est l'angle formé par les vecteurs $Y - \Pi_1 Y$ et $\Pi_X Y - \Pi_1 Y = \hat{Y} - \Pi_1 Y$.

Interprétation géométrique des cas $R^2 = 1$ et $R^2 = 0$.

1.5 Inférence sous hypothèse gaussienne

On souhaite maintenant pouvoir construire des intervalles de confiance pour les coefficients de régression, la variance du modèle, puis des tests d'hypothèses sur les coefficients de régression. On se place pour cela dans un cas "simple" (mais réaliste dans de très nombreux cas pratiques) où les bruits ε_i - comme les Y_i - sont supposés suivre une loi gaussienne.

On pose la condition (C₄) : le vecteur ε suit une loi gaussienne.

Les hypothèses (C₁) à (C₄) réunies se résument ainsi à :

$$\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n).$$

On a alors

$$Y \sim \mathcal{N}(X\beta, \sigma^2 I_n).$$

A noter que les variables ε_i sont maintenant supposées i.i.d. (mais les Y_i seulement indépendantes), et que le modèle devient paramétrique, dominé par la mesure de Lebesgue sur \mathbb{R}^n .

1.5.1 Estimateurs du maximum de vraisemblance

La vraisemblance du modèle s'écrit

$$L(\beta_0, \beta_1, \sigma^2, Y_1, \dots, Y_n) = \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n \exp \left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 x_i)^2 \right),$$

d'où

$$\ln L(\beta_0, \beta_1, \sigma^2, Y_1, \dots, Y_n) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 x_i)^2.$$

A σ^2 fixé, $\ln L(\beta_0, \beta_1, \sigma^2, Y_1, \dots, Y_n)$ est maximale si $\sum_{i=1}^n (Y_i - \beta_0 - \beta_1 x_i)^2$ est minimale. Les estimateurs du maximum de vraisemblance de β_0 et β_1 sont donc égaux aux estimateurs des moindres carrés ordinaires.

Par ailleurs, $\frac{\partial \ln L}{\partial \sigma^2}(\hat{\beta}_0, \hat{\beta}_1, \tilde{\sigma}^2, Y_1, \dots, Y_n) = 0$ si et seulement si $\tilde{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \frac{1}{n} \sum_{i=1}^n \hat{\varepsilon}_i^2$. L'estimateur du maximum de vraisemblance $\tilde{\sigma}^2$ de σ^2 est biaisé : on lui préférera en général $\hat{\sigma}^2$.

Théorème 4. Les estimateurs des MCO vérifient les propriétés suivantes :

- $\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2 V)$, où $V = \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \begin{pmatrix} \frac{\sum_{i=1}^n x_i^2}{n} & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix} = (X'X)^{-1}$,
- $\frac{n-2}{\sigma^2} \hat{\sigma}^2 \sim \chi^2(n-2)$,

- $\frac{1}{\sigma^2}(\hat{\beta} - \beta)'V^{-1}(\hat{\beta} - \beta) \sim \chi^2(2),$
- $\hat{\beta}$ et $\hat{\sigma}^2$ sont indépendants.

La preuve sera vue dans les chapitres suivants.

On en déduit en particulier les propriétés suivantes.

- $\frac{\hat{\beta}_0 - \beta_0}{\sqrt{\frac{\hat{\sigma}^2 \sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2}}} \sim \mathcal{T}(n-2),$
- $\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}} \sim \mathcal{T}(n-2),$
- $\frac{1}{2\sigma^2}(\hat{\beta} - \beta)'V^{-1}(\hat{\beta} - \beta) \sim \mathcal{F}(2, n-2).$

1.5.2 Intervalles et régions de confiance pour les coefficients de régression

Théorème 5. Soit $\alpha \in]0, 1[$. On note $t_{n-2}(u)$ et $f_{2, n-2}(u)$ les u -quantiles respectifs des lois $\mathcal{T}(n-2)$ et $\mathcal{F}(2, n-2)$.

- Un intervalle de confiance de niveau de confiance $(1 - \alpha)$ pour β_0 est donné par

$$\left[\hat{\beta}_0 - t_{n-2}(1 - \alpha/2) \sqrt{\frac{\hat{\sigma}^2 \sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2}}, \hat{\beta}_0 + t_{n-2}(1 - \alpha/2) \sqrt{\frac{\hat{\sigma}^2 \sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2}} \right].$$

- Un intervalle de confiance de niveau de confiance $(1 - \alpha)$ pour β_1 est donné par

$$\left[\hat{\beta}_1 - t_{n-2}(1 - \alpha/2) \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}, \hat{\beta}_1 + t_{n-2}(1 - \alpha/2) \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} \right].$$

- Une région de confiance simultanée pour (β_0, β_1) de niveau de confiance $(1 - \alpha)$ est donnée par

$$\left\{ (\beta_0, \beta_1), \frac{1}{2\sigma^2} \left(n(\hat{\beta}_0 - \beta_0)^2 + 2n\bar{x}(\hat{\beta}_0 - \beta_0)(\hat{\beta}_1 - \beta_1) + \sum_{i=1}^n x_i^2(\hat{\beta}_1 - \beta_1)^2 \right) \leq f_{2, n-2}(1 - \alpha) \right\}.$$

Remarque : la région de confiance simultanée pour (β_0, β_1) est une ellipse. On parlera parfois d'ellipse de confiance.

1.5.3 Tests d'hypothèses sur β_0

On souhaite tester l'hypothèse nulle $(H_0) : \beta_0 = b$ contre l'alternative $(H_1) : \beta_0 \neq b$ (test bilatère) ou $\beta_0 < b$ ou $\beta_0 > b$ (tests unilatères).

On utilise alors comme statistique de test $T_0(Y) = \frac{\hat{\beta}_0 - b}{\sqrt{\frac{\hat{\sigma}^2 \sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2}}}$ qui suit sous l'hypothèse (H_0)

la loi $\mathcal{T}(n-2)$.

On peut ensuite prendre, pour un niveau $\alpha \in]0, 1[$, comme région de rejet ou région critique dans le cas d'un test bilatère :

$$R_{(H_0)} = \{y, |T_0(y)| \geq t_{n-2}(1 - \alpha/2)\}.$$

Le test de significativité $(H_0) : \beta_0 = 0$ contre $(H_1) : \beta_0 \neq 0$ permet de tester l'utilité de la constante β_0 dans le modèle.

1.5.4 Tests d'hypothèses sur β_1

On souhaite tester l'hypothèse nulle $(H_0) : \beta_1 = b$ contre l'alternative $(H_1) : \beta_1 \neq b$ (test bilatère) ou $\beta_1 < b$ ou $\beta_1 > b$ (tests unilatères).

On utilise comme statistique de test $T_1(Y) = \frac{\hat{\beta}_1 - b}{\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}}$ qui suit sous l'hypothèse (H_0) la loi

$\mathcal{T}(n-2)$.

On peut prendre, pour un niveau $\alpha \in]0, 1[$, comme région de rejet ou région critique dans le cas d'un test bilatère :

$$R_{(H_0)} = \{y, |T_1(y)| \geq t_{n-2}(1 - \alpha/2)\}.$$

Le test de significativité $(H_0) : \beta_1 = 0$ contre $(H_1) : \beta_1 \neq 0$ permet de tester l'utilité du modèle de régression. Dans ce cas, on peut montrer que $(T_1(Y))^2 = (n-2) \frac{R^2}{1-R^2}$. On retrouve ainsi l'intérêt de l'introduction du R^2 (et l'interprétation des cas limites pour les valeurs de R^2).

Remarque : dans les sorties de logiciels, les p -valeurs des tests de significativité des variables explicatives sont données, avec un "indice" de significativité des variables explicatives.

1.5.5 Intervalles de confiance et tests d'hypothèses sur la variance

Théorème 6. Soit $\alpha \in]0, 1[$. On note $c_{n-2}(u)$ le u -quantile de la loi $\chi^2(n-2)$.

Un intervalle de confiance de niveau de confiance $(1 - \alpha)$ pour σ^2 est donné par $\left[\frac{(n-2)\hat{\sigma}^2}{c_{n-2}(1-\alpha/2)}, \frac{(n-2)\hat{\sigma}^2}{c_{n-2}(\alpha/2)} \right]$.

Si l'on souhaite tester l'hypothèse nulle $(H_0) : \sigma^2 = s^2$ contre l'alternative $(H_1) : \sigma^2 \neq s^2$ (test bilatère) ou $\sigma^2 < s^2$ ou $\sigma^2 > s^2$ (tests unilatères), on utilise comme statistique de test $S^2(Y) = \frac{n-2}{s^2} \hat{\sigma}^2$ qui suit sous l'hypothèse (H_0) la loi $\chi^2(n-2)$.

1.5.6 Intervalles de prédiction

On peut utiliser les résultats précédents pour construire des intervalles de confiance pour $\mathbb{E}[Y_{n+1}] = \beta_0 + \beta_1 x_{n+1}$, mais il est généralement plus intéressant de trouver un intervalle \hat{I} dit de prédiction tel que $P(Y_{n+1} \in \hat{I}) \geq 1 - \alpha$, pour $\alpha \in]0, 1[$.

On montre pour cela que

$$Y_{n+1} - \hat{Y}_{n+1}^p \sim \mathcal{N}\left(0, \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)\right).$$

Puisque $Y_{n+1} - \hat{Y}_{n+1}^p$ et $\hat{\sigma}^2$ sont indépendantes, le théorème suivant est vérifié.

Théorème 7. Un intervalle de prédiction de niveau $(1 - \alpha)$ est donné par

$$\hat{I} = \left[\hat{Y}_{n+1}^p - t_{n-2}(1 - \alpha/2) \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}, \hat{Y}_{n+1}^p + t_{n-2}(1 - \alpha/2) \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} \right].$$

1.6 Exercices

Exercice 1 : Questions de cours - QCM

On dispose d'observations y_1, \dots, y_n de variables aléatoires telles que $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ pour tout $i = 1 \dots n$, où les ε_i sont des variables aléatoires vérifiant les conditions standards d'un modèle de régression linéaire simple.

1. Les variables Y_i sont-elles supposées indépendantes et identiquement distribuées ?
 - a) Oui.
 - b) Non.
 - c) Pas toujours.
2. La droite de régression calculée sur les observations passe par le point (\bar{x}, \bar{y}) . Est-ce toujours le cas ?
 - a) Oui.
 - b) Non.
3. Les estimateurs des moindres carrés ordinaires des coefficients de régression sont-ils indépendants ?
 - a) Oui.
 - b) Non.
 - c) Pas toujours.
4. Est-il possible de trouver des estimateurs des coefficients de régression de plus faible variance que celle des estimateurs des moindres carrés ordinaires ?
 - a) Oui.
 - b) Non.
 - c) Peut-être.
5. Les estimateurs des moindres carrés ordinaires des coefficients de régression sont-ils égaux aux estimateurs du maximum de vraisemblance sous hypothèse gaussienne ?
 - a) Oui.
 - b) Non.
 - c) Pas toujours.
6. L'estimateur du maximum de vraisemblance de la variance des ε_i sous hypothèse gaussienne est-il sans biais ?
 - a) Oui.
 - b) Non.
7. Les résidus sont-ils indépendants ?
 - a) Oui.
 - b) Non.
 - c) Pas toujours.
8. On dispose d'une nouvelle valeur x_{n+1} et on note \hat{Y}_{n+1}^p la valeur prédite correspondante,

$\hat{\varepsilon}_{n+1}^p$ l'erreur de prédiction correspondante. La variance de $\hat{\varepsilon}_{n+1}^p$ est minimale lorsque :

- a) $x_{n+1} = 0$.
- b) $x_{n+1} = \bar{x}$.
- c) La variance ne dépend pas de la valeur de x_{n+1} .

9. Le coefficient de détermination R^2 calculé sur les observations vaut 1. Les points (x_i, y_i) sont-ils alignés ?

- a) Oui.
- b) Non.
- c) Pas nécessairement.

10. Peut-on utiliser un test d'adéquation du khi-deux pour tester la normalité des variables ε_i et Y_i ?

- a) Oui.
- b) Non.

Exercice 2 : Les graines de pois de senteur de Galton (1885)

Dans ses premiers travaux sur l'hérédité, Francis Galton a cherché à mettre en évidence un lien entre le diamètre de graines de pois de senteur et le diamètre moyen de leur descendance. Il a mesuré pour cela le diamètre de 7 graines, et le diamètre moyen de leur descendance. Les résultats qu'il a obtenus sont les suivants :

Observation	Diamètre des graines mères (en 1/100 de pouce)	Diamètre moyen de la descendance (en 1/100 de pouce)
i	x_i	y_i
1	21	17.5
2	20	17.3
3	19	16
4	18	16.3
5	17	15.6
6	16	16
7	15	15.3

Déterminer les valeurs des estimateurs des moindres carrés ordinaires des coefficients de régression et des résidus de la régression linéaire simple correspondante calculés sur ces observations. Représenter les observations, la droite de régression et les résidus calculés sur un graphique.

Exercice 3 : Données Insee sur la consommation de tabac

On dispose des données Insee sur le prix relatif (indice 100 à 1970) et la consommation de tabac (en grammes par adulte de plus de 15 ans et par jour) de 1951 à 2009. Les prix relatifs étant notés x_i pour $i = 1 \dots 59$ et les consommations correspondantes y_i pour $i = 1 \dots 59$, on a les résultats numériques suivants :

$$\sum_{i=1}^{59} x_i = 6806.5 \quad \sum_{i=1}^{59} x_i^2 = 891776 \quad \sum_{i=1}^{59} x_i y_i = 35295.02 \quad \sum_{i=1}^{59} y_i = 328.07 \quad \sum_{i=1}^{59} y_i^2 = 1895.363.$$

Déterminer les valeurs des estimateurs des moindres carrés ordinaires des coefficients de régression, de l'estimateur sans biais de la variance, et du coefficient de détermination de la régression linéaire simple correspondante calculés sur ces observations.

Exercice 4 : Estimateurs linéaires sans biais de variance minimale

On considère le modèle de régression linéaire simple suivant :

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \dots, n,$$

où les bruits ε_i sont des variables aléatoires telles que $\mathbb{E}[\varepsilon_i] = 0$ et $\text{cov}(\varepsilon_i, \varepsilon_j) = \sigma^2 \delta_{i,j}$.

On cherche des estimateurs β_0^* et β_1^* de β_0 et β_1 qui possèdent les propriétés suivantes :

- β_0^* et β_1^* sont des fonctions linéaires de Y_1, \dots, Y_n .
- β_0^* et β_1^* sont des estimateurs sans biais.
- β_0^* et β_1^* sont de variance minimale parmi les estimateurs linéaires sans biais.

1. Déterminer ces estimateurs et montrer qu'ils sont égaux aux estimateurs des moindres carrés ordinaires.
2. Quel résultat a-t-on retrouvé ici ?

Exercice 5 : Consommation de confiseries

Les données suivantes, publiées par Chicago Tribune en 1993, montrent la consommation de confiseries en million de livres (variable Y) et la population en millions d'habitants (variable X) dans 17 pays en 1991. On note y_i la consommation et x_i la population du i ème pays, $i = 1, \dots, 17$.

$$\sum_{i=1}^{17} x_i = 751.8 \quad \sum_{i=1}^{17} x_i^2 = 97913.92 \quad \sum_{i=1}^{17} y_i = 13683.8 \quad \sum_{i=1}^{17} y_i^2 = 36404096.44 \quad \sum_{i=1}^{17} x_i y_i = 1798166.66$$

Pays	Consommation	Population
i	y_i	x_i
Australia	327.4	17.3
Austria	179.5	7.7
Belgium	279.4	10.4
Denmark	139.1	5.1
Finland	92.5	5.0
France	926.7	56.9
Germany	2186.3	79.7
Ireland	96.8	3.5
Italy	523.9	57.8
Japan	935.9	124.0
Netherland	444.2	15.1
Norway	119.7	4.3
Spain	300.7	39.0
Sweden	201.9	8.7
Switzerland	194.7	6.9
United Kingdom	1592.9	57.7
United States	5142.2	252.7

On considère le modèle statistique défini par $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ pour tout $i = 1 \dots 17$, avec ε_i vérifiant les conditions standards d'un modèle de régression linéaire simple.

1. Ecrire le modèle sous forme matricielle. Donner les expressions des estimateurs des MCO $\hat{\beta}_0$ et $\hat{\beta}_1$ de β_0 et β_1 . Donner les valeurs de ces estimateurs calculés sur les observations.
2. Ecrire l'équation d'analyse de la variance et calculer le coefficient de détermination R^2 .
3. Donner l'expression de l'estimateur sans biais $\widehat{\sigma^2}$ de σ^2 . Calculer sa valeur sur les observations.

Dans les questions qui suivent, on suppose les ε_i i.i.d. de loi $N(0, \sigma^2)$.

4. Déterminer les lois du vecteur aléatoire $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)$ et des variables marginales $\hat{\beta}_0$ et $\hat{\beta}_1$.
5. Donner les expressions des estimateurs $\hat{\sigma}(\hat{\beta}_0)$ et $\hat{\sigma}(\hat{\beta}_1)$ des écart-types $\sigma(\hat{\beta}_0)$ et $\sigma(\hat{\beta}_1)$ de $\hat{\beta}_0$ et $\hat{\beta}_1$. Donner les valeurs de ces estimateurs calculés sur les observations.
6. Déterminer un intervalle de confiance à 95% pour β_1 . Tester l'hypothèse nulle (H_0) : $\beta_1 = 0$ contre l'alternative (H_1) : $\beta_1 \neq 0$ au niveau 5%. Commenter.
7. Tester l'hypothèse nulle (H_0) : $\beta_0 = 0$ contre l'alternative (H_1) : $\beta_0 \neq 0$ au niveau 5%. Commenter.

Exercice 6 : Modèle de régression linéaire simple sans constante

On considère le modèle statistique de régression linéaire simple suivant :

$$Y_i = \beta x_i + \varepsilon_i, \quad i = 1, \dots, n,$$

où les bruits ε_i sont des variables aléatoires vérifiant les conditions standards de la régression (à rappeler). On définit deux estimateurs de β :

$$\hat{\beta} = \frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2} \quad \text{et} \quad \beta^* = \frac{\sum_{i=1}^n Y_i}{\sum_{i=1}^n x_i}$$

1. Quelle est la logique de construction de ces deux estimateurs ?
2. Montrer que $\hat{\beta}$ et β^* sont des estimateurs sans biais de β .
3. Montrer que $\text{var}(\beta^*) > \text{var}(\hat{\beta})$ sauf dans le cas où les x_i sont tous égaux.
4. On note $\hat{Y}_i = \hat{\beta} x_i$. Pourquoi l'équation d'analyse de la variance $SCT = SCR + SCE$, où $SCT = \sum_{i=1}^n (Y_i - \bar{Y})^2$, $SCE = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$ et $SCR = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$ n'est-elle plus vérifiée dans ce modèle ?
5. Par quelle équation peut-elle être remplacée ?
6. Comment peut-on définir dans ce cas le coefficient de détermination ?

Exercice 7 : Consommation d'alcool et espérance de vie

On dispose des données issues du rapport publié par l'OMS en février 2011 sur la consommation d'alcool dans le monde en projection pour l'année 2008 et de l'espérance de vie à la naissance en 2009 pour 188 pays. Les consommations d'alcool (en L d'alcool pur par adulte de plus de 15 ans pour l'année 2008) pour ces 188 pays sont notées x_i pour $i = 1 \dots 188$. Les espérances de vie à la naissance en 2009 sont notées pour les mêmes pays y_i pour $i = 1 \dots 188$. On a les résultats numériques suivants :

$$\sum_{i=1}^{188} x_i = 1250.77 \quad \sum_{i=1}^{188} x_i^2 = 12699.04 \quad \sum_{i=1}^{188} x_i y_i = 88858.02 \quad \sum_{i=1}^{188} y_i = 12935 \quad \sum_{i=1}^{188} y_i^2 = 907647.$$

1. Pour un modèle de régression linéaire simple complet, déterminer les valeurs des estimateurs des moindres carrés ordinaires des coefficients de régression et du coefficient de détermination calculés sur ces données.
2. Pour un modèle de régression linéaire simple sans constante, déterminer les valeurs de l'estimateur des moindres carrés ordinaires du coefficient de régression et du coefficient de détermination (défini dans l'exercice 6) calculés sur ces données.
3. Que constate-t-on ? Faut-il pour autant préférer le modèle de régression linéaire simple sans constante au modèle de régression linéaire simple complet ?

