Master Statistique Appliquée

Mention Statistique pour l'Entreprise

Modèles de régression linéaire

Magalie Fromont Renoir

Table des matières

Chapitre 1

Le modèle de régression linéaire simple

1.1 Introduction

On dispose au point de départ des observations $(x_1, y_1), \ldots, (x_n, y_n)$ de n couples $(X_1, Y_1), \ldots, (X_n, Y_n)$ 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

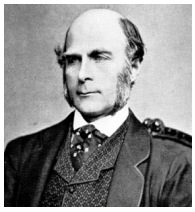
Gauss et Legendre (1795/1809, 1805) : mécanique céleste et méthode des moindres carrés ordinaires.

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.

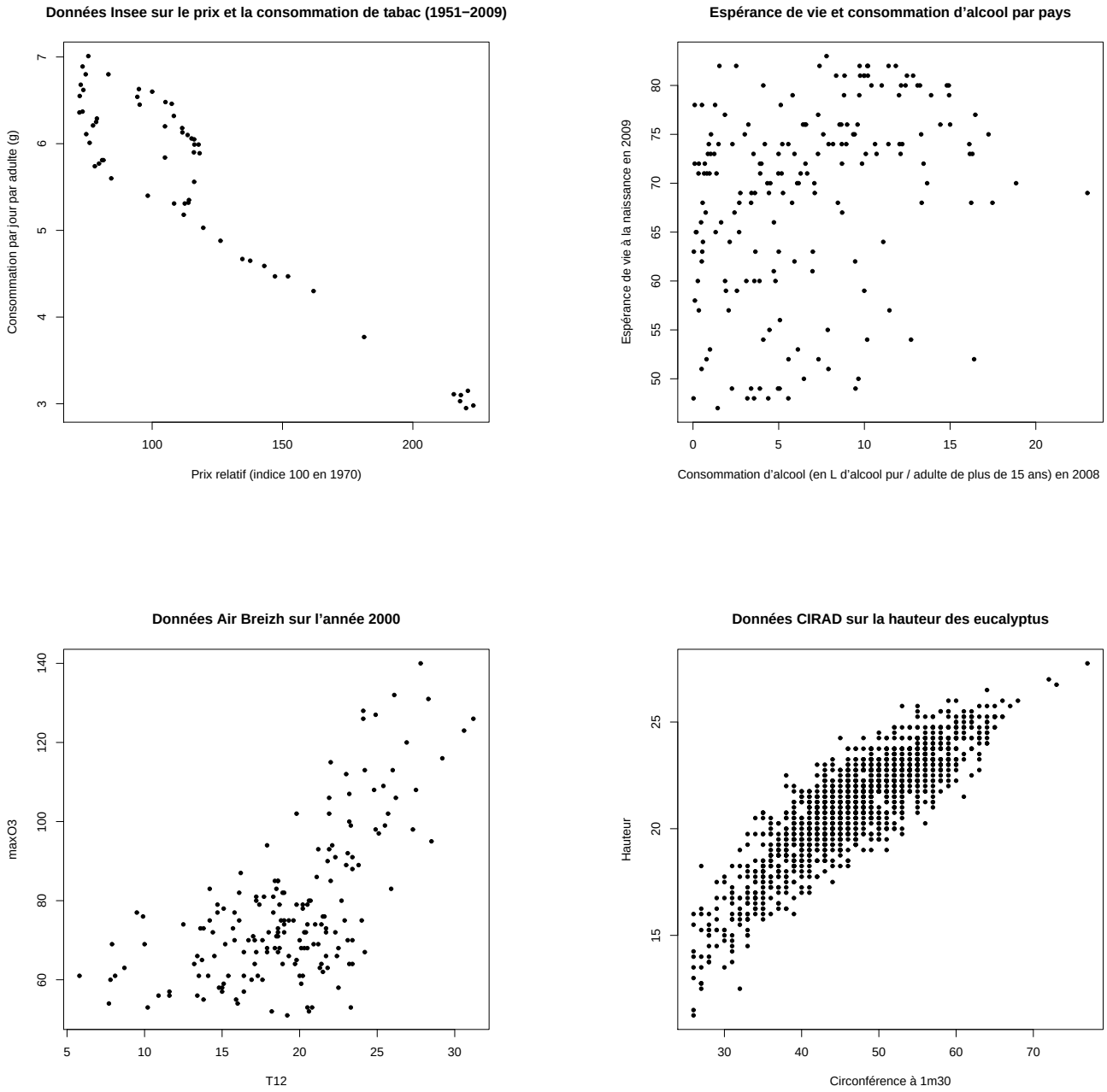
1.2.4 Qualité de l'air en Bretagne

Données fournies par Air Breizh : mesures du maximum journalier de la concentration en O_3 (en μg/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 :

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

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 *cœfficients 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 :

- (C_1) : $\mathbb{E}[\varepsilon_i] = 0$ pour tout $i = 1 \dots n$ (centrage),
- (C_2) : $\text{cov}(\varepsilon_i, \varepsilon_j) = 0$ pour tout $i \neq j$ (non corrélation),
- (C_3) : $\text{var}(\varepsilon_i) = \sigma^2$ (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 :

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

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 (C_1) à (C_3),

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

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 cœfficients 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 $(\hat{\beta}_0, \hat{\beta}_1) \in \mathrm{argmin}_{(\beta_0,\beta_1)\in\mathbb{R}^2} \sum_{i=1}^n l(Y_i, \beta_0 + \beta_1 x_i)$.

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

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

Choix usuel de la perte quadratique ⇒ 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 $(\sum_{i=1}^n x_i)^2 \leq \sum_{i=1}^n 1^2 . \sum_{i=1}^n x_i^2 \leq n \sum_{i=1}^n x_i^2$, avec égalité lorsque $(x_1, \ldots, x_n)$ est colinéaire à $(1, \ldots, 1)$, 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

Thèorème 1. *Les estimateurs des MCO vérifient les propriétés suivantes.*

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

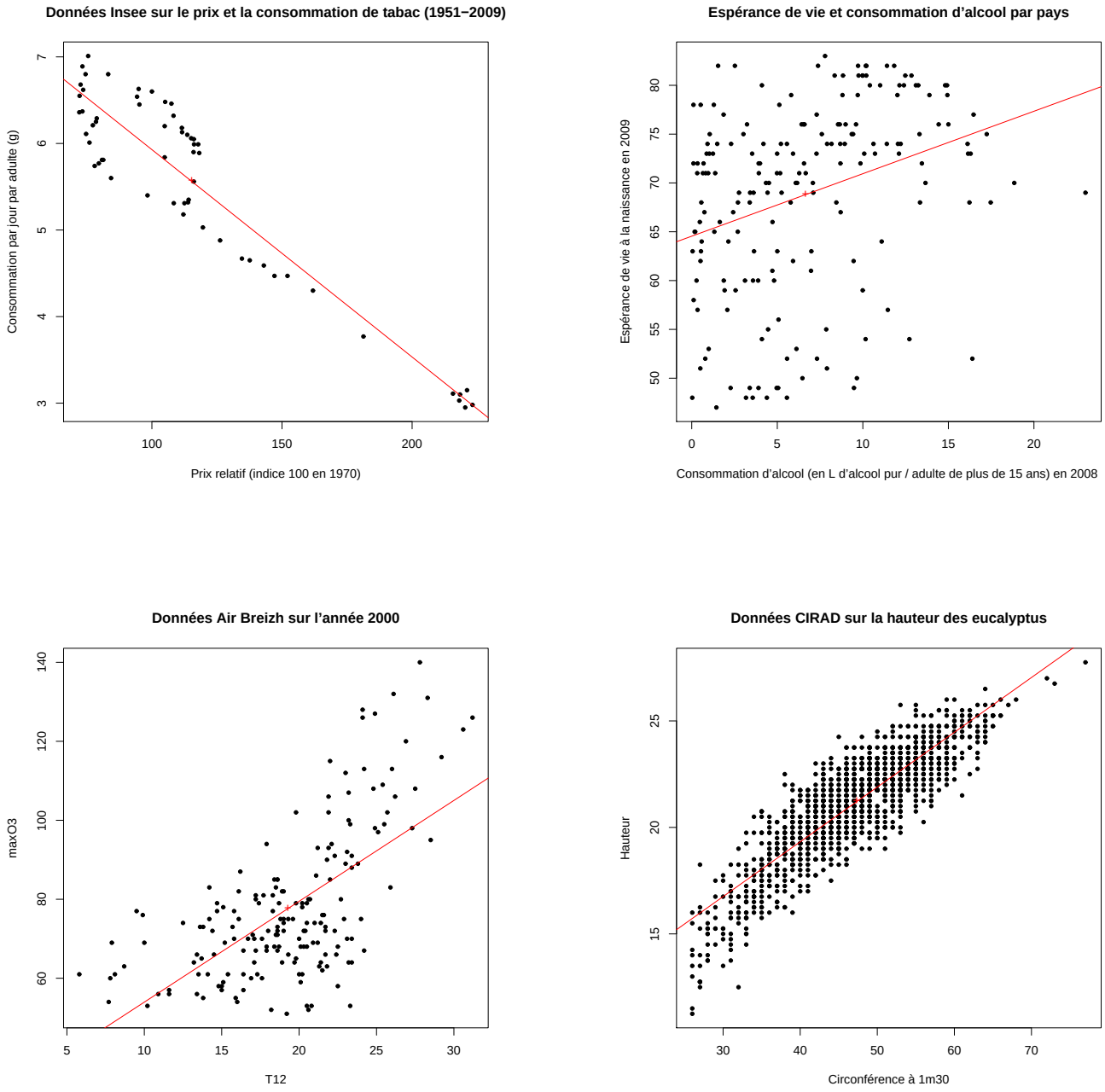
Preuve.

$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x}) Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = \sum_{i=1}^n w_i Y_i$, 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 :
$\mathbb{E}[\hat{\beta}_1] = \mathbb{E}[\sum_{i=1}^n w_i Y_i] = \sum_{i=1}^n w_i(\beta_0 + \beta_1 x_i) = \beta_1 \sum_{i=1}^n w_i x_i = \beta_1$, 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,
$\text{var}(\hat{\beta}_0) = \sum_{i=1}^n \left(\frac{1}{n^2} - 2\frac{\bar{x}w_i}{n} + \bar{x}^2 w_i^2\right)\sigma^2 = \left(\frac{1}{n} - 0 + \bar{x}^2 \sum_{i=1}^n w_i^2\right)\sigma^2 = \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i-\bar{x})^2}\right)\sigma^2 = \frac{\sum_{i=1}^n x_i^2}{n\sum_{i=1}^n (x_i-\bar{x}_i)^2}\sigma^2$,
$\text{cov}(\hat{\beta}_0, \hat{\beta}_1) = \text{cov}(\sum_{i=1}^n (\frac{1}{n} - \bar{x}w_i)Y_i, \sum_{i=1}^n w_i Y_i) = \sum_{i=1}^n \sum_{j=1}^n (\frac{1}{n} - \bar{x}w_i) w_j \text{cov}(Y_i, Y_j) = \sum_{i=1}^n (\frac{1}{n} - \bar{x}w_i) w_i \sigma^2 = -\frac{\bar{x}}{\sum_{i=1}^n (x_i-\bar{x})^2}\sigma^2$,
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.

Thèorème 2 (Gauss Markov)**.** *Parmi les estimateurs linéaires sans biais de β_0 et β_1 respectivement linéaires en $(Y_i)_{i=1\ldots n}$, $\hat{\beta}_0$ 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 $i = 1 \ldots n$, 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
$\hat{\varepsilon}_i = \beta_0+\beta_1 x_i+\varepsilon_i-\hat{\beta}_0-\hat{\beta}_1 x_i = \bar{Y}-\beta_1\bar{x}-\bar{\varepsilon}+\beta_1 x_i+\varepsilon_i-\bar{Y}+\hat{\beta}_1\bar{x}-\hat{\beta}_1 x_i = (\varepsilon_i-\bar{\varepsilon})+(\beta_1-\hat{\beta}_1)(x_i-\bar{x})$. D'où

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

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

$\sum_{i=1}^n \hat{\varepsilon}_i^2 = \sum_{i=1}^n (\varepsilon_i-\bar{\varepsilon})^2 + \sum_{i=1}^n (\beta_1-\hat{\beta}_1)^2(x_i-\bar{x})^2 - 2\sum_{i=1}^n (\beta_1-\hat{\beta}_1)^2(x_i-\bar{x})^2$

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

1.4.3 Le cœfficient 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 :

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

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

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,*

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

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 cœfficients 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 $\hat{\varepsilon}^p_{n+1} = Y_{n+1} - \hat{Y}^p_{n+1}$ (inconnue), dont la variance est égale à

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

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 $\|.\|$ 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 orthognale 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 :
— $\Pi_{\mathbb{X}}' = \Pi_{\mathbb{X}}$ (symétrie),
— $\Pi_{\mathbb{X}}\Pi_{\mathbb{X}} = \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}' = \Pi_{\mathbb{X}}$ (idempotence),
— $\Pi_{\mathbb{X}}\mathbb{X}\beta = \mathbb{X}\beta$ pour tout $\beta \in \mathbb{R}^2$.

On retrouve ainsi le résultat :
$\hat{\beta} \in \text{argmin}_{\beta\in\mathbb{R}^2} \|Y - \mathbb{X}\beta\|^2$ si et seulement si $\mathbb{X}\hat{\beta}$ est la projection orthogonale de Y sur $\mathcal{E}(\mathbb{X})$.
En utilisant ces notations, le vecteur des résidus vérifie $\hat{\varepsilon} = (\hat{\varepsilon}_1, \dots, \hat{\varepsilon}_n)' = Y - \hat{Y} = (I_n - \Pi_{\mathbb{X}})Y$, c'est-à-dire que $\hat{\varepsilon}$ est le projeté orthogonal de Y sur $\mathcal{E}(\mathbb{X})^{\perp}$.
On peut ainsi montrer facilement que
— $\mathbb{X}'\hat{\varepsilon} = (0,0)'$ i.e. $\sum_{i=1}^n \hat{\varepsilon}_i = 0$ et $\sum_{i=1}^n x_i\hat{\varepsilon}_i = 0$,
— $\mathbb{E}[\hat{\varepsilon}] = 0$,
— $\text{Var}(\hat{\varepsilon}) = \sigma^2(I_n - \Pi_{\mathbb{X}})$ (corrélation négative et hétéroscédasticité).

Représentation géométrique.

Cœfficient 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 $\mathbb{1}$,

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

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 :

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

Représentation graphique.

On a alors

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

où θ est l'angle formé par les vecteurs $Y - \Pi_{\mathbb{1}} Y$ et $\Pi_{\mathbb{X}} Y - \Pi_{\mathbb{1}} Y = \hat{Y} - \Pi_{\mathbb{1}} Y$.

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 cœfficients de régression, la variance du modèle, puis des tests d'hypothèses sur les cœfficients 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 (C_4) : le vecteur ε suit une loi gaussienne.

Les hypothèses (C_1) à (C_4) réunies se résument ainsi à :

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

On a alors

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

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

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

d'où

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

A σ^2 fixé, $\ln L(\beta_0, \beta_1, \sigma^2, Y_1, \ldots, Y_n)$ 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, $\frac{\partial \ln L}{\partial \sigma^2}(\hat{\beta}_0, \hat{\beta}_1, \widetilde{\sigma^2}, Y_1, \ldots, Y_n) = 0$ si et seulement si $\widetilde{\sigma^2} = \frac{1}{n}\sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \frac{1}{n}\sum_{i=1}^n \hat{\varepsilon}_i^2$. L'estimateur du maximum de vraisemblance $\widetilde{\sigma^2}$ de σ^2 est biaisé : on lui préfèrera en général $\widehat{\sigma^2}$.

Thèorème 4. *Les estimateurs des MCO vérifient les propriétés suivantes :*

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

— $\frac{n-2}{\sigma^2}\widehat{\sigma^2} \sim \chi^2(n-2)$,

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

La preuve sera vue dans les chapitres suivants.

On en déduit en particulier les propriétés suivantes.
— $\frac{\hat{\beta}_0-\beta_0}{\sqrt{\frac{\widehat{\sigma^2}\sum_{i=1}^n x_i^2}{n\sum_{i=1}^n (x_i-\bar{x})^2}}} \sim \mathcal{T}(n-2)$,
— $\frac{\hat{\beta}_1-\beta_1}{\sqrt{\frac{\widehat{\sigma^2}}{\sum_{i=1}^n (x_i-\bar{x})^2}}} \sim \mathcal{T}(n-2)$,
— $\frac{1}{2\widehat{\sigma^2}}(\hat{\beta}-\beta)'V^{-1}(\hat{\beta}-\beta) \sim \mathcal{F}(2,n-2)$.

1.5.2 Intervalles et régions de confiance pour les cœfficients de régression

Thèorème 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*

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

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

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

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

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

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 $T_0(Y) = \frac{\hat{\beta}_0-b}{\sqrt{\frac{\widehat{\sigma^2}\sum_{i=1}^n x_i^2}{n\sum_{i=1}^n (x_i-\bar{x})^2}}}$ 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 $T_1(Y) = \frac{\hat{\beta}_1 - b}{\sqrt{\frac{\widehat{\sigma^2}}{\sum_{i=1}^n (x_i - \bar{x})^2}}}$ 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

Thèorème 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 $\left[\frac{(n-2)\widehat{\sigma^2}}{c_{n-2}(1-\alpha/2)}, \frac{(n-2)\widehat{\sigma^2}}{c_{n-2}(\alpha/2)}\right]$.

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 $S^2(Y) = \frac{n-2}{s^2}\widehat{\sigma^2}$ 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 $\widehat{I}$ 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

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

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

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

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

1.6 Exercices

Exercice 1 : Questions de cours - QCM

On dispose d'observations $y_1, \ldots, y_n$ de variables aléatoires telles que $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ pour tout $i = 1 \ldots n$, 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 cœfficients de régression sont-ils indépendants ?
   a) Oui.
   b) Non.
   c) Pas toujours.

4. Est-il possible de trouver des estimateurs des cœfficients 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 cœfficients 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-il indépendants ?
   a) Oui.
   b) Non.
   c) Pas toujours.

8. On dispose d'une nouvelle valeur x_{n+1} et on note $\hat{Y}^p_{n+1}$ la valeur prédite correspondante,

$\hat{\varepsilon}^p_{n+1}$ l'erreur de prédiction correspondante. La variance de $\hat{\varepsilon}^p_{n+1}$ 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 cœfficients 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 :

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

Déterminer les valeurs des estimateurs des moindres carrés ordinaires des cœfficients de régression, de l'estimateur sans biais de la variance, et du cœfficient 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 :

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad i = 1, \cdots, n,$$

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 $Y_1, \cdots, Y_n$.
— 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, $i = 1 \ldots, 17$.

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

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 :

$$Y_i = \beta x_i + \varepsilon_i, \qquad i = 1, \cdots, n,$$

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 cœfficient 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 :

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

1. Pour un modèle de régression linéaire simple complet, déterminer les valeurs des estimateurs des moindres carrés ordinaires des cœfficients de régression et du cœfficient 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 cœfficient de régression et du cœfficient 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 ?

Chapitre 2

Le modèle de régression linéaire multiple

2.1 Introduction : retour sur les exemples

La modélisation des données des exemples choisis par un modèle de régression linaire simple peut être critiquée : on voit bien dans certains des exemples que ce modèle, trop simpliste, n'est pas adapté...
Des variables explicatives supplémentaires à prendre en considération, et si oui, lesquelles ?
Dans chacun des exemples, on peut envisager d'introduire de nouvelles variables explicatives, quantitatives ou qualitatives (catégorielles).
Pour les données Insee, on peut introduire des informations sur les campagnes ou lois anti-tabac mises en place ; pour les données sur l'espérance de vie, on peut imaginer beaucoup de variables explicatives, et parmi les plus pertinentes, des indicateurs de richesse, des variables sur le système de santé ; pour les données Air Breizh, par ex. la nébulosité, la vitesse et la direction du vent, les températures à différentes heures de la journée, etc. Pour les données Cirad, il semble évident que la prise en compte de la racine carrée de la circonférence serait pertinente, mais on peut aussi considérer la zone de plantation par ex.

On cherche donc à généraliser le modèle précédent, en considérant non pas une mais plusieurs variables explicatives.
On ne considère pas dans ce chapitre le caractère éventuellement aléatoire des variables explicatives, quitte à conditionner sachant les valeurs de ces variables.

2.2 Modélisation

On introduit le modèle statistique suivant :

$$Y_i = \beta_0 x_{i,0} + \beta_1 x_{i,1} + \ldots + \beta_{p-1} x_{i,p-1} + \varepsilon_i, \quad \text{pour} \quad i = 1 \ldots n,$$

où

- $p \leq n$,
- Y_i est une variable aléatoire observée, appelée *variable à expliquer*,

— $x_{i,0}, x_{i,1}, \ldots, x_{i,p-1}$ sont des valeurs réelles déterministes appelées par extension directe du cas aléatoire *variables explicatives*. Souvent $x_{i,0} = 1$ pour tout $i = 1 \ldots n$, mais PAS TOUJOURS.
— $\beta_0, \beta_1, \ldots, \beta_{p-1}$ sont des paramètres réels inconnus appelés *paramètres de régression* ou *cœfficients de régression*,
— les ε_i sont des variables aléatoires, non observées, appelées *erreurs* ou *bruits*, auxquelles on impose certaines conditions complémentaires.

Les conditions standards imposées aux ε_i sont les conditions (C_1) à (C_3) vues dans le chapitre précédent i.e.
— $(C_1) : \mathbb{E}[\varepsilon_i] = 0$ pour tout $i = 1 \ldots n$ (centrage),
— $(C_2) : \text{cov}(\varepsilon_i, \varepsilon_j) = 0$ pour tout $i \neq j$ (non corrélation),
— $(C_3) : \text{var}(\varepsilon_i) = \sigma^2$ (inconnue) pour tout $i = 1 \ldots n$ (homoscédasticité).

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

Il s'exprime de façon vectorielle :

$$Y = \mathbb{X}\beta + \varepsilon, \tag{2.1}$$

avec

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbb{X} = \begin{pmatrix} x_{1,0} & \ldots & x_{1,p-1} \\ \vdots & & \vdots \\ x_{n,0} & \ldots & x_{n,p-1} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_{p-1} \end{pmatrix} \quad \text{et} \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Sous les conditions (C_1) à (C_3), on a alors :
— $\mathbb{E}[\varepsilon] = 0$ et $\mathbb{E}[Y] = \mathbb{X}\beta$,
— $\text{Var}(\varepsilon) = \text{Var}(Y) = \sigma^2 I_n$.

La matrice $\mathbb{X}$ est appelée *matrice du plan d'expérience*.
On suppose que cette matrice est de plein rang, c'est-à-dire $\text{rang}(\mathbb{X}) = p$. Ses vecteurs colonnes sont linéairement indépendants. Cela implique en particulier que la matrice symétrique $\mathbb{X}'\mathbb{X}$ est définie positive.

2.3 Exemples de modèles de régression linéaire multiple

Les variables explicatives peuvent prendre différentes formes.

2.3.1 Variables explicatives quantitatives

Exemple des données Insee sur le tabac : Y_i = consommation de tabac en grammes par adulte par jour au cours de l'année i, $x_{i,0} = 1$, $x_{i,1}$ = prix relatif du tabac l'année i, $x_{i,2}$ = coût des campagnes publicitaires anti-tabac diffusées au cours de l'année i.

Exemple des données de l'OMS sur l'espérance de vie : Y_i = l'espérance de vie dans le $i$ème pays, $x_{i,0}$ = le PIB, $x_{i,1}$ = le revenu moyen par habitant, $x_{i,2}$ = le budget consacré à la santé.

Exemple des données Air Breizh : Y_i = maximum journalier de la concentration en ozone au jour i, $x_{i,0} = 1$, $x_{i,1}$ = la température à midi au jour i, $x_{i,2}$ = la température à 9 heures au jour i, $x_{i,3}$ = la nébulosité à midi au jour i, $x_{i,4}$ = la nébulosité à 9 heures, $x_{i,5}$ = la vitesse du vent au jour i...

2.3.2 Transformations de variables explicatives quantitatives

Exemple des données Cirad : Y_i = hauteur de l'eucalyptus i, $x_{i,0} = 1$, $x_{i,1}$ = la circonférence à 1m30 de l'eucalyptus i, $x_{i,2} = \sqrt{x_{i,1}}$.

On peut en fait considérer des transformations polynômiales, exponentielles, logarithmiques, trigonométriques... des variables explicatives quantitatives. Attention, ces transformations ne doivent pas faire intervenir de nouveaux paramètres inconnus !

2.3.3 Variables explicatives qualitatives

Dans le cas de variables explicatives qualitatives, on les représente sous la forme d'indicatrices.

Exemple des données Insee : $x_{i,3} = 1$ si une loi anti-tabac a été votée au cours de l'année i, 0 sinon.

Exemple des données de l'OMS : , $x_{i,3} = 1$ si le pays est dans une zone géographique particulière, 0 sinon, $x_{i,4} = 1$ si le pays est en guerre, 0 sinon...

Exemple des données Air Breizh : $x_{i_6} = 1$ si le vent a pour direction l'est, 0 sinon, $x_{i_7} = 1$ si le vent a pour direction l'ouest, 0 sinon, $x_{i_8} = 1$ si le vent a pour direction le nord, 0 sinon, $x_{i_9} = 1$ si le vent a pour direction le sud, 0 sinon, etc.

Exemple des données Cirad : $x_{i,3} = 1$ si l'eucalyptus i est situé dans le bloc A de la plantation, 0 sinon, $x_{i,4} = 1$ si l'eucalyptus i est situé dans le bloc B de la plantation, 0 sinon, etc.

2.3.4 Interactions

On peut envisager le cas où les variables explicatives interagissent entre elles. Ce phénomène est modélisé par des produits des différentes variables. Ces interactions peuvent être d'ordres variés.

Remarque : les modèles de régression linéaire multiple avec des variables explicatives qualitatives seront traités en cours d'ANOVA.

2.4 Estimateur des moindres carrés ordinaires

Comme pour la régression linéaire simple, on choisit ici comme fonction de perte la perte quadratique.

Définition 5. *L'estimateur des moindres carrés ordinaires de β dans le modèle de régression linéaire multiple (2.1) est défini par*

$$\hat{\beta} \in \underset{\beta\in\mathbb{R}^p}{argmin} \sum_{i=1}^{n} \left(Y_i - \sum_{j=0}^{p-1} \beta_j x_{i,j} \right)^2 = \underset{\beta\in\mathbb{R}^p}{argmin} \|Y - \mathbb{X}\beta\|^2,$$

où $\|.\|$ est la norme euclidienne de $\mathbb{R}^n$.

On montre que

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

Preuves analytique et géométrique (c.f. chapitre précédent).
Soit $L : \beta \mapsto (Y - \mathbb{X}\beta)'(Y - \mathbb{X}\beta) = Y'Y - Y'\mathbb{X}\beta - \beta'\mathbb{X}'Y + \beta'\mathbb{X}'\mathbb{X}\beta = Y'Y - 2\beta'\mathbb{X}'Y + \beta'\mathbb{X}'\mathbb{X}\beta$.

Le vecteur $\hat{\beta} = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'Y$ est bien un point critique de L puisque $\nabla L(\hat{\beta}) = 2\mathbb{X}'\mathbb{X}\hat{\beta} - 2\mathbb{X}'Y = 0$. Ce point critique correspond à un minimum. En effet, la matrice hessienne de L en $\hat{\beta}$ vaut $2\mathbb{X}'\mathbb{X}$ qui est définie positive.
On introduit maintenant, comme pour la régression linéaire simple, le sous-espace vectoriel $\mathcal{E}(\mathbb{X})$ de $\mathbb{R}^n$ engendré par les vecteurs colonnes de $\mathbb{X}$. Par définition, $\mathbb{X}\hat{\beta}$ est un vecteur de $\mathcal{E}(\mathbb{X})$ dont la distance euclidienne avec Y est la distance minimum entre Y et tout vecteur de $\mathcal{E}(\mathbb{X})$. Par conséquent, si l'on note $\Pi_{\mathbb{X}}$ la matrice de projection orthogonale sur $\mathcal{E}(\mathbb{X})$, alors $\mathbb{X}\hat{\beta} = \Pi_{\mathbb{X}}Y$. Là encore, on peut montrer que la matrice $\Pi_{\mathbb{X}}$ s'écrit aussi $\mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'$, d'où $\mathbb{X}\hat{\beta} = \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'Y$, puis $\hat{\beta} = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'Y$.

Proposition 1. *L'estimateur des MCO $\hat{\beta}$ est un estimateur linéaire sans biais de β, dont la matrice de variance covariance est donnée par*

$$Var(\hat{\beta}) = \sigma^2(\mathbb{X}'\mathbb{X})^{-1}.$$

Preuve.
Puisque $\hat{\beta} = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'Y$, il s'agit bien d'un estimateur linéaire (en Y). De plus, $\mathbb{E}[\hat{\beta}] = \mathbb{E}[(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'Y] = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\mathbb{X}\beta = \beta$, donc $\hat{\beta}$ est sans biais. Enfin, $\text{Var}(\hat{\beta}) = \text{Var}\left((\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'Y\right) = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\text{Var}(Y)\mathbb{X}(\mathbb{X}'\mathbb{X})^{-1} = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\left(\sigma^2 I_n\right)\mathbb{X}(\mathbb{X}'\mathbb{X})^{-1} = \sigma^2(\mathbb{X}'\mathbb{X})^{-1}$.

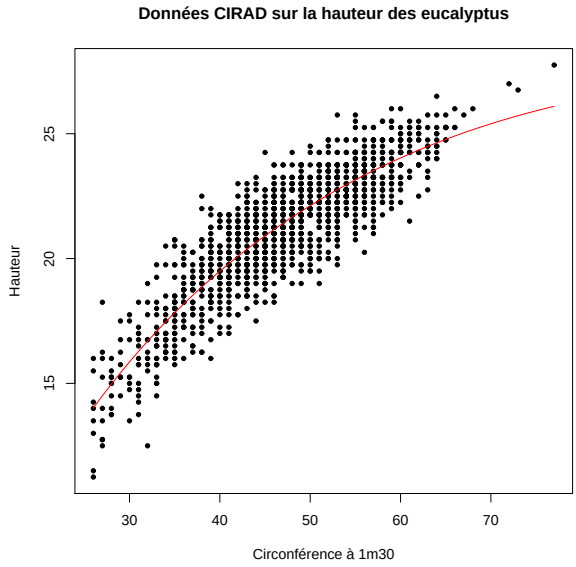
Thèorème 8 (Gauss Markov)**.** *L'estimateur $\hat{\beta}$ des moindres carrés ordinaires est l'unique estimateur linéaire sans biais de variance minimale parmi les estimateurs linéaires sans biais de β.*

Preuve (sans l'unicité).
Soit $\tilde{\beta}$ un estimateur linéaire sans biais de β. $\tilde{\beta}$ s'écrit donc $\tilde{\beta} = AY$, avec $A\mathbb{X}\beta = \beta$ pour tout β c'est-à-dire $A\mathbb{X} = I_p$.

$$\begin{aligned}
\text{Var}(AY) &= \text{Var}\left(A(I_n - \Pi_{\mathbb{X}} + \Pi_{\mathbb{X}})Y\right) \\
&= \text{Var}(A(I_n - \Pi_{\mathbb{X}})Y + A\Pi_{\mathbb{X}}Y) \\
&= A(I_n - \Pi_{\mathbb{X}})\sigma^2 I_n (I_n - \Pi_{\mathbb{X}})A' + 2A(I_n - \Pi_{\mathbb{X}})\sigma^2 I_n \Pi_{\mathbb{X}} A' + \text{Var}\left(A\mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'Y\right) \\
&= \sigma^2 A(I_n - \Pi_{\mathbb{X}})A' + 0 + \text{Var}\left((\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'Y\right) \\
&= \sigma^2 A(I_n - \Pi_{\mathbb{X}})A' + \text{Var}\left(\hat{\beta}\right).
\end{aligned}$$

Puisque la matrice $A(I_n - \Pi_{\mathbb{X}})A'$ est symétrique réelle positive (rappel sur la relation d'ordre partielle entre matrices symétriques réelles), on en conclut que $\hat{\beta}$ est de variance minimale parmi les estimateurs linéaires sans biais.

FIGURE 2.1 – Données Cirad : représentation de la courbe de régression obtenue



2.5 Valeurs ajustées, résidus

Définition 6. *Le vecteur aléatoire* $\hat{Y} = \Pi_{\mathbb{X}} Y = \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'Y$ *est appelé le* vecteur des valeurs ajustées.
Le vecteur $\hat{\varepsilon} = Y - \hat{Y} = (I_n - \Pi_{\mathbb{X}})Y$ *est appelé le* vecteur des résidus.
La matrice $\Pi_{\mathbb{X}}$ *est parfois appelée la matrice "chapeau" (hat matrix en anglais), et souvent notée dans ce cas H. Ses cœfficients sont notés* $h_{i,j}$.

Remarque : $\hat{\varepsilon}$ est orthogonal au vecteur $\hat{Y}$. Il correspond au projeté orthogonal de Y sur $\mathcal{E}(\mathbb{X})^{\perp}$.
Représentation géométrique.

On peut ensuite montrer (facilement) le résultat suivant.

Proposition 2. $\mathbb{X}'\hat{\varepsilon} = 0$, $\mathbb{E}[\hat{\varepsilon}] = 0$ *et* $Var(\hat{\varepsilon}) = \sigma^2(I_n - \Pi_{\mathbb{X}})$.

Les résidus sont centrés, corrélés et hétéroscédastiques.

2.6 Somme des carrés résiduelle et estimation ponctuelle de la variance

Comme dans le modèle de régression linéaire simple, le vecteur des résidus peut servir à l'estimation ponctuelle de la variance σ^2.
On introduit la *somme des carrés résiduelle* : SCR $=\sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \|\hat{\varepsilon}\|^2$.
On a $\mathbb{E}\left[\|\hat{\varepsilon}\|^2\right] = \mathbb{E}\left[\hat{\varepsilon}'\hat{\varepsilon}\right] = \mathbb{E}\left[tr\left(\hat{\varepsilon}'\hat{\varepsilon}\right)\right]$ (astuce de la trace).
Or pour toutes matrices A, B, C, $tr(ABC) = tr(CAB) = tr(BCA)$, d'où :
$\mathbb{E}\left[\|\hat{\varepsilon}\|^2\right] = \mathbb{E}\left[tr\left(\hat{\varepsilon}\hat{\varepsilon}'\right)\right] = tr\mathbb{E}\left[\hat{\varepsilon}\hat{\varepsilon}'\right] = tr(\mathrm{Var}(\hat{\varepsilon})) = \sigma^2 tr(I_n - \Pi_{\mathbb{X}}) = \sigma^2(n - tr(\mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'))$.
On a donc $\mathbb{E}\left[\|\hat{\varepsilon}\|^2\right] = \sigma^2(n - tr(\mathbb{X}'\mathbb{X}(\mathbb{X}'\mathbb{X})^{-1})) = \sigma^2(n - tr(I_p)) = \sigma^2(n-p)$.

Proposition 3. *Un estimateur sans biais de la variance σ^2 est donné par $\widehat{\sigma^2} = SCR/(n-p) = \|\hat{\varepsilon}\|^2/(n-p)$.*

2.7 Equation d'analyse de la variance, cœfficient de détermination

On a défini la *somme des carrés résiduelle* : SCR $=\|\hat{\varepsilon}\|^2$.
On introduit maintenant la *somme des carrés totale* : SCT $=\|Y - \bar{Y}\mathbb{1}\|^2$ si $\mathbb{1}$ est l'un des vecteurs colonnes de la matrice $\mathbb{X}$, ou la *somme des carrés totale sans constante* : $\mathrm{SCT}_{sc}=\|Y\|^2$ si le vecteur $\mathbb{1}$ n'est pas l'un des vecteurs colonnes de la matrice $\mathbb{X}$.
On introduit aussi la *somme des carrés expliquée* : SCE $=\|\hat{Y} - \bar{Y}\mathbb{1}\|^2$ si $\mathbb{1}$ est l'un des vecteurs colonnes de la matrice $\mathbb{X}$, ou la *somme des carrés expliquée sans constante* : $\mathrm{SCE}_{sc}=\|\hat{Y}\|^2$ si $\mathbb{1}$ n'est pas l'un des vecteurs colonnes de la matrice $\mathbb{X}$.
On a, par le théorème de Pythagore : SCT=SCE+SCR (équation d'analyse de la variance) si $\mathbb{1}$ est l'un des vecteurs colonnes de la matrice $\mathbb{X}$, $\mathrm{SCT}_{sc}=\mathrm{SCE}_{sc}$+SCR dans tous les cas.

Définition 7. *Le* cœfficient de détermination R^2 *est défini par :*

$$R^2 = \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT}.$$

Le cœfficient de détermination sans constante R^2_{sc} *est défini par :*

$$R^2_{sc} = \frac{SCE_{sc}}{SCT_{sc}} = 1 - \frac{SCR}{SCT_{sc}}.$$

Interprétations géométriques dans les deux cas. Interprétations des cas limites.

Proposition 4. *Le cœfficient de détermination croît avec le nombre de variables explicatives p.*

Conséquence : on ne peut pas utiliser ce critère comme critère de comparaison entre deux modèles dont les nombres de variables explicatives diffèrent... Idée du R^2 *ajusté* comme critère de comparaison dans ce cas.

2.8 Prédiction

A partir d'une nouvelle valeur explicative $x_{n+1} = (x_{n+1,0}, \ldots, x_{n+1,p-1})$, on souhaite prédire une nouvelle observation d'une variable $Y_{n+1} = \beta_0 x_{n+1,0} + \ldots + \beta_{p-1} x_{n+1,p-1} + \varepsilon_{n+1} = x_{n+1}\beta + \varepsilon_{n+1}$,

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 construire $\hat{\beta}$.

Pour cela, on introduit $\hat{Y}^p_{n+1} = x_{n+1}\hat{\beta}$.

L'erreur de prédiction est définie par $\hat{\varepsilon}^p_{n+1} = Y_{n+1} - \hat{Y}^p_{n+1}$ (inconnue).
Elle est centrée, de variance égale à $\text{var}(\hat{\varepsilon}^p_{n+1}) = \text{var}(x_{n+1}\beta + \varepsilon_{n+1} - x_{n+1}\hat{\beta}) = \text{var}(\varepsilon_{n+1}) + x_{n+1}\text{Var}(\hat{\beta})x'_{n+1} = \sigma^2(1 + x_{n+1}(\mathbb{X}'\mathbb{X})^{-1}x'_{n+1})$.

On remarque par ailleurs que : $\text{var}(\hat{\varepsilon}^p_{n+1}) = \mathbb{E}\left[\left(Y_{n+1} - \hat{Y}^p_{n+1}\right)^2\right]$ appelée aussi *erreur quadratique moyenne de prédiction* (EQMP), qu'on utilisera plus tard pour faire de la sélection de variables ou de modèle.

2.9 Estimation par intervalles de confiance et tests d'hypothèses asymptotiques

Pour construire des intervalles de confiance ou des tests d'hypothèses sur β, on a besoin de connaître la loi de $(\hat{\beta} - \beta)$. Dans le cas général, on ne fait aucune hypothèse sur la loi de ε, donc on peut éventuellement chercher à en avoir une connaissance approximative lorsque n est très grand.

2.9.1 Théorèmes limites

Pour chaque taille d'échantillon n, on précisera la dépendance en n à l'aide de $^{(n)}$: Y sera ainsi noté $Y^{(n)}$, $\mathbb{X}$, $\hat{\beta}$, $\hat{\varepsilon}$ et $\widehat{\sigma^2}$ deviennent respectivement $\mathbb{X}^{(n)}$, $\hat{\beta}^{(n)}$, $\hat{\varepsilon}^{(n)}$ et $\widehat{\sigma^2}^{(n)}$.

Thèorème 9. *Si les ε_i sont maintenant supposées i.i.d. et si $\mathbb{X}^{(n)'}\mathbb{X}^{(n)}/n \to_{n\to+\infty} A$ définie positive, alors*

$$\sqrt{n}\left(\hat{\beta}^{(n)} - \beta\right) \to^{(\mathcal{L})} \mathcal{N}(0, \sigma^2 A^{-1}).$$

Remarque : σ^2 étant inconnu, on l'estime par $\widehat{\sigma^2}^{(n)}$.

Thèorème 10. *Si les ε_i sont supposées i.i.d. alors $\widehat{\sigma^2}^{(n)} \to^{(P)} \sigma^2$.*

Le lemme de Slutsky permet de conclure que si en plus $\mathbb{X}^{(n)'}\mathbb{X}^{(n)}/n \to_{n\to+\infty} A$, la loi de $\sqrt{nA/\widehat{\sigma^2}^{(n)}}\left(\hat{\beta}^{(n)} - \beta\right)$ est asymptotiquement gaussienne centrée réduite.

2.9.2 L'idée du bootstrap non paramétrique

La loi de $\sqrt{n}\left(\hat{\beta}^{(n)} - \beta\right)$ étant inconnue, une méthode de rééchantillonnage permettra de "recréer" à partir des Y_i de nouvelles variables dont la loi conditionnelle sachant les Y_i est proche en un certain sens de cette loi inconnue.
Une méthode de rééchantillonnage classique est celle du bootstrap non paramétrique qu'on peut décrire de la façon suivante.

On suppose que le vecteur $\mathbb{1}$ est l'un des vecteurs colonnes de $\mathbb{X}^{(n)}$.

1. Calcul de l'estimateur des MCO de β, $\hat{\beta}^{(n)}$ à partir de $Y^{(n)}$, puis du vecteur des résidus $\hat{\varepsilon}^{(n)}$.
2. Tirage de n éléments notés $(\hat{\varepsilon}_1^*, \ldots, \hat{\varepsilon}_n^*)$, appelés résidus bootstrapés pris au hasard avec remise dans $\{\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n\}$.
3. A partir de $\hat{\varepsilon}^* = (\hat{\varepsilon}_1^*, \ldots, \hat{\varepsilon}_n^*)'$, calcul de $Y^* = \mathbb{X}^{(n)}\hat{\beta}^{(n)} + \hat{\varepsilon}^*$.
4. Calcul de l'estimateur bootstrapé : $\hat{\beta}^{(n)*} = \left(\mathbb{X}^{(n)'}\mathbb{X}^{(n)}\right)^{-1}\mathbb{X}^{(n)'}Y^*$.

Si d désigne une distance sur les lois de probabilité, alors :

$$d\left(\mathcal{L}\left(\sqrt{n}\left(\hat{\beta}^{(n)*} - \hat{\beta}^{(n)}\right)|Y^{(n)}\right), \mathcal{L}\left(\sqrt{n}\left(\hat{\beta}^{(n)} - \beta\right)\right)\right) \to^{(P)}_{n\to+\infty} 0.$$

Puisque les variables $\sqrt{n}\left(\hat{\beta}^{(n)*} - \hat{\beta}^{(n)}\right)$ se calculent à partir de $Y^{(n)}$, on peut simuler empiriquement la loi $\mathcal{L}\left(\sqrt{n}\left(\hat{\beta}^{(n)*} - \hat{\beta}^{(n)}\right)|Y^{(n)}\right)$ qui "approche" la loi $\mathcal{L}\left(\sqrt{n}\left(\hat{\beta}^{(n)} - \beta\right)\right)$. On peut ainsi déterminer des quantiles empiriques, etc.

2.10 Exercices

Exercice 1 : Questions de cours

On considère le modèle de régression linéaire multiple

$$Y = \mathbb{X}\beta + \varepsilon,$$

où le vecteur Y à valeurs dans $\mathbb{R}^n$ représente la variable à expliquer, $\mathbb{X}$ est une matrice réelle de taille $n \times p$ de rang p, $\beta \in \mathbb{R}^p$ (inconnu) et ε est le vecteur des bruits à valeurs dans $\mathbb{R}^n$.

1. Quelles sont les conditions standards imposées au vecteur des bruits ? Expliquer comment l'analyse du modèle est facilitée par ces conditions.
2. Rappeler les définitions de l'estimateur des moindres carrés ordinaires de β, de la valeur ajustée de Y, puis du vecteur des résidus. Quelle est l'interprétation géométrique de ces vecteurs aléatoires ?
3. Proposer un calcul matriciel de l'estimateur des moindres carrés ordinaires, et préciser les propriétés de cet estimateur. Retrouver à partir du calcul matriciel les estimateurs des moindres carrés ordinaires obtenus lorsque le modèle est un modèle de régression linéaire simple.
4. Le vecteur des résidus $\hat{\varepsilon}$ a-t-il des propriétés analogues à celles de ε ?
5. Donner un estimateur naturel de la variance du modèle. Cet estimateur est-il sans biais ?
6. Peut-on prévoir l'évolution de la somme des carrés résiduelle avec l'ajout d'une variable explicative au modèle ?
7. Préciser l'équation d'analyse de la variance et son interprétation géométrique.
8. Donner la définition du cœfficient de détermination R^2, ainsi que son interprétation géométrique. Discuter des cas limites, et de l'utilisation du R^2 comme mesure de la qualité explicative du modèle.
9. Comment peut-on mesurer la qualité prédictive du modèle ?
10. Peut-on construire des régions de confiance pour β sans faire d'hypothèse sur la loi de ε ?

Exercice 2 : Données Cirad sur la hauteur des eucalyptus

On reprend les données du Cirad présentées en cours, donnant 1429 mesures de la circonférence à 1 mètre 30 du sol et de la longueur du tronc d'eucalyptus d'une parcelle plantée. On a représenté le nuage de points sur le graphique fourni en Annexe 1.1.

1. On cherche à expliquer la longueur du tronc d'un eucalyptus comme une fonction affine de la circonférence du tronc, à une erreur aléatoire près.
a) Ecrire le modèle de régression correspondant, de façon analytique puis de façon vectorielle, en veillant à bien poser les hypothèses.
b) Les valeurs calculées sur les observations des estimateurs des moindres carrés ordinaires des cœfficients de régression sont égales à 9.04 et 0.26, celle du cœfficient de détermination R^2 à 0.7683, et celle de la somme des carrés résiduelle à 2051.457. Représenter la droite de régression obtenue sur le graphique fourni.
2. On cherche maintenant à expliquer, à une erreur aléatoire près, la longueur du tronc d'un eucalyptus comme une fonction linéaire des variables explicatives suivantes : 1, la circonférence et la racine carrée de la circonférence.

a) Ecrire le modèle de régression correspondant, de façon analytique puis de façon vectorielle, en veillant à bien poser les hypothèses : on notera Y le vecteur modélisant les longueurs des troncs des eucalyptus, β le vecteur des cœfficients de régression et $\mathbb{X}$ la matrice du plan d'expérience, supposée de plein rang. Si y désigne l'observation de Y, on a

$$\mathbb{X}'\mathbb{X} = \begin{pmatrix} 1429 & 67660 & 9791.6 \\ 67660 & 3306476 & 471237.9 \\ 9791.6 & 471237.9 & 67660 \end{pmatrix}, \quad \mathbb{X}'y = \begin{pmatrix} 30312.5 \\ 1461695.8 \\ 209685.6 \end{pmatrix}, \quad \text{SCR} = 1840.247.$$

b) Donner la valeur de l'estimateur des moindres carrés ordinaires $\hat{\beta}$ de β calculé sur les observations. Représenter sur le graphique fourni la courbe de régression obtenue.

c) Donner l'expression d'un estimateur $\widehat{\sigma^2}$ sans biais de la variance du modèle. Donner les valeurs de cet estimateur et d'un estimateur sans biais de la matrice de variance-covariance de $\hat{\beta}$ calculés sur les observations.

d) Calculer les valeurs de la somme des carrés expliquée puis du cœfficient de détermination R^2 sur les observations. Comparer ce dernier résultat à la valeur du R^2 dans le modèle de régression linéaire simple. Que peut-on en conclure ?

3. Quelle valeur peut-on prédire pour la longueur du tronc d'un eucalyptus dont la circonférence à 1m30 du sol est de 48cm dans chaque modèle ? Estimer la variance de l'erreur de prédiction correspondante dans les deux modèles. Commenter les résultats.

Exercice 3 : Rôle de la constante dans le modèle

Soit $\mathbb{X}$ une matrice $n \times p$ de rang p. Soit $\hat{Y}$ le projeté orthogonal sur l'espace engendré par les vecteurs colonnes de $\mathbb{X}$ d'un vecteur Y de $\mathbb{R}^n$.
Montrer que $\sum_{i=1}^{n}(Y_i - \hat{Y}_i) = 0$ si l'un des vecteurs colonnes de $\mathbb{X}$ est le vecteur $\mathbb{1} = (1, \ldots, 1)'$.

Exercice 4 : Cœfficient de détermination et modèles emboîtés

Soit $\mathbb{Z}$ une matrice $n \times q$ de rang q, dont le premier vecteur colonne est $\mathbb{1}$, et $\mathbb{X}$ une matrice $n \times p$ de rang p composée des q vecteurs colonnes de $\mathbb{Z}$ et de $p - q$ autres vecteurs linéairement indépendants ($q \leq p \leq n$). On considère les deux modèles de régression linéaire multiple suivants :

$$\begin{aligned} Y &= \mathbb{Z}\beta + \varepsilon \\ Y &= \mathbb{X}\tilde{\beta} + \tilde{\varepsilon}, \end{aligned}$$

où ε et $\tilde{\varepsilon}$ vérifient les conditions standards d'un modèle de régression linéaire multiple. Comparer les cœfficients de détermination R^2 dans les deux modèles. Discuter de l'utilisation du R^2 pour la sélection de modèle ou de variables explicatives.

Exercice 5 : Régression sur variables explicatives orthogonales

On considère le modèle de régression linéaire multiple :

$$Y = \mathbb{X}\beta + \varepsilon,$$

où $Y \in \mathbb{R}^n$, $\mathbb{X}$ est une matrice réelle de taille $n \times p$ ($p \leq n$) composée de p vecteurs colonnes orthogonaux, $\beta = (\beta_0, \ldots, \beta_{p-1})' \in \mathbb{R}^p$, $\varepsilon \in \mathbb{R}^n$ vérifiant les conditions standards. L'estimateur des moindres carrés ordinaires de β est noté $\hat{\beta}^{(\mathbb{X})} = (\hat{\beta}_0^{(\mathbb{X})}, \ldots, \hat{\beta}_{p-1}^{(\mathbb{X})})'$.

Soit U la matrice des q premières colonnes de $\mathbb{X}$ et V la matrice des $p-q$ dernières colonnes de $\mathbb{X}$. On définit à partir de ces matrices deux nouveaux modèles de régression linéaire multiple, l'un à q variables explicatives, l'autre à $p-q$ variables explicatives. Les estimateurs des moindres carrés ordinaires de β obtenus dans ces modèles sont respectivement notés $\hat{\beta}^{(U)} = (\hat{\beta}_0^{(U)}, \ldots, \hat{\beta}_{q-1}^{(U)})'$ et $\hat{\beta}^{(V)} = (\hat{\beta}_q^{(V)}, \ldots, \hat{\beta}_{p-1}^{(V)})'$. Les sommes des carrés expliquées dans les trois modèles sont notées SCE($\mathbb{X}$), SCE(U) et SCE(V).

1. Montrer que SCE($\mathbb{X}$)=SCE(U)+SCE(V).
2. Montrer que pour $0 \leq j \leq q-1$, $\hat{\beta}_j^{(U)} = \hat{\beta}_j^{(\mathbb{X})}$ et que pour $q \leq j \leq p-1$, $\hat{\beta}_j^{(V)} = \hat{\beta}_j^{(\mathbb{X})}$.

Chapitre 3

Le modèle de régression linéaire multiple sous hypothèse gaussienne

3.1 Introduction

On rappelle l'expression vectorielle du modèle de régression linéaire multiple ($1 \leq p \leq n$) :

$$Y = \mathbb{X}\beta + \varepsilon,$$

avec

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbb{X} = \begin{pmatrix} x_{1,0} & \dots & x_{1,p-1} \\ \vdots & & \vdots \\ x_{n,0} & \dots & x_{n,p-1} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_{p-1} \end{pmatrix} \quad \text{et} \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

On rappelle aussi les conditions standards imposées au vecteur des bruits ε :

— (C_1) : $\mathbb{E}[\varepsilon] = 0$ (centrage),
— (C_2) : $\text{cov}(\varepsilon_i, \varepsilon_j) = 0$ pour tout $i \neq j$ (non corrélation),
— (C_3) : $\text{var}(\varepsilon_i) = \sigma^2$ (inconnue) pour tout $i = 1 \dots n$ (homoscédasticité),

La matrice du plan d'expérience $\mathbb{X}$ est supposée de plein rang, c'est-à-dire $\text{rang}(\mathbb{X}) = p$. Cela implique en particulier que la matrice symétrique $\mathbb{X}'\mathbb{X}$ est définie positive.

On souhaite maintenant faire de l'inférence statistique (non asymptotique) sous une hypothèse usuelle :

— (C_4) : ε est un vecteur gaussien.

Les conditions (C_1) à (C_4) sont alors équivalentes à $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ et $Y \sim \mathcal{N}(\mathbb{X}\beta, \sigma^2 I_n)$.

On remarque que les variables ε_i sont maintenant supposées i.i.d., que les Y_i sont supposées indépendantes, et que le modèle est paramétrique, dominé par la mesure de Lebesgue sur $\mathbb{R}^n$.

3.2 Estimateurs du maximum de vraisemblance

La vraisemblance du modèle s'écrit :

$$\mathcal{L}(\beta, \sigma^2, Y_1, \dots, Y_n) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(Y_i - \sum_{j=0}^{p-1}\beta_j x_{i,j}\right)^2\right) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\|Y - \mathbb{X}\beta\|^2\right),$$

d'où

$$\ln \mathcal{L}(\beta, \sigma^2, Y_1, \dots, Y_n) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\|Y - \mathbb{X}\beta\|^2.$$

La fonction $(\beta, \sigma^2) \mapsto \ln \mathcal{L}(\beta, \sigma^2, Y_1, \dots, Y_n)$ admet un seul point critique $(\tilde{\beta}, \widetilde{\sigma^2})$ tel que

$$\tilde{\beta} = \hat{\beta} = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'Y, \text{ et } \widetilde{\sigma^2} = \frac{\|Y - \mathbb{X}\tilde{\beta}\|^2}{n} = \frac{\|\hat{\varepsilon}\|^2}{n} = \frac{n-p}{n}\widehat{\sigma^2}.$$

On vérifie facilement que ce point critique correspond bien à un maximum.

On remarque que l'estimateur du maximum de vraisemblance de β est égal à l'estimateur des MCO $\hat{\beta}$ déterminé dans le chapitre précédent. Il est donc linéaire, sans biais de variance minimale parmi les estimateurs linéaires sans biais (théorème de Gauss Markov). En revanche, l'estimateur du maximum de vraisemblance $\widetilde{\sigma^2}$ est un estimateur biaisé de σ^2. En pratique, on préfère utiliser $\widehat{\sigma^2}$.

3.3 Lois des estimateurs

Pour déterminer la loi de ces estimateurs, on se sert d'un résultat général sur les vecteurs gaussiens : le théorème de Cochran.

Thèorème 11. *Sous les conditions (C_1) à (C_4), les estimateurs $\hat{\beta}$ et $\widehat{\sigma^2}$ vérifient :*

— *Pour toute matrice réelle M de taille $q \times p$ de rang q ($q \leq p$), alors*

$$M\hat{\beta} \sim \mathcal{N}\left(M\beta, \sigma^2\left[M(\mathbb{X}'\mathbb{X})^{-1}M'\right]\right),$$

et

$$\frac{1}{\sigma^2}\left[M(\hat{\beta}-\beta)\right]'\left[M(\mathbb{X}'\mathbb{X})^{-1}M'\right]^{-1}\left[M(\hat{\beta}-\beta)\right] \sim \chi^2(q).$$

— $(n-p)\widehat{\sigma^2}/\sigma^2 \sim \chi^2(n-p)$.

— *Les estimateurs $\hat{\beta}$ et $\widehat{\sigma^2}$ sont indépendants.*

Preuve : Application directe des propriétés de base des vecteurs gaussiens puis du théorème de Cochran.
On a : $\hat{\beta} = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'(\mathbb{X}\beta + \varepsilon) = \beta + (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\varepsilon$, d'où $M\hat{\beta} = M\beta + M(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\varepsilon$. Puisque $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ on a bien $M\hat{\beta} \sim \mathcal{N}(M\beta, \sigma^2 M(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'[M(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}']') = \mathcal{N}(M\beta, \sigma^2 M(\mathbb{X}'\mathbb{X})^{-1}M')$, puis
$\left[M(\hat{\beta}-\beta)\right]'\left[\sigma^2 M(\mathbb{X}'\mathbb{X})^{-1}M'\right]^{-1}\left[M(\hat{\beta}-\beta)\right] \sim \chi^2(q)$.

Pour appliquer le théorème de Cochran, on remarque que $\hat{\beta} = \beta + (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\varepsilon = \beta + (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\Pi_{\mathbb{X}}\varepsilon$, et $(n-p)\widehat{\sigma^2} = \|\Pi_{\mathbb{X}^\perp}Y\|^2 = \|\Pi_{\mathbb{X}^\perp}(\mathbb{X}\beta+\varepsilon)\|^2 = \|\Pi_{\mathbb{X}^\perp}\varepsilon\|^2$. Le vecteur des bruits ε vérifie les hypothèses du théorème de Cochran ($\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$), donc on obtient bien la loi de $(n-p)\widehat{\sigma^2}/\sigma^2$ et l'indépendance de $\hat{\beta}$ et $\widehat{\sigma^2}$.

Le théorème 11 permettra de dégager des fonctions pivotales pour la construction d'intervalles ou de régions de confiance pour les paramètres du modèle, et des statistiques de test sur ces paramètres. Plus précisément, on utilisera les deux corollaires suivants.

Corollaire 1. *Sous les conditions* (C_1) *à* (C_4)*, pour tout* $j = 0 \dots p-1$*,*

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\widehat{\sigma^2}\left[(\mathbb{X}'\mathbb{X})^{-1}\right]_{j+1,j+1}}} \sim \mathcal{T}(n-p).$$

Preuve. Il suffit de prendre la matrice $M = (0, \dots, 0, 1, 0, \dots, 0)$ dont tous les éléments sont nuls, sauf le $(j+1)$ème élément qui est égal à 1. Alors $M(\mathbb{X}'\mathbb{X})^{-1}M' = \left[(\mathbb{X}'\mathbb{X})^{-1}\right]_{j+1,j+1}$.

Corollaire 2. *Sous les conditions* (C_1) *à* (C_4)*, si* M *est une matrice réelle de taille* $q \times p$ *de rang* q *(*$q \leq p$*),*

$$\frac{1}{q\widehat{\sigma^2}}\left[M(\hat{\beta}-\beta)\right]'\left[M(\mathbb{X}'\mathbb{X})^{-1}M'\right]^{-1}\left[M(\hat{\beta}-\beta)\right] \sim \mathcal{F}(q, n-p).$$

3.4 Intervalles et régions de confiance pour les paramètres - Intervalles de prédiction

3.4.1 Intervalles et régions de confiance pour les cœfficients de régression

On peut à partir du Corollaire 1 construire un intervalle de confiance bilatère pour le cœfficient β_j :

Proposition 5. *Soit* $\alpha \in]0,1[$*. On note* $t_{n-p}(u)$ *le u-quantile de la loi* $\mathcal{T}(n-p)$*.*
Un intervalle de confiance de niveau de confiance $(1-\alpha)$ *pour* β_j *(*$j = 0 \dots p-1$*) est donné par*

$$\hat{I}_j = \left[\hat{\beta}_j - t_{n-p}(1-\alpha/2)\sqrt{\widehat{\sigma^2}\left[(\mathbb{X}'\mathbb{X})^{-1}\right]_{j+1,j+1}}; \hat{\beta}_j + t_{n-p}(1-\alpha/2)\sqrt{\widehat{\sigma^2}\left[(\mathbb{X}'\mathbb{X})^{-1}\right]_{j+1,j+1}}\right].$$

On peut ensuite à partir du Corollaire 2 construire une région de confiance pour $m = M\beta$, où M est une matrice réelle de taille $q \times p$ de rang q ($q \leq p$).

Proposition 6. *Soit* $\alpha \in]0,1[$*. On note* $f_{q,n-p}(u)$ *le u-quantile de la loi* $\mathcal{F}(q, n-p)$*.*
Une région de confiance de niveau de confiance $(1-\alpha)$ *pour* $m = M\beta$ *est donnée par :*

$$\hat{I}^{(M)} = \left\{ m \in \overset{q}{\mathbb{R}}, \frac{1}{q\widehat{\sigma^2}}\left[M\hat{\beta}-m\right]'\left[M(\mathbb{X}'\mathbb{X})^{-1}M'\right]^{-1}\left[M\hat{\beta}-m\right] \leq f_{q,n-p}(1-\alpha)\right\}.$$

Exemples fondamentaux : région de confiance simultanée pour β (ellipsoïde de confiance) ou un "sous-vecteur" de β (ellipse de confiance pour un couple de cœfficients de régression).

3.4.2 Intervalles de confiance pour la variance σ^2

Proposition 7. *Soit $\alpha \in]0,1[$. On note $c_{n-p}(u)$ le u-quantile de la loi $\chi^2(n-p)$.*
Un intervalle de confiance de niveau de confiance $(1-\alpha)$ pour σ^2 est donné par $\left[\frac{(n-p)\widehat{\sigma^2}}{c_{n-p}(1-\alpha/2)}; \frac{(n-p)\widehat{\sigma^2}}{c_{n-p}(\alpha/2)}\right]$.

3.4.3 Intervalles de prédiction

A partir d'une nouvelle valeur explicative $x_{n+1} = (x_{n+1,0}, \ldots, x_{n+1,p-1})$, on souhaite prédire une nouvelle observation d'une variable $Y_{n+1} = \beta_0 x_{n+1,0} + \ldots + \beta_{p-1} x_{n+1,p-1} + \varepsilon_{n+1} = x_{n+1}\beta + \varepsilon_{n+1}$, avec $\varepsilon_{n+1} \sim \mathcal{N}(0, \sigma^2)$ et ε_{n+1} indépendante des ε_i, $i = 1 \ldots n$, i.e. Y_{n+1} indépendante des Y_i, $i = 1 \ldots n$, utilisées pour construire $\hat{\beta}$.

On introduit $\hat{Y}^p_{n+1} = x_{n+1}\hat{\beta} = \sum_{j=0}^{p-1} x_{n+1,j}\hat{\beta}_j$.

L'erreur de prédiction définie par $\hat{\varepsilon}^p_{n+1} = Y_{n+1} - \hat{Y}^p_{n+1}$ vérifie alors : $\hat{\varepsilon}^p_{n+1} \sim \mathcal{N}(0, \sigma^2(1 + x_{n+1}(\mathbb{X}'\mathbb{X})^{-1}x'_{n+1}))$.

Thèorème 12. *Un intervalle de prédiction pour Y_{n+1} de niveau de confiance $(1-\alpha)$ est donné par*

$$\widehat{I^p_{n+1}} = \left[\hat{Y}^p_{n+1} - t_{n-p}(1-\alpha/2)\sqrt{\widehat{\sigma^2}(1 + x_{n+1}(\mathbb{X}'\mathbb{X})^{-1}x'_{n+1})}; \hat{Y}^p_{n+1} + t_{n-p}(1-\alpha/2)\sqrt{\widehat{\sigma^2}(1 + x_{n+1}(\mathbb{X}'\mathbb{X})^{-1}x'_{n+1})}\right].$$

3.5 Tests d'hypothèses sur les cœfficients de régression

Le niveau des différents tests présentés ici est fixé à $\alpha \in]0,1[$.

3.5.1 Test de nullité d'un cœfficient ou de (non) significativité d'une variable explicative

On souhaite tester l'hypothèse nulle $(H_0) : \beta_j = 0$ contre l'alternative $(H_1) : \beta_j \neq 0$ pour $j \in \{0, \ldots, p-1\}$.
A partir du Corollaire 1, on construit une statistique de test :

$$T(Y) = \frac{\hat{\beta}_j}{\sqrt{\widehat{\sigma^2}\left[(\mathbb{X}'\mathbb{X})^{-1}\right]_{j+1,j+1}}},$$

qui suit sous l'hypothèse (H_0) la loi $\mathcal{T}(n-p)$.
On peut alors prendre comme région de rejet ou région critique :

$$\mathcal{R}_{(H_0)} = \{y, |T(y)| \geq t_{n-p}(1-\alpha/2)\}.$$

Il est aussi possible d'utiliser directement les intervalles de confiance construits ci-dessus pour retrouver ce résultat : si $0 \notin \hat{I}_j$, on rejette l'hypothèse $(H_0) : \beta_j = 0$ au profit de l'alternative $(H_1) : \beta_j \neq 0$.

3.5.2 Tests d'hypothèses linéaires sur les cœfficients

Plus généralement, si l'on souhaite tester l'hypothèse nulle $(H_0) : M\beta = m$ contre l'alternative $(H_1) : M\beta \neq m$, on peut, d'après le corollaire 2, utiliser comme statistique de test :

$$F(Y) = \frac{1}{q\widehat{\sigma^2}} \left[M\hat{\beta} - m\right]' \left[M(\mathbb{X}'\mathbb{X})^{-1}M'\right]^{-1} \left[M\hat{\beta} - m\right],$$

qui suit sous l'hypothèse (H_0) la loi $\mathcal{F}(q, n-p)$.
On prend alors comme région de rejet ou région critique :

$$\mathcal{R}_{(H_0)} = \{y, F(y) > f_{q,n-p}(1-\alpha)\}.$$

Ce résultat se retrouve aussi directement à partir de la région de confiance construite ci-dessus : si $m \notin \hat{I}^{(M)}$, on rejette l'hypothèse $(H_0) : M\beta = m$ au profit de l'alternative $(H_1) : M\beta \neq m$.

Exemples :

— Test de nullité d'un cœfficient ou test de (non) significativité d'une variable explicative. En prenant la matrice $M = (0 \dots 0\ 1\ 0 \dots 0)$ égale au vecteur ligne ne contenant que des 0 sauf le $(j+1)$ème élément qui vaut 1, on obtient la statistique :

$$F(Y) = \frac{\hat{\beta}_j^2}{\widehat{\sigma^2}[(\mathbb{X}'\mathbb{X})^{-1}]_{j+1,j+1}} = T(Y)^2,$$

sachant qu'une loi de Fisher à $(1, n-p)$ degrés de liberté est exactement la loi du carré d'une variable de loi de Student à $n-p$ degrés de liberté.
On retouve donc précisément le test introduit dans le paragraphe précédent.

— Test de nullité simultanée de plusieurs cœfficients c'est-à-dire de $(H_0) : \beta_{j_1} = \dots = \beta_{j_q} = 0$ contre (H_1) : il existe $k \in \{1, \dots, q\}$ tel que $\beta_{j_k} \neq 0$ $(j_1 \leq \dots \leq j_q)$, ou test de validité du sous-modèle

$$Y = \mathbb{X}^{(p-q)}\beta^{(p-q)} + \varepsilon^{(p-q)},$$

où $\mathbb{X}^{(p-q)}$ est la matrice formée des $p-q$ vecteurs colonnes de $\mathbb{X}$ dont on aura retiré les colonnes $j_1, \dots, j_q$, $\beta^{(p-q)}$ est un vecteur de $\mathbb{R}^{p-q}$, et où $\varepsilon^{(p-q)} \sim \mathcal{N}(0, \sigma^2 I_n)$.
Choix de la matrice M : $M_{k,l} = \delta^l_{j_k}$.

— Dans le cas où $x_{i,0} = 1$ pour tout i, test de $(H_0) : \beta_1 = \dots = \beta_{p-1} = 0$ contre (H_1) : il existe $k \in \{1, \dots, p-1\}$ tel que $\beta_k \neq 0$, ou test de validité globale du modèle (complet).
On choisit :

$$M = \begin{pmatrix} 0 & 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & 0 & \dots & 0 \\ & & & \dots & & \\ 0 & 0 & 0 & 0 & \dots & 1 \end{pmatrix}.$$

3.5.3 Test du rapport de vraisemblance maximale

On montre ici que le test construit intuitivement dans le paragraphe précédent correspond en fait au test du rapport de vraisemblance maximale.

Thèorème 13. *- Le test du rapport de vraisemblance maximale de niveau α de l'hypothèse nulle (H_0) : $M\beta = m$ contre l'alternative (H_1) : $M\beta \neq m$ a comme région critique l'ensemble $\mathcal{R}_{(H_0)} = \{y, \tilde{F}(y) > f_{q,n-p}(1-\alpha)\}$, avec*

$$\tilde{F}(Y) = \frac{\|\mathbb{X}\hat{\beta} - \mathbb{X}\hat{\beta}^{(M)}\|^2/q}{\|Y - \mathbb{X}\hat{\beta}\|^2/(n-p)} = \frac{\left(\|Y - \mathbb{X}\hat{\beta}^{(M)}\|^2 - \|Y - \mathbb{X}\hat{\beta}\|^2\right)/q}{\|Y - \mathbb{X}\hat{\beta}\|^2/(n-p)}. \tag{3.1}$$

$\hat{\beta}^{(M)}$ est l'estimateur des moindres carrés ordinaires de β sous la contrainte $M\beta = m$. Il est donné par

$$\hat{\beta}^{(M)} = \hat{\beta} + (\mathbb{X}'\mathbb{X})^{-1}M'\left[M(\mathbb{X}'\mathbb{X})^{-1}M'\right]^{-1}(m - M\hat{\beta}).$$

- La statistique du test du rapport de vraisemblance maximale $\tilde{F}(Y)$ définie par (3.1) est égale à la statistique de test $F(Y)$. Le test du rapport de vraisemblance maximale est donc équivalent au test construit intuitivement.

Preuve. On considère le test du rapport de vraisemblance maximale de l'hypothèse nulle $(H_0) : M\beta = m$ contre l'alternative $(H_1) : M\beta \neq m$, basé sur la statistique de test :

$$\rho(Y) = \frac{\sup_{(\beta,\sigma^2)\in\mathbb{R}^p\times\mathbb{R}_+, M\beta=m}\mathcal{L}(\beta,\sigma^2,Y_1,\ldots,Y_n)}{\sup_{(\beta,\sigma^2)\in\mathbb{R}^p\times\mathbb{R}_+}\mathcal{L}(\beta,\sigma^2,Y_1,\ldots,Y_n)}.$$

On a d'après ce qui précède,

$$\sup_{(\beta,\sigma^2)\in\mathbb{R}^p\times\mathbb{R}_+}\mathcal{L}(\beta,\sigma^2,Y_1,\ldots,Y_n) = \mathcal{L}(\hat{\beta},\widetilde{\sigma^2},Y_1,\ldots,Y_n) = \left(\frac{n}{2\pi\|Y-\mathbb{X}\hat{\beta}\|^2}\right)^{n/2}\exp\left(-\frac{n}{2}\right).$$

Pour déterminer $\sup_{(\beta,\sigma^2)\in\mathbb{R}^p\times\mathbb{R}_+, M\beta=m}\mathcal{L}(\beta,\sigma^2,Y_1,\ldots,Y_n)$, on minimise la fonction $\beta \mapsto \|Y - \mathbb{X}\beta\|^2$ sous la contrainte $M\beta = m$.
Le Lagrangien du problème s'écrit :

$$\Lambda(\beta,\lambda) = \|Y - \mathbb{X}\beta\|^2 - \lambda(M\beta - m).$$

Ce Lagrangien admet un unique point critique $(\hat{\beta}^{(M)}, \hat{\lambda}^{(M)})$ vérifiant :

$$\begin{cases} -2\mathbb{X}'Y + 2\mathbb{X}'\mathbb{X}\hat{\beta}^{(M)} - M'\hat{\lambda}^{(M)} = 0 \\ M\hat{\beta}^{(M)} - m = 0 \end{cases}$$

En multipliant la première égalité à gauche par $M(\mathbb{X}'\mathbb{X})^{-1}$, on obtient :

$$-2M(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'Y + 2M\hat{\beta}^{(M)} - M(\mathbb{X}'\mathbb{X})^{-1}M'\hat{\lambda}^{(M)} = 0,$$

d'où

$$-2M(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'Y + 2m - M(\mathbb{X}'\mathbb{X})^{-1}M'\hat{\lambda}^{(M)} = 0,$$

puis

$$\hat{\lambda}^{(M)} = 2\left[M(\mathbb{X}'\mathbb{X})^{-1}M'\right]^{-1}\left(m - M(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'Y\right).$$

On obtient finalement à partir de la première équation l'expression de $\hat{\beta}^{(M)}$:

$$\hat{\beta}^{(M)} = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'Y + (\mathbb{X}'\mathbb{X})^{-1}M'\left[M(\mathbb{X}'\mathbb{X})^{-1}M'\right]^{-1}(m - M(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'Y),$$

ou encore :

$$\hat{\beta}^{(M)} = \hat{\beta} + (\mathbb{X}'\mathbb{X})^{-1}M'\left[M(\mathbb{X}'\mathbb{X})^{-1}M'\right]^{-1}(m - M\hat{\beta}).$$

On a enfin

$$\sup_{(\beta,\sigma^2)\in\mathbb{R}^p\times\mathbb{R}_+, M\beta=m}\mathcal{L}(\beta,\sigma^2,Y_1,\ldots,Y_n) = \left(\frac{n}{2\pi\|Y-\mathbb{X}\hat{\beta}^{(M)}\|^2}\right)^{n/2}\exp\left(-\frac{n}{2}\right).$$

La statistique du test du rapport de vraisemblance maximale s'écrit ainsi :

$$\rho(Y) = \left(\frac{\|Y-\mathbb{X}\hat{\beta}\|^2}{\|Y-\mathbb{X}\hat{\beta}^{(M)}\|^2}\right)^{n/2}.$$

Le test du rapport de vraisemblance maximale de (H_0) contre (H_1) a une région critique de la forme $\{y, \rho(y) < \rho_\alpha\}$, ce qui est équivalent à rejeter (H_0) lorsque

$$\frac{(\|Y-\mathbb{X}\hat{\beta}^{(M)}\|^2 - \|Y-\mathbb{X}\hat{\beta}\|^2)/q}{\|Y-\mathbb{X}\hat{\beta}\|^2/(n-p)} > \rho'_\alpha.$$

Il s'agit maintenant de montrer que $\|Y-\mathbb{X}\hat{\beta}^{(M)}\|^2 - \|Y-\mathbb{X}\hat{\beta}\|^2 = \|\mathbb{X}\hat{\beta}-\mathbb{X}\hat{\beta}^{(M)}\|^2$, puis que

$$\frac{\|\mathbb{X}\hat{\beta}-\mathbb{X}\hat{\beta}^{(M)}\|^2/q}{\|Y-\mathbb{X}\hat{\beta}\|^2/(n-p)} = \frac{1}{q\widehat{\sigma^2}}\left[M\hat{\beta}-m\right]'\left[M(\mathbb{X}'\mathbb{X})^{-1}M'\right]^{-1}\left[M\hat{\beta}-m\right].$$

On considère le sous-espace affine (vectoriel si $m = 0$) de $\mathcal{E}(\mathbb{X})$ défini par $\mathcal{E}^{(M)} = \{\mathbb{X}\beta, M\beta = m\}$. On a alors la décomposition suivante :

$$\mathcal{E}^{(M)\perp} = \mathcal{E}(\mathbb{X})^\perp \overset{\perp}{\bigoplus} \left(\mathcal{E}(\mathbb{X}) \cap \mathcal{E}^{(M)\perp}\right).$$

Puisque $\hat{\beta}^{(M)}$ minimise la fonction $\beta \mapsto \|Y-\mathbb{X}\beta\|^2$ sous la contrainte $M\beta = m$, $\mathbb{X}\hat{\beta}^{(M)}$ est le projeté orthogonal de Y sur $\mathcal{E}^{(M)}$. On a alors : $Y - \mathbb{X}\hat{\beta}^{(M)} \in \mathcal{E}^{(M)\perp}$. De plus, $Y - \mathbb{X}\hat{\beta}$ est le projeté orthogonal de Y sur $\mathcal{E}(\mathbb{X})^\perp$ et $\mathbb{X}\hat{\beta} - \mathbb{X}\hat{\beta}^{(M)}$ appartient à $\mathcal{E}(\mathbb{X}) \cap \mathcal{E}^{(M)\perp}$ ($\mathbb{X}\hat{\beta}^{(M)}$ est aussi le projeté orthogonal de $\mathbb{X}\hat{\beta}$ sur $\mathcal{E}^{(M)}$.

$\hookrightarrow$ Représentation géométrique dans le cas $m = 0$.

Par le théorème de Pythagore, on a donc bien $\|Y-\mathbb{X}\hat{\beta}^{(M)}\|^2 = \|Y-\mathbb{X}\hat{\beta}\|^2 + \|\mathbb{X}\hat{\beta}-\mathbb{X}\hat{\beta}^{(M)}\|^2$ c'est-à-dire : $\|Y-\mathbb{X}\hat{\beta}^{(M)}\|^2 - \|Y-\mathbb{X}\hat{\beta}\|^2 = \|\mathbb{X}\hat{\beta}-\mathbb{X}\hat{\beta}^{(M)}\|^2$.

Par ailleurs, en reprenant l'expression matricielle de $\hat{\beta}^{(M)}$, on obtient que

$\|\mathbb{X}\hat{\beta} - \mathbb{X}\hat{\beta}^{(M)}\|^2 = \|\mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}M'\left[M(\mathbb{X}'\mathbb{X})^{-1}M'\right]^{-1}(m - M\hat{\beta})\|^2$, d'où

$\|\mathbb{X}\hat{\beta}-\mathbb{X}\hat{\beta}^{(M)}\|^2 = (m-M\hat{\beta})'\left[M(\mathbb{X}'\mathbb{X})^{-1}M'\right]^{-1}(m-M\hat{\beta})$, et on remarque que $\widehat{\sigma^2} = \|Y-\mathbb{X}\hat{\beta}\|^2/(n-p)$ pour conclure.

Enfin, la loi de $F(Y)$ sous (H_0) étant donnée par le Corollaire 2, le test du rapport de vraisemblance maximale est bien équivalent au test construit intuitivement.

Retour sur le test de validité d'un sous-modèle

On revient sur le problème de test de nullité simultanée de q cœfficients de régression i.e. $(H_0) : \beta_{j_1} = \ldots = \beta_{j_q} = 0$ contre (H_1) : il existe $k \in \{1, \ldots, q\}$ tel que $\beta_{j_k} \neq 0$, avec $j_1 \leq \ldots j_q$, ou de validité du sous-modèle

$$Y = \mathbb{X}^{(p-q)}\beta^{(p-q)} + \varepsilon^{(p-q)}.$$

D'après le théorème précédent, la statistique de test $F(Y)$ peut s'écrire également :

$$F(Y) = \frac{(SCR^{(M)} - SCR)/q}{SCR/(n-p)},$$

où $SCR^{(M)}$ est la somme des carrés résiduelle dans le sous-modèle, SCR la somme des carrés résiduelle dans le modèle complet.
Si $x_{i,0} = 1$ pour tout i, et $j_1 \geq 1$, on peut aussi écrire :

$$F(Y) = \frac{(R^2 - (R^2)^{(M)})/q}{(1-R^2)/(n-p)},$$

où $(R^2)^{(M)}$ est le cœfficient de détermination dans le sous-modèle.

Retour sur le test de validité globale du modèle

On revient sur le problème du test de validité globale du modèle dans le cas où $x_{i,0} = 1$ pour tout i, c'est-à-dire le problème de test de $(H_0) : \beta_1 = \ldots = \beta_{p-1} = 0$ contre (H_1) : il existe $k \in \{1, \ldots, p-1\}$ tel que $\beta_k \neq 0$.
On a choisi

$$M = \begin{pmatrix} 0 & 1 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 1 & 0 \ldots & 0 & \\ & & & \ldots & & \\ 0 & 0 & 0 & 0 & \ldots & 1 \end{pmatrix}.$$

Dans ce cas, $\mathbb{X}\hat{\beta}^{(M)} = \bar{Y}\mathbb{1}$, donc $F(Y)$ s'écrit aussi :

$$F(Y) = \frac{\|\hat{Y} - \bar{Y}\mathbb{1}\|^2/(p-1)}{\|Y - \hat{Y}\|^2/(n-p)} = \frac{R^2}{1-R^2}\frac{n-p}{p-1},$$

et suit sous (H_0) la loi $\mathcal{F}(p-1, n-p)$.

On remarque ici à nouveau l'intérêt de l'introduction du cœfficient de détermination R^2 (et on retrouve l'interprétation des cas limites).

3.6 Exercices

Exercice 1 : Questions de cours

On considère le modèle de régression linéaire multiple

$$Y = \mathbb{X}\beta + \varepsilon,$$

où le vecteur Y à valeurs dans $\mathbb{R}^n$ représente la variable à expliquer, $\mathbb{X}$ est une matrice réelle de taille $n \times p$ de rang p, $\beta = (\beta_0, \ldots, \beta_{p-1})' \in \mathbb{R}^p$ (inconnu) et ε est le vecteur des bruits aléatoires à valeurs dans $\mathbb{R}^n$.

On pose pour le vecteur des bruits ε les conditions standards du modèle de régression linéaire multiple, ainsi qu'une hypothèse gaussienne.

1. Comment cette hypothèse gaussienne se formule-t-elle ?
2. Quelles conditions impose-t-elle sur les variables ε_i et Y_i ?
3. Les estimateurs du maximum de vraisemblance de β et σ^2 sont-ils sans biais ? Expliquer.
4. Quelle est la loi de ces estimateurs ? Sont-ils indépendants ?
5. Le produit des intervalles de confiance (individuels) de niveau de confiance $1 - \alpha$ pour les β_j $(j = 0 \ldots p - 1)$ correspond-il à la région de confiance simultanée de niveau de confiance $(1 - \alpha)$ pour β ? Justifier la réponse.
6. Montrer que le test de Student de nullité d'un cœfficient de régression β_j est équivalent au test du rapport de vraisemblance maximale correspondant.
7. Montrer que la statistique du test du rapport de vraisemblance maximale permettant de tester la validité globale du modèle peut s'écrire en fonction du cœfficient de détermination R^2.
8. Peut-on tester la validité de n'importe quel sous-modèle à partir des intervalles de confiance individuels pour les β_j $(j = 0 \ldots p - 1)$?
9. Peut-on tester la validité d'un sous-modèle à partir des valeurs des cœfficients de détermination calculés sur les observations dans le modèle complet et dans le sous-modèle ? Expliquer.
10. Imaginer une procédure de sélection de variables explicatives basée sur les tests de validité de sous-modèles.

Exercice 2 : Théorème de Cochran (version simplifiée)

Soit $Y \sim \mathcal{N}(0, \sigma^2 I_n)$ et $\mathcal{E}$ un sous-espace vectoriel de $\mathbb{R}^n$ de dimension p. On note $\Pi_{\mathcal{E}}$ la matrice de projection orthogonale de $\mathbb{R}^n$ sur $\mathcal{E}$. Démontrer les propriétés suivantes :

1. $\Pi_{\mathcal{E}} Y \sim \mathcal{N}(0, \sigma^2 \Pi_{\mathcal{E}})$,
2. $\|\Pi_{\mathcal{E}} Y\|^2 / \sigma^2 \sim \chi^2(p)$,
3. $\|Y - \Pi_{\mathcal{E}} Y\|^2 / \sigma^2 \sim \chi^2(n - p)$,
4. Les vecteurs $\Pi_{\mathcal{E}} Y$ et $Y - \Pi_{\mathcal{E}} Y$ sont indépendants.

Expliquer en quoi le théorème de Cochran est central en modèle de régression linéaire multiple sous hypothèse gaussienne.

Exercice 3 : Variables "centrées"

On considère le modèle de régression suivant :

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \varepsilon_i, \qquad 1 \leq i \leq n.$$

où les $x_{i,j}$, $j = 1, 2, 3$, sont déterministes, et le vecteur des ε_i est un vecteur gaussien centré de matrice de variance covariance $\sigma^2 I_n$.
En posant :

$$\mathbb{X} = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} & x_{1,3} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n,1} & x_{n,2} & x_{n,3} \end{pmatrix} \quad \text{et} \quad Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix},$$

et en notant y l'observation de Y, on a

$$\mathbb{X}'\mathbb{X} = \begin{pmatrix} 50 & 0 & 0 & 0 \\ 0 & 20 & 15 & 4 \\ 0 & 15 & 30 & 10 \\ 0 & 4 & 10 & 40 \end{pmatrix}, \quad \mathbb{X}'y = \begin{pmatrix} 100 \\ 50 \\ 40 \\ 80 \end{pmatrix}, \quad y'y = 640.$$

On admettra que

$$\begin{pmatrix} 20 & 15 & 4 \\ 15 & 30 & 10 \\ 4 & 10 & 40 \end{pmatrix}^{-1} = \frac{1}{13720} \begin{pmatrix} 1100 & -560 & 30 \\ -560 & 784 & -140 \\ 30 & -140 & 375 \end{pmatrix}.$$

1. Donner la valeur de n.
2. Interpréter les 0 de la matrice $\mathbb{X}'\mathbb{X}$.
3. Estimer les paramètres β_0, β_1, β_2, β_3 par la méthode du maximum de vraisemblance et donner un estimateur sans biais de σ^2. Quelle est la loi des estimateurs obtenus ? Ces estimateurs sont-ils indépendants ?
4. Donner les valeurs de ces estimateurs calculés sur les observations.
5. Donner un intervalle de confiance de niveau de confiance 95% pour σ^2.
6. Tester la validité globale du modèle au niveau 5%.
7. Construire un test de niveau 5% de l'hypothèse $(H_0) : \beta_3 = 0$ contre $(H_1) : \beta_3 \neq 0$ de deux façons différentes. Que peut-on conclure ?
8. Construire un test de niveau 5% de l'hypothèse $(H_0) : \beta_3 = -4\beta_2$ contre $(H_1) : \beta_3 \neq -4\beta_2$ de deux façons différentes. Que peut-on conclure ?
9. On suppose que l'on dispose de nouvelles valeurs $x_{n+1,1} = 1$, $x_{n+1,2} = -1$ et $x_{n+1,3} = 0.5$. Donner un intervalle de prédiction de niveau de confiance 95% pour la variable Y_{n+1} telle que $Y_{n+1} = \beta_0 + \beta_1 x_{n+1,1} + \beta_2 x_{n+1,2} + \beta_3 x_{n+1,3} + \varepsilon_{n+1}$, avec $(\varepsilon_1, \ldots, \varepsilon_{n+1}) \sim \mathcal{N}(0, \sigma^2 I_{n+1})$.

Exercice 4 : Régression polynômiale

On considère le modèle de régression polynômiale suivant :

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i, \qquad 1 \leq i \leq n,$$

où les x_i sont déterministes, et les ε_i sont des variables aléatoires i.i.d. de loi gaussienne centrée de variance σ^2.
On observe les valeurs :

i	1	2	3	4	5	6	7	8	9	10	11
x_i	−5	−4	−3	−2	−1	0	1	2	3	4	5
y_i	−3.37	−2.11	−2.24	1.59	3.28	3.96	6.42	8.57	10.71	14.32	15.91

1. Donner une condition simple pour que le modèle soit identifiable.
2. Tester la validité globale du modèle au niveau 5%.
3. Construire un intervalle de confiance de niveau de confiance 95% pour β_2.
4. Déterminer la p-valeur du test de nullité de β_2. Le résultat est-il cohérent avec celui de la question précédente ?

Exercice 5 : Régression périodique

On considère le modèle de régression périodique suivant :

$$Y_i = \beta_0 x_i + \beta_1 \cos(x_i) + \varepsilon_i, \qquad -n \leq i \leq n \ (n \geq 1),$$

où $x_i = i\pi/n$ et les ε_i sont des variables aléatoires i.i.d. de loi gaussienne centrée de variance σ^2.

1. Vérifier que

$$\sum_{i=-n}^{n} \cos^2(x_i) = n + 1 \quad \text{et} \quad \sum_{i=-n}^{n} x_i^2 = \frac{\pi^2(n+1)(2n+1)}{3n}.$$

2. Déterminer les estimateurs du maximum de vraisemblance de β_0 et β_1, puis un estimateur sans biais de σ^2. Quelle est la loi de ces estimateurs ? Sont-il indépendants ?
3. Construire un test de niveau α de l'hypothèse nulle $(H_0) : \beta_0 = 0$ contre l'alternative $(H_1) : \beta_0 > 0$.
4. Déterminer le test du rapport de vraisemblance maximale de niveau α de l'hypothèse $(H_0) : \beta_0 - 4\beta_1 = 1$ contre $(H_1) : \beta_0 - 4\beta_1 \neq 1$.

Exercice 6 : Régression puissance

On cherche à expliquer la température absolue (en K) d'un gaz en équilibre thermodynamique en fonction de la pression (en Pa) et du volume occupé (en m^3) par une mole de molécules de ce gaz. On considère pour cela un modèle de la forme :

$$T_i = e^{\beta_0} P_i^{\beta_1} V_i^{\beta_2} \eta_i,$$

où T_i, P_i, V_i désignent respectivement les valeurs de la température, de la pression et du volume pour la $i^{\text{ème}}$ observation du gaz ($i \in \{1, \ldots, 50\}$), les η_i sont des termes d'erreur aléatoires tels que

— les η_i sont indépendants,
— $\mathbb{E}[\ln \eta_i] = 0$ pour tout i,
— $\text{var}(\ln \eta_i) = \sigma^2$ pour tout i.

Ce modèle peut s'écrire sous la forme d'un modèle de régression linéaire multiple :

$$Y = \mathbb{X}\beta + \varepsilon,$$

avec $\beta = (\beta_0, \beta_1, \beta_2)'$.

Remarque : on considère ici que les variables explicatives P_i et V_i sont déterministes.

1. Préciser Y, $\mathbb{X}$, ε et les hypothèses faites sur ε. Ces hypothèses sont-elles standards ?
2. On suppose que $\mathbb{X}$ est de plein rang avec

$$(\mathbb{X}'\mathbb{X})^{-1} = \begin{pmatrix} 1.38783 & -0.23195 & -0.03336 \\ -0.23195 & 0.04608 & -0.01167 \\ -0.03336 & -0.01167 & 0.04535 \end{pmatrix}$$

et on a pour l'observation y de Y : $\mathbb{X}'y = (283.29, 1613.49, 644.11)'$, $SCR_{obs} = 10.42653$.

a) Donner la valeur de l'estimateur des moindres carrés ordinaires $\hat{\beta}$ de β calculé sur les observations.

b) Déterminer un estimateur $\widehat{\sigma^2}$ sans biais de la variance σ^2. En déduire un estimateur sans biais de la matrice de variance-covariance $\text{Var}(\hat{\beta})$ de β.

3. On suppose désormais que $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_{50})$.

a) Quels sont les estimateurs du maximum de vraisemblance de β et de σ^2 ?

b) Donner les lois de $\hat{\beta}$ et $\widehat{\sigma^2}$. Les estimateurs $\hat{\beta}$ et $\widehat{\sigma^2}$ sont-ils indépendants ?

c) Construire une région de confiance simultanée pour (β_1, β_2) de niveau de confiance 95%. Quelle forme cette région a-t-elle ?

d) L'équation d'équilibre thermodynamique des gaz parfaits s'écrit pour une mole de molécules de gaz :

$$T = \frac{1}{R} PV,$$

où $R = 8.314$. Les observations peuvent-elles provenir d'un gaz parfait ? Répondre à cette question à l'aide d'un test d'hypothèses de niveau 5%.

Exercice 7 : Données d'Air Breizh, concentration maximale en ozone dans l'air

On souhaite expliquer le maximum journalier de la concentration en ozone dans l'air à Rennes noté O à l'aide des variables explicatives suivantes : la température à 6h, la température à 9h, la température à 12h, la nébulosité à 6h, la nébulosité à 12h, et la projection du vent sur l'axe Est-Ouest notées respectivement T_6, T_9, T_{12}, N_6, N_{12} et V. On dispose pour cela de n observations, et on introduit le modèle de régression linéaire multiple :

$$O_i = \beta + \beta_{T_6} T_{6,i} + \beta_{T_9} T_{9,i} + \beta_{T_{12}} T_{12,i} + \beta_{N_6} N_{6,i} + \beta_{N_{12}} N_{12,i} + \beta_V V_i + \varepsilon_i \quad \text{pour } i \in \{1, \dots, n\},$$

sous les conditions standards et sous hypothèse gaussienne.

Question préliminaire. Écrire le modèle de façon matricielle et rappeler les conditions imposées aux variables O_i.

On a lancé sur les données fournies par Air Breizh une procédure de régression linéaire à l'aide du logiciel R (fonction `lm`), et on a obtenu la sortie suivante :

```
Coefficients:
             Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)   63.1193      3.5993      ???    < 2e-16   ***
T6            -1.6338         ???   -5.171   2.74e-07   ***
T9             0.2184      0.4863    0.449        ???   ???
T12            2.6335      0.3299      ???        ???   ???
N6            -0.1585      0.2482   -0.639      0.523
N12           -2.2416      0.3134   -7.153   1.49e-12   ***
V              1.1124      0.1624    6.850   1.18e-11   ***
-
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 16.97 on 1179 degrees of freedom
Multiple R-Squared: 0.488, Adjusted R-squared: 0.4854
F-statistic: ??? on ??? and ??? DF, p-value: < 2.2e-16
```

1. Définir si possible les éléments fournis par cette sortie R.
2. Compléter la sortie.
3. Tester la validité globale du modèle au niveau 5%.
4. Quelles sont les variables significatives au niveau 5% individuellement ?
5. Peut-on prédire le maximum journalier de la concentration en ozone et donner un intervalle de prédiction pour des nouvelles observations de T_6, T_9, T_{12}, N_6, N_{12} et V égales respectivement à 7, 10, 20, 0, 0, 1 ? Expliquer.

Exercice 8 : Données du Cirad, hauteur des eucalyptus

On souhaite expliquer la longueur du tronc d'un eucalyptus (en m) d'une parcelle plantée à partir de sa circonférence à 1m30 du sol (en cm). On dispose pour cela de n observations, et on introduit le modèle de régression linéaire multiple :

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 \sqrt{x_i} + \varepsilon_i \quad \text{pour } i \in \{1, \dots, n\},$$

où Y_i représente la hauteur du tronc du $i$ème eucalyptus, et x_i sa circonférence à 1m30 du sol, sous les conditions standards et sous hypothèse gaussienne.

Question préliminaire. Ecrire le modèle de façon matricielle et rappeler les conditions imposées aux variables Y_i.

On a lancé sur les données fournies par le Cirad, à l'aide du logiciel SAS (`PROC REG`), une procédure de régression linéaire multiple, puis deux procédures de régression linéaires simples, avec comme seule variable explicative : soit la circonférence à 1m30, soit la racine carrée de cette circonférence. On a obtenu les sorties fournies en Annexe 1.2.

Que peut-on conclure de ces sorties ?

Chapitre 4

Détection (et correction) des écarts au modèle

Analyse des résidus, effet levier et mesure d'influence

4.1 Introduction

On rappelle l'expression vectorielle du modèle de régression linéaire multiple ($p \leq n$) :

$$Y = \mathbb{X}\beta + \varepsilon,$$

avec

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbb{X} = \begin{pmatrix} x_{1,0} & \dots & x_{1,p-1} \\ \vdots & & \vdots \\ x_{n,0} & \dots & x_{n,p-1} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_{p-1} \end{pmatrix} \quad \text{et} \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

La matrice du plan d'expérience $\mathbb{X}$ est supposée de plein rang, c'est-à-dire $\text{rang}(\mathbb{X}) = p$. Cela implique en particulier que la matrice symétrique $\mathbb{X}'\mathbb{X}$ est définie positive.

On rappelle aussi les conditions standards imposées au vecteur des bruits ε :

— (C_1) : $\mathbb{E}[\varepsilon] = 0$ (centrage),
— (C_2) : $\text{cov}(\varepsilon_i, \varepsilon_j) = 0$ pour tout $i \neq j$ (non corrélation),
— (C_3) : $\text{var}(\varepsilon_i) = \sigma^2$ (inconnue) pour tout $i = 1 \dots n$ (homoscédasticité),

ainsi que l'hypothèse gaussienne :

— (C_4) : le vecteur ε suit une loi gaussienne.

Dans ce chapitre, on s'interroge dans un premier temps sur la pertinence du modèle de façon globale. Plus précisément, on se pose les questions fondamentales suivantes :

1. La relation $\mathbb{E}[Y] = \mathbb{X}\beta$ est-elle bien vérifiée (choix des variables explicatives, linéarité, bruits centrés i.e. (C_1) satisfaite) ?
2. Les bruits sont-ils bien non-corrélés ((C_2) satisfaite) ?
3. Les bruits sont-ils bien homoscédastiques ((C_3) satisfaite) ?
4. Le vecteur des bruits est-il bien de loi gaussienne ((C_4) satisfaite) ?

Remarque : on ne se penche pas ici sur la pertinence de l'hypothèse que $\mathbb{X}$ est de plein rang, qui est vérifiable directement. On peut d'ailleurs remédier au problème dans le cas contraire ou dans un cas "presque" contraire ↪ régressions biaisées - ridge, LASSO, régression sur composantes principales, PLS, etc.

Dans un second temps, on se pose la question suivante : y a-t-il des observations "remarquables" ou "suspectes", mettant éventuellement en cause le modèle, ou en tout cas sur lesquelles l'attention devra se porter précisément ?

Etapes essentielles de la mise en œuvre d'une procédure de régression linéaire, principalement basées sur des méthodes graphiques : peu de règles strictes ⇒ regard permanent sur les données et leur signification indispensable.

4.2 Analyse des résidus

Les hypothèses du modèle portant essentiellement sur les bruits qui ne sont pas observés, il est naturel d'étudier les variables qui sont censées les approcher et qui sont calculables sur les observations : les résidus.

Rappel. Propriétés des résidus sous les hypothèses (C_1) à (C_3) : $\mathbb{E}[\hat{\varepsilon}] = 0$, $\mathrm{Var}(\hat{\varepsilon}) = \sigma^2(I_n - \Pi_{\mathbb{X}})$. De plus, $\hat{\varepsilon}$ et $\hat{Y}$ sont non corrélés. Sous l'hypothèse (C_4), le vecteur $\hat{\varepsilon}$ suit en plus une loi gaussienne.

On notera dans ce cours H ("hat matrix") la matrice de projection $\Pi_{\mathbb{X}}$, et ses éléments $h_{i,k}$, $1 \leq i \leq n, 1 \leq k \leq n$.

4.2.1 Les différents résidus

Les résidus - tels quels - sont souvent utilisés pour détecter les écarts au modèle, abstraction faite de l'hétéroscédasticité de ces résidus. Ce n'est pas à recommander.

On introduit donc deux nouveaux types de résidus.

Définition 8. *On désigne par* résidus studentisés *les variables définies par :*

$$T_i = \frac{\hat{\varepsilon}_i}{\sqrt{\widehat{\sigma^2}(1 - h_{i,i})}}.$$

Remarque : les résidus studentisés ne sont pas indépendants (à cause de $\widehat{\sigma^2}$) et sont peu robustes à des erreurs grossières sur la $i$ème observation (à cause de l'estimation de σ^2 par $\widehat{\sigma^2}$). On préfère donc en général utiliser les résidus studentisés par validation croisée.

Définition 9. *On désigne par* résidus studentisés par validation croisée *les variables définies par :*

$$T_i^* = \frac{Y_i - x_i\widehat{\beta_{(i)}}}{\sqrt{\widehat{\sigma^2_{(i)}}(1 + x_i(\mathbb{X}'_{(i)}\mathbb{X}_{(i)})^{-1}x_i')}},$$

où x_i est le ième vecteur ligne de la matrice $\mathbb{X}$, $\mathbb{X}_{(i)}$ correspond à la matrice $\mathbb{X}$ dont on a supprimé la ième ligne, $\widehat{\beta_{(i)}}$ est l'estimateur des moindres carrés ordinaires de β et $\widehat{\sigma^2_{(i)}}$ est l'estimateur sans biais de la variance, obtenus tous deux après suppression de la ième observation.

Les propriétés de ces résidus sont énoncées dans la proposition suivante.

Proposition 8. *1. Les variables T_i^* s'obtiennent à partir des résidus :*

$$T_i^* = \frac{\hat{\varepsilon}_i}{\sqrt{\widehat{\sigma^2_{(i)}}(1-h_{i,i})}}.$$

2. Les variables T_i^* s'obtiennent à partir des résidus studentisés :

$$T_i^* = T_i \sqrt{\frac{n-p-1}{n-p-T_i^2}}.$$

3. Sous les hypothèses (C_1) à (C_4), si $\mathbb{X}_{(i)}$ est de rang p, alors les T_i^* sont i.i.d. de loi $\mathcal{T}(n-p-1)$.

Preuve. Pour la troisième propriété, on montre que T_i^* correspond à la statistique d'un test de validité d'un sous-modèle.

Remarque : La deuxième propriété permet de voir que les fortes valeurs de T_i seront encore mieux repérées sur les T_i^*, et qu'en pratique, pour avoir t_i^*, on n'aura pas besoin de relancer une procédure d'estimation par moindres carrés ordinaires en supprimant la *i*ème observation.

On s'intéresse maintenant à la suppression éventuelle d'une variable explicative X_j. Pour cela, on définit la notion de résidu partiel. On note $\mathbb{X}^{(j)}$ la matrice obtenue après suppression dans $\mathbb{X}$ du $(j+1)$ème vecteur colonne, $\beta^{(j)}$ le vecteur obtenu après suppression dans β de l'élément β_j, et X_j le $(j+1)$ème vecteur colonne de $\mathbb{X}$.
On a alors :

$$Y = \mathbb{X}^{(j)}\beta^{(j)} + \beta_j X_j + \varepsilon,$$

d'où, si $\Pi_{\mathbb{X}^{(j)\perp}}$ désigne la matrice de projection orthogonale sur l'orthogonal de l'espace vectoriel engendré par les vecteurs colonnes de $\mathbb{X}^{(j)}$,

$$\Pi_{\mathbb{X}^{(j)\perp}} Y = \beta_j \Pi_{\mathbb{X}^{(j)\perp}} X_j + \Pi_{\mathbb{X}^{(j)\perp}} \varepsilon \quad \text{(régression partielle)}.$$

Définition 10. *On définit le résidu partiel $\hat{Y}^{(j)} = \hat{\varepsilon} + \hat{\beta}_j X_j$.*

4.2.2 Détection des écarts au modèle

On note t_i et t_i^* les observations respectives des variables T_i et T_i^*.
Diagnostics graphiques essentiellement.
- Oubli de variables explicatives ou non linéarité : tracés de $\hat{y}_i \mapsto t_i^*$ et/ou de $x_{i,j} \mapsto t_i^*$, tracé des t_i^* en fonction de la variable éventuellement oubliée, tracé des résidus partiels par rapport à la variable éventuellement oubliée ou de la régression partielle, utilisation des "lisseurs".

- Structure de la matrice de variance covariance (corrélation, hétéroscédasticité) : tracé de $\hat{y}_i \mapsto t_i^*$ et/ou de $x_{i,j} \mapsto t_i^*$, tracé de $i \mapsto t_i^*$, de t_i^* en fonction du temps pour détecter une auto-corrélation, tracé de t_i^* sous forme de carte pour détecter une corrélation spatiale.
- Non normalité : Histogramme des t_i^* et estimation de la densité, "boîte à moustaches" (box-plot), comparaison des quantiles empiriques associés aux t_i^* à l'espérance de ces quantiles sous hypothèse gaussienne (Q-Q plot ou droite de Henri). Ne pas oublier toutefois que la distribution des T_i^* n'est pas gaussienne (surtout quand n est petit).

Exemples : Modèle adéquat / détection d'écarts au modèle.

Mesures diagnostiques : tests de validité de sous-modèles, procédures de sélection de variables (c.f. chapitre suivant) pour détecter l'oubli de variables explicatives, test de Breusch et Pagan pour détecter l'hétéroscédasticité, test de Durbin-Watson pour détecter une auto-corrélation, test de Shapiro-Wilk pour détecter la non normalité.

4.2.3 Données aberrantes

On a vu que sous les hypothèses (C_1) à (C_4), les variables T_i^* sont i.i.d. de loi $\mathcal{T}(n-p-1)$. Même sans l'hypothèse gaussienne, si l'on a très grand nombre d'observations, on peut considérer que les T_i^* suivent approximativement cette loi.

Retour sur la signification du résidu : utilisation pour la détection de données mal ajustées par le modèle, i.e. isolées en la variable à expliquer Y.

Définition 11. *Soit $\alpha \in]0,1[$. La donnée (x_i, y_i) est dite* aberrante *au niveau α si $|t_i^*| > t_{n-p-1}(1-\alpha/2)$.*

Remarque : Attention, par définition, en moyenne approximativement αn données seront considérées comme aberrantes ! Si α est pris trop grand, on ne pourra pas vraiment parler de données aberrantes...

4.3 Analyse de la matrice de projection, effet levier

Les données aberrantes ne sont pas les seules sur lesquelles l'attention devra se porter précisément. On se concentre ici sur la détection des données isolées en les variables explicatives.

4.3.1 Propriétés de la matrice de projection

On rappelle que $H = \Pi_{\mathbb{X}} = \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'$, et que $h_{i,k}$ désigne l'élément (i,k) de H.

Proposition 9.
- $\sum_{i=1}^n h_{i,i} = p$ *(trace de H),*
- $\sum_{i=1}^n \sum_{k=1}^n h_{i,k}^2 = p$ *(idempotence),*
- $0 \leq h_{i,i} \leq 1$,
- $|h_{i,k}| \leq 0.5$ *pour tout* $k \neq i$,
- *Si* $h_{i,i} = 0$ *ou* $h_{i,i} = 1$ *alors* $h_{i,k} = 0$ *pour tout* $k \neq i$,
- $(1-h_{i,i})(1-h_{k,k}) - h_{i,k}^2 \geq 0$,
- $h_{i,i}h_{k,k} - h_{i,k}^2 \geq 0$.

4.3.2 Effet levier

On remarque ici que $\hat{y}_i = \sum_{k=1}^{n} h_{i,k} y_k$. Par conséquent, $h_{i,i}$ représente le poids de la $i$ème observation sur son ajustement. D'après la proposition précédente, si $h_{i,i} = 1$ alors $h_{i,k} = 0$ pour tout k, si au contraire $h_{i,i} = 0$. On a donc les cas extrêmes suivants : si $h_{i,i} = 1$, $\hat{y}_i$ est entièrement déterminée par y_i, si $h_{i,i} = 0$, alors y_i n'influe pas sur $\hat{y}_i$ qui est alors nul. La moyenne des $h_{i,i}$ étant par ailleurs égale à p/n, si $h_{i,i}$ est grand, cela signifie que y_i influe fortement sur $\hat{y}_i$

Définition 12. *La donnée* (x_i, y_i) *a un* effet levier *si*
- $h_{i,i} > 2p/n$ *(Hoaglin et Welsh, 1978),*
- $h_{i,i} > 3p/n$ *pour* $p > 6$ *et* $n - p > 12$ *(Velleman et Welsh, 1981),*
- $h_{i,i} > 0.2$ *ou* 0.5 *(Huber, 1981).*

Exemple de la régression linéaire simple : x_i éloignée du centre de gravité $\bar{x}$.

4.4 Mesures d'influence

On a donné des éléments de détection des données isolées en la variable à expliquer (données aberrantes), puis des données isolées en les variables explicatives (données ayant un effet levier).

Ces données influent-elles sur l'estimation des cœfficients de régression ?

Deux exemples de mesures d'influence basées sur l'estimation $\hat{\beta}_{(i)}$ après suppression de la $i$ème observation.

4.4.1 Distance de Cook

Rappel : région de confiance simultanée de niveau de confiance $1 - \alpha$ pour β :

$$\hat{I} = \left\{ \beta \in \overset{p}{\mathbb{R}}, \frac{1}{p\widehat{\sigma^2}} \left(\hat{\beta} - \beta\right)' \mathbb{X}'\mathbb{X} \left(\hat{\beta} - \beta\right) \leq f_{p,n-p}(1-\alpha) \right\}.$$

$\hookrightarrow$ ellipsoïde centré en $\hat{\beta}$.

L'influence de la $i$ème donnée peut être mesurée par le décentrage de cet ellipsoïde après suppression de cette donnée.

Définition 13. *La* distance de Cook *pour la ième donnée est définie par :*

$$C_i = \frac{(\hat{\beta} - \hat{\beta}_{(i)})'\mathbb{X}'\mathbb{X}(\hat{\beta} - \hat{\beta}_{(i)})}{p\widehat{\sigma^2}}.$$

La distance de Cook s'exprime en fonction des résidus studentisés, ou des résidus.

Proposition 10.

$$C_i = \frac{1}{p} \frac{h_{i,i}}{1 - h_{i,i}} T_i^2 = \frac{1}{p} \frac{h_{i,i}}{(1 - h_{i,i})^2 \widehat{\sigma^2}} \hat{\varepsilon}_i^2.$$

Mesure l'influence de la $i$ème donnée sur l'estimation simultanée des β_j.

Cook suggère de comparer chaque valeur observée de C_i aux quantiles de niveaux 0.1 à 0.5 d'une loi de Fisher à $(p, n-p)$ degrés de liberté, bien que les C_i ne suivent pas exactement cette loi (ce n'est pas un test exact !).
Interprétation d'une valeur observée de C_i élevée :

— Soit t_i^2 élevée : donnée aberrante,
— soit $h_{i,i}/(1-h_{i,i})$ élevée : donnée ayant un effet levier,
— soit les deux.

4.4.2 Distance de Welsh-Kuh (DFFITS)

Définition 14. *La distance de Welsh-Kuh - appelée aussi DFFITS ou DFITS - pour la ième donnée est définie par :*

$$WK_i = \frac{|x_i(\hat{\beta} - \hat{\beta}_{(i)})|}{\sqrt{\widehat{\sigma^2}_{(i)} h_{i,i}}} = |T_i^*| \sqrt{\frac{h_{i,i}}{1-h_{i,i}}}.$$

Mesure l'influence conjointe de la $i$ème donnée sur $\hat{\beta}$ et $\widehat{\sigma^2}$.

Seuils habituels pour l'observation de WK_i : $2\sqrt{p/n}$ ou $2\sqrt{p/(n-p)}$ (Chatterjee et Hadi, 1988).

Explication sur un exemple de la différence entre C_i et WK_i.

4.5 Correction des écarts au modèle

Lorsque l'on identifie une donnée "suspecte", la question de savoir si on la conserve ou non se pose nécessairement.
Il faut toujours être très vigilant lorsqu'on choisit de supprimer une donnée : cette donnée peut en effet contenir une information capitale pour la compréhension du phénomène étudié. Aussi, on recommande généralement de ne pas supprimer de données (en particulier celles qui sont influentes) sans avoir la certitude qu'elles sont le fruit d'erreurs de mesure par exemple, ou qu'une compréhension "locale" du problème est suffisante...
Si l'on souhaite conserver les données suspectes, ou si leur suppression (étudiée) ne suffit pas à corriger les écarts au modèle mis en évidence, il faut alors envisager une correction de ces écarts. Nous donnons ici quelques pistes de correction. Les corrections de la corrélation linéaire entre variables explicatives ne sont pas abordées dans ce cours : on pourra alors envisager des régression sur composantes principales ou des régressions de type PLS...

4.5.1 Ajout d'une variable explicative

Etape basée sur le diagnostic graphique pour le choix de la variable à considérer. Procédures plus "propres" de sélection de variables explicatives dans le chapitre suivant. Dans tous les cas, on devra à nouveau mettre en œuvre l'étape de détection des écarts au modèle.

4.5.2 Transformation des variables

Transformation des variables explicatives : OK
Attention à la transformation de la variable à expliquer seule, qui est en général mesurée en unités propres aux utilisateurs.
Transformation des variables à expliquer et explicatives.
Cas particuliers d'hétéroscédasticité : moindres carrés pondérés, moindres carrés généralisés.
Cas particulier de l'autocorrélation : différenciation.

4.5.3 Cas particulier d'hétéroscédasticité : Moindres Carrés Généralisés

On considère le modèle de régression linéaire multiple

$$(\mathcal{M}) \ : \ Y = \mathbb{X}\beta + \varepsilon,$$

où Y est un vecteur aléatoire à valeurs dans $\mathbb{R}^n$, $\mathbb{X}$ une matrice de plan d'expérience de taille $n \times p$ et de rang p, mais où le vecteur des bruits ε vérifie les conditions suivantes :

— $(C_1) : \mathbb{E}[\varepsilon] = 0$,
— $(C_2') : \mathrm{Var}(\varepsilon) = \sigma^2\Omega$, Ω étant une matrice symétrique définie positive connue, différente de I_n.

Exemples :

1. Régression pondérée : $\Omega = diag(w_1^2, \ldots, w_n^2)$. Les bruits sont non corrélés mais hétéroscédastiques.
2. Bruits suivant un processus AR(1) (auto-régressif d'ordre 1) : $\varepsilon_t = \varphi\varepsilon_{t-1} + \eta_t$, avec $|\varphi| < 1$, $E[\eta_t] = 0$, $\mathrm{cov}(\eta_i, \eta_k) = \sigma^2\delta_i^k$. Alors

$$\Omega = \frac{1}{1-\varphi^2}\begin{pmatrix} 1 & \varphi & \varphi^2 & \ldots & \varphi^{n-1} \\ \varphi & 1 & \varphi & \ldots & \varphi^{n-2} \\ \varphi^2 & \varphi & 1 & \ldots & \varphi^{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \varphi^{n-1} & \varphi^{n-2} & \varphi^{n-3} & \ldots & 1 \end{pmatrix}.$$

3. Données agrégées par blocs.

L'estimateur des moindres carrés ordinaires de β noté $\hat{\beta}$ est toujours défini par $\hat{\beta} = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'Y$, il reste sans biais, mais il n'est plus de variance minimale parmi les estimateurs linéaires sans biais de β.
De plus, l'estimateur $\widehat{\sigma^2}$ de σ^2 possède un biais.

La matrice Ω étant symétrique définie positive, il existe une matrice P inversible telle que $PP' = \Omega$.

On pose $Y^* = P^{-1}Y$, $\mathbb{X}^* = P^{-1}\mathbb{X}$ et $\varepsilon^* = P^{-1}\varepsilon$. En multipliant l'équation du modèle $(\mathcal{M})$ à gauche par P^{-1}, on obtient un nouveau modèle :

$$(\mathcal{M}^*) \ : \ Y^* = \mathbb{X}^*\beta + \varepsilon^*.$$

Ce nouveau modèle $(\mathcal{M}^*)$ vérifie les conditions standards d'un modèle de régression linéaire multiple et la matrice $\mathbb{X}^*$ est de rang p.

Le calcul de l'estimateur des MCO dans ce nouveau modèle donne ce qu'on appelle l'estimateur des moindres carrés généralisés $\hat{\beta}_{MCG}$.

Définition 15. *L'*estimateur des moindres carrés généralisés (ou estimateur d'Aitken) *est défini par*

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

Remarque : la matrice P n'est pas unique. On peut prendre par ex. la matrice racine carrée de Ω (obtenue par diagonalisation de Ω), mais aussi la racine carrée de Ω multipliée par une matrice Q orthogonale quelconque. Peu importe : le choix de P n'intervient pas dans l'expression de l'estimateur des moindres carrés généralisés...

Propriétés : $\hat{\beta}_{MCG}$ est un estimateur linéaire sans biais, de variance $\sigma^2(\mathbb{X}'\Omega^{-1}\mathbb{X})^{-1}$. Par le théorème de Gauss-Markov appliqué dans le modèle $(\mathcal{M}^*)$, on peut montrer que $\hat{\beta}_{MCG}$ est de variance minimale parmi les estimateurs linéaires sans biais de β.
Enfin, l'estimateur défini par

$$\widehat{\sigma^2}_{MCG} = \frac{(Y - \mathbb{X}\hat{\beta}_{MCG})'\Omega^{-1}(Y - \mathbb{X}\hat{\beta}_{MCG})}{n-p}$$

est un estimateur sans biais de σ^2.

Remarque de conclusion : si le modèle ne peut finalement pas être amélioré, on aura recours à des procédures d'estimation robuste afin de réduire l'influence des données "suspectes", ou des procédures de régression non paramétrique.

4.6 Exercice : Compléments / questions de cours

On considère une matrice de plan d'expérience $\mathbb{X}$ de taille $n \times p$. On note x_i la $i$ème ligne de la matrice $\mathbb{X}$ et $\mathbb{X}_{(i)}$ la matrice $\mathbb{X}$ privée de sa $i$ème ligne. Si H désigne la matrice de projection orthogonale sur l'espace $\mathcal{E}(\mathbb{X})$ engendré par les vecteurs colonnes de $\mathbb{X}$, $h_{i,k}$ désigne l'élément (i,k) de H. Soit Y un vecteur aléatoire de taille n, tel que $Y = \mathbb{X}\beta + \varepsilon$ où ε est un vecteur aléatoire satisfaisant les conditions standards d'un modèle de régression linéaire multiple sous hypothèse gaussienne. On note Y_i son $i$ème élément et $Y_{(i)}$ le vecteur Y privé de son $i$ème élément.

1. Résidus studentisés par validation croisée.

a) Montrer que $\mathbb{X}'\mathbb{X} = (\mathbb{X}'_{(i)}\mathbb{X}_{(i)}) + x'_i x_i$.

b) En déduire que

$$(\mathbb{X}'_{(i)}\mathbb{X}_{(i)})^{-1} = (\mathbb{X}'\mathbb{X})^{-1} + \frac{1}{1 - x_i(\mathbb{X}'\mathbb{X})^{-1}x'_i}(\mathbb{X}'\mathbb{X})^{-1}x'_i x_i(\mathbb{X}'\mathbb{X})^{-1} = (\mathbb{X}'\mathbb{X})^{-1} + \frac{1}{1-h_{i,i}}(\mathbb{X}'\mathbb{X})^{-1}x'_i x_i(\mathbb{X}'\mathbb{X})^{-1}.$$

c) Montrer que $\mathbb{X}'_{(i)}Y_{(i)} = \mathbb{X}'Y - x'_i Y_i$.

d) Montrer que les résidus studentisés par validation croisée définis par :

$$T_i^* = \frac{Y_i - x_i\widehat{\beta_{(i)}}}{\sqrt{\widehat{\sigma^2_{(i)}}(1 + x_i(\mathbb{X}'_{(i)}\mathbb{X}_{(i)})^{-1}x'_i)}},$$

où $\widehat{\beta_{(i)}}$ et $\widehat{\sigma^2_{(i)}}$ sont respectivement l'estimateur des moindres carrés ordinaires de β et l'estimateur sans biais de σ^2 obtenus après suppression de la $i$ème observation, vérifient

$$T_i^* = T_i \sqrt{\frac{n-p-1}{n-p-T_i^2}} = \frac{\hat{\varepsilon}_i}{\sqrt{\widehat{\sigma^2_{(i)}}(1-h_{i,i})}},$$

où les T_i sont les résidus studentisés, et $\hat{\varepsilon}_i$ les résidus.
e) Montrer que les T_i^* sont i.i.d. de loi $\mathcal{T}(n-p-1)$.
f) Quel est l'intérêt pratique de ces résultats ?

2. Distance de Cook.
a) Montrer que $\widehat{\beta_{(i)}} = \hat{\beta} - \frac{1}{1-h_{i,i}}(\mathbb{X}'\mathbb{X})^{-1}x_i'(Y_i - x_i\hat{\beta})$. Quel est l'impact sur la valeur de $\hat{\beta}$ de la suppression de la $i$ème observation ?
b) Montrer que la distance de Cook définie par

$$C_i = \frac{(\hat{\beta} - \widehat{\beta_{(i)}})'\mathbb{X}'\mathbb{X}(\hat{\beta} - \widehat{\beta_{(i)}})}{p\widehat{\sigma^2}}$$

s'exprime en fonction des résidus studentisés, ou des résidus comme :

$$C_i = \frac{1}{p}\frac{h_{i,i}}{1-h_{i,i}}T_i^2 = \frac{1}{p}\frac{h_{i,i}}{(1-h_{i,i})^2\widehat{\sigma^2}}\hat{\varepsilon}_i^2.$$

c) Quel est l'intérêt pratique de ce résultat ?

3. Effet levier.
Montrer que
- $tr(H) = \sum h_{i,i} = p$,
- $\sum_i \sum_k h_{i,k}^2 = p$,
- $0 \leq h_{i,i} \leq 1$ pour tout i,
- $-0.5 \leq h_{i,k} \leq 0.5$ pour tout k différent de i,
- si $h_{i,i} = 1 \quad$ ou $\quad 0$, alors $h_{i,k} = 0$ pour tout k différent de i.

Que peut-on déduire de ces propriétés ?

4. On considère un modèle de régression linéaire multiple à p variables explicatives.
a) Rappeler les méthodes algorithmiques de sélection de variables possibles et donner différents critères pouvant être utilisés pour la mise en œuvre de ces méthodes.
b) Comment ces différents critères sont-ils justifiés ?
c) Pour un même critère, toutes les méthodes algorithmiques donnent-elles le même résultat ? Pour une même méthode algorithmique, tous les critères donnent-ils le même résultat ? Expliquer.

Chapitre 5

Sélection de variables

5.1 Introduction

Dans ce chapitre, on s'intéresse à la question du choix de la matrice du plan d'expérience pour le modèle de régression linéaire, c'est-à-dire à celle du choix des variables explicatives.

Pourquoi cette question se pose-t-elle ? En pratique, on dispose de variables potentiellement explicatives seulement...
Les buts explicatif et prédictif de la modélisation statistique : problèmes liés à la multicolinéarité des variables explicatives, principe de parcimonie ⇒ critères de sélection de variables nécessaires.
Par ailleurs, on ne peut pas toujours étudier en détail TOUS les modèles de régression linéaire que l'on peut construire à partir de ces variables ⇒ méthodes algorithmiques de sélection nécessaires.

On suppose ici que l'on a un modèle de régression linéaire multiple exact à p variables explicatives ($p \leq n$) :

$$Y = \mathbb{X}\beta + \varepsilon,$$

avec

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbb{X} = \begin{pmatrix} x_{1,0} & \dots & x_{1,p-1} \\ \vdots & & \vdots \\ x_{n,0} & \dots & x_{n,p-1} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_{p-1} \end{pmatrix} \quad \text{et} \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

On rappelle les conditions standards imposées au vecteur des bruits ε :

- (C_1) : $\mathbb{E}[\varepsilon] = 0$ (centrage),
- (C_2) : $\text{cov}(\varepsilon_i, \varepsilon_j) = 0$ pour tout $i \neq j$ (non corrélation),
- (C_3) : $\text{var}(\varepsilon_i) = \sigma^2$ (inconnue) pour tout $i = 1 \dots n$ (homoscédasticité),

ainsi que l'hypothèse gaussienne :

- (C_4) : le vecteur ε suit une loi gaussienne.

On s'intéresse ici à des (sous-)modèles dont la matrice du plan d'expérience ne contient qu'une partie des variables explicatives.

Soit ξ un sous-ensemble de $\{0, \dots, p-1\}$ correspondant aux indices des variables explicatives conservées.
On note $|\xi|$ le cardinal de ξ, $\mathbb{X}_\xi$ la matrice composée des vecteurs colonnes de $\mathbb{X}$ qui sont conservés, et β_ξ le vecteur composé des éléments de β dont les indices sont dans ξ.
On note aussi $[\hat{\beta}]_\xi$ le vecteur composé des éléments de $\hat{\beta}$ (Estimateur des MCO dans le modèle exact) dont les indices sont dans ξ.
Exemple : $\xi = \{0, \dots, p-q-1\}$.

Estimateur des moindres carrés ordinaires de β_ξ dans le sous-modèle (M_ξ) $Y = \mathbb{X}_\xi \beta_\xi + \eta$: $\widehat{\beta_\xi} = (\mathbb{X}'_\xi \mathbb{X}_\xi)^{-1} \mathbb{X}'_\xi Y$.

5.2 Critères de qualité d'un modèle

Suivant que la modélisation étudiée a un but explicatif ou un but prédictif, les critères de qualité du modèle pris en compte ne seront pas les mêmes.

5.2.1 Qualité de l'estimation, erreur quadratique moyenne (EQM)

En général, $\widehat{\beta_\xi} \neq [\hat{\beta}]_\xi$ et $\widehat{\beta_\xi}$ est un estimateur biaisé de β_ξ. Mais $\mathrm{Var}([\hat{\beta}]_\xi) - \mathrm{Var}(\widehat{\beta_\xi})$ est une matrice semi-définie positive tout comme $\mathrm{Var}(\mathbb{X}\hat{\beta}) - \mathrm{Var}(\mathbb{X}_\xi \widehat{\beta_\xi})$ (estimation moins variable avec $\widehat{\beta_\xi}$) : preuve laissée en exercice.
Le biais résultant de l'abandon de variables explicatives utiles peut donc être compensé (mais pas forcément attention) par la diminution de la variance $\hookrightarrow$ compromis biais-variance, erreur quandratique moyenne.

Définition 16. *L'*erreur quadratique moyenne *associée au sous-modèle* (M_ξ) *est définie par*

$$EQM(\xi) = \mathbb{E}\left[\|X\beta - \mathbb{X}_\xi \widehat{\beta_\xi}\|^2\right].$$

Proposition 11. *On a la décomposition suivante.*

$$EQM(\xi) = \|X\beta - \mathbb{E}[\mathbb{X}_\xi \widehat{\beta_\xi}]\|^2 + \mathbb{E}\left[\|\mathbb{X}_\xi \widehat{\beta_\xi} - \mathbb{E}[\mathbb{X}_\xi \widehat{\beta_\xi}]\|^2\right] = \|(I - \Pi_{\mathbb{X}_\xi})X\beta\|^2 + |\xi|\sigma^2.$$

Preuve. Astuce de la trace.

Remarque : valable si le choix de ξ n'est pas basé sur les données utilisées pour estimer β_ξ.

Etude dans un cas simple, tendance fréquente des termes de biais et de variance : le meilleur ξ, noté ξ^*, serait celui qui réalise le meilleur compromis biais-variance.
Mais EQM(ξ) est inconnue (donc ξ^* est inaccessible, on parle d'"oracle") $\Rightarrow$ nécessité de construire des critères facilement accessibles, ou d'estimer EQM(ξ).

5.2.2 Qualité de la prédiction, erreur quadratique moyenne de prédiction (EQMP)

On cherche ici une valeur prédite $\hat{Y}^p_{n+1}$ de $Y_{n+1} = x'_{n+1,\xi}\beta_\xi + \eta_{n+1}$ (η_{n+1} vérifiant les conditions usuelles) correspondant à une nouvelle valeur des variables explicatives du sous-modèle (M_ξ) notée $x_{n+1,\xi}$.

Définition 17. *L'*erreur quadratique moyenne de prédiction *associée au sous-modèle* (M_ξ) *est définie par*

$$EQMP(\xi) = \mathbb{E}\left[(Y_{n+1} - x'_{n+1,\xi}\widehat{\beta_\xi})^2\right].$$

5.3 Critères de sélection de variables

5.3.1 Cadre général (conditions standards)

Somme des carrés résiduelle, cœfficient de détermination R^2

Le R^2 dans le sous-modèle M_ξ est défini par

$$R^2(\xi) = 1 - \frac{SCR(\xi)}{SCT}.$$

Si $\xi^- \subset \xi$ avec $|\xi^-| = |\xi| - 1$, on montre facilement que :

$$\begin{aligned} R^2(\xi) - R^2(\xi^-) &= \frac{SCR(\xi^-) - SCR(\xi)}{SCT} \\ &= \frac{\|\Pi_{\mathbb{X}_\xi} Y\|^2 - \|\Pi_{\mathbb{X}_{\xi^-}} Y\|^2}{SCT} \quad \text{par Pyhtagore} \\ &= \frac{\|\Pi_{\mathbb{X}_{\xi^-}}\Pi_{\mathbb{X}_\xi} Y + (I - \Pi_{\mathbb{X}_{\xi^-}})\Pi_{\mathbb{X}_\xi} Y\|^2 - \|\Pi_{\mathbb{X}_{\xi^-}} Y\|^2}{SCT} \\ &= \frac{\|(I - \Pi_{\mathbb{X}_{\xi^-}})\Pi_{\mathbb{X}_\xi} Y\|^2}{SCT}. \end{aligned}$$

Ainsi le cœfficient de détermination $R^2(\xi)$ est supérieur à $R^2(\xi^-)$: le cœfficient de détermination décroît à la suppression d'une variable explicative.
Choisir le (sous)-modèle dont la somme des carrés résiduelle est la plus petite ou dont le R^2 est le plus grand revient donc à choisir le modèle complet : la somme des carrés résiduelle et le R^2 sont donc globalement de mauvais critères de sélection de variables.
Ils restent cependant utiles pour choisir entre deux modèles ayant le même nombre de variables explicatives.

Deux possibilités : corriger le R^2 ou décider si l'augmentation du R^2 est statistiquement significative (tests sous hypothèse gaussienne).

Cœfficient de détermination ajusté R^2_a

Définition 18. *Le* cœfficient de détermination ajusté *dans le sous-modèle* (M_ξ) *est défini par*

$$R^2_a(\xi) = 1 - \frac{SCR(\xi)/(n - |\xi|)}{SCT/(n-1)}.$$

Remarques : Le R^2_a se calcule à partir du R^2 seul : $R^2_a(\xi) = \frac{(n-1)R^2(\xi) - (|\xi|-1)}{n-|\xi|}$. On a donc $R^2_a(\xi) < R^2(\xi)$ dès que $|\xi| \geq 2$. Il est possible que $R^2_a(\xi)$ soit négatif...

On cherche à maximiser le critère du R^2_a.

C_p de Mallows

On rappelle ici que $\text{EQM}(\xi) = \mathbb{E}[SCR(\xi)] - n\sigma^2 + 2|\xi|\sigma^2$.

Définition 19. *Le critère du C_p introduit par Mallows en 1973 est défini par :*

$$C_p(\xi) = \frac{SCR(\xi)}{\widehat{\sigma^2}} - n + 2|\xi|.$$

Représentation graphique de $C_p(\xi)$.

Proposition 12. *Si ξ ne dépend pas de Y, $\widehat{\sigma^2}C_p(\xi)$ est un estimateur sans biais de $EQM(\xi)$.*

Important : on utilise donc ce critère sur un autre jeu de données que celui utilisé pour l'estimation...

Choisir un modèle qui minimise le critère du C_p dans ces conditions revient alors à choisir un modèle qui en moyenne à une erreur quadratique moyenne minimale.

On remarque enfin que si le sous-modèle (M_ξ) est correct, et si ξ ne dépend pas de Y, $SCR(\xi)$ estime sans biais $(n - |\xi|)\sigma^2$ et $C_p(\xi) \simeq |\xi|$. Par conséquent, une règle usuelle est de retenir un sous-modèle (M_ξ) si $C_p(\xi) \leq |\xi|$.

Défaut : biais de sélection. Explication sur un exemple simple : matrice $\mathbb{X}$ orthogonale, $\mathbb{X}'\mathbb{X} = I_p$.
Dans ce cas, $SCR(\xi) = (n-p)\widehat{\sigma^2} + \sum_{j \notin \xi} \hat{\beta}_j^2$ (exercice).

Proposition 13. *Si la matrice $\mathbb{X}$ est orthogonale,*

$$\widehat{\sigma^2}C_p(\xi) = \sum_{j=0}^{p-1}(\hat{\beta}_j^2 - \widehat{\sigma^2}) - \sum_{j \in \xi}(\hat{\beta}_j^2 - 2\widehat{\sigma^2}).$$

A $|\xi|$ fixé, choisir le sous-modèle dont le C_p est minimum revient à choisir le modèle dont les paramètres estimés sont les plus grands en valeur absolue : ce plénomène est appelé *biais de sélection* $\Rightarrow$ idée du shrinkage, EMCO sous contrainte de norme (régressions ridge, PLS, lasso...).

Estimation de l'EQMP par bootstrap ou validation croisée

Si la modélisation a un but essentiellement prédictif, on peut utiliser une estimation bootstrap ou par validation croisée de l'EQMP (critère du PRESS équivalent à l'estimation par validation croisée Leave One Out).

5.3.2 Cadre gaussien

On peut dans ce cadre mettre en œuvre des tests d'hypothèses, ou considérer des critères basés sur la vraisemblance.

Tests de validité de sous-modèles

Pour choisir entre deux sous-modèles emboîtés l'un dans l'autre, l'un (M_ξ) défini par ξ de cardinal $|\xi| \geq 2$, validé, et l'autre, (M_{ξ^-}), défini par ξ^- tel que $\xi^- \subset \xi$ et $|\xi^-| = |\xi| - 1$, on peut utiliser un test de validité de sous-modèle.

Statistique de test : $F = \frac{SCR(\xi^-) - SCR(\xi)}{SCR(\xi)/(n-|\xi|)}$.
Sous l'hypothèse de validité du sous-modèle (M_{ξ^-}), $F \sim \mathcal{F}(1, n - |\xi|)$.
On rejette donc le sous-modèle (M_{ξ^-}) au profit de (M_ξ) au niveau α si la valeur f de F est supérieure au quantile de niveau $(1-\alpha)$ de la loi $\mathcal{F}(1, n-|\xi|)$, noté $f_{1,n-|\xi|,(1-\alpha)}$.

Remarque : une variante consiste à réduire par la variance estimée dans le modèle complet (plus pratique pour la mise en œuvre d'une méthode ascendante !).

Critères de vraisemblance pénalisée

Sous l'hypothèse gaussienne, la log-vraisemblance du modèle est donnée par

$$\ln \mathcal{L}(Y, \beta, \sigma^2) = -\frac{n}{2} \ln \sigma^2 - \frac{n}{2} \ln(2\pi) - \frac{1}{2\sigma^2} \|Y - \mathbb{X}\beta\|^2.$$

La log-vraisemblance maximale est obtenue dans le sous-modèle en $\tilde{\beta}_\xi = \hat{\beta}_\xi$ et $\widetilde{\sigma^2}_\xi = SCR(\xi)/n$ (estimateurs du maximum de vraisemblance) :

$$\ln \mathcal{L}(Y, \hat{\beta}_\xi, \widetilde{\sigma^2}_\xi) = -\frac{n}{2} \ln \frac{SCR(\xi)}{n} - \frac{n}{2}(1 + \ln(2\pi)).$$

Choisir un modèle maximisant la vraisemblance reviendrait donc à choisir un modèle (M_ξ) ayant la plus petite $SCR(\xi)$, donc le modèle complet $\Rightarrow$ comme pour les critères du R^2 et du C_p définis précédemment, nécessité de pénaliser les modèles ayant un grand nombre de variables explicatives.

On introduit donc des critères de la forme

$$-2 \ln \mathcal{L}(Y, \hat{\beta}_\xi, \widetilde{\sigma^2}_\xi) + |\xi| f(n),$$

où $|\xi| f(n)$ est un terme de pénalité positif et en général croissant avec n.

Définition 20. *Le* critère AIC *(Akaike's Information Criterion) introduit par Akaike en 1973 est défini par*

$$AIC(\xi) = -2 \ln \mathcal{L}(Y, \hat{\beta}_\xi, \widetilde{\sigma^2}_\xi) + 2|\xi| = n \ln \frac{SCR(\xi)}{n} + n(1 + \ln(2\pi)) + 2|\xi|.$$

Le critère BIC *(Bayesian Information Criterion) introduit par Schwarz en 1978 est défini par :*

$$BIC(\xi) = -2 \ln \mathcal{L}(Y, \hat{\beta}_\xi, \widetilde{\sigma^2}_\xi) + |\xi| \ln n = n \ln \frac{SCR(\xi)}{n} + n(1 + \ln(2\pi)) + |\xi| \ln n.$$

On cherchera à minimiser ces critères.

Remarque : Dès que $n \geq 8$, $\ln n \geq 2$ donc la pénalité du critère *BIC* est plus lourde que celle du critère *AIC* : les modèles choisis par ce critère auront moins de variables explicatives.
On peut par ailleurs montrer que lorsque n tend vers $+\infty$, la probabilité de sélectionner un modèle exact par minimisation du critère BIC tend vers 1 (pour plus de détails, on renvoie à l'ouvrage *Modèle linéaire par l'exemple*, d'Azaïs et Bardet).

5.4 Liens entre les différents critères

Sous hypothèse gaussienne, on peut comparer chaque critère à celui du test de validité de sous-modèle.
On considère deux sous-modèles emboîtés l'un dans l'autre, l'un (M_ξ) défini par ξ de cardinal $|\xi| \geq 2$, l'autre (M_{ξ^-}) défini par ξ^- tel que $\xi^- \subset \xi$ et $|\xi^-| = |\xi| - 1$.

5.4.1 R_a^2 et test de validité de sous-modèle

$R_a^2(\xi^-) < R_a^2(\xi)$ si et seulement si $\frac{SCR(\xi^-)}{n-|\xi|+1} > \frac{SCR(\xi)}{n-|\xi|}$, qui est équivalente à $\frac{(n-|\xi|)SCR(\xi^-)}{SCR(\xi)} > n-|\xi|+1$, puis $(n-|\xi|)\frac{SCR(\xi^-)-SCR(\xi)}{SCR(\xi)} > 1$ i.e. $F > 1$.
On retrouve le même critère que celui du test mais avec un seuil d'acceptation du sous-modèle (M_{ξ^-}) égal à 1 au lieu de $f_{1,n-|\xi|,(1-\alpha)}$ qui est supérieur à 3.84. On rejette le sous-modèle plus facilement avec le critère du R_a^2 qu'avec celui du test : on aura tendance à sélectionner avec le R_a^2 un modèle avec un plus grand nombre de variables explicatives.

5.4.2 C_p et test de validité de sous-modèle

$C_p(\xi^-) > C_p(\xi)$ si et seulement si $(n-p)\frac{SCR(\xi^-)-SCR(\xi)}{SCR} > 2$.
Par conséquent, si $\tilde{F}$ est la statistique de test de validité de sous-modèle variante, correspondant à une réduction par l'estimateur $\widehat{\sigma^2}$, $C_p(\xi^-) > C_p(\xi)$ si et seulement si $\tilde{F} > 2$.

5.4.3 Critères de vraisemblance pénalisée et test de validité de sous-modèle

$-2\ln\mathcal{L}(Y, \hat{\beta}_{\xi^-}\widetilde{\sigma^2}_{\xi^-}) + |\xi^-|f(n) > -2\ln\mathcal{L}(Y, \hat{\beta}_\xi\widetilde{\sigma^2}_\xi) + |\xi|f(n)$ si et seulement si $F > (n-|\xi|)\left(e^{2f(n)/n} - 1\right)$.
Si $2f(n)/n$ est proche de 0, $(n-|\xi|)\left(e^{2f(n)/n} - 1\right) \sim 2f(n)(1-|\xi|/n)$.

5.5 Méthodes algorithmiques de sélection

5.5.1 Méthode de recherche exhaustive

Le nombre de sous-modèles pouvant être construits sur les p variables explicatives est de 2^p.
Une recherche exhaustive n'est donc pas toujours possible lorsque p est grand.
De plus, ce type de recherche n'a pas de sens pour l'utilisation du critère de test de validité de sous-modèle.
On envisage alors une recherche pas à pas.

5.5.2 Méthode de recherche descendante (backward)

On part du modèle complet que l'on compare à tous les sous-modèles obtenus par suppresion d'une seule variable explicative. On retient éventuellement ensuite pour le critère de test le sous-modèle obtenu par suppression de la variable la moins significative parmi les non significatives à un niveau que l'on s'est fixé au départ (dont la valeur de la statistique du test de Student est la plus faible), et pour les autres critères le sous-modèle dont le critère est soit maximum et supérieur à celui du modèle de départ (R_a^2) soit minimum et inférieur à celui du modèle de départ (C_p, AIC, BIC).

On réitère ensuite cette étape en partant du modèle retenu, jusqu'à ce qu'on ne puisse plus supprimer de variable ou lorsqu'un certain seuil est atteint par la p valeur du test.

5.5.3 Méthode de recherche ascendante (forward)

On part du modèle avec la constante que l'on compare à tous les modèles obtenus par ajout d'une seule variable explicative. On retient éventuellement ensuite pour le critère de test le modèle obtenu par ajout de la variable la plus significative (dont la valeur de la statistique du test de Student est la plus grande), et pour les autres critères le modèle dont le critère est soit maximum et supérieur à celui du modèle de départ (R_a^2) soit minimum et inférieur à celui du modèle de départ (C_p, AIC, BIC).
On réitère ensuite cette étape en partant du modèle retenu, jusqu'à ce qu'on ne puisse plus ajouter de variable ou lorsqu'un certain seuil est atteint par la p valeur du test.

5.5.4 Méthode de recherche progressive (stepwise)

Méthode de recherche ascendante, avec possiblité de supprimer une variable précédemment introduite à chaque étape. Méthode très utile dans le cas de variables explicatives corrélées entre elles.

5.6 Exercices

Exercice 1 : Analyse de sorties R - Données Air Breizh

On souhaite pouvoir expliquer et prédire le maximum journalier de la concentration en ozone dans l'air à Rennes noté O. On considère pour cela les variables explicatives suivantes : la température à 6h, la température à 9h, la température à 12h, la nébulosité à 6h, la nébulosité à 12h, et la projection du vent sur l'axe Est-Ouest notées respectivement T_6, T_9, T_{12}, N_6, N_{12} et V.

1. Régression linéaire simple et détection des écarts au modèle.

On souhaite tout d'abord étudier un modèle de régression linéaire simple :

$$O_i = \beta_0 + \beta_1 T_{12,i} + \varepsilon_i \quad \text{pour } i \in \{1, \dots, 91\},$$

(données pour la seule année 1994, année durant laquelle certains capteurs étaient en panne) sous les conditions standards et sous hypothèse gaussienne.
On fournit en Annexe 1.3 plusieurs résultats graphiques obtenus après mise en œuvre de la régression via la fonction `lm` du logiciel R. Le nuage de points correspondant, avec la droite de régression obtenue sont représentés sur la figure 1. Les résidus studentisés sont ensuite représentés sur la figure 2 en fonction du jour, puis en fonction des valeurs ajustées. Les éléments diagonaux de la matrice de projection H sont représentés sur la figure 3 ainsi que la distance de Cook.

Que peut-on conclure de ces graphes ?

2. On considère maintenant le modèle de régression linéaire multiple complet :

$$O_i = \beta + \beta_{T_6} T_{6,i} + \beta_{T_9} T_{9,i} + \beta_{T_{12}} T_{12,i} + \beta_{N_6} N_{6,i} + \beta_{N_{12}} N_{12,i} + \beta_V V_i + \eta_i \quad \text{pour } i \in \{1, \dots, 1186\},$$

pour 1186 données ne comprenant pas celles de l'année 1994 sous les conditions standards et sous hypothèse gaussienne.

a) Après mise en œuvre de cette régression via la fonction `lm` du logiciel R, on a obtenu les résultats suivants :

```
Coefficients :
             Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)   63.1193      3.5993   17.537    < 2e-16  ***
T6            -1.6338      0.3160   -5.171   2.74e-07  ***
T9             0.2184      0.4863    0.449      0.653
T12            2.6335      0.3299    7.983   3.37e-15  ***
N6            -0.1585      0.2482   -0.639      0.523
N12           -2.2416      0.3134   -7.153   1.49e-12  ***
V              1.1124      0.1624    6.850   1.18e-11  ***
-
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 16.97 on 1179 degrees of freedom
Multiple R-Squared: 0.488, Adjusted R-squared: 0.4854
F-statistic: 187.3 on 6 and 1179 DF, p-value: < 2.2e-16
AIC: 10091.10, BIC: 10131.73.
```

Rappeler le résultat du test de validité globale du modèle au niveau 5%.

b) On a effectué les différentes régressions pour tous les sous-modèles possibles (avec la constante), et on a obtenu les résultats présentés dans les tables de l'Annexe 1.4.

Tester la validité du sous-modèle obtenu en supprimant les variables non significatives au niveau 5%. Décrire le test de façon détaillée, en justifiant la statistique de test.

À l'aide des différentes méthodes algorithmiques de sélection de variables, sur la base des critères présentés dans les tables, proposer une sélection de variables explicatives.

Exercice 2 : Analyse de sorties SAS - Consommation de confiseries

On étudie des données, publiées par Chicago Tribune en 1993, montrant la consommation de confiseries en millions 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, $i = 1 \dots, 17$. On obtient les résultats donnés à la fin de l'exercice avec :

$$\sum x_i = 751.8 \quad \sum x_i^2 = 97913.92 \quad \sum y_i = 13683.8 \quad \sum y_i^2 = 36404096.44 \quad \sum x_i y_i = 1798166.66$$

1. 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 hypothèses usuelles d'un modèle de régression linéaire simple.

a) Donner les expressions des estimateurs des MCO $\widehat{\beta_0}$ et $\widehat{\beta_1}$ de β_0 et β_1, puis des estimateurs $\hat{\sigma}(\widehat{\beta_0})$ et $\hat{\sigma}(\widehat{\beta_1})$ des écart-types $\sigma(\widehat{\beta_0})$ et $\sigma(\widehat{\beta_1})$ de $\widehat{\beta_0}$ et $\widehat{\beta_1}$. . Calculer les valeurs observées de ces estimateurs.

b) On suppose les ε_i i.i.d. de loi $N(0, \sigma^2)$. Tester l'hypothèse nulle (H_0) : $\beta_0 = 0$ contre l'alternative (H_1) : $\beta_0 \neq 0$ au niveau 5%. Commenter.

2. On considère maintenant le modèle sans constante $Y_i = \beta x_i + \varepsilon_i$, avec ε_i i.i.d. de loi $N(0, \sigma^2)$, et on cherche à diagnostiquer d'éventuels écarts au modèle. Pour cela, on met en œuvre la procédure `REG` de SAS, avec les options `influence` et `r`, avec toutes les observations dans un premier temps, on retirant deux observations dans un second temps. On obtient les sorties fournies dans l'Annexe 1.5.
Que peut-on conclure de ces sorties ?

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

Exercice 3 : Analyse de sorties SAS - La processionnaire du pin

On considère un jeu de données classique issu de l'ouvrage de Tomassone et al. *La régression* (1992). Ces données visent à une étude de la prolifération des chenilles processionnaires du pin en fonction de certaines caractéristiques de peuplements forestiers. On dispose de 33 observations, chaque observation correspondant à une parcelle forestière de 10 hectares. Une parcelle est considérée comme homogène par rapport aux variables étudiées. Les valeurs observées de ces variables ont été obtenues comme des moyennes de valeurs mesurées sur des placettes échantillon de 5 ares. La variable à expliquer est le logarithme du nombre de nids de chenilles processionnaires par arbre de la placette échantillon, les 10 variables explicatives potentielles sont :

1. L'altitude moyenne de la placette (en m).
2. La pente moyenne du terrain.
3. Le nombre de pins dans la placette.
4. La hauteur d'un arbre échantillonné au centre de la placette.
5. Le diamètre de cet arbre.
6. La note de densité du peuplement.
7. L'orientation de la placette (1= sud, 2=autre).

8. La hauteur (en m) des arbres dominants de la placette.
9. Le nombre de strates de végétation.
10. Le mélange du peuplement (1=non mélangé, 2= mélangé).

On a mis en œuvre différentes méthodes de sélection de variables à l'aide de la procédure `REG` de SAS. Les résultats obtenus sont fournis dans l'Annexe 1.6.

Analyser ces sorties.

Chapitre 6

Annales corrigées

6.1 Examens partiels

6.1.1 Sujet 1 (durée : 1h30)

Exercice : Interprétations géométriques en régression linéaire simple

1. On considère un modèle de régression linéaire simple sans constante : $Y = \mathbb{X}\beta + \varepsilon$, où :
   - Y est un vecteur aléatoire à valeurs dans $\mathbb{R}^2$,
   - $\mathbb{X} = (2, 1)'$,
   - $\beta \in \mathbb{R}$,
   - ε est un vecteur aléatoire vérifiant les conditions standards d'un modèle de régression linéaire.

a) Rappeler la définition de l'estimateur des moindres carrés ordinaires $\hat{\beta}$ de β, et donner une interprétation géométrique du vecteur des valeurs ajustées $\hat{Y} = \mathbb{X}\hat{\beta}$.

b) La valeur observée de Y est $y = (3, 4)'$. À l'aide de la question précédente exclusivement, déterminer géométriquement sans faire de calcul la valeur de $\hat{\beta}$.

2. On considère maintenant le modèle de régression linéaire simple avec constante $Y = \mathbb{X}\beta+\varepsilon$, où :
   - Y est un vecteur aléatoire à valeurs dans $\mathbb{R}^2$,
   - $\mathbb{X} = \begin{pmatrix} 1 & 2 \\ 1 & 1 \end{pmatrix}$,
   - $\beta \in \mathbb{R}^2$,
   - ε est un vecteur aléatoire vérifiant les conditions standards d'un modèle de régression linéaire.

Déterminer géométriquement sans faire de calcul la valeur de $\hat{\beta}$ dans ce cas.

3. Montrer que l'on retrouve les résultats connus dans le modèle de régression linéaire simple.

Problème : Consommation de Coca-Cola, restaurants Mac Donald's et croissance

Dans un article de recherche économique publié le 1er avril 2010 par Natixis, on étudie les effets de l'américanisation sur la croissance moyenne du PIB de 47 pays, entre 1996 et 2007 (avant la crise). Les indicateurs de l'américanisation retenus pour cette étude sont notamment

la consommation de Coca-Cola (en Ounces) par habitant et le nombre de restaurants Mac Donald's pour un million d'habitants.
On considère un modèle de la forme :

$$(\mathcal{M}) \quad P_i = \beta_0 + \beta_1 C_i + \beta_2 M_i + \varepsilon_i \quad \text{pour } i \in \{1, \dots, 47\},$$

où P_i, C_i, M_i désignent respectivement la valeur de la croissance moyenne du PIB (en %) entre 1996 et 2007, la consommation de Coca-Cola par habitant (en Ounces) et le nombre de restaurants Mac Donald's pour un million d'habitants pour le pays i, les ε_i sont des termes d'erreur aléatoires vérifiant les conditions standards d'un modèle de régression linéaire multiple.
Ce modèle peut s'écrire sous la forme :

$$P = \mathbb{X}\beta + \varepsilon,$$

avec $P = (P_1, \dots, P_{47})'$, $\beta = (\beta_0, \beta_1, \beta_2)'$ et $\varepsilon = (\varepsilon_1, \dots, \varepsilon_{47})'$.
On a alors

$$\mathbb{X}'\mathbb{X} = \begin{pmatrix} 47 & 6500 & 397.69 \\ 6500 & 1501754 & 71104.33 \\ 397.69 & 71104.33 & 7587.418 \end{pmatrix},$$

$$(\mathbb{X}'\mathbb{X})^{-1} = \begin{pmatrix} 0.0589966941 & -1.958369e-04 & -1.257021e-03 \\ -0.0001958369 & 1.847088e-06 & -7.045029e-06 \\ -0.0012570207 & -7.045029e-06 & 2.637045e-04 \end{pmatrix},$$

et si p désigne la valeur observée de P, $\mathbb{X}'p = (178.97, 22080.38, 1291.301)'$, et $p'p = 819.0221$.

1. Déterminer l'estimateur des moindres carrés ordinaires $\hat{\beta}$ de β et calculer sa valeur.

2. Déterminer un estimateur sans biais de la matrice de variance-covariance $\mathrm{Var}(\hat{\beta})$ de $\hat{\beta}$, et donner sa valeur.

3. Montrer que la valeur du coefficient de détermination R^2 est égale à 0.1302.

4. On suppose désormais que $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_{47})$.

a) Quel est l'estimateur du maximum de vraisemblance de β ?

b) Quelle est la loi de cet estimateur et quelles sont ses propriétés statistiques ?

c) Construire un test de significativité globale du modèle au niveau α ($\alpha \in]0, 1[$). A partir des tables fournies en Annexe 2, donner un encadrement de la p-valeur de ce test. Que peut-on en conclure ?

d) Construire une région de confiance simultanée pour (β_1, β_2) de niveau de confiance 95%. Quelle forme cette région a-t-elle ? Quelle(s) hypothèse(s) pourrait-on tester directement à partir de la construction de cette région de confiance ?

e) Quelles sont les variables explicatives significatives au niveau 5% ?

f) Tester l'hypothèse $(H_0) : \beta_1 \geq 0$ contre $(H_1) : \beta_1 < 0$ au niveau 5%, puis l'hypothèse $(H_0) : \beta_2 \geq 0$ contre $(H_1) : \beta_2 < 0$ au niveau 5%.

5. On a lancé des procédures de régression à l'aide de la fonction `lm` du logiciel R pour différents modèles obtenus par suppression d'une variable explicative, puis modification des variables explicatives, sur les mêmes données, et on a obtenu les sorties fournies en Annexe 2.1.

a) Est-il surprenant que la valeur du R^2 dans les modèles ($\mathcal{M}^{(1)}$) et ($\mathcal{M}^{(2)}$) soit inférieure à celle du modèle complet ($\mathcal{M}$) défini ci-dessus ? Expliquer.

b) En première analyse, quel modèle choisirait-on de considérer ? Justifier précisément la réponse.

6. On souhaite construire un intervalle de prédiction de niveau de confiance 95% pour la croissance moyenne du PIB d'un pays dont la consommation de Coca-Cola est de 300 Ounces par habitant, et le nombre de restaurants Mac Donald's par million d'habitants de 5. Quelle stratégie peut-on adopter pour construire cet intervalle au vu des résultats obtenus ? Donner la valeur de l'intervalle ainsi construit.

7. La conclusion de l'article est la suivante : "L'étude montre sans ambiguïté que l'américanisation est défavorable à la croissance.". Que pensez-vous de cette conclusion (vous aurez noté la date de parution de l'article) ?
Et de la citation d'Albert Brie (*Le mot du silencieux - L'hiver nous fait suer)* : "La statistique est moins une science qu'un art. Elle est la poésie des nombres. Chacun y trouve ce qu'il y met." ?

Extraits de tables statistiques

Table de la loi gaussienne : on donne pour différentes valeurs de $\alpha \in [0,1]$, q_α tel que $P(N \leq q_\alpha) = \alpha$ lorsque $N \sim \mathcal{N}(0,1)$.

α	0.9	0.95	0.975
q_α	1.282	1.645	1.96

Table de la loi de Student : on donne pour différentes valeurs de n et de $\alpha \in [0,1]$, $t_{n,\alpha}$ tel que $P(T \leq t_{n,\alpha}) = \alpha$ lorsque $T \sim \mathcal{T}(n)$ (on rappelle que la loi de Student est symétrique).

α	0.9	0.95	0.975
$t_{43,\alpha}$	1.302	1.681	2.017
$t_{44,\alpha}$	1.301	1.680	2.015
$t_{45,\alpha}$	1.301	1.679	2.014
$t_{46,\alpha}$	1.3	1.679	2.013
$t_{47,\alpha}$	1.3	1.678	2.012

Table de la loi de Fisher : on donne pour différentes valeurs de (n_1, n_2) et de $\alpha \in [0,1]$, $f_{n_1,n_2,\alpha}$ tel que $P(F \leq f_{n_1,n_2,\alpha}) = \alpha$ lorsque $F \sim \mathcal{F}(n_1, n_2)$.

α	0.025	0.05	0.95	0.975
$f_{1,44,\alpha}$	0.001	0.004	4.062	5.386
$f_{1,45,\alpha}$	0.001	0.004	4.057	5.377
$f_{1,46,\alpha}$	0.001	0.004	4.052	5.369
$f_{2,44,\alpha}$	0.025	0.051	3.209	4.016
$f_{2,45,\alpha}$	0.025	0.051	3.204	4.009
$f_{2,46,\alpha}$	0.025	0.051	3.2	4.001

Table de la loi de Fisher : on donne pour différentes valeurs de n_1, n_2 et q la valeur de $p_{n_1,n_2,q} = P(F \leq q)$ lorsque $F \sim \mathcal{F}(n_1, n_2)$.

q	1	2	3	3.1	3.2	3.3	3.4	3.5	3.6	3.7	3.8	3.9	4
$p_{1,44,q}$	0.677	0.836	0.910	0.915	0.919	0.924	0.928	0.932	0.936	0.939	0.942	0.945	0.948
$p_{1,45,q}$	0.677	0.836	0.910	0.915	0.920	0.924	0.928	0.932	0.936	0.939	0.942	0.946	0.948
$p_{1,46,q}$	0.677	0.836	0.910	0.915	0.920	0.924	0.928	0.932	0.936	0.939	0.943	0.946	0.949
$p_{2,44,q}$	0.624	0.853	0.940	0.945	0.950	0.954	0.958	0.961	0.964	0.967	0.970	0.972	0.975
$p_{2,45,q}$	0.624	0.853	0.940	0.945	0.950	0.954	0.958	0.961	0.965	0.967	0.970	0.973	0.975
$p_{2,46,q}$	0.624	0.853	0.940	0.945	0.950	0.954	0.958	0.962	0.965	0.968	0.970	0.973	0.975

6.1.2 Sujet 1 : éléments de correction

Exercice : Interprétations géométriques en régression linéaire simple

1. a) L'estimateur des MCO de β est défini par $\hat{\beta} = \text{argmin}_{\beta\in\mathbb{R}^p}\|Y - \mathbb{X}\beta\|^2$. Si $\mathcal{E}(\mathbb{X})$ désigne l'espace vectoriel engendré par les vecteurs colonnes de $\mathbb{X}$, le vecteur $\mathbb{X}\hat{\beta}$ est le vecteur de $\mathcal{E}(\mathbb{X})$ dont la distance à Y est minimale. En d'autres termes, $\mathbb{X}\hat{\beta}$ est le projeté orthogonal de Y sur $\mathcal{E}(\mathbb{X})$.
b) Après avoir tracé le projeté orthogonal de Y sur $\mathcal{E}(\mathbb{X})$, on trouve que $\mathbb{X}\hat{\beta}_{obs} = 2\mathbb{X}$, d'où $\hat{\beta}_{obs} = 2$.

2. On ajoute la constante au modèle. On note X_0 le vecteur constant et X_1 le vecteur $(2, 1)'$. Les vecteurs X_0 et X_1 étant non colinéaires, l'espace $\mathcal{E}(\mathbb{X})$ engendré par ces vecteurs est $\mathbb{R}^2$ tout entier. Par conséquent, le projeté orthogonal de Y sur $\mathcal{E}(X)$ est égal à Y. De plus, on peut voir que $Y = 5X_0 - X_1$. On a donc $\hat{\beta}_{0,obs} = 5$ et $\hat{\beta}_{1,obs} = -1$.

3. Le modèle s'écrit également $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, pour $i \in \{1, 2\}$, avec $y_1 = 3$, $y_2 = 4$, $x_1 = 2$, $x_2 = 1$. Avec les formules de la régression linéaire simple, on obtient bien $\hat{\beta}_{1,obs} = -1$ et $\hat{\beta}_{0,obs} = 7/2 - (-1).3/2 = 5$.

Problème : Consommation de Coca-Cola, restaurants Mac Donald's et croissance

Ici, la matrice $\mathbb{X}$ s'écrit : $\mathbb{X} = \begin{pmatrix} 1 & C_1 & M_1 \\ \vdots & \vdots & \vdots \\ 1 & C_{47} & M_{47} \end{pmatrix}$.
Les conditions standards sont les conditions (C_1) à (C_3) du cours (centrage, non corrélation et homoscédasticité des ε_i).

1. On a $\hat{\beta} = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'P$ d'où $\hat{\beta}_{obs} = (4.611292, -0.003362, -0.040004)'$.

2. Un estimateur sans biais de $\text{Var}(\hat{\beta})$ est donné par $\widehat{\sigma^2}(\mathbb{X}'\mathbb{X})^{-1}$, où $\widehat{\sigma^2} = SCR/(47 - 3) = (P'P - P'\mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'P)/44$ est un estimateur sans biais de la variance des P_i. La valeur de $\widehat{\sigma^2}$ est 2.718764, d'où la valeur de l'estimateur sans biais de $\text{Var}(\hat{\beta})$:

$$\begin{pmatrix} 0.1603980880 & -5.324344e-04 & -3.417543e-03 \\ -0.0005324344 & 5.021797e-06 & -1.915377e-05 \\ -0.0034175426 & -1.915377e-05 & 7.169504e-04 \end{pmatrix}.$$

3. $R^2 = 1 - SCR/SCT$. Or $SCT_{obs} = p'p - 47\bar{p}^2 = 819.0221 - 178.97^2/47 = 137.5272$, d'où $R^2_{obs} = 0.1302$.

4. a) L'estimateur du maximum de vraisemblance est égal à $\hat{\beta}$.
b) Il suit une loi gaussienne d'espérance β, de variance $\sigma^2(\mathbb{X}'\mathbb{X})^{-1}$. Il est sans biais, de variance minimale parmi les estimateurs linéaires sans biais de β (Th. de Gauss-Markov).
c) Test de $(H_0) : \beta_1 = \beta_2 = 0$ contre (H_1) : il existe $j \in \{1, 2\}$, $\beta_j \neq 0$.
Statistique de test : $F(P) = \frac{R^2}{1-R^2}\frac{44}{2}$ qui suit sous (H_0) la loi $\mathcal{F}(2, 44)$.
La région critique du test est donnée par $\mathcal{R}_{(H_0)} = \{p, F(p) > f_{2,44}(1 - \alpha)\}$.

Or $R^2_{obs} = 0.1302$, d'où $F(p) = 3.293$, donc la p valeur du test est comprise strictement entre 0.046 et 0.05. On rejette (H_0) pour un niveau 5%. Le modèle est tout juste globalement significatif à ce niveau !

d) Région de confiance simultanée pour (β_1, β_2) : en introduisant $M = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$, on trouve une région de confiance pour (β_1, β_2) de la forme

$$RC_{(\beta_1,\beta_2)} = \left\{(\beta_1, \beta_2),\ \frac{1}{2\widehat{\sigma^2}}(\hat{\beta}_1 - \beta_1, \hat{\beta}_2 - \beta_2)[M(\mathbb{X}'\mathbb{X})^{-1}M']^{-1}(\hat{\beta}_1 - \beta_1, \hat{\beta}_2 - \beta_2)' \leq f_{2,47}(95\%)\right\},$$

c'est-à-dire

$$RC_{(\beta_1,\beta_2)}(\omega) = \Big\{(\beta_1, \beta_2),\ 0.1839071(602817.83(\hat{\beta}_1 - \beta_1)^2 + 2 \times 16104.649(\hat{\beta}_1 - \beta_1)(\hat{\beta}_2 - \beta_2) + 4222.368(\hat{\beta}_2 - \beta_2)^2) \leq 3.209\Big\}.$$

On obtient ainsi une ellipse de confiance.

e) Individuellement, seule la constante est significative. Pour le test de $\beta_1 = 0$ contre $\beta_1 \neq 0$, la valeur de la statistique $|T_1|$ avec$T_1 = \hat{\beta}_1 / \sqrt{\widehat{\sigma^2}(\mathbb{X}'\mathbb{X})^{-1}_{2,2}}$ du test de Student est égale à $|-0.003362/\sqrt{5.021797e-06}| = |-1.500|$. Pour le test de $\beta_2 = 0$ contre $\beta_2 \neq 0$, la valeur observée de la statistique du test de Student est égale à $|-1.494|$, pour un quantile de la loi de Student à 44 degrés de liberté égal à 2.015.

f) Pour le test de $(H_0) : \beta_1 \geq 0$ contre $(H_1) : \beta_1 < 0$, on rejette (H_0) lorsque la statistique de test T_1 définie à la question précédente est inférieure à -1.68. Or la valeur de T_1 est égale à -1.5 : on ne peut pas rejeter (H_0) au niveau 5%. De la même façon, on montre qu'on ne rejette pas l'hypothèse $\beta_2 \geq 0$ au niveau 5%.

5. Le coefficient de détermination diminue nécessairement après la suppression d'une variable explicative. Sans information supplémentaire, il semble que l'on puisse préférer le modèle $(\mathcal{M}_{1^2})$, qui possède un R^2 ajusté maximum, dont le test de significativité globale possède la plus petite p valeur, et dont toutes les variables sont significatives au niveau 5%. Cependant, le R^2 observé reste très faible et, en mettant la régression en œuvre, on peut détecter plusieurs écarts au modèle... L'étude est à revoir entièrement !

6. Pour construire un intervalle de prédiction, il serait souhaitable de considérer le modèle retenu à la question précédente. Or on ne dispose pas dans les sorties R fournies de la matrice $\mathbb{X}'\mathbb{X}$ correspondante. On doit donc se contenter de construire un intervalle de prédiction sur la base du modèle $(\mathcal{M}^{(2)})$ qui est meilleur que le modèle initial.
La valeur de l'intervalle de prédiction est alors donné par

$$4.420602 - 0.004431 \times 300 \pm 2.014 \times 1.671 \sqrt{1 + (1, 300)\begin{pmatrix} 47 & 6500 \\ 6500 & 1501754 \end{pmatrix}^{-1}(1, 300)'},$$

ou encore $[-0.3811783; 6.563782]$.

7. La conclusion de l'article "Poisson d'avril" est fausse pour au moins deux raisons fondamentales.

— Le fait que les cœfficients $\hat{\beta}_1$ et $\hat{\beta}_2$ soient négatifs pourrait éventuellement indiquer que la croissance évolue dans le sens opposé à celui de la consommation de Coca-Cola ou l'implantation de restaurants Mac Donald's. En aucun cas, on a nécessairement une relation de CAUSALITÉ !
— Les conclusions se basent sur l'estimation ponctuelle... qui est clairement insuffisante au vu des résultats des tests sous hypothèse gaussienne.

Enfin, on l'a vu, l'étude doit être complétée par une étude des écarts au modèle. Le R^2 observé est en effet très petit...

6.1.3 Sujet 2 (durée : 1h30)

Données statistiques de vidéos musicales en accès sur YouTube

YouTube est un site web d'hébergement de vidéos : les internautes peuvent y déposer des vidéos, les partager et peuvent également y consulter des statistiques relatives à ces vidéos. On a choisi d'étudier certaines de ces statistiques pour n vidéos musicales, choisies parmi celles déposées en 2010 et 2011 et les plus visionnées.
On s'intéresse en particulier pour chaque vidéo au nombre de "Favoris", de "J'aime", et de "Je n'aime pas", qui correspondent respectivement aux nombres d'internautes ayant déclaré la vidéo comme une de leurs favorites, ayant déclaré l'avoir aimée, et ayant déclaré ne pas l'avoir aimée (en millions d'internautes).

On cherche ici à expliquer le nombre de "Favoris" en fonction du nombre de "J'aime" et du nombre de "Je n'aime pas".

On considère pour cela un modèle de régression linéaire multiple de la forme :

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \varepsilon_i \quad \text{pour } i \in \{1, \dots, n\},$$

où Y_i, $x_{i,1}$, $x_{i,2}$ désignent respectivement les nombres de "Favoris", de "J'aime", de "Je n'aime pas" pour la $i$ème vidéo considérée, et où les ε_i sont des termes d'erreur aléatoires vérifiant les conditions standards du modèle de régression linéaire multiple.
Ce modèle peut aussi s'écrire sous la forme matricielle : $Y = \mathbb{X}\beta + \varepsilon$, avec

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbb{X} = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} \\ \vdots & \vdots & \vdots \\ 1 & x_{n,1} & x_{n,2} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

On a alors

$$\mathbb{X}'\mathbb{X} = \begin{pmatrix} 42 & 7.763 & 0.606 \\ 7.763 & 3.843 & 0.281 \\ 0.606 & 0.281 & 0.059 \end{pmatrix}, \quad (\mathbb{X}'\mathbb{X})^{-1} = \begin{pmatrix} 0.038 & -0.074 & -0.038 \\ -0.074 & 0.544 & -1.828 \\ -0.038 & -1.828 & 26.043 \end{pmatrix},$$

et si y désigne la valeur observée de Y, $\mathbb{X}'y = (7.686, 3.682, 0.236)'$.

On suppose que le vecteur des bruits ε est de loi gaussienne.

1. Préciser l'intérêt de chacune des hypothèses du modèle. Selon vous, ces hypothèses sont-elles pertinentes pour l'étude sur les vidéos musicales de YouTube menée ici ?

2. Quel est le nombre n de vidéos considérées dans cette étude ?

3. Rappeler la définition de l'estimateur des moindres carrés ordinaires $\hat{\beta}$ de β, et donner une interprétation géométrique du vecteur des valeurs ajustées $\hat{Y} = \mathbb{X}\hat{\beta}$. En déduire une expression matricielle de $\hat{\beta}$, puis une interprétation géométrique du vecteur des résidus $\hat{\varepsilon} = Y - \hat{Y}$.

4. Donner les lois de $\hat{\beta}$, de $\hat{Y}$, et de $\hat{\varepsilon}$. Ces vecteurs aléatoires sont-ils indépendants ?

5. La valeur de l'estimateur des moindres carrés ordinaires $\hat{\beta}$ calculé sur les observations est $\hat{\beta}_{obs} = (0.011, 1.003, -0.877)'$. Expliquer comment cette valeur a pu être obtenue à partir des données fournies ci-dessus.

6. Ecrire l'équation d'analyse de la variance, et rappeler la définition du cœfficient de détermination R^2 dans le modèle considéré. Donner l'interprétation géométrique de ces deux notions.

7. Montrer que la somme des carrés expliquée est égale à $\hat{\beta}'\mathbb{X}'\mathbb{X}\hat{\beta} - n\bar{Y}^2$, et calculer sa valeur à partir des observations.

8. La valeur du cœfficient de détermination calculé sur les observations est $R^2_{obs} = 0.981$. En déduire les valeurs de la somme des carrés totale, puis de la somme des carrés résiduelle calculées sur les observations.

9. Déterminer un estimateur sans biais de la variance du modèle. Calculer sa valeur.

10. Tester la validité globale du modèle au niveau 5%.

11. On souhaite prédire le nombre de "Favoris" d'une vidéo qui aurait 0.4 (millions) de "J'aime" et 0.02 (millions) de "Je n'aime pas". Donner un intervalle de prédiction de niveau de confiance 95% pour ce nombre de "Favoris".

12. Tester l'hypothèse (H_0) $\beta_1 \leq 0$ contre (H_1) $\beta_1 > 0$ au niveau 5%, puis l'hypothèse (H_0) $\beta_2 \geq 0$ contre (H_1) $\beta_2 < 0$ au niveau 5%. Les résultats obtenus sont-ils surprenants ?

13. Donner, à partir des tables données en Annexe, une approximation ou un encadrement de la p-valeur du test de significativité de la constante ((H_0) $\beta_0 = 0$ contre (H_1) $\beta_0 \neq 0$). En déduire qu'il pourrait être opportun de considérer le modèle de régression linéaire multiple sans constante.

14. Après la mise en œuvre d'une nouvelle procédure de régression linéaire multiple sans constante, on a obtenu les résultats suivants :

	Cœfficients estimés	Ecart-type	Statistique du test de Student	p-valeur du test
"J'aime"	1.022	0.021	48.67	3.2e-37
"Je n'aime pas"	-0.874	0.171	-5.11	8.3e-06

et $R^2_{obs} = 0.987$.

Expliquer par quels calculs chacun de ces résultats a été obtenu.

15. Quel modèle choisiriez-vous finalement pour faire de la prédiction ? Justifiez votre réponse.

Extraits de tables statistiques

Table de la loi gaussienne : on donne pour différentes valeurs de $\alpha \in [0,1]$, q_α tel que $P(N \leq q_\alpha) = \alpha$ lorsque $N \sim \mathcal{N}(0,1)$.

α	0.9	0.95	0.975
q_α	1.282	1.645	1.96

Table de la loi de Student : on donne pour différentes valeurs de n et de $\alpha \in [0,1]$, $t_{n,\alpha}$ tel que $P(T \leq t_{n,\alpha}) = \alpha$ lorsque $T \sim \mathcal{T}(n)$ (on rappelle que la loi de Student est symétrique).

α	0.9	0.95	0.975
$t_{38,\alpha}$	1.304	1.686	2.024
$t_{39,\alpha}$	1.304	1.685	2.023
$t_{40,\alpha}$	1.303	1.684	2.021
$t_{41,\alpha}$	1.303	1.683	2.02
$t_{42,\alpha}$	1.302	1.682	2.018
$t_{43,\alpha}$	1.302	1.681	2.017
$t_{44,\alpha}$	1.301	1.680	2.015

Table de la loi de Student : on donne pour les valeurs de n allant de 38 à 45 et pour différentes valeurs de q la valeur de $p_q = P(T \leq q)$ lorsque $T \sim \mathcal{T}(n)$.

q	1	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9
p_q	0.838	0.861	0.881	0.899	0.915	0.929	0.941	0.951	0.96	0.968

Table de la loi de Fisher : on donne pour différentes valeurs de (n_1, n_2) et de $\alpha \in [0,1]$, $f_{n_1,n_2,\alpha}$ tel que $P(F \leq f_{n_1,n_2,\alpha}) = \alpha$ lorsque $F \sim \mathcal{F}(n_1, n_2)$.

α	0.025	0.05	0.95	0.975
$f_{1,38,\alpha}$	0.001	0.004	4.098	5.446
$f_{1,39,\alpha}$	0.001	0.004	4.091	5.435
$f_{1,40,\alpha}$	0.001	0.004	4.085	5.424
$f_{2,38,\alpha}$	0.025	0.051	3.245	4.071
$f_{2,39,\alpha}$	0.025	0.051	3.238	4.061
$f_{2,40,\alpha}$	0.025	0.051	3.232	4.051
$f_{3,38,\alpha}$	0.071	0.116	2.852	3.483
$f_{3,39,\alpha}$	0.071	0.116	2.845	3.473
$f_{3,40,\alpha}$	0.071	0.116	2.839	3.463

6.1.4 Sujet 2 : éléments de correction

1. L'hypothèse de centrage implique la relation linéaire "en moyenne" : $\mathbb{E}[Y] = \mathbb{X}\beta$. La non-corrélation, l'homoscédasticité et la loi gaussienne permettent de faire de l'inférence statistique classique : les hypothèses du théorème de Cochran sont en effet vérifiées par le vecteur des bruits ε. Toutes ces hypothèses ne sont pas forcément réalistes ici (sauf le centrage a fortiori), en particulier la non-corrélation qui a peu de chances d'être vérifiée : les vidéos les plus visionnées correspondent souvent aux mêmes catégories de musique.

2. Le premier terme de la matrice $\mathbb{X}'\mathbb{X}$ correspond au nombre de vidéos donc $n = 42$.

3. L'estimateur des MCO de β est défini par $\hat{\beta} = \text{argmin}_{\beta\in\mathbb{R}^3}\|Y - \mathbb{X}\beta\|^2$. Si $\mathcal{E}(\mathbb{X})$ désigne l'espace vectoriel engendré par les vecteurs colonnes de $\mathbb{X}$, le vecteur $\mathbb{X}\hat{\beta}$ est le vecteur de $\mathcal{E}(\mathbb{X})$ dont la distance à Y est minimale. En d'autres termes, $\mathbb{X}\hat{\beta}$ est le projeté orthogonal de Y sur $\mathcal{E}(\mathbb{X})$. La matrice de projection orthogonale sur $\mathcal{E}(\mathbb{X})$ étant égale à $\mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'$, on a $\mathbb{X}\hat{\beta} = \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'Y$, d'où $\mathbb{X}'\mathbb{X}\hat{\beta} = \mathbb{X}'Y$ et $\mathbb{X}'\mathbb{X}$ étant inversible, $\hat{\beta} = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'Y$. Par ailleurs : $\hat{\varepsilon} = Y - \hat{Y}$ est le projeté orthogonal de Y sur l'orthogonal de $\mathcal{E}(\mathbb{X})$.

4. On a $\hat{\beta} = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'Y$ et $\hat{Y} = \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'Y$. Puisque $Y \sim \mathcal{N}_n(\mathbb{X}\beta, \sigma^2 I_n)$, $\hat{\beta} \sim \mathcal{N}_n(\beta, \sigma^2(\mathbb{X}'\mathbb{X})^{-1})$, $\hat{Y} \sim \mathcal{N}_n(\mathbb{X}\beta, \sigma^2\mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}')$, et $\hat{\varepsilon} \sim \mathcal{N}_n(0, \sigma^2(I_n - \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'))$.
$\hat{\beta}$ et $\hat{Y} = \mathbb{X}\hat{\beta}$ ne sont évidemment pas indépendants. En revanche, on peut montrer en appliquant le théorème de Cochran au vecteur ε que $\hat{\beta}$ et $\hat{\varepsilon}$ sont indépendants, ainsi que $\hat{Y}$ et $\hat{\varepsilon}$ (voir cours pour une preuve complète).

5. On a vu que $\hat{\beta} = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'Y$, donc la valeur de $\hat{\beta}_{obs}$ est obtenue par le produit matriciel de $(\mathbb{X}'\mathbb{X})^{-1}$ par $\mathbb{X}'y$.

6. L'équation d'analyse de la variance s'écrit :

$$SCT = SCE + SCR,$$

avec $SCT = \|Y - \bar{Y}\mathbb{1}\|^2$, $SCE = \|\hat{Y} - \bar{Y}\mathbb{1}\|^2$ et $SCR = \|Y - \hat{Y}\|^2$. Elle correspond au théorème de Pythagore dans le triangle rectangle formé du vecteur $Y - \bar{Y}\mathbb{1}$ et des deux vecteurs orthogonaux $\hat{Y} - \bar{Y}\mathbb{1}$ et $Y - \hat{Y}$.
Le cœfficient de détermination est défini par $R^2 = SCE/SCT$ et on a géométriquement $R^2 = \cos^2\theta$ où θ est l'angle formé par les vecteurs $\hat{Y} - \bar{Y}\mathbb{1}$ et $Y - \bar{Y}\mathbb{1}$.

7. $SCE = \|\hat{Y} - \bar{Y}\mathbb{1}\|^2 = \|\hat{Y}\|^2 - \|\bar{Y}\mathbb{1}\|^2$ par Pythagore et donc $SCE = \hat{\beta}'\mathbb{X}'\mathbb{X}\hat{\beta} - n\bar{Y}^2$. On a ainsi $SCE_{obs} = (0.011, 1.003, -0.877)(\mathbb{X}'\mathbb{X})(0.011, 1.003, -0.877)' - 42(7.686/42)^2 = 2.175$.

8. $SCT = SCE/R^2$ donc $SCT_{obs} = 2.217$, et $SCR = SCT - SCE$ donc $SCR_{obs} = 0.042$.

9. Un estimateur sans biais de σ^2 est donné par $\widehat{\sigma^2} = SCR/39$, et $\widehat{\sigma^2}_{obs} = 0.042/39 = 0.0011$.

10. Test de $(H_0) : \beta_1 = \beta_2 = 0$ contre (H_1) : il existe $j \in \{1, 2\}$, $\beta_j \neq 0$.
Statistique de test : $F(Y) = \frac{R^2}{1-R^2}\frac{39}{2}$ qui suit sous (H_0) la loi $\mathcal{F}(2, 39)$.
La région critique du test est donnée par $\mathcal{R}_{(H_0)} = \{y, F(y) > f_{2,39}(0.95)\}$.
On a $f_{2,39}(0.95) = 3.238$. Or $R^2_{obs} = 0.981$, d'où $F(y) = 1006.82$: on rejette clairement (H_0) pour un niveau 5% !

11. Un intervalle de prédiction pour Y_{n+1} de niveau de confiance 95% est donné par

$$\hat{I}^p = \left[\hat{Y}^p_{n+1} - t_{39}(0.975)\sqrt{\widehat{\sigma^2}(1 + (1, 0.4, 0.02)(\mathbb{X}'\mathbb{X})^{-1}(1, 0.4, 0.02)')}; \right.$$
$$\left. \hat{Y}^p_{n+1} + t_{39}(0.975)\sqrt{\widehat{\sigma^2}(1 + (1, 0.4, 0.02)(\mathbb{X}'\mathbb{X})^{-1}(1, 0.4, 0.02)')} \right].$$

On a $(\hat{Y}^p_{n+1})_{obs} = (1, 0.4, 0.02)\hat{\beta}_{obs} = 0.39466$, d'où $\hat{I}^p_{obs} = [0.326; 0.463]$.

12. Pour le test de $(H_0) : \beta_1 \leq 0$ contre $(H_1) : \beta_1 > 0$, on rejette (H_0) lorsque la statistique de test $T_1(Y) = \hat{\beta}_1 / \sqrt{0.544\widehat{\sigma^2}}$ est supérieure à $t_{39}(0.95) = 1.685$. Or ici $T_1(y) = 41$ donc on rejette (H_0) au profit de (H_1). Pour le test de $(H_0) : \beta_2 \geq 0$ contre $(H_1) : \beta_2 < 0$, on rejette (H_0) lorsque la statistique de test $T_2(Y) = \hat{\beta}_2 / \sqrt{26.043\widehat{\sigma^2}}$ est inférieure à $t_{39}(0.05) = -1.685$. Or ici $T_2(y) = -5.18$ donc on rejette (H_0) au profit de (H_1) là aussi pour un niveau 5%.
Ces résultats ne sont certes pas très surprenants : on peut en effet imaginer que toutes choses égales par ailleurs, plus on a de "J'aime", plus on aura de "Favoris", et moins on a de "Je n'aime pas", plus on aura de "Favoris". Mais l'interprétation ne peut se faire que "toutes choses égales par ailleurs"... L'interprétation individuelle de chaque cœfficient estimé dans un modèle complet n'est pas conseillée !

13. Si T_0 désigne la statistique du test de significativité de la constante, $T_0(Y) = \hat{\beta}_0 / \sqrt{0.038\widehat{\sigma^2}}$ et $T_0(y) \simeq 1.7$ donc la p-valeur du test est donnée par $p \simeq P_{\beta_0=0}(|T_0(Y)| > 1.7) = 2*(1-P(T \leq 1.7))$, où $T \sim \mathcal{T}(39)$. Ainsi $p \simeq 0.098$, et par exemple pour un niveau 5%, la constante n'est pas significative.

14. Voir cours. En particulier, faire attention à la nouvelle définition du cœfficient de détermination.

15. Les deux cœfficients de détermination sont assez proches l'un de l'autre dans les deux modèles, très bons, mais non comparables, donc on ne peut pas se baser sur ce critère pour choisir un modèle. La constante n'étant pas significative pour un niveau classique 5%, on pourra s'autoriser à travailler dans le modèle sans constante. Le modèle avec constante reste néanmoins plus facile à interpréter.

6.1.5 Sujet 3 (durée : 2h)

Problème : Étude de la criminalité aux États Unis

Une étude sur la criminalité aux États-Unis a permis de relever dans chaque état (le district de Columbia étant considéré ici comme un état), les observations sur l'année 1997 des variables suivantes :

- `murder` : nombre de meurtres commis pour 1 million d'habitants,
- `poverty` : pourcentage de la population vivant sous le seuil de pauvreté,
- `single` : pourcentage de la population vivant dans une famille monoparentale,
- `pctmetro` : pourcentage de la population vivant en zone urbaine,
- `pcths` : pourcentage de la population ayant fait des études supérieures.

Pour chaque état ($i \in \{1, \dots, 51\}$), on note :

- Y_i : `murder` ou le nombre de meurtres commis pour un million d'habitants de l'état i,
- $x_{i,1}$: `poverty` ou le pourcentage de la population sous le seuil de pauvreté de l'état i,
- $x_{i,2}$: `single` ou le pourcentage de la population vivant dans une famille monoparentale de l'état i,
- $x_{i,3}$: `pctmetro` ou le pourcentage de la population en zone urbaine de l'état i,
- $x_{i,4}$: `pcths` ou le pourcentage de la population ayant fait des études supérieures de l'état i.

Partie I : Régression linéaire simple

On considère les modèles suivants : pour $j = 1 \dots 4$,

$(\mathcal{M}_{0,j})$ $\quad Y_i = \beta_0 + \beta_j x_{i,j} + \varepsilon_i \quad$ pour $i \in \{1, \dots, 51\}$,
$(\mathcal{M}_j)$ $\quad Y_i = \beta_j x_{i,j} + \varepsilon_i \quad$ pour $i \in \{1, \dots, 51\}$,

où les ε_i sont des termes d'erreur aléatoires supposés vérifier les conditions standards des modèles de régression linéaire sous hypothèse gaussienne.

1. Qu'induisent les hypothèses faites sur les ε_i pour les Y_i ? Qu'en pensez-vous par rapport aux données étudiées ici ?

2. Quelle condition supplémentaire impose-t-on sur les $x_{i,j}$ dans le modèle $(\mathcal{M}_{0,j})$? Dans le modèle $(\mathcal{M}_j)$? Que se passe-t-il si cette condition n'est pas vérifiée ?

3. Pour les modèles considérés, on a mis en œuvre sous le logiciel R une procédure de régression linéaire simple et on a obtenu les sorties données en Annexe 2.2.

a) Rappeler les définitions des éléments de ces sorties désignés par : `Estimate, Std.Error, t value, Multiple R-squared`.
b) Sur la base de ces sorties, quel modèle de régression linéaire simple choisiriez-vous pour ajuster les données ? Justifiez votre réponse.
c) Tester l'hypothèse (H_0) $\beta_j \leq 0$ contre (H_1) $\beta_j > 0$ au niveau 5% dans le modèle retenu à la question précédente. Que peut-on en déduire ?

Partie II : Régression linéaire multiple

On considère maintenant le modèle de régression linéaire multiple suivant :

$$(\mathcal{M}_{0,1,2}) \quad Y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \varepsilon_i \quad \text{pour } i \in \{1, \dots, 51\}.$$

Ce modèle peut aussi s'écrire sous la forme matricielle : $Y = \mathbb{X}\beta + \varepsilon$, avec

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbb{X} = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} \\ \vdots & \vdots & \vdots \\ 1 & x_{n,1} & x_{n,2} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

On a alors, si y désigne la valeur observée de Y,

$$\mathbb{X}'\mathbb{X} = \begin{pmatrix} 51 & 727.2 & 577.6 \\ 727.2 & 11419.78 & 8502.66 \\ 577.6 & 8502.66 & 6766.64 \end{pmatrix}, \quad (\mathbb{X}'\mathbb{X})^{-1} = \begin{pmatrix} 0.59053 & -0.00113 & -0.04898 \\ -0.00113 & 0.00136 & -0.00161 \\ -0.04898 & -0.00161 & 0.00636 \end{pmatrix},$$

$\mathbb{X}'y = (445.1, 7736.72, 6017.44)'$, et $y'y = 9627.91$.

On suppose que le vecteur des bruits ε est de loi gaussienne centrée de variance-covariance $\sigma^2 I_n$.

1. Rappeler la définition de l'estimateur des moindres carrés ordinaires de β, donner son expression matricielle, et calculer sa valeur sur les observations.

2. Donner une expression de la somme des carrés résiduelle en fonction de $\mathbb{X}'\mathbb{X}$, $\mathbb{X}'Y$ et $Y'Y$, et en déduire que la valeur de l'estimateur sans biais $\widehat{\sigma^2}$ de la variance σ^2 calculée sur les observations est égale à 20.8386.

3. Donner les valeurs de la somme des carrés expliquée et de la somme des carrés totale calculées sur les observations en justifiant les calculs de façon précise.

4. En déduire la valeur du R^2 calculée sur les observations.

5. Construire un test de niveau 5% de $(H_0) : \beta_1 = \beta_2 = 0$ contre $(H_1) : \beta_1 \neq 0$ ou $\beta_2 \neq 0$, et conclure à l'aide de ce test quant à la validité globale du modèle $(\mathcal{M}_{0,1,2})$.

6. Préciser les lois de $\hat{\beta}$ et de $\widehat{\sigma^2}$ et expliquer comment ces lois permettent de construire un test de (non) significativité de chaque variable explicative du modèle.

7. Tester au niveau 5% la (non) significativité de chaque variable explicative du modèle.

8. Construire des intervalles de confiance de niveau de confiance 95% pour β_j pour $j \in \{0, 1, 2\}$ et retrouver les réponses à la question 7.

9. Donner à l'aide des tables données un encadrement ou une majoration des p-valeurs des tests de (non) significativité des variables explicatives du modèle, et retrouver les réponses à la question 7.

Questionnaire à choix multiple

NOM : ..... Prénom : .....

On considère un modèle général de régression linéaire multiple s'écrivant sous la forme matricielle
$Y = \mathbb{X}\beta + \varepsilon$, avec pour $1 \leq p \leq n$:

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbb{X} = \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{1,p-1} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,p-1} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_{p-1} \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

On suppose que les **conditions standards des modèles de régression linéaire et l'hypothèse gaussienne** sont vérifiées. On reprend les notations usuelles du cours, et notamment, on note $\Pi_{\mathbb{X}}$ la matrice de projection orthogonale sur le sous-espace vectoriel $\mathcal{E}(\mathbb{X})$ de $\mathbb{R}^n$ engendré par les vecteurs colonnes composant la matrice $\mathbb{X}$, et $\mathbb{1}$ le vecteur de $\mathbb{R}^n$ égal à $(1, \ldots, 1)'$.

1. Parmi les variables suivantes, lesquelles sont de valeur observée ou calculable sur les observations :

□ Y,
□ ε,
□ $\hat{\varepsilon}$,
□ $\hat{Y}$,
□ $Y - \mathbb{X}\beta$.

2. Quelle(s) propriété(s) l'estimateur des moindres carrés ordinaires $\hat{\beta}$ vérifie-t-il ?

□ Il est linéaire en Y, sans biais.
□ Il est optimal au sens du risque quadratique parmi les estimateurs sans biais.
□ Il est égal à l'estimateur du maximum de vraisemblance.
□ C'est un vecteur gaussien dont la matrice de variance-covariance est diagonale.

3. Parmi les variables suivantes, lesquelles sont indépendantes entre elles ?

□ les Y_i,
□ les $\hat{\beta}_j$,
□ les ε_i,
□ les $\hat{\varepsilon}_i$,
□ $\hat{\beta}$ et $\hat{\varepsilon}$,
□ Y et $\hat{Y}$,
□ $\hat{Y}$ et $\hat{\varepsilon}$.

TSVP

4. L'espace vectoriel $\mathcal{E}(\mathbb{X})$ est de dimension :

- □ n,
- □ $p-1$,
- □ p,
- □ $p+1$,
- □ $n-p$.

5. Parmi les vecteurs suivants, lesquels sont orthogonaux entre eux ?

- □ Y et $\Pi_{\mathbb{X}} Y$,
- □ Y et $\hat{\varepsilon}$,
- □ $\hat{Y}$ et $\hat{\varepsilon}$,
- □ Y et $\mathbb{1}$,
- □ $Y - \bar{Y}\mathbb{1}$ et $\mathbb{1}$,
- □ $\hat{Y} - \bar{Y}\mathbb{1}$ et $\hat{\varepsilon}$.

Extraits de tables statistiques

Table de la loi gaussienne : on donne pour différentes valeurs de $\alpha \in [0,1]$, q_α tel que $P(N \leq q_\alpha) = \alpha$ lorsque $N \sim \mathcal{N}(0,1)$.

α	0.9	0.95	0.975
q_α	1.282	1.645	1.96

Table de la loi de Student : on donne pour différentes valeurs de n et de $\alpha \in [0,1]$, $t_{n,\alpha}$ tel que $P(T \leq t_{n,\alpha}) = \alpha$ lorsque $T \sim \mathcal{T}(n)$ (on rappelle que la loi de Student est symétrique).

α	0.9	0.95	0.975
$t_{47,\alpha}$	1.300	1.678	2.012
$t_{48,\alpha}$	1.299	1.677	2.011
$t_{49,\alpha}$	1.299	1.677	2.010
$t_{50,\alpha}$	1.299	1.676	2.009
$t_{51,\alpha}$	1.298	1.675	2.008

Table de la loi de Student : on donne pour différentes valeurs de q et pour $n = 47$, 48, 49, 50, 51 ou 52 la valeur de $p_{n,q} = P(T \leq q)$ lorsque $T \sim \mathcal{T}(n)$.

q	1.7	1.8	1.9	2	2.1	2.2	2.3	2.4	2.5	2.6	2.7
$p_{n,q}$	0.952	0.961	0.968	0.974	0.979	0.984	0.987	0.99	0.992	0.994	0.995

Table de la loi de Fisher : on donne pour différentes valeurs de (n_1, n_2) et de $\alpha \in [0,1]$, $f_{n_1,n_2,\alpha}$ tel que $P(F \leq f_{n_1,n_2,\alpha}) = \alpha$ lorsque $F \sim \mathcal{F}(n_1, n_2)$.

α	0.025	0.05	0.1	0.9	0.95	0.975
$f_{1,47,\alpha}$	0.001	0.004	0.016	2.815	4.047	5.361
$f_{1,48,\alpha}$	0.001	0.004	0.016	2.813	4.043	5.354
$f_{1,49,\alpha}$	0.001	0.004	0.016	2.811	4.038	5.347
$f_{1,50,\alpha}$	0.001	0.004	0.016	2.809	4.034	5.340
$f_{2,47,\alpha}$	0.025	0.051	0.106	2.419	3.195	3.994
$f_{2,48,\alpha}$	0.025	0.051	0.106	2.417	3.191	3.987
$f_{2,49,\alpha}$	0.025	0.051	0.106	2.414	3.187	3.981
$f_{2,50,\alpha}$	0.025	0.051	0.106	2.412	3.183	3.975
$f_{3,47,\alpha}$	0.071	0.116	0.194	2.204	2.802	3.409
$f_{3,48,\alpha}$	0.071	0.117	0.194	2.202	2.798	3.402
$f_{3,49,\alpha}$	0.071	0.117	0.194	2.199	2.794	3.396
$f_{3,50,\alpha}$	0.071	0.117	0.194	2.197	2.790	3.390

Table de la loi du Khi Deux : on donne pour différentes valeurs de n et de $\alpha \in [0,1]$, $k_{n,\alpha}$ tel que
$P(K \leq k_{n,\alpha}) = \alpha$ lorsque $K \sim \chi^2(n)$.

α	0.025	0.05	0.95	0.975
$k_{47,\alpha}$	29.956	32.268	64.001	67.821
$k_{48,\alpha}$	30.755	33.098	65.171	69.023
$k_{49,\alpha}$	31.555	33.93	66.339	70.222
$k_{50,\alpha}$	32.357	34.764	67.505	71.42
$k_{51,\alpha}$	33.162	35.6	68.669	72.616

6.1.6 Sujet 1 - Éléments de correction

Problème : Étude de la criminalité aux États Unis

Partie I : Régression linéaire simple

1. On suppose que les Y_i sont indépendantes, de loi gaussienne d'espérance linéaire ou affine en $x_{i,j}$, et de même variance σ^2. L'indépendance peut être vérifiée, mais elle semble peu crédible. On a un risque de corrélation spatiale entre les états par exemple. L'hypothèse de normalité peut se vérifier à l'aide d'un test, mais elle peut être crédible, contrairement à l'hypothèse linéaire qui le sera plus ou moins en fonction de j et l'hypothèse d'homoscédasticité qui elle est peu crédible (états du Sud très différents sans doute des états du Nord, états très urbains différents des états plus ruraux...).

2. Dans le modèle $(\mathcal{M}_{0,j})$, on impose qu'il existe au moins un couple (i, i') tel que $i \neq i'$, et $x_{i,j} \neq x_{i',j}$. Dans le modèle $(\mathcal{M}_j)$, on impose que $\sum x_{i,j}^2 \neq 0$ autrement dit que les $x_{i,j}$ ne soient pas tous nuls.
Si la condition n'est pas vérifiée dans le modèle $(\mathcal{M}_{0,j})$, on a un modèle dont la seule variable explicative est la constante, autrement dit on aura $\hat{Y} = \bar{Y}\mathbb{1}$.
Si la condition n'est pas vérifiée dans le modèle $(\mathcal{M}_j)$, cela signifie qu'on suppose que les Y_i constituent un échantillon de la loi gaussienne centrée, de variance σ^2, et que l'on cherche juste à estimer la variance inconnue σ^2.

3. a) Pour chaque $j = 1, \ldots, 4$, `Estimate` correspond à $\hat{\beta}_j$, `Std.Error` à $\sqrt{\widehat{\sigma^2}(\mathbb{X}'\mathbb{X})^{-1}_{j+1,j+1}}$, `t value` à la statistique du test de Student de (non) significativité de la variable $x_{.,j}$ c'est-à-dire $\hat{\beta}_j / \sqrt{\widehat{\sigma^2}(\mathbb{X}'\mathbb{X})^{-1}_{j+1,j+1}}$, `Multiple R-squared` au cœfficient de détermination $R^2 = \|\mathbb{X}\hat{\beta} - \bar{Y}\mathbb{1}\|^2 / \|Y - \bar{Y}\mathbb{1}\|^2$ dans le modèle $(\mathcal{M}_{0,j})$, $R^2_{sc} = \|\mathbb{X}\hat{\beta}\|^2 / \|Y\|^2$ dans le modèle $(\mathcal{M}_j)$.
b) On retient le modèle $\mathcal{M}_{0,2}$ qui est un modèle avec constante, dont toutes les variables sont significatives au niveau 5% et dont le R^2_{obs} est satisfaisant.
c) On teste $(H_0) : \beta_2 \leq 0$ contre $(H_1) : \beta_2 > 0$ dans ce modèle au niveau 5%. On rejette (H_0) lorsque la statistique de test $T(Y) = \hat{\beta}_j / \sqrt{\widehat{\sigma^2}(\mathbb{X}'\mathbb{X})^{-1}_{j+1,j+1}}$ est supérieure à $t_{49}(0.95) = 1.677$. Or ici $T(y) = 11.74$ (cf sortie R) donc on rejette (H_0) au profit de (H_1) pour un niveau 5%. On conclut que le nombre de meurtres dans un état évolue dans le même sens que le pourcentage de personnes vivant dans une famille mono-parentale. On ne peut cependant, sur la base de ces résultats seuls, conclure à une relation de cause à effet. Seuls des éléments de connaissance dépassant cette étude statistique pourraient permettre de conclure à une telle relation.

Partie II : Régression linéaire multiple

1. L'estimateur des MCO de β est défini par $\hat{\beta} = \text{argmin}_{\beta \in \mathbb{R}^3} \|Y - \mathbb{X}\beta\|^2$. Si $\mathcal{E}(\mathbb{X})$ désigne l'espace vectoriel engendré par les vecteurs colonnes de $\mathbb{X}$, le vecteur $\mathbb{X}\hat{\beta}$ est le vecteur de $\mathcal{E}(\mathbb{X})$ dont la distance à Y est minimale. En d'autres termes, $\mathbb{X}\hat{\beta}$ est le projeté orthogonal de Y sur $\mathcal{E}(\mathbb{X})$. La matrice de projection orthogonale sur $\mathcal{E}(\mathbb{X})$ étant égale à $\mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'$, on a $\mathbb{X}\hat{\beta} = \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'Y$, d'où $\mathbb{X}'\mathbb{X}\hat{\beta} = \mathbb{X}'Y$ et $\mathbb{X}'\mathbb{X}$ étant inversible, $\hat{\beta} = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'Y$. On obtient alors $\hat{\beta}_{obs} = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'y = (-40.6318018, 0.3308978, 4.0138012)'$.

2. Par Pythagore, on a $SCR = Y'Y - Y'\mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'Y$ et $\widehat{\sigma^2} = SCR/48$, d'où $\widehat{\sigma^2}_{obs} = y'y - y'\mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'y/48 \simeq 20.8386$.

3. Première méthode : $SCE = \|\mathbb{X}\hat{\beta} - \bar{Y}\mathbb{1}\|^2 = \|\mathbb{X}\hat{\beta}\|^2 - n\bar{Y}^2 = Y'\mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'Y - 51\bar{Y}^2$, d'où $SCE_{obs} = 4743.068$ puis on utilise l'équation d'analyse de la variance pour trouver $SCT_{obs} = SCE_{obs} + SCR_{obs} = 4743.068 + 48 \times 20.8386 = 5743.321$.
Deuxième méthode : $SCT = Y'Y - 51\bar{Y}^2$, d'où $SCT_{obs} = 5743.322$ et ensuite $SCE_{obs} = SCT_{obs} - SCR_{obs} = 5743.322 - 48 \times 20.8386 = 4743.069$.

4. Le cœfficient de détermination est défini par $R^2 = SCE/SCT$, d'où $R^2_{obs} = 0.82584$.

5. Test de $(H_0) : \beta_1 = \beta_2 = 0$ contre (H_1) : il existe $j \in \{1, 2\}$, $\beta_j \neq 0$.
Statistique de test : $F(Y) = \frac{R^2}{1-R^2}\frac{48}{2}$ qui suit sous (H_0) la loi $\mathcal{F}(2, 48)$.
La région critique du test est donnée par $\mathcal{R}_{(H_0)} = \{y, F(y) > f_{2,48}(0.95)\}$.
On a $f_{2,48}(0.95) = 3.191$. Or $R^2_{obs} = 0.82584$, d'où $F(y) = 113.8$: on rejette clairement (H_0) pour un niveau 5% !

6. On a $\hat{\beta} = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'Y$ et $\hat{Y} = \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'Y$. Puisque $Y \sim \mathcal{N}_n(\mathbb{X}\beta, \sigma^2 I_n)$, $\hat{\beta} \sim \mathcal{N}_n(\beta, \sigma^2(\mathbb{X}'\mathbb{X})^{-1})$. Puisque $48 * \widehat{\sigma^2}/\sigma^2 = \|\hat{\varepsilon}\|^2$, et que $\hat{\varepsilon} = \varepsilon - \Pi_{\mathbb{X}}\varepsilon$, par le théorème de Cochran, $48 * \widehat{\sigma^2}/\sigma^2 \sim \chi^2(48)$. De plus, $\hat{\beta} = \beta + (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\Pi_{\mathbb{X}}\varepsilon$, et comme par le théorème de Cochran, $\Pi_{\mathbb{X}}\varepsilon$ et $\varepsilon - \Pi_{\mathbb{X}}\varepsilon$ sont indépendants, on en déduit que les variables $\hat{\beta}_j$ et $\widehat{\sigma^2}$ sont indépendantes, puis que sous (H_0), $T_j(Y) = \hat{\beta}_j / \sqrt{\widehat{\sigma^2}(\mathbb{X}'\mathbb{X})^{-1}_{j+1,j+1}}$, que l'on choisira comme statistique de test, suit une loi de Student à 48 degrés de liberté.

7. Test de $(H_0) : \beta_j = 0$ contre $(H_1) : \beta_j \neq 0$.
Statistique de test : $T_j(Y) = \hat{\beta}_j / \sqrt{\widehat{\sigma^2}(\mathbb{X}'\mathbb{X})^{-1}_{j+1,j+1}}$.
Loi de $T_j(Y)$ sous (H_0) : Student à 48 degrés de liberté.
Région critique : $R_{(H_0)} = \{y, |T_j(y)| > t_{48}(0.975)\} = \{y, |T_j(y)| > 2.011\}$.
On a $T_0(y) = -11.583$, $T_1(y) = 1.966$, et $T_2(y) = 11.025$ donc seule la variable `poverty` n'est pas significative au niveau 5%.

8. Un intervalle de confiance pour β_j est donné par

$$IC(\beta_j) = \left[\hat{\beta}_j - 2.011\sqrt{\widehat{\sigma^2}(\mathbb{X}'\mathbb{X})^{-1}_{j+1,j+1}}; \hat{\beta}_j + 2.011\sqrt{\widehat{\sigma^2}(\mathbb{X}'\mathbb{X})^{-1}_{j+1,j+1}}\right],$$

doù $IC(\beta_0)_{obs} = [-47.69; -33.58]$, $IC(\beta_1)_{obs} = [-0.0076; 0.669]$, et $IC(\beta_3)_{obs} = [3.282; 4.746]$. On a $0 \notin IC(\beta_0)_{obs}$, $0 \in IC(\beta_1)_{obs}$ et $0 \notin IC(\beta_3)_{obs}$, donc on retrouve les conclusions de la question 7.

9. On a vu que $T_0(y) = -11.583$ donc la p-valeur du test correspondant est donnée par $p_0 \simeq P_{\beta_0=0}(|T_0(Y)| > 11.58) = 2 * (1 - P(T \leq 11.583))$, où $T \sim \mathcal{T}(48)$. D'après les tables on obtient $p_0 << 2 * (1 - 0.995) = 0.01$. Ainsi $p_0 < 0.05$, et on retrouve que pour un niveau 5% par exemple (même beaucoup moins), la constante est significative.
On a vu que $T_1(y) = 1.966$ donc la p-valeur du test est donnée par $p_1 \simeq P_{\beta_1=0}(|T_1(Y)| > 1.966) = 2 * (1 - P(T \leq 1.966))$, où $T \sim \mathcal{T}(48)$. D'après les tables, on obtient $0.968 \leq P(T \leq 1.966) \leq 0.974$, et $0.052 \leq p_1 \leq 0.064$ en particulier, $p > 0.05$ et on retrouve le résultat de la question 7 pour la variable `poverty`.
Enfin, on a vu que $T_2(y) = 11.025$ donc la p-valeur du test est donnée par $p_2 \simeq P_{\beta_2=0}(|T_2(Y)| > 11.025) = 2 * (1 - P(T \leq 11.025))$, où $T \sim \mathcal{T}(48)$. D'après les tables on obtient $p_2 << 2 * (1 - 0.995) = 0.01$. Ainsi $p_2 < 0.05$, et on retrouve le résultat de la question 7 pour la variable `single`.

Questionnaire à choix multiple

1. Parmi les variables suivantes, lesquelles sont de valeur observée ou calculable sur les observations :

☒ Y,
☐ ε,
☒ $\hat{\varepsilon}$,
☒ $\hat{Y}$,
☐ $Y - \mathbb{X}\beta$.

2. Quelle(s) propriété(s) l'estimateur des moindres carrés ordinaires $\hat{\beta}$ vérifie-t-il ?

☒ Il est linéaire en Y, sans biais.
☐ Il est optimal au sens du risque quadratique parmi les estimateurs sans biais.
☒ Il est égal à l'estimateur du maximum de vraisemblance.
☐ C'est un vecteur gaussien dont la matrice de variance-covariance est diagonale.

3. Parmi les variables suivantes, lesquelles sont indépendantes entre elles ?

☒ les Y_i,
☐ les $\hat{\beta}_j$,
☒ les ε_i,
☐ les $\hat{\varepsilon}_i$,
☒ $\hat{\beta}$ et $\hat{\varepsilon}$,
☐ Y et $\hat{Y}$,
☒ $\hat{Y}$ et $\hat{\varepsilon}$.

4. L'espace vectoriel $\mathcal{E}(\mathbb{X})$ est de dimension :

☐ n,
☐ $p - 1$,
☒ p,
☐ $p + 1$,
☐ $n - p$.

5. Parmi les vecteurs suivants, lesquels sont orthogonaux entre eux ?

☐ Y et $\Pi_{\mathbb{X}} Y$,
☐ Y et $\hat{\varepsilon}$,
☒ $\hat{Y}$ et $\hat{\varepsilon}$,
☐ Y et $\mathbb{1}$,
☒ $Y - \bar{Y}\mathbb{1}$ et $\mathbb{1}$,
☒ $\hat{Y} - \bar{Y}\mathbb{1}$ et $\hat{\varepsilon}$.

6.2 Examens terminaux

6.2.1 Sujet 1 (durée : 3h)

2010 : Année de la biodiversité

L'abondance de certaines espèces de papillons est reconnue comme un indicateur de la biodiversité. Pouvoir comprendre et éventuellement prédire cette abondance est donc un enjeu d'importance en écologie.

Le Moiré blanc-fascié ou *Erebia ligea* est une espèce de papillon protégée et menacée d'extinction en Wallonie notamment. On a réalisé le comptage de cette espèce dans 20 stations de Wallonie, pour lesquelles on a également relevé les valeurs de certaines variables écologiques ou géographiques comme la surface, l'humidité, l'altitude ou le type de paysage.

Partie I : Régression linéaire multiple

On considère dans cette partie le modèle suivant :

$$\ln L_i = \beta_0 + \beta_1 S_i + \beta_2 H_i + \beta_3 A_i + \varepsilon_i \quad \text{pour } i = 1 \dots 20,$$

où L_i est une variable aléatoire représentant l'abondance de l'*Erebia ligea* dans la $i^{\text{ème}}$ station, S_i, H_i et A_i désignent respectivement la surface, l'humidité et l'altitude de cette $i^{\text{ème}}$ station, les ε_i sont des termes d'erreur aléatoires tels que

- $E[\varepsilon_i] = 0$ pour tout i,
- $\text{cov}(\varepsilon_i, \varepsilon_j) = 0$ pour tout $i \neq j$,
- $\text{var}(\varepsilon_i) = \sigma^2$ pour tout i.

Ce modèle s'écrit sous la forme matricielle classique :

$$Y = \mathbb{X}\beta + \varepsilon,$$

avec $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)'$ et $\varepsilon = (\varepsilon_1, \dots, \varepsilon_{20})'$.

On a

$$\mathbb{X}'\mathbb{X} = \begin{pmatrix} 20 & 226.8 & 54.1 & 12.7 \\ 226.8 & 3135.8 & 757.06 & 133.98 \\ 54.1 & 757.06 & 190.97 & 32.339 \\ 12.7 & 133.98 & 32.339 & 9.3258 \end{pmatrix} \quad \text{et} \quad (\mathbb{X}'\mathbb{X})^{-1} = \begin{pmatrix} 0.9711 & -0.0541 & 0.0769 & -0.8115 \\ -0.0541 & 0.011 & -0.0338 & 0.0333 \\ 0.0769 & -0.0338 & 0.1281 & -0.064 \\ -0.8115 & 0.0333 & -0.064 & 0.9557 \end{pmatrix}.$$

Par ailleurs, $SCR_{obs} = 3.4499$ et si y désigne la valeur observée de Y, $\mathbb{X}'y = (40.7378, 527.2645, 111.4137, 22.89)'$.

Question préliminaire : Préciser Y, $\mathbb{X}$, et les hypothèses faites sur Y.

1. Après avoir rappelé la définition de l'estimateur des moindres carrés ordinaires $\hat{\beta}$ de β, son interprétation géométrique et son expression matricielle, calculer sa valeur.

2. Quelles sont les propriétés statistiques connues de $\hat{\beta}$ (biais, matrice de variance-covariance $\text{Var}(\hat{\beta})$, optimalité...) ?

3. Déterminer un estimateur $\widehat{\sigma^2}$ sans biais de la variance σ^2, donner sa valeur et en déduire une estimation sans biais de la matrice de variance-covariance $\text{Var}(\hat{\beta})$.

4. On suppose maintenant que $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$.

a) Quels sont les estimateurs du maximum de vraisemblance de β et de la variance σ^2 ? Quel est le lien entre ces estimateurs et les estimateurs $\hat{\beta}$ et $\widehat{\sigma^2}$?

b) Donner les lois de $\hat{\beta}$ et $\widehat{\sigma^2}$. Les estimateurs $\hat{\beta}$ et $\widehat{\sigma^2}$ sont-ils indépendants ? Justifier précisément la réponse.

c) Construire une région de confiance simultanée pour $(\beta_1, \beta_2, \beta_3)$ de niveau de confiance 95%. Quelle est la forme de cette région ?

On donne pour cette question :

$$\begin{pmatrix} 0.011 & -0.0338 & 0.0333 \\ -0.0338 & 0.1281 & -0.064 \\ 0.0333 & -0.064 & 0.9557 \end{pmatrix}^{-1} = \begin{pmatrix} 558.654 & 142.445 & -9.926 \\ 142.445 & 44.397 & -1.99 \\ -9.926 & -1.99 & 1.259 \end{pmatrix}.$$

d) Déduire de la question précédente un test de significativité globale du modèle au niveau 5%. Quelle est la conclusion du test ?

e) Donner un encadrement des p valeurs des tests de significativité des différentes variables explicatives. Quelles sont les variables explicatives significatives au niveau 5% ?

Partie II : Sélection de variables et détection des écarts au modèle

Une généralisation du modèle précédent est le modèle de régression polynômiale (sans interaction) suivant :

$$\ln L_i = \beta_0 + \beta_1 S_i + \beta_2 H_i + \beta_3 A_i + \beta_4 S_i^2 + \beta_5 H_i^2 + \beta_6 A_i^2 + \varepsilon_i \quad \text{pour } i = 1 \dots 20.$$

On suppose que $\varepsilon = (\varepsilon_1, \dots, \varepsilon_{20})' \sim \mathcal{N}(0, \sigma^2 I)$.

1. La somme des carrés résiduelle pour ce nouveau modèle a pour valeur 2.3658. Construire un test permettant de choisir entre ce nouveau modèle et le précédent. Quelle est la conclusion du test au niveau 5% ?

2. Pour pouvoir mieux prédire l'abondance de l'*Erebia Ligea*, on réalise à l'aide du logiciel SAS une procédure de sélection de variables backward à partir du modèle de régression polynômiale. Les sorties de la procédure sont fournies en Annexe 3.1.

a) Définir les différents critères de sélection de variables pouvant être utilisés pour une procédure de type backward. Quel est le critère utilisé dans la procédure SAS ?

b) Le modèle retenu par la procédure SAS est le suivant :

$$\ln L_i = \beta_0 + \beta_1 S_i + \beta_5 H_i^2 + \varepsilon_i \quad \text{pour } i = 1 \dots 20. \tag{6.1}$$

Expliquer ce résultat en justifiant et en analysant pas à pas les sorties fournies en Annexe 3.1.

3. On fournit en Annexes 3.2 et 3.3 plusieurs résultats numériques ou graphiques obtenus après mise en œuvre de la régression retenue à la question précédente.

a) Définir les différentes quantités apparaissant dans ces résultats. Dire en particulier si ces quantités permettent de détecter d'éventuels écarts au modèle, des données atypiques, ayant un effet levier ou influentes, et si le fichier étudié contient de telles données. Si oui, quelles sont ces données ?

b) Quelle décision finale prendriez-vous quant à ces données et à la validité du modèle ?

Partie III : Analyse de la variance à un facteur

On souhaite savoir si une nouvelle variable qualitative "paysage" précisant le type de paysage de la station a un effet sur l'abondance de l'*Erebia ligea*.
Le facteur paysage possède quatre niveaux, et on dispose pour ces quatre niveaux (notés de 1 à 4) de $n_1 = 7$, $n_2 = 3$, $n_3 = 5$ et $n_4 = 5$ observations respectivement. On note $(L_{ij})_{j=1\dots n_i}$ les abondances correspondant à un paysage de type i ($i = 1 \dots 4$), et $(l_{ij})_{j=1\dots n_i}$ les observations correspondantes.

On a relevé les valeurs suivantes : $\sum_{j=1}^{7} l_{1j} = 108$, $\sum_{j=1}^{3} l_{2j} = 67$, $\sum_{j=1}^{5} l_{3j} = 50$ et $\sum_{j=1}^{5} l_{4j} = 101$, $\sum_{j=1}^{7} l_{1j}^2 = 2200$, $\sum_{j=1}^{3} l_{2j}^2 = 2449$, $\sum_{j=1}^{5} l_{3j}^2 = 2032$, $\sum_{j=1}^{5} l_{4j}^2 = 3547$.

On considère un modèle d'analyse de la variance à un facteur sous contrainte identifiante de type analyse par cellule.

1. Ecrire le modèle de façon analytique, puis de façon matricielle en précisant les hypothèses faites sur ce modèle.

2. Quelles sont les valeurs des estimateurs des cœfficients du modèle ?

3. Donner le tableau d'analyse de la variance.

4. Montrer que le facteur paysage n'a pas d'effet significatif sur l'abondance au niveau 5% sous hypothèse gaussienne.

5. Cela signifie-t-il nécessairement que les abondances pour deux paysages différents pris au hasard ne sont pas significativement différentes au niveau 5% ? Expliquer.

6. Quels autres modèles d'analyse de la variance à un facteur pourrait-on considérer ici ? Le test construit dans la question 4 est-il valide dans ces modèles ? Justifier précisément la réponse.

Partie IV : Moindres carrés généralisés

On reprend le modèle défini dans la Partie II par l'équation (6.1) :

$$\ln L_i = \beta_0 + \beta_1 S_i + \beta_5 H_i^2 + \varepsilon_i \quad \text{pour } i = 1 \dots 20,$$

avec $\varepsilon = (\varepsilon_1, \dots, \varepsilon_{20})' \sim \mathcal{N}(0, \sigma^2 I)$.
On suppose ici qu'on dispose seulement des valeurs moyennes sur deux ou quatre stations des $\ln L_i$, S_i et H_i^2. On note ces moyennes $Z_j = \frac{1}{2}\sum_{i=2j-1}^{2j} \ln L_i$, $X_j^1 = \frac{1}{2}\sum_{i=2j-1}^{2j} S_i$, $X_j^2 = \frac{1}{2}\sum_{i=2j-1}^{2j} H_i^2$ pour $j \in \{1, \dots, 8\}$ et $Z_9 = \frac{1}{4}\sum_{i=17}^{20} \ln L_i$, $X_9^1 = \frac{1}{4}\sum_{i=17}^{20} S_i$, $X_9^2 = \frac{1}{4}\sum_{i=17}^{20} H_i^2$.
On a alors pour tout $j = 1 \dots 9$,

$$Z_j = \gamma_0 + \gamma_1 X_j^1 + \gamma_2 X_j^2 + \eta_j,$$

avec $\gamma_0 = \beta_0$, $\gamma_1 = \beta_1$ et $\gamma_2 = \beta_5$, $\eta_j = \frac{1}{2}\sum_{i=2j-1}^{2j} \varepsilon_i$ pour $j \in \{1, \dots, 8\}$, $\eta_9 = \frac{1}{4}\sum_{i=17}^{20} \varepsilon_i$.
Ce modèle s'écrit matriciellement

$$Z = \bar{\mathbb{X}}\gamma + \eta,$$

avec $\gamma = (\gamma_0, \gamma_1, \gamma_2)'$ et $\eta = (\eta_1, \dots, \eta_9)'$.

1. La matrice de variance-covariance de η vérifie $\text{Var}(\eta) = \sigma^2\Omega$. Donner la matrice Ω, ainsi qu'une matrice P inversible telle que $PP' = \Omega$.

2. Pourquoi n'est-il plus judicieux d'utiliser la méthode des moindres carrés ordinaires pour estimer γ dans ce modèle ?

3. En considérant $P^{-1}Z$, $P^{-1}\bar{\mathbb{X}}$ et $P^{-1}\eta$, montrer que l'on obtient un nouveau modèle vérifiant les conditions standards d'un modèle de régression linéaire multiple.

4. En déduire un estimateur pertinent de γ.

5. Montrer que cet estimateur est linéaire sans biais, de variance minimale parmi les estimateurs linéaires sans biais de γ.

Extraits de tables statistiques

Table de la loi gaussienne : on donne pour différentes valeurs de $\alpha \in [0,1]$, q_α tel que $P(N \leq q_\alpha) = \alpha$ lorsque $N \sim \mathcal{N}(0,1)$.

α	0.9	0.95	0.975
q_α	1.282	1.645	1.96

Table de la loi de Student : on donne pour différentes valeurs de n et de $\alpha \in [0,1]$, $t_{n,\alpha}$ tel que $P(T \leq t_{n,\alpha}) = \alpha$ lorsque $T \sim \mathcal{T}(n)$ (on rappelle que la loi de Student est symétrique).

α	0.9	0.95	0.975
$t_{15,\alpha}$	1.34	1.75	2.13
$t_{16,\alpha}$	1.34	1.75	2.12
$t_{17,\alpha}$	1.33	1.74	2.11
$t_{18,\alpha}$	1.33	1.73	2.10
$t_{19,\alpha}$	1.33	1.73	2.09
$t_{20,\alpha}$	1.33	1.72	2.09

Table de la loi de Student : on donne pour différentes valeurs de q et pour différentes valeurs de n la valeur de $p_{n,q} = P(T \leq q)$ lorsque $T \sim \mathcal{T}(n)$.

q	1.5	1.6	1.7	1.8	1.9	2	2.1	2.2	2.3	2.4	2.5	5
$p_{16,q}$	0.923	0.935	0.946	0.955	0.962	0.969	0.974	0.979	0.982	0.986	0.988	1
$p_{17,q}$	0.924	0.936	0.946	0.955	0.963	0.969	0.975	0.979	0.983	0.986	0.989	1
$p_{18,q}$	0.925	0.936	0.947	0.956	0.963	0.97	0.975	0.979	0.983	0.986	0.989	1
$p_{19,q}$	0.925	0.937	0.947	0.956	0.964	0.97	0.975	0.98	0.984	0.987	0.989	1
$p_{20,q}$	0.925	0.937	0.948	0.957	0.964	0.97	0.976	0.98	0.984	0.987	0.989	1

Table de la loi de Fisher : on donne pour différentes valeurs de (n_1, n_2) et de $\alpha \in [0,1]$, $f_{n_1,n_2,\alpha}$ tel que $P(F \leq f_{n_1,n_2,\alpha}) = \alpha$ lorsque $F \sim \mathcal{F}(n_1, n_2)$.

α	0.025	0.05	0.95	0.975
$f_{1,16,\alpha}$	0.001	0.004	4.494	6.115
$f_{1,17,\alpha}$	0.001	0.004	4.451	6.042
$f_{1,18,\alpha}$	0.001	0.004	4.414	5.978
$f_{1,19,\alpha}$	0.001	0.004	4.381	5.922
$f_{1,20,\alpha}$	0.001	0.004	4.351	5.871
$f_{2,16,\alpha}$	0.025	0.051	3.634	4.687
$f_{2,17,\alpha}$	0.025	0.051	3.592	4.619
$f_{2,18,\alpha}$	0.025	0.051	3.555	4.56
$f_{2,19,\alpha}$	0.025	0.051	3.522	4.508
$f_{2,20,\alpha}$	0.025	0.051	3.493	4.461
$f_{3,13,\alpha}$	0.07	0.115	3.411	4.347
$f_{3,14,\alpha}$	0.07	0.115	3.344	4.242
$f_{3,15,\alpha}$	0.07	0.115	3.287	4.153
$f_{3,16,\alpha}$	0.07	0.115	3.239	4.077
$f_{3,17,\alpha}$	0.07	0.115	3.197	4.011
$f_{3,18,\alpha}$	0.07	0.115	3.16	3.954
$f_{3,19,\alpha}$	0.071	0.115	3.127	3.903
$f_{3,20,\alpha}$	0.071	0.115	3.098	3.859
$f_{4,16,\alpha}$	0.116	0.171	3.007	3.729
$f_{4,17,\alpha}$	0.116	0.171	2.965	3.665
$f_{4,18,\alpha}$	0.116	0.172	2.928	3.608
$f_{4,19,\alpha}$	0.117	0.172	2.895	3.559
$f_{4,20,\alpha}$	0.117	0.172	2.866	3.515

6.2.2 Sujet 1 - Éléments de correction

2010 : Année de la biodiversité

Partie I : Régression linéaire multiple

Question préliminaire : $Y = (\ln L_1, \dots, \ln L_{20})'$, $\mathbb{X} = (\mathbb{1}, S, H, A)$, où $\mathbb{1}$ est le vecteur de taille 20 ne contenant que des 1, S le vecteur $(S_1, \dots, S_{20})'$, H le vecteur $(H_1, \dots, H_{20})'$, et A le vecteur $(A_1, \dots, A_{20})'$. On a supposé $\mathbb{E}[Y] = \mathbb{X}\beta$, et $\text{Var}(Y) = \sigma^2 I$ i.e. les Y_i non corrélées et homoscédastiques.

1. L'estimateur des MCO de β est défini par $\hat{\beta} = \text{argmin}_{\beta \in \mathbb{R}^p} \|Y - \mathbb{X}\beta\|^2$. Si $\mathcal{E}(\mathbb{X})$ désigne l'espace vectoriel engendré par les vecteurs colonnes de $\mathbb{X}$, le vecteur $\mathbb{X}\hat{\beta}$ est le vecteur de $\mathcal{E}(\mathbb{X})$ dont la distance à Y est minimale. En d'autres termes, $\mathbb{X}\hat{\beta}$ est le projeté orthogonal de Y sur $\mathcal{E}(\mathbb{X})$. Or on sait que lorsque $\mathbb{X}$ est de plein rang, la matrice de projection orthogonale sur $\mathcal{E}(\mathbb{X})$ est égale à $\Pi_{\mathbb{X}} = \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'$, d'où $\mathbb{X}\hat{\beta} = \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'Y$ et comme $\mathbb{X}$ est de plein rang, on a finalement $\hat{\beta} = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'Y$. D'où la valeur de $\hat{\beta}$: $\hat{\beta}_{obs} = (1.0279467, 0.5924485, -1.8816683, -0.7553206)'$.

2. $\hat{\beta}$ est un estimateur sans biais, sa matrice de variance-covariance est égale à : $\text{Var}(\hat{\beta}) = \sigma^2(\mathbb{X}'\mathbb{X})^{-1}$. Par le théorème de Gauss-Markov, cet estimateur est de variance minimale parmi les estimateurs linéaires sans biais de β.

3. Un estimateur sans biais de σ^2 est donné par $\widehat{\sigma^2} = SCR/(20-4)$ d'où $\widehat{\sigma^2}_{obs} = 0.2156187$. Une estimation sans biais de $\text{Var}(\hat{\beta})$ est donc donnée par

$$\widehat{\sigma^2}_{obs}(\mathbb{X}'\mathbb{X})^{-1} = \begin{pmatrix} 0.20939 & -0.011665 & 0.0166 & -0.175 \\ -0.011665 & 0.002372 & -0.0073 & 0.0072 \\ 0.0166 & -0.0073 & 0.0276 & -0.0138 \\ -0.175 & 0.0072 & -0.0138 & 0.206 \end{pmatrix}.$$

4. a) L'estimateur du maximum de vraisemblance est égal à $\hat{\beta}$, celui de σ^2 à $16\widehat{\sigma^2}/20 = SCR/20$.

b) $\hat{\beta}$ suit une loi gaussienne d'espérance β, de variance $\sigma^2(\mathbb{X}'\mathbb{X})^{-1}$. Par le théorème de Cochran, on montre que $20\widehat{\sigma^2}$ suit une loi du khi-deux à 16 degrés de liberté, et que $\hat{\beta}$ et $\widehat{\sigma^2}$ sont indépendants (voir cours pour la preuve).

c) Une région de confiance simultanée pour $(\beta_1, \beta_2, \beta_3)$ de niveau de confiance 95% est donnée par :

$$\Big\{(\beta_1, \beta_2, \beta_3) \in \overset{3}{\mathbb{R}},\ \frac{1}{3\widehat{\sigma^2}}\Big(558.654(\hat{\beta}_1 - \beta_1)^2 + 44.397(\hat{\beta}_2 - \beta_2)^2 + 1.259(\hat{\beta}_3 - \beta_3)^2$$
$$+ 284.89(\hat{\beta}_1 - \beta_1)(\hat{\beta}_2 - \beta_2) - 19.852(\hat{\beta}_1 - \beta_1)(\hat{\beta}_3 - \beta_3) - 3.98(\hat{\beta}_2 - \beta_2)(\hat{\beta}_3 - \beta_3)\Big) \leq f_{3,16}(0.95)\Big\},$$

avec $\hat{\beta}_{obs} = (1.0279467, 0.5924485, -1.8816683, -0.7553206)'$, $\widehat{\sigma^2}_{obs} = 0.2156187$ et $f_{3,16}(0.95) = 3.239$. Cette région de confiance est un ellipsoïde.

d) Soit $(H_0) : \beta_1 = \beta_2 = \beta_3 = 0$ contre $(H_1) : \beta_1, \beta_2$ ou $\beta_3 \neq 0$. Un test de (H_0) contre (H_1) a pour région critique $\{Y, F(Y) > 3.239\}$, avec

$$F(Y) = \frac{1}{3\widehat{\sigma^2}}\left(558.654\hat{\beta}_1^2 + 44.397\hat{\beta}_2^2 + 1.259\hat{\beta}_3^2 + 284.89\hat{\beta}_1\hat{\beta}_2 - 19.852\hat{\beta}_1\hat{\beta}_3 - 3.98\hat{\beta}_2\hat{\beta}_3\right).$$

On a $F(y) = 61.27$. On rejette donc clairement l'hypothèse nulle. Le modèle est globalement significatif au niveau 5%.

e) Pour la constante, la p valeur du test de significativité est égale à $p = P(|T| > |\hat{\beta}_{0,obs}|/\sqrt{0.20939})$ avec $T \sim \mathcal{T}(16)$ i.e. $p = 2 - 2P(T \leq 2.246) \in [0.036, 0.042]$ donc la constante est significative au niveau 5%.
Pour la première variable explicative potentielle S, la p valeur du test de significativité est égale à $p = P(|T| > |\hat{\beta}_{1,obs}|/\sqrt{0.002372})$ avec $T \sim \mathcal{T}(16)$ i.e. $p = 2 - 2P(T \leq 12.16) \simeq 0$ donc la variable surface est significative au niveau 5%.
Pour la variable humidité H, $p = 2 - 2P(T \leq 11.33) \simeq 0$ donc la variable humidité est significative au niveau 5%.
Pour la variable altitude A, $p = 2 - 2P(T \leq 1.65) \in [0.108, 0.13]$ donc la variable altitude n'est pas significative au niveau 5%.

Partie II : Sélection de variables et détection des écarts au modèle

1. Dans le nouveau modèle, on teste (H_0) : $\beta_4 = \beta_5 = \beta_6 = 0$ contre (H_1) : l'un de ces trois éléments est non nul. On montre que la statistique de test s'écrit $F(Y) = 13(SCR - SCR_c)/(3SCR_c)$, où SCR_c représente la somme des carrés résiduelle du nouveau modèle de régression polynômiale ou modèle complet. Sous (H_0), $F \sim \mathcal{F}(3, 13)$ donc la région critique du test est $\{Y, F(Y) > 3.41\}$, avec $F(y) = 13 * (3.4499 - 2.3658)/(3 * 2.3658) = 1.986$. On accepte (H_0) au niveau 5% i.e. on préfère le premier modèle au modèle complet.

2. On veut voir si un autre sous-modèle du modèle complet serait préférable au premier modèle.

a) On peut utiliser pour n'importe quel type de procédure (exhaustive, backward, forward ou stepwise) les critères du R^2 ajusté (à maximiser), de l'AIC, du BIC, du C_p de Mallows (à minimiser), mais aussi des tests de significativité de sous-modèles. La procédure SAS utilise les tests de significativité.

b) On expliquera le "fonctionnement" de la procédure backward et on pourra donner les statistiques de test à chaque étape et les conclusions des tests.

3. a) Pour les définitions, on renvoie au cours.

b) Le graphe des résidus ainsi que le QQplot ne contredisent pas a priori les hypothèses du modèle.
Une seule donnée aberrante, la 15, au niveau 5% (tous les autres résidus studentisés sont inférieurs à 2.12 en valeur absolue).
Donnée ayant un effet levier : l'observation 20, qui est peu influente.
Donnée 19 : donnée influente.
Il n'y a priori aucune raison de remettre le modèle en cause ni d'envisager de supprimer des données.

Partie III : Analyse de la variance à un facteur

1. $L_{ij} = \alpha_i + \varepsilon_{ij}, \quad \begin{cases} i = 1 \dots 4 \\ j = 1 \dots n_i \end{cases}$,

où les ε_{ij} vérifient les conditions standards :
- (C_1) $\mathbb{E}[\varepsilon_{ij}] = 0$,
- (C_2) $\text{cov}(\varepsilon_{ij}, \varepsilon_{i'j'}) = 0$,
- (C_3) $\text{var}(\varepsilon_{ij}) = \sigma^2$.

Ecriture matricielle : $L = \mathbb{X}\beta + \varepsilon$ avec $\beta = (\alpha_1, \alpha_2, \alpha_3, \alpha_4)'$ et

$$\mathbb{X} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{matrix} \\ \}7 \\ \\ \\ \\ \\ \\ \}5 \\ \\ \\ \}5 \\ \end{matrix}.$$

$\mathbb{X}$ est la matrice d'incidence du facteur paysage.

2. Pour tout $i = 1, 2, 3, 4$, $\hat{\alpha}_i = \bar{L}_{i\bullet}$, d'où $\hat{\alpha}_{1,obs} = 15.42857$, $\hat{\alpha}_{2,obs} = 22.33333$, $\hat{\alpha}_{3,obs} = 10$, $\hat{\alpha}_{4,obs} = 20.2$.

3. Puisque $\mathbb{1} \in \mathcal{E}(\mathbb{X})$ (et ce quelle que soit la contrainte), $SCE = \|\hat{L} - \bar{L}_{\bullet\bullet}\mathbb{1}\|^2 = \sum_{i=1}^4 n_i(\bar{L}_{i\bullet} - \bar{L}_{\bullet\bullet})^2 = \sum_{i=1}^4 n_i \bar{L}_{i\bullet}^2 - 20\bar{L}_{\bullet\bullet}^2$, d'où $SCE_{obs} = 389.02$.
On a $SCR = \|Y - \hat{Y}\|^2 = \sum_{i=1}^4 \sum_{j=1}^{n_i} L_{ij}^2 - \sum_{i=1}^4 n_i \bar{L}_{i\bullet}^2$, d'où $SCR_{obs} = 4525.2$.

Tableau d'analyse de la variance :

Variation	ddl	SC	CM	F	$Pr(> F)$
Facteur paysage	3	389.02	129.7	0.4585	0.715
Résiduelle	16	4525.2	282.8		

4. La p valeur du test étant égale à 0.715, pour un niveau 5%, on accepte l'hypothèse d'absence d'effet du facteur paysage.

5. Le test d'effet du facteur paysage ne débouche pas sur un test de comparaison de deux cœfficients directement.

6. Modèles d'analyse de la variance à un facteur sous d'autres contraintes identifiantes classiques (analyse par cellule de référence, contrainte d'orthogonalité ou de type somme). Le test d'effet du facteur reste le même (voir cours).

Partie IV : Moindres carrés généralisés

1. $\Omega = \text{diag}\left(\frac{1}{2}, \ldots, \frac{1}{2}, \ldots, \frac{1}{4}\right)$, donc on peut prendre $P = \text{diag}\left(\frac{1}{\sqrt{2}}, \ldots, \frac{1}{\sqrt{2}}, \ldots, \frac{1}{2}\right)$.

2. Les termes d'erreur ne sont plus homoscédastiques. On peut certes toujours calculer l'EMCO dans le nouveau modèle. Il est sans biais. Cependant, sous l'hypothèse de normalité, il n'est pas de variance minimale parmi les estimateurs linéaires sans biais dans ce modèle.

3. On a $P^{-1}Z = P^{-1}\bar{X}\beta + P^{-1}\eta$, et $\text{Var}(P^{-1}\eta) = P^{-1}\sigma^2\Omega P^{-1} = \sigma^2 I$ donc on retrouve un modèle de régression linéaire avec les conditions standards.

4. On peut alors calculer l'EMCO dans ce nouveau modèle. Cet estimateur est appelé estimateur des moindres carrés généralisés, et son expression est donnée par $\hat{\gamma}_{MCG} = (\bar{\mathbb{X}}'\Omega^{-1}\bar{\mathbb{X}})^{-1}\bar{\mathbb{X}}'\Omega^{-1}Z$.

5. Par application du théorème de Gauss-Markov dans le nouveau modèle vérifiant les conditions standards, l'estimateur des moindres carrés généralisés est bien sans biais de variance minimale parmi les estimateurs linéaires sans biais de γ.

6.2.3 Sujet 2 (durée : 3h)

Etude de l'espérance de vie

L'espérance de vie est un indicateur influençant de nombreuses décisions politiques notamment sur les plans sanitaire et économique. Pouvoir comprendre et éventuellement prédire cette espérance de vie est donc un enjeu d'importance.

Nous cherchons ici à étudier l'espérance de vie de 114 pays au travers de données sur les consommations moyennes d'alcool et de tabac, le QI moyen, ainsi que le PIB par habitant et l'indice de démocratisation de ces pays.

On introduit pour cela les variables suivantes : pour le $i$ème pays étudié ($i \in \{1, \dots, 114\}$),

- Y_i son espérance de vie,
- $x_{i,1}$ la consommation annuelle moyenne d'alcool de ses adultes de plus de 15 ans (en L d'alcool pur),
- $x_{i,2}$ le nombre annuel moyen de cigarettes consommées par ses habitants,
- $x_{i,3}$ le QI moyen de ses habitants,
- $x_{i,4}$ son PIB par habitant (en dollar international),
- $x_{i,5}$ son indice de démocratisation (note sur 10).

Partie I : Régressions linéaires simples

On considère les modèles de régression linéaire simple suivants :

$$(\mathcal{M}_p) \quad Y_i = \beta_0 + \beta_1 x_{i,p} + \varepsilon_i \quad \text{pour } i \in \{1, \dots, 114\} \quad \text{et} \quad p \in \{1, \dots, 5\}, \text{ et}$$

$$(\mathcal{M}_{\ln p}) \quad Y_i = \beta_0 + \beta_1 \ln(x_{i,p}) + \varepsilon_i \quad \text{pour } i \in \{1, \dots, 114\} \quad \text{et} \quad p \in \{2, 3, 4\},$$

sous conditions standards et sous condition gaussienne.

1. Préciser quelles sont les conditions imposées aux ε_i.

2. Donner l'expression des estimateurs des moindres carrés ordinaires $\hat{\beta}_0$ et $\hat{\beta}_1$ des coefficients de régression β_0 et β_1, puis de l'estimateur $\widehat{\sigma^2}$ usuel de la variance σ^2 des ε_i, par exemple dans le modèle $(\mathcal{M}_1)$.

3. Quelles sont les lois et les propriétés statistiques de ces estimateurs ? Sont-ils indépendants ?

4. Pour chacun des modèles considérés, on a mis en œuvre sous le logiciel R une procédure de régression linéaire simple. On a obtenu les sorties données en Annexe 3.4.

a) Quelles sont les commandes R utilisées pour obtenir ces sorties ?

b) Au vu de ces seules sorties, quels sont les modèles de régression linéaire simple les plus pertinents ? Expliquer.

5. On se penche désormais précisément sur le modèle de régression linéaire simple ($\mathcal{M}_{\ln 3}$). On a tracé le nuage de points correspondant, puis des graphes permettant une analyse précise des résidus et des écarts éventuels au modèle. Les résultats sont reportés en Annexe 3.5.

a) Pour chacun des graphes tracés, préciser :
- les éléments qu'on a exactement représentés sur ce graphe,
- l'information que ces éléments peuvent apporter sur un modèle de régression : remise en cause des hypothèses du modèle, présence de données remarquables,
- l'information réellement apportée par ces éléments sur le modèle étudié ici.

b) Expliquer graphiquement, à partir de la représentation du nuage de points, la nature des données remarquables éventuellement trouvées à la question précédente.

Partie II : Régression linéaire multiple

On considère dans cette partie le modèle de régression linéaire multiple à 2 variables explicatives suivant :

$$(\mathcal{M}_{\ln 3,\ln 2}) \quad Y_i = \beta_0 + \beta_1 \ln(x_{i,3}) + \beta_2 \ln(x_{i,2}) + \varepsilon_i \quad \text{pour } i \in \{1, \dots, 114\}.$$

1. On a mis en œuvre sous le logiciel R une procédure de régression linéaire multiple pour le modèle ($\mathcal{M}_{\ln 3,\ln 2}$). Compléter la sortie obtenue ci-dessous.

```
Coefficients:
             Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)  -179.3450     20.1054      ???   1.08e-14   ???
log(X3)           ???      5.3455   10.077    < 2e-16    ***
log(X2)        1.3953      0.7654      ???        ???   ???
-
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 5.315 on ??? degrees of freedom
Multiple R-Squared: 0.7409, Adjusted R-squared: ???
F-statistic: ??? on ??? and ??? DF, p-value: ???
```

2. Le modèle considéré est-il globalement significatif au niveau 5% ?

3. Quelles sont les variables explicatives significatives au niveau 5% ? Au niveau 10% ?

4. Si l'on considère notamment les résultats obtenus après mise en œuvre des procédures de régression linéaire simple pour les modèles ($\mathcal{M}_{\ln 2}$) et ($\mathcal{M}_{\ln 3}$) (c.f. Partie I, Annexe 3.4), en quoi les résultats obtenus ici sont-ils surprenants au premier abord ? Comment ce phénomène peut-il s'expliquer ?

5. En notant $\mathbb{X}$ la matrice du plan d'expérience associée au modèle ($\mathcal{M}_{\ln 3,\ln 2}$), et $A = (\mathbb{X}'\mathbb{X})^{-1}$, on donne

$$\begin{pmatrix} A_{2,2} & A_{2,3} \\ A_{3,2} & A_{3,3} \end{pmatrix}^{-1} = \begin{pmatrix} 2.361 & 12.573 \\ 12.573 & 115.159 \end{pmatrix}.$$

En déduire une région de confiance simultanée pour $(\beta_1, \beta_2)'$ de niveau de confiance 95% et confirmer à l'aide de cette région de confiance la réponse de la question 2.

6. Déterminer un intervalle de confiance de niveau de confiance 95% pour chaque cœfficient de régression du modèle.

7. Déterminer un intervalle de confiance de niveau de confiance 95% pour la variance des ε_i.

Partie III : Régression linéaire multiple et sélection de variables

On considère finalement le modèle de régression linéaire multiple décrit de la façon suivante :

$$(\mathcal{M}) \quad Y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 \ln(x_{i,2}) + \beta_4 x_{i,3} + \beta_5 \ln(x_{i,3}) + \beta_6 x_{i,4} + \beta_7 \ln(x_{i,4}) + \beta_8 x_{i,5} + \varepsilon_i,$$

pour $i \in \{1 \dots 114\}$.

1. On met en œuvre des procédures de sélection de variables exhaustives à l'aide des logiciels SAS et R, pour les critères du R^2 ajusté, du C_p, et du BIC. On obtient les sorties données en Annexe 3.6.

a) Quel est le lien entre l'utilisation de ces différents critères pour une sélection de variables et les tests de validité de sous-modèles ?

b) Expliquer les sorties données en Annexe 3.6.

c) Quel modèle peut-on finalement retenir ?

2. On met en œuvre une procédure de sélection de variables `backward`, puis une procédure de sélection de variables `forward` avec le logiciel SAS. On obtient les sorties données en Annexe 3.7.

a) Quel critère de sélection de variables est utilisé dans ces procédures ?

b) Quels sont les critères d'arrêt de ces procédures ?

c) Décrire pas à pas la procédure de sélection `backward`.

d) Commenter les sorties obtenues pour les modèles retenus par les deux procédures. Comparer au modèle retenu à la question 1.

3. Après une étude soigneuse des résidus et des éventuelles données remarquables pour le modèle retenu à la question 1, on ne remet pas en cause ce modèle. On obtient dans ce modèle, pour la France qui a une consommation annuelle moyenne d'alcool par adulte de plus de 15 ans de 12.48 L, un nombre annuel moyen de cigarettes consommées par habitant de 2058, un QI moyen de ses habitants de 98, un PIB par habitant de 34092.259 dollars, et un indice de démocratisation de 7.77 (seulement !), un intervalle de prédiction pour l'espérance de vie calculé sur les observations égal à $[70.2; 86.8]$.

Expliquer comment cet intervalle a été calculé.

Extraits de tables statistiques

Table de la loi gaussienne : on donne pour différentes valeurs de $\alpha \in [0,1]$, q_α tel que $P(N \leq q_\alpha) = \alpha$ lorsque $N \sim \mathcal{N}(0,1)$.

α	0.9	0.95	0.975
q_α	1.282	1.645	1.96

Table de la loi de Student : on donne pour différentes valeurs de n et de $\alpha \in [0,1]$, $t_{n,\alpha}$ tel que $P(T \leq t_{n,\alpha}) = \alpha$ lorsque $T \sim \mathcal{T}(n)$ (on rappelle que la loi de Student est symétrique).

α	0.9	0.95	0.975
$t_{105,\alpha}$	1.290	1.659	1.983
$t_{106,\alpha}$	1.290	1.659	1.983
$t_{107,\alpha}$	1.290	1.659	1.982
$t_{108,\alpha}$	1.289	1.659	1.982
$t_{109,\alpha}$	1.289	1.659	1.982
$t_{110,\alpha}$	1.289	1.659	1.982
$t_{111,\alpha}$	1.289	1.659	1.982
$t_{112,\alpha}$	1.289	1.659	1.981
$t_{113,\alpha}$	1.289	1.658	1.981
$t_{114,\alpha}$	1.289	1.658	1.981

Table de la loi de Student : on donne pour différentes valeurs de q et pour différentes valeurs de n la valeur de $p_{n,q} = P(T \leq q)$ lorsque $T \sim \mathcal{T}(n)$.

q	1.8	1.81	1.82	1.83	1.84	1.85	1.86	1.87	1.88	1.89	1.9
$p_{111,q}$	0.963	0.963	0.964	0.965	0.966	0.967	0.967	0.968	0.969	0.969	0.970
$p_{112,q}$	0.963	0.964	0.964	0.965	0.966	0.967	0.967	0.968	0.969	0.969	0.970
$p_{113,q}$	0.963	0.964	0.964	0.965	0.966	0.967	0.967	0.968	0.969	0.969	0.970

Table de la loi de Fisher : on donne pour différentes valeurs de (n_1, n_2) et de $\alpha \in [0,1]$, $f_{n_1,n_2,\alpha}$ tel que $P(F \leq f_{n_1,n_2,\alpha}) = \alpha$ lorsque $F \sim \mathcal{F}(n_1, n_2)$.

α	0.025	0.05	0.1	0.5	0.95	0.975
$f_{1,111,\alpha}$	0.001	0.004	0.016	0.458	3.927	5.163
$f_{1,112,\alpha}$	0.001	0.004	0.016	0.458	3.926	5.162
$f_{2,111,\alpha}$	0.025	0.051	0.105	0.697	3.078	3.814
$f_{2,112,\alpha}$	0.025	0.051	0.105	0.697	3.077	3.813
$f_{3,111,\alpha}$	0.072	0.117	0.194	0.794	2.686	3.236
$f_{3,112,\alpha}$	0.072	0.117	0.194	0.793	2.686	3.235

Table de la loi de Fisher : on donne pour différentes valeurs de q et pour différentes valeurs de (n_1, n_2) la valeur de $p_{n_1,n_2,q} = P(F \leq q)$ lorsque $F \sim \mathcal{F}(n_1, n_2)$.

q	10	20	30
$p_{1,111,q}$	0.99798	0.99998	1
$p_{1,112,q}$	0.99799	0.99998	1
$p_{1,113,q}$	0.99799	0.99998	1
$p_{2,111,q}$	0.9999	1	1
$p_{2,112,q}$	0.9999	1	1
$p_{2,113,q}$	0.9999	1	1
$p_{3,111,q}$	0.99999	1	1
$p_{3,112,q}$	0.99999	1	1
$p_{3,113,q}$	0.99999	1	1

Table de la loi du Khi Deux : on donne pour différentes valeurs de n et de $\alpha \in [0, 1]$, $k_{n,\alpha}$ tel que
$P(K \leq k_{n,\alpha}) = \alpha$ lorsque $K \sim \chi^2(n)$.

α	0.025	0.05	0.95	0.975
$k_{111,\alpha}$	83.735	87.681	136.591	142.049
$k_{112,\alpha}$	84.604	88.570	137.701	143.180
$k_{113,\alpha}$	85.473	89.461	138.811	144.311
$k_{114,\alpha}$	86.342	90.351	139.921	145.441

6.2.4 Sujet 2 : Éléments de correction

Partie I : Régressions linéaires simples

1. On impose $E[\varepsilon_i] = 0\ \forall i = 1 \cdots 114$ (centrage), $\text{cov}(\varepsilon_i, \varepsilon_j) = \sigma^2 \delta_i^j$ (non corrélation et homoscédasticité), puis $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_{114})$ (hypothèse gaussienne sur le vecteur ε).

2. Les estimateurs des MCO sont donnés par :

$$\left\{ \begin{array}{l} \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}_{.,1} \\ \hat{\beta}_1 = \frac{\sum_{i=1}^{114} x_{i,1} Y_i - \sum_{i=1}^{114} x_{i,1} \bar{Y}}{\sum_{i=1}^{114} x_{i,1}^2 - \sum_{i=1}^{114} x_{i,1} \bar{x}_{.,1}} = \frac{\sum_{i=1}^{114} x_{i,1}(Y_i - \bar{Y})}{\sum_{i=1}^{114}(x_{i,1} - \bar{x}_{.,1})^2} = \frac{\sum_{i=1}^{114}(x_{i,1} - \bar{x}_{.,1}) Y_i}{\sum_{i=1}^{114}(x_{i,1} - \bar{x}_{.,1})^2} = \frac{\sum_{i=1}^{114}(x_{i,1} - \bar{x}_{.,1})(Y_i - \bar{Y})}{\sum_{i=1}^{114}(x_{i,1} - \bar{x}_{.,1})^2} \end{array} \right. ,$$

l'estimateur usuel de la variance est donné par $\widehat{\sigma^2}^2 = \frac{1}{112} \sum_{i=1}^{114} \hat{\varepsilon}_i^2$, où $\hat{\varepsilon}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i,1}$.

3. Lois des estimateurs sous l'hypothèse gaussienne :
$\hat{\beta}_0 \sim \mathcal{N}\left(\beta_0, \frac{\sum_{i=1}^{114} x_{i,1}^2}{114 \sum_{i=1}^{114}(x_{i,1} - \bar{x}_{.,1})^2} \sigma^2\right)$, $\hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{1}{\sum_{i=1}^{114}(x_{i,1} - \bar{x}_{.,1})^2} \sigma^2\right)$ et $112\widehat{\sigma^2}/\sigma^2 \sim \chi^2(112)$.
Indépendance ? $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)'$ et $\widehat{\sigma^2}$ sont indépendants par le théorème de Cochran. En revanche, $\hat{\beta}_0$ et $\hat{\beta}_1$ sont indépendants si et seulement si $\bar{x}_{.,1} = 0$.
$\hat{\beta}$ est un estimateur linéaire (en les Y_i), sans biais de $\beta = (\beta_0, \beta_1)'$, de variance minimale parmi les estimateurs linéaires sans biais de β (théorème de Gauss-Markov).
$\widehat{\sigma^2}$ est un estimateur sans biais de σ^2.

4. Tous les modèles considérés contiennent le même nombre de variables explicatives. On peut donc pour les comparer utiliser le critère du R^2. Ici, les trois modèles les plus pertinents sont les modèles $(\mathcal{M}_3)$, $(\mathcal{M}_{\ln 3})$ et $(\mathcal{M}_{\ln 4})$ qui ont des R^2 respectivement égaux à 0.7205, 0.7331 et 0.716. On aura une préférence pour le modèle $(\mathcal{M}_{\ln 3})$, dont le R^2 est le plus élevé et dont les toutes les variables explicatives sont significatives à 5% par exemple.

5. a) Graphe 1 : nuage de points et droite de régression estimée. On peut y vérifier la tendance linéaire, et regarder les points qui s'écartent de la droite (valeurs aberrantes si points éloignés en la variable à expliquer, effet levier si points éloignés en la variable explicative).
Graphe 2 : résidus studentisés par validation croisée en fonction du pays, fonction de lissage, droites horizontales en les quantiles de niveaux 0.025 et 0.975 de la loi de Student à 111 degrés de liberté, c'est-à-dire ici -1.982 et 1.982 (cf tables). On peut y observer les valeurs aberrantes (résidus dont la valeur absolue dépasse 1.982), c'est-à-dire ici les pays 1, 52, 55 et dans une moindre mesure 113, y vérifier la non auto-corrélation des résidus.
Graphe 3 : résidus studentisés par validation croisée en fonction des valeurs ajustées, fonction de lissage, droites horizontales en les quantiles de niveaux 0.025 et 0.975 de la loi de Student à 111 degrés de liberté. On peut y observer les valeurs aberrantes, les mêmes bien sûr, y vérifier l'homoscédasticité des résidus (qui peut ceci dit être discutée dans le cas présent car le lisseur semble présenter une légère tendance particulière).
Graphe 4 : QQ plot des résidus = tracé des quantiles empiriques des résidus studentisés par VC en fonction des quantiles théoriques d'une loi gaussienne. Ici, vu que le nombre d'observations est grand (114), même si la loi des résidus studentisés par VC est une loi de Student, elle peut être approchée par une loi gaussienne. On valide donc ici la loi de Student avec un grand nombre de degrés de liberté.

Graphe 5 : résidus studentisés par VC en fonction des $\ln x_{i,4}$ et une fonction de lissage. Si la fonction de lissage a une tendance linéaire, cela peut indiquer l'oubli d'une variable explicative, ou au moins la pertinence de l'ajout de cette variable dans le modèle. Ici, effectivement, il peut être intéressant de rajouter la variable $\ln x_4$.
Graphe 6 : tracé des $h_{i,i}$, éléments diagonaux de la matrice chapeau $H = \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'$ et de la droite horizontale au niveau du seuil $4/114 = 0.035$. On y repère les pays à effet levier, ici 17, 23, 34, 69, 90, et dans une moindre mesure 37, 89, 114.
Graphe 7 : tracé des distances de Cook et du seuil égal au quantile de la loi de Fisher de niveau 0.1 à 2 et 112 degrés de liberté égal à 0.105. On peut y détecter les points influents. Ici, aucun point influent, donc aucune donnée n'est à étudier en particulier.

b) Pour les valeurs aberrantes, on peut voir sur le graphe 1 l'éloignement en la variable à expliquer.
Pour les données à effet levier, on peut voir l'éloignement en la variable explicative.

En conclusion, le modèle n'est pas à remettre en cause, sauf à ajouter peut-être la variable $\ln x_4$, et à examiner une éventuelle hétéroscédasticité.

Partie II : Régression linéaire multiple

1. La sortie complétée :

```
 Coefficients :
               Estimate  Std. Error  t value  Pr(>|t|)
 (Intercept)  -179.3450     20.1054   -8.920  1.08e-14  ***
 log(X3)        53.8656      5.3455   10.077   < 2e-16  ***
 log(X2)         1.3953      0.7654    1.823     0.072    .
-
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 5.315 on 111 degrees of freedom
Multiple R-Squared: 0.7409, Adjusted R-squared: 0.7362
F-statistic: 158.7 on 2 and 111 DF, p-value: < 2.2e-16
```

2. Le modèle est globalement significatif au niveau 5% (cf dernière ligne de la sortie : `p-value: < 2.2e-16`).

3. Variables significatives au niveau 5% : la constante et $\ln x_3$. Variables significatives au niveau 10% : la constante, $\ln x_3$ et $\ln x_2$.

4. La variable $\ln x_2$ n'est pas significative au niveau 5% alors qu'elle l'était dans le modèle ($\mathcal{M}_{\ln 2}$). Cela s'explique par une forte corrélation entre les variables $\ln x_2$ et $\ln x_3$. Par ailleurs, le modèle ($\mathcal{M}_{\ln 2}$) n'était pas très pertinent (R^2 environ égal à 0.5), mais l'ajout de la variable $\ln x_2$ au modèle ($\mathcal{M}_{\ln 3}$) permet néanmoins de passer d'un R^2_a de 0.7308 à 0.7362. Il peut donc être intéressant de la garder. Regarder les autres critères de sélection de variables est conseillé ici.

5. Une région de confiance simultanée pour $(\beta_1, \beta_2)'$ de niveau de confiance 95% est donnée par :

$$\left\{ \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix},\ \frac{1}{2\widehat{\sigma^2}} \begin{pmatrix} \hat{\beta}_1 - \beta_1 \\ \hat{\beta}_2 - \beta_2 \end{pmatrix}' \begin{pmatrix} 2.361 & 12.573 \\ 12.573 & 115.159 \end{pmatrix} \begin{pmatrix} \hat{\beta}_1 - \beta_1 \\ \hat{\beta}_2 - \beta_2 \end{pmatrix} \leq 3.078 \right\}.$$

Pour confirmer la réponse à la question 2, il s'agit de regarder si le vecteur $(0,0)'$ appartient à cette région de confiance, autrement dit de regarder si

$$\frac{1}{2\widehat{\sigma^2}}\begin{pmatrix}\hat{\beta}_1\\ \hat{\beta}_2\end{pmatrix}'\begin{pmatrix}2.361 & 12.573\\ 12.573 & 115.159\end{pmatrix}\begin{pmatrix}\hat{\beta}_1\\ \hat{\beta}_2\end{pmatrix} \leq 3.078.$$

On doit logiquement trouver, en le calculant, pour le premier terme, la valeur de la statistique du test de significativité globale du modèle c'est-à-dire ici 158.7 qui est très largement supérieur à 3.078. On rejette donc l'hypothèse que $(\beta_1,\beta_2)' = (0,0)'$ pour un niveau 5%.

6. Intervalles de confiance calculés à partir du quantile de niveau 0.975 de la loi de Student à 111 degrés de liberté et des valeurs données dans la sortie R uniquement (sans autre calcul ! ! !) :

— pour β_0, $\hat{I}_0 = \left[\hat{\beta}_0 - 1.982 \times 20.1054; \hat{\beta}_0 + 1.982 \times 20.1054\right]$, puis $\hat{I}_0^{obs} = [-219.194; -139.496]$,
— pour β_1, $\hat{I}_1 = \left[\hat{\beta}_1 - 1.982 \times 5.3455; \hat{\beta}_1 + 1.982 \times 5.3455\right]$, puis $\hat{I}_1^{obs} = [43.27; 64.46]$,
— pour β_2, $\hat{I}_2 = \left[\hat{\beta}_2 - 1.982 \times 0.7654; \hat{\beta}_1 + 1.982 \times 0.7654\right]$, puis $\hat{I}_2^{obs} = [-0.12; 2.91]$.

Partie III : Régression linéaire multiple et sélection de variables

1. a) Considérons un modèle $(\mathcal{M})$ et le modèle $(\mathcal{M}^-)$ contenant toutes les variables explicatives de $(\mathcal{M})$ sauf une variable, notée x. On a vu en cours que le R_a^2 du modèle $(\mathcal{M})$ est strictement supérieur à celui du modèle $(\mathcal{M}^-)$ si et seulement si la statistique du test de Fisher de significativité de la variable x est strictement supérieure à 1. De même, le C_p du modèle $(\mathcal{M})$ est strictement inférieur à celui du modèle $(\mathcal{M}^-)$ si et seulement si la statistique variante du test de Fisher de significativité de la variable x est strictement supérieure à 2. Enfin, le BIC du modèle $(\mathcal{M})$ est strictement inférieur à celui du modèle $(\mathcal{M}^-)$ si et seulement si la statistique du test de Fisher de significativité de la variable x est strictement supérieure à $(114-p)(e^{2\ln 114/114}-1)$, où p est le nombre de variables explicatives du modèle $(\mathcal{M})$.

b) La sortie SAS donne les valeurs des R_a^2, R^2, C_p et BIC pour les 10 meilleurs modèles du point de vue du R_a^2. Les modèles sont rangés par ordre décroissant des valeurs du R_a^2, la première colonne donne les nombres de variables explicatives des modèles.
La sortie R donne pour p allant de 2 à 9, les valeurs du R_a^2, du C_p ou du BIC du meilleur modèle parmi les modèles ayant un nombre p de variables explicatives. Chaque ligne correspond à un nombre p et au meilleur modèle à p variables explicatives. Les lignes sont rangées par ordre décroissant en fonction des valeurs du critère considéré, les intensités des couleurs décroissent également en fonction des valeurs de ce critère. Un bloc de couleur au niveau d'une certaine variable indique que cette variable est dans le modèle.

c) Les critères du C_p et du BIC, qui sont plus fiables que celui du R_a^2 (cf question a), indiquent que le modèle contenant la constante, x_1, $\ln x_3$, x_4, $\ln x_4$ et x_5 est préférable aux autres modèles.

2. a) Le critère utilisé par les différentes procédures algorithmiques de SAS est celui des tests de significativité des variables explicatives.

b) Dans la procédure backward, le critère d'arrêt est le suivant : on s'arrête lorsque toutes les variables explicatives du modèle sont significatives au niveau 10%. Dans la procédure forward, on s'arrête lorsqu'aucune variable significative au niveau 50% ne peut être ajoutée.

c) A chaque étape, on retire la variable dont la p-value pour le test de significativité est la plus élevée et supérieure à 10%.

d) A l'aide de la procédure backward, on retient le modèle contenant la constante, x_1, $\ln x_3$, x_4, $\ln x_4$ et x_5, qui est le même modèle que celui retenu par une procédure exhaustive à l'aide du C_p ou du BIC (cf question 1). A l'aide de la procédure forward, on retient le modèle contenant la constante, x_1, $\ln x_2$, x_3, $\ln x_3$, x_4, $\ln x_4$ et x_5, qui est le modèle retenu par une procédure exhaustive à l'aide du R_a^2 (moins fiable que les autres critères). On retiendra donc de préférence le modèle contenant la constante, x_1, $\ln x_3$, x_4, $\ln x_4$ et x_5.

3. On note $x_{115} = (1, 12.48, \ln 98, 34092.259, \ln 34092.259, 7.77)$ la nouvelle variable explicative pour le modèle retenu à la question précédente. On souhaite prédire une nouvelle observation d'une variable $Y_{115} = \beta_0 x_{115,0} + \ldots + \beta_5 x_{115,5} + \varepsilon_{115} = x_{115}\beta + \varepsilon_{115}$, avec $\varepsilon_{115} \sim \mathcal{N}(0, \sigma^2)$ et ε_{115} indépendante des ε_i, $i = 1 \ldots 114$, i.e. Y_{115} indépendante des Y_i, $i = 1 \ldots 114$, utilisées pour construire $\hat{\beta}$.
On introduit $\hat{Y}^p_{115} = x_{115}\hat{\beta} = \sum_{j=0}^{5} x_{115,j}\hat{\beta}_j$.
Un intervalle de prédiction pour Y_{115} de niveau de confiance $(1 - \alpha)$ est donné par

$$\widehat{I}^p_{115} = \left[\hat{Y}^p_{115} - t_{108}(1-\alpha/2)\sqrt{\widehat{\sigma^2}(1 + x_{115}(\mathbb{X}'\mathbb{X})^{-1}x'_{115})}; \hat{Y}^p_{115} + t_{108}(1-\alpha/2)\sqrt{\widehat{\sigma^2}(1 + x_{115}(\mathbb{X}'\mathbb{X})^{-1}x'_{115})}\right],$$

où $\mathbb{X}$ est la matrice du plan d'expérience du modèle retenu à la question précédente.

6.2.5 Sujet 2 bis (durée : 2h) - Entraînement

Etude de l'espérance de vie (suite)

L'Annexe 3.8, l'Annexe 3.11 (recto-verso), et l'Annexe 3.12 seront à rendre avec la copie.

L'espérance de vie est un indicateur influençant de nombreuses décisions politiques notamment sur les plans sanitaire et économique. Pouvoir comprendre et éventuellement prédire cette espérance de vie est donc un enjeu d'importance.

Nous cherchons ici à compléter l'étude, menée dans le sujet d'examen, sur l'espérance de vie de 114 pays au travers de données sur les consommations moyennes d'alcool et de tabac, le QI moyen, ainsi que le PIB par habitant et l'indice de démocratisation de ces pays.

On rappelle que l'on a introduit les variables suivantes : pour le $i$ème pays étudié ($i \in \{1, \dots, 114\}$),
- Y_i son espérance de vie,
- $x_{i,1}$ la consommation annuelle moyenne d'alcool de ses adultes de plus de 15 ans (en L d'alcool pur),
- $x_{i,2}$ le nombre annuel moyen de cigarettes consommées par ses habitants,
- $x_{i,3}$ le QI moyen de ses habitants,
- $x_{i,4}$ son PIB par habitant (en dollar international),
- $x_{i,5}$ son indice de démocratisation (note sur 10).

On considère le modèle de régression linéaire multiple suivant :

$$(\mathcal{M}_m) \quad Y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,3} + \beta_3 \ln(x_{i,4}) + \varepsilon_i \quad \text{pour } i \in \{1, \dots, 114\},$$

où les ε_i sont des termes d'erreur aléatoires supposés vérifier les conditions suivantes :
- $\mathbb{E}[\varepsilon_i] = 0$ pour tout i,
- $\text{cov}(\varepsilon_i, \varepsilon_j) = 0$ pour tout $i \neq j$,
- $\text{var}(\varepsilon_i) = \sigma^2$ pour tout i,
- $\varepsilon = (\varepsilon_1, \dots, \varepsilon_{114})'$ suit une loi gaussienne.

Ce modèle s'écrit sous la forme matricielle classique : $Y = \mathbb{X}\beta + \varepsilon$.

Partie I : Régression linéaire multiple

1. Estimation de β et σ^2.

a) Quelle hypothèse sur $\mathbb{X}$ doit-on faire pour pouvoir calculer l'estimateur des moindres carrés ordinaires $\hat{\beta}$ de β ? Expliquer.

b) Sous cette hypothèse, donner l'expression de $\hat{\beta}$.

c) L'estimateur $\hat{\beta}$ est-il un estimateur :
- sans biais ?
- de variance minimale parmi les estimateurs linéaires sans biais de β ?
- d'erreur quadratique moyenne minimale ?

d) Donner l'estimateur usuel $\widehat{\sigma^2}$ de la variance σ^2. Pourquoi choisit-on en général d'utiliser cet estimateur plutôt que l'estimateur du maximum de vraisemblance ?

e) Préciser les lois de $\hat{\beta}$ et $\widehat{\sigma^2}$.

2. Pour le modèle considéré, on a mis en œuvre sous le logiciel SAS une procédure de régression linéaire multiple. On a obtenu la sortie donnée en **Annexe 3.8 (à rendre)**.

Compléter la sortie obtenue sur l'**Annexe 3.8 (à rendre)**, en précisant la commande SAS utilisée pour obtenir cette sortie et les justifications nécessaires, et la rendre avec la copie (ne pas oublier d'y inscrire son nom !).

Partie II : Signe des cœfficients de régression et modèle sans constante

1. Signe des cœfficients.

a) Tester l'hypothèse $(H_0)\ \beta_1 \geq 0$ contre $(H_1)\ \beta_1 < 0$ au niveau 5% dans le modèle $(\mathcal{M}_m)$, dont la mise en œuvre sous SAS a produit la sortie complétée précédemment en Annexe 3.8.

b) Tester l'hypothèse $(H_0')\ \beta_1' \leq 0$ contre $(H_1')\ \beta_1' > 0$ au niveau 5% dans le modèle de régression linéaire simple

$$(\mathcal{M}_s)\ Y_i = \beta_0' + \beta_1' x_{i,1} + \varepsilon_i' \quad \text{pour } i \in \{1, \ldots, 114\},$$

dont la mise en œuvre sous SAS a produit la sortie donnée en Annexe 3.9.

c) Peut-on déduire des tests précédents une interprétation de l'influence de la consommation annuelle moyenne d'alcool par adulte sur l'espérance de vie des habitants d'un pays ? Si non, pourquoi ? Si oui, laquelle ? Expliquer très précisément.

2. Modèle sans constante.

La constante n'étant pas significative au niveau 5% dans le modèle de régression linéaire multiple $(\mathcal{M}_m)$, on a mis en œuvre une procédure de régression linéaire pour ce même modèle, mais sans la constante, et on a obtenu la sortie donnée en Annexe 3.10.
Que peut-on dire de cette sortie ? Donner la définition et une interprétation géométrique du cœfficient de détermination R^2 dans le modèle $(\mathcal{M}_m)$, puis dans ce nouveau modèle sans constante. Peut-on faire un choix entre $(\mathcal{M}_m)$ et le modèle sans constante sur la base du R^2 ? Du R^2 ajusté ? Expliquer.

Partie III : Détection des écarts au modèle

On a mis en œuvre une procédure de regréssion linéaire multiple pour le modèle $(\mathcal{M}_m)$ à l'aide du logiciel R, afin de pouvoir tracer des graphes complémentaires de son choix. Ces graphes sont reportés en **Annexe 3.11 (à rendre)**.

1. Tracer sur les quatre premiers graphes de l'**Annexe 3.11 (à rendre)** les droites horizontales représentant les seuils permettant de détecter les éventuelles valeurs aberrantes, les éventuelles données à effet levier et les éventuelles données influentes.

2. Compléter le questionnaire de l'**Annexe 3.11 (à rendre)** et rendre cette annexe avec la copie (ne pas oublier d'y inscrire son nom !).

Partie IV : Sélection de variables

Compléter le questionnaire de l'**Annexe 3.12 (à rendre)** et rendre cette annexe avec la copie (ne pas oublier d'y inscrire son nom !).

Extraits de tables statistiques

Table de la loi gaussienne : on donne pour différentes valeurs de $\alpha \in [0,1]$, q_α tel que $P(N \leq q_\alpha) = \alpha$ lorsque $N \sim \mathcal{N}(0,1)$.

α	0.9	0.95	0.975
q_α	1.282	1.645	1.96

Table de la loi de Student : on donne pour différentes valeurs de n et de $\alpha \in [0,1]$, $t_{n,\alpha}$ tel que $P(T \leq t_{n,\alpha}) = \alpha$ lorsque $T \sim \mathcal{T}(n)$ (on rappelle que la loi de Student est symétrique).

α	0.9	0.95	0.975
$t_{105,\alpha}$	1.290	1.659	1.983
$t_{106,\alpha}$	1.290	1.659	1.983
$t_{107,\alpha}$	1.290	1.659	1.982
$t_{108,\alpha}$	1.289	1.659	1.982
$t_{109,\alpha}$	1.289	1.659	1.982
$t_{110,\alpha}$	1.289	1.659	1.982
$t_{111,\alpha}$	1.289	1.659	1.982
$t_{112,\alpha}$	1.289	1.659	1.981
$t_{113,\alpha}$	1.289	1.658	1.981
$t_{114,\alpha}$	1.289	1.658	1.981

Table de la loi de Student : on donne pour différentes valeurs de q et pour différentes valeurs de n la valeur de $p_{n,q} = P(T \leq q)$ lorsque $T \sim \mathcal{T}(n)$.

q	2.9	2.91	2.92	2.93	2.94	2.95	2.96	2.97	2.98	2.99
$p_{110,q}$	0.99775	0.99781	0.99788	0.99794	0.99800	0.99806	0.99812	0.99817	0.99823	0.99828
$p_{111,q}$	0.99775	0.99782	0.99788	0.99794	0.99800	0.99806	0.99812	0.99817	0.99823	0.99828
$p_{112,q}$	0.99775	0.99782	0.99788	0.99795	0.99801	0.99807	0.99812	0.99818	0.99823	0.99828

Table de la loi de Fisher : on donne pour différentes valeurs de (n_1, n_2) et de $\alpha \in [0,1]$, $f_{n_1,n_2,\alpha}$ tel que $P(F \leq f_{n_1,n_2,\alpha}) = \alpha$ lorsque $F \sim \mathcal{F}(n_1, n_2)$.

α	0.025	0.05	0.1	0.5	0.95	0.975
$f_{2,110,\alpha}$	0.025	0.051	0.105	0.698	3.079	3.815
$f_{2,111,\alpha}$	0.025	0.051	0.105	0.697	3.078	3.814
$f_{2,112,\alpha}$	0.025	0.051	0.105	0.697	3.077	3.813
$f_{3,110,\alpha}$	0.072	0.117	0.194	0.794	2.687	3.237
$f_{3,111,\alpha}$	0.072	0.117	0.194	0.794	2.686	3.236
$f_{3,112,\alpha}$	0.072	0.117	0.194	0.793	2.686	3.235
$f_{4,110,\alpha}$	0.120	0.177	0.265	0.844	2.454	2.904
$f_{4,111,\alpha}$	0.120	0.177	0.265	0.844	2.453	2.903
$f_{4,112,\alpha}$	0.120	0.177	0.265	0.844	2.453	2.902

Table de la loi du Khi Deux : on donne pour différentes valeurs de n et de $\alpha \in [0,1]$, $k_{n,\alpha}$ tel que
$P(K \leq k_{n,\alpha}) = \alpha$ lorsque $K \sim \chi^2(n)$.

α	0.025	0.05	0.95	0.975
$k_{111,\alpha}$	83.735	87.681	136.591	142.049
$k_{112,\alpha}$	84.604	88.570	137.701	143.180
$k_{113,\alpha}$	85.473	89.461	138.811	144.311
$k_{114,\alpha}$	86.342	90.351	139.921	145.441

6.2.6 Sujet 3 (durée : 2h)

Exercice : Interprétations géométriques

On considère un modèle de régression linéaire avec constante : $Y = \mathbb{X}\beta + \varepsilon$, où :

- Y est un vecteur aléatoire à valeurs dans $\mathbb{R}^3$,
- $\mathbb{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \end{pmatrix}$,
- $\beta \in \mathbb{R}^2$,
- ε est un vecteur aléatoire vérifiant les conditions standards d'un modèle de régression linéaire.

On note $\mathbb{1} = (1,1,1)'$, $x = (x_1, x_2, x_3)'$, et $<,>$ le produit scalaire de $\mathbb{R}^3$.

1. Représenter graphiquement : le vecteur $\mathbb{1}$, le vecteur x, l'espace vectoriel $\mathcal{E}(\mathbb{X})$ engendré par $\mathbb{1}$ et x, Y, $\hat{Y}$, $\bar{Y}\mathbb{1}$, $\hat{\varepsilon}$. Représenter tous les angles droits visibles sur le graphe, ainsi que l'angle θ permettant de définir le critère du R^2. Quelle est l'interprétation géométrique de l'équation d'analyse de la variance ?

2. Donner une expression simplifiée de :
- $\langle Y - \hat{Y}, \hat{Y} \rangle$,
- $\langle Y, \mathbb{1} \rangle$,
- $\|Y - n^{-1}\langle Y, \mathbb{1}\rangle \mathbb{1}\|^2 - \|\Pi_{\mathbb{X}}(Y - \bar{Y}\mathbb{1})\|^2$,
- $\|Y\|^2 - \|\hat{\varepsilon}\|^2$,
- $\Pi_X(Y - \bar{Y}\mathbb{1}) - (Y - \bar{Y}\mathbb{1})$,
- $\Pi_X(Y - \mathbb{X}\beta) - (Y - \mathbb{X}\beta)$.

Problème : Modèles de Cobb-Douglas en Économie

On dispose de données correspondant aux valeurs du capital, du travail et de la valeur ajoutée sur les 15 dernières années pour divers secteurs économiques d'un pays européen.

Partie I : Régression linéaire multiple

Pour chaque secteur économique, on considère un modèle de type Cobb-Douglas :

$$V_i = A K_i^{\beta_1} L_i^{\beta_2} \eta_i, \tag{6.2}$$

où K_i, L_i, V_i désignent respectivement les valeurs du capital, du travail et de la valeur ajoutée pour la $i^{\text{ème}}$ année d'étude ($i \in \{1, \ldots, 15\}$), les η_i sont des termes d'erreur aléatoires tels que

- $\mathbb{E}[\ln \eta_i] = 0$ pour tout i,
- $\text{cov}(\ln \eta_i, \ln \eta_j) = 0$ pour tout $i \neq j$,
- $\text{var}(\ln \eta_i) = \sigma^2$ pour tout i.

La valeur A est souvent interprétée comme le niveau de technologie dans le secteur économique considéré.

Ce modèle peut s'écrire sous la forme d'un modèle de régression linéaire multiple :

$$Y = \mathbb{X}\beta + \varepsilon.$$

1. Définitions, hypothèses.
a) Préciser Y, $\mathbb{X}$, β, ε et les hypothèses faites sur ε.
b) Rappeler la définition de l'estimateur des moindres carrés ordinaires $\hat{\beta}$ de β. Sous une hypothèse nécessaire (à préciser) sur $\mathbb{X}$, donner l'expression de $\hat{\beta}$.
c) Déterminer un estimateur $\widehat{\sigma^2}$ sans biais de la variance résiduelle σ^2.

2. Pour le secteur économique de l'automobile, gravement touché par la crise économique actuelle, on a

$$(\mathbb{X}'\mathbb{X})^{-1} = \begin{pmatrix} 1496.5823 & -150.9911 & 44.04131 \\ -150.9911 & 45.58669 & -34.78859 \\ 44.04131 & -34.78859 & 31.63375 \end{pmatrix},$$

et si y désigne la valeur observée de Y, $\mathbb{X}'y = (139.269, 1949.137, 1949.658)'$, et $SCR_{obs} = 0.067416$.
Donner les valeurs de $\hat{\beta}$ et $\widehat{\sigma^2}$ calculées sur les observations.

3. On considère toujours le secteur de l'automobile, mais on suppose ici que $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$.
a) Quels sont les estimateurs du maximum de vraisemblance $\hat{\beta}_{MV}$ et $\widehat{\sigma^2}_{MV}$ de β et de σ^2 ? Ces estimateurs sont-ils optimaux ? Préciser.
b) On rappelle que l'on peut montrer à partir du théorème de Cochran les résultats suivants :
— Pour toute matrice réelle M de taille $q \times 3$ de rang q ($q \leq 3$), alors

$$M\hat{\beta} \sim \mathcal{N}(M\beta, \sigma^2 \left[M(\mathbb{X}'\mathbb{X})^{-1}M'\right]),$$

et

$$\frac{1}{\sigma^2}\left[M(\hat{\beta}-\beta)\right]' \left[M(\mathbb{X}'\mathbb{X})^{-1}M'\right]^{-1} \left[M(\hat{\beta}-\beta)\right] \sim \chi^2(q).$$

— $12\widehat{\sigma^2}/\sigma^2 \sim \chi^2(12)$.
— Les estimateurs $\hat{\beta}$ et $\widehat{\sigma^2}$ sont indépendants.
Construire de façon détaillée, à partir de ces résultats, un test de significativité globale du modèle au niveau 5%. Quelle est la conclusion du test ?
c) Construire une région de confiance simultanée pour (β_1, β_2) de niveau de confiance 95% et retrouver le résultat précédent.
d) La fonction de production donnée par (6.2) est plus facile à interpréter si elle est à rendements d'échelle constants, c'est-à-dire si $\beta_1 + \beta_2 = 1$. Construire un test de l'hypothèse (H_0) $\beta_1 + \beta_2 = 1$ contre l'alternative (H_1) $\beta_1 + \beta_2 \neq 1$ au niveau 5%. Quelle est la conclusion de ce test ?
e) Donner un encadrement de la p-valeur du test de la question précédente. Retrouver la conclusion précédente.

Partie II : Sélection de variables

Une généralisation du modèle de Cobb-Douglas défini par (6.2) est le modèle translog :

$$\ln V_i = \beta_0 + \beta_1 \ln K_i + \beta_2 \ln L_i + \beta_3 (\ln K_i)^2 + \beta_4 (\ln L_i)^2 + \beta_5 (\ln K_i)(\ln L_i) + \gamma_i.$$

On suppose que $\gamma = (\gamma_1, \ldots, \gamma_{15})' \sim \mathcal{N}(0, \sigma^2 I)$.

Toujours pour le secteur de l'automobile, la somme des carrés résiduelle pour ce nouveau modèle calculée sur les observations est égale à 0.057158. On rappelle que la somme des carrés résiduelle du modèle de Cobb-Douglas précédent calculée sur les observations est égale à 0.067416.

1. Sous quelles hypothèses sur les β_j retrouve-t-on un modèle de Cobb-Douglas ?

2. Tester l'hypothèse que le modèle de Cobb-Douglas est adéquat contre l'alternative que le modèle translog est nécessaire au niveau 5%.

3. On souhaite faire une procédure de sélection de variables de façon à simplifier au mieux le modèle.
a) Le R^2 est-il un bon critère de sélection de variables ? Expliquer.
b) On a réalisé toutes les régressions possibles et obtenu pour différents critères de sélection de variables les résultats présentés en Annexe 3.13. Quel sous-modèle choisirait-on pour chaque critère, par les méthodes exhaustive, ascendante (ou forward) et descendante (ou backward) ?
c) Tester, en utilisant les valeurs du R^2 ajusté par exemple, la validité du sous-modèle sélectionné par la méthode exhaustive par rapport au modèle translog complet au niveau 5%.
d) Le sous-modèle sélectionné peut-il contenir des variables explicatives non significatives au niveau 5% ? Expliquer.

Partie III : Détection d'éventuels écarts au modèle, données aberrantes, leviers, influentes

On fournit en Annexe 3.14 plusieurs résultats numériques ou graphiques obtenus après mise en œuvre de la régression pour le modèle de Cobb-Douglas défini par (6.2), sous hypothèse gaussienne.

1. Quels résultats peuvent éventuellement permettre de détecter des écarts au modèle ?
Que peut-on conclure ici ?

2. Quels résultats peuvent éventuellement permettre de détecter des données aberrantes ?
En détecte-t-on ici ?

3. Quels résultats peuvent éventuellement permettre de détecter des données ayant un effet levier ?
En détecte-t-on ici ?

4. Enfin, quels résultats peuvent éventuellement permettre de détecter des données influentes ?
En détecte-t-on ici ?

Partie IV

On considère maintenant trois secteurs économiques différents, considérés comme les modalités d'une variable qualitative ou facteur. On souhaite savoir si ce facteur a un effet sur le logarithme de la valeur ajoutée.

On considère les trois secteurs suivants : agriculture, industries agricoles et alimentaires, industries des biens de consommation. On dispose pour ces trois secteurs (notés de 1 à 3) de $n = 15$ observations chacun. On note $Y_{i,j} = \ln V_{i,j}$, où $V_{i,j}$ est la valeur ajoutée de la $j$ème année pour le secteur i ($i = 1, 2, 3$; $j = 1, \dots, 15$), et $y_{i,j}$ ($i = 1, 2, 3$; $j = 1, \dots, 15$) les observations correspondantes.

1. On suppose que les $Y_{i,j}$ vérifient :

$$Y_{i,j} = \beta_i + \varepsilon_{i,j}, \quad i = 1, 2, 3, \ j = 1, \dots, 15,$$

avec :
- $\mathbb{E}[\varepsilon_{i,j}] = 0$ pour tout (i, j),
- $\mathrm{cov}(\varepsilon_{i,j}, \varepsilon_{i',j'}) = 0$ pour tout $(i, j) \neq (i', j')$,
- $\mathrm{var}(\varepsilon_{i,j}) = \sigma^2$ pour tout (i, j).

En notant $Y = (Y_{1,1}, \dots, Y_{1,15}, Y_{2,1}, \dots, Y_{2,15}, Y_{3,1}, \dots, Y_{3,15})'$, $\varepsilon = (\varepsilon_{1,1}, \dots, \varepsilon_{1,15}, \varepsilon_{2,1}, \dots, \varepsilon_{2,15}, \varepsilon_{3,1}, \dots, \varepsilon_{3,15})'$, et $\beta = (\beta_1, \beta_2, \beta_3)'$, le modèle précédent peut s'écrire sous la forme classique d'un modèle de régression linéaire :

$$Y = \mathbb{X}\beta + \varepsilon.$$

1. Donner la forme de la matrice $\mathbb{X}$, et vérifier que les hypothèses standards, ainsi que l'hypothèse de rang sur la matrice du plan d'expérience sont satisfaites.

2. Donner les expressions de $\mathbb{X}'\mathbb{X}$ et de $(\mathbb{X}'\mathbb{X})^{-1}$.

3. En déduire des expressions simples de l'estimateur des moindres carrés ordinaires $\hat{\beta}$ de β, du vecteur des valeurs ajustées $\hat{Y}$, et de l'estimateur sans biais usuel $\widehat{\sigma^2}$ de la variance σ^2.

4. On a relevé les valeurs suivantes : $\sum_{j=1}^{15} y_{1j} = 156.7362$, $\sum_{j=1}^{15} y_{2j} = 151.7167$, $\sum_{j=1}^{15} y_{3j} = 155.6461$ et $\sum_{j=1}^{15} y_{1j}^2 = 1637.84$, $\sum_{j=1}^{15} y_{2j}^2 = 1534.59$, $\sum_{j=1}^{15} y_{3j}^2 = 1615.07$. Calculer les valeurs de $\hat{\beta}$ et $\widehat{\sigma^2}$ sur les observations.

5. Le vecteur $\mathbb{1} = (1, \dots, 1)$ de $\mathbb{R}^{45}$ appartient-il à l'espace engendré par les vecteurs colonnes de $\mathbb{X}$? Écrire l'équation d'analyse de la variance dans ce cas.

6. Sous hypothèse gaussienne, construire un test permettant de tester l'absence d'effet du facteur secteur économique, i.e. un test de $(H_0) : \beta_1 = \beta_2 = \beta_3$ au niveau 5%. Quelle est la conclusion de ce test ?

Extraits de tables statistiques

Table de la loi gaussienne : on donne pour différentes valeurs de $\alpha \in [0,1]$, q_α tel que $P(N \leq q_\alpha) = \alpha$ lorsque $N \sim \mathcal{N}(0,1)$.

α	0.9	0.95	0.975
q_α	1.282	1.645	1.96

Table de la loi de Student : on donne pour différentes valeurs de n et de $\alpha \in [0,1]$, $t_{n,\alpha}$ tel que $P(T \leq t_{n,\alpha}) = \alpha$ lorsque $T \sim \mathcal{T}(n)$ (on rappelle que la loi de Student est symétrique).

α	0.9	0.95	0.975
$t_{11,\alpha}$	1.36	1.80	2.20
$t_{12,\alpha}$	1.36	1.78	2.18
$t_{13,\alpha}$	1.35	1.77	2.16
$t_{14,\alpha}$	1.35	1.76	2.14
$t_{15,\alpha}$	1.34	1.75	2.13

Table de la loi de Student : on donne pour différentes valeurs de q et pour différentes valeurs de n la valeur de $p_{n,q} = P(T \leq q)$ lorsque $T \sim \mathcal{T}(n)$.

q	1	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2
$p_{11,q}$	0.831	0.853	0.872	0.890	0.905	0.919	0.931	0.941	0.950	0.958	0.965
$p_{12,q}$	0.831	0.854	0.873	0.891	0.907	0.920	0.932	0.943	0.951	0.959	0.966
$p_{13,q}$	0.832	0.854	0.874	0.892	0.908	0.921	0.933	0.944	0.952	0.960	0.967

Table de la loi de Fisher : on donne pour différentes valeurs de (n_1, n_2) et de $\alpha \in [0, 1]$, $f_{n_1,n_2,\alpha}$ tel que $P(F \leq f_{n_1,n_2,\alpha}) = \alpha$ lorsque $F \sim \mathcal{F}(n_1, n_2)$.

α	0.025	0.05	0.1	0.5	0.95	0.975
$f_{1,11,\alpha}$	0.001	0.004	0.017	0.486	4.844	6.724
$f_{1,12,\alpha}$	0.001	0.004	0.016	0.484	4.747	6.554
$f_{1,13,\alpha}$	0.001	0.004	0.016	0.481	4.667	6.414
$f_{1,14,\alpha}$	0.001	0.004	0.016	0.479	4.6	6.298
$f_{2,11,\alpha}$	0.025	0.052	0.106	0.739	3.982	5.256
$f_{2,12,\alpha}$	0.025	0.052	0.106	0.735	3.885	5.096
$f_{2,13,\alpha}$	0.025	0.051	0.106	0.731	3.806	4.965
$f_{2,14,\alpha}$	0.025	0.051	0.106	0.729	3.739	4.857
$f_{3,9,\alpha}$	0.069	0.113	0.191	0.852	3.863	5.078
$f_{3,10,\alpha}$	0.069	0.114	0.191	0.845	3.708	4.826
$f_{3,11,\alpha}$	0.07	0.114	0.191	0.840	3.587	4.63
$f_{4,9,\alpha}$	0.112	0.167	0.254	0.906	3.633	4.718
$f_{4,10,\alpha}$	0.113	0.168	0.255	0.899	3.478	4.468
$f_{4,11,\alpha}$	0.114	0.168	0.256	0.893	3.357	4.275
$f_{1,42,\alpha}$	0.001	0.004	0.016	0.463	4.073	5.404
$f_{1,43,\alpha}$	0.001	0.004	0.016	0.463	4.067	5.395
$f_{1,44,\alpha}$	0.001	0.004	0.016	0.463	4.062	5.386
$f_{2,42,\alpha}$	0.025	0.051	0.106	0.705	3.220	4.033
$f_{2,43,\alpha}$	0.025	0.051	0.106	0.704	3.214	4.024
$f_{2,44,\alpha}$	0.025	0.051	0.106	0.704	3.209	4.016

6.2.7 Sujet 3 : Éléments de correction

Exercice : interprétations géométriques

1. Pour cette question, il s'agit du cours...
2. On a :
   — $\langle Y - \hat{Y}, \hat{Y} \rangle = 0$,
   — $\langle Y, \mathbb{1} \rangle = n\bar{Y}$,
   — $\|Y - n^{-1}\langle Y, \mathbb{1}\rangle \mathbb{1}\|^2 - \|\Pi_{\mathbb{X}}(Y - \bar{Y}\mathbb{1})\|^2 = \|\hat{\varepsilon}\|^2$,
   — $\|Y\|^2 - \|\hat{\varepsilon}\|^2 = \|\hat{Y}\|^2$,
   — $\Pi_X(Y - \bar{Y}\mathbb{1}) - (Y - \bar{Y}\mathbb{1}) = -\hat{\varepsilon}$,
   — $\Pi_X(Y - \mathbb{X}\beta) - (Y - \mathbb{X}\beta) = -\hat{\varepsilon}$.

Modèles de Cobb-Douglas en Economie

Partie I

1. a) On a pour tout i, $\ln V_i = \ln A + \beta_1 \ln K_i + \beta_2 \ln L_i + \ln \eta_i$. On peut donc écrire $Y = \mathbb{X}\beta + \varepsilon$ avec

$$Y = \begin{pmatrix} \ln V_1 \\ \vdots \\ \ln V_{15} \end{pmatrix}, \quad \mathbb{X} = \begin{pmatrix} 1 & \ln K_1 & \ln L_1 \\ \vdots & \vdots & \vdots \\ 1 & \ln K_{15} & \ln L_{15} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \ln \eta_1 \\ \vdots \\ \ln \eta_{15} \end{pmatrix},$$

où $\beta_0 = \ln A$, ε est centré : $\mathbb{E}[\varepsilon] = 0$, et sa matrice de variance-covariance est $\text{var}(\varepsilon) = \sigma^2 I$.
b) L'estimateur des MCO de β est défini par $\hat{\beta} = \text{argmin}_{\beta \in \mathbb{R}^3} \|Y - \mathbb{X}\beta\|^2$. Sous l'hypothèse que $\mathbb{X}$ est de rang 3, $\hat{\beta} = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'Y$.
c) Un estimateur sans biais de la variance résiduelle est $\widehat{\sigma^2} = SCR/(15 - 3)$.

2. D'après la question précédente, $\hat{\beta}_{obs} = (-9.32697, 0.4718784, 0.8550131)'$ et $\widehat{\sigma^2}_{obs} = SCR/12 = 0.005618$.

3. a) Les estimateurs du maximum de vraisemblance de β et σ^2 sont donnés par $\hat{\beta}_{MV} = \hat{\beta}$ et $\widehat{\sigma^2}_{MV} = SCR/15$. On préfère $\widehat{\sigma^2}$ à $\widehat{\sigma^2}_{MV}$ car ce premier est sans biais (ce qui n'est pas forcément justifié puisqu'on n'évalue pas la variance de cet estimateur...). $\hat{\beta}_{MV}$ peut être dit optimal dans le sens où il est sans biais, de variance minimale parmi les estimateurs linéaires en Y sans biais de β (mais pas optimal au sens de la minimisation du risque quadratique !). On ne sait en revanche rien sur l'optimalité de $\widehat{\sigma^2}_{MV}$ (hormis sa consistance).
b) On souhaite tester $(H_0) : \beta_1 = \beta_2 = 0$ contre $(H_1) : \beta_1 \neq 0$ ou $\beta_2 \neq 0$.
En introduisant

$$M = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

on peut prendre comme statistique de test

$$F(Y) = \frac{1}{2\widehat{\sigma^2}} \left[M\hat{\beta}\right]' \left[M(\mathbb{X}'\mathbb{X})^{-1}M'\right]^{-1} \left[M\hat{\beta}\right],$$

qui suit sous l'hypothèse (H_0) la loi $\mathcal{F}(2,12)$.
Ici $F(y) = 26.27$, et le 0.95 quantile de la loi $\mathcal{F}(2,12)$ étant égal à 3.885, on rejette l'hypothèse (H_0) de non significativité globale du modèle.
c) Une région de confiance pour (β_1, β_2) est donnée par

$$\mathcal{R} = \left\{ (\beta_1, \beta_2),\ \frac{1}{2\widehat{\sigma^2}} \left[M\hat{\beta} - (\beta_1, \beta_2)' \right]' \left[M(\mathbb{X}'\mathbb{X})^{-1} M' \right]^{-1} \left[M\hat{\beta} - (\beta_1, \beta_2)' \right] \leq 3.885 \right\}.$$

On vérifie que $(0,0) \notin \mathcal{R}$.
d) Ici on prend $M = \begin{pmatrix} 0 & 1 & 1 \end{pmatrix}$ et $m = 1$, donc la statistique du test de Fisher est

$$F(Y) = \frac{1}{\widehat{\sigma^2}} (M\hat{\beta} - 1)' (M(\mathbb{X}'\mathbb{X})^{-1} M')^{-1} (M\hat{\beta} - 1)$$

qui suit sous (H_0) la loi $\mathcal{F}(1,12)$. Pour un niveau 5%, on rejette l'hypothèse (H_0) lorsque $F(y) \geq 4.747$. On a $F(y) = 2.4885$. On ne rejette donc pas (H_0) au niveau 5%.
e) La p-valeur du test est égale à $p(y) = P_{(H_0)}(F(Y) \geq 2.4885)$, or sous (H_0), $F(Y) \sim \mathcal{F}(1,12)$, donc $\sqrt{F(Y)}$ est de même loi que $|T|$, avec $T \sim \mathcal{T}(12)$. Par conséquent, $p(y) = 2(1 - P(T \leq \sqrt{2.4885}))$, et comme $0.92 \leq P(T \leq \sqrt{2.4885}) \leq 0.932$, $0.136 \leq p(y) \leq 0.16$. Puisque $p(y) > 0.05$, on ne rejette pas (H_0) au niveau 5%.

Partie II

1. On retrouve un modèle de Cobb-Douglas lorsque $\beta_3 = \beta_4 = \beta_5 = 0$.
2. On teste $(H_0) : \beta_3 = \beta_4 = \beta_5 = 0$ contre (H_1) : l'un des $(\beta_3, \beta_4, \beta_5)$ au moins est non nul au niveau 5%. La statistique de test s'écrit

$$F(Y) = \frac{(SCR - SCR_1)/(6-3)}{SCR_1/(15-6)},$$

où SCR_1 est la somme des carrés résiduelle dans le modèle translog, SCR étant la somme des carrés résiduelle dans le modèle de Cobb-Douglas. Sous (H_0), on sait que $F(Y) \sim \mathcal{F}(3,9)$, donc on rejette l'hypothèse (H_0) lorsque $F(y) \geq 3.863$. On a ici $F(y) = 3(0.067416 - 0.057158)/0.057158 = 0.5384$. On ne rejette donc pas l'hypothèse (H_0) au niveau 5%, c'est-à-dire qu'on ne rejette pas l'hypothèse que le modèle de Cobb-Douglas est adéquat.
3. a) Le R^2 est croissant en fonction du nombre de variables explicatives, donc choisir un modèle qui a le plus grand R^2 dans une collection de modèles revient à choisir systématiquement un modèle ayant le plus grand nombre de variables explicatives. Ce n'est donc pas un bon critère de sélection de variables. En revanche, le R^2 reste un bon critère pour choisir entre deux modèles ayant le même nombre de variables.
b) Il existe quatre types de procédures de sélection de variables : procédures exhaustive, backward, forward et stepwise. On peut utiliser pour chacune de ces procédures les critères du R^2 ajusté, d'AIC et de BIC, ainsi que le C_p de Mallows, mais les tests de validité de sous-modèles ne sont pas utilisables avec une procédure de sélection exhaustive.
Pour la procédure exhaustive, on choisit M5 avec les 3 critères. Pour la procédure ascendante ou forward, on choisit également M5. Enfin, pour la procédure descendante, on choisit M345 avec le R^2 ajusté, M5 avec AIC et BIC.

c) On teste (H_0) : M5 est valide contre (H_1) : il ne l'est pas au niveau 5%. La statistique de test s'écrit

$$F(Y) = \frac{(SCR_{M5} - SCR_1)/4}{SCR_1/9},$$

où SCR_1 est la somme des carrés résiduelle dans le modèle translog, SCR_{M5} est la somme des carrés résiduelle dans le modèle M5. Sous (H_0), on sait que $F(Y) \sim \mathcal{F}(4,9)$, donc on rejette l'hypothèse (H_0) lorsque $F(y) \geq 3.633$. On a ici pour le modèle M5,

$$R^2_{a,M5} = 1 - \frac{SCR_{M5}/13}{SCT/14},$$

et pour le modèle translog complet :

$$R^2_a = 1 - \frac{SCR_1/9}{SCT/14},$$

d'où

$$F(Y) = \frac{(13(1 - R^2_{a,M5}) - 9(1 - R^2_a))/4}{1 - R^2_a},$$

et $F(y) = 0.426$. On accepte donc l'hypothèse (H_0) au niveau 5%, c'est-à-dire qu'on accepte l'hypothèse que le modèle M5 est valide.
d) Ce modèle peut avoir été sélectionné à l'aide des critères ci-dessus sans que toutes ses variables explicatives soient significatives.

Partie III

1. Graphes des résidus et graphe quantiles - quantiles gaussiens. Une seule donnée aberrante ici, ce qui ne vient pas contredire le modèle. Le graphe quantiles - quantiles gaussions semble assez satisfaisant, mais on a le problème du peu de données : dans ce cas, la loi de Student des résidus n'est pas proche d'une loi gaussienne. En revanche, on observe une structuration du graphe des résidus en fonction des valeurs ajustées, mais on a aussi peu d'observations, donc il est difficile de voir précisément...
2. Graphe des résidus : une donnée aberrante, la 12ème.
3. La donnée des Hat Values peut permettre de détecter des données à effet levier : plusieurs dépassent 0.2 (seuil de Huber). Les données 3,4,7,11,15.
4. Le quantile de la loi de Fisher à 3 et 12 degrés de liberté de niveau 0.1 étant égal à 0.192, on peut considérer la donnée 12 comme influente (donnée aberrante).

Partie IV

1. Il s'agit ici d'un modèle d'ANOVA à un facteur avec une contrainte de type analyse par cellule (constante nulle).

La matrice $\mathbb{X}$ est de la forme :

$$\begin{pmatrix} 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 1 \end{pmatrix}.$$

Les hypothèses standards sont vérifiées par ε et la matrice $\mathbb{X}$ est bien de rang 3.
2. $\mathbb{X}'\mathbb{X} = 15I_3$ et $(\mathbb{X}'\mathbb{X})^{-1} = \frac{1}{15}I_3$.
3. $\hat{\beta} = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'Y = (\bar{Y}_{1\bullet}, \bar{Y}_{2\bullet}, \bar{Y}_{3\bullet})$, avec $\bar{Y}_{i\bullet} = \frac{1}{15}\sum_{j=1}^{15} Y_{i,j}$ pour tout i. Le vecteur des valeurs ajustées est défini par $\hat{Y} = (\bar{Y}_{1\bullet}, \ldots, \bar{Y}_{1\bullet}, \ldots, \bar{Y}_{3\bullet}, \ldots, \bar{Y}_{3\bullet})'$, celui des résidus par : $\hat{\varepsilon}_{i,j} = Y_{i,j} - \bar{Y}_{i\bullet}$, et $\widehat{\sigma^2} = \frac{1}{42}\sum_{i=1}^{3}\sum_{j=1}^{15}\hat{\varepsilon}_{i,j}^2 = \frac{1}{42}\sum_{i=1}^{3}\sum_{j=1}^{15}(Y_{i,j} - \bar{Y}_{i\bullet})^2$.
4. $\hat{\beta}_{obs} = (10.44933, 10.11467, 10.37667)'$, $\widehat{\sigma^2}_{obs} = 0.1732/42 = 0.004124$.
5. Le vecteur $\mathbb{1}$ appartient bien à l'espace engendré par les colonnes de $\mathbb{X}$, donc l'équation d'analyse de la variance s'écrit :

$$\|Y - \bar{Y}\mathbb{1}\|^2 = \|\hat{Y} - \bar{Y}\mathbb{1}\|^2 + \|Y - \hat{Y}\|^2.$$

$SCT_{obs} = \|y - \bar{y}\mathbb{1}\|^2 = 1.102627$. Statistique de test : $F(Y) = \frac{SCE/2}{SCR/42} = 21\left(\frac{SCT}{SCE} - 1\right)$, avec $F(y) = 112.685$. Sous (H_0), $F(Y) \sim \mathcal{F}(2, 42)$, et puisque $f_{2,42,0.95} = 3.22$, on rejette l'hypothèse (H_0) pour un niveau 5%.

Bibliographie

[1] Azaïs, J.-M., Bardet, J.-M. (2006). *Le modèle linéaire par l'exemple. Régression analyse de la variance et plans d'expérience*, Dunod, Paris.

[2] Cornillon, P.-A., Matzner-Løber, E. *Régression avec R*, Springer.

[3] Dodge, Y., Rousson, V. *Analyse de régression appliquée*, Dunod.

[4] Kleinbaum, D. et al. *Applied regression analysis and other multivariate methods*.

[5] Tomassone, R. *La régression*, Masson.