INSTITUT Mines-Télécom

Statistique pour ingénieur

Thème 4 : Régression linéaire

A. Badea, S. Mussard, F. Seyte & T. Verdel, 24 mars 2016

Table des matières

1 Introduction générale

Un modèle est une représentation simplifiée, mais la plus exhaustive possible, d'une entité donné, de nature biologique, industrielle, économique, médicale, etc. Sous sa forme la plus courante, il est présenté comme un système d'équations, le plus souvent linéaires, équations reliant entre elles deux types de variables que l'on appelle :

- variables expliquées (ou endogènes) ;
- variables explicatives (ou exogènes).

Un modèle s'écrit différemment selon la manière dont sont observées les variables du modèle :

- lorsque les observations s'effectuent au cours du temps, les variables sont des séries temporelles et le modèle porte le nom de modèle en séries temporelles ;
- lorsque les observations sont réalisées sur des échantillons d'individus, à un instant donné, le modèle porte le nom de modèle en coupe instantanée ;
- lorsque les observations portent sur des échantillons au cours du temps, on parle de modèle de panels.

Dans ce qui suit, nous considérons que la variable explicative X possède une forme déterministe, ainsi $X = x$. Le Modèle Linéaire Général Simple (MLGS) à plusieurs variables explicatives s'écrit :

$$Y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \varepsilon_i, \ \forall i \in \{1, \ldots, n\},$$

avec n le nombre d'observations des variables, Y_i la *variable expliquée*, $\beta_0, \ldots, \beta_k$ les paramètres inconnus du modèle, $x_1, \ldots, x_k$ les k *variables explicatives* (en réalité $k+1$, puisqu'on considère que $x_{0i} = 1$, ce qui permet de tenir compte d'une constante β_0) et ε_i *l'aléa* (ou terme d'*erreur*). En faisant varier i dans cette écriture, on obtient le MLGS sous sa forme matricielle :

$$\underset{(n,1)}{Y} = \underset{(n,k+1)}{\mathbf{X}} \underset{(k+1,1)}{\mathbf{B}} + \underset{(n,1)}{\varepsilon},$$

avec $\mathbf{X}$ la matrice comportant en colonne les $k+1$ variables explicatives x_k.

Dans ce cours, nous étudierons un cas particulier du MLGS, celui pour lequel une variable expliquée Y est reliée linéairement à une seule variable explicative x et à un aléa ε. On l'appelle le modèle de régression linéaire simple. Il s'écrit :

$Y_t = \beta_0 + \beta_1 x_t + \varepsilon_t, \ \forall t \in \{1, \ldots, T\}$ **pour le modèle en séries temporelles** ;
$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \ \forall i \in \{1, \ldots, n\}$ **pour le modèle en coupe instantanée** ;
$Y_{it} = \beta_0 + \beta_1 x_{it} + \varepsilon_{it}, \ \forall i \in \{1, \ldots, n\}, \ \forall t \in \{1, \ldots, T\}$ **pour le modèle de panels**.

On retiendra par la suite la seule écriture en coupe instantanée et de ce fait i représente l'individu i. Ce modèle se différencie de celui rencontré en statistique descriptive qui s'écrit sous la forme exacte : $Y = \beta_0 + \beta_1 x$.

Il est rare qu'en statistique ce type de relation exacte existe. Dans la majorité des cas, x ne fournit qu'une partie de l'explication de la variable Y. On contourne cette difficulté en introduisant, dans le second membre de la relation, une nouvelle variable appelée aléa ou erreur. Cet élément aléatoire va permettre de synthétiser l'ensemble des influences sur Y que x ne peut expliquer. On suppose qu'il rassemble un nombre important mais indépendant de fluctuations, sans qu'aucune n'ait à elle seule une importance par rapport

aux autres de telle sorte que cet élément puisse être assimilé à une variable aléatoire obéissant à une loi de probabilité définie sur un domaine.

Le modèle de régression linéaire simple rassemble plusieurs formes non linéaires que l'on transforme linéairement par *anamorphose*.

- Le modèle semi-logarithmique :

$$Y_i = \beta_0 + \beta_1 \log x_i,$$

s'étudie sur les couples $(z_i = \log x_i, Y_i)$, avec $x_i > 0$ pour tout $i \in \{1, \ldots, n\}$.

- Le modèle doublement logarithmique :

$$\log Y_i = \log \beta_0 + \beta_1 \log x_i \iff z_i = \beta_0' + \beta_1 v_i,$$

s'étudie sur les couples $(v_i = \log x_i, z_i = \log Y_i)$, avec $x_i > 0$ et $Y_i > 0$. Ce modèle a pour paramètre de pente β_1, le coefficient d'élasticité instantanée entre Y_i et x_i qui mesure la réponse, en pourcentage, de la variable Y_i suite à une modification de 1% de la variable explicative x_i.

- Le modèle logistique :

$$Y_i = \frac{K}{1 + \exp(-ax_i + b)}$$

s'écrit,

$$\ln\left(\frac{K}{Y_i} - 1\right) = Ax_i + b,$$

avec A$= -a$. Il s'étudie avec les couples :

$$\left(v_i = \ln\left(\frac{K}{Y_i} - 1\right), x_i\right).$$

Ce modèle est souvent utilisé pour modéliser la pénétration des produits nouveaux sur un marché ou encore pour calculer la part de marché K d'un produit. Il existe par ailleurs d'autres formes de modèles non linéaires transformables linéairement par anamorphose.

Le problème que nous devons résoudre dans le cadre de ce cours est celui du calcul des paramètres inconnus β_0 et β_1 à partir des couples (Y_i, x_i) : il s'agit de l'*estimation* du modèle. L'analyse et la pertinence du choix de ce modèle seront aussi analysées.

Si on appelle $\hat{\beta}_0$ et $\hat{\beta}_1$ les valeurs calculées du modèle à partir des Y_i et x_i pour tout $i \in \{1, \ldots, n\}$, on peut alors obtenir une série de valeurs notées $\hat{Y}_i$ calculées à partir de la relation :

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i.$$

Il existe une différence entre Y_i et $\hat{Y}_i$; cet écart noté $\hat{\varepsilon}_i$ est appelée *résidu* de la valeur Y_i. Il s'écrit :

$$\hat{\varepsilon}_i = Y_i - \hat{Y}_i \iff Y_i = \hat{\varepsilon}_i + \hat{Y}_i,$$

ou bien à partir des réalisations y_i de la variable aléatoire Y_i,

$$\hat{\varepsilon}_i = y_i - \hat{y}_i \iff y_i = \hat{\varepsilon}_i + \hat{y}_i.$$

Or,

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \implies Y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\varepsilon}_i.$$

Il existe donc deux écritures du modèle :

- *le modèle théorique* :

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \ ;$$

- *le modèle empirique* (ou calculé) :

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\varepsilon}_i.$$

Remarque 1

Ne pas confondre $\hat{\varepsilon}_i$ et ε_i. Le résidu $\hat{\varepsilon}_i$ est connu, alors que l'erreur ε_i est inconnue. Les informations dont on dispose concernant l'aléa ε_i sont $\hat{\varepsilon}_i$ et le fait que ε_i suive une loi normale.

2 Les estimateurs des Moindres Carrés Ordinaires (MCO)

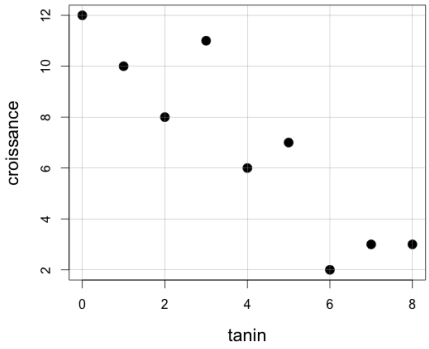
2.1 Exemple numérique

Nous allons illustrer les propos qui suivront par un exemple simple que voici [1].

Exemple 1

Le jeu de données représente la croissance des chenilles et la quantité de tanin contenue dans leurs aliments.

x	*tanin*	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
y	*croissance*	*12*	*10*	*8*	*11*	*6*	*7*	*2*	*3*	*3*



2.2 Les hypothèses de base du modèle

Le problème statistique à résoudre est le calcul des paramètres β_0 et β_1, sachant les observations de Y et de x.

Il existe, entre autres, deux méthodes bien connues permettant le calcul de β_0 et β_1 : la méthode des moindres carrés ordinaires (MCO) et la méthode du maximum de vraisemblance. L'utilisation de ces deux méthodes conduit à des valeurs calculées de β_0 et β_1 qui possèdent des propriétés statistiques remarquables. Cela implique qu'un certain nombre d'hypothèses de base soient vérifiées avant l'utilisation de ces méthodes.

- Hypothèses :

1. A partir du livre de Michael Crawley, *Statistics : An introduction using R*, Wiley (2005).

$\hookrightarrow$ x est une variable indépendante de l'erreur (aléa) : $\mathbb{C}\text{ov}\,(x_i, \varepsilon_i) = 0$.

$\hookrightarrow$ Les erreurs ε_i sont des variables aléatoires indépendantes et identiquement distribuées (i.i.d.).

$\hookrightarrow$ De plus, on suppose ici que la distribution des erreurs est normale : $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$.

Par conséquent, $\mathbb{E}\,(\varepsilon_i) = 0$ et $\mathbb{V}(\varepsilon_i) = \mathbb{E}\left[(\varepsilon_i - \mathbb{E}\,(\varepsilon_i))^2\right] = \mathbb{E}\,[\varepsilon_i^2] = \sigma^2$, $\forall i \in \{1, \ldots, n\}$ (il s'agit de l'hypothèse d'homoscédasticité : cela implique que la variance des ε_i est constante quel que soit le sous-échantillon tiré dans l'ensemble $\{1, \ldots, n\}$).
$\mathbb{C}\text{ov}\,(\varepsilon_i, \varepsilon_{i'}) = \mathbb{E}\,[\varepsilon_i - \mathbb{E}\,(\varepsilon_i)]\,\mathbb{E}\,[\varepsilon_{i'} - \mathbb{E}\,(\varepsilon_{i'})] = \mathbb{E}[\varepsilon_i \varepsilon_{i'}] = 0$, $\forall i, i' \in \{1, \ldots, n\}$ et $i \neq i'$ (hypothèse de non auto-corrélation des erreurs).
On peut donc écrire :

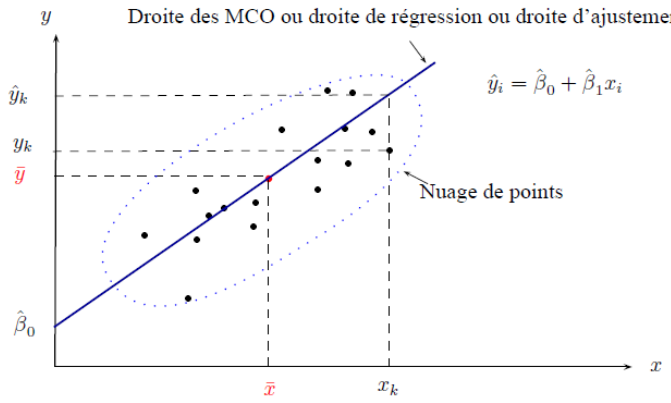
$$\mathbb{E}[\varepsilon_i \varepsilon_{i'}] = \begin{cases} \sigma^2 & \text{si } i = i' \\ 0 & \text{si } i \neq i'. \end{cases}$$

Trois quantités sont donc inconnues dans ce modèle : β_0, β_1 et σ^2. L'objectif des méthodes d'estimation est de trouver des estimateurs de ces paramètres inconnus qui possèdent les propriétés requises par la théorie statistique.

2.3 Les estimateurs des moindres carrés ordinaires

La méthode des MCO consiste à minimiser la somme des carrés des écarts, écarts entre la valeur observée de la variable expliquée (pour un point du nuage) et sa valeur calculée par le modèle. Graphiquement (voir la figure ci-dessous), il s'agit de la distance mesurée parallèlement à l'axe des ordonnées entre ces deux points. Rappelons que l'écart entre une valeur observée et une valeur calculée est appelé le *résidu*, noté $\hat{\varepsilon}_i$:

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



Formellement, la méthode des MCO est la suivante :

$$\min \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \min \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \min \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2 \equiv \min \phi(\hat{\beta}_0, \hat{\beta}_1).$$

Il s'agit de la minimisation d'une fonction ϕ à deux inconnues $\hat{\beta}_0$ et $\hat{\beta}_1$. La solution, si elle existe, est donnée par le système d'équations normales suivant :

$$\begin{cases} \frac{\partial \phi}{\partial \hat{\beta}_0} & = 0 \\ \frac{\partial \phi}{\partial \hat{\beta}_1} & = 0. \end{cases}$$

On obtient :

$$\begin{aligned} & \frac{\partial \phi}{\partial \hat{\beta}_0} = 0 \\ \iff & -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) = 0 \\ \iff & n\bar{y} - n\hat{\beta}_0 - n\hat{\beta}_1 \bar{x} = 0 \\ \iff & \hat{\beta}_0 = \bar{y} - \bar{x}\hat{\beta}_1. \end{aligned}$$

La droite passe par le *centre de gravité* (le point moyen $G(\bar{x}, \bar{y})$) du nuage de régression. Aussi :

$$\begin{aligned} & \frac{\partial \phi}{\partial \hat{\beta}_1} = 0 \\ \iff & -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) x_i = 0 \\ \iff & \sum_{i=1}^{n} \left(y_i - (\bar{y} - \bar{x}\hat{\beta}_1) - \hat{\beta}_1 x_i \right) x_i = 0 \\ \iff & \sum_{i=1}^{n} \left(y_i x_i - \bar{y} x_i + \hat{\beta}_1 \bar{x} x_i - \hat{\beta}_1 x_i^2 \right) = 0 \\ \iff & \sum_{i=1}^{n} y_i x_i - n\bar{y}\bar{x} + n\hat{\beta}_1 \bar{x}^2 - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0 \\ \iff & \sum_{i=1}^{n} y_i x_i - n\bar{y}\bar{x} - \hat{\beta}_1 \left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \right) = 0 \\ \iff & \hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - n\bar{y}\bar{x}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{\frac{1}{n}\sum_{i=1}^{n} y_i x_i - \bar{y}\bar{x}}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2} = \frac{\mathbb{C}\mathrm{ov}\,(x, y)}{s_x^2}, \end{aligned}$$

avec s_x^2 la variance empirique de x. Il faut vérifier que $\hat{\beta}_0$ et $\hat{\beta}_1$ minimisent la somme des $\hat{\varepsilon}_i^2$. Pour cela, on calcule les dérivées secondes :

$$\begin{cases} \frac{\partial^2 \phi}{\partial \hat{\beta}_0^2} & = 2 > 0 \\ \frac{\partial^2 \phi}{\partial \hat{\beta}_1^2} & = 2\sum_{i=1}^{n} x_i^2 > 0. \end{cases}$$

Les dérivées partielles secondes sont positives, ce qui indique que la solution à l'aide des dérivées partielles premières $(\hat{\beta}_0, \hat{\beta}_1)$ est bien un minimum. Les résultats finaux sont donc :

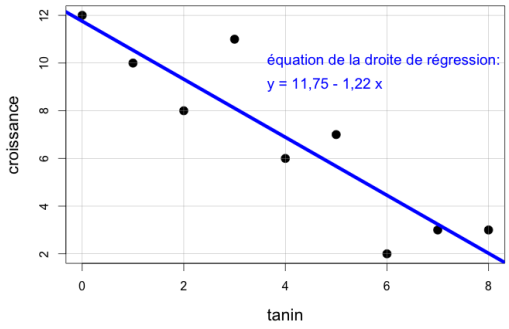
$$\boxed{\hat{\beta}_0 = \bar{y} - \bar{x}\hat{\beta}_1 \text{ et } \hat{\beta}_1 = \frac{\mathbb{C}\mathrm{ov}(x, y)}{s_x^2}.}$$

Exemple 2

En appliquant les formules précédentes à notre jeu de données, on obtient :

$$\begin{cases} \bar{x} = 4 & , \quad \bar{y} \approx 6,89 \\ s_x^2 \approx 6,67 & , \quad \mathbb{C}\text{ov}(x,y) \approx -8,11 \\ \hat{\beta}_1 = \frac{\mathbb{C}\text{ov}(x,y)}{s_x^2} \approx -1,22 & , \quad \hat{\beta}_0 = \bar{y} - \bar{x}\hat{\beta}_1 \approx 11,75 \end{cases}$$

x	0	1	2	3	4	5	6	7	8
y	12	10	8	11	6	7	2	3	3
$\hat{y}$	11,76	10,54	9,32	8,11	6,89	5,67	4,46	3,24	2,02



Il est aussi possible d'exprimer la relation entre y et x à l'aide des *données centrées* :

$$\begin{cases} x_i^c = x_i - \bar{x} & , \quad \hat{y}_i^c = \hat{y}_i - \bar{y} \\ y_i^c = y_i - \bar{y} & , \quad \hat{\varepsilon}_i = y_i^c - \hat{y}^c. \end{cases}$$

On peut calculer $\hat{\beta}_0$ et $\hat{\beta}_1$ en fonction des données centrées par une démonstration analogue à la précédente. La méthode des MCO est :

$$\min \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \min \sum_{i=1}^{n} \left(y_i^c - \hat{y}_i^c\right).$$

Puisque $\bar{x}^c = \bar{y}^c = 0$, alors :

$$\begin{cases} \hat{\beta}_0^c = \bar{y}^c - \hat{\beta}_1 \bar{x}^c = 0 \\ \hat{\beta}_1^c = \frac{\sum_{i=1}^n x_i^c y_i^c}{\sum_{i=1}^n (x_i^c)^2} = \frac{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\mathbb{C}\text{ov}(x,y)}{s_x^2} = \hat{\beta}_1. \end{cases}$$

Comme $\hat{\beta}_0^c = 0$, le modèle centré s'écrit :

$$\hat{y}_i^c = \hat{\beta}_1 x_i^c.$$

Ce changement de variable consiste à changer d'axe dans le nuage de régression et à placer l'origine au centre de gravité $G(\bar{x}, \bar{y})$.

2.4 Les propriétés des estimateurs des moindres carrés ordinaires

- Propriétés :

↪ Les estimateurs des MCO sont des fonctions linéaires des x_i, y_i.

↪ $\hat{\beta}_0$ et $\hat{\beta}_1$ sont des *estimateurs* (linéaires) sans biais de β_0 et β_1 :

$$\mathbb{E}[\hat{\beta}_0] = \beta_0,\ \mathbb{E}[\hat{\beta}_1] = \beta_1.$$

↪ $\hat{\beta}_0$ et $\hat{\beta}_1$ sont des estimateurs à *variance minimale* de β_0 et β_1 :

$$\mathbb{V}(\hat{\beta}_0) = \sigma^2 \frac{\sum_{i=1}^n x_i^2}{n^2 s_x^2} = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{n s_x^2} \right),\ \mathbb{V}(\hat{\beta}_1) = \frac{\sigma^2}{n s_x^2}.$$

Par ailleurs le lien entre les deux estimateurs est donné par :

$$\mathbb{C}\text{ov}\left(\hat{\beta}_0, \hat{\beta}_1\right) = -\frac{\sigma^2 \bar{x}}{n s_x^2}.$$

Les estimateurs des moindres carrés $\hat{\beta}_0$ et $\hat{\beta}_1$ sont parmi tous les estimateurs linéaires sans biais les meilleurs au sens de la variance minimale (Théorème de Gauss-Markov). Ils sont dits *estimateurs BLUE* (Best Linear Unbiased Estimator).

↪ $\hat{\beta}_0$ et $\hat{\beta}_1$ sont des estimateurs *convergents*. Les estimateurs sont convergents si :

$$\lim_{n\to\infty} \mathbb{V}(\hat{\beta}_0) = 0 \text{ et } \lim_{n\to\infty} \mathbb{V}(\hat{\beta}_1) = 0.$$

Pour $\hat{\beta}_0$ le résultat est immédiat puisque :

$$\mathbb{V}(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{n s_x^2} \right).$$

Pour $\hat{\beta}_1$, il suffit de récrire sa variance de la manière suivante,

$$\mathbb{V}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (x_i^c)^2} = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sigma^2}{\sum_{i=1}^n n \cdot \frac{1}{n} (x_i - \bar{x})^2} = \frac{\sigma^2}{n s_x^2}.$$

On en déduit alors que $\lim_{n\to\infty} \mathbb{V}(\hat{\beta}_1) = 0$.

↪ Estimateur de la variance de l'erreur σ^2 :

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

On obtient un estimateur $\hat{\sigma}^{*2}$ sans biais de σ^2 :

$$\mathbb{E}\left(\hat{\sigma}^{*2}\right) = \sigma^2.$$

Exemple 3

Application au jeu de données : $\hat{\sigma}^{*2} \approx 2,87$.

x	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
y	*12*	*10*	*8*	*11*	*6*	*7*	*2*	*3*	*3*
$\hat{y}$	*11,76*	*10,54*	*9,32*	*8,11*	*6,89*	*5,67*	*4,46*	*3,24*	*2,02*
$\hat{\varepsilon}$	*0,24*	*-0,54*	*-1,32*	*2,89*	*-0,89*	*1,33*	*-2,46*	*-0,24*	*0,98*

$\hookrightarrow$ Les estimateurs des moindres carrés $\hat{\beta}_0$ et $\hat{\beta}_1$ de β_0 et β_1 correspondent aux estimateurs du maximum de vraisemblance.

Pour obtenir les résultats précédents, il est possible d'utiliser les estimateurs issus de la méthode du *maximum de vraisemblance*.

Comme les différents ε_i ont des distributions identiques (supposées normales, voir les hypothèses) ont peut conclure que les couples (x_i, y_i) correspondent à des valeurs de ε_i issues d'une même distribution $f(\varepsilon_i)$ et on a :

$$\mathbb{P}\left(\varepsilon_i < \varepsilon < \varepsilon_i + \mathrm{d}\varepsilon\right) = f(\varepsilon_i)\mathrm{d}\varepsilon = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}\right).$$

La forme de la densité de probabilité étant connue, la fonction de vraisemblance de l'échantillon s'écrit :

$$L[(x_1, y_1), \ldots, (x_i, y_i); \beta_0, \beta_1] = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}\right).$$

De manière équivalente :

$$L[(x_1, y_1), \ldots, (x_i, y_i); \beta_0, \beta_1] = \frac{1}{\left(\sigma\sqrt{2\pi}\right)^n} \exp\left(-\frac{\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}\right).$$

La fonction log-vraisemblance, plus facile à minimiser, s'écrit :

$$\ln L[(x_1, y_1), \ldots, (x_i, y_i); \beta_0, \beta_1] = -n \ln \sigma - \frac{n}{2} \ln 2\pi - \frac{\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}.$$

En maximisant $\ln L[\cdot]$ par rapport à β_0, β_1 et σ^2 on trouve les estimateurs $\tilde{\beta}_0$, $\tilde{\beta}_1$ et $\tilde{\sigma}^2$ qui correspondent au maximum de vraisemblance :

$$\begin{cases} \frac{\partial \ln L[\cdot]}{\partial \tilde{\beta}_0} = \frac{1}{\sigma^2} \sum_{i=1}^{n}(y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i) = 0 \\ \frac{\partial \ln L[\cdot]}{\partial \tilde{\beta}_1} = \frac{1}{\sigma^2} x_i \sum_{i=1}^{n}(y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i) = 0 \\ \frac{\partial \ln L[\cdot]}{\partial \tilde{\sigma}^2} = \frac{1}{2\tilde{\sigma}^2} \sum_{i=1}^{n}(y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i) - \frac{2n}{2\tilde{\sigma}^2} = 0. \end{cases}$$

Les deux premières équations permettent de vérifier que les estimateurs sont les mêmes que ceux des MCO :

$$\tilde{\beta}_0 = \hat{\beta}_0 \ ; \ \tilde{\beta}_1 = \hat{\beta}_1.$$

La dernière équation conduit à :

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

Or, d'après le résultat des MCO :

$$\mathbb{E}\left[\frac{\sum_{i=1}^{n} \hat{\varepsilon}_i^2}{n-2}\right] = \sigma^2 \iff \mathbb{E}\left[\sum_{i=1}^{n} \hat{\varepsilon}_i^2\right] = \sigma^2(n-2).$$

D'où :

$$\mathbb{E}\left[\tilde{\sigma}^2\right] = \mathbb{E}\left[\frac{\sum_{i=1}^{n} \hat{\varepsilon}_i^2}{n}\right] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\hat{\varepsilon}_i^2\right] = \frac{1}{n}\sigma^2(n-2).$$

L'estimateur $\tilde{\sigma}^2$ de σ^2 issu du maximum de vraisemblance est donc biaisé. On retiendra par la suite celui des MCO.

3 Lois des estimateurs et tests des estimateurs

Les estimateurs sont des estimateurs linéaires des Y_i (qui dépendent de ε_i qui sont aléatoires et obéissent à des lois normales) alors les estimateurs sont aussi aléatoires et obéissent à des lois normales :

$$\hat{\beta}_0 \sim \mathcal{N}\left(\beta_0 \ , \ \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{ns_x^2}\right)\right)$$

$$\hat{\beta}_1 \sim \mathcal{N}\left(\beta_1 \ , \ \frac{\sigma^2}{ns_x^2}\right).$$

Ces lois contiennent l'écart type σ de l'erreur. Or, puisque ce paramètre est inconnu les deux lois précédentes ne sont pas empiriquement utilisables. On a donc besoin d'une loi qui contienne à la fois l'estimateur de la variance de l'aléa et la variance de l'aléa pour pouvoir estimer par intervalle de confiance β_0 et β_1. On utilise alors le *théorème de Fisher* :

$$Z \equiv (n-2)\frac{\hat{\sigma}^{*2}}{\sigma^2} \sim \chi^2_{n-2}.$$

A partir de ce résultat nous pourrons construire, avec la loi de Student, des intervalles de confiance de β_0, β_1 et σ^2 avant de réaliser des tests d'hypothèses sur ces paramètres.

3.1 Estimation par intervalle de confiance de β_0, β_1 et σ^2

3.1.1 Intervalle de confiance de β_1

Le problème est le suivant : on cherche les réels a et b tels que, pour un risque d'erreur donné α :

$$1 - \alpha = \mathbb{P}\left(a \leq \beta_1 \leq b\right).$$

A partir de cette expression nous pouvons déduire un *intervalle de confiance* de β_1. On sait que :

$$\hat{\beta}_1 \sim \mathcal{N}\left(\beta_1 \ , \ \frac{\sigma^2}{ns_x^2}\right).$$

Comme σ^2 est inconnu, il faut utiliser une *loi de Student*. Soit deux variables aléatoires $U \sim \mathcal{N}(0,1)$ et $Z \sim \chi^2_d$ indépendantes, alors la variable aléatoire T suivant une loi de Student à d degrés de liberté est :

$$T \equiv \frac{U}{\sqrt{\frac{Z}{d}}} \sim \mathcal{T}(n-2).$$

En utilisant la variable aléatoire normale $\hat{\beta}_1$, il vient :

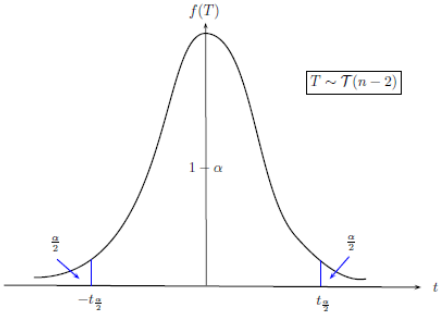
$$T_{\hat{\beta}_1} = \frac{\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{ns_x^2}}}}{\sqrt{\frac{(n-2)\hat{\sigma}^{*2}}{\frac{\sigma^2}{n-2}}}} = \frac{(\hat{\beta}_1 - \beta_1)\sqrt{ns_x^2}}{\sqrt{\frac{(n-2)\hat{\sigma}^{*2}}{(n-2)}}} \sim \mathcal{T}(n-2).$$

Finalement :

$$T_{\hat{\beta}_1} = \frac{(\hat{\beta}_1 - \beta_1)\sqrt{ns_x^2}}{\hat{\sigma}^*} \sim \mathcal{T}(n-2).$$

Afin de comprendre la construction des intervalles de confiance, commençons par représenter la loi de probabilité d'une variable aléatoire $T_{\hat{\beta}_1}$ suivant une loi de Student. Si la variable aléatoire $T_{\hat{\beta}_1}$ est comprise entre deux valeurs, les fractiles symétriques $-t_{\frac{\alpha}{2}}$ et $t_{\frac{\alpha}{2}}$ (voir la figure ci-dessous), alors l'intervalle de confiance bilatéral symétrique de $T_{\hat{\beta}_1}$ s'écrit :

$$\begin{aligned} 1-\alpha &= \mathbb{P}\left(-t_{\frac{\alpha}{2}} \leq T_{\hat{\beta}_1} \leq t_{\frac{\alpha}{2}}\right) \\ &= \mathbb{P}\left(-t_{\frac{\alpha}{2}} \leq \frac{(\hat{\beta}_1-\beta_1)\sqrt{ns_x^2}}{\hat{\sigma}^*} \leq t_{\frac{\alpha}{2}}\right). \end{aligned}$$



On en déduit alors :

$$1-\alpha = \mathbb{P}\left(-t_{\frac{\alpha}{2}} \cdot \frac{\hat{\sigma}^*}{\sqrt{ns_x^2}} \leq \hat{\beta}_1 - \beta_1 \leq t_{\frac{\alpha}{2}} \cdot \frac{\hat{\sigma}^*}{\sqrt{ns_x^2}}\right)$$

$$\iff 1-\alpha = \mathbb{P}\left(\underbrace{\hat{\beta}_1 - t_{\frac{\alpha}{2}} \cdot \frac{\hat{\sigma}^*}{\sqrt{ns_x^2}}}_{a} \leq \beta_1 \leq \underbrace{\hat{\beta}_1 + t_{\frac{\alpha}{2}} \cdot \frac{\hat{\sigma}^*}{\sqrt{ns_x^2}}}_{b}\right).$$

L'intervalle de confiance de β_1 s'écrit alors :

$$Ic_{1-\alpha}(\beta_1) = \left[\hat{\beta}_1 - t_{\frac{\alpha}{2}} \cdot \frac{\hat{\sigma}^*}{\sqrt{ns_x^2}} \ , \ \hat{\beta}_1 + t_{\frac{\alpha}{2}} \cdot \frac{\hat{\sigma}^*}{\sqrt{ns_x^2}}\right].$$

Proposition 1

Si le risque α est réparti à hauteur de $\alpha/2$ dans les queues de la distribution de $T_{\hat{\beta}_1}$ (voir la figure précédente), alors l'intervalle de confiance bilatéral de β_1 est symétrique en $\hat{\beta}_1$. On peut l'écrire :

$$Ic_{1-\alpha}(\beta_1) = \left[\hat{\beta}_1 \pm t_{\frac{\alpha}{2}} \cdot \frac{\hat{\sigma}^*}{\sqrt{ns_x^2}}\right].$$

Exemple 4

Application au jeu de données : intervalle de confiance de β_1, pour un niveau de confiance de $1-\alpha = 95\%$ (ou bien pour un risque d'erreur de $\alpha = 5\%$).

$$\begin{cases} T_{\hat{\beta}_1} \sim \mathcal{T}(7) & , \quad t_{\frac{\alpha}{2}} = t_{0,025} \approx 2,36 \\ \hat{\sigma}^* \approx 1,69 & , \quad \sqrt{ns_x^2} \approx 7,75 \end{cases}$$

$$Ic_{0,95}(\beta_1) = \left[-1,22 - 2,36 \cdot \frac{1,69}{7,75} \ ; \ -1,22 + 2,36 \cdot \frac{1,69}{7,75} \right] \approx [-1,73 \ ; \ -0,71]$$

3.1.2 Intervalle de confiance de β_0

Le problème est similaire à celui exposé précédemment. On souhaite déterminer un intervalle de confiance de β_0 en déterminant les réels a et b tels que, pour un risque d'erreur donné α :

$$1 - \alpha = \mathbb{P}\left(a \leq \beta_0 \leq b\right).$$

On sait que :

$$\hat{\beta}_0 \sim \mathcal{N}\left(\beta_0 \ , \ \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{ns_x^2}\right)\right).$$

De même ici, σ étant inconnu, il faut utiliser la loi de Student :

$$T_{\hat{\beta}_0} \equiv \frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}^* \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{ns_x^2}}} =\sim \mathcal{T}(n-2).$$

Pour construire l'intervalle de confiance de β_0, on suppose que la variable aléatoire $T_{\hat{\beta}_0}$ est comprise entre $t_{\frac{\alpha}{2}}$ et $t_{\frac{\alpha}{2}}$ pour une probabilité $1-\alpha$ donnée :

$$1 - \alpha = \mathbb{P}\left(-t_{\frac{\alpha}{2}} \leq T_{\hat{\beta}_0} \leq t_{\frac{\alpha}{2}}\right).$$

En remplaçant $T_{\hat{\beta}_0}$ par son expression, il vient :

$$1 - \alpha = \mathbb{P}\left(-t_{\frac{\alpha}{2}} \cdot \hat{\sigma}^* \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{ns_x^2}} \leq \hat{\beta}_0 - \beta_0 \leq t_{\frac{\alpha}{2}} \cdot \hat{\sigma}^* \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{ns_x^2}}\right)$$

$$\iff 1 - \alpha = \mathbb{P}\left(\underbrace{\hat{\beta}_0 - t_{\frac{\alpha}{2}} \cdot \hat{\sigma}^* \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{ns_x^2}}}_{a} \leq \beta_0 \leq \underbrace{\hat{\beta}_0 + t_{\frac{\alpha}{2}} \cdot \hat{\sigma}^* \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{ns_x^2}}}_{b}\right).$$

L'intervalle de confiance bilatéral symétrique de β_0 s'écrit alors :

$$\begin{aligned} Ic_{1-\alpha}(\beta_0) &= \left[\hat{\beta}_0 - t_{\frac{\alpha}{2}} \cdot \hat{\sigma}^* \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{ns_x^2}} \ , \ \hat{\beta}_0 + t_{\frac{\alpha}{2}} \cdot \hat{\sigma}^* \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{ns_x^2}}\right] \\ &= \left[\hat{\beta}_0 \pm t_{\frac{\alpha}{2}} \cdot \hat{\sigma}^* \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{ns_x^2}}\right]. \end{aligned}$$

Exemple 5

Application au jeu de données : intervalle de confiance de β_0, pour un niveau de confiance de $1-\alpha = 95\%$ (ou bien pour un risque d'erreur de $\alpha = 5\%$).

$$\begin{cases} T_{\hat{\beta}_0} \sim \mathcal{T}(7) & , \quad t_{\frac{\alpha}{2}} = t_{0,025} \approx 2,36 \\ \hat{\sigma}^* \approx 1,69 & , \quad \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{n s_x^2}} \approx 0,61 \end{cases}$$

$$Ic_{0,95}(\beta_0) = [11,75 - 2,36 \cdot 1,69 \cdot 0,61 \ ; \ 11,75 + 2,36 \cdot 1,69 \cdot 0,61] \approx [9,32 \ ; \ 14,18]$$

3.1.3 Intervalle de confiance de σ^2

Le problème est le suivant. On cherche un encadrement de σ^2, en déterminant les réels a et b tels que :

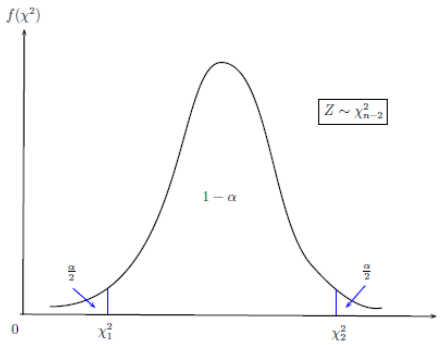
$$1 - \alpha = \mathbb{P}\left(a \leq \sigma^2 \leq b\right).$$

D'après le théorème de Fisher, on sait que :

$$Z \equiv (n-2)\frac{\hat{\sigma}^{*2}}{\sigma^2} \sim \chi^2_{n-2}.$$

La loi du χ^2 est positive, elle est notamment utilisée pour tester σ^2 ou pour en obtenir un intervalle de confiance. Si la variable aléatoire Z, suivant une loi de χ^2_{n-2}, est comprise entre les fractiles χ^2_1 et χ^2_2 tels que $\mathbb{P}\left(\chi^2_1 \leq Z\right) = 1 - \frac{\alpha}{2}$ et $\mathbb{P}\left(\chi^2_2 \leq Z\right) = \frac{\alpha}{2}$ (voir le graphique suivant), alors :

$$\begin{aligned} & 1 - \alpha = \mathbb{P}\left(\chi^2_1 \leq Z \leq \chi^2_2\right) \\ \iff & 1 - \alpha = \mathbb{P}\left(\frac{\chi^2_1}{(n-2)\hat{\sigma}^{*2}} \leq \frac{1}{\sigma^2} \leq \frac{\chi^2_2}{(n-2)\hat{\sigma}^{*2}}\right) \\ \iff & 1 - \alpha = \mathbb{P}\left(\underbrace{\frac{(n-2)\hat{\sigma}^{*2}}{\chi^2_2}}_{a} \leq \sigma^2 \leq \underbrace{\frac{(n-2)\hat{\sigma}^{*2}}{\chi^2_1}}_{b}\right). \end{aligned}$$



En utilisant la somme des carrés des résidus :

$$1-\alpha = \mathbb{P}\left(\frac{\sum_{i=1}^{n}\hat{\varepsilon}_i^2}{\chi_2^2} \leq \sigma^2 \leq \frac{\sum_{i=1}^{n}\hat{\varepsilon}_i^2}{\chi_1^2}\right).$$

L'intervalle de confiance bilatéral de σ^2 s'écrit alors :

$$Ic_{1-\alpha}(\sigma^2) = \left[\frac{(n-2)\hat{\sigma}^{*2}}{\chi_2^2}\ ,\ \frac{(n-2)\hat{\sigma}^{*2}}{\chi_1^2}\right].$$

Proposition 2

L'intervalle est bilatéral symétrique lorsque le risque α est réparti dans les mêmes proportions $(\alpha/2)$ dans les queues de la distribution de Z.

Exemple 6

Application au jeu de données : intervalle de confiance de σ^2, pour un niveau de confiance de $1-\alpha = 95\%$ (ou bien pour un risque d'erreur de $\alpha = 5\%$).

$$\left\{\begin{array}{lcl} Z \sim \chi_7^2 & , & \hat{\sigma}^{*2} \approx 2,86 \\ \chi_1^2 \approx 1,69 & , & \chi_2^2 \approx 16,01 \end{array}\right.$$

$$Ic_{0,95}(\sigma^2) = \left[\frac{7\cdot 2,86}{16,01}\ ;\ \frac{7\cdot 2,86}{1,69}\right] \approx [1,25;\ 11,85]$$

3.2 Tests d'hypothèse

3.2.1 Test sur β_1

On souhaite savoir si β_1 se rapproche d'une valeur hypothétique, fixée *a priori*, notée β_1^0 :

$$\left\|\begin{array}{l} H_0 : \beta_1 = \beta_1^0 \\ H_1 : \beta_1 \neq \beta_1^0. \end{array}\right.$$

Afin de tester la *validité du modèle*, on teste $\beta_1 = 0$. En effet, si l'hypothèse nulle $\beta_1 = 0$ est acceptée, cela signifie que qu'il n'existe aucune relation (linéaire) entre y et x. Au contraire, si l'hypothèse nulle H_0 est rejetée, le modèle est valide puisqu'une relation linéaire existe entre y et x (la qualité de cette relation sera discutée dans la section suivante).

Sous l'hypothèse H_0 :

$$T_{\hat{\beta}_1} = \frac{(\hat{\beta}_1 - \beta_1^0)\sqrt{ns_x^2}}{\hat{\sigma}^*} \sim \mathcal{T}(n-2).$$

D'où :

$$1-\alpha = \mathbb{P}\left(-t_{\frac{\alpha}{2}} \leq \frac{(\hat{\beta}_1 - \beta_1^0)\sqrt{ns_x^2}}{\hat{\sigma}^*} \leq t_{\frac{\alpha}{2}}\right) = \mathbb{P}\left(\left|T_{\hat{\beta}_1}\right| \leq t_{\frac{\alpha}{2}}\right).$$

On a alors :

$$1-\alpha = \mathbb{P}\left(\left|T_{\hat{\beta}_1}\right| \leq t_{\frac{\alpha}{2}}\right) = \mathbb{P}\left(\beta_1^0 - t_{\frac{\alpha}{2}}\cdot\frac{\hat{\sigma}^*}{\sqrt{ns_x^2}} \leq \hat{\beta}_1 \leq \beta_1^0 + t_{\frac{\alpha}{2}}\cdot\frac{\hat{\sigma}^*}{\sqrt{ns_x^2}}\right).$$

Ceci définit la *région d'acceptation* de l'hypothèse H_0.
Règle de décision :
• L'hypothèse H_0 est acceptée au risque de première espèce α si :

$$\hat{\beta}_1 \in \left[\beta_1^0 \pm t_{\frac{\alpha}{2}} \cdot \frac{\hat{\sigma}^*}{\sqrt{ns_x^2}}\right] \quad \text{ou} \quad \left|T_{\hat{\beta}_1}\right| \leq t_{\frac{\alpha}{2}}.$$

• L'hypothèse H_0 est rejetée au risque de première espèce α si :

$$\hat{\beta}_1 \notin \left[\beta_1^0 \pm t_{\frac{\alpha}{2}} \cdot \frac{\hat{\sigma}^*}{\sqrt{ns_x^2}}\right] \quad \text{ou} \quad \left|T_{\hat{\beta}_1}\right| > t_{\frac{\alpha}{2}}.$$

Lorsque la validité du modèle est testée, autrement dit $H_0 : \beta_1 = \beta_1^0 = 0$, l'intervalle d'acceptation de l'hypothèse H_0 se réécrit :

$$\hat{\beta}_1 \in \left[\pm t_{\frac{\alpha}{2}} \cdot \frac{\hat{\sigma}^*}{\sqrt{ns_x^2}}\right].$$

L'hypothèse H_0 doit être rejetée, au risque α, afin de conserver une relation linéaire entre les variables y et x.

Proposition 3

On peut tester β_1 en passant par la p-valeur (p_{val}) basée sur la région critique. En notant $t_{\hat{\beta}_1}$ la valeur calculée de la variable aléatoire $T_{\hat{\beta}_1}$:

$$p_{val} = \mathbb{P}_{H_0}\left(\left|T_{\hat{\beta}_1}\right| > \left|t_{\hat{\beta}_1}\right|\right).$$

L'hypothèse H_0 est rejetée au risque α lorsque $p_{val} \leq \alpha$.

Remarque 2

La notation $\mathbb{P}_{H_0}$ représente la probabilité calculée sous l'hypothèse H_0.

Exemple 7

Application au jeu de données : test de validité du modèle (test sur β_1), pour un risque de première espèce $\alpha = 5\%$.

$$\left\| \begin{array}{l} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0. \end{array} \right.$$

$$\left\{ \begin{array}{lll} T_{\hat{\beta}_1} \sim \mathcal{T}(7) & , & t_{\frac{\alpha}{2}} = t_{0,025} \approx 2,36 \\ \hat{\sigma}^* \approx 1,69 & , & \sqrt{ns_x^2} \approx 7,75 \end{array} \right.$$

Intervalle d'acceptation est $\left[\pm 2,36 \cdot \frac{1,69}{7,75}\right] = [-0,51\ ;\ 0,51]$. *Comme* $\hat{\beta}_1 = -1,22$ *n'appartient pas à l'intervalle d'acceptation, l'hypothèse H_0 doit être rejetée, au risque $\alpha = 5\%$.*

De manière alternative, en calculant la p-valeur, on obtient $p_{val} = \mathbb{P}_{H_0}\left(\left|T_{\hat{\beta}_1}\right| > \left|t_{\hat{\beta}_1}\right|\right) = \mathbb{P}_{H_0}\left(\left|T_{\hat{\beta}_1}\right| > |-5,6|\right) \approx 0,0008$ *(la valeur de $t_{\hat{\beta}_1}$ étant calculée, sous H_0, selon la formule $t_{\hat{\beta}_1} = \hat{\beta}_1\sqrt{ns_x^2}/\hat{\sigma}^*$). Comme $0,0008 < 5\%$, on rejette l'hypothèse H_0 au risque $\alpha = 5\%$.*

3.2.2 Test sur β_0

Il s'agit d'un test sur la nullité de la constante (l'ordonnée à l'origine) :

$$\left\| \begin{array}{l} H_0 : \beta_0 = \beta_0^0 \\ H_1 : \beta_0 \neq \beta_0^0. \end{array} \right.$$

La statistique de test est :

$$T_{\hat{\beta}_0} = \frac{(\hat{\beta}_0 - \beta_0^0)}{\hat{\sigma}^* \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{n s_x^2}}} \sim \mathcal{T}(n-2).$$

Par conséquent, sous l'hypothèse H_0, pour un risque donné α on a :

$$1 - \alpha = \mathbb{P}\left(-t_{\frac{\alpha}{2}} \leq T_{\hat{\beta}_0} \leq t_{\frac{\alpha}{2}}\right) = \mathbb{P}\left(\beta_0^0 - t_{\frac{\alpha}{2}} \hat{\sigma}^* \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{n s_x^2}} \leq \hat{\beta}_0 \leq \beta_0^0 + t_{\frac{\alpha}{2}} \hat{\sigma}^* \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{n s_x^2}}\right).$$

Ceci définit la *région d'acceptation* de l'hypothèse H_0.
Règle de décision :
• L'hypothèse H_0 est acceptée au risque de première espèce α si (pour $\beta_0^0 = 0$) :

$$\hat{\beta}_0 \in \left[\pm t_{\frac{\alpha}{2}} \cdot \hat{\sigma}^* \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{n s_x^2}}\right] \quad \text{ou} \quad \left|T_{\hat{\beta}_1}\right| \leq t_{\frac{\alpha}{2}}.$$

• L'hypothèse H_0 est rejetée au risque de première espèce α si :

$$\hat{\beta}_0 \notin \left[\pm t_{\frac{\alpha}{2}} \cdot \hat{\sigma}^* \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{n s_x^2}}\right] \quad \text{ou} \quad \left|T_{\hat{\beta}_1}\right| > t_{\frac{\alpha}{2}}.$$

Proposition 4

L'hypothèse H_0 est rejetée au risque α lorsque $p_{val} \leq \alpha$ avec :

$$p_{val} = \mathbb{P}_{H_0}\left(\left|T_{\hat{\beta}_0}\right| > \left|t_{\hat{\beta}_0}\right|\right).$$

Exemple 8

Application au jeu de données : test sur la nullité de la constante (test sur β_0), pour un risque de première espèce $\alpha = 5\%$

$$\left\| \begin{array}{l} H_0 : \beta_0 = 0 \\ H_1 : \beta_0 \neq 0. \end{array} \right.$$

$$\left\{ \begin{array}{lll} T_{\hat{\beta}_0} \sim \mathcal{T}(7) & , & t_{\frac{\alpha}{2}} = t_{0,025} \approx 2,36 \\ \hat{\sigma}^* \approx 1,69 & , & \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{n s_x^2}} \approx 0,61 \end{array} \right.$$

Intervalle d'acceptation est $[\pm 2,36 \cdot 1,69 \cdot 0,61] = [-2,43\ ;\ 2,43]$. *Comme* $\hat{\beta}_0 = 11,75$ *n'appartient pas à l'intervalle d'acceptation, l'hypothèse H_0 doit être rejetée, au risque* $\alpha = 5\%$.

De manière alternative, en calculant la p-valeur, on obtient $p_{val} = \mathbb{P}_{H_0}\left(\left|T_{\hat{\beta}_0}\right| > \left|t_{\hat{\beta}_0}\right|\right) = \mathbb{P}_{H_0}\left(\left|T_{\hat{\beta}_1}\right| > |11,3|\right) \approx 9,5 \times 10^{-6}$ *(la valeur de* $t_{\hat{\beta}_0}$ *étant calculée, sous* H_0*, selon la formule* $t_{\hat{\beta}_0} = \hat{\beta}_0 / \left(\hat{\sigma}^* \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{ns_x^2}}\right)$*). Comme la* p_{val} *est inférieure à* $\alpha = 5\%$*, on rejette* H_0 *au risque* α*.*

3.2.3 Test sur σ^2

On souhaite tester une valeur particulière de la variance de l'aléa. En effet une valeur trop importante de celle-ci serait synonyme d'un modèle qui s'ajuste mal aux données (*cf.* la section suivante sur l'analyse de la variance). Testons la variance de l'aléa à la valeur σ_0^2 :

$$\left\| \begin{array}{l} H_0 : \sigma^2 = \sigma_0^2 \\ H_1 : \sigma^2 \neq \sigma_0^2. \end{array} \right.$$

La statistique de test est :

$$Z = (n-2)\frac{\hat{\sigma}^{*2}}{\sigma_0^2} \sim \chi^2_{n-2}.$$

Pour un risque donné α la statistique de test est comprise entre deux fractiles χ_1^2 et χ_2^2 tels que $\mathbb{P}\left(\chi_1^2 \leq Z\right) = 1 - \frac{\alpha}{2}$ et $\mathbb{P}\left(\chi_2^2 \leq Z\right) = \frac{\alpha}{2}$:

$$1 - \alpha = \mathbb{P}\left(\chi_1^2 \leq (n-2)\frac{\hat{\sigma}^{*2}}{\sigma_0^2} \leq \chi_2^2\right).$$

Ainsi, la région d'acceptation de l'hypothèse H_0 est :

$$1 - \alpha = \mathbb{P}\left(\frac{\sigma_0^2\chi_1^2}{n-2} \leq \hat{\sigma}^{*2} \leq \frac{\sigma_0^2\chi_2^2}{n-2}\right).$$

Règle de décision :

- L'hypothèse H_0 est acceptée au risque de première espèce α si :

$$\hat{\sigma}^{*2} \in \left[\frac{\sigma_0^2\chi_1^2}{n-2} \ , \ \frac{\sigma_0^2\chi_2^2}{n-2}\right].$$

- L'hypothèse H_0 est rejetée au risque de première espèce α si :

$$\hat{\sigma}^{*2} \notin \left[\frac{\sigma_0^2\chi_1^2}{n-2} \ , \ \frac{\sigma_0^2\chi_2^2}{n-2}\right].$$

Exemple 9

Application au jeu de données : test pour une valeur particulière de la variance de l'aléa, pour un risque de première espèce $\alpha = 5\%$*. Prenons* $\sigma_0^2 = 3$*, par exemple.*

$$\left\| \begin{array}{l} H_0 : \sigma^2 = 3 \\ H_1 : \sigma^2 \neq 3. \end{array} \right.$$

$$\left\{ \begin{array}{lll} Z \sim \chi^2_7 & , & \hat{\sigma}^{*2} \approx 2,86 \\ \chi_1^2 \approx 1,69 & , & \chi_2^2 \approx 16,01 \end{array} \right.$$

Intervalle d'acceptation est $\left[\frac{3 \cdot 1,69}{7} \ ; \ \frac{3 \cdot 16,01}{7}\right] = [0,72 \ ; \ 6,86]$*. Comme* $\hat{\sigma}^{*2} = 2,86$ *appartient à l'intervalle d'acceptation, l'hypothèse* H_0 *ne peut pas être rejetée, au risque* $\alpha = 5\%$*.*

4 Corrélation et analyse de la variance

Le coefficient de corrélation linéaire $r_{y/x}$ mesure le degré de covariation linéaire entre deux variables, c'est-à-dire l'intensité avec laquelle les deux variables y et x vont varier conjointement :

$$r_{y/x} = \frac{\mathbb{C}\text{ov}\,(x,y)}{s_y s_x} = \frac{\frac{1}{n}\sum_{i=1}^n y_i x_i - \bar{y}\bar{x}}{\sqrt{\frac{1}{n}\sum_{i=1}^n y_i^2 - \bar{y}^2}\sqrt{\frac{1}{n}\sum_{i=1}^n x_i^2 - \bar{x}^2}}.$$

4.1 Propriétés

- $r_{y/x} \in [-1;1]$:
  $\hookrightarrow$ plus $r_{y/x}$ se rapproche de 1 (ou -1), plus l'intensité de la covariation linéaire est forte ;
  $\hookrightarrow$ plus $r_{y/x}$ s'éloigne de 1 (ou -1), plus l'intensité de la covariation linéaire est faible.

- $r_{y/x}$ est sans dimension.
- $r_{y/x}$ est symétrique : $r_{y/x} = r_{x/y}$.
- $r_{y/x}$ n'est pas affecté par un changement de variable :
  $\hookrightarrow$ $r_{y/x} = r_{y^c/x^c}$;
  $\hookrightarrow$ $\forall k, k' \in \mathbb{R}\setminus\{0\}$, $r_{ky/k'x} = r_{y/x}$.

4.2 Relation entre $\hat{\beta}_1$ et $r_{y/x}$

Puisque,

$$r_{y/x} = r_{y^c/x^c} = \frac{\mathbb{C}\text{ov}(x,y)}{\sqrt{s_y^2 s_x^2}} = \frac{\frac{1}{n}\sum_{i=1}^n (y_i-\bar{y})(x_i-\bar{x})}{\sqrt{s_y^2 s_x^2}},$$

et comme,

$$\hat{\beta}_1 = \frac{\mathbb{C}\text{ov}(x,y)}{s_x^2},$$

alors :

$$r_{y/x} = r_{y^c/x^c} = \frac{ns_x^2}{ns_x^2}\frac{(\sum_{i=1}^n y_i^c x_i^c)}{\sqrt{ns_x^2 ns_y^2}} = \hat{\beta}_1 \frac{ns_x^2}{\sqrt{ns_x^2}\sqrt{ns_y^2}} = \hat{\beta}_1 \frac{s_x}{s_y}.$$

Exemple 10

Application au jeu de données : calcul de $r_{y/x}$.

$$s_x = 2,58 \quad , \quad s_y = 3,48 \quad , \quad r_{y/x} = \hat{\beta}_1 \frac{s_x}{s_y} = -1,22 \cdot \frac{2,58}{3,48} \approx -0,9.$$

Il existe une forte corrélation négative entre y et x.

4.3 Analyse de la variance

L'analyse de la variance permet de décomposer la variance totale en variance expliquée et variance résiduelle afin de mesurer la qualité du modèle de régression. Par définition, nous avons $\hat{\varepsilon}_i = y_i - \hat{y}_i \iff y_i^c = \hat{\varepsilon}_i + \hat{y}_i^c$ avec $\hat{y}_i^c = \hat{\beta}_1 x_i^c$, d'où :

$$\sum_{i=1}^{n}(y_i^c)^2 = \sum_{i=1}^{n}(\hat{y}_i^c)^2 + \sum_{i=1}^{n}\hat{\varepsilon}_i^2 + 2\sum_{i=1}^{n}\hat{y}_i^c\hat{\varepsilon}_i.$$

Or :

$$\begin{aligned}\sum_{i=1}^{n}\hat{\varepsilon}_i^2\hat{y}_i^c &= \sum_{i=1}^{n}\left(y_i^c - \hat{y}_i^c\right)\hat{y}_i^c \\ &= \sum_{i=1}^{n}\left(y_i^c - \hat{y}_i^c\right)\hat{\beta}_1 x_i^c \\ &= \hat{\beta}_1\left(\sum_{i=1}^{n}\left(y_i^c - \hat{\beta}_1 x_i^c\right)x_i^c\right) \\ &= \hat{\beta}_1\left(\sum_{i=1}^{n} y_i^c x_i^c - \hat{\beta}_1 n s_x^2\right).\end{aligned}$$

Aussi,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i^c y_i^c}{n s_x^2} \implies \sum_{i=1}^{n} x_i^c y_i^c - \hat{\beta}_1 n s_x^2 = 0 \implies \sum_{i=1}^{n}\hat{\varepsilon}_i\hat{y}_i^c = 0.$$

Donc :

$$\sum_{i=1}^{n}(y_i^c)^2 = \sum_{i=1}^{n}(\hat{y}_i^c)^2 + \sum_{i=1}^{n}\hat{\varepsilon}_i^2. \qquad \text{(ANOVA)}$$

Il s'agit de l'équation de l'analyse de la variance qui décrit la décomposition de la variabilité totale du nuage de points en variations expliquées et variations résiduelles. En effet :

$$SCT = \sum_{i=1}^{n}(y_i^c)^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2 \quad : \text{variance de } y \text{ (à } n \text{ près)}$$

$$SCE = \sum_{i=1}^{n}(\hat{y}_i^c)^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \quad : \text{variance de } \hat{y} \text{ (à } n \text{ près)}$$

$$SCR = \sum_{i=1}^{n}\hat{\varepsilon}_i^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad : \text{variance de } \hat{\varepsilon} \text{ (à } n \text{ près),}$$

avec SCT la somme des carrés totaux, SCE la somme des carrés expliqués (par la droite de régression), et SCR la somme des carrés résiduels. On écrit l'analyse de la variance :

$$\boxed{SCT = SCE + SCR.}$$

En divisant les deux membres de l'équation (ANOVA) par $\sum_{i=1}^{n}(y_i^c)^2$ on a :

$$\begin{aligned}\frac{\sum_{i=1}^{n}(y_i^c)^2}{\sum_{i=1}^{n}(y_i^c)^2} &= \frac{\sum_{i=1}^{n}(\hat{y}_i^c)^2}{\sum_{i=1}^{n}(y_i^c)^2} + \frac{\sum_{i=1}^{n}\hat{\varepsilon}_i^2}{\sum_{i=1}^{n}(y_i^c)^2} \\ 1 &= \frac{\sum_{i=1}^{n}(\hat{y}_i^c)^2}{\sum_{i=1}^{n}(y_i^c)^2} + \frac{\sum_{i=1}^{n}\hat{\varepsilon}_i^2}{\sum_{i=1}^{n}(y_i^c)^2}.\end{aligned}$$

On appelle coefficient de détermination, noté R^2, le rapport de la somme expliquée à la somme totale :

$$\begin{aligned} R^2 &= \frac{\sum_{i=1}^n (\hat{y}_i^c)^2}{\sum_{i=1}^n (y_i^c)^2} = 1 - \frac{\sum_{i=1}^n \hat{\varepsilon}_i^2}{\sum_{i=1}^n (y_i^c)^2}. \\ &= \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT}. \end{aligned}$$

Ce coefficient de détermination s'interprète comme un pourcentage. En effet, par construction :

$$R^2 \in [0\ ;\ 1].$$

Par exemple, $R^2 = 80\%$ signifie que 80% de la variance totale (la variance de y) est expliquée par la doite de régression. Dans le cas de la régression linéaire, on peut montrer que :

$$R^2 = \left(r_{y/x}\right)^2.$$

En effet :

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i^c)^2}{ns_y^2} = \frac{\left(\hat{\beta}_1 \sum_{i=1}^n x_i^c\right)^2}{ns_y^2} = \hat{\beta}_1^2 \frac{ns_x^2}{ns_y^2} = \left(\hat{\beta}_1 \frac{s_x}{s_y}\right)^2 = \left(r_{y/x}\right)^2.$$

Exemple 11

Application au jeu de données : calcul du coefficient de détermination R^2.

$$SCT = 108,88 \quad , \quad SCE = 88,81 \quad , \quad SCR = 20,07$$

On peut vérifie qu'en effet $SCT = SCE + SCR$.

$$R^2 = \frac{SCE}{SCT} = \frac{88,81}{108,88} \approx 0,81$$

$$\left(r_{y/x}\right)^2 = (-0,9)^2 = 0,81$$

4.4 Test du coefficient de corrélation linéaire

Comme $\hat{\beta}_1 = r_{y/x} \frac{s_y}{s_x}$, l'absence de relation entre x et y, *i.e.* $\hat{\beta}_1 = 0$, se traduit par un coefficient de corrélation linéaire nul. On peut donc utiliser le test de Student du paramètre β_1 pour tester la signification du coefficient de corrélation. On sait que :

$$\frac{(\hat{\beta}_1 - \beta_1^0)\sqrt{ns_x^2}}{\hat{\sigma}^*} \sim \mathcal{T}(n-2).$$

D'où :

$$\frac{(\hat{\beta}_1 - \beta_1^0)\sqrt{ns_x^2}\sqrt{n-2}}{\sum_{i=1}^n \hat{\varepsilon}_i^2} \sim \mathcal{T}(n-2).$$

Si l'hypothèse H_0 est vraie, *i.e.* $\beta_1 = \beta_1^0 = 0$, alors :

$$\frac{\hat{\beta}_1\sqrt{ns_x^2}\sqrt{n-2}}{\sum_{i=1}^n \hat{\varepsilon}_i^2} \sim \mathcal{T}(n-2).$$

Comme,

$$r_{y/x} = \hat{\beta}_1 \frac{\sqrt{ns_x^2}}{\sqrt{ns_y^2}},$$

et que,

$$r_{y/x} = \sqrt{1 - \frac{\sum_{i=1}^n \hat{\varepsilon}_i^2}{\sum_{i=1}^n (y_i^c)^2}} \iff \sqrt{\sum_{i=1}^n \hat{\varepsilon}_i^2} = \sqrt{1-(r_{y/x})^2}\sqrt{ns_y^2},$$

alors :

$$\frac{\hat{\beta}_1\sqrt{ns_x^2}\sqrt{n-2}}{\sqrt{\sum_{i=1}^n \hat{\varepsilon}_i^2}} = \frac{r_{y/x}\sqrt{ns_y^2}\sqrt{n-2}}{\sqrt{1-(r_{y/x})^2}\sqrt{ns_y^2}} \sim \mathcal{T}(n-2).$$

Ainsi :

$$T_r = \frac{r_{y/x}\sqrt{n-2}}{\sqrt{1-R^2}} \sim \mathcal{T}(n-2).$$

Le test d'hypothèse du paramètre ρ, qui est la valeur du coefficient de corrélation linéaire du modèle théorique $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, se spécifie de la manière suivante :

$$\left\| \begin{array}{l} H_0 : \rho = 0 \\ H_1 : \rho \neq 0. \end{array} \right.$$

Pour un risque donné α, la statistique de test est comprise entre deux fractiles $-t_{\frac{\alpha}{2}}$ et $t_{\frac{\alpha}{2}}$, ainsi la région d'acceptation de l'hypothèse H_0 est donnée par :

$$1-\alpha = \mathbb{P}\left(-t_{\frac{\alpha}{2}} \leq \frac{r_{y/x}\sqrt{n-2}}{\sqrt{1-R^2}} \leq t_{\frac{\alpha}{2}}\right) = \mathbb{P}\left(\left|\frac{r_{y/x}\sqrt{n-2}}{\sqrt{1-R^2}}\right| \leq t_{\frac{\alpha}{2}}\right).$$

Règle de décision :

- L'hypothèse H_0 est acceptée au risque de première espèce α (modèle non valide) si :

$$\left|\frac{r_{y/x}\sqrt{n-2}}{\sqrt{1-R^2}}\right| \leq t_{\frac{\alpha}{2}} \quad \text{ou} \quad |T_r| \leq t_{\frac{\alpha}{2}}.$$

- L'hypothèse H_0 est rejetée au risque de première espèce α (modèle valide) si :

$$\left|\frac{r_{y/x}\sqrt{n-2}}{\sqrt{1-R^2}}\right| > t_{\frac{\alpha}{2}} \quad \text{ou} \quad |T_r| > t_{\frac{\alpha}{2}}.$$

Proposition 5

En notant t_r la valeur calculée de la variable aléatoire T_r, l'hypothèse H_0 est rejetée au risque α lorsque $p_{val} \leq \alpha$, avec :

$$p_{val} = \mathbb{P}_{H_0}\left(|T_r| > |t_r|\right).$$

Exemple 12

Application au jeu de données : test de signification du coefficient linéaire, pour un risque de première espèce $\alpha = 5\%$.

$$\left\| \begin{array}{l} H_0 : \rho = 0 \\ H_1 : \rho \neq 0. \end{array} \right.$$

$$\begin{cases} T_r \sim \mathcal{T}(7) & , \quad t_{\frac{\alpha}{2}} = t_{0,025} \approx 2,36 \\ r_{y/x} = -0,9 & , \quad R^2 = 0,81 \end{cases}$$

Calculons $\left|\frac{r_{y/x}\sqrt{n-2}}{\sqrt{1-R^2}}\right| = \left|\frac{-0,9\sqrt{7}}{\sqrt{1-0,81}}\right| = 5,46$. *Cette valeur étant supérieur au fractile* $2,36$, *on rejette l'hypothèse* H_0 *et on conclut que le modèle est valide.*

4.5 Tableau de l'analyse de la variance

ANOVA

Variations	Somme des carrés des écarts	Degré(s) de liberté	Variance ou carrés moyens
Expliquée	$SCE = \sum_{i=1}^n (\hat{y}_i^c)^2 = \hat{\beta}_1^2 \sum_{i=1}^n (x_i^c)^2$	1	$\frac{SCE}{1} = \hat{\beta}_1^2 n s_x^2$
Résiduelle	$SCR = \sum_{i=1}^n \hat{\varepsilon}_i^2$	$n-2$	$\frac{SCR}{n-2} = \frac{\sum_{i=1}^n \hat{\varepsilon}_i^2}{n-2} = \hat{\sigma}^{*2}$
Totale	$SCT = n s_y^2$	$n-1$	$\times$

Rappelons que la loi de Student se construit de la manière suivante :

$$T = \frac{U}{\sqrt{\frac{Z}{d}}} \sim \mathcal{T}(n-2),$$

avec $U \sim \mathcal{N}(0,1)$, $Z \sim \chi_d^2$ où U et Z sont des variables aléatoires indépendantes. On sait par ailleurs qu'il existe une relation entre la loi de Fisher et la loi de Student : la loi de Fisher se construit à partir du carré de la loi de Student. Soit $F \sim \mathcal{F}(1, n-2)$ et $T \sim \mathcal{T}(n-2)$, on a :

$$F = T^2.$$

A partir de ce résultat, on peut en déduire le test du coefficient de détermination.

Exemple 13

Application au jeu de données : tableau ANOVA.

Variations	*Somme des carrés des écarts*	*Degré(s) de liberté*	*Variance ou carrés moyens*
Expliquée	$SCE = 88,81$	1	$\frac{SCE}{1} = 88,81$
Résiduelle	$SCR = 20,07$	7	$\frac{SCR}{7} = \hat{\sigma}^{*2} = 2,86$
Totale	$SCT = 108,88$	8	$\times$

4.6 Test du coefficient de détermination

On a vu que :

$$\hat{\beta}_1 \frac{\sqrt{n s_x^2}\sqrt{n-2}}{\sqrt{\sum_{i=1}^n \hat{\varepsilon}_i^2}} = \frac{r_{y/x}}{\sqrt{1-R^2}}\sqrt{n-2} \sim \mathcal{T}(n-2).$$

D'où :

$$\hat{\beta}_1^2 \frac{n s_x^2 (n-2)}{\sum_{i=1}^n \hat{\varepsilon}_i^2} = \frac{R^2}{1-R^2}(n-2) = \frac{\sum_{i=1}^n \hat{y}_i^2}{\sum_{i=1}^n \hat{\varepsilon}_i^2}(n-2) \sim \mathcal{F}(1, n-2).$$

Pour tester le coefficient de détermination du modèle linéaire théorique ρ^2, on utilise la statistique de test suivante :

$$F = \frac{R^2}{1-R^2}(n-2) \sim \mathcal{F}(1, n-2).$$

Le test est spécifié de la manière suivante :

$$\left\| \begin{array}{l} H_0 : \rho^2 = 0 \\ H_1 : \rho^2 \neq 0. \end{array} \right.$$

Pour un risque donné α, il est possible de construire un intervalle unilatéral (un intervalle bilatéral conduirait à un non-sens) représentant la région d'acceptation de H_0 :

$$1 - \alpha = \mathbb{P}\left(F \leq f_{1-\alpha}\right).$$

Règle de décision :

- L'hypothèse H_0 est acceptée au risque de première espèce α (modèle non valide) si :

$$\frac{R^2}{1-R^2}(n-2) = F \leq f_{1-\alpha}.$$

- L'hypothèse H_0 est rejetée au risque de première espèce α (modèle valide) si :

$$\frac{R^2}{1-R^2}(n-2) = F > f_{1-\alpha}.$$

Proposition 6

En notant f la valeur calculée de la variable aléatoire F, l'hypothèse H_0 est rejetée au risque α lorsque $p_{val} \leq \alpha$, avec :

$$p_{val} = \mathbb{P}_{H_0}\left(F > f\right).$$

Exemple 14

Application au jeu de données : test du coefficient de détermination, pour un risque de première espèce $\alpha = 5\%$

$$\left\| \begin{array}{l} H_0 : \rho^2 = 0 \\ H_1 : \rho^2 \neq 0. \end{array} \right.$$

$$\left\{ \begin{array}{l} F \sim \mathcal{F}(1,7) \quad , \quad f_{1-\alpha} = f_{0,95} \approx 5,59 \\ R^2 = 0,81 \end{array} \right.$$

Calculons $\frac{R^2}{1-R^2}(n-2) = \frac{0,81}{1-0,81} \cdot 7 = 29,84$. Cette valeur étant supérieur au fractile $5,59$, on rejette l'hypothèse H_0 et on conclut que le modèle est valide.

5 Utilisation du modèle de régression en prévision

On peut utiliser le modèle estimé en prévision de deux façons.

• Prévoir la valeur moyenne de la variable expliquée pour une valeur donnée de la variable explicative au point i. Cette valeur moyenne est l'espérance mathématique des valeurs possibles de Y_0 (y_i à prévoir) associées à x_0 (valeur donnée de x_i) c'est-à-dire $\mathbb{E}(Y_0|x_0)$. Cette *espérance conditionnelle* est estimée par intervalle de confiance.

• Vérifier qu'une prévision ponctuelle donnée (Y_0, x_0) est compatible avec la relation linéaire estimée. Il s'agit d'un test.

5.1 Intervalle de confiance d'une valeur moyenne de Y connaissant une valeur donnée de x

Soit x_0 la valeur donnée de x. Soit la valeur x_0 correspond à une observation appartenant à l'ensemble $\{1, \ldots, n\}$, soit la valeur x_0 correspond à une observation provenant de l'ensemble $\{n+1, \ldots, n+h\}$ où h représente l'horizon de la prévision.
Le modèle théorique s'écrit :

$$Y_0 = \beta_0 + \beta_1 x_0 + \varepsilon_0.$$

Le modèle estimé s'écrit :

$$\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0.$$

La valeur moyenne de Y connaissant une valeur donnée de x est notée :

$$\mathbb{E}[Y_0|x_0] = \beta_0 + \beta_1 x_0.$$

La statistique $\hat{Y}_0$ est l'estimateur linéaire sans biais de $\mathbb{E}[Y_0|x_0]$ (*cf.* démonstration ci-dessous). Comme $\hat{\beta}_0$ et $\hat{\beta}_1$ suivent une loi normale, alors $\hat{Y}_0$ suit une loi normale :

$$\hat{Y}_0 \sim \mathcal{N}\left(\mathbb{E}[\hat{Y}_0], \mathbb{V}[\hat{Y}_0]\right).$$

On montre que :

$$\mathbb{E}[\hat{Y}_0] = \beta_0 + \beta_1 x_0 \ , \ \mathbb{V}[\hat{Y}_0] = \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{n s_x^2}\right).$$

Démonstration :

$$\mathbb{E}[\hat{Y}_0] = \mathbb{E}[\hat{\beta}_0 + \hat{\beta}_1 x_0] = \beta_0 + \beta_1 x_0 = \mathbb{E}[Y_0|x_0].$$

$$\mathbb{V}[\hat{Y}_0] = \mathbb{V}[\hat{\beta}_0 + \hat{\beta}_1 x_0] = \mathbb{V}[\hat{\beta}_0] + x_0^2 \mathbb{V}[\beta_1] + 2x_0 \mathbb{C}\text{ov}\left(\hat{\beta}_0, \hat{\beta}_1\right).$$

Or, on a vu que,

$$\mathbb{V}[\hat{\beta}_1] = \frac{\sigma^2}{n s_x^2},$$

et

$$\mathbb{V}[\hat{\beta}_0] = \sigma^2 \frac{\sum_{i=1}^n x_i^2}{n^2 s_x^2} = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{n s_x^2}\right).$$

En effet, par définition :

$$n s_x^2 = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 + \sum_{i=1}^n \bar{x}^2 + 2\sum_{i=1}^n x_i \bar{x} = \sum_{i=1}^n x_i^2 - n\bar{x}^2.$$

D'où :

$$\frac{\sum_{i=1}^n x_i^2}{n n s_x^2} = \frac{n s_x^2 + n\bar{x}^2}{n n s_x^2} = \frac{1}{n} + \frac{\bar{x}^2}{n s_x^2}.$$

Par ailleurs,

$$\mathbb{C}\text{ov}\left(\hat{\beta}_0, \hat{\beta}_1\right) = \frac{-\bar{x}}{ns_x^2}\sigma^2,$$

d'où :

$$\begin{aligned}\mathbb{V}[\hat{Y}_0] = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{ns_x^2}\right) + x_0^2\frac{\sigma^2}{ns_x^2} - 2x_0\frac{\bar{x}}{ns_x^2}\sigma^2 &= \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2 + x_0^2 - 2x_0\bar{x}}{ns_x^2}\right) \\ &= \sigma^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{ns_x^2}\right).\end{aligned}$$

Par conséquent :

$$\hat{Y}_0 \sim \mathcal{N}\left(\beta_0 + \beta_1 x_0\ ,\ \sigma^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{ns_x^2}\right)\right).$$

On peut également démontrer, que l'estimateur trouvé est un estimateur sans biais et de variance minimale. L'intervalle de confiance de la valeur moyenne de $Y|x$ peut se construire en posant le problème suivant. On cherche les réels a et b tels que, pour un risque donné α :

$$1 - \alpha = \mathbb{P}\left(a \leq \mathbb{E}[Y_0|x_0] \leq b\right) = \mathbb{P}\left(a \leq \beta_0 + \beta_1 x_0 \leq b\right).$$

Autrement dit, on veut déterminer un intervalle de confiance de la valeur moyenne de $Y|x$. On a :

$$(n-2)\frac{\hat{\sigma}^{*2}}{\sigma^2} \sim \chi^2_{n-2}.$$

Comme σ est inconnu, on utilise la loi de Student :

$$T_{\hat{Y}_0} \equiv \frac{\frac{\hat{Y}_0 - (\beta_0 + \beta_1 x_0)}{\sigma\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{ns_x^2}}}}{\sqrt{\frac{(n-2)\frac{\hat{\sigma}^{*2}}{\sigma^2}}{n-2}}} = \frac{\hat{Y}_0 - (\beta_0 + \beta_1 x_0)}{\hat{\sigma}^*\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{ns_x^2}}} \sim \mathcal{T}(n-2).$$

L'intervalle de confiance de $T_{\hat{Y}_0}$, pour un risque donné α est :

$$1 - \alpha = \mathbb{P}\left(-t_{\frac{\alpha}{2}} \leq T_{\hat{Y}_0} \leq t_{\frac{\alpha}{2}}\right),$$

Proposition 7

L'intervalle de confiance bilatéral symétrique de $\mathbb{E}[Y_0|x_0]$ est :

$$Ic_{1-\alpha}(\mathbb{E}[Y_0|x_0]) = \left[\hat{Y}_0 \pm t_{\frac{\alpha}{2}}\hat{\sigma}^*\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{ns_x^2}}\right].$$

Exemple 15

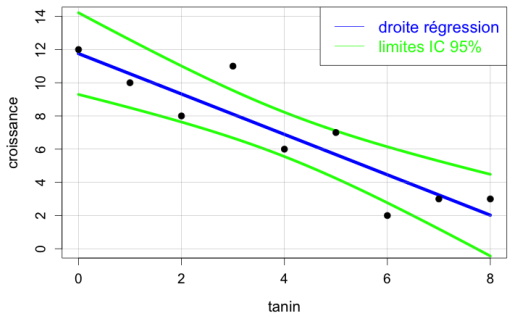
Application au jeu de données : construction de l'intervalle de confiance de la valeur moyenne de Y en sachant une valeur de x, disons $x_0 = 4,5$, pour un niveau de confiance

de $1-\alpha = 95\%$.

$$\begin{cases} \hat{y}_0 = 11,75 - 1,22 * 4,5 = 6,26 \quad , \quad t_{\frac{\alpha}{2}} = t_{0,025} \approx 2,36 \\ \hat{\sigma}^* = 1,69 \end{cases}$$

$$Ic_{0,95}(\mathbb{E}[Y_0|4,5]) = [6,26 \pm 1,35] = [4,91 \ ; \ 7,61]\,.$$

En faisant le même calcul pour toutes les valeurs x_0 *appartenant à l'intervalle* $[0\ ;\ 8]$, *on obtient une bande de confiance autour de la droite de régression, bande représentée ci-dessous.*



5.2 Vérification de la compatibilité entre la prévision ponctuelle et la relation linéaire estimée

Soit un couple de valeurs (x_0, Y_0), deux cas sont possibles : soit x_0 et Y_0 sont des valeurs dont leurs observations appartiennent à l'ensemble $\{1,\ldots,n\}$ ou à l'ensemble $\{n+1,\ldots,n+h\}$ avec h l'horizon de la prévision.

On se demande si le point du nuage de régression obtenu par ce couple de valeurs peut être considéré comme appartenant à la droite de régression estimée. On va donc accepter ou rejeter la compatibilité d'une prévision ponctuelle donnée, admettons une valeur réelle a, avec la relation estimée :

$$\begin{array}{l} H_0 : \mathbb{E}[Y_0|x_0] = a \iff \text{compatibilité de la prévision } a \text{ avec le modèle} \\ H_1 : \mathbb{E}[Y_0|x_0] \neq a \iff \text{non compatibilité de la prévision } a \text{ avec le modèle.} \end{array}$$

On peut démontrer, en passant par la loi du résidu $\hat{\varepsilon}_0$, que :

$$T_{\hat{\varepsilon}_0} \equiv \frac{\mathbb{E}\left(Y_0|x_0\right) - \hat{Y}_0}{\hat{\sigma}^* \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{n s_x^2}}} \sim \mathcal{T}(n-2).$$

L'intervalle de confiance de $T_{\hat{\varepsilon}_0}$, pour un risque donné α, est donné par :

$$1 - \alpha = \mathbb{P}\left(-t_{\frac{\alpha}{2}} \leq T_{\hat{\varepsilon}_0} \leq t_{\frac{\alpha}{2}}\right).$$

La région d'acceptation est donc :

$$1 - \alpha = \mathbb{P}\left(\hat{Y}_0 \in \left[\underbrace{(\beta_0 + \beta_1 x_0)}_{a} \pm t_{\frac{\alpha}{2}} \hat{\sigma}^* \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{n s_x^2}}\right]\right).$$

Règle de décision :

• L'hypothèse H_0 est acceptée au risque de première espèce α (compatibilité du modèle linéaire avec a) si :

$$\hat{Y}_0 \in \left[a \pm t_{\frac{\alpha}{2}} \hat{\sigma}^* \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{n s_x^2}} \right] \quad \text{ou} \quad |T_{\hat{\varepsilon}_0}| \leq t_{\frac{\alpha}{2}}.$$

• L'hypothèse H_0 est rejetée au risque de première espèce α (incompatibilité du modèle linéaire avec a) si :

$$\hat{Y}_0 \notin \left[a \pm t_{\frac{\alpha}{2}} \hat{\sigma}^* \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{n s_x^2}} \right] \quad \text{ou} \quad |T_{\hat{\varepsilon}_0}| > t_{\frac{\alpha}{2}}.$$

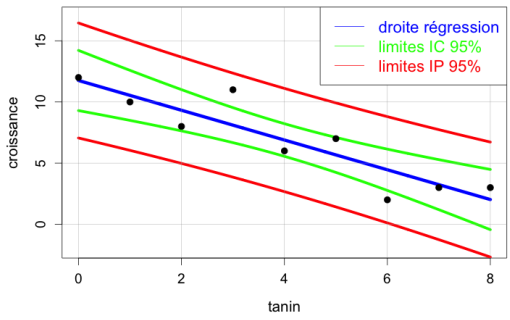
Exemple 16

Application au jeu de données : construction de l'intervalle de prévision d'une nouvelle valeur de Y pour une valeur de x, disons $x_0 = 4,5$, pour un niveau de confiance de $1 - \alpha = 95\%$.

$$\begin{cases} \hat{y}_0 = 11,75 - 1,22 \cdot 4,5 = 6,26 \quad , \quad t_{\frac{\alpha}{2}} = t_{0,025} \approx 2,36 \\ \hat{\sigma}^* = 1,69 \end{cases}$$

$$Ic_{0,95}(y_0) = \left[\hat{y}_0 \pm t_{\frac{\alpha}{2}} \hat{\sigma}^* \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{n s_x^2}} \right] = [6,26 \pm 4,21] = [2,05 \ ; \ 10,47] \, .$$

En faisant le même calcul pour toutes les valeurs x_0 appartenant à l'intervalle $[0 \ ; \ 8]$, on obtient une bande de prévision autour de la droite de régression, bande qui se rajoute à celle de confiance. Les deux bandes, ainsi que la droite de régression sont représentées ci-dessous.



6 Exercices

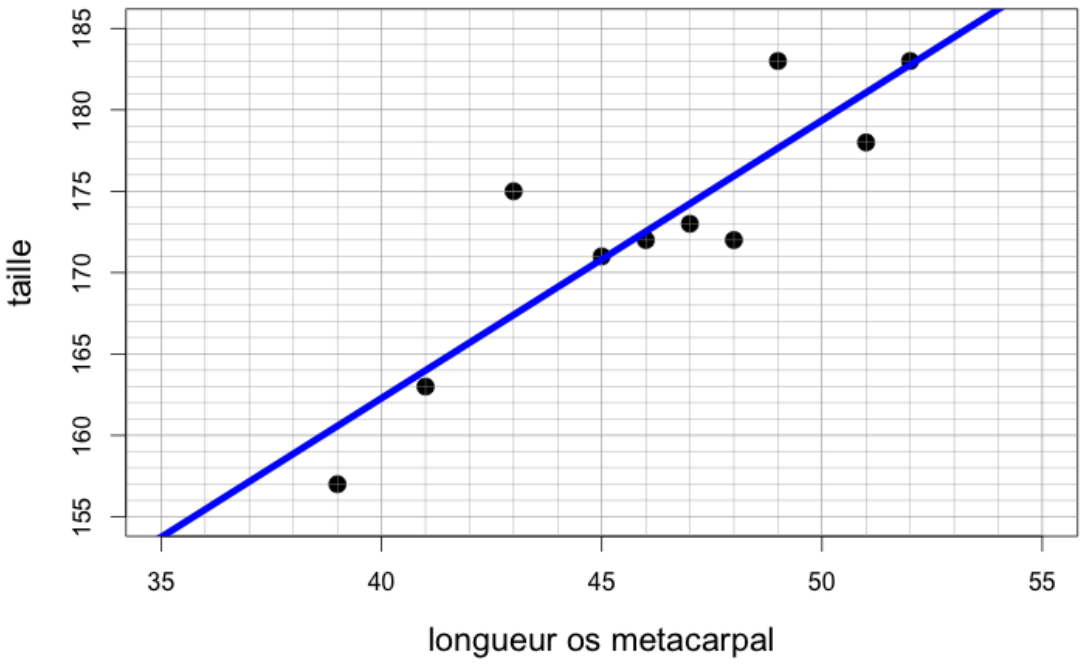
Exercice 1

Quand des anthropologues étudient des ossements humains, l'un des points importants est de déterminer la taille des individus. Comme les squelettes sont souvent incomplets,

on estime cette taille à partir de mesures sur des petits os. Dans un article intitulé *The Estimation of Adult Stature from Metacarpal Bone Length*, une équipe de chercheurs a ainsi présenté une méthode permettant d'estimer la taille d'un individu en fonction de la longueur des métacarpes, les os de la paume de main, validée sur les données suivantes où x est la longueur de l'os metacarpal du pouce et y la taille de l'individu.

x (mm)	45	51	39	41	52	48	49	46	43	47
y (cm)	171	178	157	163	183	172	183	172	175	173

On a représenté ci-après les données et la droite des moindres carrés reliant y à x.



1. Calculer les coefficients de la droite des moindres carrés. Vérifiez avec le graphique.

2. Pour quel risque minimal, peut-on considérer que la relation entre x et y est significative ?

3. Donner l'intervalle de confiance à 95% de la hauteur moyenne des individus dont l'os metacarpal du pouce serait long de 50 mm.

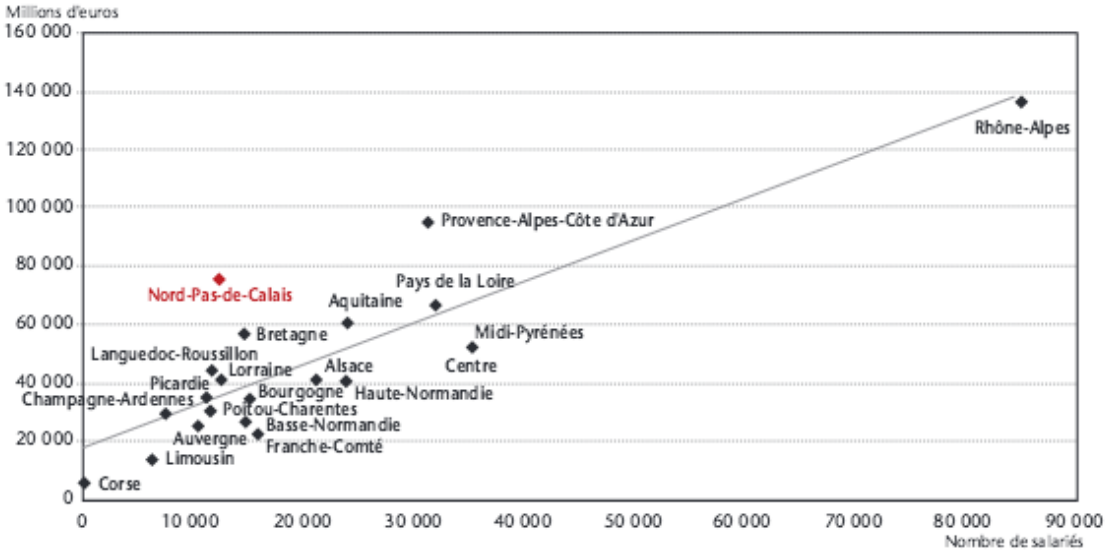
4. Des éléments anthropologiques complémentaires ont permis d'estimer à 1m90 la taille d'un individu dont l'os metacarpal du pouce est de 50 mm. Que penser de cet individu ?

5. Tracer les résidus. Qu'est-ce qu'il faut faire pour vérifier s'il s'agit de réalisations de variables aléatoires normales ?

Exercice 2

La figure suivante indique, pour les 21 régions françaises de province et de métropole (en vigueur jusqu'en 2015), le PIB (y) par région en fonction du nombre d'emplois (x) dans la haute technologie, pour l'année 2000 (source : INSEE Nord-Pas-de-Calais). Le

nuage de points, de forme allongée, suggère l'existence d'une relation linéaire (figurée par la droite des moindres carrées) entre ces deux variables.



On donne par ailleurs les résultats intermédiaires suivants :

$\sum x_i$	$\sum y_i$	$\sum x_i^2$	$\sum y_i^2$	$\sum x_i y_i$
431 200	992 600	15 078 020 000	64 038 160 000	29 144 300 000

1. Calculer les coefficients $\hat{\beta}_0$ et $\hat{\beta}_1$, estimations des paramètres β_0 et β_1 de la relation linéaire $\beta_0 + \beta_1 x$ qu'on cherche à mettre en évidence.

2. La relation obtenue est elle significative au risque 5% ?

3. Pour 12 000 emplois de haute technologie, quelle est l'espérance mathématique du PIB et son intervalle de confiance à 95 % ?

4. Dans cette étude, la région Nord-Pas-de-Calais (cliente de l'étude) affiche un PIB de 76 Milliards d'euros pour environ 12 000 emplois de haute technologie. Que pensez de cette région par rapport aux autres ?

5. La région Nord-Pas-de-Calais ainsi que la région Provence-Alpes-Côte d'Azur sont en effet assez éloignées du modèle obtenu. Selon vous, quelles raisons structurelles propres à ces régions pourraient expliquer cet écart ?

6. Quel défaut présente le modèle de régression choisi ici et comment aurait-on pu le corriger ?

Exercice 3

Les données ci-dessous sont relatives à l'étalonnage d'une méthode gravimétrique pour le dosage de la chaux en présence de magnésium. La variable en x est la teneur vraie et la variable en y est la teneur mesurée (en mg).

Vraie (x)	20	22,5	25	28,5	31	35,5	33,5	37	38	40
Mesurée (y)	19,8	22,8	24,5	27,3	31	35	35,1	37,1	38,5	39

1. Estimer par la méthode des moindres carrés les paramètres β_0 et β_1 de la relation linéaire $\beta_0 + \beta_1 x$ qu'on cherche à mettre en évidence.

2. Caractériser la précision de la méthode gravimétrique.

3. Tester l'hypothèse $\beta_0 = 0$ de telle façon que la probabilité d'accepter l'hypothèse si elle est vraie soit égale à 90%.

4. Tester l'hypothèse $\beta_1 = 1$ de telle façon que la probabilité d'accepter l'hypothèse si elle est vraie soit égale à 90%.

5. Bâtir et mettre en oeuvre un test permettant de tester simultanément que $\beta_0 = 0$ et que $\beta_1 = 1$, la probabilité d'accepter l'hypothèse si elle est vraie étant encore égale à 90%.

$\sum x_i = 311$ $\quad \sum y_i = 310,1$
$\sum x_i^2 =$10 100 $\quad \sum y_i^2 =$10 055,09 $\quad \sum x_i y_i =$10 074,8

Exercice 4

Le tableau ci-après donne les résultats d'un certain nombre de déterminations de la distance nécessaire (y en mètres) à l'arrêt par freinage d'une automobile lancée à différentes vitesses (x en km/h). Une étude graphique montre que la courbe représentant y en fonction de x est manifestement concave vers les y positifs, mais que si l'on utilise x^2 au lieu de x, la liaison apparaît sensiblement linéaire. Peut-on justifier ce fait par une loi physique ? Admettant la validité de ce type de liaison entre y et x^2, on suppose de plus que la vitesse x peut être déterminée avec une grande précision et que les écarts constatés sont dus à des fluctuations aléatoires de y autour d'une vraie valeur correspondant à une liaison linéaire représentée par l'équation $y = \beta_1 x^2 + \beta_0$.

Vitesse (x)	33	49	65	33	79	49	93
Distance (y)	5,3	14,45	20,26	6,5	38,45	11,23	50,42
x^2	1 089	2 401	4 225	1 089	6 241	2 401	8 649

$\sum y_i = 146,61$ $\quad \sum x_i^2 = 26,095$
$\sum y_i^2 =$4 836,3019 $\quad \sum x_i^4 =$145 507 351 $\quad \sum x_i^2 y_i =$836 155, 41

1. Quelle est la meilleure estimation de β_0 et β_1 ? Quelle hypothèse supplémentaire suppose cette estimation ?

2. Déterminer les limites de confiance à 95% pour les estimations précédentes.

3. Considérant le cas d'une voiture dont la vitesse est de 85 km/h, estimer la valeur moyenne correspondante de y. En donner une limite supérieure au seuil de confiance 99%.

4. On suppose que pour une voiture se déplaçant à 85 km/h, on observe une distance de freinage $y = 55$ mètres. Cette valeur peut-elle être considérée comme étant, à des fluctuations aléatoires admissibles près, d'accord avec l'équation d'estimation trouvée ?

Exercice 5

Il y a des situations où la droite de régression passe par l'origine. Le modèle devient alors $Y_i = \beta_1 x_i + \varepsilon_i$.

1. En utilisant la méthode des moindres carrés, donner les expressions de :

(a) $\hat{\beta}_1$,

(b) $\mathbb{E}\left(\hat{\beta}_1\right)$, $\mathbb{V}\left(\hat{\beta}_1\right)$, $\mathbb{V}\left(\hat{Y}_i\right)$.

2. Montrer algébriquement que $\sum \hat{\varepsilon}_i \neq 0$.