Université Rennes 2
Master de Statistique

Année 2012/2013
Premier Semestre

Régression linéaire

Arnaud Guyader

Ce cours est tiré des quatre premiers chapitres du livre de Pierre-André Cornillon et Eric Matzner-Løber, *Régression avec R*, paru chez Springer en 2010.

Table des matières

Chapitre 1

La régression linéaire simple

Introduction

Commençons par un exemple afin de fixer les idées. Pour des raisons de santé publique, on s'intéresse à la concentration d'ozone O_3 dans l'air (en microgrammes par millilitre). En particulier, on cherche à savoir s'il est possible d'expliquer le taux maximal d'ozone de la journée par la température T_{12} à midi. Les données sont :

Température à 12h	23.8	16.3	27.2	7.1	25.1	27.5	19.4	19.8	32.2	20.7
O_3 max	115.4	76.8	113.8	81.6	115.4	125	83.6	75.2	136.8	102.8

TABLE 1.1 – 10 données journalières de température et d'ozone.

D'un point de vue pratique, le but de cette régression est double :
- ajuster un modèle pour expliquer O_3 en fonction de T_{12} ;
- prédire les valeurs d'O_3 pour de nouvelles valeurs de T_{12}.

Avant toute analyse, il est intéressant de représenter les données, comme sur la figure 1.1.

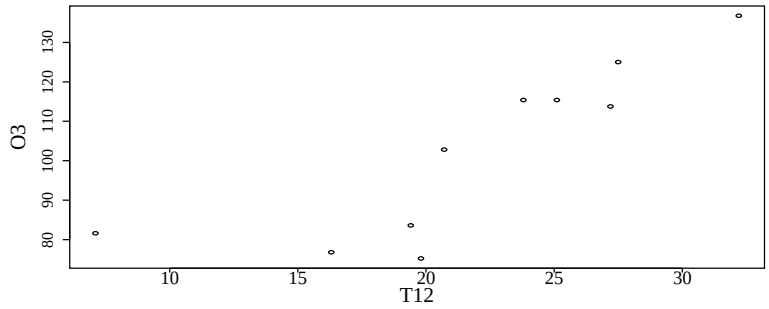


FIGURE 1.1 – 10 données journalières de température et d'ozone.

Pour analyser la relation entre les x_i (température) et les y_i (ozone), nous allons chercher une fonction f telle que :

$$y_i \approx f(x_i).$$

Pour préciser le sens de $\approx$, il faut se donner un critère quantifiant la qualité de l'ajustement de la fonction f aux données. Il conviendra aussi de se donner une classe de fonctions $\mathcal{F}$ dans laquelle est supposée vivre la vraie fonction inconnue.

Le problème mathématique peut alors s'écrire de la façon suivante :

$$\arg\min_{f\in\mathcal{F}} \sum_{i=1}^{n} L(y_i - f(x_i)),$$

où n représente le nombre de données disponibles (taille de l'échantillon) et $L(.)$ est appelée fonction de coût ou fonction de perte (*Loss* en anglais).

1.1 Modélisation

Dans de nombreuses situations, en première approche, une idée naturelle est de supposer que la variable à expliquer y est une fonction affine de la variable explicative x, c'est-à-dire de chercher f dans l'ensemble $\mathcal{F}$ des fonctions affines de $\mathbb{R}$ dans $\mathbb{R}$. C'est le principe de la régression linéaire simple. On suppose dans la suite disposer d'un échantillon de n points (x_i, y_i) du plan.

Définition 1 (Modèle de régression linéaire simple) *Un modèle de régression linéaire simple est défini par une équation de la forme :*

$$\forall i \in \{1,\dots,n\} \qquad y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$$

Les quantités ε_i viennent du fait que les points ne sont jamais parfaitement alignés sur une droite. On les appelle les erreurs (ou bruits) et elles sont supposées aléatoires. Pour pouvoir dire des choses pertinentes sur ce modèle, il faut néanmoins imposer des hypothèses les concernant. Voici celles que nous ferons dans un premier temps :

$$(\mathcal{H}) \begin{cases} (\mathcal{H}_1) : \mathbb{E}[\varepsilon_i] = 0 \text{ pour tout indice } i \\ (\mathcal{H}_2) : \mathrm{Cov}(\varepsilon_i, \varepsilon_j) = \delta_{ij}\sigma^2 \text{ pour tout couple } (i,j) \end{cases}$$

Les erreurs sont donc supposées centrées, de même variance (homoscédasticité) et non corrélées entre elles (δ_{ij} est le symbole de Kronecker, i.e. $\delta_{ij} = 1$ si $i = j$, $\delta_{ij} = 0$ si $i \neq j$). Notons que le modèle de régression linéaire simple de la définition 1 peut encore s'écrire de façon vectorielle :

$$Y = \beta_1 \mathbb{1} + \beta_2 X + \boldsymbol{\varepsilon},$$

où :

- le vecteur $Y = [y_1, \dots, y_n]'$ est aléatoire de dimension n,
- le vecteur $\mathbb{1} = [1, \dots, 1]'$ est le vecteur de $\mathbb{R}^n$ dont les n composantes valent toutes 1,
- le vecteur $X = [x_1, \dots, x_n]'$ est un vecteur de dimension n donné (non aléatoire),
- les coefficients β_1 et β_2 sont les paramètres inconnus (mais non aléatoires !) du modèle,
- le vecteur $\boldsymbol{\varepsilon} = [\varepsilon_1, \dots, \varepsilon_n]'$ est aléatoire de dimension n.

Cette notation vectorielle sera commode notamment pour l'interprétation géométrique du problème. Nous y reviendrons en Section 1.3 et elle sera d'usage constant en régression linéaire multiple, c'est pourquoi il convient d'ores et déjà de s'y habituer.

1.2 Moindres Carrés Ordinaires

Les points (x_i, y_i) étant donnés, le but est maintenant de trouver une fonction affine f telle que la quantité $\sum_{i=1}^{n} L(y_i - f(x_i))$ soit minimale. Pour pouvoir déterminer f, encore faut-il préciser la fonction de coût L. Deux fonctions sont classiquement utilisées :

- le coût absolu $L(u) = |u|$;

— le coût quadratique $L(u) = u^2$.

Les deux ont leurs vertus, mais on privilégiera dans la suite la fonction de coût quadratique. On parle alors de méthode d'estimation par moindres carrés (terminologie due à Legendre dans un article de 1805 sur la détermination des orbites des comètes).

Définition 2 (Estimateurs des Moindres Carrés Ordinaires) *On appelle estimateurs des Moindres Carrés Ordinaires (en abrégé MCO) $\hat{\beta}_1$ et $\hat{\beta}_2$ les valeurs minimisant la quantité :*

$$S(\beta_1, \beta_2) = \sum_{i=1}^{n} (y_i - \beta_1 - \beta_2 x_i)^2.$$

Autrement dit, la droite des moindres carrés minimise la somme des carrés des distances verticales des points (x_i, y_i) du nuage à la droite ajustée $y = \hat{\beta}_1 + \hat{\beta}_2 x$.

1.2.1 Calcul des estimateurs de β_1 et β_2

La fonction de deux variables S est une fonction quadratique et sa minimisation ne pose aucun problème, comme nous allons le voir maintenant.

Proposition 1 (Estimateurs $\hat{\beta}_1$ et $\hat{\beta}_2$) *Les estimateurs des MCO ont pour expressions :*

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x},$$

avec :

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Preuves. La première méthode consiste à remarquer que la fonction $S(\beta_1, \beta_2)$ est strictement convexe, donc qu'elle admet un minimum en un unique point $(\hat{\beta}_1, \hat{\beta}_2)$, lequel est déterminé en annulant les dérivées partielles de S. On obtient les "équations normales" :

$$\begin{cases} \dfrac{\partial S}{\partial \beta_1} &= -2\displaystyle\sum_{i=1}^{n}(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) = 0 \\ \dfrac{\partial S}{\partial \beta_2} &= -2\displaystyle\sum_{i=1}^{n} x_i(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) = 0 \end{cases}$$

La première équation donne :

$$\hat{\beta}_1 n + \hat{\beta}_2 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

d'où l'on déduit immédiatement :

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x}, \tag{1.1}$$

où $\bar{x}$ et $\bar{y}$ sont comme d'habitude les moyennes empiriques des x_i et des y_i. La seconde équation donne :

$$\hat{\beta}_1 \sum_{i=1}^{n} x_i + \hat{\beta}_2 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

et en remplaçant $\hat{\beta}_1$ par son expression (1.1), nous avons :

$$\hat{\beta}_2 = \frac{\sum x_i y_i - \sum x_i \bar{y}}{\sum x_i^2 - \sum x_i \bar{x}} = \frac{\sum x_i (y_i - \bar{y})}{\sum x_i (x_i - \bar{x})} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})(x_i - \bar{x})}. \tag{1.2}$$

La seconde méthode consiste à appliquer la technique de Gauss de réduction des formes quadratiques, c'est-à-dire à décomposer $S(\beta_1, \beta_2)$ en somme de carrés, carrés qu'il ne restera plus qu'à annuler pour obtenir les estimateurs $\hat{\beta}_1$ et $\hat{\beta}_2$. Dans notre cas, après calculs, ceci s'écrit :

$$\begin{aligned} S(\beta_1, \beta_2) =& n \left(\beta_1 - (\bar{y} - \beta_2 \bar{x})\right)^2 + \left(\sum_{i=1}^n (x_i - \bar{x})^2\right) \left(\beta_2 - \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)^2 \\ &+ \left(\sum_{i=1}^n (y_i - \bar{y})^2\right) \left(1 - \frac{\left(\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})\right)^2}{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}\right), \end{aligned}$$

où apparaissent deux carrés et un troisième terme indépendant de β_1 et β_2 : ce dernier est donc incompressible. Par contre, le second est nul si et seulement si $\beta_2 = \hat{\beta}_2$. Ceci étant fait, le premier est alors nul si et seulement si $\beta_1 = \hat{\beta}_1$.

■

L'expression (1.2) de $\hat{\beta}_2$ suppose que le dénominateur $\sum_{i=1}^n (x_i - \bar{x})^2$ est non nul. Or ceci ne peut arriver que si tous les x_i sont égaux, situation sans intérêt pour notre problème et que nous excluons donc a priori dans toute la suite.

Remarques :

1. La relation $\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x}$ montre que la droite des MCO passe par le centre de gravité du nuage $(\bar{x}, \bar{y})$.
2. Les expressions obtenues pour $\hat{\beta}_1$ et $\hat{\beta}_2$ montrent que ces deux estimateurs sont linéaires par rapport au vecteur $Y = [y_1, \ldots, y_n]'$.
3. L'estimateur $\hat{\beta}_2$ peut aussi s'écrire comme suit (exercice !) :

$$\hat{\beta}_2 = \beta_2 + \frac{\sum (x_i - \bar{x}) \varepsilon_i}{\sum (x_i - \bar{x})^2}. \tag{1.3}$$

   Si cette décomposition n'est pas intéressante pour le calcul effectif de $\hat{\beta}_2$ puisqu'elle fait intervenir les quantités inconnues β_2 et ε_i, elle l'est par contre pour démontrer des propriétés théoriques des estimateurs (biais et variance). Son avantage est en effet de mettre en exergue la seule source d'aléa du modèle, à savoir les erreurs ε_i.

Avant de poursuivre, notons que le calcul des estimateurs des moindres carrés est purement déterministe : il ne fait en rien appel aux hypothèses $(\mathcal{H}_1)$ et $(\mathcal{H}_2)$ sur le modèle. Celles-ci vont en fait servir dans la suite à expliciter les propriétés statistiques de ces estimateurs.

1.2.2 Quelques propriétés des estimateurs $\hat{\beta}_1$ et $\hat{\beta}_2$

Sous les seules hypothèses $(\mathcal{H}_1)$ et $(\mathcal{H}_2)$ de centrages, décorrélations et homoscédasticités des erreurs ε_i du modèle, on peut déjà donner certaines propriétés des estimateurs $\hat{\beta}_1$ et $\hat{\beta}_2$ des moindres carrés.

Théorème 1 (Estimateurs sans biais) *$\hat{\beta}_1$ et $\hat{\beta}_2$ sont des estimateurs sans biais de β_1 et β_2.*

Preuve. Partons de l'écriture (1.3) pour $\hat{\beta}_2$:

$$\hat{\beta}_2 = \beta_2 + \frac{\sum(x_i - \bar{x})\varepsilon_i}{\sum(x_i - \bar{x})^2}.$$

Dans cette expression, seuls les bruits ε_i sont aléatoires, et puisqu'ils sont centrés, on en déduit bien que $\mathbb{E}[\hat{\beta}_2] = \beta_2$. Pour $\hat{\beta}_1$, on part de l'expression :

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2\bar{x},$$

d'où l'on tire :

$$\mathbb{E}[\hat{\beta}_1] = \mathbb{E}[\bar{y}] - \bar{x}\mathbb{E}[\hat{\beta}_2] = \beta_1 + \bar{x}\beta_2 - \bar{x}\beta_2 = \beta_1.$$

■

On peut également exprimer variances et covariance de nos estimateurs.

Théorème 2 (Variances et covariance) *Les variances des estimateurs sont :*

$$\mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2\sum x_i^2}{n\sum(x_i - \bar{x})^2} = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}\right) \qquad \& \qquad \mathrm{Var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum(x_i - \bar{x})^2},$$

tandis que leur covariance vaut :

$$\mathrm{Cov}(\hat{\beta}_1, \hat{\beta}_2) = -\frac{\sigma^2\bar{x}}{\sum(x_i - \bar{x})^2}.$$

Preuve. On part à nouveau de l'expression de $\hat{\beta}_2$ utilisée dans la preuve du non-biais :

$$\hat{\beta}_2 = \beta_2 + \frac{\sum(x_i - \bar{x})\varepsilon_i}{\sum(x_i - \bar{x})^2},$$

or les erreurs ε_i sont décorrélées et de même variance σ^2 donc la variance de la somme est la somme des variances :

$$\mathrm{Var}(\hat{\beta}_2) = \frac{\sum(x_i - \bar{x})^2\sigma^2}{(\sum(x_i - \bar{x})^2)^2} = \frac{\sigma^2}{\sum(x_i - \bar{x})^2}.$$

Par ailleurs, la covariance entre $\bar{y}$ et $\hat{\beta}_2$ s'écrit :

$$\mathrm{Cov}(\bar{y}, \hat{\beta}_2) = \mathrm{Cov}\left(\frac{\sum y_i}{n}, \frac{\sum(x_i - \bar{x})\varepsilon_i}{\sum(x_i - \bar{x})^2}\right) = \frac{\sigma^2\sum(x_i - \bar{x})}{n\sum(x_i - \bar{x})^2} = 0,$$

d'où il vient pour la variance de $\hat{\beta}_1$:

$$\mathrm{Var}(\hat{\beta}_1) = \mathrm{Var}\left(\frac{\sum y_i}{n} - \hat{\beta}_2\bar{x}\right) = \frac{\sigma^2}{n} + \frac{\bar{x}^2\sigma^2}{\sum(x_i - \bar{x})^2} - 2\bar{x}\mathrm{Cov}(\bar{y}, \hat{\beta}_2),$$

c'est-à-dire :

$$\mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{n} + \frac{\bar{x}^2\sigma^2}{\sum(x_i - \bar{x})^2} = \frac{\sigma^2\sum x_i^2}{n\sum(x_i - \bar{x})^2}.$$

Enfin, pour la covariance des deux estimateurs :

$$\mathrm{Cov}(\hat{\beta}_1, \hat{\beta}_2) = \mathrm{Cov}(\bar{y} - \hat{\beta}_2\bar{x}, \hat{\beta}_2) = \mathrm{Cov}(\bar{y}, \hat{\beta}_2) - \bar{x}\mathrm{Var}(\hat{\beta}_2) = -\frac{\sigma^2\bar{x}}{\sum(x_i - \bar{x})^2}.$$

■

Remarques :

1. On a vu que la droite des MCO passe par le centre de gravité du nuage $(\bar{x}, \bar{y})$. Supposons celui-ci fixé et $\bar{x}$ positif, alors il est clair que si on augmente la pente, l'ordonnée à l'origine va baisser et vice versa, on retrouve donc bien le signe négatif pour la covariance entre $\hat{\beta}_1$ et $\hat{\beta}_2$.
2. En statistique inférentielle, la variance d'un estimateur décroît typiquement de façon inversement proportionnelle à la taille de l'échantillon, c'est-à-dire en $1/n$. En d'autres termes, sa précision est généralement en $1/\sqrt{n}$. Ceci ne saute pas aux yeux si l'on considère par exemple l'expression obtenue pour la variance de β_2 :

$$\mathrm{Var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum(x_i - \bar{x})^2}.$$

Pour comprendre que tout se passe comme d'habitude, il suffit de considérer que les x_i sont eux-mêmes aléatoires, avec écart-type σ_x. Dans ce cas très général, le dénominateur est d'ordre $n\sigma_x^2$ et l'on retrouve bien une variance en $1/n$.

Les estimateurs des moindres carrés sont en fait optimaux en un certain sens, c'est ce que précise le résultat suivant.

Théorème 3 (Gauss-Markov) *Parmi les estimateurs sans biais linéaires en y, les estimateurs $\hat{\beta}_j$ sont de variances minimales.*

Preuve. L'estimateur des MCO s'écrit $\hat{\beta}_2 = \sum_{i=1}^n p_i y_i$, avec $p_i = (x_i - \bar{x})/\sum(x_i - \bar{x})^2$. Considérons un autre estimateur $\tilde{\beta}_2$ linéaire en y_i et sans biais, c'est-à-dire :

$$\tilde{\beta}_2 = \sum_{i=1}^n \lambda_i y_i.$$

Montrons que $\sum \lambda_i = 0$ et $\sum \lambda_i x_i = 1$. L'égalité

$$\mathbb{E}(\tilde{\beta}_2) = \beta_1 \sum \lambda_i + \beta_2 \sum \lambda_i x_i + \sum \lambda_i \mathbb{E}(\varepsilon_i) = \beta_1 \sum \lambda_i + \beta_2 \sum \lambda_i x_i$$

est vraie pour tout β_2. L'estimateur $\tilde{\beta}_2$ est sans biais donc $\mathbb{E}(\tilde{\beta}_2) = \beta_2$ pour tout β_2, c'est-à-dire que $\sum \lambda_i = 0$ et $\sum \lambda_i x_i = 1$. Montrons que $\mathrm{Var}(\tilde{\beta}_2) \geq \mathrm{Var}(\hat{\beta}_2)$:

$$\mathrm{Var}(\tilde{\beta}_2) = \mathrm{Var}(\tilde{\beta}_2 - \hat{\beta}_2 + \hat{\beta}_2) = \mathrm{Var}(\tilde{\beta}_2 - \hat{\beta}_2) + \mathrm{Var}(\hat{\beta}_2) + 2\mathrm{Cov}(\tilde{\beta}_2 - \hat{\beta}_2, \hat{\beta}_2).$$

Or :

$$\mathrm{Cov}(\tilde{\beta}_2 - \hat{\beta}_2, \hat{\beta}_2) = \mathrm{Cov}(\tilde{\beta}_2, \hat{\beta}_2) - \mathrm{Var}(\hat{\beta}_2) = \frac{\sigma^2 \sum \lambda_i (x_i - \bar{x})}{\sum(x_i - \bar{x})^2} - \frac{\sigma^2}{\sum(x_i - \bar{x})^2} = 0,$$

la dernière égalité étant due aux deux relations $\sum \lambda_i = 0$ et $\sum \lambda_i x_i = 1$. Ainsi :

$$\mathrm{Var}(\tilde{\beta}_2) = \mathrm{Var}(\tilde{\beta}_2 - \hat{\beta}_2) + \mathrm{Var}(\hat{\beta}_2).$$

Une variance est toujours positive, donc :

$$\mathrm{Var}(\tilde{\beta}_2) \geq \mathrm{Var}(\hat{\beta}_2).$$

Le résultat est démontré. On obtiendrait la même chose pour $\hat{\beta}_1$.

■

Remarque. Comme nous le verrons au chapitre suivant, on peut en fait montrer un peu mieux : au sens de la relation d'ordre sur les matrices symétriques réelles, la matrice de covariance de $\hat{\beta} = [\hat{\beta}_1, \hat{\beta}_2]'$ est inférieure à celle de n'importe quel autre estimateur $\tilde{\beta} = [\tilde{\beta}_1, \tilde{\beta}_2]'$ sans biais et linéaire en y.

1.2.3 Calcul des résidus et de la variance résiduelle

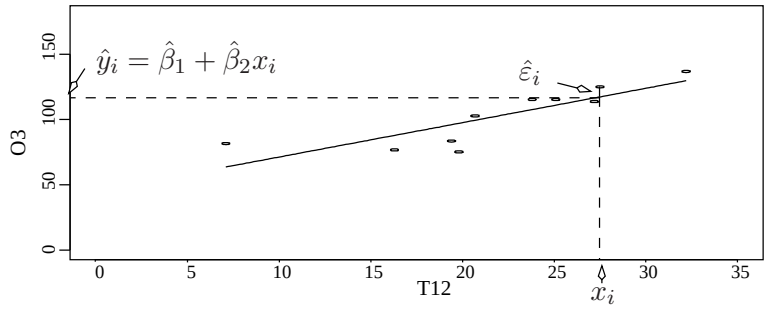


FIGURE 1.2 – Représentation des individus.

Dans $\mathbb{R}^2$ (espace des variables x_i et y_i), $\hat{\beta}_1$ est l'ordonnée à l'origine et $\hat{\beta}_2$ la pente de la droite ajustée. Cette droite minimise la somme des carrés des distances verticales des points du nuage à la droite ajustée. Notons $\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i$ l'ordonnée du point de la droite des moindres carrés d'abscisse x_i, ou valeur ajustée. les résidus sont définis par (cf. figure 1.2) :

$$\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i = (y_i - \bar{y}) - \hat{\beta}_2(x_i - \bar{x}). \tag{1.4}$$

Par construction, la somme des résidus est nulle :

$$\sum_i \hat{\varepsilon}_i == \sum_i (y_i - \bar{y}) - \hat{\beta}_2 \sum_i (x_i - \bar{x}) = 0.$$

Notons maintenant que les variances et covariance des estimateurs $\hat{\beta}_1$ et $\hat{\beta}_2$ établies en section précédente ne sont pas pratiques car elles font intervenir la variance σ^2 des erreurs, laquelle est en général inconnue. Néanmoins, on peut en donner un estimateur sans biais grâce aux résidus.

Théorème 4 (Estimateur non biaisé de σ^2) *La statistique $\hat{\sigma}^2 = \sum_{i=1}^n \hat{\varepsilon}_i^2/(n-2)$ est un estimateur sans biais de σ^2.*

Preuve. Réécrivons les résidus en constatant que $\hat{\beta}_1 = \bar{y} - \hat{\beta}_2\bar{x}$ et $\beta_1 = \bar{y} - \beta_2\bar{x} - \bar{\varepsilon}$, ce qui donne :

$$\begin{aligned}
\hat{\varepsilon}_i &= \beta_1 + \beta_2 x_i + \varepsilon_i - \hat{\beta}_1 - \hat{\beta}_2 x_i \\
&= \bar{y} - \beta_2\bar{x} - \bar{\varepsilon} + \beta_2 x_i + \varepsilon_i - \bar{y} + \hat{\beta}_2\bar{x} - \hat{\beta}_2 x_i \\
&= (\beta_2 - \hat{\beta}_2)(x_i - \bar{x}) + (\varepsilon_i - \bar{\varepsilon}).
\end{aligned}$$

En développant et en nous servant de l'écriture vue plus haut :

$$\hat{\beta}_2 = \beta_2 + \frac{\sum(x_i - \bar{x})\varepsilon_i}{\sum(x_i - \bar{x})^2},$$

nous avons :

$$\begin{aligned}
\sum \hat{\varepsilon}_i^2 &= (\beta_2 - \hat{\beta}_2)^2 \sum (x_i - \bar{x})^2 + \sum (\varepsilon_i - \bar{\varepsilon})^2 + 2(\beta_2 - \hat{\beta}_2) \sum (x_i - \bar{x})(\varepsilon_i - \bar{\varepsilon}) \\
&= (\beta_2 - \hat{\beta}_2)^2 \sum (x_i - \bar{x})^2 + \sum (\varepsilon_i - \bar{\varepsilon})^2 - 2(\beta_2 - \hat{\beta}_2)^2 \sum (x_i - \bar{x})^2.
\end{aligned}$$

Prenons-en l'espérance :

$$\mathbb{E}\left(\sum \hat{\varepsilon_i}^2\right) = \mathbb{E}\left(\sum (\varepsilon_i - \bar{\varepsilon})^2\right) - \sum (x_i - \bar{x})^2 \mathrm{Var}(\hat{\beta}_2) = (n-2)\sigma^2.$$

■

Bien sûr, lorsque n est grand, cet estimateur diffère très peu de l'estimateur empirique de la variance des résidus, à savoir $\sum_{i=1}^n \hat{\varepsilon}_i^2/n$.

1.2.4 Prévision

Un des buts de la régression est de faire de la prévision, c'est-à-dire de prévoir la variable à expliquer y en présence d'une nouvelle valeur de la variable explicative x. Soit donc x_{n+1} une nouvelle valeur, pour laquelle nous voulons prédire y_{n+1}. Le modèle est toujours le même :

$$y_{n+1} = \beta_1 + \beta_2 x_{n+1} + \varepsilon_{n+1}$$

avec $\mathbb{E}[\varepsilon_{n+1}] = 0$, $\text{Var}(\varepsilon_{n+1}) = \sigma^2$ et $\text{Cov}(\varepsilon_{n+1}, \varepsilon_i) = 0$ pour $i = 1, \ldots, n$. Il est naturel de prédire la valeur correspondante via le modèle ajusté :

$$\hat{y}_{n+1} = \hat{\beta}_1 + \hat{\beta}_2 x_{n+1}.$$

Deux types d'erreurs vont entacher notre prévision : la première est due à la non-connaissance de ε_{n+1}, la seconde à l'incertitude sur les estimateurs $\hat{\beta}_1$ et $\hat{\beta}_2$.

Proposition 2 (Erreur de prévision) *L'erreur de prévision $\hat{\varepsilon}_{n+1} = (y_{n+1} - \hat{y}_{n+1})$ satisfait les propriétés suivantes :*

$$\left\{ \begin{array}{l} \mathbb{E}[\hat{\varepsilon}_{n+1}] = 0 \\ \text{Var}(\hat{\varepsilon}_{n+1}) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_{n+1}-\bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right). \end{array} \right.$$

Preuve. Pour l'espérance, il suffit d'utiliser le fait que ε_{n+1} est centrée et que les estimateurs $\hat{\beta}_1$ et $\hat{\beta}_2$ sont sans biais :

$$\mathbb{E}[\hat{\varepsilon}_{n+1}] = \mathbb{E}[\beta_1 - \hat{\beta}_1] + \mathbb{E}[\beta_2 - \hat{\beta}_2] x_{n+1} + \mathbb{E}[\varepsilon_{n+1}] = 0.$$

Nous obtenons la variance de l'erreur de prévision en nous servant du fait que y_{n+1} est fonction de ε_{n+1} seulement tandis que $\hat{y}_{n+1}$ est fonction des autres erreurs $(\varepsilon_i)_{1 \leq i \leq n}$:

$$\text{Var}(\hat{\varepsilon}_{n+1}) = \text{Var}\left(y_{n+1} - \hat{y}_{n+1}\right) = \text{Var}(y_{n+1}) + \text{Var}(\hat{y}_{n+1}) = \sigma^2 + \text{Var}(\hat{y}_{n+1}).$$

Calculons le second terme :

$$\begin{aligned} \text{Var}\left(\hat{y}_{n+1}\right) &= \text{Var}\left(\hat{\beta}_1 + \hat{\beta}_2 x_{n+1}\right) = \text{Var}(\hat{\beta}_1) + x_{n+1}^2 \text{Var}(\hat{\beta}_2) + 2x_{n+1}\text{Cov}\left(\hat{\beta}_1, \hat{\beta}_2\right) \\ &= \frac{\sigma^2}{\sum (x_i - \bar{x})^2} \left(\frac{\sum x_i^2}{n} + x_{n+1}^2 - 2x_{n+1}\bar{x}\right) \\ &= \frac{\sigma^2}{\sum (x_i - \bar{x})^2} \left(\frac{\sum (x_i - \bar{x})^2}{n} + \bar{x}^2 + x_{n+1}^2 - 2x_{n+1}\bar{x}\right) \\ &= \sigma^2 \left(\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right). \end{aligned}$$

Au total, on obtient bien :

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

■

Ainsi la variance augmente lorsque x_{n+1} s'éloigne du centre de gravité du nuage. Autrement dit, faire de la prévision lorsque x_{n+1} est "loin" de $\bar{x}$ est périlleux, puisque la variance de l'erreur de prévision peut être très grande ! Ceci s'explique intuitivement par le fait que plus une observation x_{n+1} est éloignée de la moyenne $\bar{x}$ et moins on a d'information sur elle.

1.3 Interprétations géométriques

1.3.1 Représentation des variables

Si nous abordons le problème d'un point de vue vectoriel, nous avons deux vecteurs à notre disposition : le vecteur $X = [x_1, \ldots, x_n]'$ des n observations pour la variable explicative et le vecteur $Y = [y_1, \ldots, y_n]'$ des n observations pour la variable à expliquer. Ces deux vecteurs appartiennent au même espace $\mathbb{R}^n$: l'espace des variables.

Si on ajoute à cela le vecteur $\mathbb{1} = [1, \ldots, 1]'$, on voit tout d'abord que par l'hypothèse selon laquelle tous les x_i ne sont pas égaux, les vecteurs $\mathbb{1}$ et X ne sont pas colinéaires : ils engendrent donc un sous-espace de $\mathbb{R}^n$ de dimension 2, noté $\mathcal{M}(X)$. On peut projeter orthogonalement le vecteur Y sur le sous-espace $\mathcal{M}(X)$, notons provisoirement $\tilde{Y}$ ce projeté. Puisque $(\mathbb{1}, X)$ forme une base de $\mathcal{M}(X)$, il existe une unique décomposition de la forme $\tilde{Y} = \tilde{\beta}_1\mathbb{1} + \tilde{\beta}_2 X$. Par définition du projeté orthogonal, $\tilde{Y}$ est l'unique vecteur de $\mathcal{M}(X)$ minimisant la distance euclidienne $\|Y - \tilde{Y}\|$, ce qui revient au même que de minimiser son carré. Or, par définition de la norme euclidienne, cette quantité vaut :

$$\|Y - \tilde{Y}\|^2 = \sum_{i=1}^{n}(y_i - (\tilde{\beta}_1 + \tilde{\beta}_2 x_i))^2,$$

ce qui nous ramène à la méthode des moindres carrés ordinaires. On en déduit que $\tilde{\beta}_1 = \hat{\beta}_1$, $\tilde{\beta}_2 = \hat{\beta}_2$ et $\tilde{Y} = \hat{Y} = [\hat{y}_1, \ldots, \hat{y}_n]'$, avec les expressions de $\hat{\beta}_1$, $\hat{\beta}_2$ et $\hat{Y}$ vues précédemment.

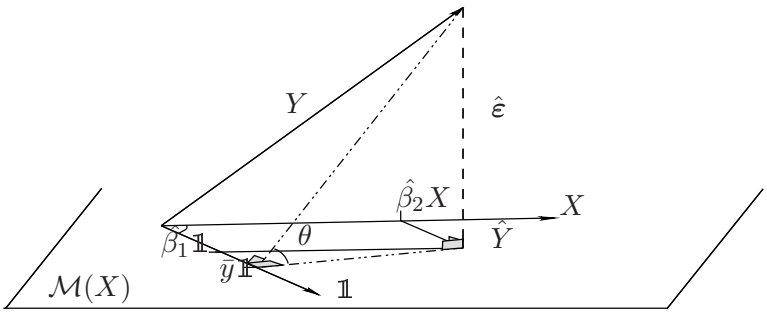


FIGURE 1.3 – Représentation de la projection dans l'espace des variables.

Autrement dit, dans $\mathbb{R}^n$, $\hat{\beta}_1$ et $\hat{\beta}_2$ s'interprètent comme les coordonnées de la projection orthogonale $\hat{Y}$ de Y sur le sous-espace de $\mathbb{R}^n$ engendré par $\mathbb{1}$ et X (voir figure 1.3).

Remarques :

1. Cette vision géométrique des choses peut sembler un peu abstraite, mais c'est en fait l'approche féconde pour comprendre la régression multiple, comme nous le verrons dans les chapitres suivants.
2. Nous avons supposé que $\mathbb{1}$ et X ne sont pas colinéaires. En général, ces vecteurs ne sont pas orthogonaux (sauf si $\bar{x} = 0$), ce qui implique que $\hat{\beta}_1\mathbb{1}$ n'est pas la projection orthogonale de Y sur $\mathbb{1}$ (laquelle vaut $\bar{y}\mathbb{1}$), et que $\hat{\beta}_2 X$ n'est pas la projection orthogonale de Y sur X (laquelle vaut $\frac{\langle Y,X\rangle}{\|X\|^2}X$).

1.3.2 Le coefficient de détermination R^2

Nous conservons les notations du paragraphe précédent, avec $\hat{Y} = [\hat{y}_1, \ldots, \hat{y}_n]'$ la projection orthogonale du vecteur Y sur $\mathcal{M}(X)$ et

$$\hat{\boldsymbol{\varepsilon}} = Y - \hat{Y} = [\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n]'$$

le vecteur des résidus déjà rencontrés en section 1.2.3. Le théorème de Pythagore donne alors directement :

$$\begin{aligned} \|Y - \bar{y}\mathbb{1}\|^2 &= \|\hat{Y} - \bar{y}\mathbb{1}\|^2 + \|\hat{\boldsymbol{\varepsilon}}\|^2 \\ \sum_{i=1}^{n}(y_i - \bar{y})^2 &= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}\hat{\varepsilon}_i^2 \\ SCT &= SCE + SCR, \end{aligned}$$

où SCT (respectivement SCE et SCR) représente la somme des carrés totale (respectivement expliquée par le modèle et résiduelle). Ceci peut se voir comme une formule typique de décomposition de la variance. Elle permet en outre d'introduire le coefficient de détermination de façon naturelle.

Définition 3 (Coefficient de détermination R^2) *Le coefficient de détermination R^2 est défini par :*

$$R^2 = \frac{SCE}{SCT} = \frac{\|\hat{Y} - \bar{y}\mathbb{1}\|^2}{\|Y - \bar{y}\mathbb{1}\|^2} = 1 - \frac{\|\hat{\boldsymbol{\varepsilon}}\|^2}{\|Y - \bar{y}\mathbb{1}\|^2} = 1 - \frac{SCR}{SCT}.$$

On voit sur la figure 1.3 que R^2 correspond au cosinus carré de l'angle θ. De façon schématique, on peut différencier les cas suivants :

- Si $R^2 = 1$, le modèle explique tout, l'angle θ vaut zéro et Y est dans $\mathcal{M}(X)$, c'est-à-dire que $y_i = \beta_1 + \beta_2 x_i$ pour tout i : les points de l'échantillon sont parfaitement alignés sur la droite des moindres carrés ;
- Si $R^2 = 0$, cela veut dire que $\sum(\hat{y}_i - \bar{y})^2 = 0$, donc $\hat{y}_i = \bar{y}$ pour tout i. Le modèle de régression linéaire est inadapté puisqu'on ne modélise rien de mieux que la moyenne ;
- Si R^2 est proche de zéro, cela veut dire que Y est quasiment dans l'orthogonal de $\mathcal{M}(X)$, le modèle de régression linéaire est inadapté, la variable x n'explique pas bien la variable réponse y (du moins pas de façon affine).

De façon générale, l'interprétation est la suivante : le modèle de régression linéaire permet d'expliquer $100 \times R^2\%$ de la variance totale des données.

Remarques :

1. On peut aussi voir R^2 comme le carré du coefficient de corrélation empirique entre les x_i et les y_i (cf. exercice 1.2) :

$$R^2 = \left(\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \right)^2 = \rho_{X,Y}^2.$$

2. Sur la figure 1.3 est noté un angle droit entre les vecteurs $\mathbb{1}$ et $\hat{Y} - \bar{y}\mathbb{1}$. On vérifie en effet facilement que ces deux vecteurs sont orthogonaux puisque $\bar{y}\mathbb{1}$ n'est rien d'autre que le projeté orthogonal de Y sur (la droite vectorielle engendrée par) le vecteur $\mathbb{1}$ (exercice).

1.4 Cas d'erreurs gaussiennes

Mieux que les expressions des estimateurs et celles de leurs variances, on aimerait connaître leurs lois : ceci permettrait par exemple d'obtenir des régions de confiance et d'effectuer des tests d'hypothèses. Dans cette optique, il faut bien entendu faire une hypothèse plus forte sur notre modèle, à savoir préciser la loi des erreurs. Nous supposerons ici que les erreurs sont gaussiennes. Les hypothèses $(\mathcal{H}_1)$ et $(\mathcal{H}_2)$ deviennent dès lors :

$$(\mathcal{H}) \begin{cases} (\mathcal{H}_1) : \varepsilon_i \sim \mathcal{N}(0, \sigma^2) \\ (\mathcal{H}_2) : \varepsilon_i \text{ mutuellement indépendants} \end{cases}$$

Le modèle de régression simple devient un modèle paramétrique, où les paramètres $(\beta_1, \beta_2, \sigma^2)$ sont à valeurs dans $\mathbb{R} \times \mathbb{R} \times \mathbb{R}_+^*$. La loi des ε_i étant connue, les lois des y_i s'en déduisent :

$$\forall i \in \{1, \dots, n\} \qquad y_i \sim \mathcal{N}(\beta_1 + \beta_2 x_i, \sigma^2),$$

et les y_i sont mutuellement indépendants puisque les ε_i le sont. Nous pouvons donc calculer la vraisemblance de l'échantillon et les estimateurs qui maximisent cette vraisemblance. C'est l'objet de la section suivante.

1.4.1 Estimateurs du maximum de vraisemblance

La vraisemblance vaut

$$\begin{aligned} \mathcal{L}(\beta_1, \beta_2, \sigma^2) &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \beta_1 - \beta_2 x_i)^2\right] \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left[-\frac{1}{2\sigma^2} S(\beta_1, \beta_2)\right] \end{aligned}$$

Ce qui donne pour la log-vraisemblance :

$$\log \mathcal{L}(\beta_1, \beta_2, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2} S(\beta_1, \beta_2).$$

Nous voulons maximiser cette quantité par rapport aux trois variables $(\beta_1, \beta_2, \sigma^2)$. Les deux premières variables n'apparaissent que dans le terme en $-S(\beta_1, \beta_2)$, qu'il faut donc minimiser. Or on a déjà vu que cette quantité est minimale lorsqu'on considère les estimateurs des moindres carrés, c'est-à-dire pour $\beta_1 = \hat{\beta}_1$ et $\beta_2 = \hat{\beta}_2$. Bilan : les estimateurs du maximum de vraisemblance de β_1 et β_2 sont égaux aux estimateurs des moindres carrés.

Ceci étant vu, il reste simplement à maximiser $\log \mathcal{L}(\hat{\beta}_1, \hat{\beta}_2, \sigma^2)$ par rapport à σ^2. Calculons donc la dérivée par rapport à σ^2 :

$$\frac{\partial \log \mathcal{L}(\hat{\beta}_1, \hat{\beta}_2, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} S(\hat{\beta}_1, \hat{\beta}_2) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^n (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i)^2$$

D'où l'on déduit que l'estimateur du maximum de vraisemblance de σ^2 est différent de l'estimateur $\hat{\sigma}^2$ vu précédemment et vaut :

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

L'estimateur du maximum de vraisemblance de σ^2 est donc biaisé. On a en effet $\mathbb{E}[\hat{\sigma}^2_{mv}] = \frac{n-2}{n}\sigma^2$, mais ce biais est d'autant plus négligeable que le nombre d'observations est grand.

Avant de passer aux lois des estimateurs et aux intervalles de confiance qui s'en déduisent, faisons quelques rappels sur les lois usuelles dans ce contexte.

1.4.2 Rappels sur les lois usuelles

Outre la sacro-sainte gaussienne, trois lois seront d'usage constant dans la suite : la loi du χ^2, la loi de Student et la loi de Fisher.

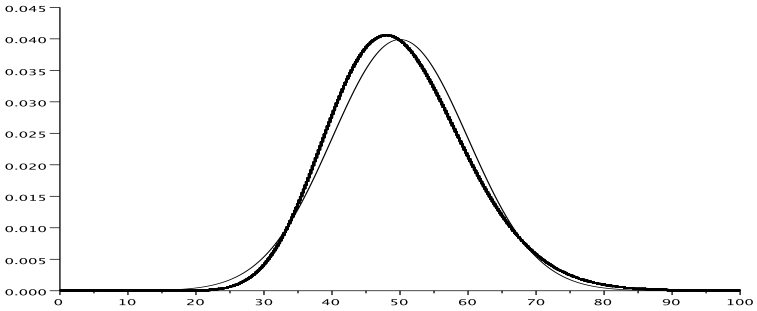


FIGURE 1.4 – Densité d'un χ^2_{50} (trait gras) et densité d'une $\mathcal{N}(50, 100)$ (trait fin).

Définition 4 (Loi du χ^2) *Soit $X_1, \ldots, X_n$ des variables aléatoires i.i.d. suivant une loi normale centrée réduite. La loi de la variable $X = \sum_{i=1}^{n} X_i^2$ est appelée loi du χ^2 à n degrés de liberté (ddl), noté $X \sim \chi^2_n$.*

On a $\mathbb{E}[X] = n$ et $\mathrm{Var}(X) = 2n$. Lorsque n est grand, on sait par le Théorème Central Limite que X suit approximativement une loi normale de moyenne n et de variance $2n$: $X \approx \mathcal{N}(n, 2n)$. Ainsi, pour n grand, environ 95% des valeurs de X se situent dans l'intervalle $[n - 2\sqrt{2n}, n + 2\sqrt{2n}]$. Ceci est illustré figure 8 pour $n = 50$ ddl.

Définition 5 (Loi de Student) *Soit Z une variable aléatoire suivant une loi normale centrée réduite et X une variable suivant une loi du χ^2 à n degrés de liberté, avec Z et X indépendantes. La loi de la variable $T = \frac{Z}{\sqrt{X/n}}$ est appelée loi de Student à n degrés de liberté et on note $T \sim \mathcal{T}_n$.*

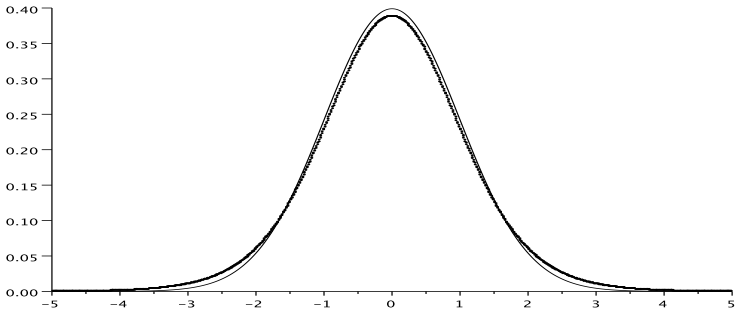


FIGURE 1.5 – Densité d'une $\mathcal{T}_{10}$ (trait gras) et densité d'une $\mathcal{N}(0, 1)$ (trait fin).

Lorsque $n = 1$, T suit une loi de Cauchy et n'a donc pas d'espérance (ni, a fortiori, de variance). Pour $n = 2$, T est centrée mais de variance infinie. Pour $n \geq 3$, T est centrée et de variance $\frac{n}{n-2}$.

D'autre part, lorsque n devient grand, on sait par la Loi des Grands Nombres que le dénominateur tend presque sûrement vers 1. De fait, on peut montrer que pour n grand, T tend en loi vers une gaussienne centrée réduite : $T \approx \mathcal{N}(0,1)$. Ceci est illustré figure 1.5 pour $n = 10$ ddl. Par conséquent, lorsque n sera grand, on pourra remplacer les quantiles d'une loi de Student $\mathcal{T}_n$ par ceux d'une loi $\mathcal{N}(0,1)$ (cf. tables en Annexe C.3).

Définition 6 (Loi de Fisher) *Soit U_1 une variable aléatoire suivant une loi du χ^2 à n_1 degrés de liberté et U_2 une variable aléatoire suivant une loi du χ^2 à n_2 degrés de liberté, avec U_1 et U_2 indépendantes. La loi de la variable $F = \frac{U_1/n_1}{U_2/n_2}$ est appelée loi de Fisher à (n_1, n_2) degrés de liberté et on note $F \sim \mathcal{F}_{n_2}^{n_1}$.*

Pour $n_2 > 2$, l'espérance d'une loi de Fisher $\mathcal{F}_{n_2}^{n_1}$ est $n_2/(n_2-2)$. Dans la suite, typiquement, n_2 sera grand, de sorte qu'à nouveau la Loi des Grands Nombres implique que U_2/n_2 tend vers 1. Dans ce cas, F peut se voir comme un chi-deux normalisé par son degré de liberté : $F \approx \chi^2_{n_1}/n_1$. Ceci est illustré figure 1.6 pour $n_1 = 2$ et $n_2 = 10$.

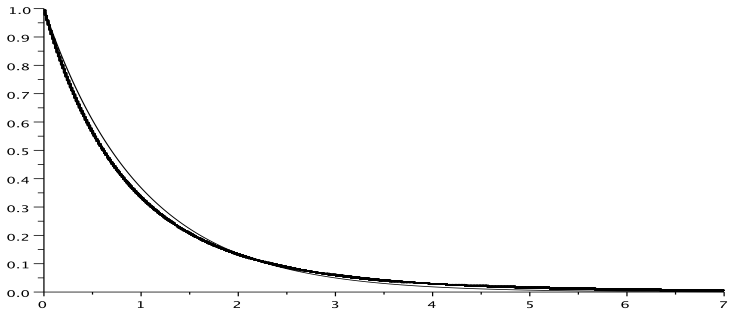


FIGURE 1.6 – Densité d'une $\mathcal{F}_{10}^2$ (trait gras) et densité d'un $\frac{\chi^2_2}{2}$ (trait fin).

1.4.3 Lois des estimateurs et régions de confiance

Nous allons maintenant voir comment les lois précédentes interviennent dans nos estimateurs. Afin de faciliter la lecture de cette partie, fixons les notations suivantes :

$$\begin{aligned} c &= \frac{-\sigma^2 \bar{x}}{\sum (x_i - \bar{x})^2} & \hat{\sigma}^2 &= \frac{1}{n-2} \sum \hat{\varepsilon}_i^2 \\ \sigma_1^2 &= \sigma^2 \left(\frac{\sum x_i^2}{n \sum (x_i - \bar{x})^2} \right) & \hat{\sigma}_1^2 &= \hat{\sigma}^2 \left(\frac{\sum x_i^2}{n \sum (x_i - \bar{x})^2} \right) \\ \sigma_2^2 &= \frac{\sigma^2}{\sum (x_i - \bar{x})^2} & \hat{\sigma}_2^2 &= \frac{\hat{\sigma}^2}{\sum (x_i - \bar{x})^2}. \end{aligned}$$

Comme nous l'avons vu, σ_1^2, σ_2^2 et c sont les variances et covariance des estimateurs des moindres carrés ordinaires. les quantités $\hat{\sigma}_1^2$ et $\hat{\sigma}_2^2$ correspondent quant à elles aux estimateurs des variances de $\hat{\beta}_1$ et $\hat{\beta}_2$.

Propriétés 1 (Lois des estimateurs avec variance connue) *Les lois des estimateurs des MCO avec variance σ^2 connue sont :*

(i) $\hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} \sim \mathcal{N}\left(\beta, \sigma^2 V\right)$ *où* $\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$ *et*

$$V = \frac{1}{\sum (x_i - \bar{x})^2} \begin{bmatrix} \sum x_i^2/n & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix} = \frac{1}{\sigma^2} \begin{bmatrix} \sigma_1^2 & c \\ c & \sigma_2^2 \end{bmatrix}.$$

(ii) $\dfrac{(n-2)}{\sigma^2}\hat{\sigma}^2 \sim \chi^2_{n-2}$*, loi du* χ^2 *à* $(n-2)$ *degrés de liberté.*
(iii) $\hat{\beta}$ *et* $\hat{\sigma}^2$ *sont indépendants.*

Remarque. Ces propriétés, comme celles à venir, ne sont pas plus faciles à montrer dans le cadre de la régression linéaire simple que dans celui de la régression linéaire multiple. C'est pourquoi nous reportons les preuves au Chapitre 3.

Le problème des propriétés ci-dessus vient de ce qu'elles font intervenir la variance théorique σ^2, généralement inconnue. La façon naturelle de procéder est de la remplacer par son estimateur $\hat{\sigma}^2$. Les lois intervenant dans les estimateurs s'en trouvent de fait légèrement modifiées.

Propriétés 2 (Lois des estimateurs avec variance estimée) *Les lois des estimateurs des MCO avec variance* $\hat{\sigma}^2$ *estimée sont :*

(i) $\dfrac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_1} \sim \mathcal{T}_{n-2}$*, où* $\mathcal{T}_{n-2}$ *est une loi de Student à* $(n-2)$ *degrés de liberté.*
(ii) $\dfrac{\hat{\beta}_2 - \beta_2}{\hat{\sigma}_2} \sim \mathcal{T}_{n-2}$.
(iii) $\dfrac{1}{2\hat{\sigma}^2}(\hat{\beta} - \beta)' V^{-1} (\hat{\beta} - \beta) \sim \mathcal{F}^2_{n-2}$*, loi de Fisher de paramètres* $(2, n-2)$.

Ces dernières propriétés nous permettent de donner des intervalles de confiance (IC) ou des régions de confiance (RC) des estimateurs. En effet, la valeur ponctuelle d'un estimateur est de peu d'intérêt en général et il est intéressant de lui associer un intervalle de confiance. Les résultats sont donnés pour un α général, en pratique on prend typiquement $\alpha = 0,05$.

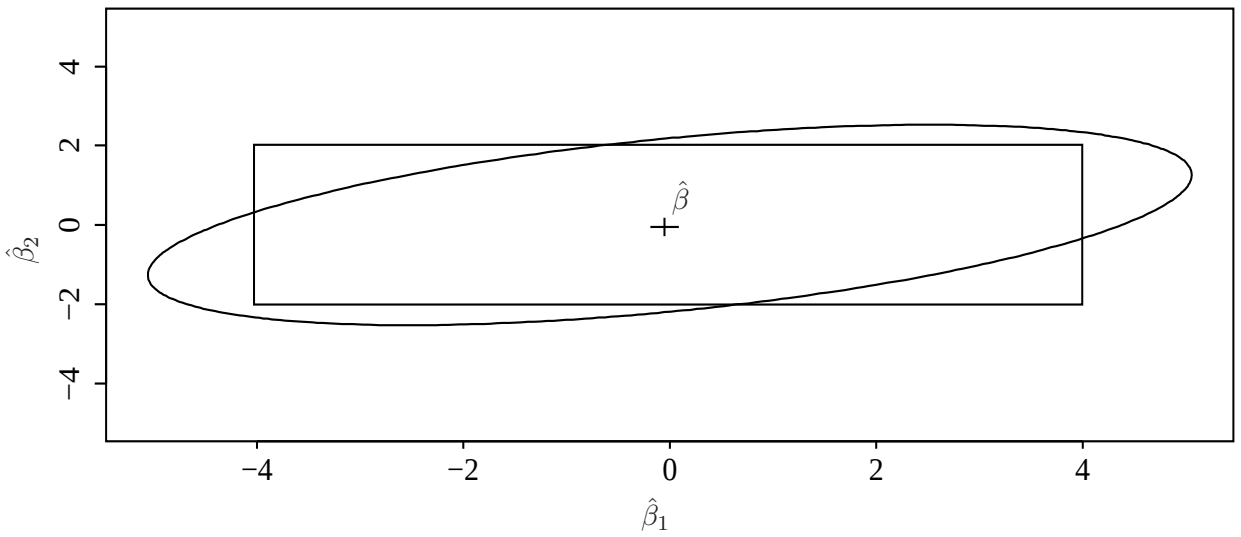


FIGURE 1.7 – Comparaison entre ellipse de confiance et rectangle de confiance.

Propriétés 3 (Intervalles et régions de confiance) *(i) IC(β_1) : $\hat{\beta}_1 \pm t_{n-2}(1-\alpha/2)\hat{\sigma}_1$, où $t_{n-2}(1-\alpha/2)$ est le quantile de niveau $(1-\alpha/2)$ d'une loi de Student $\mathcal{T}_{n-2}$.*
(ii) IC(β_2) : $\hat{\beta}_2 \pm t_{n-2}(1-\alpha/2)\hat{\sigma}_2$.
(iii) RC(β) : Une région de confiance simultanée pour β_1 et β_2 au niveau $(1-\alpha)$ est

$$\frac{1}{2\hat{\sigma}^2}\left(n(\hat{\beta}_1-\beta_1)^2+2n\bar{x}(\hat{\beta}_1-\beta_1)(\hat{\beta}_2-\beta_2)+\sum x_i^2(\hat{\beta}_2-\beta_2)^2\right) \leq f_{n-2}^2(1-\alpha),$$

où $f_{n-2}^2(1-\alpha)$ est le quantile de niveau $(1-\alpha)$ d'une loi $\mathcal{F}_{n-2}^2$.
(iv) Un intervalle de confiance de σ^2 est donné par :

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

où $c_{n-2}(1-\alpha/2)$ est le quantile de niveau $(1-\alpha/2)$ d'une loi χ_{n-2}^2.

Remarque : Le point (iii) donne la région de confiance simultanée des paramètres (β_1, β_2) de la régression, appelée ellipse de confiance, tandis que (i) et (ii) donnent des intervalles de confiance pour β_1 et β_2 pris séparément. La figure 1.7 montre la différence entre ces deux notions.

1.4.4 Prévision

En matière de prévision dans le cas d'erreurs gaussiennes, les résultats obtenus en section 1.2.4 pour l'espérance et la variance sont toujours valables. De plus, puisque $\hat{y}_{n+1}$ est linéaire en $\hat{\beta}_1$, $\hat{\beta}_2$ et ε_{n+1}, on peut préciser sa loi :

$$y_{n+1}-\hat{y}_{n+1} \sim \mathcal{N}\left(0, \sigma^2\left(1+\frac{1}{n}+\frac{(x_{n+1}-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right)\right).$$

A nouveau on ne connaît pas σ^2 et on l'estime donc par $\hat{\sigma}^2$. Comme $(y_{n+1}-\hat{y}_{n+1})$ et $\hat{\sigma}^2(n-2)/\sigma^2$ sont indépendants, on peut énoncer un résultat donnant des intervalles de confiance pour y_{n+1}.

Proposition 3 (Loi et intervalle de confiance pour la prédiction) *Avec les notations et hypothèses précédentes, on a :*

$$\frac{y_{n+1}-\hat{y}_{n+1}}{\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_{n+1}-\bar{x})^2}{\sum(x_i-\bar{x})^2}}} \sim \mathcal{T}_{n-2},$$

d'où l'on déduit l'intervalle de confiance suivant pour y_{n+1} :

$$\left[\hat{y}_{n+1} \pm t_{n-2}(1-\alpha/2)\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_{n+1}-\bar{x})^2}{\sum(x_i-\bar{x})^2}}\right].$$

Nous retrouvons ainsi la remarque déjà faite : plus le point à prévoir admet pour abscisse x_{n+1} une valeur éloignée de $\bar{x}$, plus l'intervalle de confiance sera grand.

Plus précisément, la courbe décrite pas les limites de ces intervalles de confiance lorsque x_{n+1} varie est une hyperbole d'axes $x=\bar{x}$ et $y=\hat{\beta}_1+\hat{\beta}_2 x$. Pour s'en persuader, il suffit d'effectuer le changement de variables

$$\begin{cases} X = x-\bar{x} \\ Y = y-(\hat{\beta}_1+\hat{\beta}_2 x) \end{cases}$$

d'où il ressort qu'un point (X,Y) est dans la région de confiance ci-dessus si et seulement si

$$\frac{Y^2}{b^2}-\frac{X^2}{a^2} \leq 1,$$

avec

$$\begin{cases} a^2 = \left(1 + \frac{1}{n}\right) \sum (x_i - \bar{x})^2 \\ b^2 = \left(1 + \frac{1}{n}\right) (t_{n-2}(1 - \alpha/2)\hat{\sigma})^2 \end{cases}$$

ce qui définit bien l'intérieur d'une hyperbole. En particulier, le centre de cette hyperbole est tout bonnement le centre de gravité du nuage de points.

1.5 Exemple

Nous allons traiter les 50 données journalières présentées en Annexe D. La variable à expliquer est la concentration en ozone, notée `O3`, et la variable explicative est la température à midi, notée `T12`. Les données sont traitées avec le logiciel `R`.

```
> a <- lm(O3 ~ T12)
> summary(a)
Call:
lm(formula = O3 ~ T12)


Residuals:
     Min       1Q   Median       3Q      Max
-45.256  -15.326   -3.461   17.634   40.072

Coefficients :
             Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)   31.4150     13.0584    2.406    0.0200       *
T12            2.7010      0.6266    4.311   8.04e-05     ***

--
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 20.5 on 48 degrees of freedom
Multiple R-Squared: 0.2791, Adjusted R-squared: 0.2641
F-statistic: 18.58 on 1 and 48 DF, p-value: 8.041e-05
```

Les sorties du logiciel donnent les valeurs estimées $\hat{\beta}_1$ et $\hat{\beta}_2$ des paramètres, leurs écart-types $\hat{\sigma}_1$ et $\hat{\sigma}_2$, les statistiques de tests sous l'hypothèse $H_0 : \beta_i = 0$. Nous rejetons H_0 pour les deux paramètres estimés.

1.6 Exercices

Exercice 1.1 (QCM)

1. Lors d'une régression simple, si le R^2 vaut 1, les points sont-ils alignés ?
   A. Non ;
   B. Oui ;
   C. Pas obligatoirement.
2. La droite des MCO d'une régression simple passe-t-elle par le point $(\bar{x}, \bar{y})$?
   A. Toujours ;
   B. Jamais ;
   C. Parfois.

3. Nous avons effectué une régression simple, nous recevons une nouvelle observation x_N et nous calculons la prévision correspondante $\hat{y}_N$. La variance de la valeur prévue est minimale lorsque
   A. $x_N = 0$;
   B. $x_N = \bar{x}$;
   C. Aucun rapport.
4. Le vecteur $\hat{Y}$ est-il orthogonal au vecteur des résidus estimés $\hat{\varepsilon}$?
   A. Toujours ;
   B. Jamais ;
   C. Parfois.

Exercice 1.2 (R^2 et corrélation empirique) Rappeler la formule définissant le coefficient de détermination R^2 et la développer pour montrer qu'il est égal au carré du coefficient de corrélation empirique entre x et y, noté $\rho_{x,y}$, c'est-à-dire qu'on a :

$$R^2 = \rho_{x,y}^2 = \left(\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \right)^2$$

Exercice 1.3 (Poids des pères et des fils) L'étude statistique ci-dessous porte sur les poids respectifs des pères et de leur fil aîné.

Père	65	63	67	64	68	62	70	66	68	67	69	71
Fils	68	66	68	65	69	66	68	65	71	67	68	70

Voici les résultats numériques que nous avons obtenus :

$$\sum_{i=1}^{12} p_i = 800 \qquad \sum_{i=1}^{12} p_i^2 = 53418 \qquad \sum_{i=1}^{12} p_i f_i = 54107 \qquad \sum_{i=1}^{12} f_i = 811 \qquad \sum_{i=1}^{12} f_i^2 = 54849.$$

1. Calculez la droite des moindres carrés du poids des fils en fonction du poids des pères.
2. Calculez la droite des moindres carrés du poids des pères en fonction du poids des fils.
3. Montrer que le produit des pentes des deux droites est égal au carré du coefficient de corrélation empirique entre les p_i et les f_i (ou encore au coefficient de détermination).

Exercice 1.4 (Hauteur d'un arbre) Nous souhaitons exprimer la hauteur y (en pieds) d'un arbre d'une essence donnée en fonction de son diamètre x (en pouces) à 1m30 du sol. Pour ce faire, nous avons mesuré 20 couples (diamètre,hauteur) et effectué les calculs suivants : $\bar{x} = 4.53$, $\bar{y} = 8.65$ et

$$\frac{1}{20}\sum_{i=1}^{20} (x_i - \bar{x})^2 = 10.97 \qquad \frac{1}{20}\sum_{i=1}^{20} (y_i - \bar{y})^2 = 2.24 \qquad \frac{1}{20}\sum_{i=1}^{20} (x_i - \bar{x})(y_i - \bar{y}) = 3.77$$

1. On note $y = \hat{\beta}_0 + \hat{\beta}_1 x$ la droite de régression. Calculer $\hat{\beta}_0$ et $\hat{\beta}_1$.
2. Donner et commenter une mesure de la qualité de l'ajustement des données au modèle. Exprimer cette mesure en fonction des statistiques élémentaires. Commenter le résultat.

3. On donne les estimations de l'écart-type de $\hat{\beta}_0$, $\hat{\sigma}_0 = 1.62$, et de $\hat{\beta}_1$, $\hat{\sigma}_1 = 0.05$. On suppose les perturbations ε_i gaussiennes, centrées, de même variance et indépendantes. Tester $H_0 : \beta_j = 0$ contre $H_1 : \beta_j \neq 0$ pour $j = 0, 1$. Pourquoi ce test est-il intéressant dans notre contexte ? Que pensez-vous du résultat ?

Exercice 1.5 (Droite de régression et points atypiques) Douze personnes sont inscrites à une formation. Au début de la formation, ces stagiaires subissent une épreuve A notée sur 20. A la fin de la formation, elles subissent une épreuve B de niveau identique. Les résultats sont donnés dans le tableau suivant :

Epreuve A	3	4	6	7	9	10	9	11	12	13	15	4
Epreuve B	8	9	10	13	15	14	13	16	13	19	6	19

1. Représenter le nuage de points. Déterminer la droite de régression. Calculer le coefficient de détermination. Commenter.
2. Deux stagiaires semblent se distinguer des autres. Les supprimer et déterminer la droite de régression sur les dix points restants. Calculer le coefficient de détermination. Commenter.

Exercice 1.6 (La hauteur des eucalyptus) On souhaite expliquer la hauteur y (en mètres) d'un arbre en fonction de sa circonférence x (en centimètres) à 1m30 du sol. On a relevé $n = 1429$ couples (x_i, y_i), le nuage de points étant représenté figure 1.8. On a obtenu $(\bar{x}, \bar{y}) = (47,3; 21,2)$ et :

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = 102924 \qquad \sum_{i=1}^{n}(y_i - \bar{y})^2 = 8857 \qquad \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = 26466$$

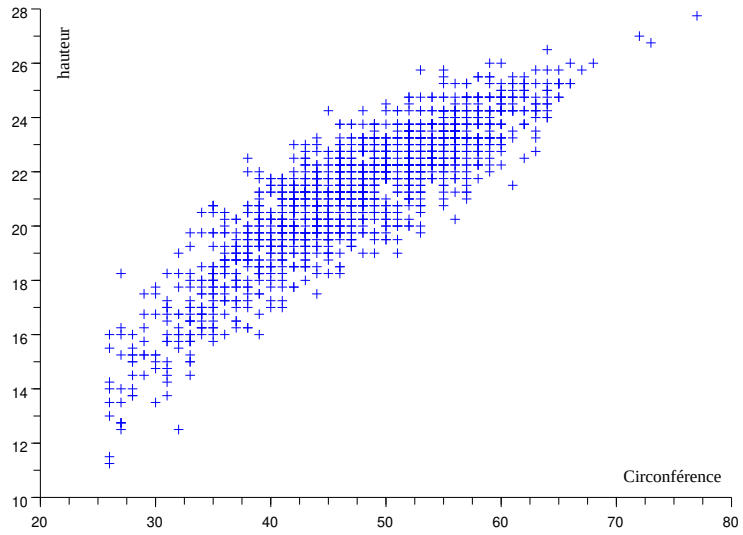


FIGURE 1.8 – Nuage de points pour les eucalyptus.

1. Calculer la droite des moindres carrés pour le modèle $y = \beta_1 + \beta_2 x + \varepsilon$ et la représenter sur la figure 1.8.
2. Calculer le coefficient de détermination R^2. Commenter la qualité de l'ajustement des données au modèle.

3. Avec ces estimateurs, la somme des carrés des résidus vaut alors $\sum_{i=1}^n (y_i - \hat{y}_i)^2 = 2052$. Si on suppose les perturbations ε_i gaussiennes, centrées, indépendantes et de même variance σ^2, en déduire un estimateur non biaisé $\hat{\sigma}^2$ de σ^2.
4. Donner un estimateur $\hat{\sigma}_1^2$ de la variance de $\hat{\beta}_1$.
5. Tester l'hypothèse $H_0 : \beta_1 = 0$ contre $H_1 : \beta_1 \neq 0$.

Exercice 1.7 (Forrest Gump for ever) On appelle "fréquence seuil" d'un sportif amateur sa fréquence cardiaque obtenue après trois quarts d'heure d'un effort soutenu de course à pied. Celle-ci est mesurée à l'aide d'un cardio-fréquence-mètre. On cherche à savoir si l'âge d'un sportif a une influence sur sa fréquence seuil. On dispose pour cela de 20 valeurs du couple (x_i, y_i), où x_i est l'âge et y_i la fréquence seuil du sportif. On a obtenu $(\bar{x}, \bar{y}) = (35,6; 170,2)$ et :

$$\sum_{i=1}^n (x_i - \bar{x})^2 = 1991 \qquad \sum_{i=1}^n (y_i - \bar{y})^2 = 189,2 \qquad \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = -195,4$$

1. Calculer la droite des moindres carrés pour le modèle $y = \beta_1 + \beta_2 x + \varepsilon$.
2. Calculer le coefficient de détermination R^2. Commenter la qualité de l'ajustement des données au modèle.
3. Avec ces estimateurs, la somme des carrés des résidus vaut $\sum_{i=1}^n (y_i - \hat{y}_i)^2 = 170$. Si on suppose les perturbations ε_i gaussiennes, centrées, indépendantes et de même variance σ^2, en déduire un estimateur non biaisé $\hat{\sigma}^2$ de σ^2.
4. Donner un estimateur $\hat{\sigma}_2^2$ de la variance de $\hat{\beta}_2$.
5. Tester l'hypothèse $H_0 : \beta_2 = 0$ contre $H_1 : \beta_2 \neq 0$ pour un risque de 5%. Conclure sur la question de l'influence de l'âge sur la fréquence seuil.

Exercice 1.8 (Comparaison d'estimateurs) Nous considérons le modèle statistique suivant :

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

où nous supposons que les perturbations ε_i sont telles que $\mathbb{E}[\varepsilon_i] = 0$ et $\mathrm{Cov}(\varepsilon_i, \varepsilon_i) = \sigma^2 \delta_{i,j}$.

1. En revenant à la définition des moindres carrés, montrer que l'estimateur des moindres carrés de β vaut

$$\hat{\beta} = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}.$$

2. Montrer que la droite passant par l'origine et le centre de gravité du nuage de points est $y = \beta^* x$, avec

$$\beta^* = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n x_i}.$$

3. Montrer que $\hat{\beta}$ et β^* sont tous deux des estimateurs sans biais de β.
4. On rappelle l'inégalité de Cauchy-Schwarz : si $u = [u_1, \ldots, u_n]'$ et $v = [v_1, \ldots, v_n]'$ sont deux vecteurs de $\mathbb{R}^n$, alors leur produit scalaire est (en valeur absolue) plus petit que le produit de leurs normes, c'est-à-dire :

$$|\langle u, v \rangle| \leq \|u\| \times \|v\| \iff \left| \sum_{i=1}^n u_i v_i \right| \leq \sqrt{\sum_{i=1}^n u_i^2} \times \sqrt{\sum_{i=1}^n v_i^2},$$

avec égalité si et seulement si u et v sont colinéaires. Grâce à cette inégalité, montrer que $V(\beta^*) > V(\hat{\beta})$ sauf dans le cas où tous les x_i sont égaux. Ce résultat était-il prévisible ?

Exercice 1.9 (Intervalles de confiance vs Région de confiance) On considère le modèle de régression linéaire simple $y = \beta_1 + \beta_2 x + \epsilon$. Soit un échantillon $(x_i, y_i)_{1 \leq i \leq 100}$ de statistiques résumées

$$\sum_{i=1}^{100} x_i = 0 \qquad \sum_{i=1}^{100} x_i^2 = 400 \qquad \sum_{i=1}^{100} x_i y_i = 100 \qquad \sum_{i=1}^{100} y_i = 100 \qquad \hat{\sigma}^2 = 1.$$

1. Exprimer les intervalles de confiance à 95% pour β_1 et β_2.
2. Donner l'équation de la région de confiance à 95% de (β_1, β_2). (Rappel : l'ensemble des points (x, y) tels que $\frac{(x-x_0)^2}{a^2} + \frac{(y-y_0)^2}{b^2} \leq 1$ est l'intérieur d'une ellipse centrée en (x_0, y_0), dont les axes sont parallèles à ceux des abscisses et des ordonnées, et de sommets $(x_0 \pm a, 0)$ et $(0, y_0 \pm b)$.)
3. Représenter sur un même graphique les résultats obtenus.

Exercice 1.10 (Régression simple) On dispose de n points $(x_i, y_i)_{1 \leq i \leq n}$ et on sait qu'il existe une relation de la forme : $y_i = ax_i + b + \varepsilon_i$, où les erreurs ε_i sont des variables centrées, décorrélées et de même variance σ^2.

1. Rappeler les formules des estimateurs des moindres carrés $\hat{a}$ et $\hat{b}$, ainsi que leurs variances respectives.
2. Dans cette question, on suppose connaître b, mais pas a.
   (a) En revenant à la définition des moindres carrés, calculer l'estimateur $\tilde{a}$ des moindres carrés de a.
   (b) Calculer la variance de $\tilde{a}$. Montrer qu'elle est inférieure à celle de $\hat{a}$.
3. Dans cette question, on suppose connaître a, mais pas b.
   (a) En revenant à la définition des moindres carrés, calculer l'estimateur $\tilde{b}$ des moindres carrés de b.
   (b) Calculer la variance de $\tilde{b}$. Montrer qu'elle est inférieure à celle de $\hat{b}$.

Exercice 1.11 (Forces de frottement et vitesse) Au $17^{\text{ème}}$ siècle, Huygens s'est intéressé aux forces de résistance d'un objet en mouvement dans un fluide (eau, air, etc.). Il a d'abord émis l'hypothèse selon laquelle les forces de frottement étaient proportionnelles à la vitesse de l'objet, puis, après expérimentation, selon laquelle elles étaient proportionnelles au carré de la vitesse. On réalise une expérience dans laquelle on fait varier la vitesse x d'un objet et on mesure les forces de frottement y. Ensuite, on teste la relation existant entre ces forces de frottement et la vitesse.

1. Quel(s) modèle(s) testeriez-vous ?
2. Comment feriez-vous pour déterminer le modèle adapté ?

Exercice 1.12 (Prix d'un appartement en fonction de sa superficie) En juin 2005, on a relevé dans les petites annonces les superficies (en m^2) et les prix (en euros) de 108 appartements de type T3 à louer sur l'agglomération de Rennes (cf. figure 1.9).

1. D'après le listing du tableau 1.2, donner une estimation du coefficient de corrélation entre le prix et la superficie d'un appartement T3.

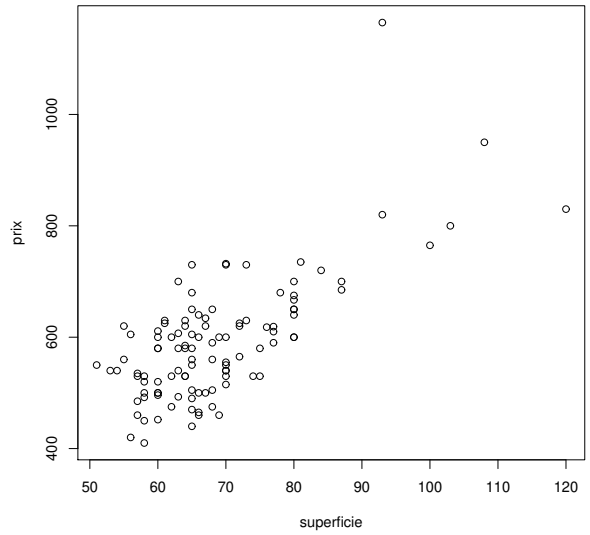


FIGURE 1.9 – Prix de location des appartements en fonction de leur superficie.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 134.3450    45.4737   2.954  0.00386
Superficie    6.6570     0.6525  10.203  < 2e-16

Residual standard error: 77.93 on 106 degrees of freedom
Multiple R-Squared: 0.4955,     Adjusted R-squared: 0.4907
F-statistic: 104.1 on 1 and 106 DF,  p-value: < 2.2e-16
```

TABLE 1.2 – Prix en fonction de la superficie : résultats de la régression linéaire simple (sortie R).

2. Proposer un modèle permettant d'étudier la relation entre le prix des appartements et leur superficie. Préciser les hypothèses de ce modèle.
3. D'après le tableau 1.2, est-ce que la superficie joue un rôle sur le prix des appartements de type 3 ? Considérez-vous ce rôle comme important ?
4. Quelle est l'estimation du coefficient β (coefficient de la superficie dans le modèle) ? Comment interprétez-vous ce coefficient ?
5. La superficie moyenne des 108 appartements est de 68.74 m^2 et le prix moyen des appartements est de 591.95 euros. Quel est le prix moyen d'un mètre carré ? Pourquoi ce prix moyen est différent de l'estimation de β ?
6. Dans l'échantillon dont on dispose, comment savoir quels sont les appartements "bon marché" du seul point de vue de la surface ?

Exercice 1.13 (Total Least Squares (TLS)) Nous avons un nuage de points observés (x_i, y_i) pour $i = 1, \cdots, n$, et nous cherchons un couple $(\hat{x}, \hat{y})$ vérifiant la relation linéaire suivante

$$\hat{y} = \alpha \hat{x},$$

tel que la norme matricielle $\|[x, y] - [\hat{x}, \hat{y}]\|_F$ soit minimale (rappel : $\|A\|_F = \sqrt{\mathrm{Tr}(AA')}$).

1. Que représente la norme matricielle $\|[x,y]-[\hat{x},\hat{y}]\|_F$ d'un point de vue géométrique ?
2. Supposons pour simplifier que $\bar{x}=\bar{y}=0$, c'est-à-dire que le centre de gravité du nuage de points est en l'origine du repère. Quel rapport voyez-vous entre TLS et ACP ?

1.7 Corrigés

Exercice 1.1 (QCM) C'est le B.A.-BA.

Exercice 1.2 (R^2 et corrélation empirique) Le coefficient R^2 s'écrit

$$\begin{aligned} R^2 &= \frac{\|\hat{Y}-\bar{y}\mathbb{1}\|^2}{\|Y-\bar{y}\mathbb{1}\|^2} = \frac{\sum_{i=1}^n \left(\hat{\beta}_1+\hat{\beta}_2 x_i-\bar{y}\right)^2}{\sum_{i=1}^n (y_i-\bar{y})^2} = \frac{\sum_{i=1}^n \left(\bar{y}-\hat{\beta}_2\bar{x}+\hat{\beta}_2 x_i-\bar{y}\right)^2}{\sum_{i=1}^n (y_i-\bar{y})^2} \\ &= \frac{\hat{\beta}_2^2\sum_{i=1}^n (x_i-\bar{x})^2}{\sum_{i=1}^n (y_i-\bar{y})^2} = \frac{\left[\sum_{i=1}^n (x_i-\bar{x})(y_i-\bar{y})\right]^2 \sum_{i=1}^n (x_i-\bar{x})^2}{\left[\sum_{i=1}^n (x_i-\bar{x})^2\right]^2 \sum_{i=1}^n (y_i-\bar{y})^2} \\ &= \frac{\left[\sum_{i=1}^n (x_i-\bar{x})(y_i-\bar{y})\right]^2}{\sum_{i=1}^n (x_i-\bar{x})^2 \sum_{i=1}^n (y_i-\bar{y})^2} = \rho_{x,y}^2, \end{aligned}$$

et la messe est dite.

Exercice 1.3 (Poids des pères et des fils)

1. La droite des moindres carrés du poids des fils en fonction du poids des pères s'écrit (cf. figure 1.10 à gauche) : $f=\hat{\alpha}_1+\hat{\alpha}_2 p = 35.8+0.48p$.
2. La droite des moindres carrés du poids des fils en fonction du poids des pères s'écrit (cf. figure 1.10 à droite) : $p=\hat{\beta}_1+\hat{\beta}_2 f = -3.38+1.03f$.
3. Le produit des pentes des deux droites est

$$\hat{\alpha}_2\hat{\beta}_2 = \frac{\left(\sum(f_i-\bar{f})(p_i-\bar{p})\right)^2}{\left(\sum(f_i-\bar{f})^2\right)\left(\sum_{i=1}^n (p_i-\bar{p})^2\right)} = R^2,$$

où R^2 est le coefficient de détermination, carré du coefficient de corrélation linéaire.

Exercice 1.4 (Hauteur d'un arbre) Nous souhaitons exprimer la hauteur y (en pieds) d'un arbre d'une essence donnée en fonction de son diamètre x (en pouces) à 1m30 du sol. Pour ce faire, nous avons mesuré 20 couples (diamètre,hauteur) et effectué les calculs suivants : $\bar{x}=4.53$, $\bar{y}=8.65$ et

$$\frac{1}{20}\sum_{i=1}^{20}(x_i-\bar{x})^2 = 10.97 \qquad \frac{1}{20}\sum_{i=1}^{20}(y_i-\bar{y})^2 = 2.24 \qquad \frac{1}{20}\sum_{i=1}^{20}(x_i-\bar{x})(y_i-\bar{y}) = 3.77$$

1. Les estimateurs de la droite des moindres carrés $y=\hat{\beta}_0+\hat{\beta}_1 x$ sont respectivement :

$$\hat{\beta}_1 = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2} \approx 0.344$$

et

$$\hat{\beta}_0 = \bar{y}-\hat{\beta}_1\bar{x} \approx 7.09$$

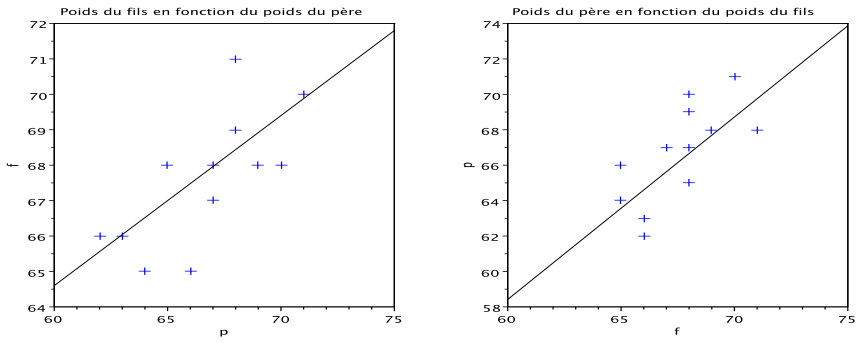


FIGURE 1.10 – Nuages de points et droites de régression pour les poids des pères et des fils.

2. Une mesure de la qualité de l'ajustement des données au modèle est donnée par le coefficient de détermination R^2, dont on a vu qu'il correspond au carré du coefficient de corrélation linéaire empirique :

$$R^2 = \left(\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \right)^2 \approx 0.58.$$

Le modèle de régression linéaire simple explique donc un peu plus de la moitié de la variance présente dans les données.

3. Sous H_0, on sait que

$$\frac{\hat{\beta}_0}{\hat{\sigma}_0} \sim \mathcal{T}_{18},$$

loi de Student à 18 degrés de liberté. Pour un niveau de confiance de 95%, on compare donc la valeur absolue obtenue dans notre cas particulier, à savoir $|\hat{\beta}_0/\hat{\sigma}_0| \approx 4.38$ au quantile $t_{18}(0.975) \approx 2.1$. On en déduit qu'on rejette l'hypothèse selon laquelle β_0 serait nul. De même pour le test d'hypothèse sur β_1, ce qui donne la statistique de test :

$$\left| \frac{\hat{\beta}_1}{\hat{\sigma}_1} \right| \approx 6.88 > 2.1$$

donc on rejette également l'hypothèse selon laquelle β_1 serait nul.

A priori, un arbre de diamètre nul a une hauteur égale à zéro, donc on aurait pu s'attendre à ce que le coefficient β_0 soit nul. Ceci est en contradiction avec le résultat du test d'hypothèse ci-dessus, mais il n'y a rien d'étonnant à ça : le modèle de régression proposé est pertinent dans l'intervalle considéré, c'est-à-dire pour des arbres de hauteur moyenne 8.65 pieds, avec un écart-type égal à $\sqrt{2.24} \approx 1.5$, non pour des arbres tout petits.

Exercice 1.5 (Droite de régression et points aberrants) Douze personnes sont inscrites à une formation. Au début de la formation, ces stagiaires subissent une épreuve A notée sur 20. A la fin de la formation, elles subissent une épreuve B de niveau identique. Les résultats sont donnés dans le tableau suivant :

Epreuve A	3	4	6	7	9	10	9	11	12	13	15	4
Epreuve B	8	9	10	13	15	14	13	16	13	19	6	19

1. Pour l'explication de la note B à partir de la note A, la droite de régression (cf. figure 1.11 à gauche) est donnée par $y = \hat{\beta}_1 + \hat{\beta}_2 x$, où :

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \approx 0.11$$

et $\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x} \approx 12.0$ Le coefficient de détermination vaut :

$$R^2 = \frac{\left(\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})\right)^2}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)\left(\sum_{i=1}^{n}(y_i - \bar{y})^2\right)} \approx 0,01$$

Le modèle de régression linéaire expliquerait donc 1% de la variance des données, ce qui est très faible.

2. Si on supprime les deux derniers stagiaires, on obtient cette fois (cf. figure 1.11 à droite) $y = \hat{\alpha}_1 + \hat{\alpha}_2 x = 5.47 + 0.90x$ et $R^2 \approx 0.81$. Sans ces deux stagiaires, le modèle de régression linéaire expliquerait donc 81% de la variance des données, ce qui le rend tout à fait pertinent. Les deux derniers stagiaires correspondent à ce qu'on appelle des points aberrants.

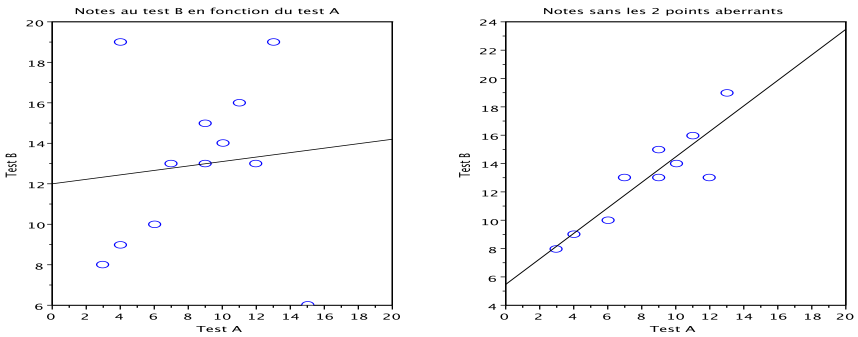


FIGURE 1.11 – Droites de régression et points aberrants.

Exercice 1.6 (La hauteur des eucalyptus) Cet exercice est corrigé en annexe (décembre 2009).

Exercice 1.7 (Forrest Gump for ever)

1. La méthode des moindres carrés ordinaires donne pour estimateur de β_2 :

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \approx -0,098.$$

Et pour estimateur de β_1 :

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x} \approx 173.7.$$

2. Le coefficient de détermination R^2 est égal au carré du coefficient de corrélation linéaire entre les variables x et y, ce qui donne :

$$R^2 = \frac{\left(\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})\right)^2}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)\left(\sum_{i=1}^{n}(y_i - \bar{y})^2\right)} \approx 0,101.$$

On en conclut que 10% de la variance des fréquences seuils y_i est expliquée par l'âge. Ce modèle de régression linéaire simple ne semble donc pas efficace.

3. Un estimateur non biaisé $\hat{\sigma}^2$ de σ^2 est tout simplement :

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2} = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{18} \approx 9.44.$$

4. Un estimateur $\hat{\sigma}_2^2$ de la variance de $\hat{\beta}_2$ est alors donné par :

$$\hat{\sigma}_2^2 = \frac{\hat{\sigma}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \approx 0,0047.$$

5. On sait que l'estimateur centré et normalisé de β_2 suit une loi de Student à $(n-2) = 18$ degrés de liberté :

$$\frac{\hat{\beta}_2 - \beta_2}{\hat{\sigma}_2} \sim \mathcal{T}_{18},$$

donc sous l'hypothèse $H_0 : \beta_2 = 0$, ceci se simplifie en $\frac{\hat{\beta}_2}{\hat{\sigma}_2} \sim \mathcal{T}_{18}$, et cette statistique de test donne ici :

$$t = T(\omega) \approx \frac{-0,098}{\sqrt{0,0047}} \approx -1.43 > -2.101 = t_{18}(0.025).$$

Ainsi on accepte l'hypothèse H_0 selon laquelle la pente de la droite de régression est nulle. Ceci signifie qu'au vu des données dont nous disposons, on serait tenté de considérer que l'âge n'a pas d'influence sur la fréquence seuil. Vu la valeur du coefficient de détermination, il faut toutefois tenir compte du fait que le modèle n'explique pas grand-chose...

Exercice 1.8 (Comparaison d'estimateurs) Nous considérons le modèle statistique

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

où nous supposons que les perturbations ε_i sont telles que $\mathbb{E}[\varepsilon_i] = 0$ et $\mathrm{Cov}(\varepsilon_i, \varepsilon_i) = \sigma^2 \delta_{i,j}$.

1. Par définition, l'estimateur des moindres carrés de β vérifie

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^n (y_i - \beta x_i)^2 = \arg\min_{\beta} S(\beta).$$

Cette fonction S est strictement convexe et admet donc un unique minimum au point où sa dérivée s'annule :

$$S'(\beta) = -2 \sum_{i=1}^n x_i (y_i - \beta x_i) = 2 \left(\beta \sum_{i=1}^n x_i^2 - \sum_{i=1}^n x_i y_i \right).$$

Ceci mène bien à :

$$\hat{\beta} = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}.$$

2. La droite passant par l'origine et le centre de gravité $(\bar{x}, \bar{y})$ du nuage de points admet pour équation $y = \beta^* x$, où

$$\beta^* = \frac{\bar{y}}{\bar{x}} = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n x_i}.$$

3. Commencons par réécrire les estimateurs obtenus grâce à la relation $y_i = \beta x_i + \varepsilon_i$. Pour le premier, ceci donne :

$$\hat{\beta} = \beta + \frac{\sum_{i=1}^n x_i \varepsilon_i}{\sum_{i=1}^n x_i^2},$$

et pour le second :

$$\beta^* = \beta + \frac{\sum_{i=1}^n \varepsilon_i}{\sum_{i=1}^n x_i}.$$

Puisque par hypothèse les erreurs sont centrées (i.e. $\mathbb{E}[\varepsilon_i] = 0$), il en découle que $\mathbb{E}[\hat{\beta}] = \mathbb{E}[\beta^*] = \beta$, c'est-à-dire que les deux estimateurs sont sans biais.

4. On réutilise les expressions précédentes des estimateurs pour cette question. Puisque les erreurs sont décorrélées, la variance de $\hat{\beta}$ vaut

$$V(\hat{\beta}) = \frac{\sum_{i=1}^n x_i^2 \sigma^2}{\left(\sum_{i=1}^n x_i^2\right)^2} = \frac{\sigma^2}{\sum_{i=1}^n x_i^2}.$$

La variance de β^* vaut quant à elle

$$V(\beta^*) = \frac{n\sigma^2}{\left(\sum_{i=1}^n x_i\right)^2}.$$

L'inégalité de Cauchy-Schwarz dit que la valeur absolue du produit scalaire de deux vecteurs est inférieure ou égale au produit de leurs normes, c'est-à-dire : pour tous vecteurs $u = [u_1, \ldots, u_n]'$ et $v = [v_1, \ldots, v_n]'$ de $\mathbb{R}^n$, $|\langle u, v\rangle| \leq \|u\| \times \|v\|$, ou encore en passant aux carrés :

$$\left(\sum_{i=1}^n u_i v_i\right)^2 \leq \left(\sum_{i=1}^n u_i^2\right)\left(\sum_{i=1}^n v_i^2\right),$$

avec égalité si et seulement si u et v sont colinéaires. En prenant $u = [x_1, \ldots, x_n]'$ et $v = [1, \ldots, 1]'$, on en déduit que $V(\beta^*) \geq V(\hat{\beta})$, avec égalité si et seulement si u et v sont colinéaires, c'est-à-dire si et seulement si tous les x_i sont égaux. Puisque les deux estimateurs sont linéaires en y et que $\hat{\beta}$ est celui des moindres carrés, ce résultat n'est pas étonnant si l'on repense au théorème de Gauss-Markov.

Exercice 1.9 (Intervalles de confiance vs Région de confiance) 1. Il sort des statistiques résumées que $\hat{\beta}_1 = \overline{y} - \hat{\beta}_2 \overline{x} = 1$ et $\hat{\beta}_2 = (\sum x_i y_i)/(\sum x_i^2) = 1/4$. La droite des moindres carrés a donc pour équation $y = 1 + x/4$. Les estimateurs des variances se calculent facilement

$$\hat{\sigma}_1^2 = \frac{\hat{\sigma}^2}{\sum (x_i - \overline{x})^2} \times \frac{\sum x_i^2}{n} = \frac{\hat{\sigma}^2}{n} = \frac{1}{100} \;\Rightarrow\; \hat{\sigma}_1 = \frac{1}{10}$$

tandis que

$$\hat{\sigma}_2^2 = \frac{\hat{\sigma}^2}{\sum (x_i - \overline{x})^2} = \frac{1}{400} \;\Rightarrow\; \hat{\sigma}_2 = \frac{1}{20}.$$

Le quantile d'ordre 0.975 d'une Student à 98 degrés de liberté est à peu près le même que celui d'une Student à 100 degrés de liberté, c'est-à-dire environ 1.984 que l'on va arrondir à 2. L'intervalle de confiance à 95% pour β_1 est donc

$$IC(\beta_1) = [\hat{\beta}_1 - 2\hat{\sigma}_1, \hat{\beta}_1 + 2\hat{\sigma}_1] = [0.8; 1.2]$$

et pour β_2

$$IC(\beta_2) = [\hat{\beta}_2 - 2\hat{\sigma}_2, \hat{\beta}_2 + 2\hat{\sigma}_2] = [0.15; 1.35]$$

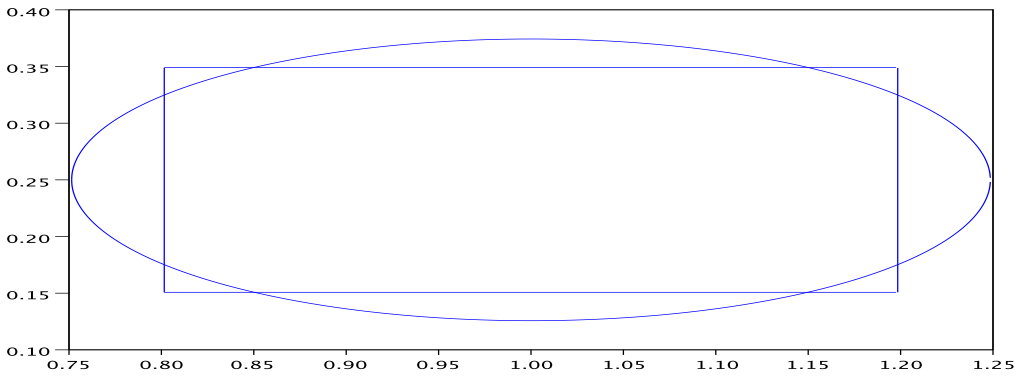


FIGURE 1.12 – Intervalles de confiance vs Région de confiance.

2. Avec les notations du cours, la région de confiance simultanée à 95% est l'ensemble des points (β_1, β_2) tels que

$$\frac{1}{2\hat{\sigma}^2}\left(n(\beta_1-\hat{\beta}_1)^2+2n\bar{x}(\beta_1-\hat{\beta}_1)(\beta_2-\hat{\beta}_2)+\sum x_i^2(\beta_2-\hat{\beta}_2)^2\right)\leq f^2_{n-2}(0.95).$$

Le quantile d'ordre 0.95 d'une loi de Fisher à (2,100) degrés de liberté étant égal à 3.09, nous arrondirons à nouveau et prendrons $f^2_{98}(0.95)\approx 3$, de sorte que nous obtenons comme région de confiance l'ensemble des points (β_1, β_2) tels que

$$\frac{1}{2}\left(100(\beta_1-1)^2+400(\beta_2-1/4)^2\right)\leq 3 \;\Leftrightarrow\; \frac{(\beta_1-1)^2}{\left(\frac{\sqrt{6}}{10}\right)^2}+\frac{(\beta_2-1/4)^2}{\left(\frac{\sqrt{6}}{20}\right)^2}\leq 1.$$

La région de confiance est donc l'intérieur d'une ellipse de centre $(\hat{\beta}_1, \hat{\beta}_2) = (1, 1/4)$ et de sommets $(1 \pm \sqrt{6}/10, 0)$ et $(0, 1/4 \pm \sqrt{6}/20)$, c'est-à-dire $(1.24, 0)$, $(0, 0.37)$, $(0.76, 0)$, $(0, 0.13)$.

3. Les résultats obtenus sont représentés figure 1.12.

Exercice 1.10 (Régression simple) Cet exercice est corrigé en annexe, sujet de décembre 2010.

Exercice 1.11 (Forces de frottement et vitesse) Cet exercice est corrigé en annexe, sujet de décembre 2010.

Exercice 1.12 (Prix d'un appartement en fonction de sa superficie) Cet exercice est corrigé en annexe, sujet de décembre 2011.

Chapitre 2

La régression linéaire multiple

Introduction

La modélisation de la concentration d'ozone dans l'atmosphère évoquée au Chapitre 1 est relativement simpliste. En effet, d'autres variables peuvent expliquer cette concentration, par exemple le vent qui pousse les masses d'air. Ce phénomène physique est connu sous le nom d'advectance (apport d'ozone) ou de dilution. D'autres variables telles le rayonnement, la précipitation, etc., ont une influence certaine sur la concentration d'ozone. L'association Air Breizh mesure ainsi en même temps que la concentration d'ozone d'autres variables susceptibles d'avoir une influence sur celle-ci (voir Annexe D). Voici quelques-unes de ces données :

T_{12}	23.8	16.3	27.2	7.1	25.1	27.5	19.4	19.8	32.2	20.7
V	9.25	-6.15	-4.92	11.57	-6.23	2.76	10.15	13.5	21.27	13.79
N_{12}	5	7	6	5	2	7	4	6	1	4
O_3	115.4	76.8	113.8	81.6	115.4	125	83.6	75.2	136.8	102.8

TABLE 2.1 – 10 données journalières de température, vent, nébulosité et ozone.

La variable V est une variable synthétique. En effet, le vent est normalement mesuré en degrés (direction) et mètres par seconde (vitesse). La variable V que nous avons créée est la projection du vent sur l'axe Est-Ouest, elle tient donc compte à la fois de la direction et de la vitesse.

Pour analyser la relation entre la température T, le vent V, la nébulosité à midi N et l'ozone O_3, nous allons chercher une fonction f telle que :

$$O_{3i} \approx f(T_i, V_i, N_i).$$

Afin de préciser $\approx$, il va falloir définir comme au Chapitre 1 un critère quantifiant la qualité de l'ajustement de la fonction f aux données, ou inversement le coût de non-ajustement. Cette notion de coût permet d'appréhender de manière aisée les problèmes d'ajustement économique dans certains modèles, d'où son nom.

Minimiser un coût nécessite aussi la connaissance de l'espace sur lequel on minimise, c'est-à-dire la classe de fonctions $\mathcal{F}$ dans laquelle nous supposerons que se trouve la vraie fonction inconnue. Le problème mathématique peut s'écrire de la façon suivante :

$$\arg\min_{f \in \mathcal{F}} \sum_{i=1}^{n} L(y_i - f(x_i)), \tag{2.1}$$

où n représente le nombre de données à analyser, $L(.)$ est appelée fonction de coût, ou de perte, et x_i est une variable vectorielle pour tout i. La fonction de coût sera la même que celle utilisée précédemment, c'est-à-dire le coût quadratique. En ce qui concerne le choix de la classe $\mathcal{F}$, par analogie avec le chapitre précédent, nous utiliserons la classe suivante :

$$\mathcal{F} = \left\{ f : \mathbb{R}^P \to \mathbb{R},\ f(x_1, \cdots, x_p) = \sum_{j=1}^{p} \beta_j x_j \right\}.$$

En général, avec cette convention d'écriture, x_1 est constant égal à 1 et β_1 correspond à l'ordonnée à l'origine. On parle de régression linéaire en raison de la linéarité de f en les paramètres $\beta_1, \ldots, \beta_p$, non en les variables explicatives x_j. Par exemple, ce modèle inclut les fonctions polynomiales d'une seule variable x si l'on prend $x_1 = 1$, $x_2 = x, \ldots, x_p = x^{p-1}$.

Ce chapitre est donc la généralisation naturelle du précédent, mais nous allons cette fois manipuler sytématiquement des vecteurs et des matrices à la place des scalaires.

2.1 Modélisation

Le modèle de régression linéaire multiple est une généralisation du modèle de régression simple lorsque les variables explicatives sont en nombre quelconque. Nous supposons donc que les données collectées suivent le modèle suivant :

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i, \qquad i = 1, \ldots, n \tag{2.2}$$

où :
- les x_{ij} sont des nombres connus, non aléatoires, la variable x_{i1} valant souvent 1 pour tout i ;
- les paramètres β_j du modèle sont inconnus, mais non aléatoires ;
- les ε_i sont des variables aléatoires inconnues.

Remarque. Du fait que la constante appartient généralement au modèle, beaucoup d'auteurs écrivent plutôt le modèle sous la forme :

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i, \qquad i = 1, \ldots, n$$

de sorte que p correspond toujours au nombre de variables explicatives. Avec notre convention d'écriture (2.2), si x_{i1} vaut 1 pour tout i, p est le nombre de paramètres à estimer, tandis que le nombre de variables explicatives est, à proprement parler, $(p-1)$.

En utilisant l'écriture matricielle de (2.2) nous obtenons la définition suivante :

Définition 7 (Modèle de régression linéaire multiple) *Un modèle de régression linéaire est défini par une équation de la forme :*

$$Y = X\beta + \boldsymbol{\varepsilon}$$

où :
- *Y est un vecteur aléatoire de dimension n,*
- *X est une matrice de taille $n \times p$ connue, appelée matrice du plan d'expérience,*
- *β est le vecteur de dimension p des paramètres inconnus du modèle,*
- *$\boldsymbol{\varepsilon}$ est le vecteur de dimension n des erreurs.*

Les hypothèses concernant le modèle sont

$$(\mathcal{H}) \begin{cases} (\mathcal{H}_1) : \mathrm{rg}(X) = p \\ (\mathcal{H}_2) : \mathbb{E}[\boldsymbol{\varepsilon}] = 0, \quad \mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2 I_n \end{cases}$$

L'hypothèse $(\mathcal{H}_2)$ signifie que les erreurs sont centrées, de même variance (homoscédasticité) et non corrélées entre elles.

Notation. On notera $X = [X_1|\dots|X_p]$, où X_j est le vecteur de taille n correspondant à la j-ème variable. La i-ème ligne de la matrice X sera quant à elle notée $x_i' = [x_{i1}, \dots, x_{ip}]$. Ainsi l'équation (2.2) s'écrit aussi :

$$\forall i \in \{1, \dots, n\} \qquad y_i = x_i'\beta + \varepsilon_i$$

2.2 Estimateurs des Moindres Carrés Ordinaires

Comme pour la régression linéaire simple, nous allons considérer ici une fonction de coût quadratique, d'où la dénomination de Moindres Carrés Ordinaires (MCO).

Définition 8 (Estimateur des MCO) *L'estimateur des moindres carrés $\hat{\beta}$ est défini comme suit :*

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 = \arg\min_{\beta \in \mathbb{R}^p} \|Y - X\beta\|^2. \tag{2.3}$$

Dans la suite de cette section, nous allons donner l'expression de l'estimateur $\hat{\beta}$ ainsi que certaines de ses propriétés.

2.2.1 Calcul de $\hat{\beta}$

Pour déterminer $\hat{\beta}$, une méthode consiste à se placer dans l'espace des variables, comme on l'a fait au Chapitre 1, Section 1.3.1. Rappelons brièvement le principe : $Y = [y_1, \dots, y_n]'$ est le vecteur des variables à expliquer. La matrice du plan d'expérience $X = [X_1|\dots|X_p]$ est formée de p vecteurs colonnes (la première colonne étant généralement constituée de 1). Le sous-espace de $\mathbb{R}^n$ engendré par les p vecteurs colonnes de X est appelé espace image, ou espace des solutions, et noté $\mathcal{M}(X)$. Il est de dimension p par l'hypothèse $(\mathcal{H}_1)$ et tout vecteur de cet espace est de la forme $X\alpha$, où α est un vecteur de $\mathbb{R}^p$:

$$X\alpha = \alpha_1 X_1 + \dots + \alpha_p X_p$$

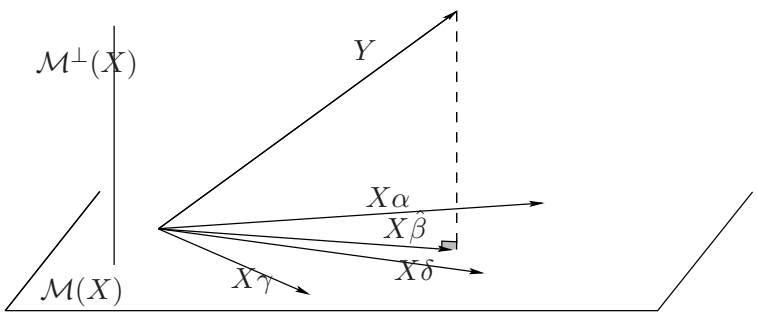


FIGURE 2.1 – Représentation de $X\hat{\beta}$ dans l'espace des variables.

Selon le modèle de la Définition 7, le vecteur Y est la somme d'un élément de $\mathcal{M}(X)$ et d'un bruit élément de $\mathbb{R}^n$, lequel n'a aucune raison d'appartenir à $\mathcal{M}(X)$. Minimiser $\|Y - X\alpha\|^2$ revient à chercher un élément de $\mathcal{M}(X)$ qui soit le plus proche de Y au sens de la norme euclidienne classique. Cet unique élément est, par définition, le projeté orthogonal de Y sur $\mathcal{M}(X)$. Il sera noté $\hat{Y} = P_X Y$, où P_X est la matrice de projection orthogonale sur $\mathcal{M}(X)$. Il peut aussi s'écrire sous la forme $\hat{Y} = X\hat{\beta}$, où $\hat{\beta}$ est l'estimateur des MCO de β. L'espace orthogonal à $\mathcal{M}(X)$, noté $\mathcal{M}^{\perp}(X)$, est souvent appelé espace des résidus. En tant que supplémentaire orthogonal, il est de dimension $n - p = \dim(\mathbb{R}^n) - \dim(\mathcal{M}(X))$.

Proposition 4 (Expression de $\hat{\beta}$) *L'estimateur $\hat{\beta}$ des Moindres Carrés Ordinaires a pour expression :*

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

et la matrice P_X de projection orthogonale sur $\mathcal{M}(X)$ s'écrit :

$$P_X = X(X'X)^{-1}X'.$$

Remarque. L'hypothèse $(\mathcal{H}_1)$ assure que la matrice $X'X$ est bien inversible. Supposons en effet qu'il existe un vecteur β de $\mathbb{R}^p$ tel que $(X'X)\beta = 0$. Ceci impliquerait que $\|X\beta\|^2 = \beta'(X'X)\beta = 0$, donc $X\beta = 0$, d'où $\beta = 0$ puisque $\mathrm{rg}(X) = p$. Autrement dit, la matrice symétrique $X'X$ est définie positive.

Preuve. On peut prouver ce résultat de plusieurs façons.

1. Par différentiation : on cherche $\beta \in \mathbb{R}^p$ qui minimise la fonction

$$S(\beta) = \|Y - X\beta\|^2 = \beta'(X'X)\beta - 2Y'X\beta + \|Y\|^2.$$

   Or S est de type quadratique en β, avec $X'X$ symétrique définie positive, donc le problème admet une unique solution $\hat{\beta}$: c'est le point où le gradient de S est nul. Ceci s'écrit (voir Annexe, section B.5) :

$$\nabla S(\hat{\beta}) = 2\hat{\beta}'X'X - 2Y'X = 0 \iff (X'X)\hat{\beta} = X'Y.$$

   La matrice $X'X$ étant inversible par $(\mathcal{H}_1)$, ceci donne $\hat{\beta} = (X'X)^{-1}X'Y$. Puisque par définition $\hat{Y} = P_X Y = X\hat{\beta} = X(X'X)^{-1}X'Y$ et que cette relation est valable pour tout $Y \in \mathbb{R}^n$, on en déduit que $P_X = X(X'X)^{-1}X'$.

2. Par projection : une autre façon de procéder consiste à dire que le projeté orthogonal $\hat{Y} = X\hat{\beta}$ est défini comme l'unique vecteur tel que $(Y - \hat{Y})$ soit orthogonal à $\mathcal{M}(X)$. Puisque $\mathcal{M}(X)$ est engendré par les vecteurs $X_1, \ldots, X_p$, ceci revient à dire que $(Y - \hat{Y})$ est orthogonal à chacun des X_i :

$$\begin{cases} \langle X_1, Y - X\hat{\beta}\rangle = 0 \\ \qquad\vdots \\ \langle X_p, Y - X\hat{\beta}\rangle = 0 \end{cases}$$

   Ces p équations se regroupent en une seule : $X'(Y - X\hat{\beta}) = 0$, d'où l'on déduit bien l'expression de $\hat{\beta}$, puis celle de P_X.

■

Dorénavant nous noterons $P_X = X(X'X)^{-1}X'$ la matrice de projection orthogonale sur $\mathcal{M}(X)$ et $P_{X^\perp} = (I - P_X)$ la matrice de projection orthogonale sur $\mathcal{M}^\perp(X)$. La décomposition

$$Y = \hat{Y} + (Y - \hat{Y}) = P_X Y + (I - P_X)Y = P_X Y + P_{X^\perp} Y$$

n'est donc rien de plus qu'une décomposition orthogonale de Y sur $\mathcal{M}(X)$ et $\mathcal{M}^\perp(X)$.

Achtung ! La décomposition

$$\hat{Y} = X\hat{\beta} = \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p$$

signifie que les $\hat{\beta}_i$ sont les coordonnées de $\hat{Y}$ dans la base $(X_1, \ldots, X_p)$ de $\mathcal{M}(X)$. Il ne faudrait pas croire pour autant que les $\hat{\beta}_i$ sont les coordonnées des projections de Y sur les X_i : ceci n'est vrai que si la base $(X_1, \ldots, X_p)$ est orthogonale, ce qui n'est pas le cas en général.

Rappels sur les projecteurs. Soit P une matrice carrée de taille n. On dit que P est une matrice de projection si $P^2 = P$. Ce nom est dû au fait que pour tout vecteur x de $\mathbb{R}^n$, Px est la projection de x sur $\mathrm{Im}(P)$ parallèlement à $\mathrm{Ker}(P)$. Si en plus de vérifier $P^2 = P$, la matrice P est symétrique, alors Px est la projection **orthogonale** de x sur $\mathrm{Im}(P)$ parallèlement à $\mathrm{Ker}(P)$, c'est-à-dire que dans la décomposition $x = Px + (x - Px)$, les vecteurs Px et $(x - Px)$ sont orthogonaux. C'est ce cas de figure qui nous concernera dans ce cours. Toute matrice symétrique réelle étant diagonalisable en base orthonormée, il existe une matrice orthogonale U (i.e. $UU' = I_n$, ce qui signifie que les colonnes de U forment une base orthonormée de $\mathbb{R}^n$) et une matrice diagonale Δ telles que $P = U\Delta U'$. On voit alors facilement que la diagonale de Δ est composée de p "1" et de $(n-p)$ "0", où p est la dimension de $\mathrm{Im}(P)$, espace sur lequel on projette. Des rappels et compléments sur les projections sont donnés en Annexe, section B.4.

Revenons à nos moutons : on a vu que $P_X = X(X'X)^{-1}X'$. On vérifie bien que $P_X^2 = P_X$ et que P_X est symétrique. Ce qui précède assure également que $\mathrm{Tr}(P_X) = p$ et $\mathrm{Tr}(P_{X^\perp}) = n - p$. Cette dernière remarque nous sera utile pour construire un estimateur sans biais de σ^2. D'autre part, la matrice P_X est souvent notée H (comme *Hat*) dans la littérature anglo-saxonne, car elle met des chapeaux sur les vecteurs : $P_X Y = \hat{Y}$. De fait, les élements de P_X sont notés $(h_{ij})_{1 \leq i,j \leq n}$.

2.2.2 Quelques propriétés

Comme en régression simple, l'estimateur obtenu est sans biais. On obtient de plus une expression très simple pour sa matrice de covariance $\mathrm{Var}(\hat{\beta})$. On rappelle que la matrice de covariance du vecteur aléatoire $\hat{\beta}$, ou matrice de variance-covariance, ou matrice de dispersion, est par définition :

$$\mathrm{Var}(\hat{\beta}) = \mathbb{E}[(\hat{\beta} - \mathbb{E}[\hat{\beta}])(\hat{\beta} - \mathbb{E}[\hat{\beta}])'] = \mathbb{E}[\hat{\beta}\hat{\beta}'] - \mathbb{E}[\hat{\beta}]\mathbb{E}[\hat{\beta}]'.$$

Puisque β est de dimension p, elle est de dimension $p \times p$. De plus, pour pour toute matrice A de taille $m \times p$ et tout vecteur B de dimension m déterministes, on a : $\mathbb{E}[A\hat{\beta} + B] = A\mathbb{E}[\hat{\beta}] + B$ et $\mathrm{Var}(A\hat{\beta} + B) = A\mathrm{Var}(\hat{\beta})A'$. Ces propriétés élémentaires seront constamment appliquées dans la suite.

Proposition 5 (Biais et matrice de covariance) *L'estimateur $\hat{\beta}$ des moindres carrés est sans biais, i.e. $\mathbb{E}[\hat{\beta}] = \beta$, et sa matrice de covariance est :*

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

Preuve. Pour le biais il suffit d'écrire :

$$\mathbb{E}[\hat{\beta}] = \mathbb{E}[(X'X)^{-1}X'Y] = (X'X)^{-1}X'\mathbb{E}[Y] = (X'X)^{-1}X'\mathbb{E}[X\beta + \boldsymbol{\varepsilon}],$$

et puisque $\mathbb{E}[\boldsymbol{\varepsilon}] = 0$, il vient :

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

Pour la variance, on procède de même :

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

or $\mathrm{Var}(Y) = \mathrm{Var}(X\beta + \boldsymbol{\varepsilon}) = \mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2 I_n$, donc :

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

■

L'estimateur des MCO est optimal en un certain sens. C'est ce que précise le résultat suivant, généralisation de celui vu en régression linéaire simple.

Théorème 5 (Gauss-Markov) *L'estimateur $\hat{\beta}$ des MCO est de variance minimale parmi les estimateurs linéaires sans biais de β.*

Remarques :

1. Linéaire signifie "linéaire par rapport à Y", c'est-à-dire de la forme AY où A est une matrice (p, n) : en ce sens, l'estimateur $\hat{\beta}$ des MCO est bien linéaire puisque $\hat{\beta} = (X'X)^{-1}X'Y$.
2. Rappelons qu'il existe une relation d'ordre partielle entre matrices symétriques réelles : dire que $S_1 \leq S_2$ signifie que $S = (S_2 - S_1)$ est une matrice symétrique réelle positive, c'est-à-dire que pour tout vecteur x, on a $x'S_1x \leq x'S_2x$. Ceci revient encore à dire que les valeurs propres de S sont toutes supérieures ou égales à 0.

Preuve. Nous allons montrer que, pour tout autre estimateur $\tilde{\beta}$ de β linéaire et sans biais, $\mathrm{Var}(\tilde{\beta}) \geq \mathrm{Var}(\hat{\beta})$, où l'inégalité entre matrices de variance-covariance est à comprendre au sens précisé ci-dessus. Rappelons la formule générale pour la matrice de covariance de la somme deux vecteurs aléatoires U et V :

$$\mathrm{Var}(U + V) = \mathrm{Var}(U) + \mathrm{Var}(V) + \mathrm{Cov}(U, V) + \mathrm{Cov}(V, U),$$

où $\mathrm{Cov}(U, V) = \mathbb{E}[UV'] - \mathbb{E}[U]\mathbb{E}[V]' = \mathrm{Cov}(V, U)'$. Décomposons ainsi la variance de $\tilde{\beta}$:

$$\mathrm{Var}(\tilde{\beta}) = \mathrm{Var}(\tilde{\beta} - \hat{\beta} + \hat{\beta}) = \mathrm{Var}(\tilde{\beta} - \hat{\beta}) + \mathrm{Var}(\hat{\beta}) + \mathrm{Cov}(\tilde{\beta} - \hat{\beta}, \hat{\beta}) + \mathrm{Cov}(\hat{\beta}, \tilde{\beta} - \hat{\beta}).$$

Les variances étant semi-définies positives, si nous montrons que $\mathrm{Cov}(\tilde{\beta} - \hat{\beta}, \hat{\beta}) = 0$, nous aurons fini la démonstration. Puisque $\tilde{\beta}$ est linéaire, $\tilde{\beta} = AY$. De plus, nous savons qu'il est sans biais, c'est-à-dire $\mathbb{E}[\tilde{\beta}] = \beta$ pour tout β, donc $AX = I$. La covariance devient :

$$\begin{aligned} \mathrm{Cov}(\tilde{\beta} - \hat{\beta}, \hat{\beta}) &= \mathrm{Cov}(AY, (X'X)^{-1}X'Y) - \mathrm{Var}(\hat{\beta}) \\ &= \sigma^2 AX(X'X)^{-1} - \sigma^2(X'X)^{-1} = 0. \end{aligned}$$

■

2.2.3 Résidus et variance résiduelle

Les résidus sont définis par

$$\hat{\boldsymbol{\varepsilon}} = [\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n]' = Y - \hat{Y} = (I - P_X)Y = P_{X^\perp}Y = P_{X^\perp}\boldsymbol{\varepsilon},$$

car $Y = X\beta + \boldsymbol{\varepsilon}$ et $X\beta \in \mathcal{M}(X)$. On peut alors énoncer les résultats suivants.

Propriétés 4 (Biais et Variance de $\boldsymbol{\varepsilon}$ et $\hat{\boldsymbol{Y}}$) *Sous le jeu d'hypothèses $(\mathcal{H})$, on a :*

1. $\mathbb{E}[\hat{\boldsymbol{\varepsilon}}] = 0$.
2. $\mathrm{Var}(\hat{\boldsymbol{\varepsilon}}) = \sigma^2 P_{X^\perp}$.
3. $\mathbb{E}[\hat{Y}] = X\beta$.
4. $\mathrm{Var}(\hat{Y}) = \sigma^2 P_X$.
5. $\mathrm{Cov}(\hat{\boldsymbol{\varepsilon}}, \hat{Y}) = 0$.

Preuve.

1. $\mathbb{E}[\hat{\boldsymbol{\varepsilon}}] = \mathbb{E}[P_{X^\perp}\boldsymbol{\varepsilon}] = P_{X^\perp}\mathbb{E}[\boldsymbol{\varepsilon}] = 0$.
2. $\mathrm{Var}(\hat{\boldsymbol{\varepsilon}}) = P_{X^\perp}\mathrm{Var}(\boldsymbol{\varepsilon})P'_{X^\perp} = P_{X^\perp}\mathrm{Var}(\boldsymbol{\varepsilon})P_{X^\perp} = \sigma^2 P_{X^\perp}P_{X^\perp} = \sigma^2 P_{X^\perp}$.
3. $\mathbb{E}[\hat{Y}] = \mathbb{E}[X\hat{\beta}] = X\beta$, car $\hat{\beta}$ est sans biais.
4. $\mathrm{Var}(\hat{Y}) = \mathrm{Var}(X\hat{\beta}) = X\mathrm{Var}(\hat{\beta})X' = \sigma^2 X(X'X)^{-1}X' = \sigma^2 P_X$.
5. Rappelons que la covariance entre deux vecteurs aléatoires est une application bilinéaire et que $\mathrm{Cov}(U, U) = \mathrm{Var}(U)$. Ici, ceci donne :

$$\mathrm{Cov}(\hat{\boldsymbol{\varepsilon}}, \hat{Y}) = \mathrm{Cov}(\hat{\boldsymbol{\varepsilon}}, Y - \hat{\boldsymbol{\varepsilon}}) = \mathrm{Cov}(\hat{\boldsymbol{\varepsilon}}, Y) - \mathrm{Var}(\hat{\boldsymbol{\varepsilon}}) = \mathrm{Cov}(P_{X^\perp}Y, Y) - \sigma^2 P_{X^\perp}$$

et puisque $\mathrm{Var}(Y) = \sigma^2 I_n$, nous avons :

$$\mathrm{Cov}(\hat{\boldsymbol{\varepsilon}}, \hat{Y}) = P_{X^\perp}\mathrm{Var}(Y) - \sigma^2 P_{X^\perp} = 0.$$

■

Comme en régression linéaire simple, un estimateur "naturel" de la variance résiduelle est donné par :

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2 = \frac{1}{n}\|\hat{\boldsymbol{\varepsilon}}\|^2.$$

Malheureusement on va voir que cet estimateur est biaisé. Ce biais est néanmoins facile à corriger, comme le montre le résultat suivant. C'est une bête généralisation du résultat obtenu en régression linéaire simple, en remplaçant $n-2$ par $n-p$.

Proposition 6 *La statistique $\hat{\sigma}^2 = \frac{\|\hat{\boldsymbol{\varepsilon}}\|^2}{n-p} = \frac{SCR}{n-p}$ est un estimateur sans biais de σ^2.*

Preuve. Nous calculons $\mathbb{E}[\|\hat{\boldsymbol{\varepsilon}}\|^2]$. Ruse de sioux : puisque c'est un scalaire, il est égal à sa trace, ce qui donne :

$$\mathbb{E}[\|\hat{\boldsymbol{\varepsilon}}\|^2] = \mathbb{E}[\mathrm{Tr}(\|\hat{\boldsymbol{\varepsilon}}\|^2)] = \mathbb{E}[\mathrm{Tr}(\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}})],$$

et puisque pour toute matrice A, on a $\mathrm{Tr}(AA') = \mathrm{Tr}(A'A) = \sum_{i,j} a_{ij}^2$, il vient :

$$\mathbb{E}[\|\hat{\boldsymbol{\varepsilon}}\|^2] = \mathbb{E}[\mathrm{Tr}(\hat{\boldsymbol{\varepsilon}}\hat{\boldsymbol{\varepsilon}}')] = \mathrm{Tr}(\mathbb{E}[\hat{\boldsymbol{\varepsilon}}\hat{\boldsymbol{\varepsilon}}']) = \mathrm{Tr}(\mathrm{Var}(\hat{\boldsymbol{\varepsilon}})) = \mathrm{Tr}(\sigma^2 P_{X^\perp}).$$

Et comme $P_{X^\perp}$ est la matrice de la projection orthogonale sur un espace de dimension $(n-p)$, on a bien :

$$\mathbb{E}[\|\hat{\boldsymbol{\varepsilon}}\|^2] = (n-p)\sigma^2.$$

■

On déduit de cet estimateur de $\hat{\sigma}^2$ de la variance résiduelle σ^2 un estimateur $\hat{\sigma}^2_{\hat{\beta}}$ de la variance $\text{Var}(\hat{\beta}) = \sigma^2(X'X)^{-1}$:

$$\hat{\sigma}^2_{\hat{\beta}} = \hat{\sigma}^2(X'X)^{-1} = \frac{\|\hat{\boldsymbol{\varepsilon}}\|^2}{n-p}(X'X)^{-1} = \frac{SCR}{n-p}(X'X)^{-1}.$$

En particulier, un estimateur de l'écart-type de l'estimateur $\hat{\beta}_j$ du j-ème coefficient de la régression est tout simplement :

$$\hat{\sigma}_{\hat{\beta}_j} = \hat{\sigma}\sqrt{[(X'X)^{-1}]_{jj}}.$$

Afin d'alléger les notations, on écrira parfois $\hat{\sigma}_j$ pour $\hat{\sigma}_{\hat{\beta}_j}$.

2.2.4 Prévision

Un des buts de la régression est de proposer des prédictions pour la variable à expliquer y lorsque nous avons de nouvelles valeurs de x. Soit donc $x'_{n+1} = [x_{n+1,1}, \cdots, x_{n+1,p}]$ une nouvelle valeur pour laquelle nous voudrions prédire y_{n+1}. Cette variable réponse est définie par $y_{n+1} = x'_{n+1}\beta + \varepsilon_{n+1}$, avec $\mathbb{E}[\varepsilon_{n+1}] = 0$, $\text{Var}(\varepsilon_{n+1}) = \sigma^2$ et $\text{Cov}(\varepsilon_{n+1}, \varepsilon_i) = 0$ pour $i = 1, \ldots, n$.

La méthode naturelle est de prédire la valeur correspondante grâce au modèle ajusté, soit : $\hat{y}_{n+1} = x'_{n+1}\hat{\beta}$. L'erreur de prévision est à nouveau définie par $\hat{\varepsilon}_{n+1} = y_{n+1} - \hat{y}_{n+1} = x'_{n+1}(\beta - \hat{\beta}) + \varepsilon_{n+1}$. Deux types d'erreurs vont alors entacher notre prévision : la première due à l'incertitude sur ε_{n+1}, l'autre à l'incertitude inhérente à l'estimateur $\hat{\beta}$.

Proposition 7 (Erreur de prévision) *L'erreur de prévision $\hat{\varepsilon}_{n+1} = (y_{n+1} - \hat{y}_{n+1})$ satisfait les propriétés suivantes :*

$$\begin{cases} \mathbb{E}[\hat{\varepsilon}_{n+1}] = 0 \\ \text{Var}(\hat{\varepsilon}_{n+1}) = \sigma^2(1 + x'_{n+1}(X'X)^{-1}x_{n+1}). \end{cases}$$

Preuve. Comme $\mathbb{E}[\varepsilon_{n+1}] = 0$ et puisque $\hat{\beta}$ est un estimateur sans biais de β, il est clair que

$$\mathbb{E}[\hat{\varepsilon}_{n+1}] = \mathbb{E}[x'_{n+1}(\beta - \hat{\beta}) + \varepsilon_{n+1}] = x'_{n+1}(\beta - \mathbb{E}[\hat{\beta}]) + \mathbb{E}[\varepsilon_{n+1}] = 0.$$

Autrement dit, en moyenne, notre estimateur ne se trompe pas. Calculons la variance de l'erreur de prévision. Puisque $\hat{\beta}$ dépend uniquement des variables aléatoires $(\varepsilon_i)_{1\leq i\leq n}$, dont ε_{n+1} est décorrélée, il vient :

$$\begin{aligned} \text{Var}\,(\hat{\varepsilon}_{n+1}) &= \text{Var}(\varepsilon_{n+1} + x'_{n+1}(\beta - \hat{\beta})) = \sigma^2 + x'_{n+1}\text{Var}(\hat{\beta})x_{n+1} \\ &= \sigma^2(1 + x'_{n+1}(X'X)^{-1}x_{n+1}). \end{aligned}$$

■

Nous retrouvons bien l'incertitude d'observation σ^2 à laquelle vient s'ajouter l'incertitude d'estimation. Enfin, comme en régression linéaire simple, on peut prouver qu'en présence de la constante, cette incertitude est minimale au centre de gravité des variables explicatives, c'est-à-dire lorsque $x'_{n+1} = [1, \bar{x}_2, \ldots, \bar{x}_p]$ et qu'elle vaut encore $\sigma^2(1 + 1/n)$ (voir exercice 2.7).

2.3 Interprétation géométrique

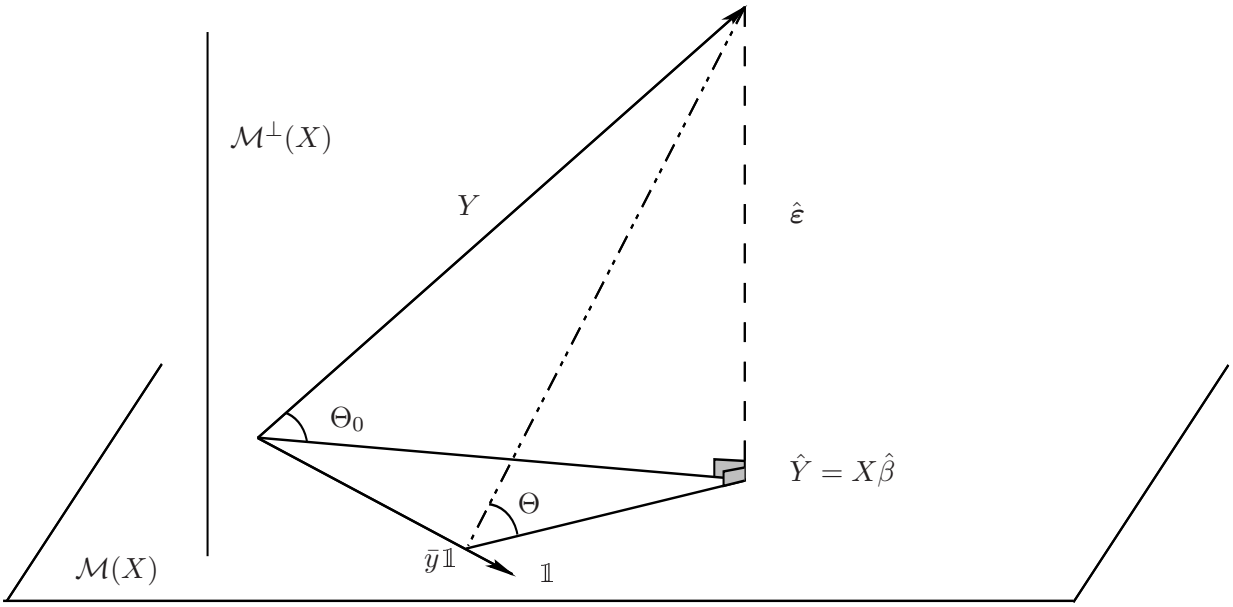


FIGURE 2.2 – Représentation des variables.

À partir de la figure 2.2, le théorème de Pythagore donne :

$$\begin{aligned} SCT &= SCE + SCR \\ \|Y\|^2 &= \|\hat{Y}\|^2 + \|\hat{\varepsilon}\|^2 \\ &= \|X\hat{\beta}\|^2 + \|Y - X\hat{\beta}\|^2. \end{aligned}$$

Si la constante fait partie du modèle (ce qui est généralement le cas), alors nous avons, toujours par Pythagore :

$$\begin{aligned} SCT &= SCE + SCR \\ \|Y - \bar{y}\mathbb{1}\|^2 &= \|\hat{Y} - \bar{y}\mathbb{1}\|^2 + \|\hat{\varepsilon}\|^2 \\ \text{Variation totale} &= \text{V. expliquée par le modèle} + \text{V. résiduelle}. \end{aligned}$$

Définition 9 *Le coefficient de détermination R^2 est défini par :*

$$R^2 = \cos^2\theta_0 = \frac{\|\hat{Y}\|^2}{\|Y\|^2} = 1 - \frac{\|\hat{\varepsilon}\|^2}{\|Y\|^2} = 1 - \frac{SCR}{SCT},$$

ou plus souvent, si la constante fait partie du modèle, par :

$$R^2 = \cos^2\theta = \frac{\textit{V. expliquée par le modèle}}{\textit{Variation totale}} = \frac{\|\hat{Y} - \bar{y}\mathbb{1}\|^2}{\|Y - \bar{y}\mathbb{1}\|^2} = 1 - \frac{\|\hat{\varepsilon}\|^2}{\|Y - \bar{y}\mathbb{1}\|^2} = 1 - \frac{SCR}{SCT}.$$

Ce coefficient mesure le cosinus carré de l'angle entre les vecteurs Y et $\hat{Y}$ pris à l'origine ou pris en $\bar{y}\mathbb{1}$. Néanmoins, on peut lui reprocher de ne pas tenir compte de la dimension de l'espace de projection $\mathcal{M}(X)$, d'où la définition du coefficient de détermination ajusté.

Définition 10 *Le coefficient de détermination ajusté R_a^2 est défini par :*

$$R_a^2 = 1 - \frac{n}{n-p}\frac{\|\hat{\varepsilon}\|^2}{\|Y\|^2} = 1 - \frac{n}{n-p}\frac{SCR}{SCT} = 1 - \frac{n}{n-p}(1-R^2),$$

ou plus souvent, si la constante fait partie du modèle, par :

$$R_a^2 = 1 - \frac{n-1}{n-p}\frac{\|\hat{\varepsilon}\|^2}{\|Y-\bar{y}\mathbb{1}\|^2} = 1 - \frac{n-1}{n-p}\frac{SCR}{SCT} = 1 - \frac{n-1}{n-p}(1-R^2).$$

Avec le logiciel R, le coefficient de détermination R^2 est appelé `Multiple R-Squared`, tandis que le coefficient de détermination ajusté R_a^2 est appelé `Adjusted R-Squared` (cf. infra).

2.4 Exemple

Nous allons traiter les 50 données journalières présentées en Annexe D. La variable à expliquer est la concentration en ozone notée `O3` et les variables explicatives sont la température `T12`, le vent `Vx` et la nébulosité `Ne12`. Les données sont traitées avec le logiciel R.

```
> a <- lm(O3 ~ T12+Vx+Ne12,data=DONNEE)
> summary(a)
Call:
lm(formula = O3 ~ T12 + Vx + Ne12, data = DONNEE)

 Residuals:
                     Min       1Q   Median       3Q      Max
                -29.0441  -8.4833   0.7857   7.7011  28.2919

 Coefficients :
              Estimate  Std. Error  t value  Pr(>|t|)
 (Intercept)   84.5483     13.6065    6.214  1.38e-07      ***
 T12            1.3150      0.4974    2.644   0.01118        *
 Vx             0.4864      0.1675    2.903   0.00565       **
 Ne12          -4.8935      1.0270   -4.765  1.93e-05      ***

--
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 13.91 on 46 degrees of freedom
Multiple R-Squared: 0.6819, Adjusted R-squared: 0.6611
F-statistic: 32.87 on 3 and 46 DF, p-value: 1.663e-11
```

Les interprétations des sorties sont similaires à celles obtenues pour la régression simple. Noter que le `Residual standard error` correspond à l'écart-type résiduel, c'est-à-dire à $\hat{\sigma}$.

2.5 Exercices

Exercice 2.1 (Régression simple et Régression multiple) Soit un échantillon de n couples $(x_i, y_i)_{1\leq i\leq n}$ pour le modèle de régression linéaire simple $y = \beta_1 + \beta_2 x + \varepsilon$.

1. Rappeler les formules de $\hat{\beta}_1$ et $\hat{\beta}_2$ vues au Chapitre 1.
2. Rappeler la formule de $\hat{\beta} = [\hat{\beta}_1, \hat{\beta}_2]'$ vue au Chapitre 2.

3. Retrouver le résultat de la question 1 à partir de celui de la question 2.
4. Rappeler les formules des variances et covariance de $\hat{\beta}_1$ et $\hat{\beta}_2$ vues au Chapitre 1.
5. Rappeler la formule de la matrice de covariance de $\hat{\beta}$ vue au Chapitre 2.
6. Retrouver le résultat de la question 4 à partir de celui de la question 5.

Exercice 2.2 (Rôle de la constante) Soit X une matrice de dimensions $n \times p$. Soit $\hat{Y}$ la projection orthogonale d'un vecteur Y de $\mathbb{R}^n$ sur l'espace engendré par les colonnes de X. On note $\mathbb{1}$ le vecteur de $\mathbb{R}^n$ uniquement composé de la valeur 1.

1. Exprimer le produit scalaire $\langle Y, \mathbb{1} \rangle$ en fonction des y_i.
2. Soit $\hat{\boldsymbol{\varepsilon}} = Y - \hat{Y}$ et supposons que la constante fait partie du modèle, c'est-à-dire que la première colonne de X est $\mathbb{1}$. Que vaut $\langle \hat{\boldsymbol{\varepsilon}}, \mathbb{1} \rangle$?
3. En déduire que lorsque la constante fait partie du modèle, $\sum_{i=1}^n y_i = \sum_{i=1}^n \hat{y}_i$.

Exercice 2.3 (Le R^2 et les modèles emboîtés) Soit Z une matrice (n, q) de rang q et soit X une matrice (n, p) de rang p composée des q vecteurs colonnes de Z et de $p - q$ autres vecteurs linéairement indépendants. Nous considérons les deux modèles suivants :

$$\begin{aligned} Y &= Z\beta + \varepsilon \\ Y &= X\beta^* + \eta \end{aligned}$$

Supposons pour simplifier que la constante ne fait partie d'aucun modèle. Notons respectivement P_X et P_Z les projections orthogonales sur les sous-espaces $\mathcal{M}(X)$ et $\mathcal{M}(Z)$ engendrés par les p colonnes de X et les q colonnes de Z. Notons enfin $P_{X \cap Z^\perp}$ la projection orthogonale sur le sous-espace $\mathcal{M}(X) \cap \mathcal{M}(Z)^\perp$, orthogonal de $\mathcal{M}(Z)$ dans $\mathcal{M}(X)$, autrement dit :

$$\mathbb{R}^n = \mathcal{M}(X) \overset{\perp}{\oplus} \mathcal{M}(X)^\perp = \Big(\mathcal{M}(Z) \overset{\perp}{\oplus} \big(\mathcal{M}(X) \cap \mathcal{M}(Z)^\perp \big) \Big) \overset{\perp}{\oplus} \mathcal{M}(X)^\perp.$$

1. Exprimer $\|P_X Y\|^2$ en fonction de $\|P_Z Y\|^2$ et $\|P_{X \cap Z^\perp} Y\|^2$.
2. Comparer alors les coefficients de détermination des deux modèles, c'est-à-dire R^2_Z et R^2_X.
3. De façon générale, qu'en déduire quant à l'utilisation du R^2 pour le choix de variables ?

Exercice 2.4 (Deux variables explicatives) On examine l'évolution d'une variable réponse y_i en fonction de deux variables explicatives x_i et z_i. Soit $X = (\mathbb{1}\ x\ z)$ la matrice $n \times 3$ du plan d'expérience.

1. Nous avons obtenu les résultats suivants :

$$X'X = \begin{pmatrix} 25 & 0 & 0 \\ ? & 9.3 & 5.4 \\ ? & ? & 12.7 \end{pmatrix} \qquad (X'X)^{-1} = \begin{pmatrix} 0.04 & 0 & 0 \\ 0 & 0.1428 & -0.0607 \\ 0 & -0.0607 & 0.1046 \end{pmatrix}.$$

   (a) Donner les valeurs manquantes.
   (b) Que vaut n ?
   (c) Calculer le coefficient de corrélation linéaire empirique entre x et z.
2. La régression linéaire de Y sur $(\mathbb{1}, x, z)$ donne

$$Y = -1.6\mathbb{1} + 0.61x + 0.46z + \hat{\varepsilon}, \qquad SCR = \|\hat{\varepsilon}\|^2 = 0.3.$$

(a) Déterminez la moyenne empirique $\bar{y}$.

(b) Calculer la somme des carrés expliquée (SCE), la somme des carrés totale (SCT), le coefficient de détermination et le coefficient de détermination ajusté.

Exercice 2.5 (Régression sur variables orthogonales) Nous considérons le modèle de régression linéaire

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

où $Y \in \mathbb{R}^n$, X est une matrice de taille $n \times p$ composée de p vecteurs orthogonaux, $\beta \in \mathbb{R}^p$ et $\boldsymbol{\varepsilon} \in \mathbb{R}^n$. Considérons Z la matrice des q premières colonnes de X et U la matrice des $(p-q)$ dernières colonnes de X. Nous avons obtenu par les MCO les estimations suivantes :

$$\begin{aligned} \hat{Y}_X &= \hat{\beta}_1^X X_1 + \cdots + \hat{\beta}_p^X X_p \\ \hat{Y}_Z &= \hat{\beta}_1^Z X_1 + \cdots + \hat{\beta}_q^Z X_q \\ \hat{Y}_U &= \hat{\beta}_{q+1}^U X_{q+1} + \cdots + \hat{\beta}_p^U X_p. \end{aligned}$$

Notons également $SCE(A)$ la norme au carré de $P_A Y$.

1. Montrer que $SCE(X) = SCE(Z) + SCE(U)$.
2. Donner l'expression de $\hat{\beta}_1^X$ en fonction de Y, X_1 et $\|X_1\|$.
3. En déduire que $\hat{\beta}_1^X = \hat{\beta}_1^Z$.

Exercice 2.6 (Régression sur variables centrées) Nous considérons le modèle de régression linéaire

$$Y = X\beta + \boldsymbol{\varepsilon}, \tag{2.4}$$

où $Y \in \mathbb{R}^n$, X est une matrice de taille $n \times p$ de rang p, $\beta \in \mathbb{R}^p$ et $\boldsymbol{\varepsilon} \in \mathbb{R}^n$. La première colonne de X est le vecteur constant $\mathbb{1}$. X peut donc s'écrire $X = [\mathbb{1}, Z]$ où $Z = [X_2, \ldots, X_p]$ est la matrice $n \times (p-1)$ des $(p-1)$ derniers vecteurs colonnes de X. Le modèle peut donc s'écrire sous la forme :

$$Y = \beta_1 \mathbb{1} + Z\beta_{(1)} + \boldsymbol{\varepsilon},$$

où β_1 est la première coordonnée du vecteur β et $\beta_{(1)}$ représente le vecteur β privé de sa première coordonnée.

1. Donner $P_{\mathbb{1}}$, matrice de projection orthogonale sur le sous-espace engendré par le vecteur $\mathbb{1}$.
2. En déduire la matrice de projection orthogonale $P_{\mathbb{1}^\perp}$ sur le sous-espace $\mathbb{1}^\perp$ orthogonal au vecteur $\mathbb{1}$.
3. Calculer $P_{\mathbb{1}^\perp} Z$.
4. En déduire que l'estimateur de β des Moindres Carrés Ordinaires du modèle (2.4) peut être obtenu en minimisant par les MCO le modèle suivant :

$$\tilde{Y} = \tilde{Z}\beta_{(1)} + \boldsymbol{\eta}, \tag{2.5}$$

où $\tilde{Y} = P_{\mathbb{1}^\perp} Y$ et $\tilde{Z} = P_{\mathbb{1}^\perp} Z$.

5. Ecrire la SCR estimée dans le modèle (2.5) en fonction des variables du modèle (2.5). Vérifier que la SCR du modèle (2.5) est identique à celle qui serait obtenue par l'estimation du modèle (2.4).

Exercice 2.7 (Minimisation de l'erreur de prévision) 1. Soit un échantillon de n couples de réels $(x_i, y_i)_{1\leq i\leq n}$ pour le modèle de régression linéaire simple $y_i = \beta_0+\beta_1 x_i+\varepsilon_i$, où les erreurs ε_i sont supposées centrées décorrélées et de même variance σ^2. On estime $\beta = (\beta_0, \beta_1)$ par la méthode des moindres carrés ordinaires, ce qui donne $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)$.

(a) Soit x_{n+1} une nouvelle valeur de la variable explicative pour laquelle on veut prédire la variable réponse y_{n+1}. Qu'appelle-t-on erreur de prévision ? Rappeler sa variance telle qu'elle est énoncée dans le chapitre sur la régression linéaire simple.

(b) Rappeler sa variance telle qu'elle est énoncée dans le chapitre sur la régression linéaire multiple.

(c) Retrouver le résultat de la question 1a à partir de celui de la question 1b.

(d) A partir du résultat de la question 1a, trouver pour quelle valeur de x_{n+1} la variance de l'erreur de prévision est minimale. Que vaut alors cette variance ?

2. Le but de cette partie est de généraliser le résultat de la question 1d. Nous considérons désormais un échantillon $(x_i', y_i)_{1\leq i\leq n}$, où $x_i' = [1, z_i']$ avec $z_i' = [x_{i1}, \ldots, x_{ip}]$. En notant $\mathbb{1}$ le vecteur de taille n uniquement composé de 1, nous adoptons l'écriture matricielle :

$$X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix} = \left[\begin{array}{cc} 1 & z_1' \\ \hline \vdots & \vdots \\ \hline 1 & z_n' \end{array}\right] = \left[\begin{array}{c|c|c|c} \mathbb{1} & Z_1 & \cdots & Z_p \end{array}\right] = \left[\begin{array}{c|c} \mathbb{1} & Z \end{array}\right],$$

où Z est donc une matrice de taille $n \times p$. Les moyennes de ses colonnes $Z_1, \ldots, Z_p$ sont regroupées dans le vecteur ligne $\bar{x}' = [\bar{x}_1, \ldots, \bar{x}_p]$. Enfin, on considère comme précédemment le modèle de régression linéaire

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = x_i'\beta + \varepsilon_i,$$

où les erreurs ε_i sont supposées centrées indépendantes et de même variance σ^2. Matriciellement, ceci s'écrit donc $Y = X\beta + \boldsymbol{\varepsilon}$, avec X donnée ci-dessus et supposée telle que $X'X$ est inversible.

(a) Ecrire la matrice $X'X$ sous forme de 4 blocs faisant intervenir Z, $\bar{x}$ et la taille n de l'échantillon.

(b) On rappelle la formule d'inversion matricielle par blocs : Soit M une matrice inversible telle que

$$M = \left[\begin{array}{c|c} T & U \\ \hline V & W \end{array}\right]$$

avec T inversible, alors $Q = W - VT^{-1}U$ est inversible et l'inverse de M est :

$$M^{-1} = \left[\begin{array}{c|c} T^{-1} + T^{-1}UQ^{-1}VT^{-1} & -T^{-1}UQ^{-1} \\ \hline -Q^{-1}VT^{-1} & Q^{-1} \end{array}\right].$$

Ecrire la matrice $(X'X)^{-1}$ sous forme de 4 blocs dépendant de n, $\bar{x}$ et Γ^{-1}, où $\Gamma = \frac{1}{n}Z'Z - \bar{x}\bar{x}'$.

(c) Soit $x_{n+1}' = [1, z_{n+1}']$ une nouvelle donnée. Montrer que la variance de l'erreur de prévision est égale à

$$\mathrm{Var}(\hat{\varepsilon}_{n+1}) = \sigma^2\left(1 + \frac{1}{n} + \frac{1}{n}(z_{n+1} - \bar{x})'\Gamma^{-1}(z_{n+1} - \bar{x})\right).$$

(d) On admet pour l'instant que $\Gamma = \frac{1}{n}Z'Z - \bar{x}\bar{x}'$ est symétrique définie positive (on rappelle que S est symétrique définie positive si $S' = S$ et si pour tout vecteur x non nul, $x'Sx > 0$). Pour quelle nouvelle donnée x'_{n+1} la variance de l'erreur de prévision est-elle minimale ? Que vaut alors cette variance ?

(e) Justifier le fait que si $X'X$ est inversible, alors Γ est bien symétrique définie positive.

Exercice 2.8 (QCM) Ce questionnaire fait appel non seulement au cours, mais également à certains des résultats vus dans les exercices qui précèdent.

1. Nous avons effectué une régression multiple, une des variables explicatives est la constante, la somme des résidus calculés vaut :
   A. 0 ;
   B. Approximativement 0 ;
   C. Parfois 0.
2. Le vecteur $\hat{Y}$ est-il orthogonal au vecteur des résidus estimés $\hat{\boldsymbol{\varepsilon}}$?
   A. Oui ;
   B. Non ;
   C. Seulement si $\mathbb{1}$ fait partie des variables explicatives.
3. Un estimateur de la variance de $\hat{\beta}$, estimateur des MC de β, vaut :
   A. $\sigma^2(X'X)^{-1}$;
   B. $\hat{\sigma}^2(X'X)^{-1}$;
   C. $\hat{\sigma}^2(XX')^{-1}$.
4. Une régression a été effectuée et le calcul de la SCR a donné la valeur notée $SCR1$. Une variable est ajoutée, le calcul de la SCR a donné une nouvelle valeur notée $SCR2$. Nous savons que :
   A. SCR1 $\leq$ SCR2 ;
   B. SCR1 $\geq$ SCR2 ;
   C. Cela dépend de la variable ajoutée.
5. Une régression a été effectuée et un estimateur de la variance résiduelle a donné la valeur notée $\hat{\sigma}_1^2$. Une variable est rajoutée et un estimateur de la variance résiduelle vaut maintenant $\hat{\sigma}_2^2$. Nous savons que :
   A. $\hat{\sigma}_1^2 \leq \hat{\sigma}_2^2$;
   B. $\hat{\sigma}_1^2 \geq \hat{\sigma}_2^2$;
   C. On ne peut rien dire.

2.6 Corrigés

Exercice 2.1 (Régression simple et Régression multiple) On dispose donc d'un échantillon de n points $(x_i, y_i)_{1\leq i\leq n}$.

1. On a vu au Chapitre 1 que les estimateurs des MCO ont pour expressions :

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2\bar{x},$$

avec

$$\hat{\beta}_2 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

2. Conformément aux conventions du Chapitre 2, on note X la matrice $n \times 2$ dont la première colonne est uniquement composée de 1 et la seconde est composée des x_i. De même, $Y =$

$[y_1, \ldots, y_n]'$ est un vecteur colonne de taille n. On a vu que l'estimateur $\hat{\beta} = [\hat{\beta}_1, \hat{\beta}_2]'$ des moindres carrés s'écrit alors :

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

3. Les calculs de $(X'X)^{-1}$ et de $X'Y$ donnent :

$$(X'X)^{-1}X'Y = \frac{1}{n\sum x_i^2 - n^2\bar{x}^2}\begin{bmatrix} \sum x_i^2 & -n\bar{x} \\ -n\bar{x} & n \end{bmatrix}\begin{bmatrix} n\bar{y} \\ \sum x_i y_i \end{bmatrix},$$

d'où :

$$(X'X)^{-1}X'Y = \frac{1}{\sum(x_i - \bar{x})^2}\begin{bmatrix} \bar{y}\sum x_i^2 - \bar{x}\sum x_i y_i \\ \sum x_i y_i - n\bar{x}\bar{y} \end{bmatrix}$$

Il suffit alors de voir que

$$\sum x_i y_i - n\bar{x}\bar{y} = \sum(x_i - \bar{x})(y_i - \bar{y})$$

pour vérifier que la seconde composante de ce vecteur correspond bien à la formule de $\hat{\beta}_2$ de la première question. Pour la première composante, on écrit :

$$\frac{\bar{y}\sum x_i^2 - \bar{x}\sum x_i y_i}{\sum(x_i - \bar{x})^2} = \frac{\bar{y}\sum(x_i - \bar{x})^2 - \bar{x}(\sum x_i y_i - n\bar{x}\bar{y})}{\sum(x_i - \bar{x})^2} = \bar{y} - \hat{\beta}_2\bar{x}$$

et la messe est dite.

4. Les formules des variances de $\hat{\beta}_1$ et $\hat{\beta}_2$ vues au Chapitre 1 sont

$$\mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2\sum x_i^2}{n\sum(x_i - \bar{x})^2} \qquad \& \qquad \mathrm{Var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum(x_i - \bar{x})^2},$$

tandis que leur covariance vaut :

$$\mathrm{Cov}(\hat{\beta}_1, \hat{\beta}_2) = -\frac{\sigma^2\bar{x}}{\sum(x_i - \bar{x})^2}.$$

5. La matrice de covariance de $\hat{\beta}$ est tout bonnement

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

6. Pour retrouver le résultat de la question 4 à partir de celui de la question 5, il suffit de voir que

$$(X'X)^{-1} = \frac{1}{\sum(x_i - \bar{x})^2}\begin{bmatrix} \sum x_i^2/n & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix}.$$

Exercice 2.2 (Rôle de la constante) Soit $X_{(n,p)}$ une matrice de rang p. Soit $\hat{Y}$ la projection d'un vecteur Y de $\mathbb{R}^n$ sur l'espace engendré par les colonnes de X. On note $\mathbb{1}$ le vecteur de $\mathbb{R}^n$ uniquement composé de 1.

1. Par définition du produit scalaire usuel dans $\mathbb{R}^n$, on a tout simplement :

$$\langle Y, \mathbb{1}\rangle = \sum y_i$$

2. Puisque $\hat{Y}$ est la projection orthogonale de Y sur le sous-espace engendré par les colonnes de X, le vecteur $\hat{\boldsymbol{\varepsilon}} = Y - \hat{Y}$ est orthogonal à toutes les colonnes de X. En particulier, si l'une d'entre elles est constante et vaut $c\mathbb{1}$ (c supposé non nul), on en déduit que :

$$\langle\hat{\boldsymbol{\varepsilon}}, c\mathbb{1}\rangle = 0 \Rightarrow \ \langle\hat{\boldsymbol{\varepsilon}}, \mathbb{1}\rangle = 0.$$

Autrement dit, lorsque la constante fait partie du modèle, la somme des résidus vaut 0.

3. Dire que la constante fait partie du modèle signifie typiquement que la première colonne de X est le vecteur $\mathbb{1}$. D'après la question précédente, on sait que dans ce cas :

$$\langle \hat{\boldsymbol{\varepsilon}}, \mathbb{1} \rangle = 0 \;\Leftrightarrow\; \sum y_i = \sum \hat{y}_i.$$

Ainsi, lorsque la constante fait partie du modèle, la moyenne des observations y_i est la même que celle de leurs valeurs ajustées.

Exercice 2.3 (Le R^2 et les modèles emboîtés)

1. Par le théorème de Pythagore, on a :

$$\|P_X Y\|^2 = \|P_Z Y\|^2 + \|P_{X \cap Z^\perp} Y\|^2.$$

2. Si la constante ne fait partie d'aucun modèle, alors dans le premier modèle, le R^2 vaut :

$$R^2_Z = \frac{\|P_Z Y\|^2}{\|Y\|^2},$$

et dans le second :

$$R^2_X = \frac{\|P_X Y\|^2}{\|Y\|^2} = \frac{\|P_Z Y\|^2 + \|P_{X \cap Z^\perp} Y\|^2}{\|Y\|^2} \geq \frac{\|P_Z Y\|^2}{\|Y\|^2} = R^2_Z.$$

3. Ceci montre la chose suivante : dès lors que deux modèles sont emboîtés, le coefficient de détermination du plus gros sera supérieur à celui du plus petit. Autrement dit, dès que l'on ajoute une ou des variables à un modèle, on améliore le pourcentage de variation expliquée, même si les variables explicatives supplémentaires ne sont pas pertinentes ! En ce sens, le coefficient de détermination ajusté est préférable, ayant au moins le mérite de tenir compte des dimensions des différents modèles. Plus précisément, nous verrons au Chapitre 3 comment effectuer des tests d'hypothèses entre modèles emboîtés.

Exercice 2.4 (Deux variables explicatives) On examine l'évolution d'une variable y en fonction de deux variables exogènes x et z. On dispose de n observations de ces variables. On note $X = [\mathbb{1}\ \boldsymbol{x}\ \boldsymbol{z}]$ où $\mathbb{1}$ est le vecteur constant et $\boldsymbol{x}$, $\boldsymbol{z}$ sont les vecteurs des variables explicatives.

1. Nous avons obtenu les résultats suivants :

$$X'X = \begin{bmatrix} 25 & 0 & 0 \\ ? & 9.3 & 5.4 \\ ? & ? & 12.7 \end{bmatrix} \qquad (X'X)^{-1} = \begin{bmatrix} 0.04 & 0 & 0 \\ 0 & 0.1428 & -0.0607 \\ 0 & -0.0607 & 0.1046 \end{bmatrix}.$$

   (a) Les 3 valeurs manquantes se déduisent de la symétrie de la matrice $X'X$.

   (b) Puisque $X = [\mathbb{1}\ \boldsymbol{x}\ \boldsymbol{z}]$, il vient $n = (X'X)_{1,1} = 25$.

   (c) Le coefficient de corrélation linéaire empirique entre x et z se déduit lui aussi de la matrice $X'X$. On remarque tout d'abord que les moyennes empiriques sont nulles puisque

$$\bar{x} = \frac{(X'X)_{1,2}}{n} = 0 = \frac{(X'X)_{1,3}}{n} = \bar{z}$$

Par conséquent

$$r_{x,z} = \frac{\sum (x_i - \bar{x})(z_i - \bar{z})}{\sqrt{(x_i - \bar{x})^2}\sqrt{(z_i - \bar{z})^2}} = \frac{\sum x_i z_i}{\sqrt{\sum x_i^2}\sqrt{\sum z_i^2}} = \frac{(X'X)_{2,3}}{\sqrt{(X'X)_{2,2}}\sqrt{(X'X)_{3,3}}}$$

ce qui donne

$$r_{x,z} = \frac{5.4}{\sqrt{9.3}\sqrt{12.7}} \approx 0.5$$

2. La régression linéaire de Y sur $(\mathbb{1}, \boldsymbol{x}, \boldsymbol{z})$ donne

$$Y = -1.6\mathbb{1} + 0.61\boldsymbol{x} + 0.46\boldsymbol{z} + \hat{\boldsymbol{\varepsilon}}, \qquad SCR = \|\hat{\boldsymbol{\varepsilon}}\|^2 = 0.3.$$

(a) Puisque la constante fait partie du modèle, la moyenne empirique des résidus est nulle : $\bar{\hat{\varepsilon}} = 0$. On en déduit que

$$\bar{y} = -1.6 + 0.61\bar{x} + 0.46\bar{z} + \bar{\hat{\varepsilon}} = -1.6$$

(b) Puisque la constante fait partie du modèle, la somme des carrés expliquée par le modèle est

$$SCE = \|\hat{Y} - \bar{y}\mathbb{1}\|^2 = \sum(\hat{y}_i - \bar{y})^2 = \sum(0.61x_i + 0.46z_i)^2$$

c'est-à-dire

$$SCE = \|\hat{Y} - \bar{y}\mathbb{1}\|^2 = 0.61^2 \sum x_i^2 + 2 \times 0.61 \times 0.46 \sum x_i z_i + 0.46^2 \sum z_i^2$$

ce qui se calcule à nouveau grâce à la matrice $X'X$:

$$SCE = \|\hat{Y} - \bar{y}\mathbb{1}\|^2 = 0.61^2 (X'X)_{2,2} + 2 \times 0.61 \times 0.46 (X'X)_{2,3} + 0.46^2 (X'X)_{3,3} = 9.18$$

La somme des carrés totale est alors immédiate, en vertu de la sacro-sainte formule de décomposition de la variance :

$$SCT = SCE + SCR = 9.18 + 0.3 = 9.48$$

Le coefficient de détermination vaut donc

$$R^2 = \frac{SCE}{SCT} \approx 0.968$$

Autrement dit, 97% de la variance des données est expliquée par ce modèle de régression. Le coefficient de détermination ajusté est à peine différent :

$$R_a^2 = 1 - \frac{n-1}{n-p}(1 - R^2) \approx 0.965$$

et on vérifie bien la relation générale selon laquelle $R_a^2 < R^2$.

Exercice 2.5 (Régression sur variables orthogonales) Nous considérons le modèle de régression linéaire

$$Y = X\beta + \varepsilon,$$

où $Y \in \mathbb{R}^n$, X est une matrice de taille $n \times p$ composée de p vecteurs orthogonaux, $\beta \in \mathbb{R}^p$ et $\varepsilon \in \mathbb{R}^n$. Considérons Z la matrice des q premières colonnes de X et U la matrice des $p-q$ dernières colonnes de X. Nous avons obtenu par les MCO les estimations suivantes :

$$\begin{aligned} \hat{Y}_X &= \hat{\beta}_1^X X_1 + \cdots + \hat{\beta}_p^X X_p \\ \hat{Y}_Z &= \hat{\beta}_1^Z X_1 + \cdots + \hat{\beta}_q^Z X_q \\ \hat{Y}_U &= \hat{\beta}_{q+1}^U X_{q+1} + \cdots + \hat{\beta}_p^U X_p. \end{aligned}$$

Notons également $SCE(A)$ la norme au carré de $P_A Y$.

1. Nous avons :
$$\hat{Y}_X = P_X Y = (P_Z + P_{Z^\perp})P_X Y = P_Z P_X Y + P_{Z^\perp} P_X Y,$$
or d'une part $P_Z P_X = P_{Z\cap X} = P_Z$, d'autre part
$$P_{Z^\perp} P_X = P_{Z^\perp \cap X} = P_U$$
projection orthogonale sur le sous-espace engendré par les colonnes de U puisque les colonnes de X sont orthogonales. Au total, on obtient la décomposition orthogonale $Y_X = Y_Z + Y_U$ et le théorème de Pythagore assure donc que $SCE(X) = SCE(Z) + SCE(U)$.
2. Pour l'expression de $\hat{\beta}_1^X$, on part tout simplement de la formule générale
$$\hat{\beta}^X = (X'X)^{-1}X'Y$$
Puisque les colonnes de X sont orthogonales, la matrice $X'X$ est diagonale, de termes diagonaux $\|X_i\|^2$. Par ailleurs, $X'Y$ est un vecteur colonne de taille p, dont les coordonnées sont les produits scalaires $X_i'Y = \langle X_i, Y\rangle$. Ainsi
$$\hat{\beta}^X = \left[\frac{\langle X_1, Y\rangle}{\|X_1\|^2}, \ldots, \frac{\langle X_p, Y\rangle}{\|X_p\|^2}\right]' \Rightarrow \hat{\beta}_1^X = \frac{\langle X_1, Y\rangle}{\|X_1\|^2}.$$
3. La première colonne de Z étant X_1, le raisonnement précédent appliqué à $\hat{\beta}_1^X$ montre que $\hat{\beta}_1^X = \hat{\beta}_1^Z$. Ainsi, lorsque les variables explicatives sont orthogonales, effectuer une régression multiple revient à effectuer p régression simples. En pratique, néanmoins, il arrive rarement que les variables explicatives soient effectivement orthogonales...

Exercice 2.6 (Régression sur variables centrées) Nous considérons le modèle de régression linéaire
$$Y = X\beta + \boldsymbol{\varepsilon}, \tag{2.6}$$
où $Y \in \mathbb{R}^n$, X est une matrice de taille $n \times p$ de rang p, $\beta \in \mathbb{R}^p$ et $\boldsymbol{\varepsilon} \in \mathbb{R}^n$. La première colonne de X est le vecteur constant $\mathbb{1}$. X peut ainsi s'écrire $X = [\mathbb{1}, Z]$, où $Z = [X_2, \ldots, X_p]$ est la matrice $n \times (p-1)$ des $(p-1)$ derniers vecteurs colonnes de X. Le modèle peut donc s'écrire sous la forme :
$$Y = \beta_1 \mathbb{1} + Z\beta_{(1)} + \boldsymbol{\varepsilon},$$
où β_1 est la première coordonnée du vecteur β et $\beta_{(1)}$ représente le vecteur β privé de sa première coordonnée.

1. La matrice de la projection orthogonale sur le sous-espace engendré par le vecteur $\mathbb{1}$ s'écrit
$$P_{\mathbb{1}} = \mathbb{1}(\mathbb{1}'\mathbb{1})^{-1}\mathbb{1}' = \frac{1}{n}\mathbb{1}\mathbb{1}' = \frac{1}{n}J,$$
où $J = \mathbb{1}\mathbb{1}'$ est la matrice $n \times n$ composée uniquement de 1.
2. La matrice de projection orthogonale $P_{\mathbb{1}^\perp}$ sur le sous-espace $\mathbb{1}^\perp$ orthogonal au vecteur $\mathbb{1}$ est donc : $P_{\mathbb{1}^\perp} = I - \frac{1}{n}J$.
3. On a ainsi $P_{\mathbb{1}^\perp} Z = Z - \frac{1}{n}JZ$. Si on note $\bar{x}_2, \ldots, \bar{x}_n$ les moyennes empiriques des colonnes $X_2, \ldots, X_n$, $P_{\mathbb{1}^\perp} Z$ est donc la matrice $n \times (p-1)$ dont les colonnes sont $X_2 - \bar{x}_2\mathbb{1}, \ldots, X_n - \bar{x}_n\mathbb{1}$. Autrement dit $P_{\mathbb{1}^\perp} Z$ est la matrice (individus $\times$ variables) pour laquelle chaque variable x_i a été centrée.

4. L'estimateur de β des Moindres Carrés Ordinaires du modèle (2.6) est défini par
$$\hat{\beta} = \arg\min_{\beta\in\mathbb{R}^p} \|Y - X\beta\|^2.$$
En décomposant le vecteur β sous la forme $\beta = [\beta_1, \beta'_{(1)}]'$, ceci peut encore s'écrire :
$$(\beta_1, \beta_{(1)}) = \arg\min_{\beta_1\in\mathbb{R},\beta_{(1)}\in\mathbb{R}^{p-1}} \|Y - \beta_1\mathbb{1} - Z\beta_{(1)}\|^2.$$
Puisque $P_{\mathbb{1}} + P_{\mathbb{1}^\perp} = I_n$, il vient :
$$(\beta_1, \beta_{(1)}) = \arg\min_{\beta_1\in\mathbb{R},\beta_{(1)}\in\mathbb{R}^{p-1}} \|(P_{\mathbb{1}}Y - \beta_1\mathbb{1} - P_{\mathbb{1}}Z\beta_{(1)}) + (P_{\mathbb{1}^\perp}Y - P_{\mathbb{1}^\perp}Z\beta_{(1)})\|^2.$$
Le premier vecteur entre parenthèses est dans le sous-espace engendré par le vecteur $\mathbb{1}$, le second dans son orthogonal, donc par Pythagore :
$$(\beta_1, \beta_{(1)}) = \arg\min_{\beta_1\in\mathbb{R},\beta_{(1)}\in\mathbb{R}^{p-1}} \|P_{\mathbb{1}}Y - \beta_1\mathbb{1} - P_{\mathbb{1}}Z\beta_{(1)}\|^2 + \|P_{\mathbb{1}^\perp}Y - P_{\mathbb{1}^\perp}Z\beta_{(1)}\|^2.$$
Or $P_{\mathbb{1}}Y = \bar{y}\mathbb{1}$, $P_{\mathbb{1}^\perp}Y = Y - \bar{y}\mathbb{1} = \tilde{Y}$ et $P_{\mathbb{1}^\perp}Z = \tilde{Z}$, donc ceci se réécrit :
$$(\beta_1, \beta_{(1)}) = \arg\min_{\beta_1\in\mathbb{R},\beta_{(1)}\in\mathbb{R}^{p-1}} \|\bar{y}\mathbb{1} - \beta_1\mathbb{1} - \overline{Z\beta_{(1)}}\mathbb{1}\|^2 + \|\tilde{Y} - \tilde{Z}\beta_{(1)}\|^2.$$
Minimiser cette somme de deux termes en $(\beta_1, \beta_{(1)})$ revient à commencer par minimiser le second terme en $\beta_{(1)}$, ce qui fournit $\hat{\beta}_{(1)}$, et à prendre ensuite
$$\hat{\beta}_1 = \bar{y} - \overline{Z\hat{\beta}_{(1)}}.$$
Or la minimisation du premier terme revient à chercher l'estimateur des moindres carrés ordinaires pour le modèle suivant :
$$\tilde{Y} = \tilde{Z}\beta_{(1)} + \boldsymbol{\eta}, \tag{2.7}$$
où $\tilde{Y} = P_{\mathbb{1}^\perp}Y$ et $\tilde{Z} = P_{\mathbb{1}^\perp}Z$.
5. La SCR estimée dans le modèle (2.7) est
$$SCR = \|\tilde{Y} - \hat{\tilde{Y}}\|^2 = \sum_{i=1}^{n}(\tilde{y}_i - \hat{\tilde{y}}_i)^2.$$
Or pour tout i, $\tilde{y}_i = y_i - \bar{y}$ et :
$$\hat{\tilde{y}}_i = \hat{\beta}_2\tilde{x}_{i2} + \cdots + \hat{\beta}_p\tilde{x}_{ip} = \hat{\beta}_2(x_{i2} - \bar{x}_2) + \cdots + \hat{\beta}_p(x_{ip} - \bar{x}_p),$$
d'où :
$$SCR = \sum_{i=1}^{n}(y_i - \bar{y} - \hat{\beta}_2(x_{i2} - \bar{x}_2) - \cdots - \hat{\beta}_p(x_{ip} - \bar{x}_p))^2.$$
Lorsque la constante appartient au modèle, la somme des résidus est nulle, donc :
$$\bar{y} = \bar{\hat{Y}} = \hat{\beta}_1 + \hat{\beta}_2\bar{x}_2 + \cdots + \hat{\beta}_p\bar{x}_p,$$
ce qui, reporté dans l'équation précédente, donne :
$$SCR = \sum_{i=1}^{n}(y_i - (\hat{\beta}_1 + \hat{\beta}_2x_{i2} + \cdots + \hat{\beta}_px_{ip})^2.$$
Autrement dit, la SCR du modèle (2.7) est identique à celle qui serait obtenue par l'estimation du modèle (2.6). Mazel tov !

Exercice 2.7 (Minimisation de l'erreur de prévision) Cet exercice est corrigé en annexe (sujet de décembre 2012).

Exercice 2.8 (QCM) AABBC.

Chapitre 3

Le modèle gaussien

Introduction

Rappelons le contexte du chapitre précédent. Nous avons supposé un modèle de la forme :

$$y_i = x_i'\beta + \varepsilon_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i, \qquad i = 1, \ldots, n$$

que nous avons réécrit en termes matriciels :

$$Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \boldsymbol{\varepsilon}_{n\times 1}$$

où les dimensions sont indiquées en indices. Les hypothèses concernant le modèle étaient :

$$(\mathcal{H}) \begin{cases} (\mathcal{H}_1) : \mathrm{rg}(X) = p \\ (\mathcal{H}_2) : \mathbb{E}[\boldsymbol{\varepsilon}] = 0, \quad \mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2 I_n \end{cases}$$

Dans tout ce chapitre, comme ce fut le cas en fin de Chapitre 1, nous allons faire une hypothèse plus forte, à savoir celle de gaussianité des résidus. Nous supposerons donc désormais :

$$(\mathcal{H}) \begin{cases} (\mathcal{H}_1) : \mathrm{rg}(X) = p \\ (\mathcal{H}_2) : \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_n) \end{cases}$$

Ceci signifie que les résidus sont indépendants et identiquement distribués. L'intérêt de supposer la gaussianité des résidus est de pouvoir en déduire les lois de nos estimateurs, donc de construire des régions de confiance et des tests d'hypothèses.

3.1 Estimateurs du Maximum de Vraisemblance

Nous allons commencer par faire le lien entre l'estimateur du maximum de vraisemblance et l'estimateur des moindres carrés vu au chapitre précédent. Commençons par remarquer que les y_i sont eux-mêmes gaussiens :

$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2) \;\Rightarrow\; y_i = x_i'\beta + \varepsilon_i \sim \mathcal{N}(x_i'\beta, \sigma^2)$$

et mutuellement indépendants puisque les erreurs ε_i le sont. La vraisemblance s'en déduit :

$$\begin{aligned} \mathcal{L}(Y, \beta, \sigma^2) &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^n \left(y_i - x_i'\beta\right)^2\right] \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left[-\frac{1}{2\sigma^2}\|Y - X\beta\|^2\right] \end{aligned}$$

D'où l'on déduit la log-vraisemblance :

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

On cherche les estimateurs $\hat{\beta}_{mv}$ et $\hat{\sigma}^2_{mv}$ qui maximisent cette log-vraisemblance. Il est clair qu'il faut minimiser la quantité $\|Y - X\beta\|^2$, ce qui est justement le principe des moindres carrés ordinaires, donc :

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

Une fois ceci fait, on veut maximiser sur $\mathbb{R}_+^*$ une fonction de la forme $\varphi(x) = a + b\log x + \frac{c}{x}$, ce qui ne pose aucun souci en passant par la dérivée :

$$\frac{\partial \log \mathcal{L}(Y, \hat{\beta}, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\|Y - X\hat{\beta}\|^2,$$

d'où il vient :

$$\hat{\sigma}^2_{mv} = \frac{\|Y - X\hat{\beta}\|^2}{n}.$$

Si l'on compare à ce qu'on a obtenu au chapitre précédent, où nous avons noté $\hat{\sigma}^2 = \frac{\|Y - X\hat{\beta}\|^2}{n-p}$ l'estimateur de la variance σ^2, nous avons donc :

$$\hat{\sigma}^2_{mv} = \frac{n-p}{n}\hat{\sigma}^2.$$

On voit donc que l'estimateur $\hat{\sigma}^2_{mv}$ du maximum de vraisemblance est biaisé, mais d'autant moins que le nombre de variables explicatives est petit devant le nombre n d'observations. Dans la suite, nous continuerons à considérer l'estimateur $\hat{\sigma}^2$ des moindres carrés vu au chapitre précédent et nous conserverons aussi la notation adoptée pour les résidus $\hat{\varepsilon}_i$, de sorte que :

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n \hat{\varepsilon}_i^2}{n-p} = \frac{\|\hat{\boldsymbol{\varepsilon}}\|^2}{n-p} = \frac{\|Y - X\hat{\beta}\|^2}{n-p}.$$

3.2 Lois des estimateurs

Nous commençons cette section par un rappel sur les vecteurs gaussiens.

3.2.1 Quelques rappels

Un vecteur aléatoire Y de $\mathbb{R}^n$ est dit gaussien si toute combinaison linéaire de ses composantes est une variable aléatoire gaussienne. Ce vecteur admet alors une espérance $\mu = \mathbb{E}[Y]$ et une matrice de variance-covariance $\Sigma_Y = \mathbb{E}[(Y-\mu)(Y-\mu)']$ qui caractérisent complètement sa loi. On note dans ce cas $Y \sim \mathcal{N}(\mu, \Sigma_Y)$. On montre alors que les composantes d'un vecteur gaussien $Y = [Y_1, \cdots, Y_n]'$ sont indépendantes si et seulement si Σ_Y est diagonale.

Soit $Y \sim \mathcal{N}(\mu, \Sigma_Y)$ un vecteur gaussien. Il admet une densité f sur $\mathbb{R}^n$ si et seulement si sa matrice de dispersion Σ_Y est inversible, auquel cas :

$$f(y) = \frac{1}{(2\pi)^{n/2}\sqrt{\det(\Sigma_Y)}} e^{-\frac{1}{2}(y-\mu)'\Sigma_Y^{-1}(y-\mu)}$$

Dans ce cas, on montre aussi la propriété suivante.

Proposition 8 (Vecteur gaussien et Loi du χ^2) *Soit $Y \sim \mathcal{N}(\mu, \Sigma_Y)$ un vecteur gaussien. Si Σ_Y est inversible, alors*

$$(Y-\mu)'\Sigma_Y^{-1}(Y-\mu) \sim \chi^2_n$$

loi du chi-deux à n degrés de liberté.

Le théorème de Cochran, très utile dans la suite, assure que la décomposition d'un vecteur gaussien sur des sous-espaces orthogonaux donne des variables indépendantes dont on peut expliciter les lois.

Théorème 6 (Cochran) *Soit $Y \sim \mathcal{N}(\mu, \sigma^2 I_n)$, $\mathcal{M}$ un sous-espace de $\mathbb{R}^n$ de dimension p, P la matrice de projection orthogonale sur $\mathcal{M}$ et $P_\perp = I_n - P$ la matrice de projection orthogonale sur $\mathcal{M}^\perp$. Nous avons les propriétés suivantes :*

(i) $PY \sim \mathcal{N}(P\mu, \sigma^2 P)$ et $P_\perp Y \sim \mathcal{N}(P_\perp \mu, \sigma^2 P_\perp)$;
(ii) les vecteurs PY et $P_\perp Y = (Y - PY)$ sont indépendants ;
(iii) $\frac{\|P(Y-\mu)\|^2}{\sigma^2} \sim \chi^2_p$ et $\frac{\|P_\perp(Y-\mu)\|^2}{\sigma^2} \sim \chi^2_{n-p}$.

Nous pouvons appliquer ce résultat dans notre cadre, comme nous allons le voir en section suivante.

3.2.2 Nouvelles propriétés

Notons au préalable que, pour ce qui nous concerne, la gaussianité des résidus implique celle du vecteur Y :

$$\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_n) \;\Rightarrow Y = X\beta + \boldsymbol{\varepsilon} \sim \mathcal{N}(X\beta, \sigma^2 I_n).$$

Propriétés 5 (Lois des estimateurs avec variance connue) *Sous les hypothèses ($\mathcal{H}$), nous avons :*

(i) $\hat\beta$ est un vecteur gaussien de moyenne β et de variance $\sigma^2(X'X)^{-1}$: $\hat\beta \sim \mathcal{N}(\beta, \sigma^2(X'X)^{-1})$;
(ii) $\hat\beta$ et $\hat\sigma^2$ sont indépendants ;
(iii) $(n-p)\frac{\hat\sigma^2}{\sigma^2} \sim \chi^2_{n-p}$.

Preuve.

(i) Nous avons vu que $\hat\beta = (X'X)^{-1}X'Y = (X'X)^{-1}X'(X\beta + \boldsymbol{\varepsilon})$, or par hypothèse $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_n)$ est un vecteur gaussien. On en déduit que $\hat\beta$ est lui aussi un vecteur gaussien, sa loi est donc entièrement caractérisée par la donnée de sa moyenne et de sa matrice de dispersion, lesquelles ont été calculées au Chapitre 2 (Proposition 5).
(ii) Comme dans le chapitre précédent, notons $\mathcal{M}(X)$ le sous-espace de $\mathbb{R}^n$ engendré par les colonnes de X et $P_X = X(X'X)^{-1}X'$ la projection orthogonale sur ce sous-espace. On peut noter que :

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

donc $\hat\beta$ est un vecteur aléatoire fonction de P_XY, tandis que :

$$\hat\sigma^2 = \frac{\|\hat{\boldsymbol{\varepsilon}}\|^2}{n-p} = \frac{\|Y - P_XY\|^2}{n-p}$$

est une variable aléatoire fonction de $(Y - P_XY)$. Par le théorème de Cochran, nous savons que les vecteurs P_XY et $(Y - P_XY)$ sont indépendants, il en va donc de même pour toutes fonctions de l'un et de l'autre.

(*iii*) En notant $P_{X^\perp}$ la projection orthogonale sur $\mathcal{M}^\perp(X)$, sous-espace de dimension $(n-p)$ de $\mathbb{R}^n$, on a :

$$\hat{\boldsymbol{\varepsilon}} = (Y - P_X Y) = P_{X^\perp} Y = P_{X^\perp}(X\beta + \boldsymbol{\varepsilon}) = P_{X^\perp}\boldsymbol{\varepsilon},$$

où $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_n)$. Il s'ensuit par le théorème de Cochran que :

$$(n-p)\frac{\hat{\sigma}^2}{\sigma^2} = \frac{\|P_{X^\perp}\boldsymbol{\varepsilon}\|^2}{\sigma^2} = \frac{\|P_{X^\perp}(\boldsymbol{\varepsilon} - \mathbb{E}[\boldsymbol{\varepsilon}])\|^2}{\sigma^2} \sim \chi^2_{n-p}.$$

■

Bien entendu le premier point du résultat précédent n'est pas satisfaisant pour obtenir des régions de confiance sur β car il suppose la variance σ^2 connue, ce qui n'est pas le cas en général. La proposition suivante pallie cette insuffisance.

Propriétés 6 (Lois des estimateurs avec variance inconnue) *Sous les hypothèses (*$\mathcal{H}$*) :*

(i) pour $j = 1, \ldots, p$*, nous avons* $T_j = \dfrac{\hat{\beta}_j - \beta_j}{\hat{\sigma}\sqrt{[(X'X)^{-1}]_{jj}}} = \dfrac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}} \sim \mathcal{T}_{n-p}$.

(ii) Soit R *une matrice de taille* $q \times p$ *de rang* q $(q \leq p)$ *alors :*

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

Cautious ! L'écriture $\left[(X'X)^{-1}\right]_{jj}$ signifie "le j-ème terme diagonal de la matrice $(X'X)^{-1}$", et non "l'inverse du j-ème terme diagonal de la matrice $(X'X)$". Afin d'alléger les écritures, nous écrirons souvent $(X'X)^{-1}_{jj}$ au lieu de $\left[(X'X)^{-1}\right]_{jj}$.

Preuve.

(*i*) D'après la proposition précédente, on sait d'une part que $\hat{\beta}_j \sim \mathcal{N}(\beta_j, \sigma^2(X'X)^{-1}_{jj})$, d'autre part que $(n-p)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p}$ et enfin que $\hat{\beta}_j$ et $\hat{\sigma}^2$ sont indépendants. Il reste alors à écrire T_j sous la forme :

$$T_j = \frac{\frac{\hat{\beta}_j - \beta_j}{\sigma\sqrt{(X'X)^{-1}_{jj}}}}{\frac{\hat{\sigma}}{\sigma}}$$

pour reconnaître une loi de Student $\mathcal{T}_{n-p}$.

(*ii*) Commençons par remarquer que la matrice carrée $R(X'X)^{-1}R'$ de taille q est inversible puisque $(X'X)^{-1}$ est de rang plein dans $\mathbb{R}^p$, avec $p \geq q$. En tant que transformée linéaire d'un vecteur gaussien, $R\hat{\beta}$ est un vecteur gaussien de moyenne $R\beta$ et de matrice de covariance $\sigma^2 R(X'X)^{-1}R'$. On en déduit que :

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

Il reste à remplacer σ^2 par $\hat{\sigma}^2$ en se souvenant que $(n-p)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p}$ et du fait que $\hat{\beta}$ et σ^2 sont indépendants. On obtient bien alors la loi de Fisher annoncée.

■

De ces résultats vont découler les régions de confiance de la section suivante. Auparavant, donnons un exemple illustrant le second point que l'on vient d'établir.

Exemple. Considérons le cas $p = q = 2$ et la matrice $R = I_2$, de sorte que

$$R(\hat{\beta} - \beta) = \begin{bmatrix} \hat{\beta}_1 - \beta_1 \\ \hat{\beta}_2 - \beta_2 \end{bmatrix}.$$

Si la constante fait partie du modèle, X est la matrice $n \times 2$ dont la première colonne est uniquement composée de 1 et la seconde est composée des x_i, si bien que

$$X'X = \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix} = \begin{bmatrix} n & n\bar{x} \\ n\bar{x} & \sum x_i^2 \end{bmatrix}$$

et le point (ii) s'écrit

$$\frac{1}{2\hat{\sigma}^2}\left(n(\hat{\beta}_1 - \beta_1)^2 + 2n\bar{x}(\hat{\beta}_1 - \beta_1)(\hat{\beta}_2 - \beta_2) + \sum x_i^2(\hat{\beta}_2 - \beta_2)^2\right) \sim \mathcal{F}^2_{n-2},$$

qui est exactement le résultat de la Propriété 3 (iii), permettant de construire une ellipse de confiance pour $\beta = (\beta_1, \beta_2)$. Plus généralement, si $p = q$ et $R = I_p$, nous avons

$$\frac{1}{p\hat{\sigma}^2}(\hat{\beta} - \beta)'(X'X)(\hat{\beta} - \beta) \sim \mathcal{F}^p_{n-p},$$

définissant un ellipsoïde de confiance centré en $\hat{\beta}$ pour β. Ce résultat est à la base de la distance de Cook définie en Chapitre 4, Section 4.3.

3.2.3 Intervalles et régions de confiance

Les logiciels et certains ouvrages donnent des intervalles de confiance (IC) pour les paramètres pris séparément. Cependant ces intervalles de confiance ne tiennent pas compte de la dépendance des paramètres, ce qui conduirait à construire plutôt des régions de confiance (RC). Nous allons donc traiter les deux cas, en considérant que σ^2 est inconnu.

Théorème 7 (Intervalles et Régions de Confiance) *(i) Pour tout $j \in \{1, \cdots, p\}$, un intervalle de confiance de niveau $(1-\alpha)$ pour β_j est :*

$$\left[\hat{\beta}_j - t_{n-p}(1-\alpha/2)\hat{\sigma}\sqrt{(X'X)^{-1}_{jj}}, \hat{\beta}_j + t_{n-p}(1-\alpha/2)\hat{\sigma}\sqrt{(X'X)^{-1}_{jj}}\right],$$

où $t_{n-p}(1-\alpha/2)$ est le quantile de niveau $(1-\alpha/2)$ d'une loi de Student $\mathcal{T}_{n-p}$.
(ii) Un intervalle de confiance de niveau $(1-\alpha)$ pour σ^2 est :

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

où $c_{n-p}(1-\alpha/2)$ est le quantile de niveau $(1-\alpha/2)$ d'une loi χ^2_{n-p}.
(iii) Une région de confiance de niveau $(1-\alpha)$ pour q $(q \leq p)$ paramètres β_j notés $(\beta_{j_1}, \cdots, \beta_{j_q})$ est l'ensemble des $(\beta_{j_1}, \cdots, \beta_{j_q})$ tels que

$$\frac{1}{q\hat{\sigma}^2}(R(\hat{\beta} - \beta))'(R(X'X)^{-1}R')^{-1}(R(\hat{\beta} - \beta)) \leq f^q_{n-p}(1-\alpha), \tag{3.1}$$

où R est la matrice de taille $q \times p$ dont tous les éléments sont nuls sauf les R_{i,j_i}, qui valent 1, et $f^q_{n-p}(1-\alpha)$ est le quantile de niveau $(1-\alpha)$ d'une loi de Fisher $\mathcal{F}^q_{n-p}$.

Preuve. Il suffit d'appliquer les résultats de la Proposition 6.

■

Exemple. Considérons $p \geq 2$, $q = 2$ et la matrice R définie comme suit :

$$R = \begin{bmatrix} 1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & 0 & \cdots & 0 \end{bmatrix},$$

de sorte que

$$R(\hat{\beta} - \beta) = \begin{bmatrix} \hat{\beta}_1 - \beta_1 \\ \hat{\beta}_2 - \beta_2 \end{bmatrix}.$$

Si on note c_{ij} le terme général de $(X'X)^{-1}$, le point (iii) permet d'obtenir une région de confiance simultanée $RC(\beta_1, \beta_2)$ pour (β_1, β_2) :

$$\left\{ (\beta_1, \beta_2) \in \mathbb{R}^2 \ : \ \frac{c_{22}(\hat{\beta}_1 - \beta_1)^2 - 2c_{12}(\hat{\beta}_1 - \beta_1)(\hat{\beta}_2 - \beta_2) + c_{11}(\hat{\beta}_2 - \beta_2)^2}{2\hat{\sigma}^2(c_{11}c_{22} - c_{12}^2)} \leq f^2_{n-p}(1-\alpha) \right\}.$$

Cette région de confiance est une ellipse qui tient compte de la corrélation entre $\hat{\beta}_1$ et $\hat{\beta}_2$. La figure 3.1 permet de faire le distinguo entre intervalles de confiance considérés séparément pour $\hat{\beta}_1$ et $\hat{\beta}_2$ et région de confiance simultanée pour (β_1, β_2).

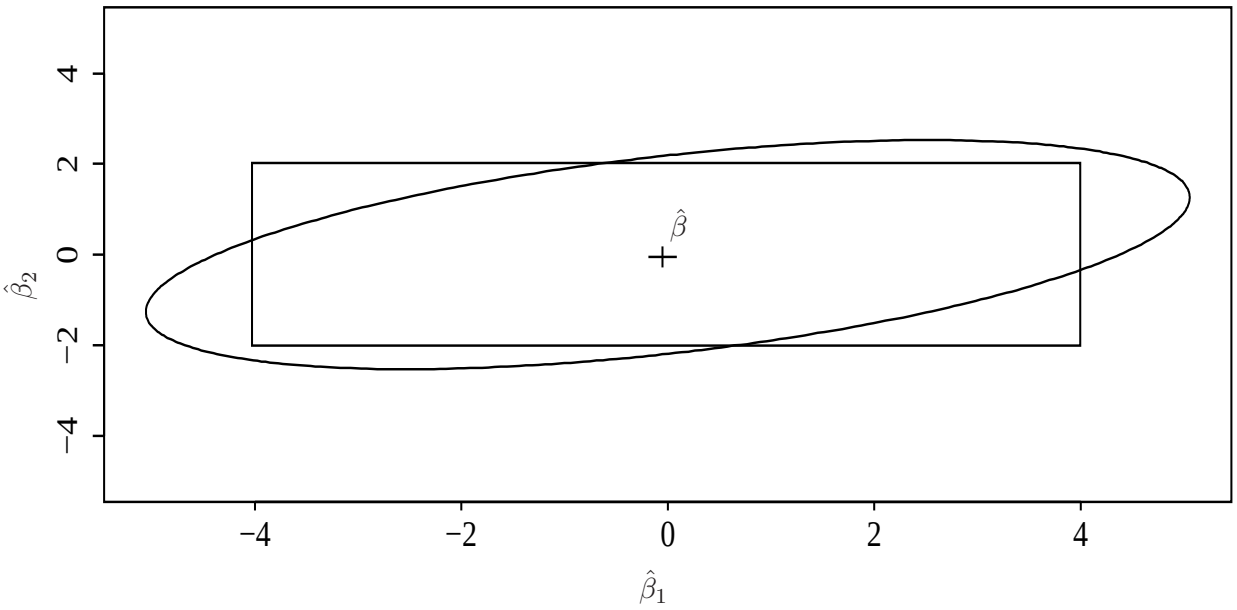


FIGURE 3.1 – Comparaison entre ellipse et rectangle de confiance.

3.2.4 Prévision

Soit $x'_{n+1} = [x_{n+1,1}, \cdots, x_{n+1,p}]$ une nouvelle valeur pour laquelle nous voulons prédire la variable à expliquer y_{n+1} définie par :

$$y_{n+1} = x'_{n+1}\beta + \varepsilon_{n+1}$$

avec $\varepsilon_{n+1} \sim \mathcal{N}(0, \sigma^2)$ indépendant des $(\varepsilon_i)_{1 \leq i \leq n}$. A partir des n observations précédentes, nous avons pu calculer un estimateur $\hat{\beta}$ de β. Nous nous servons de cet estimateur pour prévoir y_{n+1} par :

$$\hat{y}_{n+1} = x'_{n+1}\hat{\beta}.$$

Pour quantifier l'erreur de prévision $(y_{n+1} - \hat{y}_{n+1})$, on utilise la décomposition :

$$y_{n+1} - \hat{y}_{n+1} = x'_{n+1}(\beta - \hat{\beta}) + \varepsilon_{n+1},$$

qui est la somme de deux variables gaussiennes indépendantes puisque $\hat{\beta}$ est construit à partir des $(\varepsilon_i)_{1 \leq i \leq n}$. On en déduit que $(y_{n+1} - \hat{y}_{n+1})$ est une variable gaussienne, dont moyenne et variance ont été calculées au chapitre précédent, ce qui donne :

$$y_{n+1} - \hat{y}_{n+1} \sim \mathcal{N}(0, \sigma^2(1 + x'_{n+1}(X'X)^{-1}x_{n+1}))$$

Mieux, nous pouvons maintenant donner un intervalle de confiance pour y_{n+1}.

Proposition 9 (Intervalle de Confiance pour la prévision) *Un intervalle de confiance de niveau $(1-\alpha)$ pour y_{n+1} est donné par :*

$$\left[x'_{n+1}\hat{\beta} + t_{n-p}(\alpha/2)\hat{\sigma}\sqrt{1 + x'_{n+1}(X'X)^{-1}x_{n+1}}, x'_{n+1}\hat{\beta} - t_{n-p}(\alpha/2)\hat{\sigma}\sqrt{1 + x'_{n+1}(X'X)^{-1}x_{n+1}}\right]$$

Preuve. D'après ce qui a été dit auparavant, on a :

$$\frac{y_{n+1} - \hat{y}_{n+1}}{\sigma\sqrt{1 + x'_{n+1}(X'X)^{-1}x_{n+1}}} \sim \mathcal{N}(0, 1).$$

On procède donc comme d'habitude en faisant intervenir $\hat{\sigma}$:

$$\frac{y_{n+1} - \hat{y}_{n+1}}{\hat{\sigma}\sqrt{1 + x'_{n+1}(X'X)^{-1}x_{n+1}}} = \frac{\frac{y_{n+1} - \hat{y}_{n+1}}{\sigma\sqrt{1 + x'_{n+1}(X'X)^{-1}x_{n+1}}}}{\frac{\hat{\sigma}}{\sigma}}.$$

On remarque que le numérateur suit une loi normale centrée réduite, le dénominateur est la racine d'un chi-deux à $(n-p)$ ddl divisé par $(n-p)$. Il reste à voir que numérateur et dénominateur sont indépendants, or $y_{n+1} - \hat{y}_{n+1} = x'_{n+1}(\beta - \hat{\beta}) + \varepsilon_{n+1}$ et $\hat{\sigma}$ est indépendant à la fois de $\hat{\beta}$ (cf. Propriétés 5) et de ε_{n+1} (puisque $\hat{\sigma}$ ne dépend que des $(\varepsilon_i)_{1 \leq i \leq n}$). On en conclut que :

$$\frac{y_{n+1} - \hat{y}_{n+1}}{\hat{\sigma}\sqrt{1 + x'_{n+1}(X'X)^{-1}x_{n+1}}} \sim \mathcal{T}_{n-p},$$

d'où se déduit l'intervalle de confiance de l'énoncé.

■

Après avoir explicité les lois de nos estimateurs et les intervalles ou régions de confiance associés, tout est prêt pour construire des tests d'hypothèses. C'est ce que nous allons faire dans la section suivante.

3.3 Tests d'hypothèses

3.3.1 Introduction

Reprenons l'exemple de la prévision des pics d'ozone vu en début de Chapitre 2. Nous avons décidé de modéliser les pics d'ozone O_3 par la température à midi T, le vent V (ou plus précisément sa projection sur l'axe Est-Ouest) et la nébulosité à midi N. Il paraît alors raisonnable de se poser par exemple les questions suivantes :

1. Est-ce que la valeur de O_3 est influencée par la variable vent V ?
2. Y a-t-il un effet nébulosité ?
3. Est-ce que la valeur de O_3 est influencée par le vent V ou la température T ?

Rappelons que le modèle utilisé est le suivant :

$$O_{3i} = \beta_1 + \beta_2 T_i + \beta_3 V_i + \beta_4 N_i + \varepsilon_i$$

En termes de tests d'hypothèses, les questions ci-dessus se traduisent comme suit :

1. correspond à $\mathrm{H}_0 : \beta_3 = 0$, contre $\mathrm{H}_1 : \beta_3 \neq 0$.
2. correspond à $\mathrm{H}_0 : \beta_4 = 0$, contre $\mathrm{H}_1 : \beta_4 \neq 0$.
3. correspond à $\mathrm{H}_0 : \beta_2 = \beta_3 = 0$, contre $\mathrm{H}_1 : \ \beta_2 \neq 0$ ou $\beta_3 \neq 0$.

Ces tests d'hypothèses reviennent à tester la nullité d'un ou plusieurs paramètres en même temps. Si l'on teste plusieurs paramètres à la fois, on parle de nullité simultanée des coefficients. Ceci signifie que, sous l'hypthèse H_0, certains coefficients sont nuls, donc les variables correspondant à ceux-ci ne sont pas utiles pour la modélisation du phénomène. Ce cas de figure revient à comparer deux modèles emboîtés, l'un étant un cas particulier de l'autre.

Le plan d'expérience privé de ces variables sera noté X_0 et les colonnes de X_0 engendreront un sous-espace noté $\mathcal{M}_0 = \mathcal{M}(X_0)$. De même, pour alléger les notations, nous noterons $\mathcal{M} = \mathcal{M}(X)$ l'espace engendré par les colonnes de X. Le niveau de risque des tests sera fixé de façon classique à α.

3.3.2 Tests entre modèles emboîtés

Rappelons tout d'abord le modèle :

$$Y = X\beta + \boldsymbol{\varepsilon} \quad \text{sous les hypothèses } (\mathcal{H}) \begin{cases} (\mathcal{H}_1) : \mathrm{rg}(X) = p \\ (\mathcal{H}_2) : \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_n) \end{cases}$$

En particulier, cela veut dire que $\mathbb{E}[Y] = X\beta \in \mathcal{M}$, sous-espace de dimension p de $\mathbb{R}^n$ engendré par les p colonnes de X. Pour faciliter les notations, on suppose vouloir tester la nullité simultanée des $q = (p - p_0)$ derniers coefficients du modèle (avec $q \leq p$ of course !). Le problème s'écrit alors de la façon suivante :

$$\mathrm{H}_0 : \ \beta_{p_0+1} = \cdots = \beta_p = 0 \quad \text{contre} \quad \mathrm{H}_1 : \ \exists j \in \{p_0 + 1, \cdots, p\} : \beta_j \neq 0.$$

Que signifie $\mathrm{H}_0 : \ \beta_{p_0+1} = \cdots = \beta_p = 0$ en termes de modèle ? Si les q derniers coefficients sont nuls, le modèle devient

$$Y = X_0\beta_0 + \boldsymbol{\varepsilon}_0 \quad \text{sous les hypothèses } (\mathcal{H}) \begin{cases} (\mathcal{H}_1) : \mathrm{rg}(X_0) = p_0 \\ (\mathcal{H}_2) : \boldsymbol{\varepsilon}_0 \sim \mathcal{N}(0, \sigma^2 I_n) \end{cases}$$

La matrice X_0, de taille $n \times p_0$, est composée des p_0 premières colonnes de X et β_0 est un vecteur colonne de taille p_0. Puisque X est supposée de rang p, il est clair que X_0 est de rang p_0, donc les colonnes de X_0 engendrent un sous-espace $\mathcal{M}_0$ de $\mathbb{R}^n$ de dimension p_0. Ce sous-espace $\mathcal{M}_0$ est bien évidemment aussi un sous-espace de $\mathcal{M}$. Sous l'hypothèse nulle H_0, l'espérance de Y, à savoir $\mathbb{E}[Y] = X_0\beta_0$, appartiendra à ce sous-espace $\mathcal{M}_0$.

Maintenant que les hypothèses du test sont fixées, il faut proposer une statistique de test. Nous allons voir une approche géométrique et intuitive de l'affaire.

Approche géométrique

Considérons le sous-espace $\mathcal{M}_0$. Nous avons écrit que sous $\mathrm{H}_0 : \mathbb{E}[Y] = X_0\beta_0 \in \mathcal{M}_0$. Dans ce cas, la méthode des moindres carrés consiste à projeter Y non plus sur $\mathcal{M}$ et à obtenir $\hat{Y}$, mais sur $\mathcal{M}_0$ et à obtenir $\hat{Y}_0$. Visualisons ces différentes projections sur la figure 3.2.

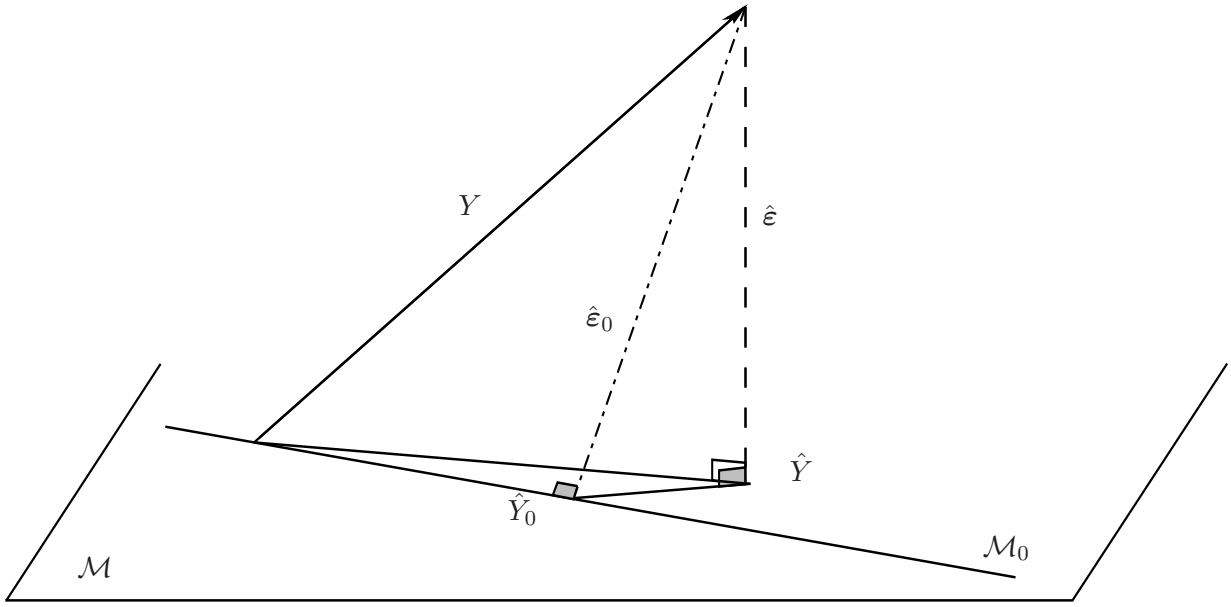


FIGURE 3.2 – Représentation des projections.

L'idée intuitive du test, et donc du choix de conserver ou non H_0, est la suivante : si la projection $\hat{Y}_0$ de Y dans $\mathcal{M}_0$ est "proche" de la projection $\hat{Y}$ de Y dans $\mathcal{M}$, alors il semble intuitif de conserver l'hypothèse nulle. En effet, si l'information apportée par les deux modèles est "à peu près la même", il vaut mieux conserver le modèle le plus petit : c'est le principe de parcimonie.

Il faut évidemment quantifier le terme "proche". Pour ce faire, nous pouvons utiliser la distance euclidienne entre $\hat{Y}_0$ et $\hat{Y}$, ou son carré $\|\hat{Y} - \hat{Y}_0\|^2$. Mais cette distance sera variable selon les données et les unités de mesures utilisées. Pour nous affranchir de ce problème d'échelle, nous allons "standardiser" cette distance en la divisant par la norme au carré de l'erreur estimée $\|\hat{\boldsymbol{\varepsilon}}\|^2 = \|Y - \hat{Y}\|^2 = (n-p)\hat{\sigma}^2$. Les vecteurs aléatoires $(\hat{Y} - \hat{Y}_0)$ et $\hat{\boldsymbol{\varepsilon}}$ n'appartenant pas à des sous-espaces de même dimension, il faut encore diviser chaque terme par son degré de liberté respectif, soit $q = p - p_0$ et $n - p$. Toute cette tambouille nous mène à la statistique de test suivante :

$$F = \frac{\|\hat{Y} - \hat{Y}_0\|^2/q}{\|Y - \hat{Y}\|^2/(n-p)} = \frac{\|\hat{Y} - \hat{Y}_0\|^2/(p-p_0)}{\|Y - \hat{Y}\|^2/(n-p)}.$$

Pour utiliser cette statistique de test, il faut connaître au moins sa loi sous H_0. Remarquons qu'elle correspond au rapport de deux normes au carré. Nous allons déterminer la loi du numérateur, celle du dénominateur et constater leur indépendance. En notant P (resp. P_0) la matrice de projection orthogonale sur $\mathcal{M}$ (resp. $\mathcal{M}_0$), nous savons que :

$$\hat{Y} - \hat{Y}_0 = PY - P_0 Y,$$

or $\mathcal{M}_0 \subset \mathcal{M}$ donc $P_0 Y = P_0 PY$ et :

$$\hat{Y} - \hat{Y}_0 = \quad PY - P_0 PY = (I_n - P_0)PY = P_0^{\perp} PY.$$

Nous en déduisons que $(\hat{Y} - \hat{Y}_0) \in \mathcal{M}_0^{\perp} \cap \mathcal{M}$, donc que $(\hat{Y} - \hat{Y}_0) \perp (Y - \hat{Y})$ puisque $(Y - \hat{Y}) \in \mathcal{M}^{\perp}$. La figure 3.2 permet de visualier ces notions d'orthogonalité de façon géométrique. Les vecteurs aléatoires $(\hat{Y} - \hat{Y}_0)$ et $(Y - \hat{Y})$ sont éléments d'espaces orthogonaux, c'est-à-dire qu'ils ont une covariance nulle. Puisque tout est gaussien, ils sont donc indépendants et les normes du numérateur et du dénominateur sont indépendantes également.

Le théorème de Cochran nous renseigne par ailleurs sur les lois des numérateur et dénominateur. Pour le dénominateur :

$$\frac{1}{\sigma^2}\|Y - \hat{Y}\|^2 = \frac{1}{\sigma^2}\|P^{\perp} Y\|^2 = \frac{1}{\sigma^2}\|P^{\perp}(X\beta + \varepsilon)\|^2 = \frac{1}{\sigma^2}\|P^{\perp}\varepsilon\|^2 \sim \chi^2_{n-p},$$

et pour le numérateur :

$$\frac{1}{\sigma^2}\|P_0^{\perp} P(Y - X\beta)\|^2 \sim \chi^2_q.$$

Sous H_0, le paramètre de décentrage $\|P_0^{\perp} PX\beta\|^2$ est nul puisque dans ce cas $X\beta \in \mathcal{M}_0$.

Nous avons alors la loi de F sous H_0 :

$$F \sim \mathcal{F}^q_{n-p}.$$

Notons une écriture équivalente souvent utilisée, donc importante :

$$F \quad = \quad \frac{n-p}{q} \times \frac{SCR_0 - SCR}{SCR} \sim \mathcal{F}^q_{n-p}.$$

La relation $\|\hat{Y} - \hat{Y}_0\|^2 = (SCR_0 - SCR)$ peut se voir facilement en utilisant la figure 3.2 et en appliquant le théorème de Pythagore :

$$\begin{aligned} \|Y - \hat{Y}_0\|^2 &= \|Y - PY + PY - P_0 Y\|^2 = \|P^{\perp} Y + (I_n - P_0)PY\|^2 = \|P^{\perp} Y + P_0^{\perp} PY\|^2 \\ &= \|P^{\perp} Y\|^2 + \|P_0^{\perp} PY\|^2 \\ &= \|Y - \hat{Y}\|^2 + \|\hat{Y} - \hat{Y}_0\|^2, \end{aligned}$$

c'est-à-dire :

$$\|\hat{Y} - \hat{Y}_0\|^2 = \|Y - \hat{Y}_0\|^2 - \|Y - \hat{Y}\|^2 = SCR_0 - SCR.$$

Résumons ce qui précède.

Proposition 10 (Test entre modèles emboîtés) *Sous l'hypothèse* H_0*, on a la statistique de test suivante*

$$F = \frac{n-p}{q} \times \frac{\|\hat{Y} - \hat{Y}_0\|^2}{\|Y - \hat{Y}\|^2} = \frac{n-p}{q} \times \frac{SCR_0 - SCR}{SCR} \sim \mathcal{F}^q_{n-p},$$

loi de Fisher à $(q, n-p)$ *degrés de liberté.*

Preuve. Alternativement à la preuve géométrique ci-dessus, il est possible de démontrer ce résultat en appliquant brutalement la Propriété 6 (ii) avec pour R la matrice $q \times p$ définie par blocs comme suit : $R = [0|I_q]$. On sait en effet que

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

Sous l'hypothèse H_0, il vient $R\beta = 0$, donc

$$R(\hat{\beta}-\beta) = [\hat{\beta}_{p_0+1}, \ldots, \hat{\beta}_p]'.$$

D'autre part, si l'on note $X = [X_0|\bar{X}_0]$, la formule d'inversion matricielle par blocs (B.2) rappelée en Annexe assure que

$$\left[R(X'X)^{-1}R'\right]^{-1} = \bar{X}_0'(I_n - P_0)\bar{X}_0$$

de sorte que

$$(R(\hat{\beta}-\beta))' \left[R(X'X)^{-1}R'\right]^{-1} R(\hat{\beta}-\beta) = [\hat{\beta}_{p_0+1}, \ldots, \hat{\beta}_p]\bar{X}_0'(I-P_0)\bar{X}_0[\hat{\beta}_{p_0+1}, \ldots, \hat{\beta}_p]'$$

Puisque $(I_n - P_0)$ est le projecteur (orthogonal) sur $\mathcal{M}_0^{\perp}$, il est idempotent, donc

$$(R(\hat{\beta}-\beta))' \left[R(X'X)^{-1}R'\right]^{-1} R(\hat{\beta}-\beta) = \|(I_n - P_0)\bar{X}_0[\hat{\beta}_{p_0+1}, \ldots, \hat{\beta}_p]'\|^2$$

Il faut maintenant voir que

$$\hat{Y} = X\hat{\beta} = \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p = (\hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_{p_0}) + (\hat{\beta}_{p_0+1} X_{p_0+1} + \cdots + \hat{\beta}_p X_p)$$

tandis que, via $\mathcal{M}_0 \subset \mathcal{M}$, on a $P_0 P = P_0$ donc

$$\hat{Y}_0 = P_0 Y = P_0 P Y = P_0 \hat{Y} = P_0(X\hat{\beta}) = (\hat{\beta}_1 P_0 X_1 + \cdots + \hat{\beta}_p P_0 X_{p_0}) + (\hat{\beta}_{p_0+1} P_0 X_{p_0+1} + \cdots + \hat{\beta}_p P_0 X_p)$$

et puisque $P_0 X_1 = X_1, \ldots, P_0 X_{p_0} = X_{p_0}$, il vient

$$\hat{Y}_0 = (\hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_{p_0}) + (\hat{\beta}_{p_0+1} P_0 X_{p_0+1} + \cdots + \hat{\beta}_p P_0 X_p)$$

Ainsi

$$\hat{Y} - \hat{Y}_0 = (\hat{\beta}_{p_0+1} X_{p_0+1} + \cdots + \hat{\beta}_p X_p) - (\hat{\beta}_{p_0+1} P_0 X_{p_0+1} + \cdots + \hat{\beta}_p P_0 X_p)$$

ou encore

$$\hat{Y} - \hat{Y}_0 = (I_n - P_0)(\hat{\beta}_{p_0+1} X_{p_0+1} + \cdots + \hat{\beta}_p X_p)$$

c'est-à-dire

$$\hat{Y} - \hat{Y}_0 = (I_n - P_0)\bar{X}_0[\hat{\beta}_{p_0+1}, \ldots, \hat{\beta}_p]'.$$

Il reste à se souvenir que

$$\hat{\sigma}^2 = \frac{\|Y - \hat{Y}\|^2}{n-p}$$

pour arriver au résultat voulu.

■

Remarque. En supposant que la constante fait partie des deux modèles (ou ne fait partie d'aucun d'entre eux), la statistique de test précédente peut aussi s'écrire en fonction des coefficients de détermination respectifs R^2 et R_0^2 comme suit (exercice) :

$$F = \frac{n-p}{q} \times \frac{R^2 - R_0^2}{1 - R^2}.$$

Ainsi, si l'on dispose des coefficients de détermination dans deux modèles emboîtés, il suffit de calculer cette statistique et de la comparer au quantile d'une loi de Fisher pour effectuer le test d'hypothèse.

Nous allons maintenant expliciter cette statistique de test dans deux cas particuliers.

3.3.3 Test de Student de signification d'un coefficient

Nous voulons tester $H_0 : \beta_j = 0$ contre $H_1 : \beta_j \neq 0$, appelé test bilatéral de significativité de β_j. Selon ce qu'on vient de voir, la statistique de test est :

$$F = \frac{\|\hat{Y} - \hat{Y}_0\|^2}{\hat{\sigma}^2}.$$

Nous rejetons H_0 si l'observation de la statistique de test, notée $F(w)$, est telle que :

$$F(w) > f^1_{n-p}(1-\alpha),$$

où $f^1_{n-p}(1-\alpha)$ est le quantile d'ordre $(1-\alpha)$ d'une loi de Fisher à 1 et $(n-p)$ degrés de liberté.

Ce test est en fait équivalent au test de Student à $(n-p)$ degrés de liberté qui permet de tester $H_0 : \beta_j = 0$ contre $H_1 : \beta_j \neq 0$, avec cette fois la statistique de test :

$$T = \frac{\hat{\beta}_j}{\hat{\sigma}_j},$$

où $\hat{\sigma}_j = \hat{\sigma}_{\hat{\beta}_j} = \hat{\sigma}\sqrt{(X'X)^{-1}_{jj}}$ est l'écart-type estimé de $\hat{\beta}_j$. On peut en effet montrer que $F = T^2$ (voir exercice 3.3). Nous rejetons H_0 si l'observation de la statistique de test, notée $T(w)$, est telle que :

$$|T(w)| > t_{n-p}(1-\alpha/2),$$

où $t_{n-p}(1-\alpha/2)$ est le quantile d'ordre $(1-\alpha/2)$ d'une loi de Student à $(n-p)$ degrés de liberté. C'est sous cette forme que le test de significativité d'un coefficient apparaît dans tous les logiciels de statistique. Il est donc complètement équivalent au test général que nous avons proposé, lorsqu'on spécialise celui-ci à la nullité d'un seul coefficient.

3.3.4 Test de Fisher global

Si des connaissances a priori du phénomène assurent l'existence d'un terme constant dans la régression, alors pour tester l'influence des autres régresseurs (non constants) sur la réponse Y, on regarde si $\mathbb{E}[Y] = \beta_1$. En d'autres termes, on teste si tous les coefficients sont nuls, excepté la constante.

Ce test est appelé test de Fisher global. Dans ce cas $\hat{Y}_0 = \bar{y}\mathbb{1}$ et nous avons la statistique de test suivante :

$$F = \frac{\|\hat{Y} - \bar{y}\mathbb{1}\|^2/(p-1)}{\|Y - \hat{Y}\|^2/(n-p)} \sim \mathcal{F}^{p-1}_{n-p}.$$

On peut aussi l'exprimer à partir du coefficient de détermination R^2 vu au Chapitre 2 :

$$F = \frac{n-p}{p-1} \times \frac{R^2}{1-R^2}.$$

Ce test est appelé le test du R^2 par certains logiciels statistiques.

3.3.5 Lien avec le Rapport de Vraisemblance Maximale

Nous allons maintenant faire le lien entre le test général que nous avons proposé et le test du rapport de vraisemblance maximale. Nous avons vu en début du chapitre que la vraisemblance s'écrit de la façon suivante :

$$\mathcal{L}(Y, \beta, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left[-\frac{1}{2\sigma^2}\|Y - X\beta\|^2\right].$$

Cette vraisemblance est maximale lorsque $\beta = \hat{\beta}$ est l'estimateur des MCO et que $\sigma^2 = \hat{\sigma}^2_{mv} = ||Y - X\hat{\beta}||^2/n$. Nous avons alors :

$$\begin{aligned}\sup_{\beta,\sigma^2} \mathcal{L}(Y,\beta,\sigma^2) &= \mathcal{L}(Y,\hat{\beta},\hat{\sigma}^2_{mv}) \\ &= \left(\frac{n}{2\pi||Y - X\hat{\beta}||^2}\right)^{n/2} e^{-\frac{n}{2}} \\ &= \left(\frac{n}{2\pi SCR}\right)^{n/2} e^{-\frac{n}{2}},\end{aligned}$$

où SCR correspond à la somme des carrés résiduels, c'est-à-dire $SCR = ||Y - X\hat{\beta}||^2$. Sous l'hypothèse H_0, nous obtenons de façon évidente le résultat suivant :

$$\sup_{\beta,\sigma^2} \mathcal{L}_0(Y,\beta_0,\sigma^2) = \left(\frac{n}{2\pi SCR_0}\right)^{n/2} e^{-\frac{n}{2}} = \mathcal{L}_0(Y,\hat{\beta}_0,\hat{\sigma}^2_0),$$

où SCR_0 correspond à la somme des carrés résiduels sous H_0, c'est-à-dire $SCR_0 = ||Y - X_0\hat{\beta}_0||^2$, et $\hat{\sigma}^2_0 = SCR_0/n$. On définit alors le test du Rapport de Vraisemblance Maximale par la région critique :

$$\mathcal{D}_\alpha = \left\{ Y \in \mathbb{R}^n : \lambda = \frac{\mathcal{L}_0(Y,\hat{\beta}_0,\hat{\sigma}^2_0)}{\mathcal{L}(Y,\hat{\beta},\hat{\sigma}^2_{mv})} < \lambda_0 \right\}.$$

La statistique du Rapport de Vraisemblance Maximale vaut donc ici :

$$\lambda = \left(\frac{SCR_0}{SCR}\right)^{-n/2}.$$

Le test du Rapport de Vraisemblance Maximale rejette H_0 lorsque la statistique λ est inférieure à une valeur λ_0 définie de façon à avoir le niveau du test égal à α. Il reste à connaître la distribution (au moins sous H_0) de λ. Définissons, pour λ positif, la fonction g suivante :

$$g(\lambda) = \lambda^{-2/n} - 1.$$

La fonction g est décroissante donc $\lambda < \lambda_0$ si et seulement si $g(\lambda) > g(\lambda_0)$. Cette fonction g va nous permettre de nous ramener à des statistiques dont la loi est connue. Nous avons en effet :

$$g(\lambda) > g(\lambda_0) \iff \frac{SCR_0 - SCR}{SCR} > g(\lambda_0) \iff \frac{n-p}{p-p_0} \times \frac{SCR_0 - SCR}{SCR} > f_0,$$

où f_0 est déterminé par :

$$\mathbb{P}_{\mathrm{H}_0}\left(\frac{n-p}{p-p_0} \times \frac{SCR_0 - SCR}{SCR} > f_0\right) = \alpha,$$

c'est-à-dire $f_0 = f^q_{n-p}(1-\alpha)$, quantile de la loi de Fisher $\mathcal{F}^q_{n-p}$ (cf. section précédente). Le test du Rapport de Vraisemblance Maximale est donc équivalent au test qui rejette H_0 lorsque la statistique :

$$F = \frac{n-p}{p-p_0} \times \frac{SCR_0 - SCR}{SCR}$$

est supérieure à f_0, où f_0 la valeur du quantile d'ordre $(1-\alpha)$ de la loi de Fisher à $(p-p_0, n-p)$ degrés de liberté. Ainsi le test géométrique que nous avons proposé est équivalent au test du Rapport de Vraisemblance Maximale.

3.4 Estimation sous contraintes

L'espace des solutions est $\mathcal{M}$. Tous les vecteurs de $\mathcal{M}$ peuvent s'écrire comme combinaisons linéaires des vecteurs colonnes de X. Il arrive parfois que nous souhaitions imposer des contraintes linéaires à β, par exemple que la première coordonnée de β soit égale à 1. Nous supposerons en général que nous imposons q contraintes linéairement indépendantes à β, ce qui s'écrit sous la forme : $R\beta = r$, où $R_{q\times p}$ est une matrice de rang $q < p$ et r un vecteur de taille q.

Propriétés 7 *L'estimateur des Moindres Carrés Ordinaires sous contrainte, noté $\hat{\beta}_c$, vaut :*

$$\hat{\beta}_c = \hat{\beta} + (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}(r - R\hat{\beta}).$$

Preuve. Nous voulons minimiser $S(\beta)$ sous la contrainte $R\beta = r$. Ecrivons le lagrangien :

$$\mathcal{L} = S(\beta) - \lambda'(R\beta - r).$$

Les conditions de Lagrange permettent d'obtenir un minimum :

$$\left\{ \begin{array}{l} \dfrac{\partial \mathcal{L}}{\partial \beta} = -2X'Y + 2X'X\hat{\beta}_c - R'\hat{\lambda} = 0, \\ \dfrac{\partial \mathcal{L}}{\partial \lambda} = R\hat{\beta}_c - r = 0, \end{array} \right.$$

Multiplions à gauche la première égalité par $R(X'X)^{-1}$, nous obtenons

$$\begin{array}{rcl} -2R(X'X)^{-1}X'Y + 2R(X'X)^{-1}X'X\hat{\beta}_c - R(X'X)^{-1}R'\hat{\lambda} & = & 0 \\ -2R(X'X)^{-1}X'Y + 2R\hat{\beta}_c - R(X'X)^{-1}R'\hat{\lambda} & = & 0 \\ -2R(X'X)^{-1}X'Y + 2r - R(X'X)^{-1}R'\hat{\lambda} & = & 0. \end{array}$$

Nous obtenons alors pour $\hat{\lambda}$:

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

Remplaçons ensuite $\hat{\lambda}$ par cette expression dans la première équation :

$$-2X'Y + 2X'X\hat{\beta}_c - 2R'\left[R(X'X)^{-1}R'\right]^{-1}\left[r - R(X'X)^{-1}X'Y\right] = 0,$$

d'où nous déduisons $\hat{\beta}_c$:

$$\begin{array}{rcl} \hat{\beta}_c & = & (X'X)^{-1}X'Y + (X'X)^{-1}R'\left[R(X'X)^{-1}R'\right]^{-1}(r - R\hat{\beta}) \\ & = & \hat{\beta} + (X'X)^{-1}R'\left[R(X'X)^{-1}R'\right]^{-1}(r - R\hat{\beta}). \end{array}$$

■

3.5 Exemple

Nous allons traiter 50 données journalières présentées en annexe. La variable à expliquer est la concentration en ozone notée `O3` et les variables explicatives sont la température `T12`, le vent `Vx` et la nébulosité `Ne12`.

```
> a <- lm(O3 ~ T12+Vx+Ne12,data=DONNEE)
> summary(a)
Call:
```

```
lm(formula = O3 ~ T12 + Vx + Ne12, data = DONNEE))


 Residuals:
                    Min          1Q      Median          3Q         Max
                -29.0441     -8.4833      0.7857      7.7011     28.2919

 Coefficients :
                Estimate  Std. Error  t value  Pr(>|t|)
 (Intercept)     84.5483     13.6065    6.214  1.38e-07      ***
 T12              1.3150      0.4974    2.644   0.01118        *
 Vx               0.4864      0.1675    2.903   0.00565       **
 Ne12            -4.8935      1.0270   -4.765  1.93e-05      ***


--
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 13.91 on 46 degrees of freedom
Multiple R-Squared: 0.6819, Adjusted R-squared: 0.6611
F-statistic: 32.87 on 3 and 46 DF, p-value: 1.663e-11
```

Pour tous les coefficients pris séparément, nous refusons au seuil $\alpha = 5\%$ l'hypothèse $H_0 : \beta_j = 0$. La dernière ligne de la sortie du logiciel donne la statistique du test de Fisher global : "Tous les coefficients sont nuls sauf la constante". Nous avions 50 observations, nous avons estimé 4 paramètres et donc les degrés de liberté de la loi de Fisher sont bien (3,46). Nous refusons à nouveau H_0. De façon générale, il est clair qu'à moins d'avoir proposé n'importe quoi comme régresseurs, ce test est toujours rejeté...

3.6 Exercices

Exercice 3.1 (QCM)

1. Nous pouvons justifier les MC quand $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$ via l'application du maximum de vraisemblance :
   A. Oui ;
   B. Non ;
   C. Aucun rapport entre les deux méthodes.
2. Y a-t-il une différence entre les estimateurs $\hat{\beta}$ des MC et $\widetilde{\beta}$ du maximum de vraisemblance ?
   A. Oui ;
   B. Non ;
   C. Pas toujours, cela dépend de la loi des erreurs.
3. Y a-t-il une différence entre les estimateurs $\hat{\sigma}^2$ des MC et $\widetilde{\sigma}^2$ du maximum de vraisemblance quand $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$?
   A. Oui ;
   B. Non ;
   C. Pas toujours, cela dépend de la loi des erreurs.
4. Le rectangle formé par les intervalles de confiance de niveau α individuels de β_1 et β_2 correspond à la région de confiance simultanée de niveau α de la paire (β_1, β_2).
   A. Oui ;
   B. Non ;
   C. Cela dépend des données.

5. Nous avons n observations et p variables explicatives, nous supposons que ε suit une loi normale, nous voulons tester $\mathcal{H}_0 : \beta_2 = \beta_3 = \beta_4 = 0$. Quelle va être la loi de la statistique de test ?
   A. $\mathcal{F}_{p-3,n-p}$;
   B. $\mathcal{F}_{3,n-p}$;
   C. Une autre loi.

Exercice 3.2 (Analyse de sorties logiciel) Nous voulons expliquer la concentration de l'ozone sur Rennes en fonction des variables `T9`, `T12`, `Ne9`, `Ne12` et `Vx`. Les sorties données par R sont (à une vache près) :

```
Coefficients :
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	62	10	[1]	0
T9	-4	[2]	-5	0
T12	5	0.75	[3]	0
Ne9	-1.5	1	[4]	0.13
Ne12	-0.5	0.5	[5]	0.32
Vx	0.8	0.15	5.3	0

```
--
Multiple R-Squared: 0.6233, Adjusted R-squared: 0.6081
Residual standard error: 16 on 124 degrees of freedom
F-statistic: [6] on [7] and [8] DF, p-value: 0
```

1. Compléter approximativement la sortie ci-dessus.
2. Rappeler la statistique de test et tester la nullité des paramètres séparément au seuil de 5 %.
3. Rappeler la statistique de test et tester la nullité simultanée des paramètres autres que la constante au seuil de 5 %.
4. Les variables `Ne9` et `Ne12` ne semblent pas influentes et nous souhaitons tester la nullité simultanée de $\beta_{\texttt{Ne9}}$ et $\beta_{\texttt{Ne12}}$. Proposer un test et l'effectuer à partir des résultats numériques suivants :

```
Coefficients :
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	66	11	6	0
T9	-5	1	-5	0
T12	6	0.75	8	0
Vx	1	0.2	5	0

```
--
Multiple R-Squared: 0.5312, Adjusted R-squared: 0.52
Residual standard error: 16.5 on 126 degrees of freedom
```

Exercice 3.3 (Equivalence du test T et du test F) On souhaite montrer l'équivalence entre les tests de Student et de Fisher pour la nullité d'un paramètre. On considère donc le modèle

$Y = X\beta + \boldsymbol{\varepsilon}$ sous les hypothèses classiques, pour lequel on veut tester la nullité du dernier coefficient β_p.

1. Rappeler la statistique T du test de Student sous l'hypothèse $\mathrm{H}_0 : \beta_p = 0$.
2. Donner la statistique F du test de Fisher pour les modèles emboîtés correspondants.
3. Soit T_n une variable suivant une loi de Student à n degrés de liberté. Rappeler sa définition et en déduire la loi suivie par la variable $F_n = T_n^2$.
4. On note la matrice du plan d'expérience sous forme bloc $X = [X_0|X_p]$, où $X_0 = [X_1|\dots|X_{p-1}]$ est la matrice $n \times (p-1)$ des $(p-1)$ premières colonnes de X, et X_p est sa dernière colonne. Ecrire la matrice $X'X$ sous forme de 4 blocs.
5. Grâce à la formule d'inversion matricielle par blocs, en déduire que
$$[(X'X)^{-1}]_{pp} = (X_p'(I_n - P_0)X_p)^{-1}$$
où P_0 est la matrice $n \times n$ de projection orthogonale sur l'espace $\mathcal{M}_0$ engendré par les colonnes de X_0.
6. En notant $\hat{Y}$ et $\hat{Y}_0$ les projetés orthogonaux de Y sur $\mathcal{M}$ et $\mathcal{M}_0$, et en justifiant le fait que
$$\hat{Y}_0 = P_0 Y = P_0 P Y = P_0 \hat{Y}$$
montrer que
$$\hat{Y} - \hat{Y}_0 = \hat{\beta}_p (I_n - P_0) X_p$$
7. En déduire que $F = T^2$ et conclure.

Exercice 3.4 (Un modèle à 3 variables explicatives) On considère un modèle de régression de la forme :

$$y_i = \beta_1 + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_4 x_{i,4} + \varepsilon_i, \qquad 1 \leq i \leq n.$$

Les $x_{i,j}$ sont supposées non aléatoires. Les erreurs ε_i du modèle sont supposées aléatoires indépendantes gaussiennes centrées de même variance σ^2. On pose comme d'habitude :

$$X = \begin{bmatrix} 1 & x_{1,2} & x_{1,3} & x_{1,4} \\ \vdots & \vdots & \vdots & \\ 1 & x_{n,2} & x_{n,3} & x_{n,4} \end{bmatrix}, \quad Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix}.$$

Un calcul préliminaire a donné

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

On admettra que

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

1. Calculer $\hat{\beta}$, estimateur des moindres carrés de β, la somme des carrés des résidus $\sum_{i=1}^{50} \hat{\varepsilon}_i^2$, et donner l'estimateur de σ^2.

2. Donner un intervalle de confiance pour β_2, au niveau 95%. Faire de même pour σ^2 (on donne $c_1 = 29$ et $c_2 = 66$ pour les quantiles d'ordre 2,5% et 97,5% d'un chi-deux à 46 ddl).
3. Tester la "validité globale" du modèle ($\beta_2 = \beta_3 = \beta_4 = 0$) au niveau 5% (on donne $f^3_{46}(0,95) = 2.80$ pour le quantile d'ordre 95% d'une Fisher à (3,46) ddl).
4. On suppose $x_{51,2} = 1$, $x_{51,3} = -1$ et $x_{51,4} = 0,5$. Donner un intervalle de prévision à 95% pour y_{51}.

Exercice 3.5 (Modèle de Cobb-Douglas) Nous disposons pour n entreprises de la valeur du capital K_i, de l'emploi L_i et de la valeur ajoutée V_i. Nous supposons que la fonction de production de ces entreprises est du type Cobb-Douglas :

$$V_i = \lambda L_i^{\beta} K_i^{\gamma},$$

soit en passant en logarithmes :

$$\log V_i = \alpha + \beta \log L_i + \gamma \log K_i. \tag{3.2}$$

Le modèle linéaire associé est :

$$\log V_i = \alpha + \beta \log L_i + \gamma \log K_i + \varepsilon_i,$$

où les ε_i sont supposées i.i.d. de loi $\mathcal{N}(0, \sigma^2)$.

1. Ecrivez le modèle sous la forme matricielle $Y = Xb + \boldsymbol{\varepsilon}$ en précisant Y, X et b. Rappelez l'expression de l'estimateur des MCO $\hat{b}$. Donnez sa matrice de variance-covariance. Donnez un estimateur sans biais de σ^2 et un estimateur sans biais de $\mathrm{Var}(\hat{b})$.
2. Pour 1658 entreprises, nous avons obtenu par les MCO les résultats suivants :

$$\begin{cases} \log V_i = 3.136 + 0.738 \log L_i + 0.282 \log K_i \\ R^2 = 0.945 \\ SCR = 148.27. \end{cases}$$

   Nous donnons aussi :

$$(X'X)^{-1} = \begin{bmatrix} 0.0288 & 0.0012 & -0.0034 \\ 0.0012 & 0.0016 & -0.0010 \\ -0.0034 & -0.0010 & 0.0009 \end{bmatrix}$$

   Calculez $\hat{\sigma}^2$ et une estimation de $\mathrm{Var}(\hat{b})$.
3. Donnez un intervalle de confiance au niveau 95% pour β. Même question pour γ.
4. Testez au niveau 5% $H_0 : \gamma = 0$, contre $H_1 : \gamma > 0$.
5. Nous voulons tester l'hypothèse selon laquelle les rendements d'échelle sont constants (une fonction de production F est à rendement d'échelle constant si $\forall \alpha \in \mathbb{R}^+$, $F(\alpha L, \alpha K) = \alpha F(L, K)$). Quelles sont les contraintes vérifiées par le modèle lorsque les rendements d'échelle sont constants ? Tester au niveau 5% H_0 : les rendements sont constants, contre H_1 : les rendements sont croissants.

Exercice 3.6 (Modèle à deux variables explicatives) On considère le modèle de régression suivant :

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

Les $x_{i,j}$ sont des variables exogènes du modèle, les ε_i sont des variables aléatoires indépendantes, de loi normale centrée admettant la même variance σ^2. En posant :

$$X = \begin{bmatrix} 1 & x_{1,2} & x_{1,3} \\ \vdots & \vdots & \vdots \\ 1 & x_{n,2} & x_{n,3} \end{bmatrix} \quad \text{et} \quad Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix},$$

on a observé :

$$X'X = \begin{bmatrix} 30 & 20 & 0 \\ 20 & 20 & 0 \\ 0 & 0 & 10 \end{bmatrix}, \quad X'Y = \begin{bmatrix} 15 \\ 20 \\ 10 \end{bmatrix}, \quad Y'Y = 59.5.$$

1. Déterminer la valeur de n, la moyenne des $x_{i,3}$, le coefficient de corrélation des $x_{i,2}$ et des $x_{i,3}$.
2. Estimer β_1, β_2, β_3, σ^2 par la méthode des moindres carrés ordinaires.
3. Calculer pour β_2 un intervalle de confiance à 95% et tester l'hypothèse $\beta_3 = 0.8$ au niveau 10%.
4. Tester $\beta_2 + \beta_3 = 3$ contre $\beta_2 + \beta_3 \neq 3$, au niveau 5%.
5. Que vaut $\bar{y}$, moyenne empirique des y_i ? En déduire le coefficient de détermination ajusté R_a^2.
6. Construire un intervalle de prévision à 95% de y_{n+1} connaissant : $x_{n+1,2} = 3$ et $x_{n+1,3} = 0,5$.

Exercice 3.7 (Modèle hétéroscédastique) On considère n observations $y_1, \ldots, y_n$ d'une variable définie sur une certaine population, et n $k-$uplets x_i ($x_i' = [x_{i1}, \ldots, x_{ik}]$) correspondant aux valeurs prises par k autres variables sur les mêmes éléments de cette population. On suppose que pour tout i, y_i est la valeur prise par une variable aléatoire Y_i, et qu'il existe $\beta \in \mathbb{R}^k$ pour lequel :

$$Y_i \sim \mathcal{N}\left(x_i'\beta, \sigma_i^2\right) \qquad 1 \leq i \leq n,$$

où :

- β représente un vecteur de $\mathbb{R}^k$: $\beta = [\beta_1, \ldots, \beta_k]'$,
- Les Y_i sont supposées indépendantes entre elles.

Enfin, les valeurs σ_i^2 des variances dépendent de l'appartenance à p sous-populations des éléments sur lesquels les variables sont observées. En regroupant les indices des Y_i selon ces sous-populations, on posera :

— $I_1 = \{1, \ldots, n_1\}$, indices des n_1 éléments de la première sous-population ;
— $I_2 = \{n_1 + 1, \ldots, n_1 + n_2\}$, indices des n_2 éléments de la deuxième sous-population ;
— ... ;
— $I_\ell = \{n_1 + \ldots + n_{\ell-1} + 1, \ldots, n_1 + \ldots + n_{\ell-1} + n_\ell\}$, indices des n_ℓ éléments de la ℓ-ème sous-population ;
— ... ;
— $I_p = \{n_1 + \ldots + n_{p-1} + 1, \ldots, n\}$, indices des n_p éléments de la dernière sous-population.

On admettra l'hypothèse suivante : si $i \in I_\ell$, $\sigma_i^2 = \ell\sigma^2$. Autrement dit, pour les n_1 variables correspondant aux éléments de la première sous-population la valeur est σ^2, pour les n_2 variables correspondant aux éléments de la deuxième sous-population la valeur est $2\sigma^2$, etc., jusqu'à $p\sigma^2$ pour la variance des variables correspondant aux éléments de la dernière sous-population. On veut estimer β et σ^2 par la méthode du maximum de vraisemblance. On notera $\hat{\beta}$, $\hat{\sigma}^2$ ces estimateurs.

1. Que vaut $f_{Y_i}(y_i)$, f_{Y_i} représentant la densité de la loi normale $\mathcal{N}\left(x_i'\beta, \sigma_i^2\right)$?
2. Montrer que $\hat{\beta}$ et $\hat{\sigma}^2$ sont solutions du système d'équations :
$$\left\{ \begin{array}{cc} \sum_{\ell=1}^{p} \frac{1}{\ell} \sum_{i \in I_\ell} (y_i - x_i'\beta)^2 = n\sigma^2 \\ \forall j = 1, \ldots, k \qquad \sum_{\ell=1}^{p} \frac{1}{\ell} \sum_{i \in I_\ell} (y_i - x_i'\beta)\, x_{ij} = 0. \end{array} \right. \tag{3.3}$$
3. Montrer que le système (3.3) équivaut à :
$$\left\{ \begin{array}{l} \|A(Y - X\beta)\|^2 = n\sigma^2 \\ X'A^2(Y - X\beta) = 0. \end{array} \right. \tag{3.4}$$
où $\|.\|^2$ représente la norme euclidienne usuelle dans $\mathbb{R}^n$, X la matrice $(n \times k)$ du plan d'expérience, Y le vecteur $(n \times 1)$ des observations y_i, A la matrice $(n \times n)$ diagonale dont l'élément (i,i) vaut $\frac{1}{\sqrt{l}}$ si $i \in I_l$.
4. En supposant que $\left(X'A^2X\right)$ est inversible, exprimer $\hat{\beta}$ et $\hat{\sigma}^2$.
5. Montrer que $n\hat{\sigma}^2 = \|V\|^2$, où V suit une loi gaussienne centrée.
6. En déduire que $\mathbb{E}\left[\|V\|^2\right]$ est la trace de la matrice de variance-covariance de V.
7. Montrer que $n\hat{\sigma}^2/(n-k)$ est un estimateur sans biais de σ^2.
8. On note X_ℓ la matrice $(n_\ell \times k)$ formée par les lignes d'indices I_ℓ de X, supposée de rang plein, Y_ℓ le vecteur colonne $(n_\ell \times 1)$ des composantes d'indices I_ℓ de Y. En posant $\hat{\beta}_\ell = (X_\ell' \cdot X_\ell)^{-1} X_\ell' Y_\ell$, montrer que $\hat{\beta}_\ell$ est un estimateur sans biais de β.
9. Que peut-on dire de la différence des matrices de variance-covariance de $\hat{\beta}_\ell$ et de $\hat{\beta}$?

Exercice 3.8 (La hauteur des eucalyptus) On souhaite expliquer la hauteur y (en mètres) d'un arbre en fonction de sa circonférence x (en centimètres) à 1m30 du sol et de la racine carrée de celle-ci. On a relevé $n = 1429$ couples (x_i, y_i), le nuage de points étant représenté figure 3.3. On

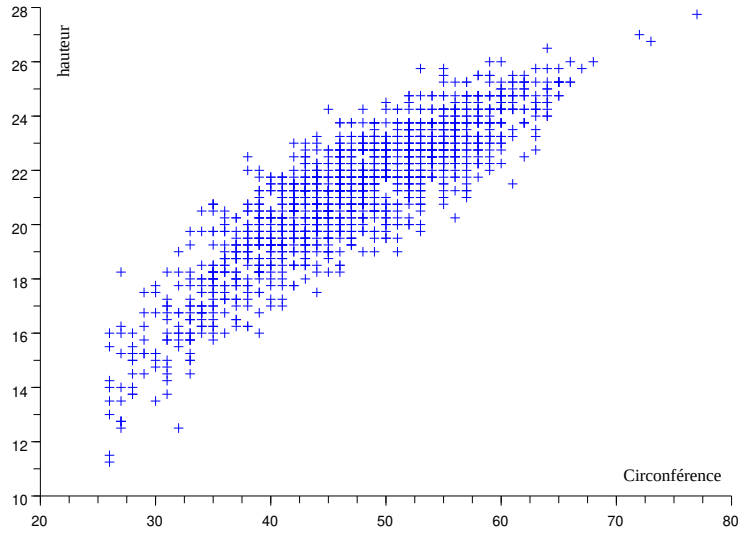


FIGURE 3.3 – Nuage de points pour les eucalyptus.

considère donc le modèle de régression suivant :
$$y_i = \beta_1 + \beta_2 x_i + \beta_3 \sqrt{x_i} + \varepsilon_i, \qquad 1 \leq i \leq n.$$

Les ε_i sont des variables aléatoires indépendantes, de loi normale centrée admettant la même variance σ^2. En posant :

$$X = \begin{bmatrix} 1 & x_1 & \sqrt{x_1} \\ \vdots & \vdots & \vdots \\ 1 & x_n & \sqrt{x_n} \end{bmatrix} \quad \text{et} \quad Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix},$$

on a observé :

$$X'X = \begin{bmatrix} ? & ? & 9792 \\ ? & 3306000 & ? \\ ? & 471200 & 67660 \end{bmatrix}, \quad X'Y = \begin{bmatrix} 30310 \\ 1462000 \\ 209700 \end{bmatrix}, \quad Y'Y = 651900.$$

1. Déterminer les '?' dans la matrice $X'X$.
2. Que vaut la circonférence moyenne empirique $\bar{x}$?
3. Le calcul donne (en arrondissant !)

$$\left(X'X\right)^{-1} = \begin{bmatrix} 4.646 & 0.101 & -1.379 \\ 0.101 & 0.002 & -0.030 \\ -1.379 & -0.030 & 0.411 \end{bmatrix} \quad \text{et} \quad (X'X)^{-1}X'Y = \begin{bmatrix} -16.8 \\ -0.30 \\ 7.62 \end{bmatrix}.$$

   Que valent les estimateurs $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_3$ par la méthode des moindres carrés ? Grâce au calcul de quelques points, représenter la courbe obtenue sur la figure 3.3.
4. Calculer l'estimateur de σ^2 pour les moindres carrés.
5. Calculer pour β_3 un intervalle de confiance à 95%.
6. Tester l'hypothèse $\beta_2 = 0$ au niveau de risque 10%.
7. Que vaut la hauteur moyenne empirique $\bar{y}$? En déduire le coefficient de détermination ajusté R^2_a.
8. Construire un intervalle de prévision à 95% de y_{n+1} connaissant $x_{n+1} = 49$.
9. Construire un intervalle de prévision à 95% de y_{n+1} connaissant $x_{n+1} = 25$.
10. Des deux intervalles précédents, lequel est le plus grand ? Pouvait-on s'y attendre ?

Exercice 3.9 (Consommation de gaz) Mr Derek Whiteside de la *UK Building Research Station* a collecté la consommation hebdomadaire de gaz et la température moyenne externe de sa maison au sud-est de l'Angleterre pendant une saison. Une régression pour expliquer la consommation de gaz en fonction de la température est réalisée avec le logiciel R. Les résultats numériques sont les suivants.

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.97802 -0.11082  0.02672  0.25294  0.63803

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.72385    0.12974    ?     < 2e-16 ***
Temp        -0.27793       ?     -11.04 1.05e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.3548 on 28 degrees of freedom
Multiple R-Squared: 0.8131,     Adjusted R-squared: 0.8064
F-statistic: 121.8 on 1 and 28 DF,  p-value: 1.046e-11
```

1. Donner le modèle et les hypothèses de la régression.
2. Compléter le tableau.
3. Soit Z une variable aléatoire de loi de Student de degré de liberté 28. Quelle est la probabilité que $|Z|$ soit supérieure à 11.04 ?
4. Préciser les éléments du test correspondant à la ligne "`Temp`" du tableau (H_0, H_1, la statistique de test, sa loi sous H_0, la règle de décision).
5. Interpréter le nombre "`Multiple R-Squared: 0.8131`" du tableau.
6. Donner une estimation de la variance du terme d'erreur dans le modèle de régression simple.
7. Expliquer et interpréter la dernière ligne du tableau :
   "`F-statistic: 121.8 on 1 and 28 DF, p-value: 1.046e-11`".
   Voyez-vous une autre façon d'obtenir cette p-value ?
8. Pensez-vous que la température extérieure a un effet sur la consommation de gaz ? Justifiez votre réponse.

Exercice 3.10 (Tests) Nous nous intéressons au modèle $Y = X\beta + \boldsymbol{\varepsilon}$ sous les hypothèses classiques. Nous avons obtenu sur 21 données :

$$\begin{aligned} \hat{y} &= 6.683_{(2.67)} + 0.44_{(2.32)}x_1 + 0.425_{(2.47)}x_2 + 0.171_{(2.09)}x_3 + 0.009_{(2.24)}x_4, \\ R^2 &= 0.54 \end{aligned}$$

où, pour chaque coefficient, le nombre entre parenthèses représente la valeur absolue de la statistique de test.

1. Quelles sont les hypothèses utilisées ?
2. Tester la nullité de β_1 au seuil de 5%.
3. Pouvez-vous tester $\mathrm{H}_0 : \beta_3 = 1$ contre $\mathrm{H}_1 : \beta_3 \neq 1$?
4. Tester la nullité simultanée des paramètres associés aux variables $x_1, \ldots, x_4$ au seuil de 5%.

Exercice 3.11 (Moindres carrés ordinaires)

1. Nous considérons le modèle de régression linéaire

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

   où $Y \in \mathbb{R}^n$, X est une matrice de taille $n \times p$ de rang p, $\beta \in \mathbb{R}^p$ et $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_n)$.
   (a) Qu'appelle-t-on estimateur des moindres carrés $\hat{\beta}$ de β ? Rappeler sa formule.
   (b) Quelle est l'interprétation géométrique de $\hat{Y} = X\hat{\beta}$ (faites un dessin) ?
   (c) Rappeler espérances et matrices de covariance de $\hat{\beta}$, $\hat{Y}$ et $\hat{\boldsymbol{\varepsilon}}$.
2. Nous considérons dorénavant un modèle avec 4 variables explicatives (la première variable étant la constante). Nous avons observé :

$$X'X = \begin{bmatrix} 100 & 20 & 0 & 0 \\ 20 & 20 & 0 & 0 \\ 0 & 0 & 10 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad X'Y = \begin{bmatrix} -60 \\ 20 \\ 10 \\ 1 \end{bmatrix}, \quad Y'Y = 159.$$

(a) Estimer β et σ^2.

(b) Donner un estimateur de la variance de $\hat{\beta}$.

(c) Donner un intervalle de confiance pour β_2, au niveau 95%.

(d) Calculer un intervalle de prévision de y_{n+1} au niveau 95% connaissant : $x_{n+1,2} = 3$, $x_{n+1,3} = 0.5$ et $x_{n+1,4} = 2$.

Exercice 3.12 (Moindres carrés pondérés) On suppose le modèle suivant

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

où X est la matrice $(n \times p)$ du plan d'expérience, $\beta = [\beta_1, \ldots, \beta_p]'$ un vecteur de $\mathbb{R}^p$, Y le vecteur $(n \times 1)$ des observations y_i, $\boldsymbol{\varepsilon}$ le vecteur $(n \times 1)$ des erreurs ε_i supposées centrées et de matrice de covariance $\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2\Omega^2$, où Ω est une matrice $(n \times n)$ diagonale dont l'élément (i, i) vaut $_i > 0$. Dans ce modèle, les valeurs $_i$ sont supposées connues, mais les paramètres β et σ^2 sont inconnus.

1. On considère le modèle transformé $Y^* = X^*\beta + \boldsymbol{\varepsilon}^*$, où :
   — $Y^* = [y_1^*, \ldots, y_n^*]'$, avec $y_i^* = y_i/\omega_i$;
   — X^* est la matrice $(n \times p)$ de terme générique $x_{ij}^* = x_{ij}/\omega_i$;
   — $\boldsymbol{\varepsilon}^* = [\varepsilon_1^*, \ldots, \varepsilon_n^*]'$, avec $\varepsilon_i^* = \varepsilon_i/\omega_i$;

   (a) Donner les relations entre X^* (respectivement Y^*, $\boldsymbol{\varepsilon}^*$), X (respectivement Y, $\boldsymbol{\varepsilon}$) et Ω.

   (b) Déterminer la moyenne et la matrice de covariance du vecteur aléatoire $\boldsymbol{\varepsilon}^*$.

   (c) En supposant $\left(X'\Omega^{-2}X\right)$ inversible, déterminer l'estimateur des moindres carrés $\hat{\beta}^*$ de β. Préciser son biais et sa matrice de covariance.

   (d) Proposer un estimateur sans biais $\hat{\sigma}_*^2$ de σ^2.

2. En revenant au modèle initial $Y = X\beta+\boldsymbol{\varepsilon}$, on suppose maintenant les erreurs ε_i gaussiennes, plus précisément $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2\Omega^2)$.

   (a) Donner la vraisemblance $\mathcal{L}(Y, \beta, \sigma^2)$ du modèle.

   (b) En déduire que les estimateurs au maximum de vraisemblance $\hat{\beta}_{mv}$ et $\hat{\sigma}_{mv}^2$ sont solutions de :
   $$\begin{cases} \left\|\Omega^{-1}\left(Y - X\beta\right)\right\|^2 = n\sigma^2 \\ X'\Omega^{-2}(Y - X\beta) = 0. \end{cases}$$

   (c) En déduire les relations entre $\hat{\beta}_{mv}$ et $\hat{\beta}^*$ d'une part, entre $\hat{\sigma}_{mv}^2$ et $\hat{\sigma}_*^2$ d'autre part.

   (d) Préciser alors la loi de $\hat{\beta}^*$. Que dire de celle de $\hat{\sigma}_*^2$?

3. Supposons maintenant le modèle classique de régression linéaire $Y = X\beta+\boldsymbol{\varepsilon}$, avec les erreurs centrées et de matrice de covariance $\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2 I_n$. Néanmoins, on n'observe pas comme d'habitude les x_i' et y_i, mais des moyennes par classe. Spécifiquement, les n données sont réparties en L classes $C_1, \ldots, C_L$ d'effectifs respectifs connus $n_1, \ldots, n_L$ et on a seulement accès aux moyennes par classe, à savoir pour tout $\ell \in \{1, \ldots, L\}$:
   $$\bar{y}_\ell = \frac{1}{n_\ell}\sum_{i \in C_\ell} y_i \qquad \& \qquad \bar{x}_{\ell j} = \frac{1}{n_\ell}\sum_{i \in C_\ell} x_{ij}$$

   (a) En notant $\bar{\varepsilon}_\ell = \frac{1}{n_\ell}\sum_{i \in C_\ell} \varepsilon_i$, vérifier que le modèle peut se mettre sous la forme $\bar{Y} = \bar{X}\beta + \bar{\boldsymbol{\varepsilon}}$.

   (b) Donner la moyenne et la matrice de covariance de $\bar{\boldsymbol{\varepsilon}}$.

   (c) Déduire des questions précédentes des estimateurs de β et σ^2.

Exercice 3.13 (Octopus's Garden) On cherche à mettre en œuvre une stratégie de prédiction du poids utile du poulpe, c'est-à-dire son poids éviscéré, à partir de son poids non éviscéré. C'est en effet le poulpe éviscéré qui est commercialisé. Pour cela, un échantillon de poulpes a été collecté en 2003 lors des opérations de pêche dans les eaux mauritaniennes. Vu l'importante différence de poids entre les poulpes mâles et les poulpes femelles, on étudie ici uniquement les données concernant 240 poulpes femelles.

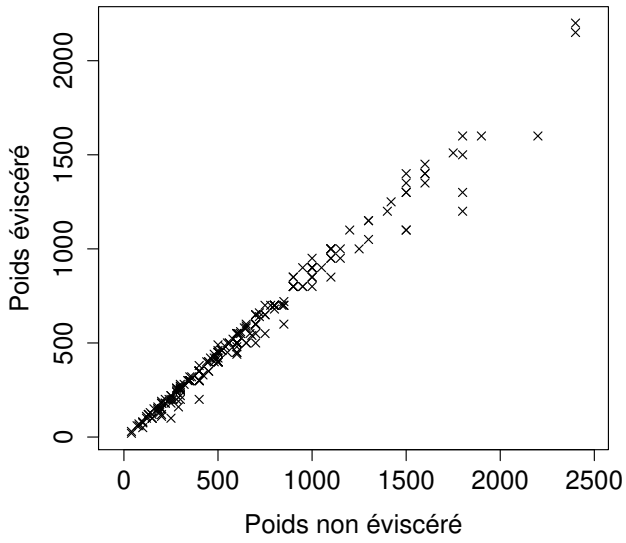


FIGURE 3.4 – Poids de poulpe éviscéré en fonction du poids non éviscéré (en grammes).

1. L'ensemble de ces données est représenté figure 3.4.
   (a) Proposer un modèle reliant le poids éviscéré et le poids non éviscéré d'un poulpe.
   (b) Rappeler les formules des estimateurs des paramètres du modèle.

```
                    Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)        -2.312146     5.959670    -0.388    0.698
Poids non éviscéré  0.853169     0.007649   111.545   <2e-16

Residual standard error: 52.73 on 238 degrees of freedom
Multiple R-Squared: 0.9812,     Adjusted R-squared: 0.9812
F-statistic: 1.244e+04 on 1 and 238 DF,  p-value: < 2.2e-16
```

TABLE 3.1 – Poids de poulpes éviscérés et non éviscérés : résultats de la régression linéaire simple (sortie R).

   (c) A partir du tableau 3.1, donner les estimations numériques des paramètres du modèle.
   (d) Que représente la valeur 0.698 du tableau 3.1 ? Comment la retrouver (à peu près) à partir de -0.388 et de la table de la loi normale donnée en annexe (faire un dessin).
   (e) Au vu de cette valeur 0.698, proposer un autre modèle reliant les poids éviscéré et non éviscéré.
2. De façon générale, considérons un échantillon de n couples de réels (x_i, y_i) suivant le modèle $y_i = \beta x_i + \varepsilon_i$, où les erreurs ε_i sont supposées gaussiennes indépendantes centrées et de même variance σ^2.
   (a) Déterminer l'estimateur $\tilde{\beta}$ de β minimisant la somme des carrés des écarts au modèle.

(b) Retrouver le résultat précédent à partir de la formule générale de l'estimateur de régression linéaire multiple en considérant la projection du vecteur $Y = [y_1, \ldots, y_n]'$ sur la droite vectorielle engendrée par le vecteur $X = [x_1, \ldots, x_n]'$.

(c) En déduire la variance de $\tilde{\beta}$. Proposer un estimateur non biaisé $\tilde{\sigma}^2$ de σ^2.

```
                       Estimate Std. Error t value Pr(>|t|)
Poids non éviscéré      0.85073    0.00436   195.1   <2e-16

Residual standard error: 52.63 on 239 degrees of freedom
Multiple R-Squared: 0.9938,     Adjusted R-squared: 0.9937
F-statistic: 3.807e+04 on 1 and 239 DF,  p-value: < 2.2e-16
```

TABLE 3.2 – Poids de poulpes éviscérés et non éviscérés : résultats de la régression linéaire simple avec le modèle simplifié (sortie R).

(d) Les résultats de l'analyse de ce nouveau modèle sont fournis dans le tableau 3.2. Localiser $\tilde{\beta}$ et $\tilde{\sigma}^2$ dans ce tableau.

(e) On veut prédire le poids éviscéré d'un poulpe de poids non éviscéré x_0. Quelle est la variance de l'erreur de prévision ? Donner un intervalle de confiance à 90% autour de la prévision.

Exercice 3.14 (Comparaison de modèles) On effectue une régression de y sur deux variables explicatives x et z à partir d'un échantillon de n individus, c'est-à-dire que $X = [\mathbb{1}, \boldsymbol{x}, \boldsymbol{z}]$, où $\mathbb{1}$ est le vecteur de taille n composé de 1. On a obtenu le résultat suivant :

$$X'X = \begin{bmatrix} 5 & 3 & 0 \\ 3 & 3 & 1 \\ 0 & 1 & 1 \end{bmatrix}.$$

1. Que vaut n ?
2. Que vaut le coefficient de corrélation linéaire empirique entre x et z ?
3. La régression par moindres carrés ordinaires a donné le résultat suivant

$$\hat{y}_i = -1 + 3x_i + 4z_i + \hat{\varepsilon}_i$$

et la somme des carrés résiduelle vaut $\|\hat{\boldsymbol{\varepsilon}}\|^2 = 3$.

(a) Exprimer $X'Y$ en fonction de $(X'X)$ et $\hat{\beta}$, et calculer $X'Y$. En déduire $\bar{y}$.

(b) Calculer $\|\hat{Y}\|^2$. En déduire $\|Y\|^2$.

(c) Calculer la somme des carrés totale $\|Y - \bar{y}\mathbb{1}\|^2$, le coefficient de détermination R^2 et le coefficient de détermination ajusté.

4. On s'intéresse maintenant au modèle privé du régresseur z, c'est-à-dire $Y = X_0\beta_0 + \boldsymbol{\varepsilon}_0$, où $X_0 = [\mathbb{1}, \boldsymbol{x}]$.

(a) Déterminer $X_0'X_0$ et $X_0'Y$. En déduire $\hat{\beta}_0$.

(b) Calculer $\|\hat{Y}_0\|^2$.

(c) Justifier l'égalité $\|\hat{Y}_0\|^2 + \|\hat{\boldsymbol{\varepsilon}}_0\|^2 = \|\hat{Y}\|^2 + \|\hat{\boldsymbol{\varepsilon}}\|^2$. En déduire $\|\hat{\boldsymbol{\varepsilon}}_0\|^2$, le coefficient de détermination R_0^2 et le coefficient de détermination ajusté.

5. On veut maintenant comparer les deux modèles précédents.

(a) Effectuer un test de Fisher entre ces deux modèles grâce aux coefficients de détermination. Qu'en concluez-vous au niveau de risque 5% ?

(b) Proposer un autre moyen d'arriver au même résultat.

3.7 Corrigés

Exercice 3.1 (QCM) ACABB.

Exercice 3.2 (Analyse de sorties logiciel)

1. Les résultats sont dans l'ordre :

$$6.2, 0.8, 6.66, -1.5, -1, 41, 5, 124.$$

2. La statistique de test de nullité du paramètre se trouve dans la troisième colonne, nous conservons H_0 pour les paramètres associés à `Ne9` et `Ne12`, et la rejetons pour les autres.
3. La statistique de test de nullité simultanée des paramètres autres que la constante vaut

$$F(\omega) = \frac{n-p}{p-1} \times \frac{R^2}{1-R^2} = \frac{124}{5} \times \frac{0.6233}{1-0.6233} \approx 41$$

Nous rejetons H_0.

4. Nous sommes en présence de modèles emboîtés, nous pouvons appliquer la formule vue dans le cours :

$$\begin{aligned} F &= \frac{n-p}{p-p_0} \times \frac{R^2 - R_0^2}{1-R^2} \\ &= \frac{124}{2} \times \frac{0.6233 - 0.5312}{1-0.6233} \approx 15. \end{aligned}$$

Nous rejettons H_0, i.e. nous conservons le premier modèle. Ainsi, bien que considérés séparément, les paramètres associés à `Ne9` et `Ne12` n'étaient pas significativement différents de 0, ce test montre qu'il serait imprudent de rejeter ces deux variables en même temps.

Ceci n'est pas étonnant : les variables `Ne9` et `Ne12` sont fortement corrélées (peu de changement de nébulosité en 3 heures), si bien que lorsque l'une est dans le modèle, l'autre apporte peu d'information supplémentaire. Or le test de Student de nullité d'un coefficient teste justement la pertinence d'une variable lorsque toutes les autres sont présentes. Le test de Fisher, par contre, nous apprend que l'information apportée par ces variables n'est pas négligeable. Au total, la solution serait donc de conserver l'une des deux variables et de supprimer l'autre.

Dernière remarque : la preuve de la Proposition 10 montre que le test de Fisher entre modèles emboîtés est lié aux régions de confiance simultanées définies par la Propriété 6 (*ii*). Dans notre cas précis, la conclusion est la suivante : si l'on traçait le rectangle de confiance à 95% issu des intervalles de confiance de $\beta_{\texttt{Ne9}}$ et $\beta_{\texttt{Ne12}}$, alors le point $(0,0)$ serait dans ce rectangle. Par contre, il ne serait pas dans l'ellipse de confiance à 95%. On voit donc sur cet exemple la pertinence des régions de confiance simultanées lorsqu'on a affaire à des variables très corrélées.

Exercice 3.3 (Equivalence du test T et du test F)

1. Sous l'hypothèse $H_0 : \beta_p = 0$, le test de Student s'écrit

$$T = \frac{\hat{\beta}_p}{\hat{\sigma}_p} \sim \mathcal{T}_{n-p}$$

où $\hat{\sigma}_p$ est l'estimateur de l'écart-type de $\hat{\beta}_p$, c'est-à-dire

$$\hat{\sigma}_p = \hat{\sigma}_{\hat{\beta}_p} = \hat{\sigma}\sqrt{(X'X)^{-1}_{p,p}}.$$

2. Sous l'hypothèse $\mathrm{H}_0 : \beta_p = 0$, le test de Fisher s'écrit

$$F = \frac{\|\hat{Y} - \hat{Y}_0\|^2}{\hat{\sigma}^2} \sim \mathcal{F}^1_{n-p}$$

où $\hat{Y}_0$ est le projeté orthogonal de Y sur le sous-espace engendré par les $(p-1)$ premières colonnes de X.

3. Soit $Z \sim \mathcal{N}(0,1)$ indépendante de $S_n \sim \chi^2_n$, alors par définition

$$T_n = \frac{Z}{\sqrt{S_n/n}} \sim \mathcal{T}_n,$$

loi de Student à n degrés de liberté. Il suffit alors de voir que Z^2 suit une loi du chi-deux à un seul degré de liberté pour en déduire que

$$F_n = T_n^2 = \frac{Z^2}{S_n/n} \sim \mathcal{F}^1_n,$$

loi de Fisher à 1 et n degrés de liberté. En particulier, les quantiles d'une loi de Fisher à 1 et n degrés de liberté sont les carrés des quantiles d'une loi de Student à n degrés de liberté.

4. Avec les notations de l'énoncé, la matrice $X'X$ sous forme blocs comme suit

$$X'X \quad = \quad \left(\begin{array}{c|c} X_0'X_0 & X_0'X_p \\ \hline X_p'X_0 & X_p'X_p \end{array}\right).$$

5. La formule d'inversion matricielle par blocs (B.2) rappelée en Annexe donne alors pour le dernier coefficient diagonal

$$[(X'X)^{-1}]_{pp} = \left(X_p'X_p - X_p'X_0(X_0'X_0)^{-1}X_0'X_p\right)^{-1} = \left(X_p'(I_n - X_0(X_0'X_0)^{-1}X_0')X_p\right)^{-1}$$

d'où

$$\frac{1}{[(X'X)^{-1}]_{pp}} = X_p'(I_n - P_0)X_p$$

où $P_0 = X_0(X_0'X_0)^{-1}X_0'$ est la matrice $n \times n$ de projection orthogonale sur l'espace $\mathcal{M}_0$ engendré par les $(p-1)$ colonnes de X_0.

6. Puisque les $(p-1)$ colonnes de X_0 correspondent aux p premières colonnes de X, il est clair que $\mathcal{M}_0$ est un sous-espace de $\mathcal{M}$ donc que $P_0P = P_0$. Puisque par définition $\hat{Y}_0 = P_0Y$ et $\hat{Y} = PY$, on en déduit que

$$\hat{Y}_0 = P_0Y = P_0PY = P_0\hat{Y}.$$

Décomposons $\hat{Y}$ sur la base des p vecteurs de $\mathcal{M}$

$$\hat{Y} = X\hat{\beta} = \hat{\beta}_1X_1 + \cdots + \hat{\beta}_{p-1}X_{p-1} + \hat{\beta}_pX_p$$

alors par linéarité de P_0 et puisque $P_0X_j = X_j$ pour tout $j \in \{1, \ldots, p-1\}$

$$\hat{Y}_0 = P_0\hat{Y} = P_0(\hat{\beta}_1X_1 + \cdots + \hat{\beta}_{p-1}X_{p-1} + \hat{\beta}_pX_p) = \hat{\beta}_1X_1 + \cdots + \hat{\beta}_{p-1}X_{p-1} + \hat{\beta}_pP_0X_p$$

de sorte que

$$\hat{Y} - \hat{Y}_0 = \hat{\beta}_pX_p - \hat{\beta}_pP_0X_p = \hat{\beta}_p(I_n - P_0)X_p$$

7. La question précédente permet d'écrire

$$\|\hat{Y} - \hat{Y}_0\|^2 = \|\hat{\beta}_p(I_n - P_0)X_p\|^2 = \hat{\beta}_p^2((I_n - P_0)X_p)'((I_n - P_0)X_p) = \hat{\beta}_p^2 X_p'(I_n - P_0)X_p$$

la dernière égalité venant de ce que $(I_n - P_0)$ est la projection orthogonale sur $\mathcal{M}_0^\perp$

$$(I_n - P_0)'(I_n - P_0) = (I_n - P_0)^2 = (I_n - P_0).$$

La comparaison montre bien que $F = T^2$. En d'autres termes, les deux tests de nullité du coefficient β_p sont complètement équivalents. Notons cependant que si on veut effectuer un test unilatéral, c'est Student qui s'impose.

Exercice 3.4 (Un modèle à 3 variables explicatives) 1. L'estimateur des moindres carrés de β est donné par $\hat{\beta} = (X'X)^{-1}X'Y$. La matrice $(X'X)^{-1}$ a la même forme que $X'X$, c'est-à-dire diagonale par blocs avec pour premier bloc diagonal le coefficient $((X'X)^{-1})_{1,1} = 1/50$ et comme second bloc diagonal la matrice 3×3 donnée dans l'énoncé :

$$\frac{1}{13720}\begin{bmatrix} 1100 & -560 & 30 \\ -560 & 784 & -140 \\ 30 & -140 & 375 \end{bmatrix}.$$

Il en résulte que

$$\hat{\beta} \approx \begin{bmatrix} 2 \\ 2.55 \\ -0.57 \\ 1.89 \end{bmatrix}.$$

La somme des carrés des résidus $\sum_{i=1}^{50} \hat{\varepsilon}_i^2$ s'écrit encore

$$\|\hat{\boldsymbol{\varepsilon}}\|^2 = \|Y - \hat{Y}\|^2 = \|Y\|^2 - \|\hat{Y}\|^2,$$

cette dernière relation découlant de Pythagore. Le premier terme ne pose pas problème puisque $\|Y\|^2 = Y'Y = 640$. Pour le second, il suffit de remarquer que

$$\|\hat{Y}\|^2 = \|X\hat{\beta}\|^2 = \hat{\beta}'(X'X)\hat{\beta} = Y'X\hat{\beta} \approx 458.$$

Ainsi la somme des carrés des résidus vaut $\|\hat{\boldsymbol{\varepsilon}}\|^2 \approx 182$. On en déduit que l'estimateur de σ^2 vaut

$$\hat{\sigma}^2 = \frac{\|\hat{\boldsymbol{\varepsilon}}\|^2}{46} \approx 3.96.$$

2. Puisqu'on sait que :

$$\frac{\hat{\beta}_2 - \beta_2}{\hat{\sigma}_2} = \frac{\hat{\beta}_2 - \beta_2}{\hat{\sigma}\sqrt{(X'X)^{-1}_{2,2}}} \sim \mathcal{T}_{n-4} = \mathcal{T}_{46},$$

on en déduit qu'un intervalle de confiance à 95% pour β_2 est :

$$I(\beta_2) = \left[\hat{\beta}_2 - t_{46}(0.975)\hat{\sigma}\sqrt{(X'X)^{-1}_{2,2}}; \hat{\beta}_2 + t_{46}(0.975)\hat{\sigma}\sqrt{(X'X)^{-1}_{2,2}}\right],$$

c'est-à-dire :

$$I(\beta_2) \approx \left[2.55 - 2.0\sqrt{3.96}\sqrt{1100/13720}; 2.55 + 2.0\sqrt{3.96}\sqrt{1100/13720}\right] \approx [1.42; 3.68].$$

Un intervalle de confiance à 95% pour σ^2 est :

$$I(\sigma^2) = \left[\frac{46\ \hat{\sigma}^2}{c_2}, \frac{46\ \hat{\sigma}^2}{c_1}\right] = \left[\frac{\|\hat{\varepsilon}\|^2}{c_2}, \frac{\|\hat{\varepsilon}\|^2}{c_1}\right],$$

où c_1 et c_2 sont tels que $\mathbb{P}(c_1 \leq \chi^2_{46} \leq c_2) = 0.95$. En l'occurrence, on trouve $c_1 \approx 29$ et $c_2 \approx 66$, ce qui donne $I(\sigma^2) \approx [2.76; 6.28]$.

3. Le test de "validité globale" du modèle au niveau 5% peut se faire via la statistique de Fisher. Sous l'hypothèse de nullité de tous les coefficients sauf la constante, on a en effet :

$$F(\omega) = \frac{n-p}{p-1} \times \frac{\|\hat{Y} - \bar{y}\mathbb{1}\|^2}{\|Y - \hat{Y}\|^2} = \frac{46}{3} \times \frac{\|\hat{Y} - \bar{y}\mathbb{1}\|^2}{\|Y - \hat{Y}\|^2} \sim \mathcal{F}^3_{46}$$

Or $\bar{y} = 100/50 = 2$ se lit sur la première coordonnée du vecteur $X'Y$, et la constante faisant partie du modèle, il vient

$$\|\hat{Y} - \bar{y}\mathbb{1}\|^2 = \|\hat{Y}\|^2 - \|\bar{y}\mathbb{1}\|^2 = \|\hat{Y}\|^2 - 50\bar{y}^2 \approx 458 - 200 = 258.$$

D'autre part, on a déjà vu que $\|Y - \hat{Y}\|^2 \approx 182$. D'où $F(\omega) \approx 21.7$. Or le quantile d'ordre 0.95 d'une Fisher à $(3, 46)$ degrés de liberté vaut environ 2.81. L'hypothèse $(\beta_2 = \beta_3 = \beta_4 = 0)$ est donc rejetée.

4. En notant $x'_{51} = [1, 1, -1, 0.5]$, la valeur prédite pour y_{51} est :

$$\hat{y}_{51} = x'_{51}\hat{\beta} \approx 6.07$$

et un intervalle de prévision à 95% pour y_{51} est :

$$I(y_{51}) = \left[\hat{y}_{51} - t_{46}(0.975)\hat{\sigma}\sqrt{1 + x'_{51}(X'X)^{-1}x_{51}}; \hat{y}_{51} + t_{46}(0.975)\hat{\sigma}\sqrt{1 + x'_{51}(X'X)^{-1}x_{51}}\right],$$

soit $I \approx [1.61; 10.53]$.

Exercice 3.5 (Modèle de Cobb-Douglas) 1. Avec les notations

$$Y = \begin{bmatrix} \log V_1 \\ \vdots \\ \log V_n \end{bmatrix} \qquad X = \begin{bmatrix} 1 & \log L_1 & \log K_1 \\ \vdots & \vdots & \vdots \\ 1 & \log L_n & \log K_n \end{bmatrix} \qquad b = \begin{bmatrix} \alpha \\ \beta \\ \gamma \end{bmatrix} \qquad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

le modèle s'écrit sous la forme matricielle $Y = Xb + \boldsymbol{\varepsilon}$. L'estimateur des MCO $\hat{b}$ s'écrit alors comme d'habitude $\hat{b} = (X'X)^{-1}XY$. Sa matrice de variance-covariance est $\text{Var}(\hat{b}) = \sigma^2(X'X)^{-1}$. En notant $\hat{Y} = X\hat{b}$, un estimateur sans biais de σ^2 est

$$\hat{\sigma}^2 = \frac{\|Y - \hat{Y}\|^2}{n-3},$$

et un estimateur sans biais de $\text{Var}(\hat{b})$ est $\hat{\sigma}^2(X'X)^{-1}$.

2. L'estimateur $\hat{\sigma}^2$ se déduit de la somme des carrés résiduelle :

$$\hat{\sigma}^2 = \frac{SCR}{n-3} = \frac{148.27}{1655} \approx 0.09$$

Une estimation de $\text{Var}(\hat{b})$ est donc

$$\hat{V}(\hat{b}) = 0.09\ (X'X)^{-1} = 0.09 \begin{bmatrix} 0.0288 & 0.0012 & -0.0034 \\ 0.0012 & 0.0016 & -0.0010 \\ -0.0034 & -0.0010 & 0.0009 \end{bmatrix}.$$

3. Puisqu'on sait que :
$$\frac{\hat{\beta}-\beta}{\hat{\sigma}_{\hat{\beta}}} = \frac{\hat{\beta}-\beta}{\hat{\sigma}\sqrt{(X'X)^{-1}_{2,2}}} \sim \mathcal{T}_{n-3} = \mathcal{T}_{1655},$$
on en déduit qu'un intervalle de confiance à 95% pour β est :
$$I(\beta) = \left[\hat{\beta} - t_{1655}(0.975)\hat{\sigma}\sqrt{(X'X)^{-1}_{2,2}}; \hat{\beta} + t_{1655}(0.975)\hat{\sigma}\sqrt{(X'X)^{-1}_{2,2}}\right],$$
c'est-à-dire :
$$I(\beta) \approx \left[0.738 - 1.96\sqrt{0.09}\sqrt{0.0016}; 0.738 + 1.96\sqrt{0.09}\sqrt{0.0016}\right] \approx [0.71; 0.76].$$
De même, un intervalle de confiance à 95% pour γ est
$$I(\gamma) \approx \left[0.282 - 1.96\sqrt{0.09}\sqrt{0.0009}; 0.282 + 1.96\sqrt{0.09}\sqrt{0.0009}\right] \approx [0.26; 0.30].$$

4. Sous $H_0 : \gamma = 0$, on sait que
$$\frac{\hat{\gamma}}{\hat{\sigma}_{\hat{\gamma}}} = \frac{\hat{\gamma}}{\hat{\sigma}\sqrt{(X'X)^{-1}_{3,3}}} \sim \mathcal{T}_{n-3} = \mathcal{T}_{1655}.$$
On obtient une statistique de test égale à
$$T(\omega) = \frac{0.282}{\sqrt{0.09}\sqrt{0.0009}} \approx 31.3 > 1.645 = t_{1655}(0.95),$$
quantile d'ordre 0.95 d'une loi de Student à 1655 degrés de liberté. Nous rejetons donc l'hypothèse H_0.

5. Puisque $V = F(L,K) = \lambda L^\beta K^\gamma$, il vient directement
$$F(\alpha L, \alpha K) = \alpha^{\beta+\gamma}\lambda L^\beta K^\gamma = \alpha^{\beta+\gamma}F(L,K).$$
Donc dire que le rendement d'échelle est constant, c'est encore dire que $\beta + \gamma = 1$. A contrario, les rendements sont croissants si $F(\alpha L, \alpha K) > \alpha F(L,K)$, c'est-à-dire que $\beta+\gamma > 1$. Nous allons donc tester au niveau 5% $H_0 : \beta + \gamma = 1$, contre $H_1 : \beta + \gamma > 1$.
Sous l'hypothèse H_0, nous savons que
$$\frac{\hat{\beta}+\hat{\gamma}-1}{\hat{\sigma}_{\hat{\beta}+\hat{\gamma}}} \sim \mathcal{T}_{n-3} = \mathcal{T}_{1655},$$
dont le quantile d'ordre 0.95 est 1.645. Il nous suffit donc de calculer $\hat{\sigma}_{\hat{\beta}+\hat{\gamma}}$. Or de façon générale, on a la décomposition :
$$\mathrm{Var}(\hat{\beta}+\hat{\gamma}) = \mathrm{Var}(\hat{\beta}) + \mathrm{Var}(\hat{\gamma}) + 2\mathrm{Cov}(\hat{\beta},\hat{\gamma}),$$
donc l'estimateur cherché est $\hat{\sigma}_{\hat{\beta}+\hat{\gamma}} = \sqrt{\hat{\mathrm{Var}}(\hat{\beta}+\hat{\gamma})}$, où :
$$\hat{\mathrm{Var}}(\hat{\beta}+\hat{\gamma}) = \hat{\mathrm{Var}}(\hat{\beta}) + \hat{\mathrm{Var}}(\hat{\gamma}) + 2\hat{\mathrm{Cov}}(\hat{\beta},\hat{\gamma}),$$
quantités qui se déduisent de la matrice $\hat{V}(\hat{b})$ calculée précédemment. Ceci donne
$$\hat{\mathrm{Var}}(\hat{\beta}+\hat{\gamma}) \approx 0.09(0.0016 + 0.0009 - 2\times 0.001) = 4.5\times 10^{-5}.$$

On en déduit que la statistique de test vaut

$$T(\omega) = \frac{0.738 + 0.282 - 1}{\sqrt{4.5 \times 10^{-5}}} \approx 2.98 > 1.645$$

En conclusion, l'hypothèse selon laquelle les rendements seraient constants est refusée. Au niveau 5%, on accepte l'hypothèse selon laquelle ils sont croissants.

Exercice 3.6 (Modèle à deux variables explicatives) Cet exercice est corrigé en annexe (sujet de décembre 2009).

Exercice 3.7 (Modèle hétéroscédastique) Cet exercice est corrigé en annexe (sujet de décembre 2009).

Exercice 3.8 (La hauteur des eucalyptus) Cet exercice est corrigé en annexe (sujet de décembre 2010).

Exercice 3.9 (Consommation de gaz) Cet exercice est corrigé en annexe (sujet de décembre 2010).

Exercice 3.10 (Tests) Cet exercice est corrigé en annexe (sujet de décembre 2011).

Exercice 3.11 (Moindres carrés ordinaires) Cet exercice est corrigé en annexe (sujet de décembre 2011).

Exercice 3.12 (Moindres carrés pondérés) Cet exercice est corrigé en annexe (sujet de décembre 2011).

Exercice 3.13 (Octopus's Garden) Cet exercice est corrigé en annexe (sujet de décembre 2012).

Exercice 3.14 (Comparaison de modèles) Cet exercice est corrigé en annexe (sujet de décembre 2012).

Chapitre 4

Validation du modèle

Introduction

En présence d'un échantillon de n observations $(x_i, y_i)_{1\leq i\leq n}$ à valeurs dans $\mathbb{R}^p \times \mathbb{R}$, les grandes étapes de la régression linéaire sont les suivantes :

1. **Modélisation.** Nous considérons un modèle de la forme :

$$\forall i \in \{1, \cdots, n\} \qquad y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i = x_i'\beta + \varepsilon_i,$$

qui se réécrit sous forme matricielle :

$$Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \boldsymbol{\varepsilon}_{n\times 1},$$

sous les hypothèses :

$$(\mathcal{H}) \begin{cases} (\mathcal{H}_1) : \mathrm{rg}(X) = p \\ (\mathcal{H}_2) : \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_n) \end{cases}$$

2. **Estimation.** Nous estimons alors les paramètres β et σ^2 par la méthode des moindres carrés, laquelle est grosso modo équivalente à la méthode du maximum de vraisemblance, ce qui donne les estimateurs $\hat{\beta}$ et $\hat{\sigma}^2$. Des lois de $\hat{\beta}$ et $\hat{\sigma}^2$, nous avons déduit des intervalles et/ou régions de confiance pour β et σ^2, et avons pu construire des tests d'hypothèses.
3. **Validation.** Les deux premiers points étant acquis, il s'agit dans ce chapitre de valider nos hypothèses. Autant la vérification de $(\mathcal{H}_1)$ ne pose pas problème, autant celle de $(\mathcal{H}_2)$ s'avère délicate. Nous nous contenterons donc de donner quelques pistes.

4.1 Analyse des résidus

L'examen des résidus constitue une étape primordiale de la régression linéaire. Cette étape étant essentiellement fondée sur des méthodes graphiques, il est difficile d'avoir des règles strictes de décision. L'objectif de cette partie est de présenter ces méthodes graphiques. Commençons par rappeler les définitions des différents résidus.

4.1.1 Résidus et valeurs aberrantes

Les erreurs ε_i sont estimées par $\hat{\varepsilon}_i = y_i - \hat{y}_i$. En notant $H = P_X = X(X'X)^{-1}X'$ la matrice de projection et h_{ij} son terme générique, nous avons :

Erreurs	Résidus
$\mathbb{E}[\varepsilon_i] = 0$	$\mathbb{E}[\hat{\varepsilon}_i] = 0$
$\text{Var}(\boldsymbol{\varepsilon}) = \sigma^2 I$	$\text{Var}(\hat{\boldsymbol{\varepsilon}}) = \sigma^2(I - H)$

Il s'ensuit que la variance de $\hat{\varepsilon}_i$ est $\text{Var}(\hat{\varepsilon}_i) = \sigma^2(1 - h_{ii})$, qui dépend donc de i. Afin d'éliminer cette non-homogénéité des variances des résidus, nous préférerions utiliser les résidus **normalisés** :

$$r_i = \frac{\hat{\varepsilon}_i}{\sigma\sqrt{1 - h_{ii}}}.$$

Mais σ est inconnu, il convient donc de le remplacer par $\hat{\sigma}$, ce qui donne des résidus dits **standardisés** :

$$t_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}.$$

Puisqu'on a simplement remplacé σ par son estimée $\hat{\sigma}$, on pourrait croire que ces résidus suivent une loi de Student : patatras, il n'en est rien ! C'est pourquoi nous utiliserons plutôt les résidus **studentisés**, souvent appelés **studentized residuals** dans les logiciels et définis par :

$$t_i^* = \frac{\hat{\varepsilon}_i}{\hat{\sigma}_{(i)}\sqrt{1 - h_{ii}}},$$

où $\hat{\sigma}_{(i)}$ est l'estimateur de σ dans le modèle linéaire privé de l'observation i.

Ces résidus t_i^* suivent bien une loi de Student (cf. Théorème 8 ci-après). Ils sont construits selon la logique de validation croisée (en abrégé VC, ou plus précisément méthode du *leave-one-out*), c'est-à-dire comme suit :

1. Dans un premier temps, nous estimons les paramètres β et σ^2 à l'aide de tous les individus sauf le $i^{ème}$, nous obtenons ainsi les estimateurs $\hat{\beta}_{(i)}$ et $\hat{\sigma}^2_{(i)}$;
2. Dans un second temps, nous considérons que la $i^{ème}$ observation $x_i' = [x_{i1}, \ldots, x_{ip}]$ est une nouvelle observation et nous prévoyons y_i par $\hat{y}_i^p$ de façon classique : $\hat{y}_i^p = x_i'\hat{\beta}_{(i)}$.

Le chapitre précédent permet alors de préciser la loi suivante :

$$\frac{y_i - \hat{y}_i^p}{\hat{\sigma}_{(i)}\sqrt{1 + x_i'(X_{(i)}'X_{(i)})^{-1}x_i}} \sim \mathcal{T}_{n-p-1},$$

loi de Student à $(n-p-1)$ ddl puisque les estimateurs $\hat{\beta}_{(i)}$ et $\hat{\sigma}^2_{(i)}$ sont construits à partir de $(n-1)$ observations. Nous allons maintenant montrer que les résidus studentisés par validation croisée t_i^* correspondent exactement à ces erreurs de prévision normalisées.

Théorème 8 *Si la matrice X est de plein rang et si la suppression de la ligne i ne modifie pas le rang de la matrice, alors les résidus studentisés par validation croisée vérifient :*

$$t_i^* = \frac{\hat{\varepsilon}_i}{\hat{\sigma}_{(i)}\sqrt{1 - h_{ii}}} = \frac{y_i - \hat{y}_i}{\hat{\sigma}_{(i)}\sqrt{1 - h_{ii}}} = \frac{y_i - \hat{y}_i^p}{\hat{\sigma}_{(i)}\sqrt{1 + x_i'(X_{(i)}'X_{(i)})^{-1}x_i}} \sim \mathcal{T}_{n-p-1}.$$

Preuve. Nous considérons la matrice X du plan d'expérience, de taille $n \times p$, $X_{(i)}$ la matrice X privée de la $i^{ème}$ ligne x_i', donc de taille $(n-1) \times p$, et $Y_{(i)}$ le vecteur Y privé de sa $i^{ème}$ coordonnée, donc de taille $(n-1) \times 1$. Nous aurons alors besoin des ingrédients matriciels suivants, dont la vérification est laissée au lecteur :

1. **Lemme d'inversion matricielle** : Soit M une matrice symétrique inversible de taille $p \times p$ et u et v deux vecteurs de taille p, alors :

$$\left(M + uv'\right)^{-1} = M^{-1} - \frac{M^{-1}uv'M^{-1}}{1 + u'M^{-1}v}.$$

2. $X'X = X'_{(i)}X_{(i)} + x_i x'_i$.
3. $X'Y = X'_{(i)}Y_{(i)} + x_i y_i$.
4. $h_{ii} = x'_i(X'X)^{-1}x_i$.

Dans notre situation, le lemme d'inversion matricielle s'écrit :

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

et la relation sur h_{ii} ci-dessus donne :

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

Calculons alors la prévision $\hat{y}_i^p$, où $\hat{\beta}_{(i)}$ est l'estimateur de β obtenu sans la $i^{ème}$ observation :

$$\begin{aligned}
\hat{y}_i^p = x'_i\hat{\beta}_{(i)} &= x'_i(X'_{(i)}X_{(i)})^{-1}X'_{(i)}Y_{(i)} \\
&= x'_i\left[(X'X)^{-1} + \frac{(X'X)^{-1}x_i x'_i(X'X)^{-1}}{1 - h_{ii}}\right]\left(X'Y - x_i y_i\right) \\
&= x'_i\hat{\beta} + \frac{h_{ii}}{1 - h_{ii}}x'_i\hat{\beta} - h_{ii}y_i - \frac{h_{ii}^2}{1 - h_{ii}}y_i \\
&= \frac{1}{1 - h_{ii}}\hat{y}_i - \frac{h_{ii}}{1 - h_{ii}}y_i.
\end{aligned}$$

On déduit de cette dernière relation que $\hat{\varepsilon}_i = y_i - \hat{y}_i = (1 - h_{ii})(y_i - \hat{y}_i^p)$, d'où il vient :

$$t_i^* = \frac{\hat{\varepsilon}_i}{\hat{\sigma}_{(i)}\sqrt{1 - h_{ii}}} = \frac{\sqrt{(1 - h_{ii})}(y_i - \hat{y}_i^p)}{\hat{\sigma}_{(i)}}.$$

Pour terminer, remarquons qu'en multipliant la relation obtenue ci-dessus pour $(X'_{(i)}X_{(i)})^{-1}$ à gauche par x'_i et à droite par x_i, on obtient :

$$\begin{aligned}
x'_i(X'_{(i)}X_{(i)})^{-1}x_i &= h_{ii} + \frac{h_{ii}^2}{1 - h_{ii}}. \\
1 + x'_i(X'_{(i)}X_{(i)})^{-1}x_i &= 1 + \frac{h_{ii}}{1 - h_{ii}} = \frac{1}{1 - h_{ii}},
\end{aligned}$$

ce qui permet d'établir l'égalité :

$$t_i^* = \frac{\hat{\varepsilon}_i}{\hat{\sigma}_{(i)}\sqrt{1 - h_{ii}}} = \frac{y_i - \hat{y}_i}{\hat{\sigma}_{(i)}\sqrt{1 - h_{ii}}} = \frac{y_i - \hat{y}_i^p}{\hat{\sigma}_{(i)}\sqrt{1 + x'_i(X'_{(i)}X_{(i)})^{-1}x_i}}.$$

Le résultat sur la loi de l'erreur de prévision vu au chapitre précédent s'applique alors directement et ceci achève la preuve.

■

En conclusion, bien que les résidus utilisés soient souvent les $\hat{\varepsilon}_i$, ceux-ci n'ont pas la même variance selon l'observation i et sont donc à déconseiller. Afin de remédier à cette hétéroscédasticité, nous préférerons utiliser les résidus studentisés t_i^* pour détecter des valeurs aberrantes.

Remarque. D'un point de vue algorithmique, et contrairement aux t_i, les t_i^* semblent coûteux puisque chacun nécessite le calcul de $\hat{\sigma}_{(i)}$. On peut en fait montrer la relation :

$$t_i^* = t_i \sqrt{\frac{n-p-1}{n-p-t_i^2}},$$

qui assure qu'on ne paie rien de plus en temps de calcul à remplacer les t_i par les t_i^* (voir par exemple l'article d'Atkinson [2]). Notons aussi sur cette formule que les t_i^* sont une fonction croissante des t_i. En d'autres termes, les plus grandes valeurs des résidus studentisés correspondent aux plus grandes valeurs des résidus standardisés.

Une valeur aberrante est une observation qui est mal expliquée par le modèle et qui conduit à un résidu élevé en ce point. Nous pouvons donc la définir grâce aux résidus studentisés t_i^*.

Définition 11 *Une donnée aberrante est un point (x_i, y_i) pour lequel le résidu studentisé par validation croisée t_i^* est élevé comparé au seuil donné par la loi de Student : $|t_i^*| \gg t_{n-p-1}(1-\alpha/2)$.*

Remarque. En pratique, si $\alpha = 5\%$ et $(n-p-1) \geq 30$, alors $t_{n-p-1}(1-\alpha/2) \approx 2$.

Généralement, les données aberrantes sont détectées en traçant les t_i^* séquentiellement ou en fonction d'autres variables ($y_i, x_i, \hat{y}_i$, etc.). La détection des données aberrantes ne dépend que de la valeur des résidus. Ces représentations graphiques permettent de s'assurer aussi de la validité du modèle.

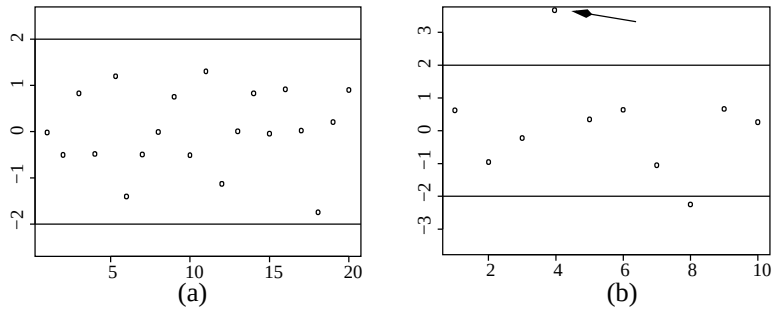


FIGURE 4.1 – Résidus studentisés corrects (figure a) et résidus studentisés avec un individu aberrant à vérifier, signalé par une flèche, et un second moins important (figure b).

La figure 4.1(a) montre un ajustement satisfaisant où aucune structure ne se dégage des résidus et où aucun résidu n'est plus grand que la valeur test 2. Remarquons qu'en théorie $\alpha\%$ des individus possèdent des valeurs aberrantes. Nous cherchons donc plutôt les résidus dont les valeurs absolues sont nettement au-dessus de $t_{n-p-1}(1-\alpha/2) \approx 2$. Ainsi, en figure 4.1(b), nous nous intéresserons seulement à l'individu désigné par une flèche.

4.1.2 Analyse de la normalité

L'hypothèse de normalité est difficile à vérifier. Notons déjà que si les erreurs ε_i sont indépendantes de loi normale $\mathcal{N}(0, \sigma^2)$, les résidus studentisés t_i^* suivent eux une loi de Student et ne sont pas

indépendants. Néanmoins, si $n \gg p$, cette loi de Student est quasiment une loi normale.

L'aspect "quasi" gaussien des t_i^* peut alors être examiné de plusieurs façons. Un histogramme est la méthode la plus grossière. Citons aussi le graphique comparant les quantiles des résidus estimés à l'espérance des mêmes quantiles sous l'hypothèse de normalité. Ce type de graphique est appelé Q-Q plot (ou diagramme quantile-quantile).

4.1.3 Analyse de l'homoscédasticité

Il n'existe pas de procédure précise pour vérifier l'hypothèse d'homoscédasticité. Nous proposons plusieurs graphiques possibles pour détecter une hétéroscédasticité. Il est recommandé de tracer les résidus studentisés t_i^* en fonction des valeurs ajustées $\hat{y}_i$, c'est-à-dire tracer les couples de points $(\hat{y}_i, t_i^*)$. Si une structure apparaît (tendance, cône, vagues), l'hypothèse d'homoscédasticité risque fort de ne pas être vérifiée. Voyons cela sur un graphique.

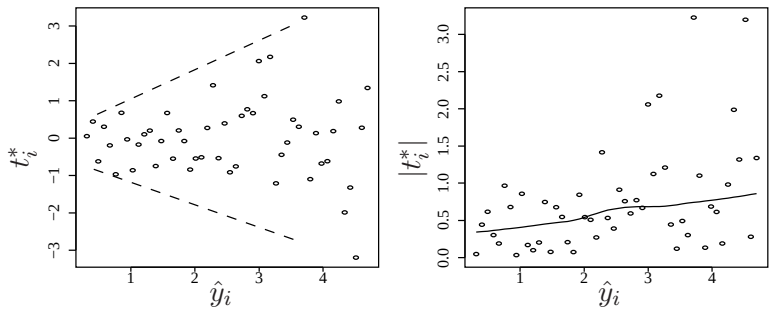


FIGURE 4.2 – Hétéroscédasticité des résidus.

Sur la figure 4.2, l'ajustement n'est pas satisfaisant car la variabilité des résidus augmente avec la valeur de $\hat{y}_i$, on parle de cône de variance croissante. Le second graphique représente la valeur absolue du résidu avec une estimation de la tendance des résidus. Cette estimation de la tendance est obtenue par un lisseur, ici `lowess`. Ce lisseur, qui est aussi nommé `loess`, est le plus utilisé pour obtenir ce type de courbe. Il consiste en une régression par polynômes locaux itérée.

Nous voyons que la tendance est croissante, donc que la variance des résidus augmente le long de l'axe des abscisses. Ce deuxième graphique permet de repérer plus facilement que le premier les changements de variance éventuels dans les résidus. Le choix de l'axe des abscisses est très important et permet (ou non) de détecter une hétéroscédasticité. D'autres choix que $\hat{y}_i$ en abscisse peuvent s'avérer plus pertinents selon le problème : ce peuvent être le temps, l'indice...

4.1.4 Analyse de la structure des résidus

Par l'hypothèse $(\mathcal{H}_2)$, les erreurs ε_i sont supposées être indépendantes, mais ceci est bien sûr impossible à vérifier puisque ces erreurs sont inconnues : nous n'avons accès qu'aux résidus $\hat{\varepsilon}_i$, or ceux-ci ne sont pas indépendants, ils ne sont même pas décorrélés puisque $\mathrm{Var}(\hat{\boldsymbol{\varepsilon}}) = \sigma^2(I - H)$.

D'un point de vue graphique, une représentation des résidus judicieuse pourra néanmoins permettre de suspecter quelques cas de non-indépendance et de compléter l'analyse obtenue par des tests. Si l'on soupçonne une structuration temporelle (autocorrélation des résidus), un graphique temps en abscisse, résidus en ordonnée sera tout indiqué. Si l'on soupçonne une structuration spatiale, un graphique possible consiste en une carte sur laquelle en chacun des points de mesure, on représente un cercle ou un carré (selon le signe du résidu estimé) de taille variable (selon la

valeur absolue du résidu estimé). Ce type de graphique (voir figure 4.3) permettra peut-être de détecter une structuration spatiale (agrégats de ronds ou de carrés, ou au contraire alternance des ronds/carrés). Si une structuration est observée, un travail sur les résidus et en particulier sur leur covariance est alors nécessaire.

Exemple. Le but ici est d'expliquer une variable Y, le nombre de plantes endémiques observées, par trois variables : la surface de l'unité de mesure, l'altitude et la latitude. Les résidus studentisés sont représentés sur la carte géographique des emplacements de mesure (figure 4.3). On observe des agrégats de résidus positifs ou négatifs qui semblent indiquer qu'une structuration spatiale reste présente dans les résidus.

Sur cet exemple, une simple représentation des résidus en fonction de $\hat{y}_i$ ou de l'indice i de l'observation n'apporte que peu d'information. Il importe donc d'insister ici sur le choix adéquat de la représentation graphique des résidus.

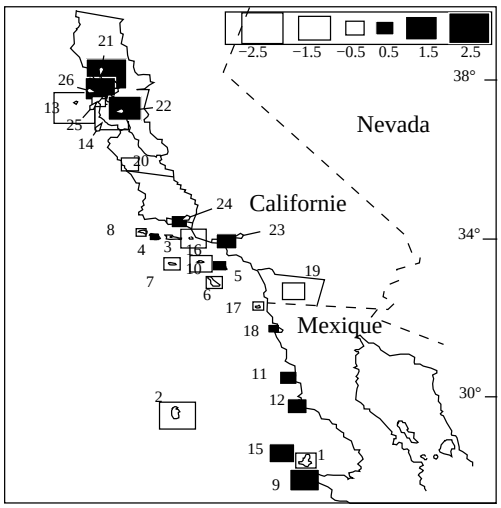


FIGURE 4.3 – Exemple de résidus studentisés structurés spatialement.

L'utilisation d'un lisseur peut permettre de dégager une éventuelle structuration dans les résidus (voir figure 4.4) et ce de manière aisée et rapide, ce qui est primordial. Il est cependant difficile, voire impossible, de discerner entre une structuration due à un oubli dans la modélisation de la moyenne et une structuration due à une mauvaise modélisation de la variance (voir figure 4.4).

Un autre type de structuration des résidus peut être dû à une mauvaise modélisation. Supposons que nous ayons oublié une variable intervenant dans l'explication de la variable Y. Cet oubli se retrouvera forcément dans les résidus, qui sont par définition les observations moins les estimations par le modèle. L'hypothèse d'absence de structuration ($\text{Cov}(\varepsilon_i, \varepsilon_j) = 0 \ \forall i \neq j$) risque de ne pas être vérifiée. En effet, la composante oubliée dans le modèle va s'additionner au vrai bruit et devrait apparaître dans le dessin des résidus.

Une forme quelconque de structuration dans le graphe des résidus sera annonciatrice d'un mauvais ajustement du modèle. Une fois détectée une structuration, il suffit, si l'on peut dire, d'ajouter au modèle une variable explicative possédant la même structuration. Voyons cela sur un exemple graphique.

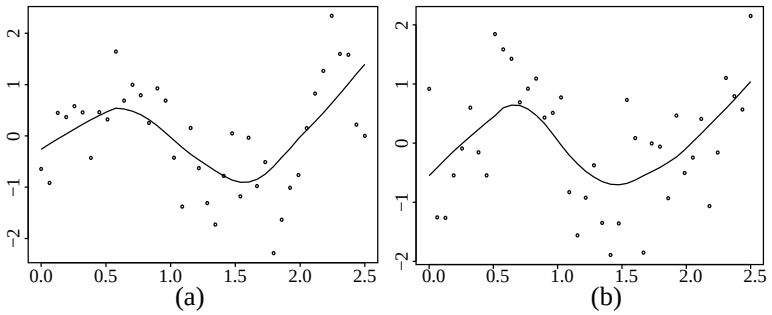


FIGURE 4.4 – Tendance sinusoïdale due à des bruits autorégressifs d'ordre 1, $\varepsilon_i = \rho\varepsilon_{i-1} + \eta_i$ (variance mal modélisée, graphique a) ou à une composante explicative non prise en compte : $x_2 = 0.2\sin(3x)$ (moyenne mal modélisée, graphique b).

La figure (4.5) montre les graphiques d'un modèle linéaire $y = \alpha + \beta_1 x_1 + \varepsilon$ alors que le vrai modèle est à deux variables $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$. L'ajustement n'est pas satisfaisant puisqu'une tendance linéaire décroissante se dégage des résidus de la troisième représentation. Notons l'importance du choix de l'axe des abscisses : les deux premiers graphiques, représentant les mêmes résidus, ne laissent pas soupçonner cette tendance décroissante. Le modèle linéaire proposé n'est donc pas judicieux, il serait bon d'ajouter la variable "oubliée" x_2.

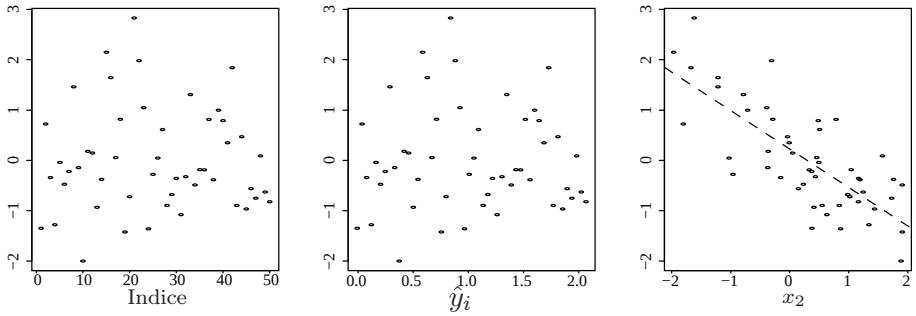


FIGURE 4.5 – Résidus studentisés avec une tendance décroissante due à l'oubli d'une variable x_2 dans le modèle. Les résidus studentisés sont représentés comme fonctions du numéro de l'observation (indice), de l'estimation du modèle $\hat{y}_i$ et comme fonction de x_2.

Malgré tout, ce type de diagnostic peut être insuffisant. Une autre méthode plus précise, mais fastidieuse, consiste à regarder, variable explicative par variable explicative, si la variable considérée agit bien de manière linéaire sur la variable à expliquer. Ce type d'analyse sera mené avec des résidus appelés résidus partiels (ou résidus partiels augmentés) ou encore via des régressions partielles. Ces graphiques permettent de constater si une variable candidate est bien utile au modèle et, le cas échéant, de trouver d'éventuelles fonctions non linéaires de variables explicatives déjà présentes. Rappelons qu'une fonction non linéaire f fixée d'une variable explicative x_j est considérée comme une variable explicative à part entière $x_{p+1} = f(x_j)$.

En conclusion, il est impératif de tracer un graphique avec en ordonnée les résidus et en abscisse : soit $\hat{y}_i$, soit le numéro i de l'observation, soit le temps ou tout autre facteur potentiel de non-indépendance. Idéalement, ce type de graphique permettra : de vérifier l'ajustement global, de repérer les points aberrants, de vérifier les hypothèses concernant la structure de variance du

vecteur $\boldsymbol{\varepsilon}$.

D'autres graphiques, tels ceux présentant la valeur absolue des résidus en ordonnée, permettront de regarder la structuration de la variance. L'analyse des résidus permet de détecter des différences significatives entre les valeurs observées et les valeurs prédites. Cela permet donc de connaître les points mal prédits et les faiblesses du modèle en termes de moyenne ou de variance.

Cependant, ceci ne nous renseigne nullement sur la robustesse des estimateurs par rapport à l'ajout ou à la suppression d'une observation. La section suivante propose quelques critères en ce sens.

4.2 Analyse de la matrice de projection

Nous souhaiterions maintenant avoir une mesure synthétique du poids d'une observation sur sa propre prévision par le modèle. Cette prévision utilise la matrice de projection orthogonale sur l'espace engendré par les colonnes de X, à savoir $P_X = H = X(X'X)^{-1}X'$. En effet, nous avons vu que $\hat{Y} = P_X Y = HY$. Commençons par donner quelques propriétés très générales sur les matrices de projection orthogonale.

Propriétés 8 (Propriétés d'une matrice de projection orthogonale) *Soit $H = P_X$ la matrice $n \times n$ de projection orthogonale sur le sous-espace $\mathcal{M}$ de dimension p engendré par les colonnes de X. Alors :*

1. $\mathrm{Tr}(H) = \sum_{i=1}^{n} h_{ii} = p$.
2. $\sum_i \sum_j h_{ij}^2 = p$.
3. *Pour tout* $i \in \{1, \ldots, n\}$, $0 \leq h_{ii} \leq 1$.
4. *Si* $h_{ii} = 0$ *ou 1, alors* $h_{ij} = 0$ *pour tout* j *différent de* i.
5. *pour tout* j *différent de* i, $-0.5 \leq h_{ij} \leq 0.5$.

Preuve.

1. La trace d'un projecteur vaut la dimension de l'espace sur lequel s'effectue la projection, donc $\mathrm{Tr}(H) = p$.
2. Ce second point découle de la propriété $H^2 = H$, d'où $\mathrm{Tr}(H^2) = p$, de la symétrie de H et du fait que pour toute matrice A, $\mathrm{Tr}(AA') = \mathrm{Tr}(A'A) = \sum_i \sum_j a_{ij}^2$.
3. Puisque les matrices H et H^2 sont égales, nous avons en particulier $h_{ii} = (H^2)_{ii}$. Cela s'écrit, en utilisant la symétrie de H :
$$h_{ii} = \sum_{j=1}^{n} h_{ij} h_{ji} = h_{ii}^2 + \sum_{j \neq i} h_{ij}^2 \;\Rightarrow\; h_{ii}(1 - h_{ii}) = \sum_{j \neq i} h_{ij}^2.$$
La quantité de droite de la dernière égalité est positive, donc le troisième point est démontré.
4. Cette propriété se déduit directement de l'équation précédente.
5. Nous pouvons écrire : $h_{ii}(1 - h_{ii}) = h_{ij}^2 + \sum_{k \neq i,j} h_{ik}^2$. La quantité de gauche est maximum lorsque $h_{ii} = 0.5$ et vaut alors 0.25. Le dernier point est ainsi prouvé.

■

Il suffit maintenant de remarquer que :

$$\hat{y}_i = \sum_{j=1}^{n} h_{ij} y_j = h_{ii} y_i + \sum_{j \neq i} h_{ij} y_j,$$

pour s'apercevoir que h_{ii} représente en quelque sorte le "poids" de l'observation y_i sur sa propre prédiction $\hat{y}_i$. Ainsi :

— si $h_{ii} = 1$, $h_{ij} = 0$ pour tout $j \neq i$, et $\hat{y}_i$ est entièrement déterminé par y_i, puisque $\hat{y}_i = y_i$;
— si $h_{ii} = 0$, $h_{ij} = 0$ pour tout $j \neq i$ donc $\hat{y}_i = 0$, et y_i n'a aucune influence sur $\hat{y}_i$;
— plus généralement, si h_{ii} est "grand", y_i influe fortement sur $\hat{y}_i$, comme en témoigne la formule précédemment établie :

$$y_i - \hat{y}_i = (1 - h_{ii})(y_i - \hat{y}_i^{p}),$$

qui montre la variation dans la prédiction de y_i selon que l'on prend en compte ou non la $i^{ème}$ observation.

Puisque $\mathrm{Tr}(P_X) = \sum h_{ii} = p$, la moyenne des h_{ii} est égale à p/n. Ceci permet de quantifier quelque peu la notion de "grand".

Définition 12 (Point levier) *Un point (x_i, y_i) est appelé point levier si :*
— $h_{ii} > 2p/n$ *selon Hoaglin & Welsch (1978) ;*
— $h_{ii} > 3p/n$ *pour* $p > 6$ *et* $n - p > 12$ *selon Velleman & Welsch (1981) ;*
— $h_{ii} > 0.5$ *selon Huber (1981).*

Remarque. Si la constante fait partie du modèle (i.e. la plupart du temps), on peut affiner la Propriété 8, puisque les termes diagonaux h_{ii} sont en fait tous supérieurs à $1/n$. Il est également possible de prouver que h_{ii} correspond d'une certaine façon à la distance du point x_i au centre de gravité $\bar{x}$ du nuage de points $(x_i)_{1 \leq i \leq n}$ de l'échantillon. Pour plus de détails sur ces points, on pourra consulter le livre de Antoniadis, Berruyer et Carmona, *Régression non linéaire et applications*, Economica (1992), pages 36-40.

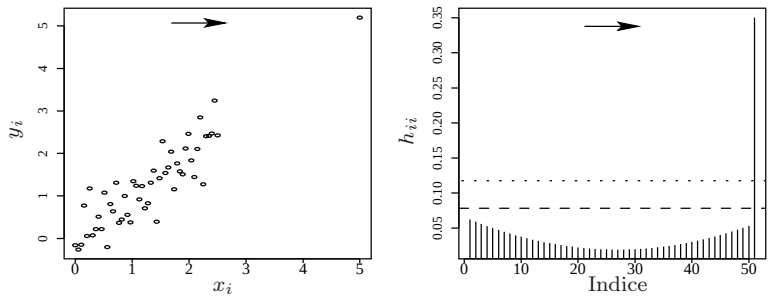


FIGURE 4.6 – Exemple d'un point levier, figuré par la flèche, pour un modèle de régression simple. Quantification par h_{ii} de la notion de levier. La ligne en pointillés longs représente le seuil de $2p/n$, celle en pointillés courts le seuil de $3p/n$.

Pour un modèle de régression simple dont le nuage de points est représenté sur la figure 4.6, le point désigné par une flèche est un point levier. Sa localisation sur l'axe x diffère des autres points et son poids h_{ii} est prépondérant et supérieur aux valeurs seuils de $2p/n$ et $3p/n$.

Remarque. Le point de la figure 4.6 est levier mais pas aberrant puisqu'il se situe dans le prolongement de la droite de régression et sera donc proche de sa prévision par le modèle (résidu faible).

En conclusion, l'analyse des résidus permet de trouver des valeurs atypiques en fonction de la valeur de la variable à expliquer, tandis que l'analyse de la matrice de projection permet de trouver des individus atypiques en fonction des valeurs des variables explicatives (observations éloignées de $\bar{x}$). D'autres critères vont combiner ces deux analyses, c'est ce que nous allons voir maintenant.

4.3 Autres mesures diagnostiques

La distance de Cook mesure l'influence de l'observation i sur l'estimation du paramètre β. Pour bâtir une telle mesure, il suffit de considérer la distance entre le coefficient estimé $\hat{\beta}$ et le coefficient $\hat{\beta}_{(i)}$ que l'on estime en enlevant l'observation i (cf. Section 4.1.1, méthode du *leave-one-out*). Si la distance est grande, alors l'observation i influence beaucoup l'estimation de β, puisque la laisser ou l'enlever conduit à des estimations très différentes l'une de l'autre. De manière générale, $\hat{\beta}$ et $\hat{\beta}_{(i)}$ étant dans $\mathbb{R}^p$, une distance bâtie sur un produit scalaire s'écrit :

$$d(\hat{\beta}_{(i)}, \hat{\beta}) \quad = \quad \sqrt{(\hat{\beta}_{(i)} - \hat{\beta})' Q (\hat{\beta}_{(i)} - \hat{\beta})},$$

où Q est une matrice symétrique définie positive. De nombreux choix sont possibles. Si nous revenons à la région de confiance simultanée de β donnée au Chapitre 3, nous obtenons en prenant $R = I_p$ et $\alpha = 5\%$:

$$RC_\alpha(\beta) \quad = \quad \left\{ \beta \in \mathbb{R}^p \ : \ \frac{1}{p\hat{\sigma}^2} (\hat{\beta} - \beta)' (X'X) (\hat{\beta} - \beta) \leq f^p_{n-p}(0.95) \right\}.$$

Cette équation donne une région de confiance pour β autour de $\hat{\beta}$ et permet de dire que, en moyenne, dans 95% des cas, la distance entre β et $\hat{\beta}$ (selon la matrice $Q = (X'X)/p\hat{\sigma}^2$) est inférieure à $f^p_{n-p}(0.95)$. Par analogie, nous pouvons utiliser cette distance, appelée distance de Cook, pour mesurer l'influence de l'observation i sur le modèle.

Définition 13 (Distance de Cook) *La distance de Cook pour la $i^{ème}$ observation est définie par :*

$$C_i \quad = \quad \frac{1}{p\hat{\sigma}^2} (\hat{\beta}_{(i)} - \hat{\beta})' (X'X) (\hat{\beta}_{(i)} - \hat{\beta}).$$

Il est possible de l'exprimer de manière plus concise comme suit :

$$C_i = \frac{h_{ii}(y_i - \hat{y}_i^p)^2}{p\hat{\sigma}^2} = \frac{h_{ii}}{p(1-h_{ii})^2} \frac{\hat{\varepsilon}_i^2}{\hat{\sigma}^2} = \frac{1}{p} \frac{h_{ii}}{1-h_{ii}} t_i^2.$$

Remarque. Il y a dans cette terminologie un léger abus de langage, puisque la distance de Cook est en fait le carré d'une distance.

Preuve. Nous allons utiliser les résultats établis dans la preuve du théorème 8. Par définition, nous avons :

$$\hat{\beta}_{(i)} = (X'_{(i)} X_{(i)})^{-1} X'_{(i)} Y_{(i)},$$

or en utilisant le lemme d'inversion matricielle pour $(X'_{(i)}X_{(i)})^{-1}$ et le fait que $X'_{(i)}Y_{(i)} = X'Y - x_i y_i$, on obtient :

$$\hat{\beta}_{(i)} = \left[(X'X)^{-1} + \frac{(X'X)^{-1} x_i x_i' (X'X)^{-1}}{1 - h_{ii}} \right] \left(X'Y - x_i y_i \right),$$

ce qui donne en développant :

$$\hat{\beta}_{(i)} = \hat{\beta} - (X'X)^{-1} x_i y_i + \frac{1}{1-h_{ii}} (X'X)^{-1} x_i x_i' \hat{\beta} - \frac{h_{ii}}{1-h_{ii}} (X'X)^{-1} x_i y_i,$$

c'est-à-dire tout simplement :

$$\hat{\beta}_{(i)} - \hat{\beta} = \frac{-\hat{\varepsilon}_i}{1-h_{ii}} (X'X)^{-1} x_i,$$

et puisqu'on a vu dans la preuve du théorème 8 que $\hat{\varepsilon}_i = (1 - h_{ii})(y_i - \hat{y}_i^p)$, on en déduit que :

$$\hat{\beta}_{(i)} - \hat{\beta} = -(y_i - \hat{y}_i^p)(X'X)^{-1}x_i.$$

Il suffit d'appliquer cette expression et le fait que $h_{ii} = x_i'(X'X)^{-1}x_i$ pour obtenir la deuxième expression de la distance de Cook :

$$C_i = \frac{h_{ii}(y_i - \hat{y}_i^p)^2}{p\hat{\sigma}^2}.$$

La troisième expression de la distance de Cook découle alors de la relation déjà mentionnée $\hat{\varepsilon}_i = (1 - h_{ii})(y_i - \hat{y}_i^p)$. Pour la dernière expression, il suffit d'appliquer la définition de t_i.

■

Une observation influente est donc une observation qui, enlevée, conduit à une grande variation dans l'estimation des coefficients, c'est-à-dire à une distance de Cook élevée. Pour juger si la distance C_i est élevée, Cook (1977) propose le seuil $f_{n-p}^p(0.1)$ comme souhaitable et le seuil $f_{n-p}^p(0.5)$ comme préoccupant. Certains auteurs citent comme seuil la valeur 1, qui est une approximation raisonnable de $f_{n-p}^p(0.5)$ lorsque p et $n - p$ sont tous deux grands.

Remarquons sur l'expression

$$C_i = \frac{1}{p}\frac{h_{ii}}{1 - h_{ii}}t_i^2.$$

que la distance de Cook peut être vue comme la contribution de deux termes. Le premier, $h_{ii}/(1 - h_{ii})$, est d'autant plus grand que le point est levier tandis que le second, t_i^2, est d'autant plus grand que le point est aberrant.

Exemple. Pour le modèle de régression simple de la figure 4.6, nous avons tracé sur la figure 4.7 : la droite des moindres carrés, les résidus studentisés par validation croisée, les distances de Cook. Nous voyons que des points ayant de forts résidus (éloignés de la droite) possèdent des distances de Cook élevées (cas des points 4, 6, 12, 29, 44 et 45). Le point 51, bien qu'ayant un résidu faible puisqu'il se situe dans le prolongement de l'axe du nuage, apparaît comme ayant une distance de Cook relativement forte (la $8^{ème}$ plus grande). Ceci illustre bien que la distance de Cook opère un compromis entre points aberrants et points leviers. Notons enfin que, dans notre cas précis, les seuils de la distance de Cook sont $f_{49}^2(0.5) \approx 0.7$ et $f_{49}^2(0.1) \approx 0.11$, ce dernier figurant en pointillé sur la figure 4.7. Sur ce graphique, les distances de Cook semblent assez bien réparties au niveau hauteur et aucun point ne se détache nettement.

Exemple (suite). En utilisant les mêmes 50 points, on remplaçe simplement le point levier 51 par un point franchement aberrant(cf. figure 4.8 au centre, son résidu t_{51}^* étant très élevé). Malgré la position de ce point 51 à l'intérieur du nuage des x_i, la distance de Cook est élevée et ceci uniquement à cause de son caractère aberrant. Bien entendu un point peut être à la fois levier et aberrant. Le seuil de $f_{n-p}^p(0.5)$, ici égal à $f_{49}^2(0.5) \approx 0.7$, semble assez conservateur : en pratique, on pourrait en effet se poser la question de la suppression de ce point 51.

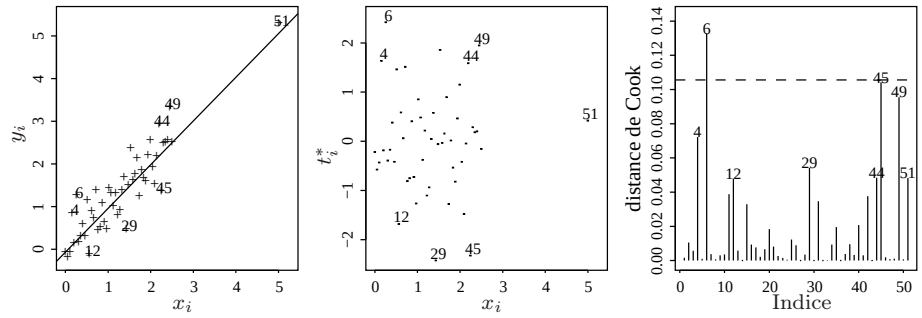


FIGURE 4.7 – Exemple du point levier (numéro 51). Les points associés aux 8 plus grandes valeurs de la distance de Cook sont numérotés ainsi que leurs distances de Cook et leurs résidus studentisés. La droite en trait plein est la droite ajustée par MCO.

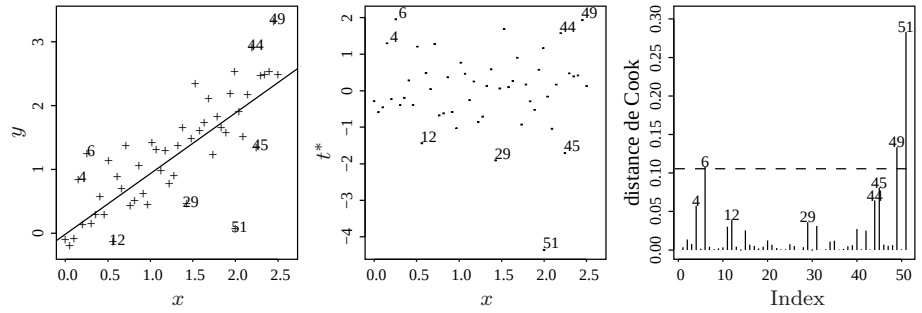


FIGURE 4.8 – Exemple de point fortement aberrant (numéro 51). Les points associés aux 8 plus grandes valeurs de la distance de Cook sont numérotés ainsi que leurs distances de Cook et leurs résidus studentisés (par VC). La droite en trait plein est la droite ajustée par MCO.

Une autre mesure d'influence est donnée par la distance de Welsh-Kuh. La définition de la distance de Cook pour l'observation i fait intervenir la variance estimée de l'erreur $\hat{\sigma}^2$. Il faut donc utiliser un estimateur de σ^2. Si l'on utilise l'estimateur classique $\hat{\sigma}^2$, alors une observation influente risque de "perturber" l'estimation $\hat{\sigma}^2$. Il est donc préférable d'utiliser $\hat{\sigma}^2_{(i)}$, obtenu pas validation croisée. L'écart de Welsh-Kuh, souvent appelé DFFITS (pour *DiFference in FITs, Standardized*) par les logiciels, est donc défini par

$$Wk_i = |t_i^*|\sqrt{\frac{h_{ii}}{1-h_{ii}}},$$

et permet d'évaluer l'écart standardisé entre l'estimation bâtie sur toute les observations et l'estimation bâtie sur toutes les observations sauf la $i^{ème}$. Cet écart de Welsh-Kuh mesure ainsi l'influence simultanée d'une observation sur l'estimation des paramètres β et σ^2. Si l'écart de Welsh-Kuh est supérieure à $2\sqrt{p+1}/\sqrt{n}$ en valeur absolue, alors il est conseillé d'analyser les observations correspondantes.

Annexe A

Annales

Université de Rennes 2
Master de Statistiques
Durée : 2 heures

Vendredi 18 Décembre 2009
Calculatrice autorisée
Aucun document

Contrôle de Régression Linéaire

I. La hauteur des eucalyptus

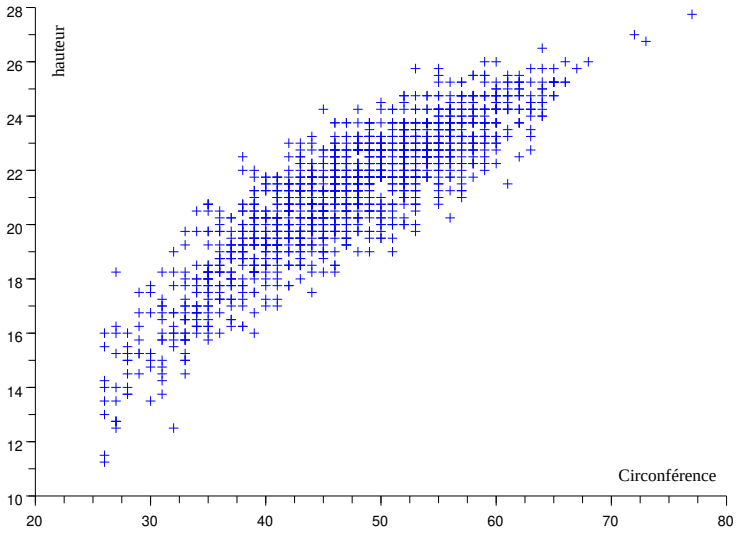


FIGURE A.1 – Nuage de points pour les eucalyptus.

On souhaite expliquer la hauteur y (en mètres) d'un arbre en fonction de sa circonférence x (en centimètres) à 1m30 du sol. On a relevé $n = 1429$ couples (x_i, y_i), le nuage de points étant représenté figure A.1. On a obtenu $(\bar{x}, \bar{y}) = (47,3; 21,2)$ et :

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = 102924 \qquad \sum_{i=1}^{n}(y_i - \bar{y})^2 = 8857 \qquad \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = 26466$$

1. Calculer la droite des moindres carrés pour le modèle $y = \beta_1 + \beta_2 x + \varepsilon$ et la représenter sur la figure A.1.
2. Calculer le coefficient de détermination R^2. Commenter la qualité de l'ajustement des données au modèle.
3. Avec ces estimateurs, la somme des carrés des résidus vaut alors $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 2052$. Si on suppose les perturbations ε_i gaussiennes, centrées, indépendantes et de même variance σ^2, en déduire un estimateur non biaisé $\hat{\sigma}^2$ de σ^2.
4. Donner un estimateur $\hat{\sigma}_1^2$ de la variance de $\hat{\beta}_1$.
5. Tester l'hypothèse $H_0 : \beta_1 = 0$ contre $H_1 : \beta_1 \neq 0$.
6. Etant donné la forme du nuage de points, on veut expliquer la hauteur non seulement par la circonférence, mais aussi par la racine carrée de celle-ci :

   $$y_i = \alpha_1 + \alpha_2 x_i + \alpha_3 \sqrt{x_i} + \varepsilon_i.$$

   Pour α_3, on a obtenu $\hat{\alpha}_3 = 10$ et $\hat{\sigma}_3 = 0,78$. Tester l'hypothèse $H_0 : \alpha_3 = 0$ contre $H_1 : \alpha_3 \neq 0$.

II. Modèle à deux variables explicatives

On considère le modèle de régression suivant :

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

Les $x_{i,j}$, sont des variables exogènes du modèle, les ε_i sont des variables aléatoires indépendantes, de loi normale centrée admettant la même variance σ^2. En posant :

$$X = \begin{bmatrix} 1 & x_{1,2} & x_{1,3} \\ \vdots & \vdots & \vdots \\ 1 & x_{n,2} & x_{n,3} \end{bmatrix} \quad \text{et} \quad Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix},$$

on a observé :

$$X'X = \begin{bmatrix} 30 & 20 & 0 \\ 20 & 20 & 0 \\ 0 & 0 & 10 \end{bmatrix}, \quad X'Y = \begin{bmatrix} 15 \\ 20 \\ 10 \end{bmatrix}, \quad Y'Y = 59.5.$$

1. Déterminer la valeur de n, la moyenne des $x_{i,3}$, le coefficient de corrélation des $x_{i,2}$ et des $x_{i,3}$.
2. Estimer β_1, β_2, β_3, σ^2 par la méthode des moindres carrés ordinaires.
3. Calculer pour β_2 un intervalle de confiance à 95% et tester l'hypothèse $\beta_3 = 0.8$ au niveau 10%.
4. Tester $\beta_2 + \beta_3 = 3$ contre $\beta_2 + \beta_3 \neq 3$, au niveau 5%.
5. Que vaut $\bar{y}$, moyenne empirique des y_i ? En déduire le coefficient de détermination ajusté R_a^2.

6. Construire un intervalle de prévision à 95% de y_{n+1} connaissant : $x_{n+1,2} = 3$ et $x_{n+1,3} = 0,5$.

III. Modèle hétéroscédastique

On considère n observations $y_1, \ldots, y_n$ d'une variable définie sur une certaine population, et n $k-$uplets x_i ($x_i' = [x_{i1}, \ldots, x_{ik}]$) correspondant aux valeurs prises par k autres variables sur les mêmes éléments de cette population. On suppose que pour tout i, y_i est la valeur prise par une variable aléatoire Y_i, et qu'il existe $\beta \in \mathbb{R}^k$ pour lequel :

$$Y_i \sim \mathcal{N}\left(x_i'\beta, \sigma_i^2\right) \qquad 1 \leq i \leq n,$$

où :

- β représente un vecteur de $\mathbb{R}^k$: $\beta = [\beta_1, \ldots, \beta_k]'$,
- Les Y_i sont supposées indépendantes entre elles.

Enfin, les valeurs σ_i^2 des variances dépendent de l'appartenance à p sous-populations des éléments sur lesquels les variables sont observées. En regroupant les indices des Y_i selon ces sous-populations, on posera :

— $I_1 = \{1, \ldots, n_1\}$, indices des n_1 éléments de la première sous-population ;
— $I_2 = \{n_1 + 1, \ldots, n_1 + n_2\}$, indices des n_2 éléments de la deuxième sous-population ;
— ... ;
— $I_\ell = \{n_1 + \ldots + n_{\ell-1} + 1, \ldots, n_1 + \ldots + n_{\ell-1} + n_\ell\}$, indices des n_ℓ éléments de la ℓ-ème sous-population ;
— ... ;
— $I_p = \{n_1 + \ldots + n_{p-1} + 1, \ldots, n\}$, indices des n_p éléments de la dernière sous-population.

On admettra l'hypothèse suivante : si $i \in I_\ell$, $\sigma_i^2 = \ell\sigma^2$. Autrement dit, pour les n_1 variables correspondant aux éléments de la première sous-population la valeur est σ^2, pour les n_2 variables correspondant aux éléments de la deuxième sous-population la valeur est $2\sigma^2$, etc. , jusqu'à $p\sigma^2$ pour la variance des variables correspondant aux éléments de la dernière sous-population. On veut estimer β et σ^2 par la méthode du maximum de vraisemblance. On notera $\hat{\beta}$, $\hat{\sigma}^2$ ces estimateurs.

1. Que vaut $f_{Y_i}(y_i)$, f_{Y_i} représentant la densité de la loi normale $\mathcal{N}\left(x_i'\beta, \sigma_i^2\right)$?
2. Montrer que $\hat{\beta}$ et $\hat{\sigma}^2$ sont solutions du système d'équations :

$$\left\{ \begin{array}{cc} \sum_{\ell=1}^{p} \frac{1}{\ell} \sum_{i \in I_\ell} (y_i - x_i'\beta)^2 = n\sigma^2 \\ \forall j = 1, \ldots, k \qquad \sum_{\ell=1}^{p} \frac{1}{\ell} \sum_{i \in I_\ell} (y_i - x_i'\beta)\, x_{ij} = 0. \end{array} \right. \tag{A.1}$$

3. Montrer que le système (*A.3*) équivaut à :

$$\left\{ \begin{array}{l} \|A(Y - X\beta)\|^2 = n\sigma^2 \\ X'A^2(Y - X\beta) = 0. \end{array} \right. \tag{A.2}$$

où $\|.\|^2$ représente la norme euclidienne usuelle dans $\mathbb{R}^n$, X la matrice $(n \times k)$ du plan d'expérience, Y le vecteur $(n \times 1)$ des observations y_i, A la matrice $(n \times n)$ diagonale dont l'élément (i, i) vaut $\frac{1}{\sqrt{l}}$ si $i \in I_l$.

4. En supposant que $\left(X'A^2X\right)$ est inversible, exprimer $\hat{\beta}$ et $\hat{\sigma}^2$.
5. Montrer que $n\hat{\sigma}^2 = \|V\|^2$, où V suit une loi gaussienne centrée.
6. En déduire que $\mathbb{E}\left[\|V\|^2\right]$ est la trace de la matrice de variance-covariance de V.
7. Montrer que $n\hat{\sigma}^2/(n-k)$ est un estimateur sans biais de σ^2.

8. On note X_ℓ la matrice $(n_\ell \times k)$ formée par les lignes d'indices I_ℓ de X, supposée de rang plein, Y_ℓ le vecteur colonne $(n_\ell \times 1)$ des composantes d'indices I_ℓ de Y. En posant $\hat{\beta}_\ell = (X'_\ell \cdot X_\ell)^{-1} X'_\ell Y_\ell$, montrer que $\hat{\beta}_\ell$ est un estimateur sans biais de β.
9. (Bonus) Que peut-on dire de la différence des matrices de variance-covariance de $\hat{\beta}_\ell$ et de $\hat{\beta}$?

Université de Rennes 2
Master de Statistiques
Durée : 2 heures

Vendredi 18 Décembre 2009
Calculatrice autorisée
Aucun document

Corrigé du Contrôle

I. La hauteur des eucalyptus

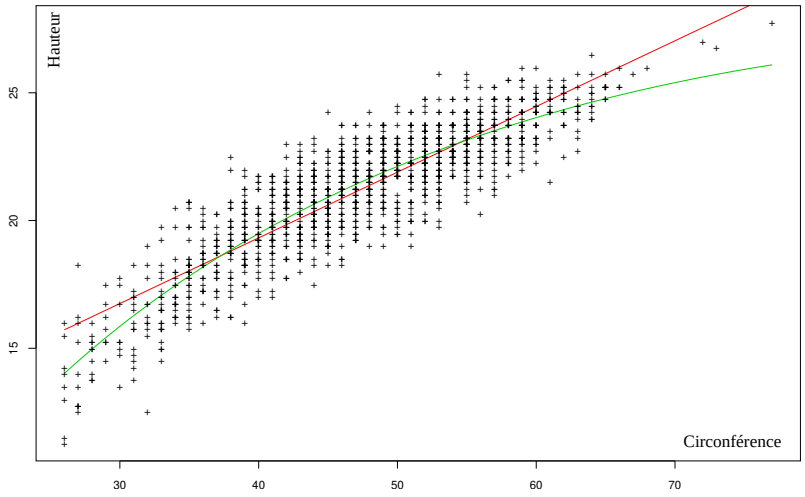


FIGURE A.2 – Nuage de points, droite de régression et courbe de régression.

1. La méthode des moindres carrés ordinaires donne pour estimateur de β_2 :

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \approx 0,257.$$

Et pour estimateur de β_1 :

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2\bar{x} \approx 9,04.$$

La droite des moindres carrés est représentée figure A.2.

2. Le coefficient de détermination R^2 est égal au carré du coefficient de corrélation entre les variables x et y, ce qui donne :

$$R^2 = \frac{\left(\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})\right)^2}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)\left(\sum_{i=1}^{n}(y_i - \bar{y})^2\right)} \approx 0,768.$$

On en conclut que 77% de la variance des hauteurs y_i des eucalyptus est expliquée par la circonférence à 1m30 du sol. Ce modèle de régression linéaire simple semble donc efficace.

3. Un estimateur non biaisé $\hat{\sigma}^2$ de σ^2 est tout simplement :

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2} = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{1427} \approx 1,438.$$

4. Un estimateur $\hat{\sigma}_1^2$ de la variance de $\hat{\beta}_1$ est alors donné par :

$$\hat{\sigma}_1^2 = \frac{\hat{\sigma}^2 \sum_{i=1}^n x_i^2}{n\sum_{i=1}^n (x_i - \bar{x})^2} = \hat{\sigma}^2 \frac{n\bar{x}^2 + \sum_{i=1}^n (x_i - \bar{x})^2}{n\sum_{i=1}^n (x_i - \bar{x})^2} \approx 0,032.$$

5. On sait que l'estimateur centré et normalisé de β_1 suit une loi de Student à $(n-2) = 1427$ degrés de liberté :

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_1} \sim \mathcal{T}_{1427},$$

donc sous l'hypothèse H_0 : $\beta_1 = 0$, ceci se simplifie en $\frac{\hat{\beta}_1}{\hat{\sigma}_1} \sim \mathcal{T}_{1427}$, et cette statistique de test donne ici :

$$t = T(\omega) \approx \frac{9,04}{\sqrt{0,032}} \approx 50,5 \gg 2.$$

Une loi de Student à 1427 degrés de libertés se comportant comme une loi normale centrée réduite, il est clair que la probabilité critique associée au quantile $50,5$ est infinitésimale, donc on rejette l'hypothèse H_0 selon laquelle l'ordonnée à l'origine serait nulle.

6. De même, on sait que sous H_0 :

$$\frac{\hat{\alpha}_3}{\hat{\sigma}_3} \sim \mathcal{T}_{n-3} = \mathcal{T}_{1426},$$

ce qui donne ici :

$$t = T(\omega) = \frac{10}{0,78} \approx 12,8.$$

Ici encore, on rejette H_0 sans hésiter. A titre indicatif, la courbe des moindres carrés est représentée figure A.2.

II. Modèle à deux variables explicatives

On considère le modèle de régression suivant :

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

Les $x_{i,j}$, sont des variables exogènes du modèle, les ε_i sont des variables aléatoires indépendantes, de loi normale centrée admettant la même variance σ^2. En posant :

$$X = \begin{bmatrix} 1 & x_{1,2} & x_{1,3} \\ \vdots & \vdots & \vdots \\ 1 & x_{n,2} & x_{n,3} \end{bmatrix} \quad \text{et} \quad Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix},$$

on a observé :

$$X'X = \begin{bmatrix} 30 & 20 & 0 \\ 20 & 20 & 0 \\ 0 & 0 & 10 \end{bmatrix}, \quad X'Y = \begin{bmatrix} 15 \\ 20 \\ 10 \end{bmatrix}, \quad Y'Y = 59.5.$$

1. La valeur de n se lit en haut à gauche de la matrice $X'X$, c'est-à-dire $n = (X'X)_{1,1} = 30$. De même, la moyenne des $x_{i,3}$ correspond à :

$$\frac{1}{30}\sum_{i=1}^{30} x_{i,3} = \frac{(X'X)_{1,3}}{30} = 0.$$

Puisque les $x_{i,3}$ sont centrés, le coefficient de corrélation entre les deux variables x_2 et x_3 est alors :

$$r_{2,3} = \frac{\sum_{i=1}^{30} x_{i,2}x_{i,3}}{\sqrt{\sum_{i=1}^{30}(x_{i,2}-\bar{x}_{i,2})^2}\sqrt{\sum_{i=1}^{30} x_{i,3}^2}} = \frac{(X'X)_{2,3}}{\sqrt{\sum_{i=1}^{30}(x_{i,2}-\bar{x}_{i,2})^2}\sqrt{\sum_{i=1}^{30} x_{i,3}^2}} = 0.$$

2. La méthode des moindres carrés ordinaires donne pour $\beta = [\beta_1, \beta_2, \beta_3]'$ l'estimateur suivant :

$$\hat{\beta} = (X'X)^{-1}X'Y = \begin{bmatrix} 0.1 & -0.1 & 0 \\ -0.1 & 0.15 & 0 \\ 0 & 0 & 0.1 \end{bmatrix} \begin{bmatrix} 15 \\ 20 \\ 10 \end{bmatrix} = \begin{bmatrix} -0.5 \\ 1.5 \\ 1 \end{bmatrix}.$$

Un estimateur non biaisé $\hat{\sigma}^2$ de σ^2 s'écrit :

$$\hat{\sigma}^2 = \frac{\|Y - X\hat{\beta}\|^2}{n-3} = \frac{\|Y\|^2 - \|X\hat{\beta}\|^2}{27},$$

ce qui s'écrit encore :

$$\hat{\sigma}^2 = \frac{Y'Y - Y'X(X'X)^{-1}X'Y}{27} = 1.$$

3. Puisqu'on sait que :

$$\frac{\hat{\beta}_2 - \beta_2}{\hat{\sigma}_2} = \frac{\hat{\beta}_2 - \beta_2}{\hat{\sigma}\sqrt{(X'X)^{-1}_{2,2}}} \sim \mathcal{T}_{n-3} = \mathcal{T}_{27},$$

on en déduit qu'un intervalle de confiance à 95% pour β_2 est :

$$I(\beta_2) = \left[\hat{\beta}_2 - t_{27}(0.975)\hat{\sigma}\sqrt{(X'X)^{-1}_{2,2}}; \hat{\beta}_2 + t_{27}(0.975)\hat{\sigma}\sqrt{(X'X)^{-1}_{2,2}}\right],$$

c'est-à-dire :

$$I(\beta_2) \approx \left[1.5 - 2.05\sqrt{0.15}; 1.5 + 2.05\sqrt{0.15}\right] \approx [0.71; 2.29].$$

Pour tester l'hypothèse $H_0 : \beta_3 = 0.8$ contre $H_1 : \beta_3 \neq 0.8$ au niveau 10%, on calcule de même un intervalle de confiance à 90% de β_3 :

$$I(\beta_3) = \left[\hat{\beta}_3 - t_{27}(0.95)\hat{\sigma}\sqrt{(X'X)^{-1}_{3,3}}; \hat{\beta}_3 + t_{27}(0.95)\hat{\sigma}\sqrt{(X'X)^{-1}_{3,3}}\right],$$

ce qui donne :

$$I(\beta_3) \approx \left[1 - 1.70\sqrt{0.1}; 1 + 1.70\sqrt{0.1}\right] \approx [0.46; 1.54],$$

donc on accepte au niveau 10% l'hypothèse selon laquelle $\beta_3 = 0.8$.

4. On sait que
$$\frac{(\hat{\beta}_2+\hat{\beta}_3)-(\beta_2+\beta_3)}{\hat{\sigma}_{\hat{\beta}_2+\hat{\beta}_3}} \sim \mathcal{T}_{27},$$
avec :
$$\hat{\sigma}_{\hat{\beta}_2+\hat{\beta}_3} = \sqrt{\hat{\sigma}_2^2 + 2\text{Cov}(\hat{\beta}_2,\hat{\beta}_3) + \hat{\sigma}_3^2} = \hat{\sigma}\sqrt{(X'X)^{-1}_{2,2} + 2(X'X)^{-1}_{2,3} + (X'X)^{-1}_{3,3}},$$
c'est-à-dire $\hat{\sigma}_{\hat{\beta}_2+\hat{\beta}_3} = 0.5$. Donc un intervalle de confiance à 95% pour $\beta_2+\beta_3$ est :
$$I(\beta_2+\beta_3) = [2.5 - 0.5t_{27}(0.975); 2.5 + 0.5t_{27}(0.975)] \approx [1.47; 3.53].$$
Par conséquent, au niveau 5%, on accepte $H_0 : \beta_2+\beta_3 = 3$ contre $H_1 : \beta_2+\beta_3 \neq 3$.
5. La moyenne empirique des y_i se déduit de la première composante du vecteur $X'Y$, donc $\bar{y} = 15/30 = 0.5$. Par définition, le coefficient de détermination ajusté R_a^2 vaut :
$$R_a^2 = 1 - \frac{n-1}{n-p}\frac{\|Y-\hat{Y}\|^2}{\|Y-\bar{y}\mathbb{1}\|^2} = 1 - (n-1)\frac{\hat{\sigma}^2}{\|Y-\bar{y}\mathbb{1}\|^2},$$
donc :
$$R_a^2 = 1 - \frac{29}{Y'Y - 30\bar{y}^2} \approx 0.44.$$
6. En notant $x'_{n+1} = [1, 3, 0.5]$, la valeur prédite pour y_{n+1} est :
$$\hat{y}_{n+1} = x'_{n+1}\hat{\beta} = \frac{9}{2},$$
et un intervalle de prévision à 95% pour y_{n+1} est :
$$IC(y_{n+1}) = \left[\hat{y}_{n+1} \pm t_{27}(0.975)\hat{\sigma}\sqrt{1 + x'_{n+1}(X'X)^{-1}x_{n+1}}\right],$$
ce qui donne numériquement $IC(y_{n+1}) \approx [1.69; 7.31]$.

III. Modèle hétéroscédastique

1. Par définition de la loi normale $\mathcal{N}\left(x_i'\beta, \sigma_i^2\right)$, on a tout simplement :
$$f_{Y_i}(y_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}}\exp\left(-\frac{(y_i - x_i'\beta)^2}{2\sigma_i^2}\right).$$
2. Les variables Y_i étant indépendantes, la densité jointe $f_Y(y)$ du n-uplet $Y = (Y_1, \ldots, Y_n)$ est le produit des densités $f_{Y_i}(y_i)$, ce qui donne pour la vraisemblance :
$$\mathcal{L}(y,\beta,\sigma^2) = f_Y(y) = \frac{1}{(2\pi)^{n/2}\sigma_1^{n_1}\ldots\sigma_p^{n_p}}\exp\left(-\sum_{\ell=1}^{p}\sum_{i\in I_\ell}\frac{(y_i - x_i'\beta)^2}{2\sigma_\ell^2}\right),$$
qui s'écrit encore :
$$\mathcal{L}(y,\beta,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}1^{n_1}\ldots p^{n_p}}\exp\left(-\frac{1}{2\sigma^2}\sum_{\ell=1}^{p}\frac{1}{\ell}\sum_{i\in I_\ell}(y_i - x_i'\beta)^2\right),$$

d'où pour la log-vraisemblance :

$$\log \mathcal{L}(y, \beta, \sigma^2) = c - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{\ell=1}^{p} \frac{1}{\ell} \sum_{i \in I_\ell} (y_i - x_i' \beta)^2,$$

où c est une constante. Les estimateurs du maximum de vraisemblance sont obtenus en annulant les dérivées partielles de cette log-vraisemblance par rapport à $\beta_1, \ldots, \beta_k$ et σ^2. Pour tout $j \in \{1, \ldots, k\}$, le calcul donne :

$$\frac{\partial \log \mathcal{L}}{\partial \beta_j}(y, \beta, \sigma^2) = \frac{1}{\sigma^2} \sum_{\ell=1}^{p} \frac{1}{\ell} \sum_{i \in I_\ell} (y_i - x_i' \beta) x_{ij}.$$

La dérivée partielle par rapport à σ^2 s'écrit elle :

$$\frac{\partial \log \mathcal{L}}{\partial \sigma^2}(y, \beta, \sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{\ell=1}^{p} \frac{1}{\ell} \sum_{i \in I_\ell} (y_i - x_i' \beta)^2.$$

On en déduit bien que $\hat{\beta}$ et $\hat{\sigma}^2$ sont les solutions du système d'équations :

$$\left\{ \begin{array}{cc} \sum_{\ell=1}^{p} \frac{1}{\ell} \sum_{i \in I_\ell} (y_i - x_i' \beta)^2 = n\sigma^2 \\ \forall j = 1, \ldots, k \qquad \sum_{\ell=1}^{p} \frac{1}{\ell} \sum_{i \in I_\ell} (y_i - x_i' \beta)\, x_{ij} = 0. \end{array} \right. \tag{A.3}$$

3. En notant A la matrice $(n \times n)$ diagonale dont l'élément (i, i) vaut $\frac{1}{\sqrt{l}}$ si $i \in I_l$, et en remarquant que A est symétrique, il vient :

$$\|A\,(Y - X\beta)\|^2 = (Y - X\beta)' A' A (Y - X\beta) = (Y - X\beta)' A^2 (Y - X\beta),$$

c'est-à-dire :

$$\|A\,(Y - X\beta)\|^2 = [y_1 - x_1'\beta, \ldots, y_n - x_n'\beta] A^2 [y_1 - x_1'\beta, \ldots, y_n - x_n'\beta]' = \sum_{\ell=1}^{p} \frac{1}{\ell} \sum_{i \in I_\ell} \left(y_i - x_i'\beta\right)^2.$$

On en déduit :

$$\sum_{\ell=1}^{p} \frac{1}{\ell} \sum_{i \in I_\ell} \left(y_i - x_i'\beta\right)^2 = n\sigma^2 \iff \|A\,(Y - X\beta)\|^2 = n\sigma^2.$$

De la même façon, on peut remarquer que :

$$X' A^2 (Y - X\beta) = \left[\sum_{\ell=1}^{p} \frac{1}{\ell} \sum_{i \in I_\ell} \left(y_i - x_i'\beta\right) x_{i1}, \ldots, \sum_{\ell=1}^{p} \frac{1}{\ell} \sum_{i \in I_\ell} \left(y_i - x_i'\beta\right) x_{ik} \right]'.$$

Au final, le système (A.3) équivaut bien à :

$$\left\{ \begin{array}{c} \|A\,(Y - X\beta)\|^2 = n\sigma^2 \\ X' A^2 (Y - X\beta) = 0. \end{array} \right. \tag{A.4}$$

4. La seconde équation du système (A.4) s'écrit encore :

$$\left(X' A^2 X\right) \beta = X' A^2 Y.$$

Puisque $\left(X' A^2 X\right)$ est supposée inversible, l'estimateur $\hat{\beta}$ vaut :

$$\hat{\beta} = \left(X' A^2 X\right)^{-1} X' A^2 Y.$$

L'estimateur $\hat{\sigma}^2$ s'en déduit immédiatement via la première équation du système (A.4) :

$$\hat{\sigma}^2 = \frac{1}{n} \left\| A \left(Y - X\hat{\beta}\right) \right\|^2.$$

5. D'après la question précédente, on a :

$$n\hat{\sigma}^2 = \left\| A\left(Y - X\hat{\beta}\right)\right\|^2 = \|V\|^2,$$

en notant $V = A\left(Y - X\hat{\beta}\right) = AY - AX\hat{\beta}$. Il suffit alors d'écrire :

$$(AX)\hat{\beta} = AX\left(X'A^2X\right)^{-1}X'A^2Y = (AX)((AX)'(AX))^{-1}(AX)'(AY),$$

pour comprendre que le vecteur $(AX)\hat{\beta}$ n'est rien d'autre que la projection orthogonale du vecteur AY sur le sous-espace $\mathcal{M}$ de $\mathbb{R}^n$ engendré par les colonnes de la matrice AX. Notons au passage que ce sous-espace est de dimension k puisque, par hypothèse, la matrice $\left(X'A^2X\right)$ est inversible. Le vecteur AY étant de loi $\mathcal{N}(AX\beta, \sigma^2 I_n)$, nous sommes exactement dans le cadre d'application du théorème de Cochran. En notant respectivement P et $P^\perp$ les matrices de projection sur $\mathcal{M}$ et $\mathcal{M}^\perp$, celui-ci assure que :

$$V = P^\perp AY \sim \mathcal{N}(P^\perp AX\beta, \sigma^2 P^\perp) = \mathcal{N}(0, \sigma^2 P^\perp).$$

Ainsi V suit bien une loi gaussienne centrée.

6. Puisque $\|V\|^2$ est un scalaire, il est égal à sa trace, ce qui donne :

$$\mathbb{E}\left[\|V\|^2\right] = \mathbb{E}\left[\mathrm{Tr}\left(\|V\|^2\right)\right] = \mathbb{E}\left[\mathrm{Tr}\left(V'V\right)\right],$$

et puisque pour toute matrice A, $\mathrm{Tr}(A'A) = \mathrm{Tr}(AA')$, il en découle :

$$\mathbb{E}\left[\|V\|^2\right] = \mathbb{E}\left[\mathrm{Tr}\left(VV'\right)\right].$$

Il reste à noter d'une part que les opérateurs de trace et d'espérance commutent, et d'autre part que V est centré pour obtenir :

$$\mathbb{E}\left[\|V\|^2\right] = \mathrm{Tr}\left(\mathbb{E}\left[VV'\right]\right) = \mathrm{Tr}\left(\mathrm{Var}(V)\right).$$

7. On déduit des deux questions précédentes que :

$$\mathbb{E}\left[\hat{\sigma}^2\right] = \frac{1}{n}\mathbb{E}\left[\|V\|^2\right] = \frac{1}{n}\mathrm{Tr}\left(\mathrm{Var}(V)\right),$$

or $V \sim \mathcal{N}(0, \sigma^2 P^\perp)$, où $P^\perp$ est la matrice de projection orthogonale sur un sous-espace de dimension $(n-k)$, donc $\mathrm{Tr}(P^\perp) = n - k$, et :

$$\mathbb{E}\left[\hat{\sigma}^2\right] = \frac{n-k}{n}\sigma^2,$$

ce qui revient à dire que $n\hat{\sigma}^2/(n-k)$ est un estimateur sans biais de σ^2.

8. Avec les notations du texte, on a $Y_\ell = X_\ell\beta + \boldsymbol{\varepsilon}$, où $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \ell\sigma^2 I_\ell)$. Il vient donc :

$$\mathbb{E}\left[\hat{\beta}_\ell\right] = \left(X_\ell'X_\ell\right)^{-1}X_\ell'\mathbb{E}\left[Y_\ell\right] = \left(X_\ell'X_\ell\right)^{-1}X_\ell'X_\ell\beta = \beta.$$

Ainsi, pour tout $\ell \in \{1, \ldots, p\}$, $\hat{\beta}_\ell$ est un estimateur sans biais de β.

9. Puisque $AX\hat{\beta}$ est la projection orthogonale du vecteur $AY \sim \mathcal{N}(AX\beta, \sigma^2 I_n)$ sur le sous-espace $\mathcal{M}$, nous savons que :

$$\mathrm{Var}(\hat{\beta}) = \sigma^2((AX)'(AX))^{-1} = \sigma^2 \left(X'A^2X\right)^{-1}.$$

De la même façon, puisque $X\hat{\beta}_\ell$ est la projection orthogonale du vecteur $Y_\ell \sim \mathcal{N}(X_\ell\beta, \ell\sigma^2 I_n)$ sur le sous-espace $\mathcal{M}_\ell$ engendré par les colonnes de X_ℓ, la matrice de covariance de l'estimateur $\hat{\beta}_\ell$ vaut :

$$\mathrm{Var}(\hat{\beta}_\ell) = \ell\sigma^2 \left(X_\ell' X_\ell\right)^{-1} = \sigma^2 \left(\frac{X_\ell}{\sqrt{\ell}}' \frac{X_\ell}{\sqrt{\ell}}\right)^{-1}.$$

La matrice $X_\ell/\sqrt{\ell}$ correspondant aux n_ℓ lignes d'indices I_ℓ de la matrice AX, notons Z_ℓ la matrice $(n - n_\ell) \times k$ des autres lignes de AX. On a donc :

$$(AX)'(AX) = \frac{X_\ell}{\sqrt{\ell}}' \frac{X_\ell}{\sqrt{\ell}} + Z_\ell' \cdot Z_\ell.$$

En particulier, pour tout vecteur u de $\mathbb{R}^k$, on a :

$$u' Z_\ell' Z_\ell u = \|Z_\ell u\|^2 \geq 0,$$

donc :

$$u' \frac{X_\ell}{\sqrt{\ell}}' \frac{X_\ell}{\sqrt{\ell}} u \leq u'(AX)'(AX)u,$$

ce qui s'écrit en terme de relation d'ordre pour les matrices symétriques :

$$\frac{X_\ell}{\sqrt{\ell}}' \frac{X_\ell}{\sqrt{\ell}} \leq (AX)'(AX),$$

les matrices des deux membres étant toutes deux symétriques définies positives.

Il reste maintenant à remarquer que, de façon générale, si B et C sont deux matrices symétriques définies positives, avec $B \leq C$, alors $C^{-1} \leq B^{-1}$. En effet, dire que $B \leq C$ revient à dire que les valeurs propres de $(C - B)$ sont toutes supérieures ou égales à 0, donc il en va de même pour la matrice $B^{-1/2}(C - B)B^{-1/2} = B^{-1/2}CB^{-1/2} - I$. Ceci signifie que les valeurs propres de la matrice $B^{-1/2}CB^{-1/2}$ sont toutes supérieures ou égales à 1, ce qui implique que celles de sa matrice inverse sont toutes inférieures ou égales à 1, ce qui s'écrit encore $B^{1/2}C^{-1}B^{1/2} \leq I$. Or cette dernière relation a pour conséquence $C^{-1} \leq B^{-1}$.

Appliqué dans notre contexte, ce résultat donne :

$$((AX)'(AX))^{-1} \leq \left(\frac{X_\ell}{\sqrt{\ell}}' \frac{X_\ell}{\sqrt{\ell}}\right)^{-1},$$

d'où l'on déduit l'inégalité entre matrices de covariance : $\mathrm{Var}(\hat{\beta}) \leq \mathrm{Var}(\hat{\beta}_\ell)$. En d'autres termes, $\hat{\beta}$ est un estimateur plus précis que $\hat{\beta}_\ell$, ce qui n'a rien d'étonnant vu que sa construction utilise $(n - n_\ell)$ observations de plus que celle de $\hat{\beta}_\ell$. Happy end !

Université de Rennes 2
Master de Statistiques
Durée : 2 heures

Mercredi 14 Décembre 2010
Calculatrice autorisée
Aucun document

Contrôle de Régression Linéaire

I. La hauteur des eucalyptus

On souhaite expliquer la hauteur y (en mètres) d'un arbre en fonction de sa circonférence x (en centimètres) à 1m30 du sol et de la racine carrée de celle-ci. On a relevé $n = 1429$ couples (x_i, y_i), le nuage de points étant représenté figure A.3. On considère donc le modèle de régression suivant :

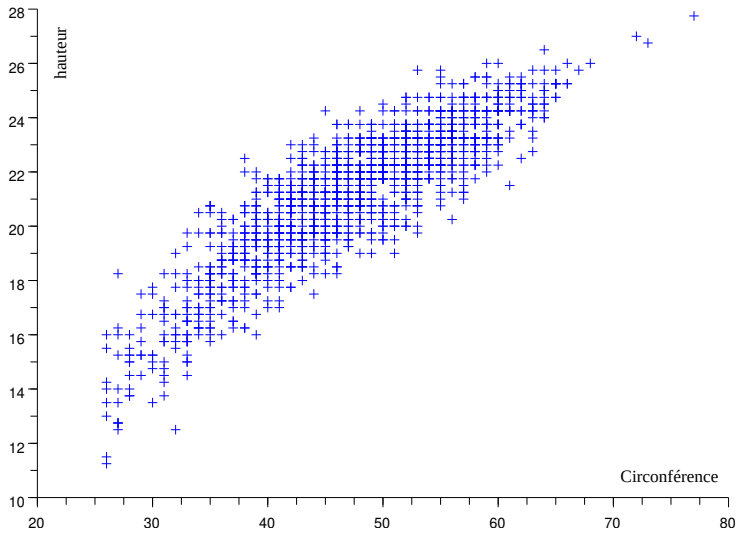


FIGURE A.3 – Nuage de points pour les eucalyptus.

$$y_i = \beta_1 + \beta_2 x_i + \beta_3 \sqrt{x_i} + \varepsilon_i, \qquad 1 \leq i \leq n.$$

Les ε_i sont des variables aléatoires indépendantes, de loi normale centrée admettant la même variance σ^2. En posant :

$$X = \begin{bmatrix} 1 & x_1 & \sqrt{x_1} \\ \vdots & \vdots & \vdots \\ 1 & x_n & \sqrt{x_n} \end{bmatrix} \quad \text{et} \quad Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix},$$

on a observé :

$$X'X = \begin{bmatrix} ? & ? & 9792 \\ ? & 3306000 & ? \\ ? & 471200 & 67660 \end{bmatrix}, \quad X'Y = \begin{bmatrix} 30310 \\ 1462000 \\ 209700 \end{bmatrix}, \quad Y'Y = 651900.$$

1. Déterminer les '?' dans la matrice $X'X$.
2. Que vaut la circonférence moyenne empirique $\bar{x}$?
3. Le calcul donne (en arrondissant !)

$$\left(X'X\right)^{-1} = \begin{bmatrix} 4.646 & 0.101 & -1.379 \\ 0.101 & 0.002 & -0.030 \\ -1.379 & -0.030 & 0.411 \end{bmatrix} \quad \text{et} \quad (X'X)^{-1}X'Y = \begin{bmatrix} -16.8 \\ -0.30 \\ 7.62 \end{bmatrix}.$$

   Que valent les estimateurs $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_3$ par la méthode des moindres carrés ? Grâce au calcul de quelques points, représenter la courbe obtenue sur la figure A.3.
4. Calculer l'estimateur de σ^2 pour les moindres carrés.
5. Calculer pour β_3 un intervalle de confiance à 95%.
6. Tester l'hypothèse $\beta_2 = 0$ au niveau de risque 10%.
7. Que vaut la hauteur moyenne empirique $\bar{y}$? En déduire le coefficient de détermination ajusté R^2_a.
8. Construire un intervalle de prévision à 95% de y_{n+1} connaissant $x_{n+1} = 49$.
9. Construire un intervalle de prévision à 95% de y_{n+1} connaissant $x_{n+1} = 25$.
10. Des deux intervalles précédents, lequel est le plus grand ? Pouvait-on s'y attendre ?

II. Consommation de gaz

Mr Derek Whiteside de la *UK Building Research Station* a collecté la consommation hebdomadaire de gaz et la température moyenne externe de sa maison au sud-est de l'Angleterre pendant une saison. Une régression pour expliquer la consommation de gaz en fonction de la température est réalisée avec le logiciel R. Les résultats numériques sont les suivants.

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.97802 -0.11082  0.02672  0.25294  0.63803

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.72385    0.12974     ?     < 2e-16 ***
Temp        -0.27793          ?   -11.04 1.05e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3548 on 28 degrees of freedom
Multiple R-Squared: 0.8131,     Adjusted R-squared: 0.8064
F-statistic: 121.8 on 1 and 28 DF,  p-value: 1.046e-11
```

1. Donner le modèle et les hypothèses de la régression.
2. Compléter le tableau.
3. Soit Z une variable aléatoire de loi de Student de degré de liberté 28. Quelle est la probabilité que $|Z|$ soit supérieure à 11.04 ?
4. Préciser les éléments du test correspondant à la ligne "`Temp`" du tableau (H_0, H_1, la statistique de test, sa loi sous H_0, la règle de décision).
5. Interpréter le nombre "`Multiple R-Squared: 0.8131`" du tableau.

6. Donner une estimation de la variance du terme d'erreur dans le modèle de régression simple.
7. Expliquer et interpréter la dernière ligne du tableau :
   "`F-statistic: 121.8 on 1 and 28 DF, p-value: 1.046e-11`".
   Voyez-vous une autre façon d'obtenir cette p-value ?
8. Pensez-vous que la température extérieure a un effet sur la consommation de gaz ? Justifiez votre réponse.

III. Régression simple

On dispose de n points $(x_i, y_i)_{1\leq i\leq n}$ et on sait qu'il existe une relation de la forme : $y_i = ax_i + b + \varepsilon_i$, où les erreurs ε_i sont des variables centrées, décorrélées et de même variance σ^2.

1. Rappeler les formules des estimateurs des moindres carrés $\hat{a}$ et $\hat{b}$, ainsi que leurs variances respectives.
2. Dans cette question, on suppose connaître b, mais pas a.
   (a) En revenant à la définition des moindres carrés, calculer l'estimateur $\tilde{a}$ des moindres carrés de a.
   (b) Calculer la variance de $\tilde{a}$. Montrer qu'elle est inférieure à celle de $\hat{a}$.
3. Dans cette question, on suppose connaître a, mais pas b.
   (a) En revenant à la définition des moindres carrés, calculer l'estimateur $\tilde{b}$ des moindres carrés de b.
   (b) Calculer la variance de $\tilde{b}$. Montrer qu'elle est inférieure à celle de $\hat{b}$.

IV. Forces de frottement et vitesse

Au $17^{\text{ème}}$ siècle, Huygens s'est intéressé aux forces de résistance d'un objet en mouvement dans un fluide (eau, air, etc.). Il a d'abord émis l'hypothèse selon laquelle les forces de frottement étaient proportionnelles à la vitesse de l'objet, puis, après expérimentation, selon laquelle elles étaient proportionnelles au carré de la vitesse. On réalise une expérience dans laquelle on fait varier la vitesse x d'un objet et on mesure les forces de frottement y. Ensuite, on teste la relation existant entre ces forces de frottement et la vitesse.

1. Quel(s) modèle(s) testeriez-vous ?
2. Comment feriez-vous pour déterminer le modèle adapté ?

Université de Rennes 2
Master de Statistiques
Durée : 2 heures

Mercredi 14 Décembre 2010
Calculatrice autorisée
Aucun document

Corrigé du Contrôle

I. La hauteur des eucalyptus

On souhaite expliquer la hauteur y (en mètres) d'un arbre en fonction de sa circonférence x (en centimètres) à 1m30 du sol et de la racine carrée $\sqrt{x}$ de cette circonférence. On a relevé 1429 couples (x_i, y_i). On considère donc le modèle de régression suivant :

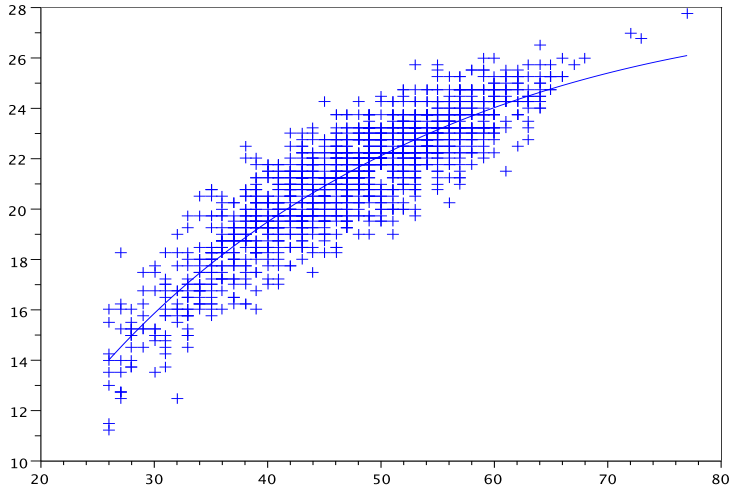


FIGURE A.4 – Nuage de points et courbe de régression pour les eucalyptus.

$$y_i = \beta_1 + \beta_2 x_i + \beta_3 \sqrt{x_i} + \varepsilon_i, \qquad 1 \leq i \leq n.$$

Les ε_i sont des variables aléatoires indépendantes, de loi normale centrée admettant la même variance σ^2. En posant :

$$X = \begin{bmatrix} 1 & x_1 & \sqrt{x_1} \\ \vdots & \vdots & \vdots \\ 1 & x_n & \sqrt{x_n} \end{bmatrix} \quad \text{et} \quad Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix},$$

on a observé :

$$X'X = \begin{bmatrix} ? & ? & 9792 \\ ? & 3306000 & ? \\ ? & 471200 & 67660 \end{bmatrix}, \quad X'Y = \begin{bmatrix} 30310 \\ 1462000 \\ 209700 \end{bmatrix}, \quad Y'Y = 651900.$$

1. La matrice $X'X$ se complète comme suit :
$$X'X = \begin{bmatrix} 1429 & 67660 & 9792 \\ 67660 & 3306000 & 471200 \\ 9792 & 471200 & 67660 \end{bmatrix}$$
2. La circonférence moyenne empirique vaut donc :
$$\bar{x} = \frac{67660}{1429} \approx 47.3 \text{ cm}.$$
3. La méthode des moindres carrés ordinaires donne pour $\beta = [\beta_1, \beta_2, \beta_3]'$ l'estimateur suivant :
$$\hat{\beta} = (X'X)^{-1}X'Y = \begin{bmatrix} -16.8 \\ -0.30 \\ 7.62 \end{bmatrix}.$$
La courbe obtenue est représentée figure A.4.
4. Un estimateur non biaisé $\hat{\sigma}^2$ de σ^2 s'écrit :
$$\hat{\sigma}^2 = \frac{\|Y - X\hat{\beta}\|^2}{n-3} = \frac{\|Y\|^2 - \|X\hat{\beta}\|^2}{1426}.$$
Puisque $\|X\hat{\beta}\|^2 = \hat{\beta}'X'X\hat{\beta} = \hat{\beta}'X'Y$, ceci s'écrit encore :
$$\hat{\sigma}^2 = \frac{Y'Y - \hat{\beta}'X'Y}{1426} \approx 1,26.$$
5. Puisqu'on sait que :
$$\frac{\hat{\beta}_3 - \beta_3}{\hat{\sigma}_3} = \frac{\hat{\beta}_3 - \beta_3}{\hat{\sigma}\sqrt{(X'X)^{-1}_{3,3}}} \sim \mathcal{T}_{n-3} = \mathcal{T}_{1426},$$
on en déduit qu'un intervalle de confiance à 95% pour β_3 est :
$$I(\beta_3) = \left[\hat{\beta}_3 - t_{1426}(0.975)\hat{\sigma}\sqrt{(X'X)^{-1}_{33}}; \hat{\beta}_3 + t_{1426}(0.975)\hat{\sigma}\sqrt{(X'X)^{-1}_{3,3}}\right],$$
c'est-à-dire en considérant que $t_{1426}(0.975) = 1.96$ comme pour une loi normale centrée réduite :
$$I(\beta_3) \approx \left[7.62 - 1.96\sqrt{0.72}; 7.62 + 1.96\sqrt{0.72}\right] \approx [6.21; 9.03].$$
6. On veut tester l'hypothèse H_0 : $\beta_2 = 0$ contre H_1 : $\beta_2 \neq 0$ au niveau de risque 10%. Sous H_0, on sait que
$$\frac{\hat{\beta}_2}{\hat{\sigma}_2} = \frac{\hat{\beta}_2}{\hat{\sigma}\sqrt{(X'X)^{-1}_{22}}} \sim \mathcal{T}_{n-3} = \mathcal{T}_{1426} \approx \mathcal{N}(0,1).$$
Il nous suffit donc de comparer la valeur absolue de la statistique de test obtenue ici au quantile d'ordre 0.95 d'une loi normale centrée réduite, c'est-à-dire à 1.645. Or
$$|T(\omega)| = \frac{|-0.30|}{\sqrt{1.26} \times \sqrt{0.002}} \approx 5.98 > 1.645.$$
Par conséquent on rejette l'hypothèse selon laquelle $\beta_2 = 0$.

7. La moyenne empirique des y_i se déduit de la première composante du vecteur $X'Y$:

$$\bar{y} = 30310/1429 \approx 21.2 \text{ m}.$$

Par définition, le coefficient de détermination ajusté R_a^2 vaut :

$$R_a^2 = 1 - \frac{n-1}{n-p}\frac{\|Y-\hat{Y}\|^2}{\|Y-\bar{y}\mathbb{1}\|^2} = 1 - (n-1)\frac{\hat{\sigma}^2}{\|Y-\bar{y}\mathbb{1}\|^2},$$

donc :

$$R_a^2 = 1 - 1428\frac{1.26}{Y'Y - 1429\bar{y}^2} \approx 0.81.$$

8. En notant $x'_{n+1} = [1, 49, 7]$, la valeur prédite pour y_{n+1} est :

$$\hat{y}_{n+1} = x'_{n+1}\hat{\beta} \approx 21.8,$$

et un intervalle de prévision à 95% pour y_{n+1} est :

$$IC(y_{n+1}) = \left[\hat{y}_{n+1} - t_{1426}(0.975)\hat{\sigma}\sqrt{1 + x'_{n+1}(X'X)^{-1}x_{n+1}}; \hat{y}_{n+1} + \dots\right],$$

ce qui donne numériquement $IC(y_{n+1}) \approx [20.1; 23.5]$.

9. De même, en posant $x'_{n+1} = [1, 25, 5]$, la valeur prédite pour y_{n+1} est :

$$\hat{y}_{n+1} = x'_{n+1}\hat{\beta} \approx 13.8,$$

et un intervalle de prévision à 95% pour y_{n+1} est $IC(y_{n+1}) \approx [11.7; 15.9]$.

10. On constate que c'est le second intervalle de prévision qui est le plus grand : ceci est dû au fait que le second point est plus éloigné du centre de gravité du nuage. On prévoit donc moins bien sa valeur.

II. Consommation de gaz

Mr Derek Whiteside, de la *UK Building Research Station*, a collecté la consommation hebdomadaire de gaz et la température moyenne externe de sa maison au sud-est de l'Angleterre pendant une saisons. Une régression pour expliquer la consommation de gaz en fonction de la température est réalisée avec le logiciel R. Les résultats numériques sont les suivants.

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.97802 -0.11082  0.02672  0.25294  0.63803

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.72385    0.12974    36.41     < 2e-16 ***
Temp        -0.27793    0.0252     -11.04 1.05e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3548 on 28 degrees of freedom
Multiple R-Squared: 0.8131,     Adjusted R-squared: 0.8064
F-statistic: 121.8 on 1 and 28 DF,  p-value: 1.046e-11
```

1. Le modèle considéré ici est : pour tout $i \in \{1, \dots, 30\}$
$$C_i = \beta_1 + \beta_2 T_i + \varepsilon_i,$$
avec les erreurs ε_i gaussiennes centrées, indépendantes et de même variance σ^2.
2. cf. ci-dessus.
3. Soit Z une variable aléatoire de loi de Student de degré de liberté 28. D'après le tableau, la probabilité que $|Z|$ soit supérieure à 11.04 est de l'ordre de 1.05×10^{-11}.
4. Pour la ligne "`Temp`" du tableau, l'hypothèse H_0 correspond à $\beta_2 = 0$ contre $H_1 : \beta_2 \neq 0$. Sous H_0, $\hat{\beta}_2/\hat{\sigma}_{\hat{\beta}_2}$ suit une loi de Student à 28 degrés de liberté. On décide de rejeter H_0 si la statistique de test $|T(\omega)| = |\hat{\beta}_2/\hat{\sigma}_{\hat{\beta}_2}|$ correspond à une p-value très faible (typiquement inférieure à 5%). En l'occurence, la règle de décision ci-dessus est calculée à partir des valeurs obtenues $\hat{\beta}_2 = -0.27793$, $\hat{\sigma}_{\hat{\beta}_2} = 0.0252$, $|T(\omega)| = |\hat{\beta}_2/\hat{\sigma}_{\hat{\beta}_2}| = 11.04$ et la p-value correspondante pour une loi de Student à 28 degrés de liberté est : $\mathbb{P}(|T| > 11.04) = 1.05 \times 10^{-11}$.
5. Le nombre `Multiple R-Squared: 0.8131` correspond au coefficient de détermination R^2 du modèle. Il signifie qu'environ 81% de la variation des données de consommation est expliquée par ce modèle de régression linéaire simple.
6. Un estimateur de la variance σ^2 du terme d'erreur est donné par le carré du terme `Residual standard error` du tableau, à savoir $\hat{\sigma}^2 = 0.3548^2 \approx 0.126$.
7. La dernière ligne du tableau correspond au test de Fisher de validité globale du modèle. Avec les notations du cours, on sait que sous l'hypothèse $H_0 : \beta_2 = 0$, nous avons
$$F = \frac{n-p}{q} \times \frac{SCR_0 - SCR}{SCR} = \frac{\|\hat{Y} - \hat{Y}_0\|^2}{\hat{\sigma}^2} \sim \mathcal{F}^1_{28},$$
loi de Fisher à 1 et 28 degrés de liberté. La statistique de test donne ici $F(\omega) = 121.8$, ce qui correspond à une p-value de 1.046×10^{-11}. Nous rejetons donc l'hypothèse selon laquelle β_2 serait nul. Remarquons que ce test correspond au test de Student effectué dans la ligne "`Temp`" du tableau.
8. Au vu des résultats du test de Student (ou de l'équivalent Test de Fisher de la dernière ligne), il est clair que la température a un impact sur la consommation de gaz. Ceci est tout à fait naturel, puisque plus il fait froid, plus on chauffe.

III. Régression simple

On dispose de n points $(x_i, y_i)_{1 \leq i \leq n}$ et on sait qu'il existe une relation de la forme : $y_i = ax_i + b + \varepsilon_i$, où les erreurs ε_i sont des variables centrées, décorrélées et de même variance σ^2.

1. Les formules des estimateurs des moindres carrés $\hat{a}$ et $\hat{b}$ sont
$$\hat{a} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\sum (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x})^2}.$$
et
$$\hat{b} = \bar{y} - \hat{a}\bar{x}.$$
Leurs variances respectives sont données par
$$\mathrm{Var}(\hat{a}) = \frac{\sigma^2}{\sum (x_i - \bar{x})^2} \qquad \& \qquad \mathrm{Var}(\hat{b}) = \frac{\sigma^2 \sum x_i^2}{n \sum (x_i - \bar{x})^2}.$$
2. Dans cette question, on suppose connaître b, mais pas a.

(a) L'estimateur $\tilde{a}$ des moindres carrés correspond à l'argmin de la quantité :

$$S(a) = \sum (y_i - (ax_i + b))^2,$$

ce qui s'obtient en annulant la dérivée de S :

$$S'(\tilde{a}) = -2\sum x_i(y_i - (\tilde{a}x_i + b)) = 0 \;\Leftrightarrow\; \tilde{a} = \frac{\sum x_i(y_i - b)}{\sum x_i^2}.$$

(b) Pour calculer la variance de $\tilde{a}$, on commence par l'exprimer différemment. Grâce à la relation $y_i = ax_i + b + \varepsilon_i$, on déduit :

$$\tilde{a} = a + \frac{\sum x_i\varepsilon_i}{\sum x_i^2}.$$

Puisque les erreurs ε_i sont décorrélées et de même variance σ^2, il vient :

$$\mathrm{Var}(\tilde{a}) = \frac{\sigma^2}{\sum x_i^2}.$$

Puisque $\sum (x_i - \bar{x})^2 = \sum x_i^2 - n\bar{x}^2 \leq \sum x_i^2$, il est alors clair que $\mathrm{Var}(\tilde{a}) \leq \mathrm{Var}(\hat{a})$.

3. Dans cette question, on suppose connaître a, mais pas b.

(a) L'estimateur $\tilde{b}$ des moindres carrés correspond cette fois à l'argmin de la quantité :

$$S(b) = \sum (y_i - (ax_i + b))^2,$$

ce qui s'obtient en annulant la dérivée de S :

$$S'(\tilde{b}) = -2\sum (y_i - (\tilde{a}x_i + b)) = 0 \;\Leftrightarrow\; \tilde{b} = \bar{y} - a\bar{x}.$$

(b) Pour calculer la variance de $\tilde{b}$, on commence à nouveau par l'exprimer différemment via la relation $y_i = ax_i + b + \varepsilon_i$:

$$\tilde{b} = b + \frac{1}{n}\sum \varepsilon_i.$$

Puisque les erreurs ε_i sont décorrélées et de même variance σ^2, il vient :

$$\mathrm{Var}(\tilde{b}) = \frac{\sigma^2}{n}.$$

Puisque $\sum (x_i - \bar{x})^2 = \sum x_i^2 - n\bar{x}^2 \leq \sum x_i^2$, il est alors clair que $\mathrm{Var}(\tilde{b}) \leq \mathrm{Var}(\hat{b})$.

IV. Forces de frottement et vitesse

1. Le premier modèle, supposant que les forces de frottement sont proportionnelles à la vitesse de l'objet, s'écrit : pour tout $i \in \{1, \ldots, n\}$

$$f_i = \alpha v_i + \varepsilon_i,$$

où n est le nombre d'observations et les ε_i représentent les erreurs du modèle, typiquement supposées centrées, décorrélées et de même variance σ^2.

Le second modèle, supposant que les forces de frottement sont proportionnelles au carré de la vitesse de l'objet, s'écrit : pour tout $i \in \{1, \ldots, n\}$

$$f_i = \beta v_i^2 + \eta_i,$$

où n est le nombre d'observations et les η_i représentent les erreurs du modèle, typiquement supposées centrées, décorrélées et de même variance s^2.

2. Pour déterminer le modèle adapté, une méthode élémentaire consiste à comparer les pourcentages de variation des données $(f_i)_{1\leq i\leq n}$ expliqués par chacun des modèles. Ceci se fait en calculant les coefficients de détermination respectifs R_1^2 et R_2^2 pour chaque modèle. On optera pour celui qui a le R^2 le plus grand.

Université Rennes 2
Master de Statistiques
Durée : 2 heures

Mardi 6 Décembre 2011
Calculatrice autorisée
Aucun document

Contrôle de Régression Linéaire

I. Prix d'un appartement en fonction de sa superficie
En juin 2005, on a relevé dans les petites annonces les superficies (en m^2) et les prix (en euros) de 108 appartements de type T3 à louer sur l'agglomération de Rennes (cf. figure A.5).

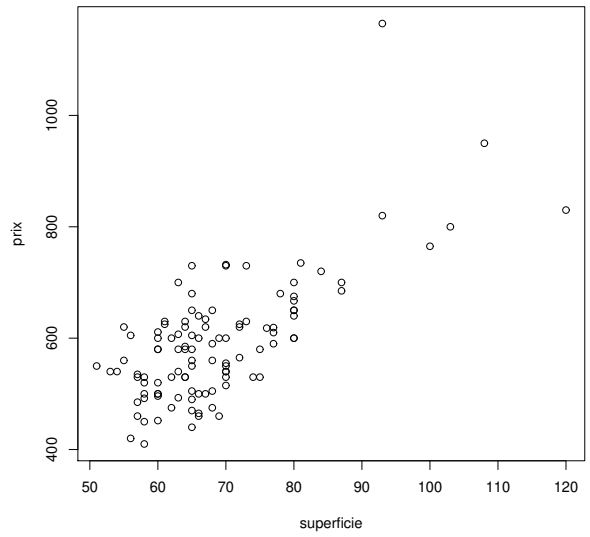


FIGURE A.5 – Prix de location des appartements en fonction de leur superficie.

1. D'après le listing du tableau A.1, donner une estimation du coefficient de corrélation entre le prix et la superficie d'un appartement T3.
2. Proposer un modèle permettant d'étudier la relation entre le prix des appartements et leur superficie. Préciser les hypothèses de ce modèle.
3. D'après le tableau A.1, est-ce que la superficie joue un rôle sur le prix des appartements de type 3 ? Considérez-vous ce rôle comme important ?
4. Quelle est l'estimation du coefficient β (coefficient de la superficie dans le modèle) ? Comment interprétez-vous ce coefficient ?

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 134.3450    45.4737   2.954  0.00386
Superficie    6.6570     0.6525  10.203  < 2e-16

Residual standard error: 77.93 on 106 degrees of freedom
Multiple R-Squared: 0.4955,     Adjusted R-squared: 0.4907
F-statistic: 104.1 on 1 and 106 DF,  p-value: < 2.2e-16
```

TABLE A.1 – Prix en fonction de la superficie : résultats de la régression linéaire simple (sortie R).

5. La superficie moyenne des 108 appartements est de 68.74 m^2 et le prix moyen des appartements est de 591.95 euros. Quel est le prix moyen d'un mètre carré ? Pourquoi ce prix moyen est différent de l'estimation de β ?
6. Dans l'échantillon dont on dispose, comment savoir quels sont les appartements "bon marché" du seul point de vue de la surface ?

II. Tests

Nous nous intéressons au modèle $Y = X\beta + \boldsymbol{\varepsilon}$ sous les hypothèses classiques. Nous avons obtenu sur 21 données :

$$\begin{aligned} \hat{y} &= 6.683_{(2.67)} + 0.44_{(2.32)}x_1 + 0.425_{(2.47)}x_2 + 0.171_{(2.09)}x_3 + 0.009_{(2.24)}x_4, \\ R^2 &= 0.54 \end{aligned}$$

où, pour chaque coefficient, le nombre entre parenthèses représente la valeur absolue de la statistique de test.

1. Quelles sont les hypothèses utilisées ?
2. Tester la nullité de β_1 au seuil de 5%.
3. Pouvez-vous tester $\mathrm{H}_0 : \beta_3 = 1$ contre $\mathrm{H}_1 : \beta_3 \neq 1$?
4. Tester la nullité simultanée des paramètres associés aux variables $x_1, \ldots, x_4$ au seuil de 5%.

III. Moindres carrés ordinaires

1. Nous considérons le modèle de régression linéaire

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

où $Y \in \mathbb{R}^n$, X est une matrice de taille $n \times p$ de rang p, $\beta \in \mathbb{R}^p$ et $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_n)$.

   (a) Qu'appelle-t-on estimateur des moindres carrés $\hat{\beta}$ de β ? Rappeler sa formule.

   (b) Quelle est l'interprétation géométrique de $\hat{Y} = X\hat{\beta}$ (faites un dessin) ?

   (c) Rappeler espérances et matrices de covariance de $\hat{\beta}$, $\hat{Y}$ et $\hat{\boldsymbol{\varepsilon}}$.

2. Nous considérons dorénavant un modèle avec 4 variables explicatives (la première variable étant la constante). Nous avons observé :

$$X'X = \begin{bmatrix} 100 & 20 & 0 & 0 \\ 20 & 20 & 0 & 0 \\ 0 & 0 & 10 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad X'Y = \begin{bmatrix} -60 \\ 20 \\ 10 \\ 1 \end{bmatrix}, \quad Y'Y = 159.$$

   (a) Estimer β et σ^2.

(b) Donner un estimateur de la variance de $\hat{\beta}$.

(c) Donner un intervalle de confiance pour β_2, au niveau 95%.

(d) Calculer un intervalle de prévision de y_{n+1} au niveau 95% connaissant : $x_{n+1,2} = 3$, $x_{n+1,3} = 0.5$ et $x_{n+1,4} = 2$.

IV. Moindres carrés pondérés

On suppose le modèle suivant

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

où X est la matrice $(n \times p)$ du plan d'expérience, $\beta = [\beta_1, \ldots, \beta_p]'$ un vecteur de $\mathbb{R}^p$, Y le vecteur $(n \times 1)$ des observations y_i, $\boldsymbol{\varepsilon}$ le vecteur $(n \times 1)$ des erreurs ε_i supposées centrées et de matrice de covariance $\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2\Omega^2$, où Ω est une matrice $(n \times n)$ diagonale dont l'élément (i, i) vaut $_i > 0$. Dans ce modèle, les valeurs $_i$ sont supposées connues, mais les paramètres β et σ^2 sont inconnus.

1. On considère le modèle transformé $Y^* = X^*\beta + \boldsymbol{\varepsilon}^*$, où :
   — $Y^* = [y_1^*, \ldots, y_n^*]'$, avec $y_i^* = y_i/\omega_i$;
   — X^* est la matrice $(n \times p)$ de terme générique $x_{ij}^* = x_{ij}/\omega_i$;
   — $\boldsymbol{\varepsilon}^* = [\varepsilon_1^*, \ldots, \varepsilon_n^*]'$, avec $\varepsilon_i^* = \varepsilon_i/\omega_i$;

   (a) Donner les relations entre X^* (respectivement Y^*, $\boldsymbol{\varepsilon}^*$), X (respectivement Y, $\boldsymbol{\varepsilon}$) et Ω.

   (b) Déterminer la moyenne et la matrice de covariance du vecteur aléatoire $\boldsymbol{\varepsilon}^*$.

   (c) En supposant $\left(X'\Omega^{-2}X\right)$ inversible, déterminer l'estimateur des moindres carrés $\hat{\beta}^*$ de β. Préciser son biais et sa matrice de covariance.

   (d) Proposer un estimateur sans biais $\hat{\sigma}_*^2$ de σ^2.

2. En revenant au modèle initial $Y = X\beta+\boldsymbol{\varepsilon}$, on suppose maintenant les erreurs ε_i gaussiennes, plus précisément $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2\Omega^2)$.

   (a) Donner la vraisemblance $\mathcal{L}(Y, \beta, \sigma^2)$ du modèle.

   (b) En déduire que les estimateurs au maximum de vraisemblance $\hat{\beta}_{mv}$ et $\hat{\sigma}_{mv}^2$ sont solutions de :
   $$\begin{cases} \left\|\Omega^{-1}(Y - X\beta)\right\|^2 = n\sigma^2 \\ X'\Omega^{-2}(Y - X\beta) = 0. \end{cases}$$

   (c) En déduire les relations entre $\hat{\beta}_{mv}$ et $\hat{\beta}^*$ d'une part, entre $\hat{\sigma}_{mv}^2$ et $\hat{\sigma}_*^2$ d'autre part.

   (d) Préciser alors la loi de $\hat{\beta}^*$. Que dire de celle de $\hat{\sigma}_*^2$?

3. Supposons maintenant le modèle classique de régression linéaire $Y = X\beta+\boldsymbol{\varepsilon}$, avec les erreurs centrées et de matrice de covariance $\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2 I_n$. Néanmoins, on n'observe pas comme d'habitude les x_i' et y_i, mais des moyennes par classe. Spécifiquement, les n données sont réparties en L classes $C_1, \ldots, C_L$ d'effectifs respectifs connus $n_1, \ldots, n_L$ et on a seulement accès aux moyennes par classe, à savoir pour tout $\ell \in \{1, \ldots, L\}$:
   $$\bar{y}_\ell = \frac{1}{n_\ell}\sum_{i \in C_\ell} y_i \qquad \& \qquad \bar{x}_{\ell j} = \frac{1}{n_\ell}\sum_{i \in C_\ell} x_{ij}$$

   (a) En notant $\bar{\varepsilon}_\ell = \frac{1}{n_\ell}\sum_{i \in C_\ell} \varepsilon_i$, vérifier que le modèle peut se mettre sous la forme $\bar{Y} = \bar{X}\beta + \bar{\boldsymbol{\varepsilon}}$.

   (b) Donner la moyenne et la matrice de covariance de $\bar{\boldsymbol{\varepsilon}}$.

   (c) Déduire des questions précédentes des estimateurs de β et σ^2.

Université Rennes 2
Master de Statistiques
Durée : 2 heures

Mardi 6 Décembre 2011
Calculatrice autorisée
Aucun document

Corrigé du Contrôle

I. Prix d'un appartement en fonction de sa superficie

1. Le coefficient de corrélation entre le prix et la superficie d'un appartement T3 correspond à la racine carré du coefficient de détermination multiple (`Multiple R-squared` dans le listing)
$$r = \sqrt{0.496} = 0.704.$$
2. Le modèle s'écrit :
$$\forall i \quad y_i = \alpha + \beta x_i + \varepsilon_i$$
avec y_i le prix de l'appartement i en euros, x_i sa superficie en m^2 et ε_i l'erreur. Les hypothèses usuelles sont de supposer les erreurs gaussiennes, centrées, indépendantes et de même variance : $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_{108})$.
3. Pour tester si la superficie joue un rôle sur le prix des appartements, on teste l'hypothèse $H_0 : \beta = 0$ contre $H_1 : \beta \neq 0$. La statistique de ce test est $T = \hat{\beta}/\hat{\sigma}_{\hat{\beta}}$ et, sous l'hypothèse H_0, cette statistique de test suit une loi de Student à $n - 2 = 106$ degrés de liberté. La probabilité critique associée à ce test est inférieure à 2×10^{-16}. Cette probabilité critique étant inférieure à 5 %, on rejette l'hypothèse H_0 : on considère que la superficie d'un appartement de type T3 influe sur son prix.
La surface a donc une influence significative sur le prix. Mais cette influence est-elle importante ? Le coefficient de détermination R^2, qui s'interprète comme le pourcentage de variabilité expliquée par le modèle, vaut 0.495 : l'influence est donc importante mais d'autres facteurs, difficiles à quantifier, agissent (emplacement, qualité des prestations, avidité du propriétaire, etc.).
4. L'estimation de la pente de la droite de régression est : $\hat{\beta} = 6.657$. Ce coefficient est significativement différent de 0 (voir question précédente) et s'interprète de la façon suivante : un appartement coûtera en moyenne 6.657 euros supplémentaires pour une augmentation de la superficie de 1 m^2.
5. Le prix moyen d'un mètre carré se calcule comme le rapport entre 591.95 et 68.74 soit 8.61 euros le mètre carré. Ce prix est différent de l'estimation de β car le prix des appartements n'est pas strictement proportionnel à leur surface. Comme $\hat{\beta}$ est inférieur au prix moyen d'un mètre carré, proportionnellement à la surface, les petits appartements sont plus chers que les grands. Le modèle de régression stipule qu'il faut d'abord une mise de fond α pour louer un T3, et qu'ensuite le prix d'1 m^2 est β euros (en moyenne). Remarquons que ce coefficient α est significatif, il n'est donc pas souhaitable de le retirer du modèle.
6. Pour déterminer les appartements "bon marché", on peut se fonder sur l'estimation des résidus du modèle : plus le résidu est faible (négatif et avec une forte valeur absolue), plus l'appartement a un prix faible par rapport à celui attendu pour sa superficie.

II. Tests

1. Les hypothèses utilisées sont : $Y = X\beta + \varepsilon$, avec $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_{21})$.
2. Nous savons que
$$T = \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\hat{\beta}_1}} \sim \mathcal{T}_{16}$$
loi de Student à 16 degrés de liberté. Sous l'hypothèse $\beta_1 = 0$, nous avons donc
$$T = \frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}} \sim \mathcal{T}_{16}$$
Or, d'après l'énoncé, la valeur absolue de la statistique de test vaut ici
$$|T(\omega)| = 2.32 > 2.12 = t_{16}(0.975)$$
Donc, au seuil de 5%, on rejette l'hypothèse selon laquelle β_1 serait nul.
3. Par le même raisonnement
$$T = \frac{\hat{\beta}_3 - \beta_3}{\hat{\sigma}_{\hat{\beta}_3}} \sim \mathcal{T}_{16}$$
Or, sous l'hypothèse $\beta_3 = 0$, la valeur absolue de la statistique de test vaut d'après l'énoncé
$$|T(\omega)| = \left| \frac{\hat{\beta}_3}{\hat{\sigma}_{\hat{\beta}_3}} \right| = 2.09$$
Puisque $\hat{\beta}_3 = 0.171$, on en déduit donc que $\hat{\sigma}_{\hat{\beta}_3} \approx 0.082$. Ainsi, sous l'hypothèse $\beta_3 = 1$, nous avons
$$T = \frac{\hat{\beta}_3 - 1}{\hat{\sigma}_{\hat{\beta}_3}} \sim \mathcal{T}_{16}$$
Or la statistique de test donne ici
$$|T(\omega)| = \left| \frac{0.171 - 1}{0.082} \right| \approx 10.1 \gg 2.12 = t_{16}(0.975)$$
donc on rejette l'hypothèse H_0 selon laquelle β_3 serait égal à 1.
4. Nous effectuons un test de Fisher global : $\mathrm{H}_0 : \beta_1 = \cdots = \beta_4 = 0$, contre $H_1 : \exists j \in \{1, \ldots, 4\}, \beta_j \neq 0$. Avec les notations du cours, nous savons que sous l'hypothèse H_0, nous avons
$$F = \frac{21 - 5}{4} \times \frac{SCR_0 - SCR}{SCR} \sim \mathcal{F}_{16}^4$$
loi de Fisher à $(4, 16)$ degrés de liberté. Cette statistique de test s'exprime aussi en fonction du coefficient de détermination comme suit :
$$F = \frac{21 - 5}{4} \times \frac{R^2}{1 - R^2} \sim \mathcal{F}_{16}^4$$
La statistique de test donne donc ici
$$F(\omega) \approx 4.7 > 3.01 = f_{16}^4(0.95)$$
ce qui nous amène à rejeter l'hypothèse H_0 au seuil de 5%.

III. Moindres carrés ordinaires

1. Nous considérons le modèle de régression linéaire

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

où $Y \in \mathbb{R}^n$, X est une matrice de taille $n \times p$ de rang p, $\beta \in \mathbb{R}^p$ et $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_n)$.

(a) L'estimateur des moindres carrés $\hat{\beta}$ de β est défini par

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \|Y - X\beta\|^2$$

où $\|.\|$ est la norme euclidienne usuelle sur $\mathbb{R}^p$. Un calcul classique permet de montrer que $\hat{\beta} = (X'X)^{-1}X'Y$.

(b) Dans ce cadre, $\hat{Y} = X\hat{\beta}$ est tout simplement la projection orthogonale de Y sur le sous-espace de $\mathbb{R}^n$ engendré par les p colonnes de X.

(c) Pour ce qui concerne $\hat{\beta}$, il est facile de montrer que $\mathbb{E}[\hat{\beta}] = \beta$ et $\mathrm{Var}(\hat{\beta}) = \sigma^2(X'X)^{-1}$. De la même façon, nous avons $\mathbb{E}[\hat{Y}] = X\beta$ et $\mathrm{Var}(\hat{Y}) = \sigma^2 P_X$, où $P_X = X(X'X)^{-1}X'$ est la matrice de projection orthogonale sur le sous-espace de $\mathbb{R}^n$ engendré par les p colonnes de X. Enfin, $\mathbb{E}[\hat{\boldsymbol{\varepsilon}}] = 0$ et $\mathrm{Var}(\hat{\boldsymbol{\varepsilon}}) = \sigma^2 P_{X^\perp}$, où $P_{X^\perp} = I_n - P_X$ est la matrice de projection orthogonale sur l'orthogonal du sous-espace de $\mathbb{R}^n$ engendré par les p colonnes de X.

2. Nous considérons dorénavant un modèle avec 4 variables explicatives (la première variable étant la constante).

(a) Nous avons

$$(X'X)^{-1} = \frac{1}{80}\begin{bmatrix} 1 & -1 & 0 & 0 \\ -1 & 5 & 0 & 0 \\ 0 & 0 & 8 & 0 \\ 0 & 0 & 0 & 80 \end{bmatrix}$$

Ce qui donne :

$$\hat{\beta} = (X'X)^{-1}X'Y = \begin{bmatrix} -1 \\ 2 \\ 1 \\ 1 \end{bmatrix}$$

Par Pythagore nous obtenons

$$\|\hat{\boldsymbol{\varepsilon}}\|^2 = \|Y\|^2 - \|\hat{Y}\|^2$$

or

$$\|\hat{Y}\|^2 = \|X\hat{\beta}\|^2 = \cdots = Y'X\hat{\beta} = 111$$

d'où $\|\hat{\boldsymbol{\varepsilon}}\|^2 = 48$ et

$$\hat{\sigma}^2 = \frac{\|\hat{\boldsymbol{\varepsilon}}\|^2}{n-p} = \frac{48}{96} = \frac{1}{2}$$

(b) Un estimateur de la variance de $\hat{\beta}$ est

$$\widehat{\mathrm{Var}}(\hat{\beta}) = \hat{\sigma}^2(X'X)^{-1} = \frac{1}{160}\begin{bmatrix} 1 & -1 & 0 & 0 \\ -1 & 5 & 0 & 0 \\ 0 & 0 & 8 & 0 \\ 0 & 0 & 0 & 80 \end{bmatrix}$$

(c) Nous savons que

$$\frac{\beta_2 - \hat{\beta}_2}{\hat{\sigma}_{\hat{\beta}_2}} \sim \mathcal{T}_{96},$$

loi de Student à 96 degrés de libertés, laquelle peut être assimilée à la loi normale centrée réduite. Puisque

$$\hat{\sigma}_{\hat{\beta}_2} = \sqrt{\widehat{\mathrm{Var}}(\hat{\beta})_{2,2}} = \frac{1}{4\sqrt{2}}$$

un intervalle de confiance à 95% pour β_2 est

$$I = [\hat{\beta}_2 - 1.96\hat{\sigma}_{\hat{\beta}_2}; \hat{\beta}_2 + 1.96\hat{\sigma}_{\hat{\beta}_2}] \approx [1.65; 2.35]$$

(d) Un intervalle de prévision de niveau 95% pour y_{n+1} est donné par

$$I = \left[\hat{y}_{n+1} - t_{96}(0.975)\hat{\sigma}\sqrt{1 + x'_{n+1}(X'X)^{-1}x_{n+1}}; \hat{y}_{n+1} + t_{96}(0.975)\hat{\sigma}\sqrt{1 + x'_{n+1}(X'X)^{-1}x_{n+1}}\right]$$

avec $t_{n-p}(0.975) \approx 1.96$, $x'_{n+1} = [1, 3, 0.5, 2]$, $\sqrt{1 + x'_{n+1}(X'X)^{-1}x_{n+1}} \approx 2.35$ et

$$\hat{y}_{n+1} = x'_{n+1}\hat{\beta} = 7.5$$

Finalement $I \approx [4.24; 10.76]$.

IV. Moindres carrés pondérés

1. On considère le modèle transformé $Y^* = X^*\beta + \boldsymbol{\varepsilon}^*$, où :
   — $Y^* = [y_1^*, \ldots, y_n^*]'$, avec $y_i^* = y_i/\omega_i$;
   — X^* est la matrice $(n \times p)$ de terme générique $x_{ij}^* = x_{ij}/\omega_i$;
   — $\boldsymbol{\varepsilon}^* = [\varepsilon_1^*, \ldots, \varepsilon_n^*]'$, avec $\varepsilon_i^* = \varepsilon_i/\omega_i$;

   (a) Il est clair que $X^* = \Omega^{-1}X$, $Y^* = \Omega^{-1}Y$ et $\boldsymbol{\varepsilon}^* = \Omega^{-1}\boldsymbol{\varepsilon}$.

   (b) Puisque $\mathbb{E}[\boldsymbol{\varepsilon}] = 0$ et $\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2\Omega^2$, on a $\mathbb{E}[\boldsymbol{\varepsilon}^*] = \mathbb{E}[\Omega^{-1}\boldsymbol{\varepsilon}] = \Omega^{-1}\mathbb{E}[\boldsymbol{\varepsilon}] = 0$ et

$$\mathrm{Var}(\boldsymbol{\varepsilon}^*) = \mathrm{Var}(\Omega^{-1}\boldsymbol{\varepsilon}) = \Omega^{-1}\mathrm{Var}(\boldsymbol{\varepsilon})(\Omega^{-1})' = \Omega^{-1}(\sigma^2\Omega^2)\Omega^{-1} = \sigma^2 I_n$$

   (c) D'après la question précédente, le modèle transformé obéit aux hypothèses usuelles du modèle linéaire (centrage, homoscédasticité et décorrélation des erreurs). L'estimateur des moindres carrés $\hat{\beta}^*$ de β est donc

$$\hat{\beta}^* = ((X^*)'(X^*))^{-1}(X^*)'Y^* = \left(X'\Omega^{-2}X\right)^{-1}X'\Omega^{-2}Y$$

   Par les propriétés classiques de l'estimateur des moindres carrés, on sait qu'il est non biaisé et que sa matrice de covariance est

$$\mathrm{Var}(\hat{\beta}^*) = \sigma^2((X^*)'(X^*))^{-1} = \sigma^2\left(X'\Omega^{-2}X\right)^{-1}.$$

   (d) Toujours par la théorie de l'estimation aux moindres carrés, on sait qu'un estimateur non biaisé de σ^2 est

$$\hat{\sigma}_*^2 = \frac{\|Y^* - X^*\hat{\beta}^*\|^2}{n-p} = \frac{\|\Omega^{-1}(Y - X\hat{\beta}^*)\|^2}{n-p}.$$

2. En revenant au modèle initial $Y = X\beta + \boldsymbol{\varepsilon}$, on suppose maintenant les erreurs ε_i gaussiennes, plus précisément $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2\Omega^2)$.

(a) Comme d'habitude, nous notons $x_i' = [x_{i1}, \ldots, x_{ip}]$ la ligne i de la matrice X du plan d'expérience. De par l'indépendance des y_i, la vraisemblance du modèle s'écrit

$$\mathcal{L}(Y, \beta, \sigma^2) = \prod_{i=1}^{n} f_i(y_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2\omega_i^2}} e^{-\frac{(y_i - x_i'\beta)^2}{2\sigma^2\omega_i^2}} = \frac{1}{(2\pi)^{\frac{n}{2}} \det\Omega} \times \frac{1}{(\sigma^2)^{\frac{n}{2}}} e^{-\frac{\|\Omega^{-1}(Y - X\beta)\|^2}{2\sigma^2}}$$

(b) La log-vraisemblance est donc

$$\log \mathcal{L}(Y, \beta, \sigma^2) = C - \frac{n}{2}\log(\sigma^2) - \frac{\|\Omega^{-1}(Y - X\beta)\|^2}{2\sigma^2},$$

où $C = -\log((2\pi)^{\frac{n}{2}}\det\Omega)$ est une constante indépendante des paramètres β et σ^2. Pour toute valeur de σ^2, le maximum en β est atteint en minimisant $\|\Omega^{-1}(Y - X\beta)\| = \|Y^* - X^*\beta\|$, or ceci a été fait précédemment, d'où il vient

$$\hat{\beta}_{mv} = \hat{\beta}^* = \left(X'\Omega^{-2}X\right)^{-1} X'\Omega^{-2}Y$$

ce qui s'écrit de façon équivalente, en prémultipliant les deux membres par $X'\Omega^{-2}X$ et en passant tout à droite

$$X'\Omega^{-2}(Y - X\hat{\beta}_{mv}) = 0$$

Une fois $\hat{\beta}_{mv}$ déterminé, il suffit de maximiser en σ^2 la fonction d'une seule variable

$$\log \mathcal{L}(Y, \hat{\beta}_{mv}, \sigma^2) = C - \frac{n}{2}\log(\sigma^2) - \frac{\|\Omega^{-1}(Y - X\hat{\beta}_{mv})\|^2}{2\sigma^2},$$

ce qui se fait en annulant sa dérivée

$$\frac{\partial \log \mathcal{L}(Y, \hat{\beta}_{mv}, \sigma^2)}{\partial\sigma^2} = -\frac{n}{2} \times \frac{1}{\sigma^2} + \frac{\|\Omega^{-1}(Y - X\hat{\beta}_{mv})\|^2}{2\sigma^4}$$

On obtient bien

$$\hat{\sigma}^2_{mv} = \frac{\|\Omega^{-1}(Y - X\hat{\beta}_{mv})\|^2}{n}.$$

(c) Nous avons donc $\hat{\beta}_{mv} = \hat{\beta}^*$ d'une part, et $\hat{\sigma}^2_{mv} = \frac{(n-p)\hat{\sigma}^2_*}{n}$ d'autre part.

(d) Par les propriétés classiques de l'estimateur du maximum de vraisemblance dans le cas du modèle linéaire gaussien, nous avons donc

$$\hat{\beta}^* \sim \mathcal{N}(\beta, \sigma^2((X^*)'(X^*))^{-1}) \sim \mathcal{N}(\beta, \sigma^2\left(X'\Omega^{-2}X\right)^{-1}).$$

De même, le théorème de Cochran permet de montrer que

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

loi du chi-deux à $(n-p)$ degrés de liberté.

3. Supposons maintenant le modèle classique de régression linéaire $Y = X\beta + \boldsymbol{\varepsilon}$, avec les erreurs centrées et de matrice de covariance $\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2 I_n$. Néanmoins, on n'observe pas comme d'habitude les x_i' et y_i, mais des moyennes par classe. Spécifiquement, les n données sont réparties en L classes $C_1, \ldots, C_L$ d'effectifs respectifs connus $n_1, \ldots, n_L$ et on a seulement accès aux moyennes par classe, à savoir pour tout $\ell \in \{1, \ldots, L\}$:

$$\bar{y}_\ell = \frac{1}{n_\ell}\sum_{i \in C_\ell} y_i \qquad \& \qquad \bar{x}_{\ell j} = \frac{1}{n_\ell}\sum_{i \in C_\ell} x_{ij}$$

(a) Dans ce contexte, il suffit de noter $\bar{Y} = [\bar{y}_1, \ldots, \bar{y}_L]'$, $\bar{X}$ la matrice $L \times p$ de terme générique $\bar{x}_{\ell j}$ et $\bar{\boldsymbol{\varepsilon}} = [\bar{\varepsilon}_1, \ldots, \bar{\varepsilon}_L]'$ pour obtenir l'écriture matricielle

$$\bar{Y} = \bar{X}\beta + \bar{\boldsymbol{\varepsilon}}$$

(b) Le vecteur aléatoire $\bar{\boldsymbol{\varepsilon}}$ est de moyenne nulle et de matrice de covariance diagonale, ses termes diagonaux étant égaux à $\frac{\sigma^2}{n_1}, \ldots, \frac{\sigma^2}{n_L}$.

(c) Tous les calculs précédents s'appliquent en remplaçant n par L et ω_i par $\frac{1}{\sqrt{n_i}}$. On obtient donc pour estimateur de β

$$\hat{\beta}^* = \left(\bar{X}'\Omega^{-2}\bar{X}\right)^{-1}\bar{X}'\Omega^{-2}\bar{Y}$$

et pour estimateur de σ^2

$$\hat{\sigma}^2_* = \frac{\|\Omega^{-1}(\bar{Y} - \bar{X}\hat{\beta}^*)\|^2}{L - p}.$$

Université Rennes 2
Master de Statistiques
Durée : 2 heures

Mardi 18 Décembre 2012
Calculatrice autorisée
Aucun document

Contrôle de Régression Linéaire

I. Octopus's Garden

On cherche à mettre en œuvre une stratégie de prédiction du poids utile du poulpe, c'est-à-dire son poids éviscéré, à partir de son poids non éviscéré. C'est en effet le poulpe éviscéré qui est commercialisé. Pour cela, un échantillon de poulpes a été collecté en 2003 lors des opérations de pêche dans les eaux mauritaniennes. Vu l'importante différence de poids entre les poulpes mâles et les poulpes femelles, on étudie ici uniquement les données concernant 240 poulpes femelles.

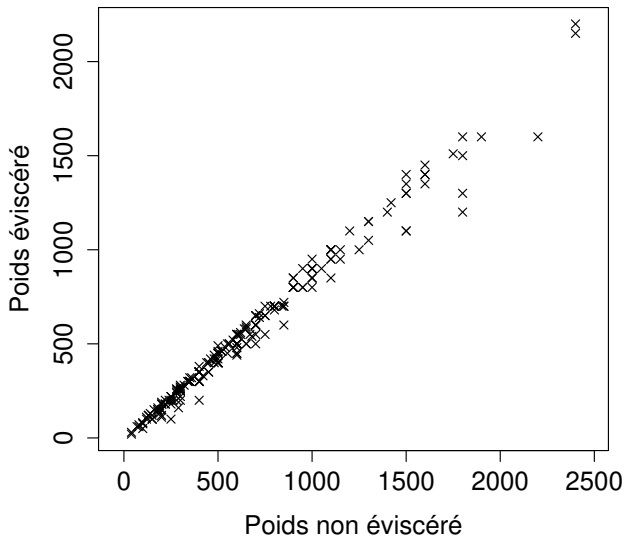


FIGURE A.6 – Poids de poulpe éviscéré en fonction du poids non éviscéré (en grammes).

1. L'ensemble de ces données est représenté figure A.6.
   (a) Proposer un modèle reliant le poids éviscéré et le poids non éviscéré d'un poulpe.
   (b) Rappeler les formules des estimateurs des paramètres du modèle.
   (c) A partir du tableau A.2, donner les estimations numériques des paramètres du modèle.
   (d) Que représente la valeur 0.698 du tableau A.2 ? Comment la retrouver (à peu près) à partir de -0.388 et de la table de la loi normale donnée en annexe (faire un dessin).
   (e) Au vu de cette valeur 0.698, proposer un autre modèle reliant les poids éviscéré et non éviscéré.
2. De façon générale, considérons un échantillon de n couples de réels (x_i, y_i) suivant le modèle $y_i = \beta x_i + \varepsilon_i$, où les erreurs ε_i sont supposées gaussiennes indépendantes centrées et de même variance σ^2.

```
                        Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)            -2.312146     5.959670    -0.388    0.698
Poids non éviscéré      0.853169     0.007649   111.545   <2e-16

Residual standard error: 52.73 on 238 degrees of freedom
Multiple R-Squared: 0.9812,     Adjusted R-squared: 0.9812
F-statistic: 1.244e+04 on 1 and 238 DF,  p-value: < 2.2e-16
```

TABLE A.2 – Poids de poulpes éviscérés et non éviscérés : résultats de la régression linéaire simple (sortie R).

(a) Déterminer l'estimateur $\tilde{\beta}$ de β minimisant la somme des carrés des écarts au modèle.

(b) Retrouver le résultat précédent à partir de la formule générale de l'estimateur de régression linéaire multiple en considérant la projection du vecteur $Y = [y_1, \ldots, y_n]'$ sur la droite vectorielle engendrée par le vecteur $X = [x_1, \ldots, x_n]'$.

(c) En déduire la variance de $\tilde{\beta}$. Proposer un estimateur non biaisé $\tilde{\sigma}^2$ de σ^2.

```
                      Estimate Std. Error t value Pr(>|t|)
Poids non éviscéré     0.85073    0.00436   195.1   <2e-16

Residual standard error: 52.63 on 239 degrees of freedom
Multiple R-Squared: 0.9938,     Adjusted R-squared: 0.9937
F-statistic: 3.807e+04 on 1 and 239 DF,  p-value: < 2.2e-16
```

TABLE A.3 – Poids de poulpes éviscérés et non éviscérés : résultats de la régression linéaire simple avec le modèle simplifié (sortie R).

(d) Les résultats de l'analyse de ce nouveau modèle sont fournis dans le tableau A.3. Localiser $\tilde{\beta}$ et $\tilde{\sigma}^2$ dans ce tableau.

(e) On veut prédire le poids éviscéré d'un poulpe de poids non éviscéré x_0. Quelle est la variance de l'erreur de prévision ? Donner un intervalle de confiance à 90% autour de la prévision.

II. Comparaison de modèles

On effectue une régression de y sur deux variables explicatives x et z à partir d'un échantillon de n individus, c'est-à-dire que $X = [\mathbb{1}, \boldsymbol{x}, \boldsymbol{z}]$, où $\mathbb{1}$ est le vecteur de taille n composé de 1. On a obtenu le résultat suivant :

$$X'X = \begin{bmatrix} 5 & 3 & 0 \\ 3 & 3 & 1 \\ 0 & 1 & 1 \end{bmatrix}.$$

1. Que vaut n ?
2. Que vaut le coefficient de corrélation linéaire empirique entre x et z ?
3. La régression par moindres carrés ordinaires a donné le résultat suivant

$$\hat{y}_i = -1 + 3x_i + 4z_i + \hat{\varepsilon}_i$$

et la somme des carrés résiduelle vaut $\|\hat{\boldsymbol{\varepsilon}}\|^2 = 3$.

(a) Exprimer $X'Y$ en fonction de $(X'X)$ et $\hat{\beta}$, et calculer $X'Y$. En déduire $\bar{y}$.

(b) Calculer $\|\hat{Y}\|^2$. En déduire $\|Y\|^2$.

(c) Calculer la somme des carrés totale $\|Y - \bar{y}\mathbb{1}\|^2$, le coefficient de détermination R^2 et le coefficient de détermination ajusté.

4. On s'intéresse maintenant au modèle privé du régresseur z, c'est-à-dire $Y = X_0\beta_0 + \boldsymbol{\varepsilon}_0$, où $X_0 = [\mathbb{1}, \boldsymbol{x}]$.
   (a) Déterminer $X_0'X_0$ et $X_0'Y$. En déduire $\hat{\beta}_0$.
   (b) Calculer $\|\hat{Y}_0\|^2$.
   (c) Justifier l'égalité $\|\hat{Y}_0\|^2 + \|\hat{\boldsymbol{\varepsilon}}_0\|^2 = \|\hat{Y}\|^2 + \|\hat{\boldsymbol{\varepsilon}}\|^2$. En déduire $\|\hat{\boldsymbol{\varepsilon}}_0\|^2$, le coefficient de détermination R_0^2 et le coefficient de détermination ajusté.
5. On veut maintenant comparer les deux modèles précédents.
   (a) Effectuer un test de Fisher entre ces deux modèles grâce aux coefficients de détermination. Qu'en concluez-vous au niveau de risque 5% ?
   (b) Proposer un autre moyen d'arriver au même résultat.

III. Minimisation de l'erreur de prévision

1. Soit un échantillon de n couples de réels $(x_i, y_i)_{1\leq i\leq n}$ pour le modèle de régression linéaire simple $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, où les erreurs ε_i sont supposées centrées décorrélées et de même variance σ^2. On estime $\beta = (\beta_0, \beta_1)$ par la méthode des moindres carrés ordinaires, ce qui donne $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)$.
   (a) Soit x_{n+1} une nouvelle valeur de la variable explicative pour laquelle on veut prédire la variable réponse y_{n+1}. Qu'appelle-t-on erreur de prévision ? Rappeler sa variance telle qu'elle est énoncée dans le chapitre sur la régression linéaire simple.
   (b) Rappeler sa variance telle qu'elle est énoncée dans le chapitre sur la régression linéaire multiple.
   (c) Retrouver le résultat de la question 1a à partir de celui de la question 1b.
   (d) A partir du résultat de la question 1a, trouver pour quelle valeur de x_{n+1} la variance de l'erreur de prévision est minimale. Que vaut alors cette variance ?
2. Le but de cette partie est de généraliser le résultat de la question 1d. Nous considérons désormais un échantillon $(x_i', y_i)_{1\leq i\leq n}$, où $x_i' = [1, z_i']$ avec $z_i' = [x_{i1}, \ldots, x_{ip}]$. En notant $\mathbb{1}$ le vecteur de taille n uniquement composé de 1, nous adoptons l'écriture matricielle :

$$X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix} = \left[\begin{array}{cc} 1 & z_1' \\ \hline \vdots & \vdots \\ \hline 1 & z_n' \end{array}\right] = \left[\begin{array}{c|c|c|c} \mathbb{1} & Z_1 & \cdots & Z_p \end{array}\right] = \left[\begin{array}{c|c} \mathbb{1} & Z \end{array}\right],$$

   où Z est donc une matrice de taille $n \times p$. Les moyennes de ses colonnes $Z_1, \ldots, Z_p$ sont regroupées dans le vecteur ligne $\bar{x}' = [\bar{x}_1, \ldots, \bar{x}_p]$. Enfin, on considère comme précédemment le modèle de régression linéaire

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = x_i'\beta + \varepsilon_i,$$

   où les erreurs ε_i sont supposées centrées indépendantes et de même variance σ^2. Matriciellement, ceci s'écrit donc $Y = X\beta + \boldsymbol{\varepsilon}$, avec X donnée ci-dessus et supposée telle que $X'X$ est inversible.
   (a) Ecrire la matrice $X'X$ sous forme de 4 blocs faisant intervenir Z, $\bar{x}$ et la taille n de l'échantillon.

(b) On rappelle la formule d'inversion matricielle par blocs : Soit M une matrice inversible telle que

$$M = \left[\begin{array}{c|c} T & U \\ \hline V & W \end{array}\right]$$

avec T inversible, alors $Q = W - VT^{-1}U$ est inversible et l'inverse de M est :

$$M^{-1} = \left[\begin{array}{c|c} T^{-1} + T^{-1}UQ^{-1}VT^{-1} & -T^{-1}UQ^{-1} \\ \hline -Q^{-1}VT^{-1} & Q^{-1} \end{array}\right].$$

Ecrire la matrice $(X'X)^{-1}$ sous forme de 4 blocs dépendant de n, $\bar{x}$ et Γ^{-1}, où $\Gamma = \frac{1}{n}Z'Z - \bar{x}\bar{x}'$.

(c) Soit $x'_{n+1} = [1, z'_{n+1}]$ une nouvelle donnée. Montrer que la variance de l'erreur de prévision est égale à

$$\mathrm{Var}(\hat{\varepsilon}_{n+1}) = \sigma^2 \left(1 + \frac{1}{n} + \frac{1}{n}(z_{n+1} - \bar{x})'\Gamma^{-1}(z_{n+1} - \bar{x})\right).$$

(d) On admet pour l'instant que $\Gamma = \frac{1}{n}Z'Z - \bar{x}\bar{x}'$ est symétrique définie positive (on rappelle que S est symétrique définie positive si $S' = S$ et si pour tout vecteur x non nul, $x'Sx > 0$). Pour quelle nouvelle donnée x'_{n+1} la variance de l'erreur de prévision est-elle minimale ? Que vaut alors cette variance ?

(e) (Bonus) Justifier le fait que si $X'X$ est inversible, alors Γ est bien symétrique définie positive.

Université Rennes 2
Master de Statistiques
Durée : 2 heures

Mardi 18 Décembre 2012
Calculatrice autorisée
Aucun document

Corrigé du Contrôle

I. Octopus's Garden

1. (a) Vu la forme du nuage de points, il semble raisonnable de proposer un modèle de régression linéaire simple : en notant x le poids non éviscéré et y le poids éviscéré, on suggère donc $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$, avec comme d'habitude les erreurs ε_i supposées gaussiennes indépendantes centrées et de même variance σ^2.

   (b) Les formules des estimateurs des moindres carrés du modèle sont :

   $$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x},$$

   avec :

   $$\hat{\beta}_2 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2},$$

   où $n = 240$ sur notre exemple. Un estimateur non biaisé de σ^2 est quant à lui

   $$\hat{\sigma}^2 = \frac{1}{n-2} \sum (y_i - (\hat{\beta}_1 + \hat{\beta}_2 x_i))^2$$

   (c) Du tableau 1, on déduit que $\hat{\beta}_1 \approx -2.31$, $\hat{\beta}_2 \approx 0.85$, et $\hat{\sigma} \approx 52.7$.

   (d) Sous l'hypothèse $H_0 : \beta_1 = 0$, nous savons que $T = \hat{\beta}_1 / \hat{\sigma}_{\hat{\beta}_1} \sim \mathcal{T}_{238}$, loi de Student à 238 degrés de libertés. La statistique de test est ici $T(\omega) = -0.368$, et la probabilité que la valeur absolue d'une loi Student à 238 ddl dépasse 0.368 est environ 0.698. Pour retrouver ce résultat à partir de la table de la loi normale : si $X \sim \mathcal{T}_{238}$, alors par symétrie de la loi de Student et son approximation par une loi normale centrée réduite on obtient successivement

   $$\mathbb{P}(|X| > 0.388) = 2 \times (1 - \mathbb{P}(X \leq 0.388)) \approx 2 \times (1 - \mathbb{P}(\mathcal{N}(0,1) \leq 0.388))$$

   et d'après la table de la loi normale

   $$\mathbb{P}(\mathcal{N}(0,1) \leq 0.388) \approx \mathbb{P}(X \leq 0.39) \approx 0.652$$

   d'où $\mathbb{P}(|X| > 0.388) \approx 0.696$, qui n'est pas bien loin du 0.698 du listing.

   (e) Ceci nous amène à accepter H_0 et à proposer un modèle sans la constante, à savoir : $y_i = \beta x_i + \varepsilon_i$, où les erreurs ε_i sont supposées gaussiennes indépendantes centrées et de même variance σ^2.

2. De façon générale, considérons un échantillon de n couples de réels (x_i, y_i) suivant le modèle $y_i = \beta x_i + \varepsilon_i$, où les erreurs ε_i sont supposées gaussiennes indépendantes centrées et de même variance σ^2.

(a) L'estimateur $\tilde{\beta}$ s'obtient en minimisant :

$$S(\beta) = \sum_{i=1}^{n}(y_i - \beta x_i)^2 \;\Rightarrow\; S'(\beta) = -2\sum_{i=1}^{n} x_i(y_i - \beta x_i) = 0 \;\Leftrightarrow\; \tilde{\beta} = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}$$

(b) L'estimateur précédent revient à considérer la projection du vecteur $Y = [y_1, \ldots, y_n]'$ sur la droite vectorielle engendrée par le vecteur $X = [x_1, \ldots, x_n]'$. Nous pouvons donc appliquer la formule générale :

$$\tilde{\beta} = (X'X)^{-1}X'Y = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}.$$

(c) La variance de $\tilde{\beta}$ se déduit elle lui aussi de la formule générale :

$$\mathrm{Var}(\tilde{\beta}) = \sigma^2(X'X)^{-1} = \frac{\sigma^2}{\sum_{i=1}^n x_i^2}.$$

Puisque le nombre de paramètres est égal à $p = 1$, un estimateur non biaisé σ^2 est :

$$\tilde{\sigma}^2 = \frac{\|Y - X\tilde{\beta}\|^2}{n-1} = \frac{1}{n-1}\sum(y_i - \tilde{\beta}x_i)^2$$

(d) Le tableau 2 indique que $\tilde{\beta} \approx 0.85$ et $\tilde{\sigma} \approx 52.6$.

(e) On veut prédire le poids éviscéré y_0 d'un poulpe de poids non éviscéré x_0. La variance de l'erreur de prévision $\tilde{\varepsilon}_0 = y_0 - \tilde{\beta}x_0$ est elle aussi donnée par la formule générale :

$$\mathrm{Var}(\tilde{\varepsilon}_0) = \sigma^2(1 + x_0(X'X)^{-1}x_0) = \sigma^2\left(1 + \frac{x_0^2}{\sum_{i=1}^n x_i^2}\right).$$

Puisqu'on ne connaît pas l'écart-type σ, on le remplace par son estimation $\tilde{\sigma}$ et on sait alors que

$$\frac{y_0 - \tilde{\beta}x_0}{\tilde{\sigma}\sqrt{1 + \frac{x_0^2}{\sum_{i=1}^n x_i^2}}} \sim \mathcal{T}_{239}$$

d'où l'on déduit un intervalle de prévision à 90%

$$IP(y_0) = \left[\tilde{\beta}x_0 - t_{239}(0.95)\tilde{\sigma}\sqrt{1 + \frac{x_0^2}{\sum_{i=1}^n x_i^2}}, \tilde{\beta}x_0 + t_{239}(0.95)\tilde{\sigma}\sqrt{1 + \frac{x_0^2}{\sum_{i=1}^n x_i^2}}\right],$$

où $t_{239}(0.95)$ représente le quantile d'ordre 0.95 d'une Student à 239 ddl, soit environ 1.653.

II. Comparaison de modèles

Puisque $X = [\mathbb{1}, \boldsymbol{x}, \boldsymbol{z}]$, on a :

$$X'X = \begin{bmatrix} 5 & 3 & 0 \\ 3 & 3 & 1 \\ 0 & 1 & 1 \end{bmatrix} = \begin{bmatrix} n & n\bar{x} & n\bar{z} \\ n\bar{x} & \sum x_i^2 & \sum x_i z_i \\ n\bar{z} & \sum x_i z_i & \sum z_i^2 \end{bmatrix}.$$

1. Il en découle que $n = 5$.

2. Le coefficient de corrélation linéaire empirique entre x et z s'écrit

$$\rho_{x,z} = \frac{\sum x_i z_i - n\bar{x}\bar{z}}{\sqrt{\sum x_i^2 - n\bar{x}^2}\sqrt{\sum z_i^2 - n\bar{z}^2}} = \sqrt{\frac{5}{6}} \approx 0.91.$$

3. (a) Puisque $\hat{\beta} = (X'X)^{-1}X'Y = [-1, 3, 4]'$, on en déduit que

$$X'Y = (X'X)\hat{\beta} = \begin{bmatrix} 5 & 3 & 0 \\ 3 & 3 & 1 \\ 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} -1 \\ 3 \\ 4 \end{bmatrix} = \begin{bmatrix} 4 \\ 10 \\ 7 \end{bmatrix} = \begin{bmatrix} n\bar{y} \\ \sum x_i y_i \\ \sum z_i y_i \end{bmatrix}.$$

En particulier, on a donc $\bar{y} = 4/5$.

(b) Un calcul direct donne

$$\|\hat{Y}\|^2 = \hat{\beta}' X'Y = 54.$$

On applique alors Pythagore :

$$\|Y\|^2 = \|\hat{Y}\|^2 + \|\hat{\varepsilon}\|^2 = 57.$$

(c) Puisque $\bar{y}\mathbb{1}$ est le projeté orthogonal de Y sur la droite vectorielle $\mathbb{R}\mathbb{1}$, la somme des carrés totale est alors immédiate, toujours par Pythagore :

$$\|Y - \bar{y}\mathbb{1}\|^2 = \|Y\|^2 - \|\bar{y}\mathbb{1}\|^2 = \|Y\|^2 - n\bar{y}^2 = \frac{269}{5} = 53.8.$$

Par définition, le coefficient de détermination s'écrit

$$R^2 = 1 - \frac{\|\hat{\varepsilon}\|^2}{\|Y - \bar{y}\mathbb{1}\|^2} = \frac{254}{269} \approx 0.94$$

et le coefficient de détermination ajusté tient compte des dimensions, soit

$$R_a^2 = 1 - \frac{n-1}{n-3} \times \frac{\|\hat{\varepsilon}\|^2}{\|Y - \bar{y}\mathbb{1}\|^2} = \frac{239}{269} \approx 0.89$$

4. On s'intéresse maintenant au modèle privé du régresseur z, c'est-à-dire $Y = X_0\beta_0 + \boldsymbol{\varepsilon}_0$, où $X_0 = [\mathbb{1}, \boldsymbol{x}]$.

(a) La matrice $X_0'X_0$ se déduit de $X'X$:

$$X_0'X_0 = \begin{bmatrix} n & n\bar{x} \\ n\bar{x} & \sum x_i^2 \end{bmatrix} = \begin{bmatrix} 5 & 3 \\ 3 & 3 \end{bmatrix}.$$

Idem pour le vecteur $X_0'Y$ à partir de $X'Y$:

$$X_0'Y = \begin{bmatrix} n\bar{y} \\ \sum x_i y_i \end{bmatrix} = \begin{bmatrix} 4 \\ 10 \end{bmatrix}.$$

Il vient donc

$$\hat{\beta}_0 = (X_0'X_0)^{-1}X_0'Y = \begin{bmatrix} -3 \\ 19/3 \end{bmatrix}.$$

(b) Nous avons comme précédemment

$$\|\hat{Y}_0\|^2 = \hat{\beta}_0' X_0'Y = \frac{154}{3} \approx 51.3$$

(c) Un coup de Pythagore dans chaque modèle donne

$$\|Y\|^2 = \|\hat{Y}_0\|^2 + \|\hat{\varepsilon}_0\|^2 = \|\hat{Y}\|^2 + \|\hat{\varepsilon}\|^2,$$

d'où l'on tire

$$\|\hat{\varepsilon}_0\|^2 = \|Y\|^2 - \|\hat{Y}_0\|^2 = \frac{17}{3} \approx 5.7$$

Le coefficient de détermination vaut donc

$$R_0^2 = 1 - \frac{\|\hat{\varepsilon}_0\|^2}{\|Y - \bar{y}\mathbb{1}\|^2} = \frac{722}{807} = \approx 0.89$$

et le coefficient de détermination ajusté

$$R_{a,0}^2 = 1 - \frac{n-1}{n-2} \times \frac{\|\hat{\varepsilon}_0\|^2}{\|Y - \bar{y}\mathbb{1}\|^2} = \frac{2081}{2421} \approx 0.86$$

5. On veut maintenant comparer les deux modèles précédents.

(a) Sous $H_0 : \beta_z = 0$, le test de Fisher entre les deux modèles s'écrit

$$F = \frac{n-p}{p-p_0} \times \frac{R^2 - R_0^2}{1 - R^2} \sim \mathcal{F}_{n-p}^{p-p_0} = \mathcal{F}_2^1$$

La statistique de test vaut ici

$$F(\omega) = \frac{16}{9} \approx 1.78 \ll 18.5 \approx f_2^1(0.95)$$

et on accepte donc l'hypothèse selon laquelle $\beta_z = 0$.

(b) Nous aurions pu tester cette hypothèse grâce à un test de Student sur le modèle initial, puisque sous H_0, on sait que

$$T = \frac{\hat{\beta}_z}{\hat{\sigma}_{\hat{\beta}_z}} \sim \mathcal{T}_{n-p} = \mathcal{T}_2,$$

or $\hat{\sigma}_{\hat{\beta}_z} = \hat{\sigma}\sqrt{[(X'X)^{-1}]_{3,3}}$, avec

$$\hat{\sigma} = \sqrt{\frac{\|\hat{\varepsilon}\|^2}{n-p}} = \sqrt{\frac{3}{2}}$$

et

$$\left[(X'X)^{-1}\right]_{3,3} = \frac{1}{\det(X'X)} \begin{vmatrix} 5 & 3 \\ 3 & 3 \end{vmatrix} = 6$$

d'où

$$|T(\omega)| = \frac{4}{3} \ll 4.303 \approx t_2(0.975)$$

Ces deux tests reviennent au même puisque $F(\omega) = T^2(\omega)$ et $f_2^1(0.95) = (t_2(0.975))^2$.

III. Minimisation de l'erreur de prévision

1. (a) L'erreur de prévision est par définition

$$\hat{\varepsilon}_{n+1} = y_{n+1} - \hat{y}_{n+1} = y_{n+1} - (\hat{\beta}_0 + \hat{\beta}_1 x_{n+1}).$$

On montre que sa variance vaut

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

(b) En notant X la matrice $n \times 2$ définie par

$$X = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

nous avons de façon générale

$$\mathrm{Var}(\hat{\varepsilon}_{n+1}) = \sigma^2 \left(1 + [1, x_{n+1}](X'X)^{-1}[1, x_{n+1}]'\right).$$

(c) Puisque

$$X'X = \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix} = \begin{bmatrix} n & n\bar{x} \\ n\bar{x} & \sum x_i^2 \end{bmatrix}$$

son inversion donne

$$(X'X)^{-1} = \frac{1}{n\sum x_i^2 - n^2\bar{x}^2} \begin{bmatrix} \sum x_i^2 & -n\bar{x} \\ -n\bar{x} & n \end{bmatrix} = \frac{1}{\sum (x_i - \bar{x})^2} \begin{bmatrix} \frac{\sum x_i^2}{n} & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix},$$

d'où

$$[1, x_{n+1}](X'X)^{-1}[1, x_{n+1}]' = \frac{1}{\sum (x_i - \bar{x})^2} \left(\frac{\sum x_i^2}{n} - 2\bar{x}x_{n+1} + x_{n+1}^2\right) = \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

et l'on retrouve bien que

$$\sigma^2 \left(1 + [1, x_{n+1}](X'X)^{-1}[1, x_{n+1}]'\right) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right),$$

i.e. le résultat de la question 1a.

(d) A partir de cette formule, il est clair que l'erreur de prévision est minimale (en moyenne) lorsque $x_{n+1} = \bar{x}$, la variance de l'erreur valant alors $\sigma^2(1 + 1/n)$.

2. (a) La matrice $X'X$ s'écrit sous forme de 4 blocs comme suit

$$X'X = \left[\begin{array}{c|c} n & n\bar{x}' \\ \hline n\bar{x} & Z'Z \end{array}\right] = n \left[\begin{array}{c|c} 1 & \bar{x}' \\ \hline \bar{x} & \frac{Z'Z}{n} \end{array}\right].$$

(b) Avec les notations de la formule d'inversion matricielle par blocs appliquée à $X'X$, nous posons $T = 1$, $U = \bar{x}'$, $V = \bar{x}$ et $W = \frac{Z'Z}{n}$. Toujours avec les notations de l'énoncé, nous avons donc

$$Q = \frac{1}{n} Z'Z - \bar{x}\bar{x}' = \Gamma$$

et

$$(X'X)^{-1} = \frac{1}{n} \left[\begin{array}{c|c} 1 + \bar{x}'\Gamma^{-1}\bar{x} & -\bar{x}'\Gamma^{-1} \\ \hline -\Gamma^{-1}\bar{x} & \Gamma^{-1} \end{array}\right]$$

(c) Soit $x'_{n+1} = [1, z'_{n+1}]$ une nouvelle donnée. La variance de l'erreur de prévision est comme ci-dessus

$$\mathrm{Var}(\hat{\varepsilon}_{n+1}) = \sigma^2 \left(1 + x'_{n+1}(X'X)^{-1}x_{n+1}\right).$$

En utilisant l'écriture par blocs de $(X'X)^{-1}$ et $x'_{n+1} = [1, z'_{n+1}]$, on arrive à

$$x'_{n+1}(X'X)^{-1}x_{n+1} = \frac{1}{n}\left(1 + \bar{x}'\Gamma^{-1}\bar{x} - z'_{n+1}\Gamma^{-1}\bar{x} - \bar{x}'\Gamma^{-1}z_{n+1} + z'_{n+1}\Gamma^{-1}z_{n+1}\right)$$

La matrice Γ est symétrique et un réel est égal à sa transposée (!) donc $z'_{n+1}\Gamma^{-1}\bar{x} = \bar{x}'\Gamma^{-1}z_{n+1}$ et ceci se réécrit

$$x'_{n+1}(X'X)^{-1}x_{n+1} = \frac{1}{n}\left(1 + z'_{n+1}\Gamma^{-1}z_{n+1} - 2\bar{x}'\Gamma^{-1}z'_{n+1} + \bar{x}'\Gamma^{-1}\bar{x}\right)$$

ou encore

$$x'_{n+1}(X'X)^{-1}x_{n+1} = \frac{1}{n}\left(1 + (z_{n+1} - \bar{x})'\Gamma^{-1}(z_{n+1} - \bar{x})\right)$$

si bien que

$$\mathrm{Var}(\hat{\varepsilon}_{n+1}) = \sigma^2\left(1 + \frac{1}{n} + \frac{1}{n}(z_{n+1} - \bar{x})'\Gamma^{-1}(z_{n+1} - \bar{x})\right),$$

qui est la formule escomptée.

(d) Dire que Γ est symétrique définie positive revient à dire qu'elle est symétrique avec toutes ses valeurs propres strictement positives, donc il en va de même pour Γ^{-1}. De fait, le dernier terme de la formule précédente est toujours positif ou nul. Il est nul si et seulement si $z_{n+1} = \bar{x}$, c'est-à-dire lorsque $x'_{n+1} = [1, \bar{x}']$. La variance de l'erreur de prévision vaut alors $\sigma^2(1+1/n)$. Ceci généralise bien le résultat vu en régression linéaire simple : il faut se placer au centre de gravité du nuage de points des variables explicatives pour prévoir au mieux.

(e) Notons Z_c la matrice $n \times p$ dont les colonnes sont les versions centrées des colonnes de Z, c'est-à-dire respectivement $X_1 - \bar{x}_1\mathbb{1}, \ldots, X_p - \bar{x}_p\mathbb{1}$. On vérifie sans trop se faire de nœuds que $\Gamma = \frac{1}{n}Z'_cZ_c$, si bien que pour tout vecteur u de $\mathbb{R}^p$

$$u'\Gamma u = \frac{1}{n}u'Z'_cZ_cu = \frac{1}{n}\|Z_cu\|^2 \geq 0,$$

avec nullité si et seulement si $Z_cu = 0$. Or $Z_cu = 0$ signifie que

$$u_1(X_1 - \bar{x}_1\mathbb{1}) + \cdots + u_p(X_p - \bar{x}_p\mathbb{1}) = 0 \iff u_1X_1 + \cdots + u_pX_p = \left(\sum_{j=1}^{p} u_j\bar{x}_j\right)\mathbb{1},$$

c'est-à-dire que la première colonne de X peut s'écrire comme une combinaison linéaire non triviale des p dernières. En particulier X serait alors de rang inférieur ou égal à p, ce qui serait en contradiction avec l'hypothèse d'inversibilité de $X'X$. Ainsi Γ est bien symétrique définie positive et la messe est dite.

Annexe B

Rappels d'algèbre

Nous ne considérons ici que des matrices réelles. Nous notons A une matrice et A' sa transposée.

B.1 Quelques définitions

Une matrice carrée A est inversible s'il existe une matrice B telle que $AB = BA = I$. On note $B = A^{-1}$.
La matrice carrée A est dite : symétrique si $A' = A$; singulière si $\det(A) = 0$; inversible si $\det(A) \neq 0$; idempotente si $A^2 = A$; orthogonale si $A' = A^{-1}$.
Le polynôme caractéristique de la matrice carrée A est défini par $P_A(\lambda) = \det(\lambda I - A)$. Les valeurs propres sont les solutions de $\det(\lambda I - A) = 0$. Le vecteur x est un vecteur propre associé à la valeur propre λ s'il est non nul et vérifie $Ax = \lambda x$.

B.2 Quelques propriétés

B.2.1 Les matrices $n \times p$

- $(A + B)' = A' + B'$ et $(AB)' = B'A'$.
- Le rang d'une matrice $A_{n\times p}$ est la plus petite des dimensions des deux sous-espaces engendrés respectivement par les lignes et par les colonnes de A.
- $0 \leq \text{rang}(A) \leq \min(n, p)$.
- $\text{rang}(A) = \text{rang}(A')$.
- $\text{rang}(AB) \leq \min(\text{rang}(A), \text{rang}(B))$.
- $\text{rang}(BAC) = \text{rang}(A)$ si B et C sont inversibles.
- $\text{rang}(AA') = \text{rang}(A'A) = \text{rang}(A)$.
- Pour $p \leq n$, si A est de rang p, alors $A'A$ est inversible.

B.2.2 Les matrices carrées $n \times n$

Soit A et B des matrices carrées de taille $n \times n$ de termes courants a_{ij} et b_{ij}.

- $\text{tr}(A) = \sum_{i=1}^{n} a_{ii}$.
- $\text{tr}(A + B) = \text{tr}(A) + \text{tr}(B), \text{tr}(AB) = \text{tr}(BA)$ et $\text{tr}(\alpha A) = \alpha \,\text{tr}(A)$.
- $\text{tr}(AA') = \text{tr}(A'A) = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij}^2$.
- $\det(AB) = \det(A) \det(B)$.
- Si $\det(A) \neq 0$, la matrice A est inversible, d'inverse notée A^{-1}, vérifiant $(A^{-1})' = (A')^{-1}$ et $\det(A^{-1}) = 1/\det(A)$. De plus, si B est inversible, alors $(AB)^{-1} = B^{-1}A^{-1}$.
- La trace et le déterminant ne dépendent pas des bases choisies.

B.2.3 Les matrices symétriques

Soit A une matrice carrée symétrique de taille $n \times n$:

- les valeurs propres de A sont réelles.
- les vecteurs propres de A associés à des valeurs propres différentes sont orthogonaux.
- si une valeur propre λ est de multiplicité k, il existe k vecteurs propres orthogonaux qui lui sont associés.
- la concaténation de l'ensemble des vecteurs propres orthonormés forme une matrice orthogonale U. Comme $U' = U^{-1}$, la diagonalisation de A s'écrit simplement $A = U\Delta U'$, où $\Delta = \text{diag}(\lambda_1, \ldots, \lambda_n)$. Pour résumer, on dit qu'une matrice symétrique réelle est diagonalisable en base orthonormée.
- $\text{tr}(A) = \sum_{i=1}^{n} \lambda_i$ et $\det(A) = \prod_{i=1}^{n} \lambda_i$.
- $\text{rang}(A)$ = nombre de valeurs propres λ_i non nulles.
- les valeurs propres de A^2 sont les carrés des valeurs propres de A et ces 2 matrices ont les mêmes vecteurs propres.
- les valeurs propres de A^{-1} (si cette matrice existe) sont les inverses des valeurs propres de A et ces 2 matrices ont les mêmes vecteurs propres.

B.2.4 Les matrices semi-définies positives

Soit A une matrice carrée symétrique de taille $n \times n$:

- La matrice A est semi-définie positive (SDP) si $\forall x \in \mathbb{R}^n$, $x'Ax \geq 0$.
- La matrice A est définie positive (DP) si $\forall x \in \mathbb{R}^n - \{0\}$, $x'Ax > 0$.
- Les valeurs propres d'une matrice SDP sont toutes positives ou nulles (et réciproquement).
- La matrice A est SDP et inversible si et seulement si A est DP.
- Toute matrice A de la forme $A = B'B$ est SDP. En effet $\forall x \in \mathbb{R}^n$, $x'Ax = x'B'Bx = (Bx)'Bx = \|Bx\|^2 \geq 0$, où $\|\cdot\|$ correspond à la norme euclidienne de $\mathbb{R}^n$.
- Toute matrice de projecteur orthogonal est SDP. En effet, les valeurs propres d'un projecteur valent 0 ou 1.
- Si B est SDP, alors $A'BA$ est SDP.
- Si A est DP et si B est SDP, alors $A + B$ est inversible et $A^{-1} - (A + B)^{-1}$ est SDP.

B.3 Propriétés des inverses

Soit M une matrice symétrique inversible de taille $p \times p$, soit u et v deux vecteurs de taille p. Si $u'M^{-1}v \neq -1$, alors nous avons l'inverse suivante :

$$\left(M + uv'\right)^{-1} = M^{-1} - \frac{M^{-1}uv'M^{-1}}{1 + u'M^{-1}v}. \tag{B.1}$$

Soit M une matrice inversible telle que :

$$M = \left(\begin{array}{c|c} T & U \\ \hline V & W \end{array}\right)$$

avec T inversible, alors $Q = W - VT^{-1}U$ est inversible et l'inverse de M est :

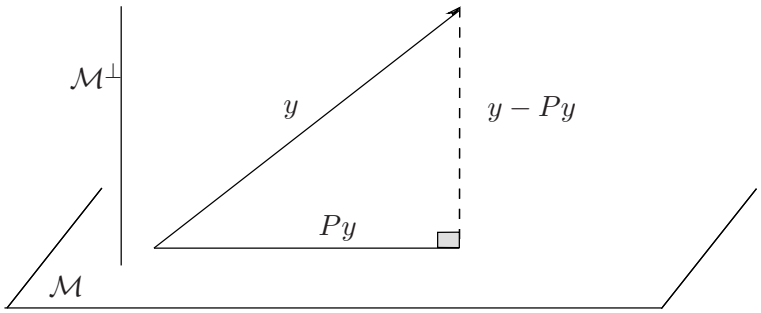
$$M^{-1} = \left(\begin{array}{c|c} T^{-1} + T^{-1}UQ^{-1}VT^{-1} & -T^{-1}UQ^{-1} \\ \hline -Q^{-1}VT^{-1} & Q^{-1} \end{array}\right).$$

B.4 Propriétés des projections

B.4.1 Généralités

Une matrice carrée P idempotente (i.e. $P^2 = P$) correspond à une projection. Si de plus P est symétrique (i.e. $P' = P$) , alors c'est une projection orthogonale sur le sous-espace $\mathcal{M} = \text{Im}(P)$ parallèlement à $\mathcal{M}^\perp = \text{Ker}(P)$.

— P est un projecteur orthogonal si le produit scalaire $\langle Py, y - Py\rangle = 0$ pour tout y.
— les valeurs propres d'une matrice idempotente ne peuvent être égales qu'à 0 ou 1.
— le rang d'une matrice idempotente est égal à sa trace, i.e. $\text{rang}(P) = \dim(\mathcal{M}) = \text{tr}(P)$.
— la matrice $(I - P)$ est la matrice de projection orthogonale sur $\mathcal{M}^\perp = \text{Ker}(P)$.



B.4.2 Exemple de projection orthogonale

Soit $X = [X_1, \cdots, X_p]$ la matrice (n, p), de rang p, des p variables explicatives du modèle linéaire. Soit $\mathcal{M}(X)$ le sous-espace engendré par ces p vecteurs linéairement indépendants et P_X la matrice de projection orthogonale sur $\mathcal{M}(X)$. Le vecteur $(y - P_X y)$ doit être orthogonal à tout vecteur de $\mathcal{M}(X)$, or tous les vecteurs de $\mathcal{M}(X)$ sont de la forme Xu. En particulier il existe un vecteur b tel que $P_X y = Xb$. Il faut donc que $\langle Xu, y - P_X y\rangle = 0$ pour tout vecteur u. En développant, nous obtenons $X'y = X'P_X y = X'Xb$. $X'X$ est inversible donc $b = (X'X)^{-1}X'y$. Ainsi

$$P_X = X(X'X)^{-1}X'$$

est la matrice de projection orthogonale sur $\mathcal{M}(X)$.

B.4.3 Trace et éléments courants

Soit P_X, de terme courant h_{ij}, la matrice $p \times p$ de la projection orthogonale sur l'espace engendré par les p colonnes de X, nous avons alors :

1. $\text{tr}(P_X) = \sum h_{ii} = p$.
2. $\text{tr}(P_X) = \text{tr}(P_X P_X)$, c'est-à-dire $\sum_i \sum_j h_{ij}^2 = p$.
3. $0 \leq h_{ii} \leq 1$ pour tout i.
4. $-0.5 \leq h_{ij} \leq 0.5$ pour tout j différent de i.
5. si $h_{ii} = 1$ alors $h_{ij} = 0$ pour tout j différent de i.
6. si $h_{ii} = 0$, alors $h_{ij} = 0$ pour tout j différent de i.

B.5 Dérivation matricielle

Soit f une fonction de $\mathbb{R}^p$ dans $\mathbb{R}$ différentiable. Le gradient de f au point x est par définition :

$$\nabla f(x) = \text{grad}(f)(x) = \left[\frac{\partial f}{\partial x_1}(x), \cdots, \frac{\partial f}{\partial x_p}(x)\right].$$

Si f est de classe $\mathcal{C}^2$, le hessien de f au point x est la matrice carrée de dimension $p \times p$, souvent notée $\nabla^2 f(x)$ ou $Hf(x)$, de terme générique $[Hf(x)]_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}(x)$. Le théorème de Schwarz assure que cette matrice est symétrique.

Exemples :

- Si $f : \mathbb{R}^p \to \mathbb{R}$ est une forme linéaire, c'est-à-dire s'il existe un vecteur colonne a de taille p tel que $f(x) = a'x$, alors son gradient est constant : $\nabla f = a'$, et sa matrice hessienne est nulle en tout point : $Hf = 0$. Ceci n'est rien d'autre que la généralisation multidimensionnelle des dérivées première et seconde de la fonction $f : \mathbb{R} \to \mathbb{R}$ définie par $f(x) = ax$.
- Si f est quadratique, par exemple si $f(x) = x'Ax$, alors son gradient est une forme linéaire : $\nabla f(x) = x'(A + A')$, et sa hessienne est constante $Hf(x) = A + A'$. A nouveau, ceci n'est rien d'autre que la généralisation multidimensionnelle des dérivées première et seconde de la fonction $f : \mathbb{R} \to \mathbb{R}$ définie par $f(x) = ax^2$.

Annexe C

Rappels de probabilité

C.1 Généralités

$Y = [Y_1, \ldots, Y_n]'$ est un vecteur aléatoire de $\mathbb{R}^n$ si toutes ses composantes $Y_1, \ldots, Y_n$ sont des variables aléatoires réelles.
L'espérance du vecteur aléatoire Y est $\mathbb{E}[Y] = [\mathbb{E}[Y_1], \cdots, \mathbb{E}[Y_n]]'$, vecteur de $\mathbb{R}^n$. La matrice de variance-covariance de Y a pour terme général $\mathrm{Cov}(Y_i, Y_j)$. C'est une matrice de taille $n \times n$, qui s'écrit encore :

$$\mathrm{Var}(Y) = \Sigma_Y = \mathbb{E}\left[(Y - \mathbb{E}[Y])\,(Y - \mathbb{E}[Y])'\right] = \mathbb{E}[YY'] - \mathbb{E}[Y](\mathbb{E}[Y])'.$$

Considérons une matrice (déterministe) A de taille $m \times n$ et un vecteur (déterministe) b de $\mathbb{R}^m$. Soit Y un vecteur aléatoire de $\mathbb{R}^n$, nous avons les égalités suivantes :

$$\begin{aligned} \mathbb{E}[AY + b] &= A\mathbb{E}[Y] + b \\ \mathrm{Var}(AY + b) &= \mathrm{Var}(AY) = A\mathrm{Var}(Y)A' \end{aligned}$$

Si Y est un vecteur aléatoire de $\mathbb{R}^n$ de matrice de variance-covariance Σ_Y, alors pour la norme euclidienne :

$$\mathbb{E}[\|Y - \mathbb{E}(Y)\|^2] = \mathbb{E}\left[\sum_{i=1}^{n}(Y_i - \mathbb{E}[Y_i])^2\right] = \sum_{i=1}^{n}\mathrm{Var}(Y_i) = \mathrm{tr}(\Sigma_Y).$$

Nous avons les égalités utiles suivantes :

$$\mathrm{tr}(\mathbb{E}[YY']) = \mathbb{E}[\mathrm{tr}(YY')] = \mathbb{E}[\mathrm{tr}(Y'Y)] = \mathrm{tr}(\Sigma_Y) + \mathbb{E}[Y]'\mathbb{E}[Y].$$

C.2 Vecteurs aléatoires gaussiens

Un vecteur aléatoire Y est dit gaussien si toute combinaison linéaire de ses composantes est une variable aléatoire gaussienne. Ce vecteur admet alors une espérance μ et une matrice de variance-covariance Σ_Y, et on note $Y \sim \mathcal{N}(\mu, \Sigma_Y)$.
Un vecteur gaussien Y de $\mathbb{R}^n$ d'espérance μ et de matrice de variance-covariance Σ_Y inversible admet pour densité la fonction

$$f(y) = \frac{1}{(2\pi)^{n/2}}\frac{1}{\sqrt{\det(\Sigma_Y)}}\exp\left(-\frac{1}{2}(y-\mu)'\Sigma_Y^{-1}(y-\mu)\right), \quad \text{où } y = [y_1, \ldots, y_n]'.$$

Les composantes d'un vecteur gaussien $Y = [Y_1, \cdots, Y_n]'$ sont indépendantes si et seulement si Σ_Y est diagonale. D'autre part, soit $Y \sim \mathcal{N}(\mu, \Sigma_Y)$, avec Σ_Y inversible, alors

$$(Y - \mu)'\Sigma_Y^{-1}(Y - \mu) \sim \chi^2_n$$

Enfin, le Théorème de Cochran explicite les lois obtenues après projection orthogonale d'un vecteur gaussien.

Théorème 9 (Cochran) *Soit $Y \sim \mathcal{N}(\mu, \sigma^2 I_n)$, $\mathcal{M}$ un sous-espace de $\mathbb{R}^n$ de dimension p et P la matrice de projection orthogonale de $\mathbb{R}^n$ sur $\mathcal{M}$. Nous avons les propriétés suivantes :*

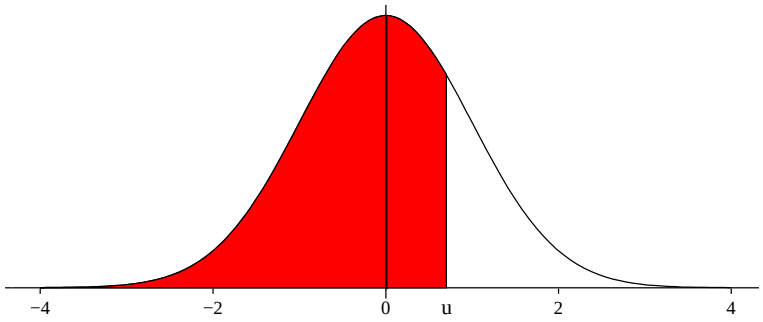
(i) $PY \sim \mathcal{N}(P\mu, \sigma^2 P)$;

(ii) les vecteurs PY et $Y - PY$ sont indépendants ;

(iii) $\|P(Y-\mu)\|^2/\sigma^2 \sim \chi^2_p$.

C.3 Tables des lois usuelles

C.3.1 Loi Normale $X \sim \mathcal{N}(0,1)$



Valeurs de $\Pr(X \leq u)$ en fonction de u.

u	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995

C.3.2 Loi de Student $X \sim \mathcal{T}_\nu$

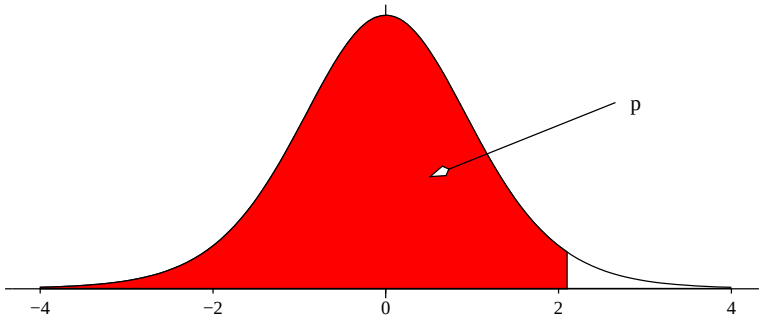


Table des fractiles $t_\nu(p)$ pour une loi de $\mathcal{T}_\nu$: $p = \Pr\{X \leq t_\nu(p)\}$

ν \ p	0.5	0.6	0.7	0.8	0.9	0.95	0.975	0.99	0.995	0.999	0.9995
1	0.000	0.325	0.727	1.376	3.078	6.314	12.706	31.821	63.656	318.289	636.578
2	0.000	0.289	0.617	1.061	1.886	2.920	4.303	6.965	9.925	22.328	31.600
3	0.000	0.277	0.584	0.978	1.638	2.353	3.182	4.541	5.841	10.214	12.924
4	0.000	0.271	0.569	0.941	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.267	0.559	0.920	1.476	2.015	2.571	3.365	4.032	5.894	6.869
6	0.000	0.265	0.553	0.906	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.263	0.549	0.896	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.262	0.546	0.889	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.261	0.543	0.883	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.260	0.542	0.879	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.260	0.540	0.876	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.259	0.539	0.873	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.259	0.538	0.870	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.258	0.537	0.868	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.258	0.536	0.866	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.258	0.535	0.865	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.257	0.534	0.863	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.257	0.534	0.862	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.257	0.533	0.861	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.257	0.533	0.860	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.257	0.532	0.859	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.256	0.532	0.858	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.256	0.532	0.858	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.256	0.531	0.857	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.256	0.531	0.856	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.256	0.531	0.856	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.256	0.531	0.855	1.314	1.703	2.052	2.473	2.771	3.421	3.689
28	0.000	0.256	0.530	0.855	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.256	0.530	0.854	1.311	1.699	2.045	2.462	2.756	3.396	3.660
30	0.000	0.256	0.530	0.854	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.255	0.529	0.851	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.254	0.527	0.848	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.254	0.526	0.846	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.254	0.526	0.845	1.290	1.660	1.984	2.364	2.626	3.174	3.390
200	0.000	0.254	0.525	0.843	1.286	1.653	1.972	2.345	2.601	3.131	3.340
∞	0.000	0.253	0.524	0.842	1.282	1.645	1.960	2.326	2.576	3.090	3.290

C.3.3 Loi du Khi-deux à ν ddl $X \sim \chi^2_\nu$

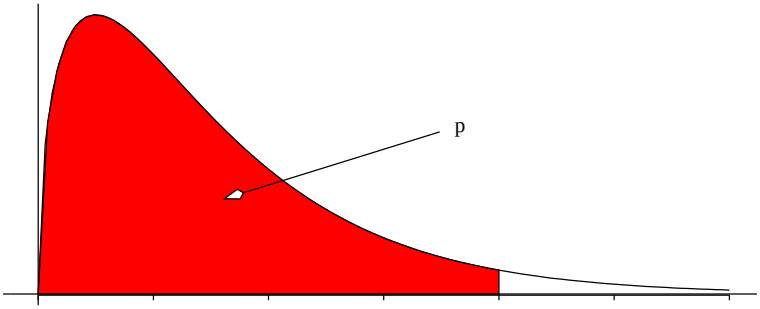


Table des fractiles $c_\nu(p)$ pour une loi du χ^2_ν : $p = \Pr\{X \leq c_\nu(p)\}$

ν \ p	0.001	0.005	0.01	0.025	0.05	0.1	0.9	0.95	0.975	0.99	0.995	0.999
1	0.000	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879	10.827
2	0.002	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597	13.815
3	0.024	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838	16.266
4	0.091	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860	18.466
5	0.210	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.832	15.086	16.750	20.515
6	0.381	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548	22.457
7	0.599	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278	24.321
8	0.857	1.344	1.647	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955	26.124
9	1.152	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589	27.877
10	1.479	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188	29.588
11	1.834	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757	31.264
12	2.214	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300	32.909
13	2.617	3.565	4.107	5.009	5.892	7.041	19.812	22.362	24.736	27.688	29.819	34.527
14	3.041	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319	36.124
15	3.483	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801	37.698
16	3.942	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267	39.252
17	4.416	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718	40.791
18	4.905	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156	42.312
19	5.407	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582	43.819
20	5.921	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997	45.314
21	6.447	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401	46.796
22	6.983	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796	48.268
23	7.529	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181	49.728
24	8.085	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.558	51.179
25	8.649	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928	52.619
26	9.222	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290	54.051
27	9.803	11.808	12.878	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645	55.475
28	10.391	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.994	56.892
29	10.986	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.335	58.301
30	11.588	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672	59.702
40	17.917	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766	73.403
50	24.674	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490	86.660
60	31.738	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952	99.608
70	39.036	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.42	104.21	112.32
80	46.520	51.172	53.540	57.153	60.391	64.278	96.578	101.88	106.63	112.33	116.32	124.84
90	54.156	59.196	61.754	65.647	69.126	73.291	107.56	113.14	118.14	124.12	128.30	137.21
100	61.918	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169	149.449

C.3.4 Loi de Fisher à ν_1, ν_2 ddl $X \sim \mathcal{F}_{\nu_2}^{\nu_1}$

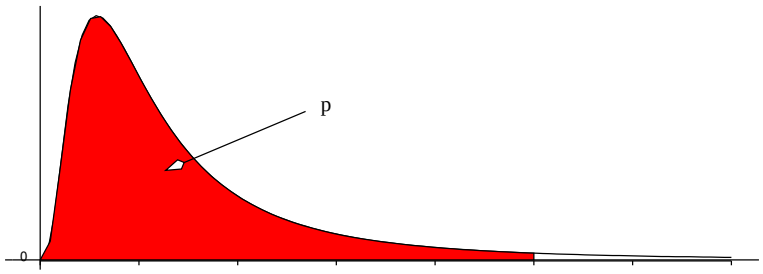


Table des fractiles $f_{(\nu_1,\nu_2)}$) pour une loi $\mathcal{F}_{(\nu_1,\nu_2)}$: $0.95 = \Pr\left\{X \leq f_{(\nu_1,\nu_2)}(p)\right\}$

ν_2 \ ν_1	1	2	3	4	5	6	7	8	9	10	15	20	30	40	50	60	80	100
1	161	199	216	225	230	234	237	239	241	242	246	248	250	251	252	252	253	253
2	18.5	19	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.5	19.5	19.5	19.5	19.5	19.5
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.7	8.66	8.62	8.59	8.58	8.57	8.56	8.55
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6	5.96	5.86	5.8	5.75	5.72	5.7	5.69	5.67	5.66
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.62	4.56	4.5	4.46	4.44	4.43	4.41	4.41
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.1	4.06	3.94	3.87	3.81	3.77	3.75	3.74	3.72	3.71
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.51	3.44	3.38	3.34	3.32	3.3	3.29	3.27
8	5.32	4.46	4.07	3.84	3.69	3.58	3.5	3.44	3.39	3.35	3.22	3.15	3.08	3.04	3.02	3.01	2.99	2.97
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.01	2.94	2.86	2.83	2.8	2.79	2.77	2.76
10	4.96	4.1	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.85	2.77	2.7	2.66	2.64	2.62	2.6	2.59
11	4.84	3.98	3.59	3.36	3.2	3.09	3.01	2.95	2.9	2.85	2.72	2.65	2.57	2.53	2.51	2.49	2.47	2.46
12	4.75	3.89	3.49	3.26	3.11	3	2.91	2.85	2.8	2.75	2.62	2.54	2.47	2.43	2.4	2.38	2.36	2.35
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.53	2.46	2.38	2.34	2.31	2.3	2.27	2.26
14	4.6	3.74	3.34	3.11	2.96	2.85	2.76	2.7	2.65	2.6	2.46	2.39	2.31	2.27	2.24	2.22	2.2	2.19
15	4.54	3.68	3.29	3.06	2.9	2.79	2.71	2.64	2.59	2.54	2.4	2.33	2.25	2.2	2.18	2.16	2.14	2.12
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.35	2.28	2.19	2.15	2.12	2.11	2.08	2.07
17	4.45	3.59	3.2	2.96	2.81	2.7	2.61	2.55	2.49	2.45	2.31	2.23	2.15	2.1	2.08	2.06	2.03	2.02
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.27	2.19	2.11	2.06	2.04	2.02	1.99	1.98
19	4.38	3.52	3.13	2.9	2.74	2.63	2.54	2.48	2.42	2.38	2.23	2.16	2.07	2.03	2	1.98	1.96	1.94
20	4.35	3.49	3.1	2.87	2.71	2.6	2.51	2.45	2.39	2.35	2.2	2.12	2.04	1.99	1.97	1.95	1.92	1.91
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.18	2.1	2.01	1.96	1.94	1.92	1.89	1.88
22	4.3	3.44	3.05	2.82	2.66	2.55	2.46	2.4	2.34	2.3	2.15	2.07	1.98	1.94	1.91	1.89	1.86	1.85
23	4.28	3.42	3.03	2.8	2.64	2.53	2.44	2.37	2.32	2.27	2.13	2.05	1.96	1.91	1.88	1.86	1.84	1.82
24	4.26	3.4	3.01	2.78	2.62	2.51	2.42	2.36	2.3	2.25	2.11	2.03	1.94	1.89	1.86	1.84	1.82	1.8
25	4.24	3.39	2.99	2.76	2.6	2.49	2.4	2.34	2.28	2.24	2.09	2.01	1.92	1.87	1.84	1.82	1.8	1.78
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.07	1.99	1.9	1.85	1.82	1.8	1.78	1.76
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.2	2.06	1.97	1.88	1.84	1.81	1.79	1.76	1.74
28	4.2	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.04	1.96	1.87	1.82	1.79	1.77	1.74	1.73
29	4.18	3.33	2.93	2.7	2.55	2.43	2.35	2.28	2.22	2.18	2.03	1.94	1.85	1.81	1.77	1.75	1.73	1.71
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.01	1.93	1.84	1.79	1.76	1.74	1.71	1.7
32	4.15	3.29	2.9	2.67	2.51	2.4	2.31	2.24	2.19	2.14	1.99	1.91	1.82	1.77	1.74	1.71	1.69	1.67
34	4.13	3.28	2.88	2.65	2.49	2.38	2.29	2.23	2.17	2.12	1.97	1.89	1.8	1.75	1.71	1.69	1.66	1.65
36	4.11	3.26	2.87	2.63	2.48	2.36	2.28	2.21	2.15	2.11	1.95	1.87	1.78	1.73	1.69	1.67	1.64	1.62
38	4.1	3.24	2.85	2.62	2.46	2.35	2.26	2.19	2.14	2.09	1.94	1.85	1.76	1.71	1.68	1.65	1.62	1.61
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	1.92	1.84	1.74	1.69	1.66	1.64	1.61	1.59
42	4.07	3.22	2.83	2.59	2.44	2.32	2.24	2.17	2.11	2.06	1.91	1.83	1.73	1.68	1.65	1.62	1.59	1.57
44	4.06	3.21	2.82	2.58	2.43	2.31	2.23	2.16	2.1	2.05	1.9	1.81	1.72	1.67	1.63	1.61	1.58	1.56
46	4.05	3.2	2.81	2.57	2.42	2.3	2.22	2.15	2.09	2.04	1.89	1.8	1.71	1.65	1.62	1.6	1.57	1.55
48	4.04	3.19	2.8	2.57	2.41	2.29	2.21	2.14	2.08	2.03	1.88	1.79	1.7	1.64	1.61	1.59	1.56	1.54
50	4.03	3.18	2.79	2.56	2.4	2.29	2.2	2.13	2.07	2.03	1.87	1.78	1.69	1.63	1.6	1.58	1.54	1.52
60	4	3.15	2.76	2.53	2.37	2.25	2.17	2.1	2.04	1.99	1.84	1.75	1.65	1.59	1.56	1.53	1.5	1.48
70	3.98	3.13	2.74	2.5	2.35	2.23	2.14	2.07	2.02	1.97	1.81	1.72	1.62	1.57	1.53	1.5	1.47	1.45
80	3.96	3.11	2.72	2.49	2.33	2.21	2.13	2.06	2	1.95	1.79	1.7	1.6	1.54	1.51	1.48	1.45	1.43
90	3.95	3.1	2.71	2.47	2.32	2.2	2.11	2.04	1.99	1.94	1.78	1.69	1.59	1.53	1.49	1.46	1.43	1.41
100	3.94	3.09	2.7	2.46	2.31	2.19	2.1	2.03	1.97	1.93	1.77	1.68	1.57	1.52	1.48	1.45	1.41	1.39
500	3.86	3.01	2.62	2.39	2.23	2.12	2.03	1.96	1.9	1.85	1.69	1.59	1.48	1.42	1.38	1.35	1.3	1.28
∞	3.84	3	2.6	2.37	2.21	2.1	2.01	1.94	1.88	1.83	1.67	1.57	1.46	1.39	1.35	1.32	1.27	1.24

Annexe D

Quelques données

Date	"maxO3"	"T12"	"T15"	"Ne12"	"N12"	"S12"	"E12"	"W12"	"Vx"	"maxO3v"
"19960422"	63.6	13.4	15	7	0	0	3	0	9.35	95.6
"19960429"	89.6	15	15.7	4	3	0	0	0	5.4	100.2
"19960506"	79	7.9	10.1	8	0	0	7	0	19.3	105.6
"19960514"	81.2	13.1	11.7	7	7	0	0	0	12.6	95.2
"19960521"	88	14.1	16	6	0	0	0	6	-20.3	82.8
"19960528"	68.4	16.7	18.1	7	0	3	0	0	-3.69	71.4
"19960605"	139	26.8	28.2	1	0	0	3	0	8.27	90
"19960612"	78.2	18.4	20.7	7	4	0	0	0	4.93	60
"19960619"	113.8	27.2	27.7	6	0	4	0	0	-4.93	125.8
"19960627"	41.8	20.6	19.7	8	0	0	0	1	-3.38	62.6
"19960704"	65	21	21.1	6	0	0	0	7	-23.68	38
"19960711"	73	17.4	22.8	8	0	0	0	2	-6.24	70.8
"19960719"	126.2	26.9	29.5	2	0	0	4	0	14.18	119.8
"19960726"	127.8	25.5	27.8	3	0	0	5	0	13.79	103.6
"19960802"	61.6	19.4	21.5	7	6	0	0	0	-7.39	69.2
"19960810"	63.6	20.8	21.4	7	0	0	0	5	-13.79	48
"19960817"	134.2	29.5	30.6	2	0	3	0	0	1.88	118.6
"19960824"	67.2	21.7	20.3	7	0	0	0	7	-24.82	60
"19960901"	87.8	19.7	21.7	5	0	0	3	0	9.35	74.4
"19960908"	96.8	19	21	6	0	0	8	0	28.36	103.8
"19960915"	89.6	20.7	22.9	1	0	0	4	0	12.47	78.8
"19960923"	66.4	18	18.5	7	0	0	0	2	-5.52	72.2
"19960930"	60	17.4	16.4	8	0	6	0	0	-10.8	53.4
"19970414"	90.8	16.3	18.1	0	0	0	5	0	18	89
"19970422"	104.2	13.6	14.4	1	0	0	1	0	3.55	97.8
"19970429"	70	15.8	16.7	7	7	0	0	0	-12.6	61.4

TABLE D.1 – Quelques données journalières sur Rennes.

Date	"maxO3"	"T12"	"T15"	"Ne12"	"N12"	"S12"	"E12"	"W12"	"Vx"	"maxO3v"
"19970708"	96.2	26	27.3	2	0	0	5	0	16.91	87.4
"19970715"	65.6	23.5	23.7	7	0	0	0	3	-9.35	67.8
"19970722"	109.2	26.3	27.3	4	0	0	5	0	16.91	98.6
"19970730"	86.2	21.8	23.6	6	4	0	0	0	2.5	112
"19970806"	87.4	24.8	26.6	3	0	0	0	2	-7.09	49.8
"19970813"	84	25.2	27.5	3	0	0	0	3	-10.15	131.8
"19970821"	83	24.6	27.9	3	0	0	0	2	-5.52	113.8
"19970828"	59.6	16.8	19	7	0	0	0	8	-27.06	55.8
"19970904"	52	17.1	18.3	8	5	0	0	0	-3.13	65.8
"19970912"	73.8	18	18.3	7	0	5	0	0	-11.57	90.4
"19970919"	129	28.9	30	1	0	0	3	0	8.27	111.4
"19970926"	122.4	23.4	25.4	0	0	0	2	0	5.52	118.6
"19980504"	106.6	13	14.3	3	7	0	0	0	12.6	84
"19980511"	121.8	26	28	2	0	4	0	0	2.5	109.8
"19980518"	116.2	24.9	25.8	2	0	0	5	0	18	142.8
"19980526"	81.4	18.4	16.8	7	0	0	0	4	-14.4	80.8
"19980602"	88.6	18.7	19.6	5	0	0	0	5	-15.59	60.4
"19980609"	63	20.4	16.6	7	0	0	0	8	-22.06	79.8
"19980617"	104	19.6	21.2	6	0	0	0	3	-10.8	84.6
"19980624"	88.4	23.2	23.9	4	0	4	0	0	-7.2	92.6
"19980701"	83.8	19.8	20.3	8	0	0	5	0	17.73	40.2
"19980709"	56.4	18.9	19.3	8	0	0	0	4	-14.4	73.6
"19980716"	50.4	19.7	19.3	7	0	0	0	5	-17.73	59
"19980724"	79.2	21.1	21.9	3	4	0	0	0	9.26	55.2

TABLE D.2 – Quelques données journalières sur Rennes.

Bibliographie

[1] A. Antoniadis, J. Berruyer, and R. Carmona. *Régression non linéaire et applications*. Economica, 1992.

[2] A.C. Atkinson. Two graphical displays for outlying and influential observations in regression. *Biometrika*, 68 :13–20, 1981.

[3] B. Bercu and D. Chafaï. *Modélisation stochastique et simulation*. Dunod, Paris, 2007.

[4] R. D. Cook. Detection of influential observation in linear regression. *Technometrics*, 19 :15–18, 1977.

[5] P.-A. Cornillon and E. Matzner-Løber. *Régression avec R*. Springer, Paris, 2010.

[6] Y. Dodge and V. Rousson. *Analyse de régression appliquée*. Dunod, 2004.

[7] G. H. Golub and C. F. Van Loan. *Matrix computations*. John Hopkins university press, 3rd edition, 1996.

[8] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning - data mining, inference and prediction*. Springer, New-York, 2001.

[9] D. C. Hoaglin and R. E. Welsch. The hat matrix in regression and anova. *The American Statistician*, 32 :17–22, 1978.

[10] P. Huber. *Robust Statistics*. J. Wiley & Sons, New-York, 1981.

[11] F. Husson and J. Pagès. *Statistiques générales pour utilisateurs (2. Exercices et corrigés)*. Presses Universitaires de Rennes, 2005.

[12] E. L. Lehmann and G. Casella. *Theory of point estimation*. Springer, New-York, 1998.

[13] M. Lejeune. *Statistique. La théorie et ses applications*. Springer, Paris, 2004.

[14] D. C. Montgomery, E. A. Peck, and G. Geoffrey Vining. *Introduction to linear regression analysis*. John Wiley, New-York, 3 edition, 2001.

[15] A. Sen and M. Srivastava. *Regression Analysis : Theory, Methods, and Applications*. Springer, 1990.