République Algériennne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Mohamed Khider, Biskra

Faculté des Sciences Exactes et des Sciences de la Nature et de la Vie

Départementde Mathématiques

Mémoire présenté pour obtenir le diplôme de

Master en "**Mathematiques Appliquées**"

Option : **Statistique**

Par

REFRAFI Asma

Titre :

Régression Linéaire Bayésienne

Devant le Jury :

YAHIA Djabrane	**Pr.**	UMKB	Président
BENAMEUR Sana	**M.C.B**	UMKB	Encadreur
TOUBA SONIA	**M.C.B**	UMKB	Examinateur

Soutenu le 28/06/2022

Dédicace

Je dédie cet humble travail à :

À deux personnes qui m'ont donné leur confiance et leur soutien tout au long de mes études, mon père et ma mère.

À mes très chers frères et mes très chères sœurs qui m'ont encouragé sur le long de mon parcours universitaire.

À mon cher fiancé, pour le soutien dont il a fait preuve pendentif toute la durée de ce travail.

À tous mes professeurs et tous ceux qui ont contribué à mon éducation.

À tous ceux que j'aime.

Remerciements

Tout d'abord je tiens à remercier **ALLAH** de m'avoir donné le courage, la patience et la santé pour accomplir ce Modeste travail.

Je remercie mon encadreur **Dr. BENAMEUR SANA**, pour ses efforts et surtout ses conseils précieux et aussi pour sa guidance et sa patience pendant la durée de ce modeste travail.

Je remercie les membres de jury : Le **Pr. YAHIA Djabrane** et le **Dr. TOUBA SONIA** d'avoir accepté d'examiner et d'évaluer ce travail.

J'aimerais remercier toute ma famille dans laquelle j'ai toujours trouvé le support dans toutes mes activités.

Sans oublier d'exprimer nos vifs remerciements et toute notre gratitude à tous nos enseignants qui ont contribués à notre formation pendant cinq ans.

Nos derniers remerciements et ce ne sont pas les moindres, vont à tous ceux qui ont contribué de près ou de loin, directement ou indirectement pour l'aboutissement de ce travail.

Table des matières

Table des figures

Liste des tableaux

Introduction

L'origine du mot régression vient de Sir Francis Galton. En 1885, travaillant sur l'hérédité, il chercha à expliquer la taille des fils en fonction de celle des pères, les résultats obtenus l'ont conduit à considérer sa théorie dite théorie de régression.

Une régression est l'une des méthodes statistiques les plus utilisées dans les sciences appliquées, grâce à son objectif double. Elle permet tout d'abord d'établir la relation entre une variable quantitative a expliqué (ou dépendante) à une ou plusieurs autres variables quantitatives explicatives (ou indépendantes) sous la forme d'un modèle et aussi d'effectuer des prévisions de l'une des variables en fonction de l'autre. Elle est dite linéaire si elle impose une forme fonctionnelle linéaire dans les paramètres du modèle.

Les paramètres de ce modèle sont estimés par deux méthodes différentes : L'approche fréquentiste où chaque paramètre est considéré comme un point fixe inconnu, et une approche bayésienne où les paramètres eux-mêmes sont considérés comme une variable aléatoire.

L'objectif de ce mémoire est de présenter les différentes méthodes de régression linéaire en mettant un accent particulier sur la régression linéaire bayésienne.

Le présent travail est réparti en trois chapitres.

Chapitre 1 (Régression Linéaire Simple) : Ce chapitre consiste à présenter les principales notions et propriétés de la régression linéaire simple.

Chapitre 2 (Régression Linéaire Multiple) : Dans ce chapitre le modèle de régression

linéaire multiple est abordé. Il constitue une généralisation naturelle de la régression linéaire simple, dans ce cas le nombre des variables explicatives est supérieure ou égal à deux.

Chapitre 3 (Régression Linéaire Bayésienne) : Ce dernier chapitre est la partie la plus importante dans ce travail où le modèle bayésien a été présenté avec les estimations bayésienne des deux modèles mentionnés précédemment. On termine ce chapitre avec un exemple illustratif sur un modèle de régression linéaire simple, tant par l'approche fréquentiste et l'approche bayésienne, à l'aide du logiciel R.

On arrive finalement à une conclusion suivi de la liste des références bibliographiques.

Chapitre 1

Régression Linéaire Simple

Ce chapitre est une introduction à la modélisation linéaire par le modèle le plus élémentaire, la régression linéaire simple. Le principe de ce modèle est de supposer qu'une variable Y est expliquée, modélisée par une fonction affine d'une seule variable explicative X. Après avoir explicité les hypothèses nécessaires et les termes du modèle, on discute les méthodes d'estimation des paramètres, les lois des estimateurs, les intervalles de confiance puis la signification des tests d'hypothèse et la qualité d'ajustement du modèle dont le but est la prévision. Pour plus de détaille consulter les livres suivants : Cornillon [7] et Bourbonnais [3].

1.1 Modèle de régression linéaire simple

1.1.1 Modèle

Un modèle de régression linéaire simple est défini par une équation de la forme :

$$y_i = b_0 + b_1 x_i + \varepsilon_i \ ; \ i = 1, \ldots, n,$$

où :

y_i : la $i^{ème}$ observation de la variable aléatoire à expliquer Y,

x_i : la $i^{ème}$ observation de la variable explicative X,

ε_i : L'erreur (ou bruit) aléatoire du modèle,

b_0 et b_1 : sont des constantes inconnues appelées paramètres du modèle,

n : la taille de l'échantillon.

1.1.2 Ecriture matricielle

Notons que le modèle de régression linéaire peut encore s'écrire sous forme matricielle, comme suit :

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

telle que :

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \; X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \; \beta = \begin{pmatrix} b_0 \\ b_1 \end{pmatrix} \text{ et } \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

1.1.3 Hypothèses sur le modèle

Les hypothèses relatives à ce modèle sont les suivantes :

$(\mathcal{H}_1)$ Les erreurs ε_i sont centrées, ont la même variances, et non corrélées entre elles, c-à-d :

$$\forall i = 1, ..., n : E(\varepsilon_i) = 0, \; E(\varepsilon_i^2) = \sigma_\varepsilon^2 < \infty,$$
$$\text{et } \forall (i, j) \text{ tel que } i \neq j : cov(\varepsilon_i, \varepsilon_j) = 0.$$

$(\mathcal{H}_2)$ L'erreur est indépendante de X :

$$cov(\varepsilon, X) = 0.$$

1.2 Estimation des paramètres

Pour estimer les paramètres du modèle, on peut utiliser la méthode des moindres carrés ordinaires (MCO) qui ne nécessite pas d'hypothèse supplémentaire sur la distribution de ε_i, contrairement on peut utiliser la méthode de Maximum de Vraisemblance (MV) qui est fondée sur la normalité de ε_i.

1.2.1 Estimation par la méthode des MCO

Définition 1.2.1 (Estimateurs des MCO)

On appelle estimateurs des MCO de b_0 et b_1, les estimateurs $\hat{b}_0$ et $\hat{b}_1$ obtenus par minimisation de la quantité :

$$\Psi(b_0, b_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2 .$$

Les estimateurs des MCO peuvent également s'écrire sous la forme suivante :

$$\left(\hat{b}_0, \hat{b}_1\right) = \underset{(b_0,b_1)\in\mathbb{R}\times\mathbb{R}}{\arg\min}\ \Psi(b_0, b_1) .$$

Le minimum est atteint pour

$$\left\{ \begin{array}{l|l} \dfrac{\partial \Psi(b_0, b_1)}{\partial b_0} = 0, & b_0 = \hat{b}_0 \\ \dfrac{\partial \Psi(b_0, b_1)}{\partial b_1} = 0, & b_1 = \hat{b}_1 \end{array} \right. .$$

On en déduit après quelques calculs :

$$\left\{ \begin{array}{l} \hat{b}_0 = \bar{y} - \hat{b}_1 \bar{x}, \\ \hat{b}_1 = \dfrac{S_{xy}}{S_x}, \end{array} \right. \tag{1.1}$$

où

$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ et $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ les moyennes empiriques des x_i et des y_i (respectivement),

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

$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}.$

Propriétés des estimateurs

Théorème 1.2.1 (Biais des estimateurs)

Sous les hypothèses $(\mathcal{H}_1)$ et $(\mathcal{H}_2)$. Les estimateurs $\hat{b}_0$ et $\hat{b}_1$ sont sans biais de b_0 et b_1, c-à-d

$$E(\hat{b}_0) = b_0 \text{ et } E(\hat{b}_1) = b_1.$$

Théorème 1.2.2 (Variance et covariance)

Sous les hypothèses $(\mathcal{H}_1)$ et $(\mathcal{H}_2)$. Les variances et la covariance des estimateurs sont

$$Var\left(\hat{b}_0\right) = \frac{\sigma_\varepsilon^2 \sum_{i=1}^{n} x_i^2}{nS_x},\ Var\left(\hat{b}_1\right) = \frac{\sigma_\varepsilon^2}{S_x} \text{ et } Cov\left(\hat{b}_0, \hat{b}_1\right) = \frac{-\sigma_\varepsilon^2 \bar{x}}{S_x}.$$

La matrice de variance covariance de $\hat{b}_0$ et $\hat{b}_1$ s'écrire comme suit :

$$\Gamma\left(\hat{b}_0, \hat{b}_1\right) = \frac{\sigma_\varepsilon^2}{S_x} \begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n} x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix},$$

qui est estimée en remplaçant σ_ε^2 par son estimateur $\hat{\sigma}_\varepsilon^2$.

Théorème 1.2.3 (Biais de l'estimateur de σ_ε^2)

L'estimateur sans biais de la variance résiduelle (variance des erreurs) σ_ε^2 est :

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

avec les résidus sont définis par $\hat{\varepsilon}_i = y_i - \hat{y}_i$ *où* $\hat{y}_i$ *est la valeur ajustée de* y_i *par le modèle, c'est-à-dire*

$$\hat{y}_i = \hat{b}_0 + \hat{b}_1 x_i.$$

1.2.2 Estimation par la méthode de MV

L'estimation des paramètres b_0, b_1 et σ_ε^2 est obtenue encore par maximisant la vraisemblance par rapport à ces paramètres, sous l'hypothèse supplémentaire de normalité des erreurs :

1) $\forall i = \overline{1,n}$: $\varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon^2)$.

2) $\forall i = \overline{1,n}$: ε_i sont mutuellement indépendants.

La fonction de vraisemblance de l'échantillon $(y_1, y_2, \ldots, y_n)$ s'écrit

$$L(b_0, b_1, \sigma_\varepsilon^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}} \exp\left\{ -\frac{1}{2\sigma_\varepsilon^2} \left[y_i - (b_0 + b_1 x_i) \right]^2 \right\}.$$

Ce qui donne pour la fonction log-vraisemblance

$$\log L(b_0, b_1, \sigma_\varepsilon^2) = -\frac{n}{2} \ln\left(2\pi\sigma_\varepsilon^2\right) - \frac{1}{2\sigma_\varepsilon^2} \sum_{i=1}^{n} \left(y_i - b_0 - b_1 x_i \right)^2,$$

puis on maximise cette quantité, ce qui revient à annuler ses dérivées premières par rapport les paramètres $b_0, b_1, \sigma_\varepsilon^2$.

Les estimateurs de MV $\hat{b}_0$ et $\hat{b}_1$ sont égaux aux estimateurs de MCO et ils ont les mêmes propriétés (voir théorème 1.2.1 et 1.2.2)

$$\hat{b}_0 = \bar{y} - \hat{b}_1 \bar{x},$$
$$\hat{b}_1 = \frac{S_{xy}}{S_x}.$$

L'estimateur du MV de σ_ε^2 est différent de l'estimateur qui vu précédemment et vaut :

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

L'estimateur du MV de σ_ε^2 est donc biaisé. En effet, $E\left(\hat{\sigma}_{mv}^2\right) = \dfrac{n-2}{n}\sigma_\varepsilon^2$, mais ce biais est d'autant plus négligeable quand le nombre d'observations est grand.

1.3 Interprétation géométrique

1.3.1 Représentation des individus

La représentation graphique des observations (x_i, y_i) dans le plan $\mathbb{R}^2$ est appelé nuage de points. Ce dernier peut être résumé par une droite qui s'appelle la droite de régression linéaire simple. Cette droite minimise la somme des carrés des distances verticales des points du nuage à la droite ajustée, où $\hat{b}_0$ correspond à l'ordonnée à l'origine et $\hat{b}_1$ représente la pente de la droite ajustée d'équation $\hat{y}_i = \hat{b}_0 + \hat{b}_1 x_i$.

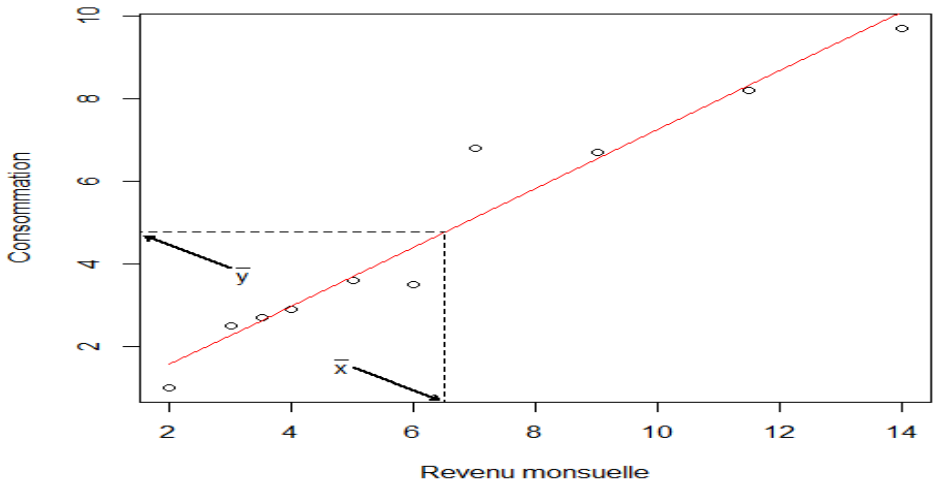

FIG. 1.1 – Ajustement du nuage de point par une droit de régression

1.3.2 Représentation des variables

On peut envisager ce problème d'une autre manière. on a deux vecteurs à la disposition : le vecteur $X'' = (x_1, \ldots, x_n)^t$ des n observations pour la variable explicative et le vecteur $Y = (y_1, \ldots, y_n)^t$ des n observations pour la variable expliquée. Ces deux vecteurs appartiennent au même espace $\mathbb{R}^n$: l'espace des variables.

On peut donc représenter les données dans l'espace des variables, soit I est le vecteur unitaire de $\mathbb{R}^n$. Les deux vecteurs I et X'' génèrent un sous-espace de $\mathbb{R}^n$ de dimension 2, on suppose que I et X'' ne sont pas colinéaires (il existe au moins deux points d'abscisses différentes) mais ces vecteurs ne sont pas nécessairement orthogonaux. Ces vecteurs sont orthogonaux lorsque la moyenne des observations $x_1, \ldots, x_n$ vaut zéro.

Autrement dit, la régression linéaire peut être vue comme la projection orthogonale du vecteur Y sur le sous espace de $\mathbb{R}^n$ engendré par et I et X''.

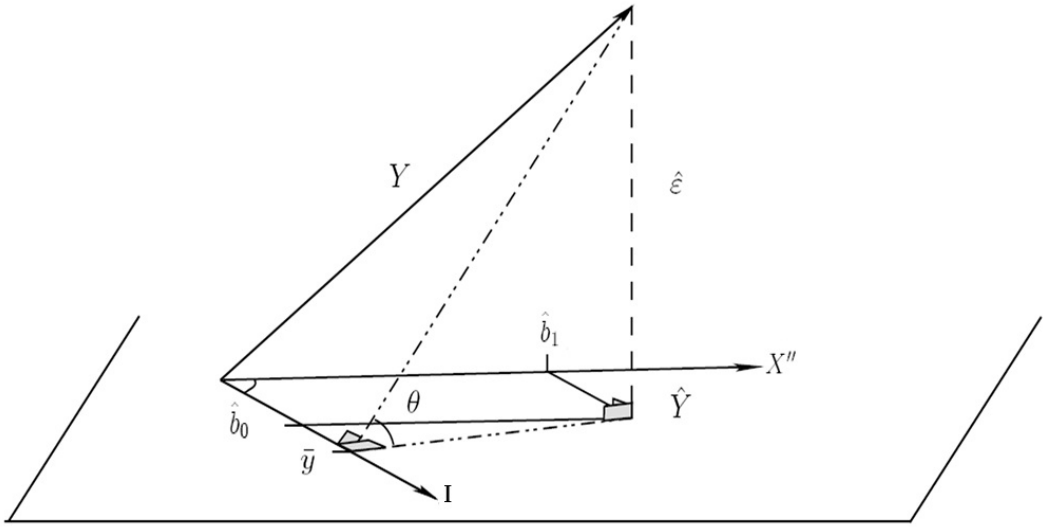

FIG. 1.2 – Représentation des variables

1.4 Lois des Estimateurs

Pour simplifier, on considère les notations suivantes :

$$\sigma^2_{\hat{b}_0} = Var\left(\hat{b}_0\right), \quad \hat{\sigma}^2_{\hat{b}_0} = \frac{\hat{\sigma}^2_\varepsilon \sum_{i=1}^{n} x_i^2}{nS_x},$$
$$\sigma^2_{\hat{b}_1} = Var\left(\hat{b}_1\right), \quad \hat{\sigma}^2_{\hat{b}_1} = \frac{\hat{\sigma}^2_\varepsilon}{S_x}.$$

Théorème 1.4.1 (Lois des estimateurs si σ^2_ε est connue)

Sous l'hypothèse de normalité des résidus avec σ^2_ε est connue, les lois des estimateurs sont :

(i) $\hat{b}_j \sim \mathcal{N}\left(b_j, \sigma^2_{\hat{b}_j}\right)$, $j = 0, 1$.

(ii) *La loi jointe des estimateurs* $\left(\hat{b}_0, \hat{b}_1\right)$ *:* $\begin{pmatrix} \hat{b}_0 \\ \hat{b}_1 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} b_0 \\ b_1 \end{pmatrix}, \Gamma\left(\hat{b}_0, \hat{b}_1\right)\right)$.

Théorème 1.4.2 (Lois des estimateurs si σ^2_ε est inconnue)

Si σ^2_ε est inconnue, dans ce cas σ^2_ε est estimé par S^2, nous avons :

(i) $\dfrac{\hat{b}_j - b_j}{\hat{\sigma}^2_{\hat{b}_j}} \sim \mathcal{T}(n-2)$, $j = 0, 1$ *où* $\mathcal{T}(n-2)$ *est une loi de Student à* $(n-2)$ *ddl.*

(ii) $\dfrac{1}{2}\begin{pmatrix} \hat{b}_0 - b_0 \\ \hat{b}_1 - b_1 \end{pmatrix}^t \Gamma^{-1}\left(\hat{b}_0, \hat{b}_1\right) \begin{pmatrix} \hat{b}_0 - b_0 \\ \hat{b}_1 - b_1 \end{pmatrix} \sim \mathcal{F}(2, n-2)$, *où* $\mathcal{F}(2, n-2)$ *est la loi de Fisher à 2 et* $(n-2)$ *ddl.*

(iii) $\dfrac{(n-2)}{\sigma^2_\varepsilon}\hat{\sigma}^2_\varepsilon \sim \mathcal{X}^2_{n-2}$, *où* $\mathcal{X}^2_{n-2}$ *est la loi de Khi-deux (ou encore noté khi2) à* $(n-2)$ *degrés de liberté* (ddl).

(iv) $\left(\hat{b}_0, \hat{b}_1\right)$ *et* $\hat{\sigma}^2_\varepsilon$ *sont indépendants.*

Remarque 1.4.1 *Ces théorèmes nous permettent de donner des intervalles de confiance (IC), des régions de confiance (RC) des paramètres inconnus et de tester la signification des paramètres et la validation du modèle.*

1.5 Intervalles de confiance

Soit $\alpha \in [0,1]$, on cherche les intervalles de confiance des paramètres b_0 et b_1 au niveau de confiance $(1-\alpha)$.

- D'après le théorème 1.4.2, on a :

$$P\left(\frac{\left|\hat{b}_j - b_j\right|}{\hat{\sigma}^2_{\hat{b}_j}} < t\right) = 1-\alpha,\ j=0,1,$$

on obtient les IC des paramètres b_0 et b_1

$$b_j \in \left]\hat{b}_j - t\hat{\sigma}_{\hat{b}_j}, \hat{b}_j + t\hat{\sigma}_{\hat{b}_j}\right[,$$

où t est le fractile d'ordre $(1-\alpha/2)$ de la loi de Student à $(n-2)$ *ddl*.

- De même, pour l'IC de σ^2_ε

$$P\left(t_2 < \frac{(n-2)}{\sigma^2_\varepsilon}\hat{\sigma}^2_\varepsilon < t_1\right) = 1-\alpha,$$

donc

$$\sigma^2_\varepsilon \in \left]\frac{(n-2)\,\hat{\sigma}^2_\varepsilon}{t_2}, \frac{(n-2)\,\hat{\sigma}^2_\varepsilon}{t_1}\right[,$$

où t_1 (respt t_2) est le fractile d'ordre $(1-\alpha/2)$ (respt $\alpha/2$) de la loi de $\mathcal{X}^2_{n-2}$.

- Finalement, la région de confiance de $\begin{pmatrix}\hat{b}_0\\ \hat{b}_1\end{pmatrix}$ est

$$\frac{1}{2\hat{\sigma}^2_\varepsilon}\left\{n\left(\hat{b}_0 - b_0\right)^2 + 2n\bar{x}\left(\hat{b}_0 - b_0\right)\left(\hat{b}_1 - b_1\right) + \left(\hat{b}_1 - b_1\right)^2\sum_{i=1}^{n}x_i^2\right\} < f,$$

où f est le fractile d'ordre $(1-\alpha)$ d'une loi de Fisher $\mathcal{F}(2, n-2)$.

1.6 Qualité d'ajustement

Une fois le modèle est établi, il faut juger sa qualité d'ajustement et sa fiabilité, on utilise l'équation de l'analyse de la variance, c-à-d cherchons tout d'abord à décomposer la variance des y_i autour de leur moyenne en une somme de deux autre variances.

$$SCT = SCR + SCE.$$

Cette équation s'écrit

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 ,$$

où

SCT : somme des carrés totale (variation totale des y_i),

SCR : somme des carrés des résidus (variation des résidus $\hat{\varepsilon}_i$),

SCE : somme des carrés expliqués (variation expliquée).

Un indicateur de la qualité d'ajustement est le coefficient de détermination R^2 qui exprime le rapport entre la variation expliquée et la variation totale, dont

$$R^2 = \frac{SCE}{SCT} = \frac{S_{xy}^2}{S_x S_y} \leq 1.$$

Dans le cas où R^2 est proche de 1 alors l'ajustement est meilleure, c-à-d la variation expliquée est proche de la variation totale alors la variation résiduelle est proche de 0.

1.7 Tests d'hypothèses

1.7.1 Test de signification des paramètres

On définit les hypothèses du test de Student (test bilatéral de significativité)

$$\begin{cases} H_0 : b_j = 0, \\ H_1 : b_j \neq 0, \end{cases} j = 0, 1.$$

On accepte l'hypothèse (H_0) au niveau de signification α, $\alpha \in]0, 1[$, on dit dans ce cas que le paramètre b_j est significativement nulle si

$$T_{obs} = \frac{\left|\hat{b}_j\right|}{\hat{\sigma}_{\hat{b}_j}} \leq t,$$

où la valeur critique t est le fractile d'ordre $(1 - \alpha/2)$ de la loi de Student à $(n - 2)$ ddl (lue à partir la table de Student).

1.7.2 Test de la signification globale du modèle

Soit les hypothèses du test :

$$\begin{cases} H_0 : b_0 = 0 \text{ et } b_1 = 0, \\ H_1 : b_0 \neq 0 \text{ et } b_1 \neq 0. \end{cases}$$

On accepte l'hypothèse (H_0) et on dit que le modèle est non valide, si

$$\frac{1}{2\hat{\sigma}_\varepsilon^2} \left\{ n \left(\hat{b}_0 - b_0\right)^2 + 2n\bar{x} \left(\hat{b}_0 - b_0\right) \left(\hat{b}_1 - b_1\right) + \left(\hat{b}_1 - b_1\right)^2 \sum_{i=1}^{n} x_i^2 \right\} < f,$$

où f est le fractile d'ordre $(1 - \alpha)$ de la loi de Fisher $\mathcal{F}(2, n - 2)$ (lue à partir la table de Fisher).

On construisant le tableau d'analyse de la variance comme suit (la table d'ANOVA)

Source de variation	**ddl**	**Somme des carrés**	**Moyenne des carrés**	**F**
Expliquée	1	SCE	$MCE = SCE/1$	$\frac{MCE}{MCR}$
Résiduelle	$n-2$	SCR	$MCR = SCR/(n-2)$	
Totale	$n-1$	SCT		

On accepte H_0 si :

$$F = \frac{MCE}{MCR} \leq f.$$

où f est le fractile d'ordre $(1-\alpha)$ de la loi de Fisher $\mathcal{F}(1, n-2)$

1.8 Prévision

Un des buts de la régression est de proposer des prévisions pour la variable à expliquer Y. Soit x_p une nouvelle valeur de la variable explicative X, peut-on prévoir la valeur y_p?. Le modèle indique que

$$y_p = b_0 + b_1 x_p + \varepsilon_p.$$

On estime y_p par un estimateur sans bais, donné par :

$$\hat{y}_p = \hat{b}_0 + \hat{b}_1 x_p.$$

De plus, l'erreur de la prévision est :

$$\hat{y}_p - y_p = \left(\hat{b}_0 - b_0\right) + \left(\hat{b}_1 - b_1\right) x_p - \varepsilon_p.$$

Cette quantité est la somme des variables aléatoires normales, alors elle suit la loi normale centrée de variance donnée dans la proposition suivante.

Proposition 1.8.1 (Variance de l'erreur de la prévision) .

- *La variance de la valeur prévue $\hat{y}_p$ vaut :*

$$Var(\hat{y}_p) = \sigma^2 \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right).$$

- *La variance de l'erreur de la prévision est :*

$$\sigma_e^2 = Var(\hat{y}_p - y_p) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right).$$

- *On en déduit l'intervalle de prédiction pour y_p au niveau de confiance $(1 - \alpha)$:*

$$y_p \in \left[\hat{y}_p - t\hat{\sigma}_e^2, \hat{y}_p + t\hat{\sigma}_e^2 \right],$$

où t est le fractile d'ordre $(1 - \alpha/2)$ de la loi de Student $\mathcal{T}(n-2)$.

Chapitre 2

Régression Linéaire Multiple

Le modèle de régression linéaire multiple est l'outil statistique le plus habituellement mis en œuvre pour l'étude des données multidimensionnelles. Dans ce chapitre nous considérons que la variable aléatoire à expliquer Y est modélisée par plusieurs variables explicatives X_j $(j = 1, ..., p)$, il constitue la généralisation naturelle de la régression linéaire simple. Pour plus de détaille consulter les livres suivants : Cornillon [7] et Bourbonnais [3].

2.1 Modèle de régression linéaire multiple

Le modèle de régression linéaire multiple se base sur l'écriture de chaque observation y_i, il est définit par la formule suivante :

$$y_i = b_0 + \sum_{j=1}^{p} b_j x_{i,j} + \varepsilon,\ i = 1, \ldots, n,\ n > p$$

où :

- $x_{i,j}$ est le $i^{éme}$ individu pour la $j^{éme}$ variable explicative (non aléatoire),
- Les erreurs ε_i sont des variables aléatoires inconnues,
- Les paramètres du modèle b_j $(j = 0, \ldots, p)$ sont des constantes inconnus.

Notons que le modèle de régression linéaire multiple peut encore s'écrire sous forme matricielle, comme suit :

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

avec :

- Y est un vecteur aléatoire de dimension n,
- Les données sont rangées dans une matrice X de dimension $(n, p+1) \in \mathbb{R}^n \times \mathbb{R}^{p+1}$ dont la première colonne contient le vecteur unitaire indique la constante b_0 dans l'équation,
- β est le vecteur des paramètres du modèle de dimension $p+1$,
- ε est le vecteur des erreurs de dimension n.

D'où :

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & \dots & x_{1,p} \\ 1 & x_{2,1} & \dots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \dots & x_{n,p} \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Les hypothèses relatives à ce modèle sont :

$(\mathcal{H}_1)$ La matrice X est de plein rang, c-à-d

$$rang(X) = p + 1.$$

$(\mathcal{H}_2)$ Les erreurs sont centrées, ont la même variance, et non corrélées entre elles :

$$E(\varepsilon) = 0_n,\ Var\,(\varepsilon) = \sigma_\varepsilon^2 I_n,$$

avec I_n la matrice identité d'ordre n et $0_n = (0, 0, \dots, 0)^t \in \mathbb{R}^n$.

$(\mathcal{H}_3)$ Les erreurs ε_i sont indépendantes des X_j $\left(j = \overline{1, p}\right)$.

2.2 Estimation des paramètres

2.2.1 Estimation par méthode de MCO

Les paramètres β et σ_{ε}^2 sont inconnus et doivent être estimés à partir de la connaissance des X_j. On peut utiliser la méthode des moindres carrés qui cherche une estimation des paramètres $\hat{\beta}$ en minimisant la quantité

$$\hat{\beta} = \underset{\beta \in R^{p+1}}{\arg\min} S(\beta),$$

telle que :

$$S(\beta) = S(b_0, ..., b_p) = \sum_{i=1}^{n} \varepsilon_i^2 = \|Y - X\beta\|^2 = (Y - X\beta)^t (Y - X\beta).$$

Le minimum de cette somme est atteint lorsque toutes les dérivées partielles de $S(\beta)$ par rapport aux différents b_j s'annulent, c-à-d

$$\frac{\partial S(\beta)}{\partial \beta} = -2X^tY + 2X^tX\hat{\beta} = 0,$$

d'où l'expression de l'estimateur est

$$\hat{\beta} = \left(X^tX\right)^{-1} X^tY,$$

où X^t signifie la matrice X transposée et X^{-1} la matrice inverse.

Remarque 2.2.1

1) La matrice X^tX est carrée d'ordre $p+1$, symétrique et inversible car X est de rang $p+1$.

2) X^tX est définie positive.

3) La valeur ajustée de Y est $\hat{Y} = X\hat{\beta}$.

4) Les résidus sont définis par la relation suivante :

$$\begin{aligned}\hat{\varepsilon} &= Y - \hat{Y} \\ &= Y - X\hat{\beta} \\ &= Y - X\left(X^tX\right)^{-1}X^tY \\ &= \left[I_n - X\left(X^tX\right)^{-1}X^t\right]Y \\ &= \left(I_n - H\right)Y = PY,\end{aligned}$$

telle que les matrices $H = X\left(X^tX\right)^{-1}X^t$ et $P = I_n - H$ sont idempotentes, symétriques et du rang $(p+1)$ et $(n-p-1)$ respectivement.

Propriétés des estimateurs

Proposition 2.2.1 (Biais et matrice de variance covariance)

Sous les hypothèses $(\mathcal{H}_1)$ et $(\mathcal{H}_2)$, l'estimateur $\hat{\beta}$ est un estimateur sans biais de β, c-à-d

$$E\left(\hat{\beta}\right) = \beta,$$

et sa matrice de variance-covariance vaut

$$Var\left(\hat{\beta}\right) = \sigma_\varepsilon^2\left(X^tX\right)^{-1}.$$

Preuve.

On trouve la preuve dans [3] page 51-52. ■

Proposition 2.2.2 (Propriétés de $\hat{\varepsilon}$ et $\hat{Y}$)

Sous les hypothèses du modèle $(\mathcal{H}_1)$ et $(\mathcal{H}_2)$, on a :

1. $E\left(\hat{\varepsilon}\right) = 0$,

2. $Var\left(\hat{\varepsilon}\right) = \sigma_\varepsilon^2 P$,

3. $E\left(\hat{Y}\right) = X\beta$,

4. $Var\left(\hat{Y}\right) = \sigma_\varepsilon^2 H$,

5. $Cov\left(\hat{\beta}, \hat{\varepsilon}\right) = 0$.

Proposition 2.2.3 (Biais de $\hat{\sigma}_\varepsilon^2$)

L'estimateur sans biais de la variance résiduelle σ_ε^2 est donnée par :

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

2.2.2 Estimation par la méthode de MV

La méthode du MV consiste à estimer les paramètres inconnus β et σ_ε^2, sous l'hypothèse d'indépendance et de normalité des erreurs

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

où $\mathcal{N}_n$ désigne la loi normale multidimensionnelle (de dimension n).

Cette méthode est basée sur la maximisation de la fonction de vraisemblance, définie comme suit :

$$L\left(\beta, \sigma_\varepsilon^2\right) = \left(\frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}}\right)^n \exp\left[-\frac{1}{2\sigma_\varepsilon^2}\left(Y - X\beta\right)^t\left(Y - X\beta\right)\right],$$

ce qui revient à maximiser la fonction log-vraisemblance, donnée par :

$$\log L\left(\beta, \sigma_\varepsilon^2\right) = -\frac{n}{2}\ln\left(2\pi\sigma_\varepsilon^2\right) + \frac{1}{2\sigma_\varepsilon^2}\left(Y - X\beta\right)^t\left(Y - X\beta\right).$$

Alors les estimateurs $\hat{\beta}, \hat{\sigma}_\varepsilon^2$ du MV des coefficients $\beta, \sigma_\varepsilon^2$ sont respectivement

$$\hat{\beta} = \left(X^t X\right)^{-1} X^t Y,$$
$$\hat{\sigma}_\varepsilon^2 = \frac{1}{n}\hat{\varepsilon}^t\hat{\varepsilon}.$$

Remarque 2.2.2 *On remarque que les deux méthodes MCO et MV donnent les mêmes estimateurs pour β, alors que les estimateurs de σ_ε^2 sont différents.*

2.3 Interprétation géométrique

Dans la régression linéaire simple, on a vu une seule variable explicative X' mais dans le cas de la régression linéaire multiple, on avait plusieurs variables $X'_1, \ldots, X'_p$, alors le plan de projection (X', I) serait remplacé par l'hyperplan formé par les vecteurs $X'_1, \ldots . X'_p$, I. Régresser Y sur les p variables explicatives consisterait à projeter orthogonalement Y sur l'hyperplan déterminé par $X'_1, \ldots . X'_p$, I.

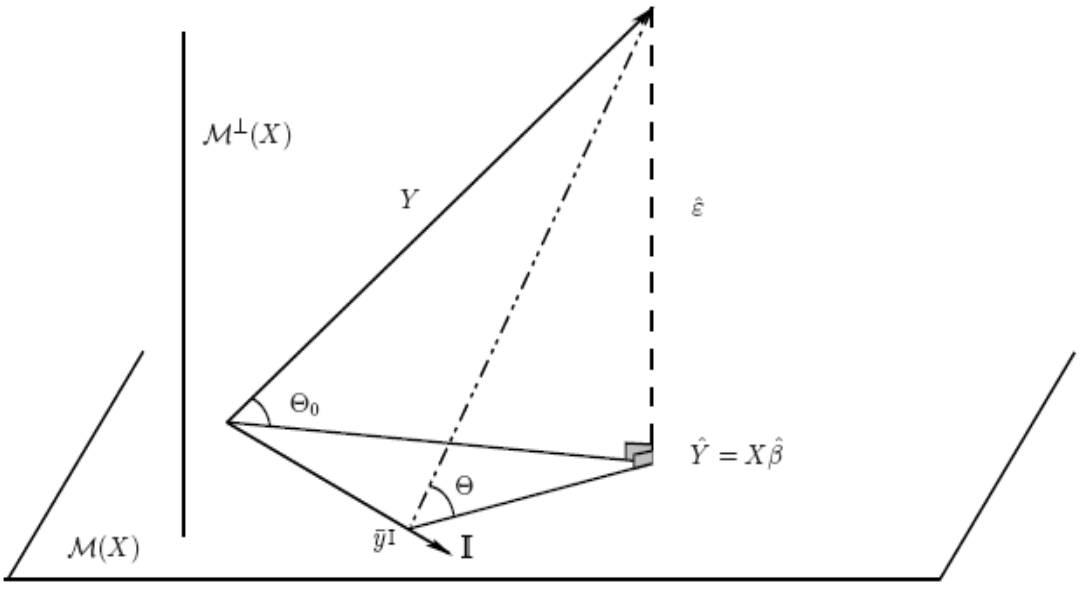

FIG. 2.1 – Représentation des variables et interprétation géométrique.

P est la matrice de projection orthogonale sur $\mathcal{M}^\perp(X)$ (l'espace orthogonal à $\mathcal{M}(X)$), et H est la matrice de projection orthogonale sur $\mathcal{M}(X)$ (le sous espace de $\mathbb{R}^n$ engendré

par les $(p+1)$ vecteurs de la matrice X).

2.4 Qualité d'ajustement

De même que la régression linéaire simple, on a une triangulaire généralisée :

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

La qualité d'ajustement peut être déterminée par le coefficient de détermination R^2, défini par :

$$R^2=\frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}=\frac{SCE}{SCT}=1-\frac{SCR}{SCT}.$$

Ce R^2 est un coefficient réel toujours compris entre 0 et 1.

Puisque R^2 dépend fortement de p, on ne peut pas être utilisé pour comparer la qualité de deux modèles de la régression linéaire multiple qui diffèrent quant aux nombre de variables explicatives, alors on lui préfère d'utiliser le coefficient de détermination ajusté $\overline{R^2}$, qui définit par :

$$\overline{R^2}=1-\frac{\frac{1}{n-p-1}SCR}{\frac{1}{n-1}SCT}=1-\frac{n-1}{n-p-1}\left(1-R^2\right).$$

2.5 Lois des estimateurs

Notons au préalable que, pour ce qui nous concerne, la gaussianité des erreurs :

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

donc Y suit la loi normale dans $\mathbb{R}^n$

$$Y \sim \mathcal{N}_n \left(X\beta, \sigma_\varepsilon^2 I_n \right).$$

Théorème 2.5.1

Sous les hypothèses précédentes, nous avons

i) $\hat{\beta} \sim \mathcal{N}_{p+1} \left(\beta, \sigma_\varepsilon^2 \left(X^t X \right)^{-1} \right)$.

ii) $(n-p-1) \dfrac{\hat{\sigma}_\varepsilon^2}{\sigma_\varepsilon^2} \sim \mathcal{X}^2_{n-p-1}$.

iii) $\hat{\beta}$ *et* $\hat{\sigma}_\varepsilon^2$ *sont deux variable aléatoire indépendantes.*

Proposition 2.5.1

i) *Si* σ_ε^2 *est connue. Notons* V_j *est le* $j^{ème}$ *terme diagonal de la matrice* $\left(X^t X \right)^{-1}$*, alors*

$$\frac{\hat{b}_j - b_j}{\sigma_\varepsilon \sqrt{V_j}} \sim \mathcal{N}(0,1), \ j = \overline{0,p}.$$

ii) *Si* σ_ε^2 *est inconnue. Pour* $j = \overline{0,p}$

$$\frac{\hat{b}_j - b_j}{\hat{\sigma}_\varepsilon \sqrt{V_j}} \sim \mathcal{T}_{n-p-1},$$

où $\mathcal{T}_{n-p-1}$ *est une loi de Student à* $(n-p-1)$ *ddl.*

iii) *Si* σ_ε^2 *est inconnue, alors :*

$$\frac{1}{(p+1)\,\hat{\sigma}_\varepsilon^2} \left(R\left(\hat{\beta} - \beta \right) \right)^t \left(R \left(X^t X \right)^{-1} R^t \right)^{-1} \left(R\left(\hat{\beta} - \beta \right) \right) \sim \mathcal{F}(p+1, n-p-1),$$

où R *est la matrice de taille* $p+1 \times p$ *dont tous les éléments sont nuls sauf les* R_{ij} *qui valent* 1*, et* $\mathcal{F}(p+1, n-p-1)$ *est une loi de Fisher à* $(p+1)$ *et* $(n-p-1)$ *ddl.*

2.6 Intervalles et régions de Confiance

D'après la proposition 2.5.1, on a :

$$P\left(\frac{\left|\hat{b}_j - b_j\right|}{\hat{\sigma}_\varepsilon\sqrt{V_j}} < t\right) = 1-\alpha.$$

L'intervalle de confiance au niveau $(1-\alpha)$ pour chaque coefficient b_j du modèle est :

$$b_j \in \left]\hat{b}_j - t\hat{\sigma}_\varepsilon\sqrt{V_j}, \hat{b}_j + t\hat{\sigma}_\varepsilon\sqrt{V_j}\right[,$$

où t étant le fractile d'ordre $(1-\alpha/2)$ de la loi de Student à $(n-p-1)$ *ddl*.

De même, pour l'intervalle de confiance de σ_ε^2

$$P\left(t_2 < \frac{(n-p)}{\sigma_\varepsilon^2}\hat{\sigma}_\varepsilon^2 < t_1\right) = 1-\alpha,$$

donc

$$\sigma_\varepsilon^2 \in \left]\frac{(n-p)\,\hat{\sigma}_\varepsilon^2}{t_1}, \frac{(n-p)\,\hat{\sigma}_\varepsilon^2}{t_2}\right[,$$

où t_1 (respt t_2) est le fractile d'ordre $(1-\alpha/2)$ (respt $\alpha/2$) de la loi de $\mathcal{X}^2_{n-2}$.

Finalement, la région de confiance au niveau $(1-\alpha)$ pour $p+1$ $(p+1 \leq p)$ paramètres b_j notés $(b_{j1}, \ldots, b_{jq})$ est

$$RC = \left\{R\beta \in \mathbb{R}^{p+1}, \frac{1}{(p+1)\,\hat{\sigma}_\varepsilon^2}\left(R\left(\hat{\beta}-\beta\right)\right)^t\left(R\left(X^tX\right)^{-1}R^t\right)^{-1}\left(R\left(\hat{\beta}-\beta\right)\right) < f\right\},$$

où f est le fractile d'ordre $(1-\alpha)$ d'une loi de Fisher $\mathcal{F}(p+1, n-p)$.

2.7 Tests d'hypothèses

2.7.1 Test de signification des paramètres

Le test de Student permet de tester si un coefficient donné a une influence sur Y, les hypothèses à tester sont :

$$\begin{cases} H_0 : b_j = 0, \\ H_1 : b_j \neq 0, \end{cases} j = \overline{0, p}.$$

On accepte H_0 au niveau de signification $\alpha \in]0, 1[$, et on dit que le paramètre b_j est non significatif si :

$$\frac{\left|\hat{b}_j - b_j\right|}{\hat{\sigma}_\varepsilon \sqrt{V_j}} \leq t,\ \alpha \in]0, 1[\,,$$

où t le fractile d'ordre $(1 - \alpha/2)$ de la loi de Student à $(n - p - 1)$ ddl.

2.7.2 Test de la signification globale du modèle

On test l'hypothèse globale

$$\begin{cases} H_0 : \beta = \beta_0, \\ H_1 : \beta \neq \beta_0, \end{cases} \beta_0 \in \mathbb{R}^{p+1}.$$

La statistique de test est :

$$Q\left(\hat{\beta}, \hat{\sigma}_\varepsilon^2\right) = \frac{\left(\hat{\beta} - \beta\right)^t (X^t X) \left(\hat{\beta} - \beta\right)}{(p+1)\,\hat{\sigma}_\varepsilon^2},$$

avec $\hat{\beta}$ et $\hat{\sigma}_\varepsilon^2$ sont indépendante, on accepte H_0 au niveau α, si

$$Q\left(\hat{\beta}, \hat{\sigma}_\varepsilon^2\right) \leq f,$$

où f est le fractile d'ordre $(1 - \alpha)$ de la loi de Fisher $\mathcal{F}(p + 1, n - p - 1)$.

Les résultats sont présentés dans un tableau appelé le tableau d'analyse de la variance (ANOVA) sous forme suivante :

Source de variation	**ddl**	**Somme des carrés**	**Moyenne des carrés**	**F**
Expliquée	p	SCE	$MCE = SCE/p$	$\frac{MCE}{MCR}$
Résiduelle	$n-p-1$	SCR	$MCR = \frac{SCR}{n-p-1}$	
Totale	$n-1$	SCT		

On accepte H_0 si

$$F = \frac{MCE}{MCR} \leq f,$$

où f est le fractile d'ordre $(1-\alpha)$ de la loi de Fisher $\mathcal{F}(p, n-p-1)$.

2.7.3 Test sur le modèle réduit

On cherche à tester la nullité de certains coefficients, le problème sera donc à tester :

$$\begin{cases} H_0 : b_1 = b_2 = \cdots = b_q = 0, \\ H_1 : b_j \neq 0,\ j = \overline{1,q},\ q < p, \end{cases}$$

où p est le nombre exact des paramètres du modèle. Soit R_q^2 le coefficient de détermination du modèle réduit à $(p-q)$ variables.

Sous H_0, la statistique de test est :

$$Q_q = \frac{\left(R^2 - R_q^2\right)(n-p-1)}{(1-R^2)\,q} = \frac{(SCE - SCE_q)/q}{SCR/(n-p-1)},$$

où SCE_q est la somme des carrées expliquées du modèle réduit.

On accepte H_0 si

$$Q_q \leq f',$$

où f' est le fractile d'ordre $(1-\alpha)$ de la loi de Fisher $\mathcal{F}(q, n-p-1)$.

Remarque 2.7.1 *Ce test est utile pour faire de la modélisation pas à pas et sélectionner un ensemble optimal de régresseurs nécessaires à la reconstruction de Y.*

2.8 Prévision

La prévision est l'un des buts de la régression, soit $X_0 = (1, X_0^1, X_0^2, \ldots, X_0^p)$ le vecteur des variables explicatives pour un nouvel individu, et Y_0 la valeur correspondante de la variable à expliquer. On peut prédire Y_0 grâce au modèle ajusté, comme suit :

$$\hat{Y}_0 = \hat{b}_0 + \hat{b}_1 X_0^1 + \hat{b}_2 X_0^2 + \cdots + \hat{b}_p X_0^p.$$

Les intervalles de confiance des prévisions de Y et $E(Y)$ au niveau de confiance $(1-\alpha)\,\%$, sont données respectivement par :

$$prév(Y) = \hat{Y}_0 \mp t\hat{\sigma}_\varepsilon \left(1 + X_0 \left(X^t X\right)^{-1} X_0^t\right)^{1/2},$$
$$prév(E(Y)) = \hat{Y}_0 \mp t\hat{\sigma}_\varepsilon \left(X_0 \left(X^t X\right)^{-1} X_0^t\right)^{1/2},$$

où t est le fractile d'ordre $(1-\alpha/2)$ de la loi de Student $\mathcal{T}(n-p-1)$.

Chapitre 3

Régression Linéaire Bayésienne

Dans ce chapitre, on s'intéresse à la régression linéaire dans le cadre bayésien. On va d'abord présenter successivement les lois de probabilités intervenant au modèle bayésien. Par la suite, on va aborder l'estimation bayésienne d'un modèle de régression linéaire simple. Ainsi, lorsque le nombre de variables explicatives est important, on considère les aspects algorithmiques. Enfin, on va exposer un exemple d'application sous logiciel R. Pour plus d'informations vous pouvez consulter les livres de Robert (2006) [6] et Birkes et Dodge (1993) [4].

3.1 Théorème de Bayes

Soit $(\mathcal{X}, \mathcal{A}, P)$ un espace probabilisé et A, B deux évènements tel que $P(B) \neq 0$.

Définition 3.1.1 (Probabilité conditionnelle)

La probabilité conditionnelle de A sachant B est définie comme suit

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

Théorème 3.1.1 (Probabilités totales)

Soit $(A_j)_{j\in n}$ *est une partition (c-à-d* $A_i \cap A_j \neq \phi$, $\forall i \neq j$ *et* $\overset{n}{\underset{j=1}{\cup}} A_j = \mathcal{X}$*) de l'évènement* A*, alors*

$$P(B) = \sum_{j=1}^{n} P(B|A_j)\, P(A_j).$$

Théorème 3.1.2 (Bayes (1763) **[1])**

Le théorème de Bayes est une conséquence immédiate des probabilités conditionnelles et des probabilités totales, alors

$$P(A_i|B) = \frac{P(B|A_i)\, P(A_i)}{\sum_{j=1}^{n} P(B|A_j)\, P(A_j)}.$$

3.2 Modèle Bayésien

Les espaces intervenant dans un modèle Bayésien sont les suivants

- **Espace des observations** (noté $\mathcal{X}$) : il représente l'ensemble des résultats suit à une étude d'un phénomène.
- **Espace des actions** (noté A) : il représente l'ensemble des actions ou décisions à prendre après l'obtention de l'information.
- **Espace des états de la nature** (noté Θ) : c'est l'espace des paramètres inconnus θ.
- **Espace des règles de décisions** (noté $\mathcal{D}$) : il représente l'ensemble des règles de décisions $\delta \in \mathcal{D}$ qu'on définit comme une application de $\mathcal{X}$ dans A :

$$\delta : \mathcal{X} \rightarrow A$$
$$x_i \rightarrow \delta(x_i) = a_i,$$

où a_i étant la $i^{éme}$ action.

Nous définissons les lois de probabilité qui interviennent en statistique bayésienne.

Définition 3.2.1 (Vraisemblance)

La vraisemblance (notée $f(x\,|\theta)$) est la loi des observations ou encore la loi de la variable aléatoire $x = (x_1, \ldots, x_n)$ en fonction du paramètre θ, donnée par

$$f(x\,|\theta) = \prod_{i=1}^{n} f(x_i\,|\theta).$$

Définition 3.2.2 (Loi a priori)

La loi a priori est une loi de probabilité qui résume toute l'information disponible sur le paramètre inconnu θ, notée par $\pi(\theta)$. L'appellation a priori exprime le fait qu'elle a été établie préalablement à l'observation des données x.

Définition 3.2.3 (Loi jointe du couple)

La loi joint du couple, généralement notée par $f(\theta, x)$, est donnée par

$$f(\theta, x) = f(x\,|\theta)\pi(\theta).$$

Définition 3.2.4 (Loi marginale)

On appelle loi marginale de x la loi définie par

$$f(x) = \int_{\Theta} f(\theta, x)d\theta = \int_{\Theta} f(x\,|\theta)\pi(\theta)\,d\theta.$$

Définition 3.2.5 (Loi a posteriori)

La loi a posteriori est la loi conditionnelle de θ sachant x, notée $\pi(\theta\,|x)$. En vertu de la formule de Bayes, on a

$$\pi(\theta\,|x) = \frac{\textit{Loi jointe du couple}}{\textit{Loi marginale}} = \frac{f(\theta, x)}{\displaystyle\int_{\Theta} f(\theta, x)d\theta} = \frac{f(x\,|\theta)\pi(\theta)}{\displaystyle\int_{\Theta} f(x\,|\theta)\pi(\theta)\,d\theta},$$

ou encore, on peut écrire

$$\pi(\theta\,|x) \propto f(x\,|\theta)\pi(\theta)\,, \tag{3.1}$$

où $\propto$ *signifie que les fonctions* $\pi(\theta\,|x)$ *et* $f(x\,|\theta)\pi(\theta)$ *sont proportionnelles.*

Définition 3.2.6 (Modèle Bayésien)

Un modèle statistique bayésien est constitué d'un modèle statistique paramétrique, $f(x\,|\theta)$*, et d'une distribution a priori pour les paramètres,* $\pi(\theta)$*.*

3.3 Lois a priori

L'élément le plus disputé de l'analyse bayésienne est le choix de la loi a priori, sa détermination est donc l'étape la plus importante dans la mise en œuvre de cette inférence. Il existe deux types des lois a priori : La loi a priori conjuguées introduite par Raiffa et Schlaifer (1961) [14], et la loi a priori non informative.

3.3.1 Lois a priori conjuguées

L'avantage des familles conjuguées est de faciliter le calcul de la loi a posteriori. Elles se produisent lorsque la loi a posteriori a la même forme que la loi a priori.

Définition 3.3.1 (Loi a priori conjuguée)

Une famille $\mathcal{F}$ *de loi sur* Θ *est dite conjuguée par une fonction de vraisemblance* $f(x\,|\theta)$ *si, pour tout* $\pi \in \mathcal{F}$*, la loi a posteriori* $\pi(\theta\,|x)$ *appartient également à* $\mathcal{F}$*.*

Les lois a priori conjuguées sont généralement associées à un type particulier de lois qui permet toujours leur obtention, ces lois constituent ce qu'on appelle une famille des exponentielles et qui sont étudiées en détail dans Brown (1986) [5].

Définition 3.3.2 (Famille exponentielle)

On appelle famille exponentielle de dimension k, toute famille de lois dont la densité a la forme suivante

$$f(x\,|\theta) = C(\theta)\, h(x) \exp\{R(\theta).T(x)\},$$

où C et h des fonctions respectivement de Θ et $\mathcal{X}$ dans $\mathbb{R}_+$, et R et T des fonctions de Θ et $\mathcal{X}$ dans $\mathbb{R}^k$.

Exemple 3.3.1

On dispose d'un vecteur d'observations $x = (x_1, \ldots, x_n)$, on suppose $x_i\,|\theta$ suit la loi de Bernoulli de paramètre θ et que la loi a priori est une loi Bêta, on a

$$f(x\,|\theta) = \prod_{i=1}^{n} P(X = x_i\,|\theta) = \theta^s (1-\theta)^{n-s},$$

où $s = \sum_{i=1}^{n} x_i$, comme $\theta \sim \mathcal{B}e(\alpha, \beta)$, on a

$$\pi(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)} \mathbb{1}_{[0,1]}(\theta),$$

où $B(\alpha,\beta) = \dfrac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$, $\Gamma(\alpha) = (k-1)!$.

Il est facile de vérifier que

$$\int_{\Theta} f(x\,|\theta)\pi(\theta)\, d\theta = B(a,b),$$

où $a = \alpha + s$ et $b = \beta + n - s$.

D'où

$$\pi(\theta\,|x) = \frac{f(x\,|\theta)\pi(\theta)}{\displaystyle\int_{\Theta} f(x\,|\theta)\pi(\theta)\, d\theta} = \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)} \mathbb{1}_{[0,1]}(\theta).$$

Par conséquent

$$\theta \,|x \sim \mathcal{B}e\,(\alpha + s, \beta + n - s)\,.$$

Exemples de lois conjuguées

$f(x\,\|\theta)$	$\pi\,(\theta)$	$\pi\,(\theta\,\|x\,)$
Normale $\mathcal{N}\,(\theta, \sigma^2)$	Nomale $\mathcal{N}\,(\theta, \tau^2)$	$\mathcal{N}\left(\dfrac{\sigma^2\mu + \tau^2 x}{\sigma^2 + \tau^2}, \dfrac{\sigma^2\tau^2}{\sigma^2 + \tau^2}\right)$
Poisson $\mathcal{P}\,(\theta)$	Gamma $\mathcal{G}\,(\alpha, \beta)$	$\mathcal{G}\,(\alpha + x, \beta + 1)$
Gamma $\mathcal{G}\,(v, \theta)$	Gamma $\mathcal{G}\,(\alpha, \beta)$	$\mathcal{G}\,(\alpha + v, \beta + x)$
Binomiale $\mathcal{B}\,(n, \theta)$	Bêta $\mathcal{B}e\,(\alpha, \beta)$	$\mathcal{B}e\,(\alpha + x, \beta + n - x)$
Binomiale Négative $\mathcal{N}eg\,(m, \theta)$	Bêta $\mathcal{B}e\,(\alpha, \beta)$	$\mathcal{B}e\,(\alpha + m, \beta + x)$

TAB. 3.1 – Quelques exemples de lois a priori conjuguées usuelles

3.3.2 Lois a priori non informative

Dans le cas où on dispose d'aucune information sur la loi a priori, on utilise la loi dite a priori non informative.

Définition 3.3.3 (Loi impropre)

Une loi a priori $\pi\,(\theta)$ est dite impropre si

$$\int_{\Theta} \pi\,(\theta)\,d\theta = +\infty.$$

Cette terminologie est bien sur un abus de langage puisque $\pi\,(\theta)$ n'est pas une densité de probabilité. Ce type de lois n'est utile que si la loi a posteriori $\pi\,(\theta\,|x\,)$ existe, aussi on se limite aux lois impropres, telles que

$$\int_{\Theta} f(x\,|\theta)\pi\,(\theta)\,d\theta < +\infty.$$

Nous donnons dans ce qui suit deux techniques de construction de lois non-informatives.

Lois a priori de Laplace

Laplace [11], [12] et [13] fut le premier à utiliser des techniques non-informatives en l'absence d'information, puis il a utilisé la loi uniforme qui est l'une des lois les plus simples et les plus utilisées parmi les lois a priori pour l'approche non-informative. En effet, ce choix est basé sur l'équiprobabilité de la valeur du paramètre θ dans son domaine de définition. Supposons que Θ est un ensemble de taille k, alors

$$\pi(\theta) = \frac{1}{k}.$$

Alors il est possible de travailler avec des lois a priori impropres, du moment que nous n'essayons pas de les interpréter comme des lois de probabilité.

Lois a priori de Jeffreys

Les lois a priori de Jeffreys [10] sont fondées sur l'information de Fisher, qui mesure la quantité d'information sur θ, donnée par

$$I(\theta) = E\left[\left(\frac{\partial}{\partial\theta}\log f(x\,|\theta)\right)^2\right],$$

dans le cas unidimensionnel. Sous certaines conditions de régularité, l'information peut se réécrire

$$I(\theta) = -E\left(\frac{\partial^2}{\partial^2\theta}\log f(x\,|\theta)\right).$$

Dans le cas multidimensionnel, $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$, la matrice d'information de Fisher est définie par

$$I(\theta)\,|_{i,j} = -E\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f(x\,|\theta)\right],\ 1 \leq i, j \leq k$$

Définition 3.3.4 (Loi a priori de Jeffreys)

On appelle loi a priori non informative de Jeffreys, la loi (éventuellement impropre) de

densité

$$\pi(\theta) = c\sqrt{I(\theta)},$$

dans le cas unidimensionnel, où c est une constante de normalisation.

Dans le cas multidimensionnel, la loi a priori non informative est de la forme

$$\pi(\theta) = \left[\det\left(I(\theta)|_{i,j}\right)\right]^{1/2}.$$

Exemple 3.3.2 (Loi exponentielle)

Soit $x \sim \exp(\theta)$. *Alors*

$$f(x|\theta) = \theta \exp(-\theta x)\, \mathbb{1}_{[0,+\infty[}(x),$$

On calcule l'information de Fischer

$$\frac{\partial}{\partial\theta}\log f(x|\theta) = \frac{1}{\theta} - x, \quad \frac{\partial^2}{\partial^2\theta}\log f(x|\theta) = -\frac{1}{\theta^2},\ I(\theta) = \frac{1}{\theta^2},$$

d'où la loi non informative de Jeffreys est

$$\pi(\theta) = c\frac{1}{\theta},$$

qui est une loi a priori impropre qu'on peut normaliser en considérant que $\theta \in [a, b]$ *avec* $a > 0$. *On détermine alors la constante c, tel que*

$$c\int_a^b \pi(\theta)\, d\theta = 1,\ et \int_a^b \frac{1}{\theta} d\theta = \log b - \log a.$$

Finalement, on obtient la loi a priori de Jeffreys

$$\pi(\theta) = \frac{1}{\theta}\log\left(\frac{b}{a}\right).$$

3.4 Estimation Bayésienne

La méthode Bayésienne est un ensemble de techniques statistiques utilisées pour modéliser des problèmes, extraire de l'information de données brutes et prendre des décisions de façon cohérente et rationnelle. Son cadre d'application est général, mais ses avantages sont déterminants lorsque l'information disponible est incertaine ou incomplète. Cette dernière s'appuie principalement sur le théorème de Bayes. Dans cette section, on va appliquer la méthode Bayésienne aux modèles de régression.

3.4.1 Régression Linéaire Simple

Soit Y une variable dépendante à expliquer, et soit X une variable explicative, on souhaite estimer la droite de la régression linéaire simple pour un ensemble des données impliquant les variables x_i et y_i, on peut utiliser le modèle

$$y_i \,|x_i = b_0 + b_1 x_i + \varepsilon_i,\ i = \overline{1,n}.$$

où x_i est la $i^{ème}$ observation de X, y_i est la $i^{ème}$ observation de Y et ε_i est la $i^{ème}$ observation d'une variable aléatoire désignant l'erreur. On suppose que les ε_i sont indépendants de loi normale ($\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$). b_0, b_1 et σ_ε^2 sont des paramètres inconnus.

La variable conditionnelle $y \,|x_i, b_0, b_1, \sigma_\varepsilon^2$ est de loi normale telle que $E\,(y \,|x_i\,) = b_0 + b_1 x_i$ et $Var\,(y \,|x_i\,) = \sigma_\varepsilon^2$.

La fonction de vraisemblance des paramètres inconnus est

$$\begin{aligned} f(x, y \,|b_0, b_1, \sigma_\varepsilon\,) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}} \exp\left\{-\frac{1}{2\sigma_\varepsilon^2}\left(y_i - (b_0 + b_1 x_i)\right)^2\right\} \\ &= \frac{1}{\left(2\pi\sigma_\varepsilon^2\right)^{n/2}} \exp\left\{-\frac{1}{2\sigma_\varepsilon^2}\left(y_i - (b_0 + b_1 x_i)\right)^2\right\} \\ &\propto \frac{1}{\sigma_\varepsilon^n} \exp\left\{-\frac{1}{2\sigma_\varepsilon^2}\sum_{i=1}^{n}\left(y_i - (b_0 + b_1 x_i)\right)^2\right\}, \end{aligned}$$

où l'on ignore la constante de proportionnalité $(2\pi)^{-n/2}$ dans l'équation.

Loi a priori non informative

Dans ce modèle, on dispose d'aucune information sur les paramètres b_0 et b_1 et σ_ε autres que σ est positif. Alors, on a besoin d'une loi a priori non informative. A cet effet, on utilise une loi a priori dite loi a priori standard, bien qu'elle soit impropre, la loi a posteriori qui en résulte est une densité de probabilité propre.

On suppose alors que les paramètres b_0, b_1 et $\log\sigma$ sont uniformément distribués et indépendants, ce qui donne la loi a priori jointe de ces paramètres comme suit (Voir [6] page 139,141-142)

$$\pi\left(b_0, b_1, \sigma_\varepsilon\right) \propto \frac{1}{\sigma}.$$

Loi a posteriori jointe

La loi a posteriori jointe est la base de l'inférence Bayésienne, pour l'obtenir on combine la fonction de vraisemblance avec la densité a priori

$$\begin{aligned}\pi\left(b_0, b_1, \sigma_\varepsilon \left| x, y\right.\right) &\propto f(x, y \left| b_0, b_1, \sigma_\varepsilon^2\right.)\pi\left(b_0, b_1, \sigma_\varepsilon^2\right) \\ &\propto \frac{1}{\sigma_\varepsilon^{n+1}} \exp\left\{-\frac{1}{2\sigma_\varepsilon^2}\sum_{i=1}^{n}\left(y_i - b_0 - b_1 x_i\right)^2\right\}. \qquad (3.2)\end{aligned}$$

Nous prenons les estimateurs des paramètres b_0, b_1, σ_ε^2, comme suit

$$\hat{b}_0 = \bar{y} - \hat{b}_1\bar{x},\ \hat{b}_1 = \frac{S_{xy}}{S_x} \text{ et } \hat{\sigma}_\varepsilon^2 = S^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(y_i - \hat{b}_0 - \hat{b}_1 x_i\right)^2.$$

Ces estimateurs sont sans biais, elles sont exactement les mêmes que les estimateurs des MCO (voir (1.1), le théorème 1.2.1 et 1.2.3).

Avant de poursuivre nos calculs, il est d'abord important de noter que

$$\sum_{i=1}^{n}\left(b_0-\hat{b}_0\right)\left(y_i-\hat{b}_0-\hat{b}_1x_i\right)=0,$$
$$\sum_{i=1}^{n}\left(y_i-\hat{b}_0-\hat{b}_1x_i\right)\left(b_1-\hat{b}_1\right)x_i=0.$$

Considérons alors l'équation algébrique suivante

$$\begin{aligned}\sum_{i=1}^{n}\left(y_i-b_0-b_1x_i\right)^2&=\sum_{i=1}^{n}\left[\left(y_i-\hat{b}_0-\hat{b}_1x_i\right)-\left(b_0-\hat{b}_0\right)-\left(b_1-\hat{b}_1\right)x_i\right]^2\\&=(n-2)\,S^2+n\left(b_0-\hat{b}_0\right)^2+\left(b_1-\hat{b}_1\right)^2\sum_{i=1}^{n}x_i^2\\&\quad+2\left(b_0-\hat{b}_0\right)\left(b_1-\hat{b}_1\right)\sum_{i=1}^{n}x_i,\end{aligned}$$

on pose $l=\dfrac{1}{2}\sum\limits_{i=1}^{n}\left(y_i-b_0-b_1x_i\right)^2$, en remplaçant cette égalité dans la formule (3.2), on a

$$\pi\left(b_0,b_1,\sigma_\varepsilon^2\,|x,y\right)\propto\frac{1}{\sigma_\varepsilon^{n+1}}\exp\left\{-\frac{l}{\sigma_\varepsilon^2}\right\}.$$

En intégrant par rapport à σ, on obtient la loi a posteriori jointe de b_0 et de b_1

$$\begin{aligned}\pi\left(b_0,b_1\,|x,y\right)&=\int_0^\infty\pi\left(b_0,b_1,\sigma_\varepsilon\,|x,y\right)d\sigma=\int_0^\infty\frac{1}{\sigma_\varepsilon^{n+1}}\exp\left\{-\frac{l}{\sigma_\varepsilon^2}\right\}d\sigma\\&\propto l^{-\frac{n}{2}}.\end{aligned}$$

Alors

$$\begin{aligned}\pi\left(b_0,b_1\,|x,y\right)\propto\Bigg[&(n-2)\,S^2+n\left(b_0-\hat{b}_0\right)^2+\left(b_1-\hat{b}_1\right)^2\sum_{i=1}^{n}x_i^2\\&+2\left(b_0-\hat{b}_0\right)\left(b_1-\hat{b}_1\right)\sum_{i=1}^{n}x_i\Bigg]^{-\frac{n}{2}}.\end{aligned}\tag{3.3}$$

Loi a posteriori marginale

La loi a posteriori marginale de b_0 (respectivement de b_1) est obtenue en intégrant par rapport à b_0 (respectivement à b_1)

$$\pi\left(b_0\left|x,y\right.\right) \propto \left[(n-2)+\frac{S_x}{\frac{S^2}{n}\sum_{i=1}^{n}x_i^2}\left(b_0-\hat{b}_0\right)^2\right]^{-((n-2)+1)/2}, \quad -\infty < b_0 < +\infty,$$

et

$$\pi\left(b_1\left|x,y\right.\right) \propto \left[(n-2)+\frac{S_x}{S^2}\left(b_1-\hat{b}_1\right)^2\right]^{-((n-2)+1)/2}, \quad -\infty < b_1 < +\infty,$$

où $\left(\frac{S_x}{\frac{S^2}{n}\sum x_i^2}\right)^{1/2}\left(b_0-\hat{b}_0\right)$ et $\frac{S_x^{1/2}}{S}\left(b_1-\hat{b}_1\right)$ suivent une loi de Student à $(n-2)$ ddl.

La loi a posteriori marginale de σ_ε est obtenue en intégrant la loi a posteriori jointe (3.2) par rapport à b_0 et b_1

$$\pi\left(\sigma_\varepsilon\left|x,y\right.\right) \propto \frac{1}{\sigma^{(n-2)+1}}\exp\left\{-\frac{n-2}{2\sigma^2}\right\},\ 0<\sigma<+\infty,$$

qui est une gamma inverse de paramètres $\left(\frac{n}{2}-1\right), \frac{(n-2)\,S^2}{2}$.

Intervalle de crédibilité

Dans l'analyse bayésienne, on utilise l'intervalle de crédibilité au lieu de l'intervalle de confiance. L'interprétation de ces deux derniers est légèrement différente : Les intervalles de crédibilité traitent leurs limites comme fixes et le paramètre estimé comme variable aléatoire, les intervalles de confiances traitent leurs limites comme variables aléatoire et le paramètre estimé comme fixe.

- L'intervalle de crédibilité du paramètre b_0 est

$$P\left(\xi\left|\hat{b}_0 - b_0\right| \leq t\right) = 1 - 2\alpha,$$

alors

$$b_0 \in \left[\hat{b}_0 - \frac{t}{\xi}, \hat{b}_0 + \frac{t}{\xi}\right],$$

où $\xi = \left(\frac{S_x}{\frac{S^2}{n}\sum x_i^2}\right)^{1/2}$, t est le fractile d'ordre $(1-2\alpha)$ de la loi de Student à $(n-2)$ ddl.

- L'intervalle de crédibilité du paramètre b_1 est

$$P\left(\frac{\left|\hat{b}_1 - b_1\right|}{\delta} \leq t\right) = 1 - 2\alpha,$$

alors

$$b_1 \in \left[\hat{b}_1 - \delta t, \hat{b}_1 + \delta t\right],$$

où $\delta = \frac{S}{S_x^{1/2}}$, t est le fractile d'ordre $(1-2\alpha)$ de la loi de Student à $(n-2)$ ddl.

Les résultats de la loi a posteriori de b_0 et b_1 montrent que sous la connaissance de la loi a priori, les intervalles de crédibilité sont en fait numériquement équivalents aux intervalles de confiance de la méthode MCO fréquentiste classique.

Prévision

Le modèle de prévision du résultat de y^* compte tenu d'une nouvelle observation x^* est

$$y_i^* \,|x_i^* = b_0 + b_1 x_i^* + \varepsilon_i,\ i = \overline{1,n},$$

où $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$, pour faire une prédiction, on calcule une densité prédictive, étant donné

que les paramètres ont déjà été estimés.

Définition 3.4.1 (Densité prédite)

Soit la loi a posteriori $\pi(\theta|x)$ *donnée par (3.1). Si on souhaite prédire une nouvelle observation* y^**, on utilise une densité prédite pour* y^* *donnée par*

$$p(y^*|x^*) = \int f(y^*|\theta)\pi(\theta|x^*)\, d\theta.$$

Alors, dans la régression linéaire simple on a

$$\begin{aligned} p(y^*|x^*) &= \int\int\int f(y^*|x^*, b_0, b_1, \sigma)\pi(b_0, b_1, \sigma|x, y)\, db_0 db_1 d\sigma \\ &= \int\int\int \frac{1}{\sigma_\varepsilon^{2n+1}} \exp\left\{-\frac{1}{2\sigma_\varepsilon^2}\left[\sum_{i=1}^{n}(y^* - b_0 - b_1 x^*)^2\right.\right. \\ &+ \left.\left.\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)^2\right]\right\} db_0 db_1 d\sigma. \end{aligned}$$

Loi a priori conjuguée

On suppose qu'on dispose d'informations a priori sur les paramètres inconnus b_0, b_1 et σ^2, il est pratique d'exprimer ces informations sous forme de lois a priori conjuguées.

La manière la plus adoptée est d'admettre qu'il existe une indépendance a priori des paramètres b_0, b_1 et σ^2, on suppose par suite que b_0 et b_1 sont de loi gaussienne et σ^2 suit une loi gamma inverse (c-à-d $1/\sigma^2$ suit une loi gamma), cependant, le résultat de ces trois fonctions de densités a priori n'est pas une densité conjuguée puisque la densité a posteriori obtenue n'est pas un produit de trois densités indépendantes de la même famille.

3.4.2 Régression Linéaire Multiple

Dans un cadre bayésien, il est impossible de calculer explicitement les lois a posteriori pour un modèle de régression linéaire multiple, car les calculs peuvent devenir de plus en plus difficiles, lorsque le nombre p de variables explicatives est grand. C'est là que la méthode de Monte-Carlo par chaîne de Markov (MCMC) devient très utile pour aider à ces calculs. Cette dernière simule la loi a posteriori afin de pouvoir l'analyser. Les résultats peuvent ensuite être utilisés pour tirer des conclusions sur le modèle et les paramètres. Un des algorithmes de la MCMC, largement utilisé dans les applications de l'analyse bayésienne, est l'échantillonneur de Gibbs.

Echantillonnage de Gibbs

L'échantillonnage de Gibbs a été utilisé par Geman et Geman (1984) [9], il s'agit d'une méthode alternative de génération (simulation) des variables aléatoires indirectement à partir d'une distribution.

Après avoir choisi un point initial, l'idée principale de l'échantillonneur de Gibbs est de générer les d composantes du vecteur des paramètres θ les unes après les autres conditionnellement à toutes les autres composantes. Si $\pi(\theta\,|x)$ est la loi a posteriori jointe des d composantes du vecteur θ, conditionnellement aux données observées $(X_1, X_2, \ldots, X_n)$, on utilise alors les densités conditionnelles $\pi(\theta_1\,|\theta_2, \theta_3, \ldots, \theta_d, x)$, $\pi(\theta_2\,|\theta_1, \theta_3, \ldots, \theta_d, x)$ et ainsi de suite. A chaque $k^{ème}$ étape, la loi conditionnelle utilise les valeurs générées les plus récentes parmi toutes les autres composantes. Par la théorie des chaînes de Markov, il s'avère que, comme $k \to \infty$, la loi des réalisation obtenues tend vers $\pi(\theta\,|x)$.

L'algorithme d'échantillonnage de Gibbs est le suivant

1. Fixer $k = 0$,
2. - Générer $\theta_1^{(k+1)}$ selon $\pi\left(\theta_1 \left| \theta_2^{(k)}, \theta_3^{(k)}, \ldots, \theta_d^{(k)}, x \right.\right)$,
   - Générer $\theta_2^{(k+1)}$ selon $\pi\left(\theta_2 \left| \theta_1^{(k+1)}, \theta_3^{(k)}, \ldots, \theta_d^{(k)}, x \right.\right)$,

$\vdots$

- Générer $\theta_{d-1}^{(k+1)}$ selon $\pi\left(\theta_{d-1}\left|\theta_1^{(k+1)},\theta_2^{(k+1)},\ldots,\theta_d^{(k)},x\right.\right)$,
- Générer $\theta_d^{(k+1)}$ selon $\pi\left(\theta_d\left|\theta_1^{(k+1)},\theta_3^{(k)},\ldots,\theta_d^{(k+1)},x\right.\right)$,

3. Si la convergence est obtenue, alors

   -retenir $\theta=\theta^{(k+1)}$ si non fixer $k=k+1$ et retourner à l'étape 1.

3.5 Exemple illustratif sous R

Dans cette section, on va examiner un exemple de régression linéaire simple à l'aide du logiciel R.

3.5.1 Régression Linéaire Simple

On va étudier la relation entre le rendement de maïs X (en quintal) et la quantité d'engrais Y (en kilogrammes) sur des parcelles de terre similaires, les données consignés dans (3.2). Ce tableau est tiré du livre de Bourbonnais (1998) [3].

X	20	24	28	22	32	28	32	36	41	41
Y	16	18	23	24	28	29	26	31	32	34

TAB. 3.2 – Rendement de maïs et quantité d'engrais

Le nuage de points, illustré dans la figure (3.1), indique qu'il existe une relation linéaire entre les deux variables X et Y. On peut donc appliquer le modèle de régression linéaire simple

$$Y=b_0+b_1X+\varepsilon,$$

Cette figure est obtenue à l'aide des commandes suivantes

```
x=c(20,24,28,22,32,28,32,36,41,41)
y=c(16,18,23,24,28,29,26,31,32,34)
plot(x,y,xlab="Rendement de maïs",ylab="Quantité d'engrais")
```

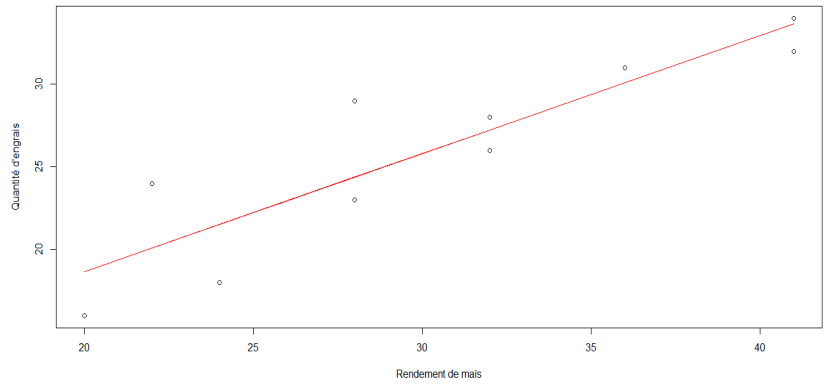

FIG. 3.1 – Nuage de points et droite de régression

Approche fréquentiste

Pour effectuer les calculs nécessaires par la méthode de MCO, on utilise les commandes suivantes sous logiciel R.

```
RLS=lm(y~x)
ye=4.3928+0.7141*x
abline(RLS,col="red")
summary(RLS)
confint(RLS)
anova(RLS)
```

Les résultats obtenus sur les estimateurs des paramètres $\hat{b}_j$, les écarts-types $\hat{\sigma}_{\hat{b}_j}$, les statistiques observées de Student T_{obs} et les intervalles de confiances IC, sont résumés respectivement dans le tableau 3.5.1.

Coefficients	$\hat{b}_j$	$\hat{\sigma}_{\hat{b}_j}$	T_{obs}	IC
b_0	4.3928	3.9718	1.106	$]-4.766, 13.551[$
b_1	0.7141	0.1273	5.609	$]0.420, 1.005[$

TAB. 3.3 – Résultats d'estimation des paramètres

Le coefficient de détermination vaut

$$R^2 = 0.7973,$$

qui est une valeur satisfaisante.

L'estimateur sans biais de la variance résiduelle est

$$\hat{\sigma}_{\varepsilon}^2 = (2.825)^2 = 7.980\,6.$$

On a donc la droite ajustée de la régression

$$\hat{y}_i = 4.3928 + 0.7141 x_i, i = \overline{1, 10}.$$

que l'on peut tracer avec le nuage de point voir la figure 3.1, la ligne en rouge.

Au seuil de signification 5%, on teste les hypothèses bilatérales de signification des paramètres b_0 et b_1

$$\begin{cases} H_0 : b_j = 0, \\ H_1 : b_j \neq 0, \end{cases} j = 0, 1.$$

Pour b_0, on remarque que $T_{obs} = 1.106 < t_{0.975}\,(8) = 2.306$, H_0 est acceptée, alors le paramètre b_0 est significativement nulle. H_0 est rejetée pour b_1 car $T_{obs} = 5.609 > t_{0.975}\,(8) = 2.306$, alors le paramètre b_1 est significatif.

Pour la validation du modèle, on considère les hypothèses suivantes

$$\begin{cases} H_0 : b_0 = 0 \text{ et } b_1 = 0, \\ H_1 : b_0 \neq 0 \text{ et } b_1 \neq 0. \end{cases}$$

Le tableau de l'analyse de variance 3.5.1 résume les résultats obtenus à l'aide de la commande `anova()`.

Source de variation	**ddl**	**SC**	**MC**	**F**
Expliquée	1	251.061	251.06	31.46
Résiduelle	8	63.839	7.98	
Totale	9	314.9		

TAB. 3.4 – Résultats de la table d'anova pour la régressionn linéaire simple

On remarque que la statistique de Fisher F est supérieur à la valeur $f_{0.95}(1,8) = 5.32$, on rejete donc H_0 ce qui confirme la signification globale du modèle.

Approche Bayésienne

Étant donné que nous n'avons aucune information a priori sur les données du tableau 3.2, nous utilisons une loi a priori jointe uniforme

$$\pi(b_0, b_1, \sigma_\varepsilon) \propto \frac{1}{\sigma_\varepsilon}.$$

Ainsi, nous pouvons utiliser la fonction de vraisemblance dans l'équation (3.2) de la loi a posteriori jointe $\pi(b_0, b_1, \sigma_\varepsilon \,|x, y)$.

Comme les estimations des paramètres du modèle bayésien sont sans biais, on peut utiliser les estimateurs de MCO trouvées dans la table 3.5.1

$$\hat{b}_0 = 4.3928 \text{ et } \hat{b}_1 = 0.7141,$$

pour obtenir les lois a posteriori marginales de b_0 et de b_1.

En utilisant la fonction `bayes.lin.reg` du package `Bolstad` sous R, avec quelques modification, les densités a posteriori de ces paramètres sont représentées dans la figure 3.2.

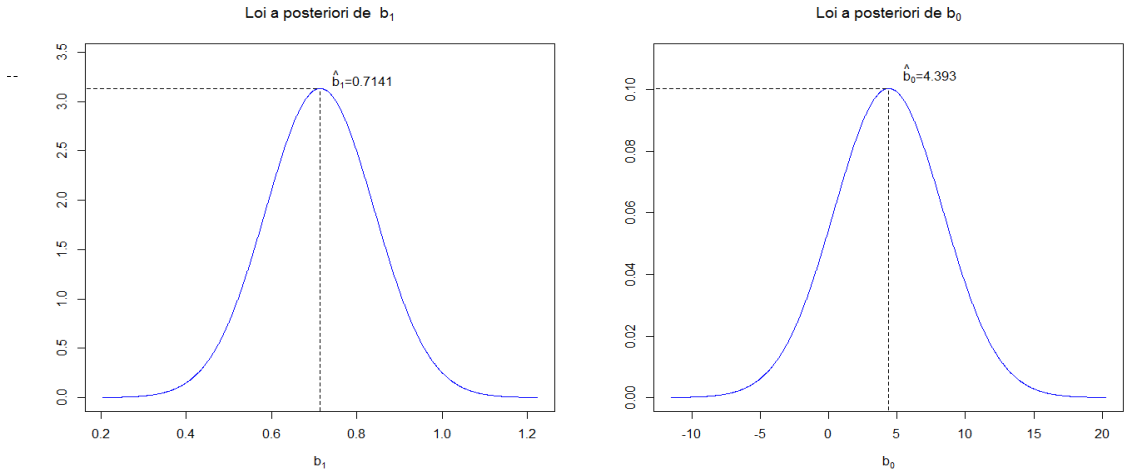

FIG. 3.2 – Lois a posteriori des paramètres b_0 et b_1

On remarque que les moyennes a posteriori sont approximativement égales aux estimations des paramètres de MCO, telles que

$$E\left(b_0 \mid x, y\right) = \hat{b}_0 = 4.3928,$$
$$E\left(b_1 \mid x, y\right) = \hat{b}_1 = 0.7141.$$

Étant donné que les intervalles de crédibilités sont numériquement les mêmes que les intervalles de confiance de l'approche fréquentiste, nous pouvons utiliser la fonction `lm` pour obtenir les estimations de MCO et construire les intervalles de crédibilités, de b_0 et b_1 au niveau de confiance 95%.

$$b_0 \in [-4.766, 13.551] \text{ et } b_1 \in [0.420, 1.005]\,.$$

La différence principale est l'interprétation. Par exemple, sur la base des données, nous pensons qu'il y a 95% de chance que la quantité d'engrais Y augmente de 0.420 jusqu'à 1.005 pour chaque augmentation de 1 quintal du rendement de maïs X et qu'il y a 95% de chance que la quantité d'engrais moyenne augmente de -4.766 à 13.551quant le rendement de maïs est égale à 0 quintal.

L'intervalle de crédibilité pour la moyenne

$$E\left(y_i \left| x_i \right.\right) = \hat{b}_0 + \hat{b}_1 x_i \pm t\hat{\sigma}^2 \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right),$$

et l'intervalle de prévision

$$y_p \left| x_p \right. = \hat{b}_0 + \hat{b}_1 x_i \pm t\hat{\sigma}^2 \left(1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right),$$

sont illustrés dans la figure 3.3.

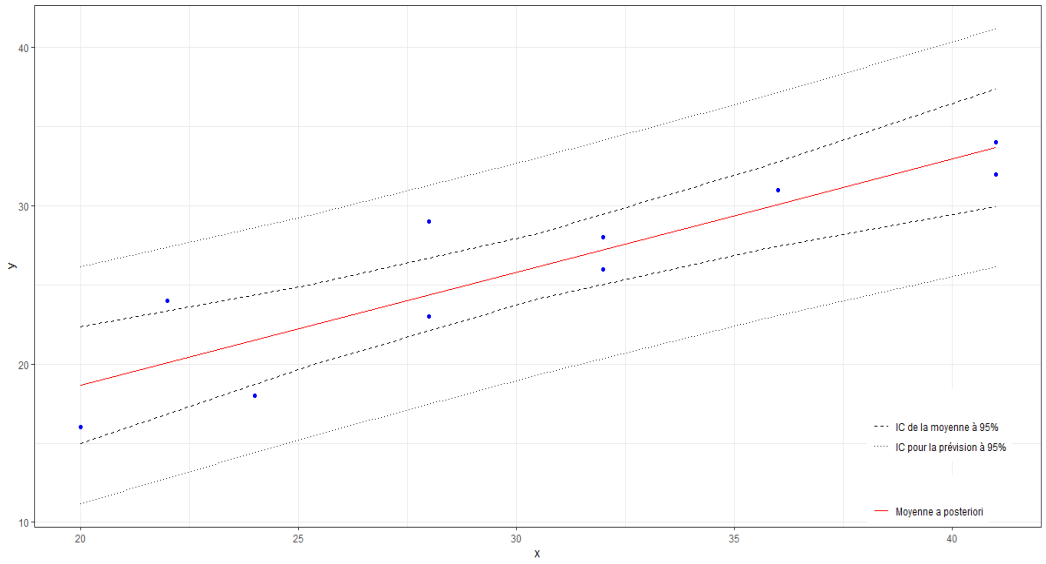

FIG. 3.3 – Intervalle de crédibilité de la moyenne et de prévision

Cette figure est la sortie du code suivant, en utilisant le package `ggplot2`.

```
library(ggplot2)
ye=b0+b1*xp
ymean=data.frame(predict(RLS,newdata=data.frame(x=xp),
            interval="confidence",level=0.95))
ypred=data.frame(predict(RLS,newdata=data.frame(x=xp),
            interval="prediction",level=0.95))
output=data.frame(x=xp,y_hat=ye,ymean_lwr=ymean$lwr,
ymean_upr=ymean$upr,ypred_lwr=ypred$lwr,ypred_upr=ypred$upr)
plot1=ggplot(data=data.frame(y,x),aes(x=x,y=y))+
            geom_point(color="blue")
plot2 = plot1 +
geom_line(data =output,aes(x=xp,y=ye,color="first"),lty=1)+
geom_line(data=output,aes(x=xp,y=ymean_lwr,lty="2"))+
geom_line(data=output,aes(x=xp,y=ymean_upr,lty="2"))+
geom_line(data=output,aes(x=xp,y=ypred_lwr,lty="3"))+
scale_colour_manual(values=c("red"),labels="Moyenne a posteriori",
            name="")+
scale_linetype_manual(values=c(2,3),labels=c("IC de la moyenne à 95%",
"IC pour la prévision à 95%"),name="")+
theme_bw() +
theme(legend.position = c(1, 0), legend.justification = c(1.2, 0))
```

Conclusion

A travers ce travail, nous avons présentés les différents types de régression linéaire permettant de préciser la liaison entre les variables d'un phénomène statistique, à savoir la régression linéaire simple et la régression linéaire multiple. Nous avons non seulement présenté ces méthodes sous une approche fréquentiste, mais nous nous sommes intéressées à une approche bayésienne qui représente le cadre général de l'estimation de paramètre basé sur le théorème de Bayes.

En conclusion, la méthode de régression linéaire par l'approche fréquentiste classique repose sur une hypothèse de base et une hypothèse alternative, pour obtenir les estimateurs des paramètres inconnus. D'autre part, l'approche bayésienne repose sur l'information contenue dans les données (la fonction de vraisemblance) associée aux connaissances à priori (loi à priori), pour aboutir à un résultat à postériori.

Bibliographie

[1] Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. Philosophical Transactions of the Royal Society of London, 53.

[2] Berger, J.O.(1985). Statistical decision theory and Bayesian analysis, 2nd edition, Springer-Verlag.

[3] Bourbonnais, R. (1998) Économétrie. Dunod.

[4] Birkes, D and Dodge, Y. (1993). Alternative Methods of Regression. John Wiley and Sons, Inc.

[5] Brown, L. (1986). Foundations of Exponential Families, volume 6 dans IMS lecture notes Monograph Series. Hayward California.

[6] Robert, C.P. (2006) Le choix bayésien-Principe et pratique. Springer.

[7] Cornillon, Pierre-André.and Eric Matzner-Lober (2010).Régression avec R. Springer Science & Business Media.

[8] Cowles M.K. (2013). Applied Bayesian statistics. Springer-Verlag .

[9] Geman, S. & Geman, D. (1984). Stochastics relaxation, Gibbs distribution and the bayesian restoration of images. IEEE transaction on Pattern Analysis and Machine Intilligence 6 : 721-741.

[10] Jeffreys, H. (1961). Theory of Probability (3rd edition). Oxford University Press, Oxford.

[11] Laplace, P. (1773). Mémoire sur la probabilité des causes par les événements. Mémoires de l'Académie Royale des Sciences présentés par divers savants, Reprinted in Laplace (1878).

[12] Laplace, P. (1786). Sur les naissances, les mariages et les morts à Paris depuis 1771 jusqu'à 1784 et dans toute l'étendue de la France, pendant les années 1781 et 1782. Mémoires de l'Académie Royale des Sciences présentés pardivers savants, 11, 35–46. Reprinted in Laplace (1878).

[13] Laplace, P. (1795). Essai Philosophique sur les Probabilités. Epistémé. Christian Bourgeois, Paris. Reprinted in 1986.

[14] Raiffa, H. et Schlaifer, R. (1961). Applied Statistical decision theory. Technical report, Division of Research, Graduate School of Business Administration, Harvard Univ.

Annexe : Abréviations et Notations

Y	$\rightarrow$	Vecteur aléatoire a expliqué de dimension n.
X	$\rightarrow$	Matrice des variables explicative de dimension $(n, p+1)$.
ε	$\rightarrow$	L'erreur.
b_0, b_1	$\rightarrow$	Les paramètres du modèle.
β	$\rightarrow$	Vecteur des paramètres du modèle de dimension $(p+1)$.
$E(.)$	$\rightarrow$	L'esperance.
$\sigma^2, Var(.)$	$\rightarrow$	La variance.
$Cov(.)$	$\rightarrow$	La covariance.
$\mathcal{X}_n^2$	$\rightarrow$	Loi de Khi-deux.
$\mathcal{T}(v)$	$\rightarrow$	Loi de Student.
$\mathcal{F}(v_1, v_2)$	$\rightarrow$	Loi de Fisher.
$\mathcal{N}(\mu, \sigma^2)$	$\rightarrow$	Loi Normal.
MCO	$\rightarrow$	Méthode des moindres carrés ordinaires.
MV	$\rightarrow$	Méthode de Maximum de Vraisemblance.
ddl	$\rightarrow$	Degrés de liberté.
IC	$\rightarrow$	L'intervalle de confiance.
RC	$\rightarrow$	La région de confiance
SCE	$\rightarrow$	Somme des carrés expliqués.
SCR	$\rightarrow$	Somme des carrés des résidus.
SCT	$\rightarrow$	Somme des carrés totale.

$\mathcal{X}$	$\rightarrow$	Ensemble fondamentale.
$\mathcal{A}$	$\rightarrow$	Tribu.
P	$\rightarrow$	Probabilité.
$(\mathcal{X}, \mathcal{A}, P)$	$\rightarrow$	Espace probabilisé.
$\propto$	$\rightarrow$	Proportionnelle.
$f(x \mid \theta)$	$\rightarrow$	La vraisemblance.
$\pi(\theta)$	$\rightarrow$	La loi a priori.
$\pi(\theta \mid x)$	$\rightarrow$	La loi a posteriori.

Résumé

L'objectif de ce travail est de présenter les différents modèles de régression linéaire, avec un intérêt particulier sur la régression bayésienne. Dans un premier temps, nous exposons les principales notions et propriétés d'un modèle de régression linéaire simple. Nous passons ensuite au modèle de régression linéaire multiple. Finalement, nous nous plaçons dans un cadre bayésien en explorant les différentes lois intervenant dans ce texte. On outre, on enrichi ce travail par un exemple illustratif à l'aide du logiciel R.

Mots-clés : Régression linéaire simple, Régression linéaire multiple, Régression bayésienne, Moindres carrées, Maximum de vraisemblance, Loi a priori, Loi a posteriori.

ملخص

الهدف من هذا العمل هو تقديم نماذج الإنحدار الخطي المختلفة ، مع اهتمام خاص بالإنحدار البايزياني. أولاً ، نعرض المفاهيم والخصائص الرئيسية لنموذج الانحدار الخطي البسيط. ثم ننتقل إلى نموذج الانحدار الخطي المتعدد. أخيرًا ، نضع أنفسنا في إطار بايزياني من خلال استكشاف التوزيعات المختلفة المتضمنة في هذا السياق. بالإضافة إلى ذلك، تم إثراء هذا العمل بمثال توضيحي باستخدام برنامج R.

الكلمات الرئيسية : الإنحدار الخطي البسيط ، الانحدار الخطي المتعدد ، الإنحدار البايزياني ، المربعات الصغرى ، الإحتمال الأقصى ، التوزيع القبلي ، التوزيع البعدي.

Abstract

The aim of this work is to present the different linear regression models, with a particular interest in Bayesian regression. First, we expose the main concepts and properties of a simple linear regression model. Then we move on to the multiple linear regression model. Finally, we place ourselves in a Bayesian framework by exploring the different distributions involved in this context. In addition, this work is enriched by an illustrative example using the R software.

Keywords : Simple linear regression, Multiple linear regression, Bayesian regression, Least squares, Maximum likelihood, Prior distribution, Posterior distribution.