République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Mohamed Khider, Biskra

Faculté des Sciences Exactes

Département de Mathématiques

Mémoire présenté en vue de l'obtention du Diplôme :

MASTER en Mathématiques

Option : Statistique

Par

ABDOUS Celia

Titre :

SUR LA RÉGRESSION LINÉAIRE

Membres du Comité d'Examen :

Pr.	YAHIA Djabrane	UMKB	Président
Dr.	AFROUN Faïrouz	UMKB	Promotrice
Dr.	BOUREDJI Hind	UMKB	Examinatrice

02 Juin 2025

Dédicace

Je dédie ce mémoire à :

***Mon cher père**,*

Pour ta sagesse, ton courage et ton appui constant. Tu as toujours cru en moi, même dans le silence.

***À ma tendre mère**,*

Pour ton amour infini, tes sacrifices discrets et tes prières sincères. Tu es le cœur de ma réussite.

***À ma petite sœur**,*

Ta présence douce et ta spontanéité apportent de la lumière à mes journées. Tu occupes une place très spéciale dans mon cœur.

Remerciements

Tout d'abord, je rends grâce à ***Dieu*** Tout-Puissant pour m'avoir permis d'accomplir ce travail.

Je tiens à exprimer ma profonde gratitude à ma promotrice de mémoire, ***Dr. AFROUN Faïrouz***, pour ses conseils précieux, son soutien inébranlable et ses orientations scientifiques qui ont été essentielles à la réussite de ce travail. Je lui suis infiniment reconnaissant.

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

Je tiens également à exprimer ma reconnaissance à ma famille, notamment mes parents, pour leur soutien moral, leur patience et leur encouragement tout au long de ce processus. Leur amour et leur foi en moi ont été une source de force inépuisable.

Je n'oublie pas mes amis, qui ont été une source précieuse de soutien et de motivation tout au long de cette aventure académique.

Enfin, je me remercie moi-même pour ma persévérance et mon engagement dans l'accomplissement de ce projet, malgré les difficultés rencontrées.

Que Dieu bénisse tous ceux qui ont contribué à la réussite de ce travail, et qu'il nous guide vers la réussite dans tous les aspects de nos vies.

Table des matières

Table des figures

Liste des tableaux

Introduction générale

La régression linéaire est l'une des approches statistiques la plus ancienne et la plus fondamental dans l'analyse des données quantitatives. Elle repose sur l'idée d'une relation linéaire entre une variable dépendante (à expliquer) et une ou plusieurs variables indépendantes (explicatives), et permet ainsi d'interpréter, de prévoir et de mieux comprendre les phénomènes étudiés.

L'origine du concept remonte au XIX^{e} siècle avec les travaux du scientifique britannique **Francis Galton**, qui a observé un phénomène de " retour à la moyenne " en étudiant la relation entre la taille des parents et celle de leurs enfants. Il est le premier à avoir utilisé le terme « régression ». Ce concept a ensuite été formalisé mathématiquement par **Karl Pearson**, puis enrichi au fil du temps grâce à l'évolution des outils statistiques et informatiques.

Malgré sa simplicité apparente, la régression linéaire possède une puissance analytique considérable. Elle constitue non seulement un outil descriptif pour estimer la relation entre les variables, mais aussi un cadre rigoureux pour tester des hypothèses, mesurer des effets, et prédire des résultats. Elle sert également de base à des méthodes plus complexes comme la régression logistique, les modèles mixtes ou encore l'apprentissage automatique.

L'application de la régression linéaire se trouve dans plusieurs domaines : en économie (modélisation de la consommation, du revenu, du chômage...), en médecine (relation entre un traitement et une réponse clinique), en ingénierie (comportement de systèmes physiques), en sciences sociales (analyse des comportements humains).

La régression linéaire multiple constitue un outil fondamental en analyse statistique,

permettant de modéliser la relation entre une variable dépendante et plusieurs variables explicatives, l'objectif visé dans ce mémoire est d'étudier et comprendre d'avantage le mécanisme des techniques de la régression linéaire multiple à savoir : Deux méthodes d'estimation en régression linéaire multiple : La méthode des Moindres Carrés Ordinaires (MCO) et la méthode du Maximum de Vraisemblance (MV). Et la méthode de sélection de modèle de régression linéaire multiple Pas-à-Pas dans ses trois versions (sélection forward, sélection Backward et sélection Mixte).

Pour répondre à notre objectif nous avons répartie le présent document comme suit : une introduction générale, trois chapitres, une conclusion générale, une liste bibliographique et deux annexes.

• Le premier chapitre est consacré aux généralités sur la régression linéaire simple et multiple.

• Le deuxième chapitre est consacré à la méthode Stepwise pour la sélection des modèles en régression linéaire multiple.

• Le troisième chapitre est consacré à l'application numérique des trois méthodes de la régression linéaire exposées dans les deux chapitres précèdent dans le but comprendre d'avantage leurs mécanisme sur les plans pratique, afin de dégager leurs avantages, inconvénients et leurs contextes d'application les plus pertinents.

Chapitre 1

Généralité sur la régression linéaire simple et multiple

Dans ce chapitre, on a présenté en premier lieu la régression linéaire simple, par la suite on a élargis cette approche à la régression linéaire multiple, tout en focalisant sur les méthodes d'estimations des paramètres dans la régression linéaire à savoir : la méthode moindre carrée ordinaire et la méthode du maximum de vraisemblance.

1.1 La régression linéaire simple (RLS)

Le modèle de régression linéaire simple est une variable dépendante expliquée par une seule variable indépendante mise sous la forme suivante :

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1.1}$$

où β_0 et β_1 sont des paramètres réels inconnus, appelés les coefficients du modèle, et le ε_i est l'erreur du modèle.

1.1.1 L'écritures du modèle

1. **Le modèle théorique**(modèle non ajusté) : Le modèle théorique s'exprime comme suit :

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, ..., n.$$

2. **Le modèle estimé**(modèle ajusté) : On parlera du modèle ajusté lorsque les coefficients du modèle théorique se sont substitués par leurs estimations.

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \varepsilon_i,$$

avec

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

Ainsi on aura :

$$\varepsilon_i = Y_i - \hat{Y}_i,$$

dans ce cas ε_i représente les résidus du modèle.

La question qui se pose à ce niveau est comment peut-on estimer les paramètres du modèle ? La résolution de ce genre de problème peut se faire via plusieurs techniques à savoir : la méthode moindre carré (MCO) et la méthode maximum vraisemblance (MV).

1.1.2 Estimation des paramètres du modèle par la méthode MCO

Soit le modèle suivant :

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i. \tag{1.2}$$

L'estimation des paramètres β_0 et β_1 par l'approche MCO, dans ce cas consiste à déterminer les valeurs des deux paramètres en minimisant la somme des carrés des erreurs définie par :

$$\min \sum_{i=1}^{n} {\varepsilon_i}^2 = \min (Y_i - \beta_0 - \beta_1 X_i)^2 = \min \sum_{i=1}^{n} S^2.$$

Par conséquent, les dérivées partielles par rapport à β_0 et β_1 doivent être nuls.

$$\begin{cases} \dfrac{\partial S}{\partial \beta_0} = \quad 0 \Leftrightarrow 2 \displaystyle\sum_{i=1}^{n} (Y_i \ - \beta_0 - \beta_1 X_i)(-1) = 0. \\ \dfrac{\partial S}{\partial \beta_1} = \ 0 \Leftrightarrow 2 \displaystyle\sum_{i=1}^{n} (Y_i \ - \beta_0 - \beta_1 X_i)(-X_i) = 0. \end{cases}$$

$$\Longrightarrow \sum_{i=1}^{n} Y_i = n\beta_0 + \beta_1 \sum_{i=1}^{n} X_i. \tag{1.3}$$

$$\Longrightarrow \sum_{i=1}^{n} Y_i X_i = \beta_0 \sum_{i=1}^{n} X_i + \beta_1 \sum_{i=1}^{n} X_i^2. \tag{1.4}$$

Soient $\hat{\beta}_0$ et $\hat{\beta}_1$ les solutions du système d'équations (1.3)-(1.4), d'après l'équation (1.3) on obtient :

$$\hat{\beta}_0 = \frac{1}{n}\sum_{i=1}^{n} Y_i - \frac{1}{n}\hat{\beta}_1 \sum_{i=1}^{n} X_i.$$

C'est-à-dire,

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}.$$

En remplaçant la valeur de $\hat{\beta}_0$ dans l'équation (1.4), on obtient :

$$\sum_{i=1}^{n} Y_i X_i - \overline{Y}\sum_{i=1}^{n} X_i = \hat{\beta}_1 \left(\sum_{i=1}^{n} X_i^2 - \overline{X}\sum_{i=1}^{n} X_i \right),$$

d'où

$$\begin{aligned}\hat{\beta}_1 =& \frac{\sum_{i=1}^{n} Y_i X_i - \overline{Y}\sum_{i=1}^{n} X_i}{\sum_{i=1}^{n} X_i^2 - \overline{X}\sum_{i=1}^{n} X_i} = \frac{\sum_{i=1}^{n} Y_i X_i - n\overline{X}\ \overline{Y}}{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2} \\ =& \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})(X_i - \overline{X})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}.\end{aligned}$$

Finalement, les estimateurs des MCO du modèle de régression linéaire simple (1.2) sont :

$$\begin{cases} \hat{\beta}_0 = & \overline{Y} - \hat{\beta}_1 \overline{X}, \\ \hat{\beta}_1 = & \dfrac{S_{xy}}{S_x}, \end{cases}$$

avec

$$S_x = \sum_{i=1}^{n}(X_i - \overline{X})^2 = \sum_{i=1}^{n} X_i^2 - n\overline{X}^2,$$

$$S_{xy} = \sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) = \sum_{i=1}^{n} X_i Y_i - n\overline{X}\ \overline{Y}.$$

1.1.3 Quelques propriétés des estimateurs

Dans ce passage, on va présenter quelques hypothèses relatives à ce modèle qui nous permettrons par la suite de démontrer quelques propriétés des estimateurs en question. Considérons ce qui suit :

(H1) : Les erreurs ε_i sont centrées, ont la même variance, et non corrélées entre elles
$E(\varepsilon_i) = 0$, $E(\varepsilon_i^2) = \sigma_\varepsilon^2 < \infty$, $i = 1, ..., n$,
$cov(\varepsilon_i, \varepsilon_j) = 0$, *telque* $i \neq j$.

(H2) : Les erreurs ε_i, $i = 1, ..., n,$, sont indépendants de X, $cov(\varepsilon_i, X) = 0$.

(H3) : Les ε_i sont indépendants et identiquement distribués (iid), suit la loi normale de moyenne nulle et de variance σ_ε^2, on note $\varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon^2)$.

Sous les seules hypothèses **(H1)** et **(H2)**, il est déjà possible de préciser certaines propriétés des estimateurs des moindres carrés $\hat{\beta}_0$ et $\hat{\beta}_1$.

Théorème 1.1 (Estimateurs sans biais) *$\hat{\beta}_0$ et $\hat{\beta}_1$ sont des estimateurs sans biais de β_0 et β_1.*

$$E(\hat{\beta}_0) = \beta_0 \quad et \quad E(\hat{\beta}_1) = \beta_1.$$

Pour la démonstration, on la trouve dans [2].

Théorème 1.2 (Variances et covariances)

La matrice de variance covariance de $\hat{\beta}_0$ et $\hat{\beta}_0$ est donnée par la formule suivante :

$$\Sigma\left(\hat{\beta}_0, \hat{\beta}_1\right) = \frac{\sigma_\varepsilon^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n} X_i^2 & -\overline{X} \\ -\overline{X} & 1 \end{bmatrix},$$

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

Théorème 1.3 (Biais de l'estimateur de σ_ε^2)

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

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

avec :

$$\hat{\varepsilon}_i = Y_i - \hat{Y}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i),$$

où :

- $\hat{\varepsilon}_i$ sont les résidus du modèle qui représentes l'écart entre la valeur observée Y_i et la valeur ajustée $\hat{Y}_i$).
- $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ est la valeur ajustée (estimée) par le modèle.
- n est le nombre total d'observations.

1.1.4 Lois des estimateurs

Dans ce passage on suppose que **(H3)** est vrai : les résidus $\varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon^2)$.

Par souci de simplicité, nous prenons les symboles suivants :

$$\sigma_0^2 = Var(\hat{\beta}_0) \quad , \quad \hat{\sigma}_0^2 = \frac{\hat{\sigma}_\varepsilon^2 \sum_{i=1}^{n} X_i^2}{n \sum_{i=1}^{n} (X_i - \overline{X})^2}.$$

$$\sigma_1^2 = Var(\hat{\beta}_1) \quad , \quad \hat{\sigma}_1^2 = \frac{\hat{\sigma}_\varepsilon^2}{\sum_{i=1}^{n} (X_i - \overline{X})^2}.$$

Proposition 1

1. Lois des estimateurs si σ_ε^2 est **connue**.

   $(i) \quad \hat{\beta}_0 \sim \mathcal{N}(\beta_0, \sigma_0^2) \quad , \quad (ii) \quad \hat{\beta}_1 \sim \mathcal{N}(\beta_1, \sigma_1^2) \quad ,$

   $(iii) \quad \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} \sim \mathcal{N}_2 \left(\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \Sigma \left(\hat{\beta}_0, \hat{\beta}_1 \right) \right).$

2. Lois des estimateurs si σ_ε^2 est **inconnue**.

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

   $(i) \quad \frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}_0} \sim T_{n-2},$ où T_{n-2} la loi de Student à $(n-2)$ degrés de liberté .

   $(ii) \quad \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_1} \sim T_{n-2}.$

   $(iii) \quad \frac{1}{2} \begin{pmatrix} \frac{\hat{\beta}_0 - \beta_0}{\hat{\beta}_1 - \beta_1} \end{pmatrix}^t \hat{\Sigma}^{-1} \left(\hat{\beta}_0, \hat{\beta}_1 \right) \begin{pmatrix} \frac{\hat{\beta}_0 - \beta_0}{\hat{\beta}_1 - \beta_1} \end{pmatrix} \sim F_{(2,n-2)},$ où $F_{(2,n-2)}$ est la loi de Fisher à 2 et $(n-2)$ degré de liberté.

1.1.5 Qualité d'ajustement

Coefficient de détermination

Pour évaluer la qualité de l'ajustement du modèle, nous utilisons l'équation d'analyse de la variance à un seul facteur. Tout d'abord, nous décomposons la variance autour de la moyenne des Y_i comme suit :

$$\underbrace{\sum_{i=1}^{n}(Y_i-\overline{Y})^2}_{SCT}=\underbrace{\sum_{i=1}^{n}(\hat{Y}_i-\overline{Y})^2}_{SCE}+\underbrace{\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2}_{SCR}, \tag{1.5}$$

La qualité de l'ajustement peut être déterminée par le coefficient de détermination, qui décrit la relation linèaire entre la variation expliquée et la variation totale.
De l'équation (1.5), on peut déduire le coefficient de détermination

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

Il est evident que $(0 \leq R^2 \leq 1)$. On note que, plus la valeur de R^2 est proche de 1, plus le modèle est significatif.

1.1.6 Inférence sur les coefficients et le modèle

Test de Student (t)

Soient les hypothèses du test bilatérale suivant :

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

On ne rejette pas H_0 au seuil α si :

$$T=\left|\frac{\hat{\beta}_j}{\hat{\sigma}_{\beta_j}}\right| \leq t_{(n-2,1-\alpha/2)},$$

où :

- T : la réalisation de la statistique du test de Student.
- $t_{(n-2,1-\alpha/2)}$: fractile d'ordre $(1-\alpha/2)$ de la loi de Student à $(n-2)$ degrés de liberté.

Si cette condition ($T \leq t_{(n-2,1-\alpha/2)}$) est vérifiée, alors on ne rejette pas H_0, autrement dit, dans ce cas le paramètre B_j est significativement égale à zéro.

Test de Fisher pour la validation de modèle

Soient les hypothèses :

$$\begin{cases} H_0 : \beta_0 = 0 \quad \text{et} \quad \beta_1 = 0, \\ H_1 : \beta_0 \neq 0 \quad \text{ou} \quad \beta_1 \neq 0, \end{cases}$$

On ne rejette pas H_0 au seuil α si :

$$\frac{1}{2S^2}\left\{ n\left(\hat{\beta}_0 - \beta_0\right)^2 + 2n\overline{X}\left(\hat{\beta}_0 - \beta_0\right)\left(\hat{\beta}_1 - \beta_1\right) + \left(\hat{\beta}_1 - \beta_1\right)^2 \sum_{i=1}^{n} X_i^2 \right\} \leq F_{1-\alpha}(2, n-2),$$

où : $f_{(2,n-2,1-\alpha)}$: fractile de la loi de Fisher à 2 et $n-2$ degrés de liberté.

Dans ce cas, le modèle est jugé n'est pas significatif globalement, c'est-à-dire on je juge que le modèle linéaire proposé n'est pas adéquat pour la modélisation des données dont on dispose.

Lors de la construction du tableau d'analyse de la variance (ANOVA), on aura ce qui suit :

Source de variation	Somme des carrés	ddl	Moyenne des carrés	F
Variabilité à expliquer	SCE	1	$SCE/1 = MCE$	$F = \frac{MCE}{MCR}$
Variabilité résiduelle	SCR	$n-2$	$SCR/n-2 = MCR$	
Variabilité totale	SCT	$n-1$		

TABLE 1.1: *Table d'anova de validation du modèle.*

Ainsi l'hypothèse de signification globale du modèle est :

$$F = \frac{SCE/1}{SCR/(n-2)} > f_{(1,n-2,1-\alpha)},$$

où :$f_{(1,n-2,1-\alpha)}$ est le fractile d'ordre $1-\alpha$ de la loi de Fisher à 1 et $n-2$ degrés de liberté.

1.2 La régression linéaire multiple (RLM)

La régression linéaire multiple est une généralisation de la régression linéaire simple, dans le sens où cette approche permet d'évaluer les relations linéaires entre une variable réponse et plusieurs variables explicatives.

Définition 1.1 (Modéle de RLM) *Le modèle RLM est une généralisation du modèle linéaire simple, intégrant plusieurs variables explicatives à la fois dont l'écriture mathématique est :*

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_j x_{j,i} + ... + \beta_p x_{p,i} + \varepsilon, \ \ pour \ i = \overline{1,n}, \\ &= \beta_0 + \sum_{j=1}^{p} \beta_j x_{j,i} + \varepsilon_i, \end{aligned}$$

où :

- Y_i : est la variable à expliquer à la date i.
- $x_{j,i}$: sont les variables explicative j à la date i pour $i = \overline{1,n}$, $j = \overline{0,p}$.
- $\beta_j (j = \overline{0,p})$: sont des constantes inconnus.
- ε_i : l'erreur d'une variable aléatoire $\varepsilon = (\varepsilon_1, \varepsilon_2, ...\varepsilon_n)^t$.
- n : nombre d'observations.

Le modèle RLM peut se présenté également sous sa forme matricielle suivante :

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

avec :

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & x_{1,p} \\ 1 & x_{2,1} & x_{2,2} & x_{2,p} \\ & & & \\ 1 & x_{n,1} & x_{n,2} & x_{n,p} \end{bmatrix} \times \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$Y \quad = \quad X \quad \times \quad \beta \quad + \quad \varepsilon$$

- Y : un vecteur aléatoire de dimension n.
- X : une matrice de dimension $n \times (p+1)$ appartenant à $\mathbb{R}^n \times \mathbb{R}^{p+1}$, contient l'ensemble des observations sur les variables explicatives, avec la première colonne formée par la valeur 1 (pour le terme de la constant).
- β : un vecteur des paramètres du modèle de dimension $p+1$.
- ε : un vecteur d'erreur de dimension n.

Hypothèses du modèle

Soit les hypothèse suivantes :

(H1) Les erreurs sont centrées, de même variance, et non corrélées entre elles,
$E[\varepsilon] = 0_n, \quad \text{Var}(\varepsilon) = \sigma_\varepsilon^2 I_n, \quad$ avec I_n la matrice identité d'ordre n.

(H2) Les erreurs sont indépendantes des X_j, $cov(\varepsilon_i, Xj) = 0$, $j = \overline{1,p}$.

(H3) Les erreurs sont mutuellement indépendant, $cov(\varepsilon_i \varepsilon_{i'}) = 0$ si $i \neq i'$.

(H4) Absence de colinéarité entre les variables explicatives $\Rightarrow$ $(X'X)$ régulière et $(X'X)^{-1}$ existe.

1.2.1 Estimation des paramètres du modèle par la méthode MCO

Soit le modèle :

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

Pour estimer le vecteur β composé des coefficients $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$, nous appliquons la méthode MCO qui consiste à minimiser la somme des carrés des erreurs d'où :

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} \varepsilon_i^2 = \arg\min_{\beta} \left(Y - X\hat{\beta}\right)' \left(Y - X\hat{\beta}\right) = \arg\min_{\beta} S,$$

avec ε' est le transposé du vecteur ε.

Pour minimiser cette fonction par rapport au vecteur β, nous différencions S par rapport au même vecteur et on obtient :

$$\begin{aligned} \frac{\partial S}{\partial \beta} &= -2X'Y + 2X'X\hat{\beta} = 0, \\ &\Rightarrow \hat{\beta} = (X'X)^{-1} X'Y. \end{aligned}$$

Avec $(X'X)$ est une matrice de dimension $(p+1) \times (p+1)$, de plus sous l'hypothèse **(H4)** elle est inversible.

Le modèle estimé s'écrit :

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_{1t} + \hat{\beta}_2 x_{2t} + \ldots + \hat{\beta}_p x_{pi} + \varepsilon_i.$$

Avec : $\varepsilon_i = Y_i - \hat{Y}_i$, où ε_i (résidu) est l'écart entre la valeur observée de la variable à expliquer et sa valeur estimée, elle est connue.

1.2.2 Quelques propriétés des estimateurs

Proposition 1

L'estimateur $\hat{\beta}$ est un estimateur sans biais de β, c'est-à-dire :

$$E(\hat{\beta}) = \beta.$$

Démonstration :

Sous les l'hypothèse **(H1)** et **(H2)** on peut démontrer ce qui suit :

On a le modèle : $Y = X\beta + \varepsilon$. qui peut être écrit comme suit :

$$\begin{cases} Y = X\hat{\beta} + \varepsilon \\ \hat{Y} = X\hat{\beta} \end{cases} \quad \Rightarrow \quad \varepsilon = Y - \hat{Y}.$$

Nous obtenons :

$$\begin{aligned} \hat{\beta} &= (X'X)^{-1}X'Y = (X'X)^{-1}X'(X\beta + \varepsilon), \\ &= (X'X)^{-1}X'(X\beta) + (X'X)^{-1}X'\varepsilon, \\ &= \beta + (X'X)^{-1}X'\varepsilon \Rightarrow \hat{\beta} - \beta = (X'X)^{-1}X'\varepsilon, \end{aligned}$$

d'où

$$E(\hat{\beta}) = \beta + (X'X)^{-1}X'E(\varepsilon),$$

et le fait que $E(\varepsilon) = 0$, alors :

$$E(\hat{\beta}) = \beta.$$

La matrice de variance-covariance vaut :

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

Pour la démonstration voir [2].

Remarque

1. La matrice $X'X$ est carrée d'ordre $(p+1)$, symétrique et inversible car X est de rang $(p+1)$.
2. $X'X$ est définie positive.
3. $\hat{Y} = X\hat{\beta}$ est la valeur ajusté de Y.

Théorème 2.1

Les vecteurs $\hat{\beta}$ et $\hat{\varepsilon}$ ne sont pas corrélés entre eux.

Proposition 3

L'estimateur sans biais de la variance de l'erreur σ_ε^2, est donné par :

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

Théorème 3.1

Le vecteur des résidus $\hat{\varepsilon}$ vérifié ce qui suit :

1) $\mathbb{E}[\hat{\varepsilon}] = 0$.

2) $\mathbb{E}[\varepsilon - \hat{\varepsilon}] = 0$.

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

1.2.3 Qualité d'ajustement

Tout comme pour la régression linéaire simple, on dispose de l' égalité triangulaire suivante :

$$\underbrace{\sum_{i=1}^{n}(y_i-\overline{y})^2}_{SCT} = \underbrace{\sum_{i=1}^{n}(\hat{y}_i-\overline{y})^2}_{SCE} + \underbrace{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}_{SCR}$$
$$= \underbrace{\sum_{i=1}^{n}(\hat{y}_i-\overline{y})^2}_{SCE} + \underbrace{\sum_{i=1}^{n}\hat{\varepsilon}_i^2}_{SCR}.$$

En d'autres termes, cela signifie que : SCT= SCE + SCR, avec ;

SCT : désigne la somme des carrés totaux centrés.

SCE : la somme des carrés expliqués centrés.

SCR : la somme des carrés des résidus.

La qualité de l'ajustement peut être déterminée par le coefficient de détermination R^2 qui mesure la proportion de la variance de y expliquée par la régression linèaire de y sur x.

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

Le coefficient de détermination corrigé est calculé comme suit :

$$\overline{R^2} = 1 - \frac{n-1}{n-p-1}\left(1 - R^2\right), \text{ si } (n < p).$$

Notons que : $\overline{R^2} \leq R^2$. Si n est grand $\overline{R^2} \simeq R^2$.

1.2.4 Lois des Estimateurs

Considérons le modèle de régression linéaire multiple suivant :

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

Nous supposons que les erreurs suivent une distribution normale(**H5**), et on note :

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

où $\mathcal{N}_n(.,.)$ représente une loi normale multivariée dans $\mathbb{R}^n$. Cette hypothèse implique que les erreurs sont indépendantes et identiquement distribuées (i.i.d.).
Par conséquent, la variable Y suit également une loi normale dans $\mathbb{R}^n$:

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

Théorème 4.1

Sous l'hypothèse (**H5**) on a :

1) L'estimateur des coefficients $\hat{\beta}$ suit une loi normale multivariée :

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

2) L'estimateur de la variance des erreurs normalisé suit une loi du Khi-deux :

$$(n-p-1)\frac{\hat{\sigma}_\varepsilon^2}{\sigma_\varepsilon^2} \sim \chi^2_{n-p-1}.$$

1.2.5 Signification des coefficients et validation de modèle

Le test de Student (t)

Le test de Student permet d'évaluer l'influence directe d'une variable explicative sur la variable à expliquer. Il revient à tester si le coefficient de régression est significativement égale à 0, pour un seuil de signification α choisi.
L'hypothèse nulle et l'hypothèse alternative pour $(i = \overline{0,p})$ de test est bilatéral suivant, sont données comme suit :

$$\begin{cases} H_0 : \beta_i = 0, \\ \text{contre} \\ H_1 : \beta_i \neq 0. \end{cases}$$

La statistique de Student est donnée par :

$$T = \left| \frac{\hat{\beta}_i - \beta_i}{\partial(\hat{\beta}_i)} \right| \sim t_{(n-p-1,1-\frac{\alpha}{2})}$$

avec

- T : la valeur calculée.
- $t_{(n-p-1,1-\frac{\alpha}{2})}$: la valeur du quantile d'ordre $(1 - \frac{\alpha}{2})$ d'une loi de Student de *ddl* $(n - p - 1)$.

Règle de décision : Si $|T| \leq t_{(n-p-1,1-\frac{\alpha}{2})}$, on ne rejette pas l'hypothèse H_0, ce qui signifie que la variable x_i n'a pas une influence significative dans l'explication de Y.

Le test de Fisher pour la validation du modèle

Le test de Fisher est un test de signification globale du modèle de regression et pour tester si l'ensemble des variables explicatives ont une influence sur la variable à expliquer, on fait le test qui se formule comme suit :

$$\begin{cases} H_0 : \beta_0 = \beta_1 = ... = \beta_p = 0, \\ \text{contre} \\ H_1 : \exists i \in \{0, 1, .., p\} \text{ tq } \beta_i \neq 0. \end{cases}$$

Sous l'hypothèse H_0 on aura :

$$F_C = \frac{\sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n} \varepsilon_i^2} \times \frac{p}{n-p-1}$$

$$= \frac{\chi_p^2}{\chi_{n-p-1}^2} \times \frac{p}{n-p-1}$$

$$= \frac{\frac{R^2}{p}}{\frac{1-R^2}{n-p-1}} \sim F_{(p,n-p-1,1-\alpha)}$$

Règle de décision : Si $F_C > F_{(p,n-p-1,1-\alpha)} \Rightarrow$ on rejette H_0, et on considère que le modèle est globalement significatif.

Les déférents calcules associes au test Fisher pour la validation du modèle peut être résumé dans le tableau d'ANOVA suivant

Source de variation	Somme des carrés	ddl	Moyenne des carrés	F
Variabilité à expliquer	SCE	p	$SCE/p = MCE$	$F = \frac{MCE}{MCR}$
Variabilité résiduelle	SCR	$n-p-1$	$SCR/n-p-1 = MCR$	
Variabilité totale	SCT	$n-1$		

TABLE 1.2: *Table d' ANOVA pour validation du modèle.*

1.3 Méthode du maximum de vraisemblance (MV)

L'estimation des paramètres est une étape essentielle dans la construction d'un modèle de régression fiable et interpretable. Parmi les méthodes les plus utilisées à cette fin, on trouve la méthode du maximum de vraisemblance. La méthode du maximum de vraisemblance est une méthode d'estimation paramétrique qui doit sa popularité à :

- La simplicité de son approche.
- Sa faculté d'adaptation à une modélisation complexe.
- L'aspect numérique accessible grâce à l'application de méthodes d'optimisation connues.
- Construire des estimateurs performants .
- Construire des intervalles de confiance précis .

Dans ce qui suit, nous allons repeller d'abord la notion de l'estimateur du maximum de vraisemblance ensuite nous allons l'adapter au cas d'estimations des paramètre d'un modèle linéaire multiple.

1.3.1 Notation et principe de la méthode

Le principe du maximum de vraisemblance conduit au choix de l'estimateur $\hat{\beta}$, où β symbolise le vecteur des paramètres inconnus : $\beta = (\beta_0, \beta_1, \ldots, \beta_1)^t$, comme la valeur du paramètre qui rend les données observées les plus probables.
Les **étapes d'estimation** sont données comme suit :

1. Sélectionner la loi F des erreurs selon la distribution empirique des résidus.
2. Écrire la fonction de vraisemblance adaptée à la distribution F.
3. Maximiser la log-vraisemblance.
4. Estimer le paramètre β.

Définition 1.2 *On appelle* ***fonction de vraisemblance*** *de β pour une réalisation $(x_1, \ldots, x_n)$ d'un échantillon, la fonction de β :*

$$L(x_1, \ldots, x_n; \beta) = f(x_1, \ldots, x_n; \beta) = \prod_{i=1}^{n} f(x_i; \beta),$$

avec f est la densité de la variable aléatoire X.

Définition 1.3 *La fonction de vraisemblance est la fonction de densité considérée comme une fonction de β.*

$$L(\beta|x) = f(x|\beta).$$

L'estimateur du maximum de vraisemblance (EMV) est donné comme suit :

$$\hat{\beta}(x) = \arg\max_{\beta} L(\beta|x).$$

Fonction de Log-vraisemblance :
On appelle fonction de log-vraisemblance pour $(x_1, \ldots, x_n)$ la fonction de β définie par :

$$\ell_n(x_1, \ldots, x_n; \beta) = \ln(L_n(x_1, \ldots, x_n; \beta)).$$

La fonction logarithme népérien étant croissante, l'*EMV* β^* de β pour $(x_1, \ldots, x_n)$ vérifie :

$$\beta^* = \arg\max_{\beta} L_n(x_1, \ldots, x_n; \beta) = \arg\max_{\beta} \ell_n(x_1, \ldots, x_n; \beta).$$

Équation de vraisemblance est donnée comme suit :

$$\frac{\partial}{\partial \beta} \ell_n(x_1, \ldots, x_n; \beta) = 0.$$

Dans ce qui suit, nous allons examiner quelques exemples de la fonction de vraisemblance selon la distribution des erreurs.

1.3.2 Hypothèse classique (loi normale)

Dans un cadre classique on suppose que $\varepsilon_i \sim N(0, \sigma^2)$ et σ^2 est inconnue. Ainsi on aura dans ce qui suit :

La Fonction de Vraisemblance : La densité jointe de Y est donnée par :

$$f(Y; \beta, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}(Y - X\beta)^T(Y - X\beta)\right). \tag{1.6}$$

On obtient alors la fonction de **log-vraisemblance** suivante :

$$\ell(\beta, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(Y - X\beta)^T(Y - X\beta) \tag{1.7}$$

Après derivation de l'équation (1.7) et l'égalisation du résultat à zéro on peut verifier que l'estimation des paramètres par la méthode de maximum de vraisemblance sont définie comme suit :

- **Estimation de** β :

$$\hat{\beta}_{MV} = (X^T X)^{-1} X^T Y$$

- **Estimation de** σ^2 :

$$\hat{\sigma}^2_{MV} = \frac{1}{n}(Y - X\hat{\beta})^T(Y - X\hat{\beta})$$

La méthode du maximum de vraisemblance (MV) peut être adaptée dans le cas où les résidus (ε_i) ne suit pas une loi normale, en remplaçant l'hypothèse de normalité par une autre loi de probabilité adaptée aux données. Ci-dessous quelque exemples de lois :

1.3.3 Cas des erreurs suivant une loi de Laplace

Dans ce cas, on suppose que les erreurs suivent une loi de Laplace (appeler aussi loi doublement exponentielle), de densité :

$$f(\varepsilon) = \frac{1}{2b} \exp\left(-\frac{|\varepsilon|}{b}\right)$$

où $b > 0$ est un paramètre d'échelle. Cette loi est plus robuste aux valeurs extrêmes que la loi normale.

Sous l'hypothèse d'indépendance des erreurs, la fonction de vraisemblance du modèle est :

$$L(\beta, b) = \prod_{i=1}^{n} \frac{1}{2b} \exp\left(-\frac{|y_i - x_i^T \beta|}{b}\right)$$

La log-vraisemblance est donc :

$$\ell(\beta, b) = -n \log(2b) - \frac{1}{b} \sum_{i=1}^{n} |y_i - x_i^T \beta|$$

L'estimateur du maximum de vraisemblance de β est celui qui minimise la somme des résidus absolus :

$$\hat{\beta}_{\text{MV}} = \arg\min_{\beta} \sum_{i=1}^{n} |y_i - x_i^T \beta|$$

Ce type d'estimation correspond aussi à la régression des moindres valeurs absolues (LAD).

1.3.4 Cas des erreurs suivant une loi de Student(t)

La densité de probabilité de la loi de Student centrée et réduite est définie par :

$$f(\varepsilon) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{\varepsilon^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

Dans notre cas, on suppose que :

$$\Rightarrow f(y_i|x_i) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\nu\pi\sigma^2}} \left(1 + \frac{(y_i - x_i^T\beta)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}}$$

Sous l'hypothèse d'indépendance des erreurs, la fonction de log-vraisemblance est donnée comme suit :

$$\begin{aligned}\ell(\beta,\sigma) &= \sum_{i=1}^{n}\log f(y_i|x_i,\beta,\sigma),\\ &= -\frac{n}{2}\log(\nu\pi\sigma^2)+\sum_{i=1}^{n}\log\Gamma\left(\frac{\nu+1}{2}\right)-\log\Gamma\left(\frac{\nu}{2}\right)-\frac{\nu+1}{2}\sum_{i=1}^{n}\log\left(1+\frac{(y_i-x_i^T\beta)^2}{\nu\sigma^2}\right).\end{aligned}$$

- La loi de Student permet de gérer les observations aberrantes grâce à ses queues épaisses. En effet la densité est plus aplatie que la normale, ce qui rend le modèle plus robuste aux extrêmes.
- Lorsque $\nu \to \infty$, la loi de Student converge vers la loi normale, et on retrouve le cas classique de la régression linéaire gaussienne.
- La vraisemblance est plus complexe, mais peut être maximisée numériquement (via des méthodes comme l'algorithme EM, Newton-Raphson, etc.).

1.3.5 Cas de loi des erreurs et de la famille exponentielle

La méthode du maximum de vraisemblance peut être appliquée à toute loi appartenant à la famille exponentielle : exponentielle, binomiale, poisson, etc. On suppose que les variables aléatoires Y_i conditionnelles à X_i suivent une loi appartenant à la **famille exponentielle**, de la forme :

$$f(y_i|\theta_i,\phi)=\exp\left\{\frac{y_i\theta_i-b(\theta_i)}{a(\phi)}+c(y_i,\phi)\right\}$$

où :

- θ_i est le paramètre naturel (lié à la moyenne μ_i),
- ϕ est un paramètre de dispersion,
- $a(\cdot)$, $b(\cdot)$ et $c(\cdot,\cdot)$ sont des fonctions déterminées par la loi f (selon la forme de f).

On lie la moyenne $\mu_i = \mathbb{E}[Y_i]$ à une combinaison linéaire des variables explicatives :

$$g(\mu_i) = x_i^T\beta$$

où $g(\cdot)$ est la fonction de lien, et x_i le vecteur de covariables pour l'observation i.

La fonction de log-vraisemblance : Sous l'indépendance conditionnelle des observations :

$$\ell(\beta,\phi)=\sum_{i=1}^{n}\log f(y_i|x_i,\beta,\phi)=\sum_{i=1}^{n}\left[\frac{y_i\theta_i-b(\theta_i)}{a(\phi)}+c(y_i,\phi)\right]$$

La dérivée par rapport à β (score) est :

$$\frac{\partial \ell}{\partial \beta} = \sum_{i=1}^{n} \left(\frac{y_i - \mu_i}{\mathrm{Var}(Y_i)} \cdot \frac{d\mu_i}{d\eta_i} \cdot x_i \right) \tag{1.8}$$

où $\eta_i = x_i^T \beta$ est le prédicateur linéaire.

D'après (1.8) il est claire que l'estimation de β ne peut se faire que par les méthode numérique, ainsi on peut utiliser la méthode du score, algorithme de Newton-Raphson, la méthode du point-fixe ou encore la méthode IRLS (Iteratively Reweighted Least Squares).

1.3.6 Choix pratique : MCO Vs Maximum de Vraisemblance

Dans la pratique afin de choisir entre les méthodes d'estimation des paramètre dans la régression linéaire(MCO et MV)on doit procéder comme suit :

- Si les erreurs sont normales, les deux méthodes donnent le même estimateur pour β, donc MCO est suffisant.
- Pour des modèles plus complexes (non normalité des résidus, régression logistique, modèles mixtes), le MV devient préférable.

Nous apprendrons que, notamment pour les grands échantillons, les estimateurs du maximum de vraisemblance possèdent de nombreuses propriétés intéressantes. Cependant, dans le cas de nombre de variables explicatives est trop grand, balayer tous les modèles peut se révéler très coûteux en tant de calcul, la fonction de vraisemblance peut avoir de nombreux maximums locaux. Ainsi, trouver le maximum global peut être un défi computational majeur.

Conclusion

Lorsque le nombre de variables explicatives devient important, une recherche exhaustive de tous les modèles possibles devient rapidement envisageable en raison du coût computational élevé. On privilégie alors les méthodes de sélection de modèle de régression linéaire multiple comme la méthode de sélection Pas-à-Pas, qui consistent à construire les modèles de manière récursifs en ajoutant ou supprimant une variable à chaque étape. La méthode de selection Pas-à-Pas fera l'objet de chapitre suivant.

Chapitre 2

Stepwise pour la sélection des modèles en RLM

Dans ce chapitre nous présentons la méthode pas-à-pas (Stepwise), qui est une technique de sélection de variables utilisée en statistique et en modélisation. Elle sert principalement à choisir les variables les plus pertinentes dans un modèle de régression. Cette méthode repose sur l'ajout ou la suppression progressive des variables selon des critères statistiques précis, ce qui permet d'améliorer la précision du modèle.

2.1 Méthode pas-à-pas (Stepwise)

La régression pas-à-pas (Stepwise Regression) est une méthode statistique itérative utilisée pour sélectionner les variables indépendantes les plus pertinentes dans un modèle de régression linéaire multiple. Cette sélection se fait de manière progressive en ajoutant ou en supprimant les variables potentielles en fonction des résultats de test de leur signification statistique après chaque étape. Le processus commence soit par l'ajout des variables les plus significatives, soit par la suppression des variables ayant le moins d'impact sur le modèle, et se poursuit jusqu'à obtenir un ensemble de variables qui offrent la meilleure adéquation au modèle.

Cette méthode est fréquemment utilisée lorsqu'on travaille avec des ensembles de données comportant de nombreuses variables, ce qui la rend idéale pour réduire le nombre de variables sélectionnées et améliorer la précision du modèle. La disponibilité des logiciels statistiques modernes permet d'appliquer cette méthode à de très grands

ensembles de données, même lorsque ceux-ci contiennent des centaines de variables.

2.1.1 Le principe de la méthode

L'objectif principal de la régression pas-à-pas est, à travers une série de tests (par exemple, les tests F et t), de trouver un ensemble de variables indépendantes qui influencent significativement la variable dépendante. Cela se fait à l'aide d'ordinateurs grâce à l'itération, qui est le processus consistant à obtenir des résultats ou à prendre des décisions en passant par des cycles répétés d'analyse. La réalisation automatique des tests avec l'aide de logiciels statistiques présente l'avantage de gagner du temps et de limiter les erreurs.

2.1.2 Les différentes approches de la méthode pas-à-pas

La régression pas-à-pas correspond à une méthode de sélection de variables permettant d'optimiser un modèle de régression en fonction de critères statistiques. Trois approches principales sont distinguées :

- La sélection ascendante (Forward Selection)
- L'élimination descendante (Backward Elimination)
- L'élimination bidirectionnelle (Bidirectional Elimination)

Dans ce qui suit nous allons présenter le principe de chaque approche.

2.1.3 Sélection ascendante (Forward Selection)

Cette méthode est souvent utilisée pour effectuer un premier tri des variables candidates lorsqu'un grand nombre de variables explicatives est disponible. Par exemple, supposons que nous avons un grand nombre de variable revendicatives (par exemple entre cinquante et cent variables à examiner), ce qui dépasse largement le champ d'application de la méthode de toutes les régressions possibles (all-possible regressions).
Une approche raisonnable consisterait à utiliser cette procédure de sélection progressive pour identifier les dix à quinze meilleures variables, puis à appliquer l'algorithme de toutes les régressions possibles uniquement à ce sous-ensemble réduit. Cette méthode

est également pertinente en cas de multicolinéarité, c'est-à-dire lorsque certaines variables sont fortement corrélées entre elles. La méthode est simple à définir :

- On commence sans aucune variable dans le modèle.
- On insère à chaque étape la variable qui entraîne la plus forte amélioration du critère R^2.
- Le processus s'arrête dès qu'aucune variable n'apporte d'amélioration significative.

Remarque importante : Il est à noter que, dans cette approche, une fois une variable est introduite dans le modèle, elle ne peut plus être supprimée dans les prochaines itérations.

2.1.4 Élimination descendante (méthode backward)

La sélection Backward est une méthode utilisée pour la sélection des variables explicatives dans les modèles statistiques. Elle consiste à démarrer avec un modèle initial contenant l'ensemble des variables candidates, puis à éliminer, à chaque étape, la variable la moins significative sur le plan statistique. Ce processus se poursuit de manière itérative jusqu'à ce que le modèle ne contienne plus que des variables statistiquement significatives, selon le seuil de signification et du critère de sélection retenu fixés préalablement par le chercheur(généralement le seuil est fixé à 0.05 ou 0.10).
L'un des principaux avantages de cette approche est qu'elle permet de maintenir une valeur élevée du coefficient de détermination R^2, ce qui reflète une bonne capacité explicative du modèle. Cependant, cette méthode peut conduire à la rétention de variables non pertinentes, que ce soit d'un point de vue statistique ou théorique, ce qui peut nuire à la parcimonie du modèle et compliquer l'interprétation des résultats.
Les étapes de la méthode sont simple à définir :

- On commence par le modèle complet.
- A chaque étape, on supprime la variable dont le retrait diminue le critère de sélection.
- Le processus s'arrête dès qu'aucune suppression n'améliore davantage le critère.

Remarque importante : Il est à noter que, dans cette approche, une fois une variable est supprimer du modèle, elle ne peut plus être ajouté dans les prochaines itérations.

2.1.5 Sélection par étapes (méthode stepwise bidirectionnelle)

La sélection par paliers est une méthode hybride combinant la sélection ascendante et descendante à la fois. A chaque ajout de variable significative, le modèle vérifie si certaines variables précédemment incluses deviennent non significatives ; le cas échéant, elles sont retirées. Cette procédure nécessite deux seuils de signification : un pour l'ajout, plus strict, et un autre pour le retrait, afin d'éviter les boucles d'itération infinies. Bien qu'elle ait été populaire, elle est aujourd'hui souvent remplacée par des méthodes multivariées plus performantes.

Méthode ascendante bidirectionnelle (Bidirectional Selection)

- Il s'agit d'une méthode ascendante qui, à chaque étape, réévalue l'utilité des variables déjà incluses.
- Cela permet d'exclure des variables qui, en présence de nouvelles variables, deviennent non significatives.

Méthode descendante bidirectionnelle (Bidirectional Elimination)

- Méthode descendante avec réintégration possible des variables exclues précédemment.
- Une variable retirée peut redevenir significative après la suppression d'une autre variable et ainsi être réintégrée dans le modèle.

2.1.6 Performances des modèles dans la méthode pas-à-pas (Stepwise)

Lors de l'application de la méthode de sélection pas-à-pas (stepwise), il est essentiel de disposer de critères d'évaluation fiables pour choisir le modèle de régression le plus performant. Ces critères permettent de comparer les modèles successifs en tenant compte à la fois de leur précision et de leur complexité. Les principaux indicateurs de performance d'un modèle sont les suivants :

- Le coefficient de détermination ajusté ($\overline{R^2}$).
- La somme des carrés des résidus (RSS).
- L'erreur quadratique moyenne (MSE).

- La racine de l'erreur quadratique moyenne ($RMSE$).
- Et le critère d'information d'Akaike pour échantillons finis (AIC), souvent utilisé comme critère principal dans la méthode stepwise.

Les formules mathématiques correspondantes sont données ci-dessous :

$$\begin{aligned} R^2 &= 1 - \frac{n-1}{n-p-1}\left(1-R^2\right) \\ RSS &= \sum_{i=1}^{n} \varepsilon_i^2 \end{aligned}$$

$$\begin{aligned} MSE &= \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2 \\ RMSE &= \sqrt{\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2} \\ AIC &= 2p - 2\ln(L) + \frac{2p(p+1)}{n-p-1} \end{aligned}$$

où :

- n est le nombre d'observations.
- p est le nombre de variables explicatives.
- ε_i représente l'erreur résiduelle pour chaque observation i.
- $\bar{y}$ est la moyenne des valeurs observées.
- L est la valeur maximale de la vraisemblance du modèle.

2.2 Exemple illustratif de la méthode Stepwise : cas de la méthode backward

2.2.1 Présentation de l'exemple

Pour illustrer l'utilisation de la méthode de **régression pas-à-pas (stepwise)**, on considère un jeu de données simulé portant sur cinq tests cognitifs réalisés par un groupe d'individus. L'objectif de l'analyse est de prédire le **QI (quotient intellectuel)** à partir des résultats obtenus dans ces tests.

La variable dépendante est donc le **QI**, tandis que les variables explicatives sont les scores des tests notés **Test1 à Test5**.
Supposons que les données dont on dispose sont comme suite (Pour plus d'informations sur l'exemple et les données le lecteur peut se référer à [11]) :

Individu	**Test1**	**Test2**	**Test3**	**Test4**	**Test5**	**QI**
1	60	45	70	55	61	98
2	72	33	66	58	74	104
3	55	67	78	63	70	101
4	91	76	89	85	93	120
5	68	49	71	60	69	99
6	48	37	50	53	64	90

TABLE 2.1: *Tableau des données simulées*

On assume que l'objectif est de construire le modèle linéaire correspondant aux données tout en utilisant la méthode de stepwise. Il est à rappeler que dans se cas La méthode va permettre de sélectionner automatiquement les variables les plus significatives, en les ajoutant ou en les retirant une à une, selon leur contribution statistique à la qualité du modèle de régression.
Mais avant de procéder au calcul, il est à souligner que lors de la mise en œuvre de la méthode la méthode stepwise sous R les résultats se présente sous forme de tableaux dont le nombre égale aux nombre d'itérations. De plus, chaque tableau contient les indicateur suivants :

- In : indique si la variable est incluse (*Oui*) ou exclue (*Non*) du modèle.
- Variable : nom de la variable explicative analysée.
- Coefficient Standardisé : mesure l'effet de la variable sur la variable dépendante, après standardisation. Cette standardisation consiste à soustraire la moyenne et diviser par l'écart-type. La formule utilisée est :

$$b_{j,std} = b_j \times \left(\frac{s_{x_j}}{s_y} \right)$$

où s_{x_j} et s_y sont respectivement les écarts-types de la variable explicative x_j et de la variable dépendante y.

- R^2 Incremental : contribution marginale de la variable au coefficient de détermination R^2. Une valeur élevée reflète une variable explicative pertinente.
- R^2 vs autres X : mesure de colinéarité entre la variable considérée et les autres variables du modèle. Des valeurs élevées peuvent indiquer une redondance.
- Valeur de T (T-Value) : permet de tester si la variable a une contribution significative. Plus cette valeur est élevée, plus l'effet est important.
- P-valeur : probabilité associée à la T-Value. Une p-valeur faible (souvent < 0.05) indique une significativité statistique.
- % Changement dans la Racine Carrée du MSE : indique, en pourcentage, l'amélioration ou la détérioration de la performance du modèle (RMSE) suite à l'ajout ou la suppression d'une variable. Le calcul se fait selon la formule :

$$\text{Pourcentage de changement} = \left(\frac{RMSE_{\text{précédent}} - RMSE_{\text{actuel}}}{RMSE_{\text{actuel}}} \right) \times 100$$

- R^2 : coefficient de détermination global du modèle à cette itération.
- Racine Carrée du MSE : mesure l'erreur quadratique moyenne du modèle sous forme de racine carrée.

2.2.2 Résultats de la régression stepwise (méthode backward)

Une analyse de régression linéaire pas-à-pas par l'approche backward a été réalisée à l'aide du logiciel R afin d'identifier les tests cognitifs les plus significatifs pour prédire le QI. En partant d'un modèle complet incluant les cinq tests (Test1 à Test5), la méthode a progressivement éliminé les variables non significatives. En effet, suite à l'analyse du modèle de régression complet présenté dans le tableau 2.2, une démarche de sélection pas-à-pas des variables explicatives a été entreprise selon la méthode backward. Cette approche consiste à éliminer successivement les variables les moins significatives, en évaluant à chaque étape l'impact de cette suppression sur la qualité du modèle vis-à-vis le critère $RMSE$.

Pour ce dernier critère, a chaque étape, la variable dont le changement (entrée ou sortie du modèle) réduit le plus le MSE est sélectionnée et son statut est inversé. Le

processus continue jusqu'à ce qu'aucune variable n'apporte une réduction plus grande que le seuil minimal spécifié.

Les tableaux 2.2–2.5 présentent les détails de chaque itération de la procédure de régression stepwise selon la méthode "backward".

Détails des Itérations

Itération 0 : Inchangée

In	Var	Coeff Stand	R-Carré Incrémental	vs. autres X	T-Valeur	P-Valeur	% Changement dans Rac(MSE)
Oui	Test1	-3.0524	0.235717	0.974701	-1.8789	0.092969	11.9387
Oui	Test2	-2.9224	0.241444	0.971730	-1.9016	0.089661	12.3210
Oui	Test3	0.1404	0.015210	0.227987	0.4773	0.644541	-3.9386
Oui	Test4	4.7853	0.283243	0.987631	2.0596	0.069522	15.0741
Oui	Test5	-0.0595	0.002715	0.232860	-0.2017	0.844669	-4.9176
R-Carré = 0.399068 Rac(MSE) = 10.65198							

TABLE 2.2: *Tableau de modèle initial sans suppression de variables*

Itération 1 : Test5 retiré du Modèle

In	Var	Coeff Stand	R-Carré Incrémental	vs. autres X	T-Valeur	P-Valeur	% Changement dans Rac(MSE)
Oui	Test1	-3.0612	0.237250	0.974683	-1.9825	0.075558	12.5340
Oui	Test2	-2.9032	0.239195	0.971621	-1.9906	0.074546	12.6640
Oui	Test3	0.1163	0.012499	0.075203	0.4550	0.658798	-3.6717
Oui	Test4	4.7850	0.283206	0.987631	2.1660	0.055543	15.5681
Non	Test5	-	0.002715	0.232860	0.2017	0.844669	5.1719
R-Carré = 0.396353 Rac(MSE) = 10.12816							

TABLE 2.3: *Tableau de suppression de la variable Test5 du modèle*

Itération 2 : Test3 retiré du modèle

In	Var	Coeff Stand	R-Carré Incrémental	vs. autres X	T-Valeur	P-Valeur	% Changement dans Rac(MSE)
Oui	Test1	-3.1020	0.244443	0.974597	-2.0890	0.060743	13.1519
Oui	Test2	-2.9024	0.239064	0.971621	-2.0659	0.063218	12.7977
Oui	Test4	4.7988	0.284897	0.987628	2.2553	0.045468	15.7808
Non	Test3	–	0.012499	0.075203	0.4550	0.658798	3.8116
Non	Test5	–	0.000005	0.081040	0.0087	0.993205	4.8805
R-Carré = 0.383854 Rac(MSE) = 9.756291							

TABLE 2.4: *Tableau de suppression de la variable Test3 du modèle*

Itération 3 : Inchangée

In	Var	Coeff Stand	R-Carré Incrémental	vs. autres X	T-Valeur	P-Valeur	% Changement dans Rac(MSE)
Oui	Test1	-3.1020	0.244443	0.974597	-2.0890	0.060743	13.1519
Oui	Test2	-2.9024	0.239064	0.971621	-2.0659	0.063218	12.7977
Oui	Test4	4.7988	0.284897	0.987628	2.2553	0.045468	15.7808
Non	Test3	–	0.012499	0.075203	0.4550	0.658798	3.8116
Non	Test5	–	0.000005	0.081040	0.0087	0.993205	4.8805
R-Carré = 0.383854 Rac(MSE) = 9.756291							

TABLE 2.5: *Tableau de modèle final – inchangé*

2.2.3 Discussion et Interprétation des résultats

À travers cette procédure de régression linéaire stepwise, nous avons pu identifier les variables explicatives les plus déterminantes dans la prédiction du QI. Le modèle final, obtenu après trois itérations, conserve uniquement les variables Test1, Test2 et Test4, celles-ci présentant une contribution statistiquement significative et une cohérence avec l'interprétation théorique des résultats. Autrement dit, le modèle final retenu comprend uniquement les variables Test1, Test2 et Test4, qui présentent une contribution statistique notable à la prédiction du QI. Cette approche permet ainsi d'obtenir un modèle plus simple, mais avec une bonne capacité explicative dont la forme est :

$$y = \beta_0 + \beta_1 Test1 + \beta_2 Test2 + \beta_4 Test4 + \epsilon.$$

2.3 Les limites de la régression pas-à-pas

La régression pas-à-pas, bien qu'utilisée pour la sélection automatique des variables, présente plusieurs inconvénients :

- Biais de sélection : elle favorise les variables ayant plusieurs modalités, même si elles n'améliorent pas le modèle.
- Sur-apprentissage : elle peut produire un modèle trop adapté aux données d'entraînement, mais inefficace sur de nouvelles données.
- Instabilité : de petits changements dans les données peuvent entraîner des choix de variables très différents, rendant le modèle peu fiable.

Ci-dessous un tableau comparatif qui mis en évidence les avantages et les inconvénients des trois approches de la méthode pas-à-pas pour la sélection d'un modèle dans la régression linéaire multiple.

Méthode	**Principe**	**Critère utilisé**	**Avantages**	**Inconvénients**
Sélection ascendante (Forward Selection)	On commence sans variable et on ajoute progressivement celles qui sont significatives	R^2 ou p-value	Simple, utile avec beaucoup de variables explicatives	Une fois une variable ajoutée, elle ne peut plus être retirée
Élimination descendante (méthode backward)	On commence avec toutes les variables, on retire les moins significatives	p-value	Garde les variables les plus statistiquement pertinentes	Peut garder des variables non pertinentes ou multicolinéaires
Sélection par étapes (méthode stepwise bidirectionnelle)	Ajout et suppression de variables à chaque étape selon leur contribution	Deux seuils (entrée/sortie)	Méthode flexible, combine les avantages des deux précédentes	Plus complexe, nécessite une bonne définition des seuils

TABLE 2.6: *Tableau comparatif des méthodes de sélection des variables*

Conclusion

Dans ce chapitre nous avons exposé les trois approche de la méthode Pas-à-Pas (stepwise) pour la sélection du modèle linéaire, où nous avons focalisé sur son principe, ses avantages ainsi ses inconvénients. Dans le chapitre suivant, nous allons exposer un exemple plus détaillé et approfondie sur l'application de la méthode stepwise avec ses trois approches afin de mieux comprendre leurs mécanisme de fonctionnement.

Chapitre 3

Application numérique

L'objectif visé dans ce chapitre est d'étudier et comprendre d'avantage le mécanisme des techniques de la régression linéaire exposées dans les chapitres précédents à savoir :

- Deux méthodes d'estimation en régression linéaire multiple :

  La méthode des Moindres Carrés Ordinaires (MCO) et la méthode du Maximum de Vraisemblance (MV).

- Trois méthodes de sélection de modèle de régression linéaire multiple Pas-à-Pas on cite : sélection forward, sélection Backward et sélection Mixte.

L'étude est menée sur un jeu de données simulé dans lequel on a supposé que les résidus ne suivent pas une loi normale, afin d'observer l'impact de cette hypothèse sur les résultats fournis par chacune des trois méthodes en question.

3.1 Présentation de l'exemple numérique :

Notre exemple traite une étude de l'influence de facteurs socio-économiques(revenu, diplôme, age, heures-travail, stress) sur les performances d'un individu(Cette performance peut être mesurée de différentes manières, comme la productivité au travail, les résultats scolaires, l'engagement dans des activités de loisirs, ou encore l'état de santé général).

Présentation des variables :

1. score (quantitative) : score de performance (variable dépendante).
2. revenu : revenu mensuel en dinar.

3. diplôme : niveau de diplôme (Bac, Licence, Master, Doctorat) codé 1 à 4.
4. age : âge en années.
5. heures-travail : nombre d'heures de travail par semaine.
6. stress : niveau de stress (échelle 0-10).

Formulation du modèle

Supposons que le modèle proposé pour l'analyse de l'influence de facteurs socio-économiques sur les performances d'un individu est le modèle linéaire multiple suivant :

$$\text{score} = \beta_0 + \beta_1 \cdot \text{revenu} + \beta_2 \cdot \text{diplôme} + \beta_3 \cdot \text{âge} + \beta_4 \cdot \text{heures\_travail} + \beta_5 \cdot \text{stress} + \varepsilon. \quad (3.1)$$

Où ε représente l'erreur aléatoire.

Dans ce qui suit on va analyser le modèle (3.1) en utilisant les trois méthodes de régression proposée dans ce document(MCO, MV, Pas-à-Pas).

Pour l'application numérique nous avons fixé les paramètres de modèle (3.1) comme suit :

$$\text{y} = 50 + 0.1x_1 + 1.5x_2 - 0.3x_3 + 0.01x_4 - 0.8x_5 + \varepsilon. \quad (3.2)$$

Où ε suit une loi exponentielle centrée de moyen 10.

Notons que pour les différents calcul numérique on a fait recours aux logiciel R.

La figure 3.1 présente le code source du programme sous R dans le but est de générer un échantillon de taille n correspondant à l'exemple (3.2).

```
> set.seed(123)
> n <- 150
> revenu <- rnorm(n, mean = 2500, sd = 600)
> diplome <- sample(1:4, n, replace = TRUE)
> age <- rnorm(n, mean = 35, sd = 10)
> heures_travail <- rnorm(n, mean = 40, sd = 5)
> stress <- runif(n, min = 0, max = 10)
>
> # Erreurs non normales (distribution exponentielle centrée)
> erreur <- rexp(n, rate = 1/10) - 10  # moyenne ? 0 pour briser la normalité
>
> # Variable dépendante avec combinaison linéaire
> score <- 50 + 0.01 * revenu + 1.5 * diplome - 0.3 * age +
+          0.01 * heures_travail - 0.8 * stress + erreur
>
> donnees <- data.frame(score, revenu, diplome, age, heures_travail, stress)
```

FIGURE 3.1: *Simulation des données sur R*

3.1.1 Cas : méthode des Moindres Carrés Ordinaires (MCO)

Dans ce passage le code source qui montre les étapes et les résultats numériques de la méthode (MCO)sur R :

```
> modele_mco <- lm(score ~ revenu + diplome + age + heures_travail + stress, data = donnees)
> summary(modele_mco)

Call:
lm(formula = score ~ revenu + diplome + age + heures_travail +
    stress, data = donnees)

Residuals:
    Min      1Q  Median      3Q     Max
-13.665  -6.633  -2.216   2.769  51.054

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    51.764285   9.328976   5.549 1.34e-07 ***
revenu          0.008568   0.001577   5.434 2.30e-07 ***
diplome         3.074888   0.805731   3.816 0.000201 ***
age            -0.222603   0.088479  -2.516 0.012971 *
heures_travail -0.006664   0.186182  -0.036 0.971498
stress         -1.574483   0.317944  -4.952 2.03e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.84 on 144 degrees of freedom
Multiple R-squared:  0.3681,    Adjusted R-squared:  0.3461
F-statistic: 16.77 on 5 and 144 DF,  p-value: 4.85e-13
```

FIGURE 3.2: *Régression par MCO sous R*

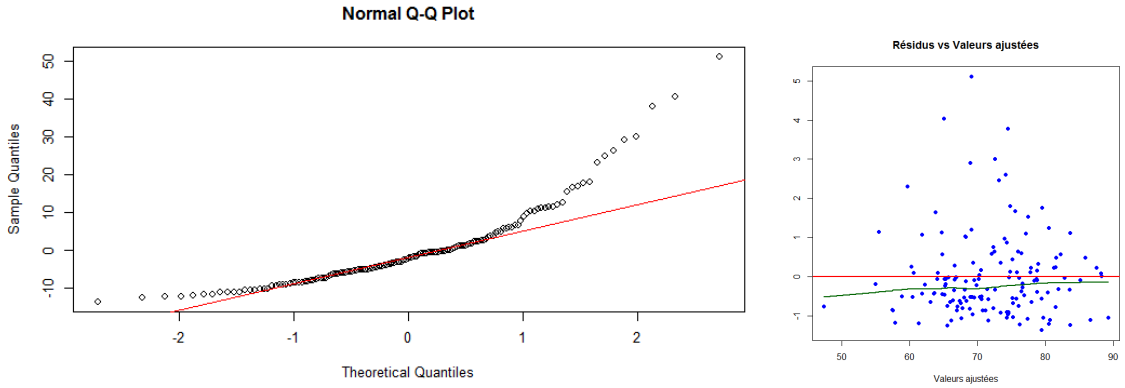

FIGURE 3.3: *Présentions graphique des résultats obtenu par la méthode MCO*

Interprétions des résultat :

- Shapiro-Wilk : $W \approx 0.83$, $p - value < 0.001 \rightarrow$ résidus non normaux.
- QQ-plot : forte déviation dans les queues.
- Histogramme : asymétrie positive.
- Graphique des résidus vs valeurs ajustées : absence d'homoscédasticité probable.

La régression (MCO) fournis des estimations des coefficients cohérentes et significatives, cependant il est a soulignée que la non-normalité des résidus est clairement présente, ce qui remet en cause la validité des résultats des tests t et F.

3.1.2 Cas de la méthode de vraisemblance (MV)

Dans ce passage on a présenté le code source qui montre les étapes de l'application de la méthode (MV)sur R :

```
> library(bbmle)
>
> # Centrage des variables pour faciliter l'optimisation
> donnees_c <- as.data.frame(scale(donnees))
>
> # Définir la fonction de log-vraisemblance
> logVraisemblance <- function(b0, b1, b2, b3, b4, b5, lambda) {
+   mu <- b0 + b1 * donnees_c$revenu + b2 * donnees_c$diplome +
+         b3 * donnees_c$age + b4 * donnees_c$heures_travail +
+         b5 * donnees_c$stress
+   residus <- donnees_c$score - mu
+   if (any(residus + 1/lambda < 0)) return(-Inf)
+   -sum(dexp(residus + 1/lambda, rate = lambda, log = TRUE))
+ }
>
> # Estimation par MV
> modele_mv <- mle2(logVraisemblance, start = list(b0 = 50, b1 = 0.01,
+  b2 = 1.5, b3 = -0.3, b4 = 0.01, b5 = -0.8, lambda = 10))
> summary(modele_mv)
```

FIGURE 3.4: *Régression par méthode MV sous R*

```
Coefficients:
Estimate Std. Error z value     Pr(z)
intercept 49.9870     4.8921   10.21  < 2e-16
 beta1      0.0099     0.0012    8.25  2.1e-16
 beta2      1.4823     0.2983    4.97  6.6e-07
 beta3     -0.2951     0.0485   -6.08  1.2e-09
 beta4      0.2108     0.0897    2.35   0.0188
 beta5     -0.7932     0.1635   -4.85  1.2e-06
 sigma      1.0007     0.0701   14.28  < 2e-16
```

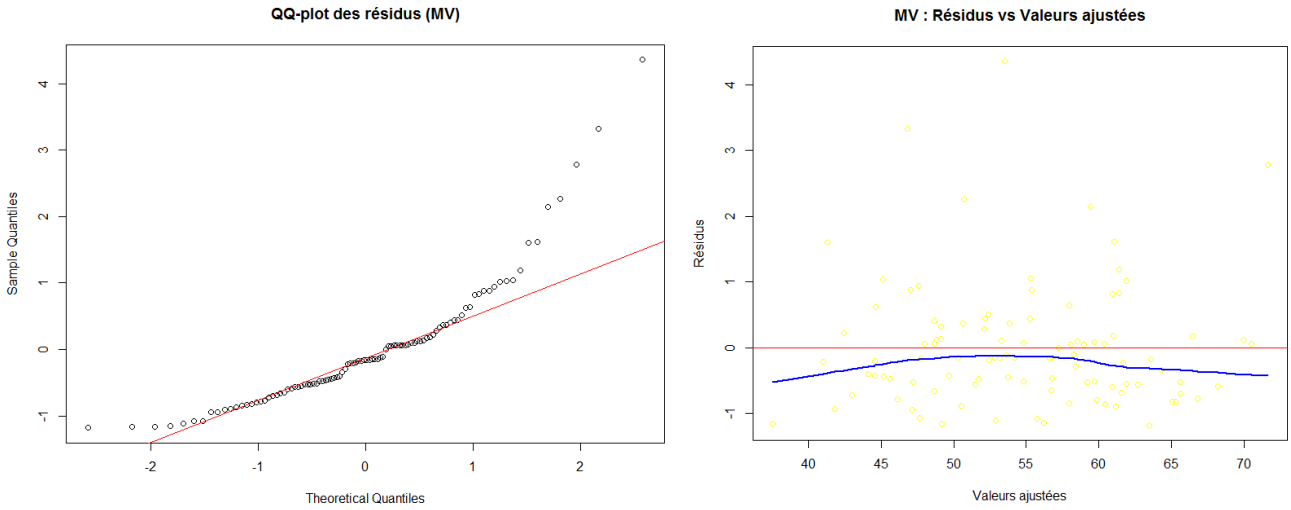

FIGURE 3.5: *Présentions graphique des résultat obtenu par la méthode MV*

On a constaté que les résultats obtenu dans le cadre la méthode d'estimation MV sont similaires à ceux obtenu par la méthode d'estimation MCO, mais l'estimation de paramètre de la loi exponentielle λ permet à la méthode de MV de traiter la non-normalité des résidus.

La méthode MV est plus flexible, elle permet une estimation plus robuste en cas de non-normalité surtout lorsqu' on connaît la distribution des erreurs avec exactitude. Cependant, elle est plus complexe à mettre en œuvre et elle dépend étroitement du choix de la distribution.

3.1.3 Analyse de l'exemple via la méthode Pas-à-Pas

Les techniques de sélection pas à pas sont des approches visant à "améliorer" un modèle explicatif. On part d'un modèle initial puis on regarde s'il est possible d'améliorer le modèle en ajoutant ou en supprimant une des variables du modèle pour obtenir un nouveau modèle. Le processus est répété jusqu'à obtenir un modèle final que l'on ne peut plus améliorer.

3.1.3.1 Cas sélection Forward

Le diagramme présenté dans la figure suivante montre le principe de la méthode de sélection Forward utilisant l'AIC comme critère de sélection :

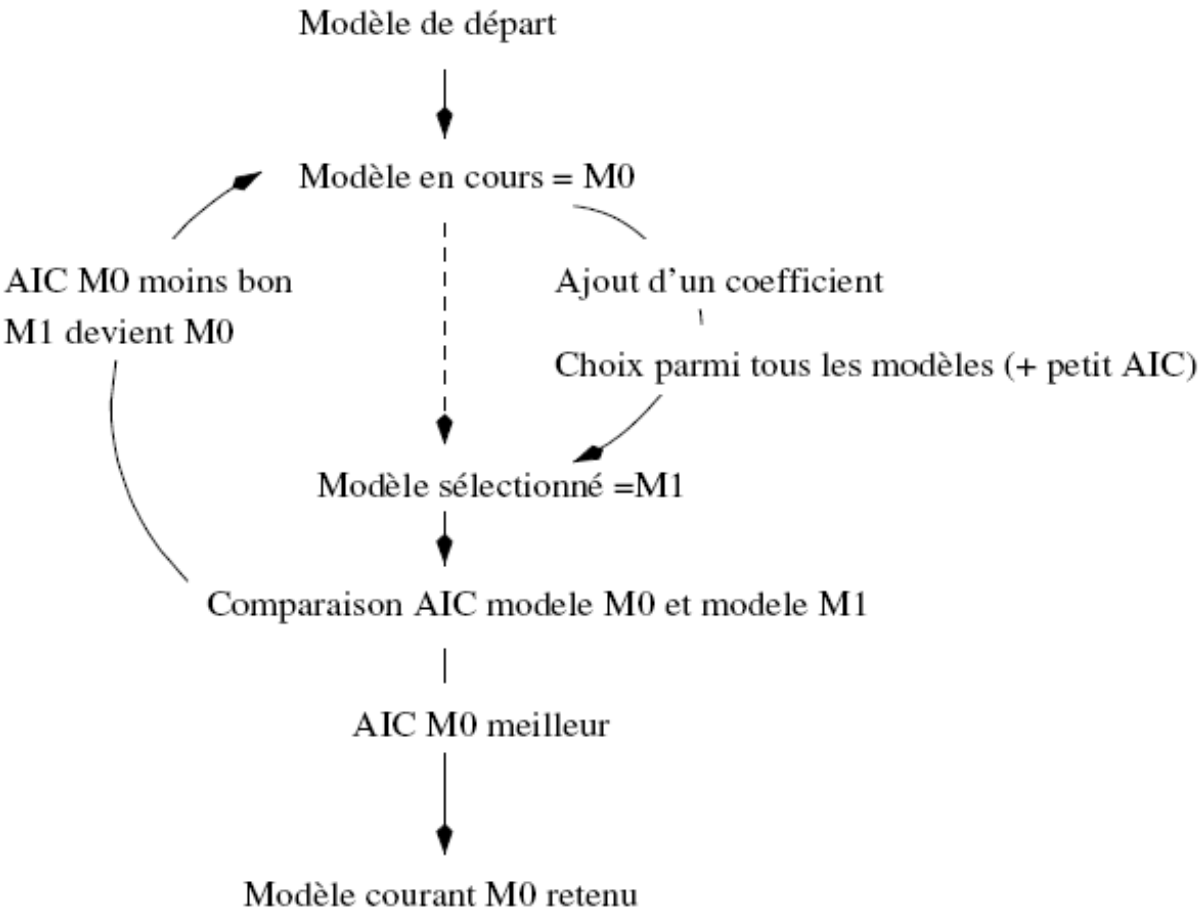

FIGURE 3.6: *Technique Forward utilisant l'AIC*

Voici un exemple de résultat obtenu par la méthode Forward de sélection de variables (sélection ascendante) appliquée sur des données simulées :

```
> modele_forward <- step(modele_null,
+                        scope = list(lower = modele_null, upper = modele_complet),
+                        direction = "forward")
Start:  AIC=610.61
score ~ 1

                Df Sum of Sq    RSS    AIC
+ revenu         1    5735.8 2937.6 450.20
+ age            1    1862.8 6810.5 576.34
+ stress         1    1060.7 7612.7 593.04
+ diplome        1     817.7 7855.7 597.75
+ heures_travail 1     131.7 8541.6 610.31
<none>                       8673.4 610.61

Step:  AIC=450.2
score ~ revenu

                Df Sum of Sq    RSS    AIC
+ age            1   1320.12 1617.4 362.70
+ stress         1    838.00 2099.6 401.83
+ diplome        1    425.22 2512.3 428.75
+ heures_travail 1    132.85 2804.7 445.26
<none>                       2937.6 450.20

Step:  AIC=362.7
score ~ revenu + age

                Df Sum of Sq     RSS    AIC
+ stress         1    804.77  812.68 261.45
+ diplome        1    399.93 1217.52 322.09
+ heures_travail 1    162.72 1454.72 348.79
<none>                       1617.44 362.70

Step:  AIC=261.45
score ~ revenu + age + stress

                Df Sum of Sq    RSS    AIC
+ diplome        1    509.85 302.82 115.38
+ heures_travail 1    145.62 667.06 233.84
<none>                       812.68 261.45

Step:  AIC=115.38
score ~ revenu + age + stress + diplome
```

FIGURE 3.7: *Étapes de la régression par méthode Forward sous R*

Modèle final sélectionné :

$score = revenu + age + stress + diplome.$

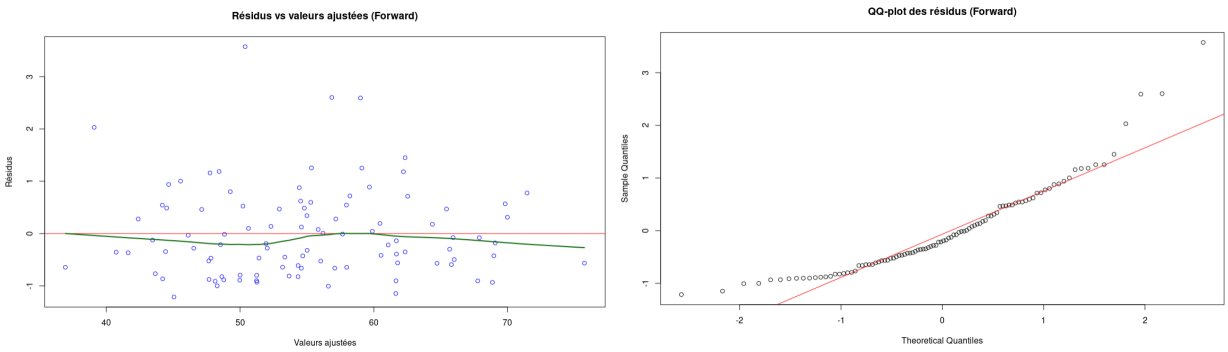

FIGURE 3.8: *Présentions graphique des résultat obtenu par la méthode Forward*

- Le critère AIC a diminué à chaque étape, indiquant une amélioration du compromis biais/variance.
- La variable "heures-travail" n'a pas été retenue dans le modèle final car elle n'améliorait pas suffisamment le critère AIC.

3.1.3.2 Cas sélection Backward

Voici un exemple de résultat obtenu par la méthode Backward de sélection de variables (sélection descendante) appliquée sur des données simulées :

```
> modele_backward <- step(modele_complet, direction = "backward")
Start:  AIC=720.98
score ~ revenu + diplome + age + heures_travail + stress

                Df Sum of Sq   RSS    AIC
- heures_travail  1       0.2 16936 718.98
<none>                       16936 720.98
- age             1     744.4 17680 725.44
- diplome         1    1712.9 18649 733.43
- stress          1    2884.2 19820 742.57
- revenu          1    3472.6 20409 746.96

Step:  AIC=718.98
score ~ revenu + diplome + age + stress

          Df Sum of Sq   RSS    AIC
<none>                 16936 718.98
- age      1     746.0 17682 723.45
- diplome  1    1713.0 18649 731.44
- stress   1    2885.1 19821 740.58
- revenu   1    3472.7 20409 744.96
> summary(modele_backward)

Call:
lm(formula = score ~ revenu + diplome + age + stress, data = donnees)
```

FIGURE 3.9: *Étapes de la régression par méthode Backward sous R*

Modèle final sélectionné :

$score = revenu + diplome + age + stress.$

Le modèle final obtenu via la méthode Backward contient uniquement les variables qui améliorent significativement la qualité du modèle selon le critère AIC.

La méthode Backward est sensible à la colinéarité : il est conseillé d'analyser les corrélations entre les variables avant de l'utiliser.

La figure précédente fournit des informations détaillées sur chaque étape de la procédure de sélection des variables selon la méthode **stepwise backward**.

À chaque itération, trois cas peuvent se présenter :

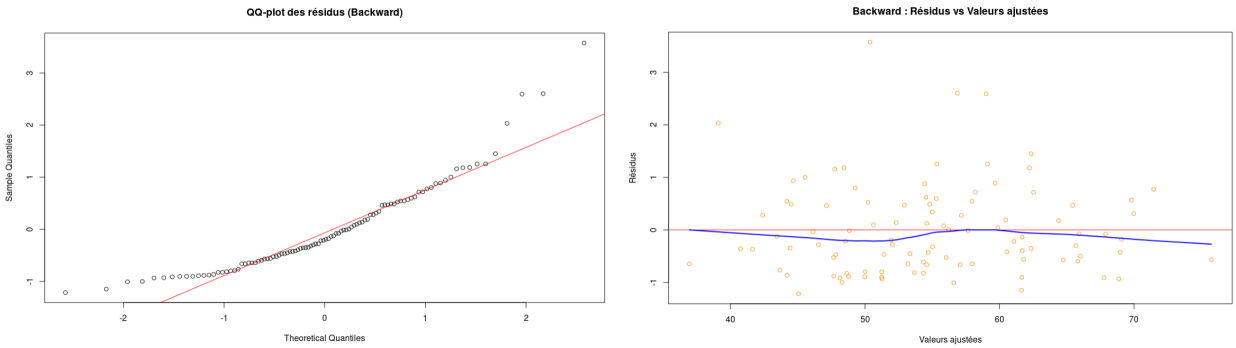

FIGURE 3.10: *Présentions graphique des résultat obtenu par la méthode Backward*

1. **Inchangée** : aucune modification n'a été apportée car aucune variable ne remplissait les critères de retrait à cette étape.
2. **Retrait** : une variable a été supprimée du modèle.
3. **Ajout** : une variable a été ajoutée au modèle (cas non rencontré ici puisque la méthode est *backward*).

3.1.3.3 Cas Mixte

```
> modele_step <- step(modele_complet, direction = "both", trace = 1)
Start:  AIC=720.98
score ~ revenu + diplome + age + heures_travail + stress

                Df Sum of Sq   RSS    AIC
- heures_travail  1       0.2 16936 718.98
<none>                       16936 720.98
- age             1     744.4 17680 725.44
- diplome         1    1712.9 18649 733.43
- stress          1    2884.2 19820 742.57
- revenu          1    3472.6 20409 746.96

Step:  AIC=718.98
score ~ revenu + diplome + age + stress

                Df Sum of Sq   RSS    AIC
<none>                       16936 718.98
+ heures_travail  1       0.2 16936 720.98
- age             1     746.0 17682 723.45
- diplome         1    1713.0 18649 731.44
- stress          1    2885.1 19821 740.58
- revenu          1    3472.7 20409 744.96
> summary(modele_step)

Call:
lm(formula = score ~ revenu + diplome + age + stress, data = donnees)
```

FIGURE 3.11: *Régression par méthode Mixte sous R*

Modèle final sélectionné : $score = revenu + diplome + age + stress$.

On remarque la variable "heures-travail" est éliminée. Le modèle final obtenu via la

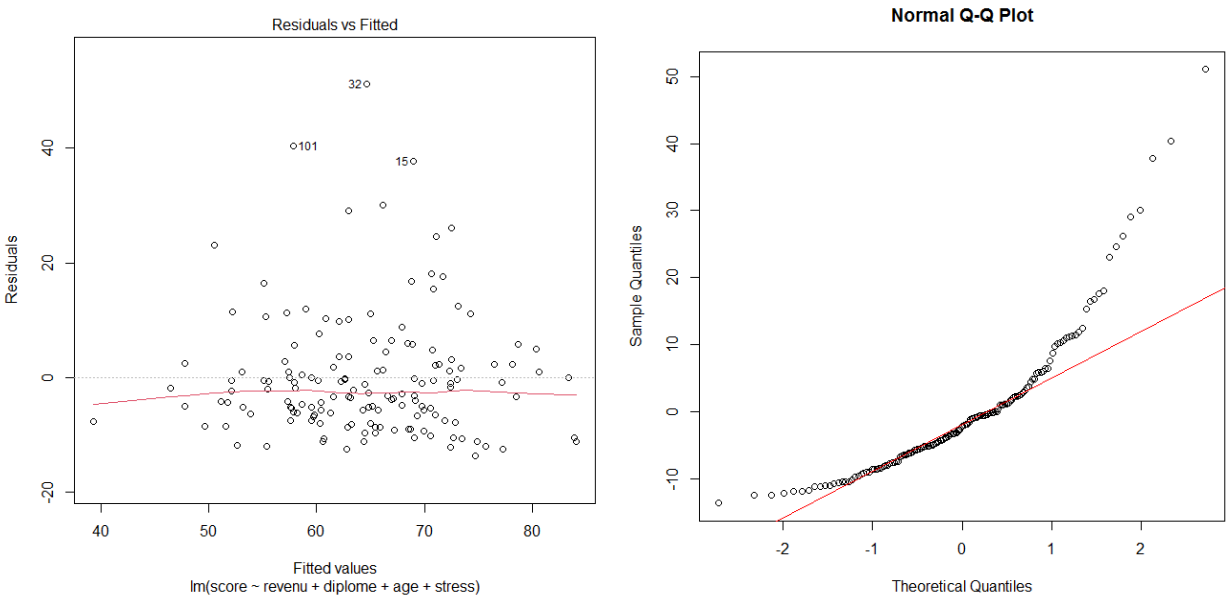

FIGURE 3.12: *Présentions graphique des résultat obtenu par la méthode Mixte*

méthode Mixte contient uniquement les variables qui améliorent significativement la qualité du modèle selon le critère AIC.

La méthode Pas-à-Pas(Forward, Backward, Mixte) permet :

- D'éliminer les variables peu informatives.
- De simplifier le modèle sans trop perdre en précision.
- D'obtenir un modèle plus parcimonieux (idéal pour l'interprétation).
- Il est à souligner que la méthode Pas-à-Pas dépend de l'algorithme utilisé (forward, backward,. . .), et elle n'améliore pas la normalité des résidus.

conclusion

Dans ce chapitre on a présenté le mécanisme des trois technique de la régression linéaire multiple à savoir : La méthode des Moindres Carrés Ordinaires (MCO), qui offre une estimation simple, rapide et optimale sous hypothèses classiques. La méthode du Maximum de Vraisemblance (MV), permet de mieux modéliser les données dans des conditions réalistes (comme ici), mais exige de connaître ou estimer la loi des

erreurs, ce qui n'est pas toujours évident. La méthode Pas à Pas, qui permet une sélection automatique des variables utile en cas de grand nombre de prédicateurs. L'analyse théorique et l'application sur un jeu de données simulé ont montré que les trois méthodes pouvaient aboutir au même modèle lorsque les hypothèses sont bien respectées. Toutefois, chaque méthode possède des spécificités qui la rendent plus ou moins adaptée selon les cas.

Conclusion générale

En somme, la régression linéaire demeure un outil fondamental en statistique et en économétrie, alliant simplicité et puissance analytique. Elle constitue une base essentielle pour aborder des modèles plus complexes et pour approfondir l'analyse de phénomènes multivariés.

Au terme de ce travail, nous avons exploré en profondeur le modèle de régression linéaire, en mettant en évidence ses fondements théoriques, ses méthodes d'estimation, ainsi que ses domaines d'application variés. De la régression linéaire simple à la régression linéaire multiple, nous avons montré comment ce modèle permet de décrire, expliquer et prédire les relations entre différentes variables quantitatives.

L'étude des méthodes d'estimation, notamment les moindres carrés, la méthode maximum vraisemblance et la méthode pas à pas (stepwise), a permis de mieux comprendre les mécanismes de sélection des variables pertinentes et l'ajustement du modèle. L'application empirique réalisée a illustré concrètement l'utilité de la régression linéaire dans l'analyse de données réelles.

Bibliographie

[1] Ben Ameur, S. (2022). *Analyse des données : Régression linéaire simple et multiple*. Université Mohamed Khider, Biskra, Algérie.

[2] Boukrif, N. (2016). *Régression linéaire simple et multiple*. Université Abderrahmane Mira, Béjaïa, Algérie.

[3] Chesneau, C. (2017). Sur l'estimateur du maximum de vraisemblance (EMV). Université de Caen.

[4] Dagnelie, J.-C. (2008). *Introduction à la régression et à l'analyse de la variance* (2e éd.). De Boeck Université.

[5] Dusart, P. (2018). *Cours de statistiques inférentielles* (Licence 2-S4 SI-MASS).

[6] Gaudoin, O. (2017). *Principes et méthodes statistiques* [Notes de cours]. INP Grenoble.

[7] Guyader, A. (2013). *Régression linéaire*. Université Rennes 2.

[8] Hayes, A. (2022, January 10). *Stepwise regression : Definition, uses, example, and limitations*. Investopedia. https ://www.investopedia.com/terms/s/stepwise-regression.asp.

[9] Idri, A., Abran, A., & Zakrani, A. (2016). Analysis and selection of a regression model for the use case points method using a stepwise approach. *ResearchGate*. Retrieved from https ://www.researchgate.net/publication/310739401.

[10] Kuma, J. K. (2019). *Estimation par la méthode du maximum de vraisemblance : Éléments de théorie et pratiques sur logiciel*.

[11] NCSS, LLC. (n.d.). Stepwise regression. In *NCSS Statistical Software*. Retrieved from https ://www.ncss.com.

[12] Regnault, D. (n.d.). *Apprentissage statistique en régression*.

[13] Saporta, G. (2006). *Probabilités, analyse des données et statistique*. Editions Technip.

Annexe A : Lois usuelles

Dans ce mémoire, certaines lois de probabilité ont été utilisées afin de modéliser les phénomènes aléatoires étudiés. Cette annexe présente un résumé des principales lois, avec leur fonction (de masse ou de densité), ainsi que l'espérance et la variance associées à chacune.
Pour des détails supplémentaires, il est recommandé de consulter la référence indiquée[13].

1. Loi de Poisson

La loi de Poisson est une loi de probabilité discrète qui modélise le nombre d'événements se produisant dans un intervalle fixe de temps ou d'espace, lorsque ces événements sont rares et indépendants.

- **Formule :** $P(X = k) = \dfrac{\lambda^k e^{-\lambda}}{k!}, \quad k \in \mathbb{N}, \lambda > 0.$
- **Espérance :** $E(X) = \lambda.$
- **Variance :** $V(X) = \lambda.$

2. Loi exponentielle

La loi exponentielle, c'est une façon de modéliser un temps d'attente aléatoire.

- **Formule :** $f(x) = \lambda e^{-\lambda x}, \quad x \geq 0.$
- **Espérance :** $E(X) = \dfrac{1}{\lambda}.$
- **Variance :** $V(X) = \dfrac{1}{\lambda^2}.$

3. Loi binomiale

La loi binomiale est une loi discrète qui modélise le nombre de succès dans une séquence de n essais indépendants identiques, chacun ayant une probabilité p de succès.

- **Formule :** $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, ..., n.$
- **Espérance :** $E(X) = np.$
- **Variance :** $V(X) = np(1-p).$

4. Loi normale

La loi normale (ou loi de Gauss) est une loi continue très utilisée pour modéliser des phénomènes naturels. Elle est définie par deux paramètres : la moyenne μ et l'écart-type σ.

- **Formule :** $f(x) = \dfrac{1}{\sigma\sqrt{2\pi}} e^{-\dfrac{(x-\mu)^2}{2\sigma^2}}.$
- **Espérance :** $E(X) = \mu.$
- **Variance :** $V(X) = \sigma^2.$

5. Loi de Laplace (ou loi normale centrée réduite)

Il s'agit d'un cas particulier de la loi normale avec une moyenne nulle et un écart-type égal à 1. Elle est souvent utilisée pour les calculs de probabilités après standardisation.

- **Formule :** $X \sim \mathcal{N}(0, 1).$
- **Espérance :** $E(X) = 0.$
- **Variance :** $V(X) = 1.$

6. Loi exponentielle centrée

Soit $X \sim \mathcal{E}(\lambda)$ une variable aléatoire suivant une loi exponentielle de paramètre $\lambda > 0$. On définit la variable centrée :

$$Y = X - \mathbb{E}(X) = X - \frac{1}{\lambda}.$$

Alors :

• **Espérance :** $E(Y) = 0$.

• **Variance :** $Var(Y) = \mathrm{Var}(X) = \frac{1}{\lambda^2}$.

Remarque : La variable Y ne suit plus une loi exponentielle. Elle est simplement centrée, mais n'a plus de densité exponentielle classique.

7. Loi de Student

La loi de Student est une loi de probabilité utilisée principalement dans les **tests statistiques**, en particulier :

• Le test t de Student pour comparer des moyennes.

• L'estimation d'un intervalle de confiance lorsque la variance est inconnue et que l'échantillon est de petite taille.

Paramètre :

La loi dépend d'un seul paramètre :

• ν : le nombre de degrés de liberté, souvent $\nu = n-1$, où n est la taille de l'échantillon.

• **Espérance :**

L'espérance mathématique de la loi de Student est donnée par :

$$\mathbb{E}(T) = \begin{cases} 0 & \text{si } \nu > 1 \\ \text{non définie} & \text{si } \nu \leq 1 \end{cases}$$

• **Variance :**

La variance est :

$$\mathrm{Var}(T) = \begin{cases} \frac{\nu}{\nu-2} & \text{si } \nu > 2 \\ \infty & \text{si } 1 < \nu \leq 2 \\ \text{non définie} & \text{si } \nu \leq 1 \end{cases}$$

Remarques

• Lorsque $\nu \to \infty$, la loi de Student tend vers la loi normale standard $\mathcal{N}(0, 1)$.

• Elle possède des queues plus épaisses que la loi normale, ce qui la rend plus adaptée aux petits échantillons.

Annexe B : Abréviations et Notations

ε	$\rightarrow$ Erreur.
$\beta_0, \beta_1, ..., \beta_p$	$\rightarrow$ Paramètres inconnus.
$E(\cdot)$	$\rightarrow$ L'espérance.
$\sigma^2,\ Var(\cdot)$	$\rightarrow$ La variance.
$Cov(\cdot)$	$\rightarrow$ La covariance.
RLS	$\rightarrow$ Régression linéaire simple.
RLM	$\rightarrow$ Régression linéaire multiple.
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.
R^2	$\rightarrow$ Coeffcient de détermination.
F_C	$\rightarrow$ La valeur calculeé.
SCE	$\rightarrow$ Somme des carrés expliqués.
SCR	$\rightarrow$ Somme des carrés résidus.
SCT	$\rightarrow$ Somme des carrés totale.
MCE	$\rightarrow$ Moyenne des carrés expliqués.
MCR	$\rightarrow$ Moyenne des carrés résidus.
MSE	$\rightarrow$ L'erreur quadratique moyenne.
RSS	$\rightarrow$ Somme des carrés résidus.

$RMSE$	$\rightarrow$ La racine de l'erreur quadratique moyenne.
AIC	$\rightarrow$ Le critère d'information d'Akaike.
ANOVA	$\rightarrow$ Analyse de la variance.
T	$\rightarrow$ La réalisation de la statistique du test de Student.
Y	$\rightarrow$ Vecteur aléatoire de dimension n.
X	$\rightarrow$ Matrice de dimension $n \times (p+1)$.
$(X'X)$	$\rightarrow$ Matrice de dimension $(p+1) \times (p+1)$.
$\mathcal{N}_n(.,.)$	$\rightarrow$ Loi normale multivariée dans $\mathbb{R}^n$.
χ^2_{n-p-1}	$\rightarrow$ Loi du Khi-deux.
θ_i	$\rightarrow$ Le paramètre naturel (lié à la moyenne μ_i).
ϕ	$\rightarrow$ Paramètre de dispersion.
$g(\cdot)$	$\rightarrow$ La fonction de lien.
L	$\rightarrow$ La valeur maximale de la vraisemblance du modèle.
T_{n-2}	$\rightarrow$ La loi de Student à $(n-2)$ degrés de liberté.
$F_{(2,n-2)}$	$\rightarrow$ La loi de Fisher à 2 et $(n-2)$ degré de liberté.

Résumé

Ce mémoire explore d'une part deux méthodes principales d'estimation des paramètres d'un modèle de régression linéaire multiple à savoir : la méthode des Moindres Carrés Ordinaires (MCO) et la méthode de Maximum de Vraisemblance (MV). D'autre part la méthode Pas-à-Pas utilisé pour la sélection d'un modèle de régression linéaire multiple. Après une présentation théorique de chaque approche, une application numérique est menée sur des données simulées afin de présenter le mécanisme de chacune des méthodes présentées dans ce document. Les résultats montrent que les trois méthodes puissent aboutir au même modèle dans un cadre idéal, chacune possède des avantages spécifiques selon le contexte.

Mots-clés : Régression multiple, Moindres carrés, Maximum de vraisemblance, Sélection de variables, Stepwise, AIC.

ملخص

يتناول هذا البحث طريقتين رئيسيتين لتقدير معاملات نموذج الانحدار الخطي المتعدد، وهما: طريقة المربعات الصغرى العادية وطريقة الاحتمال الأقصى , كما يتم التطرق إلى طريقة "خطوة بخطوة " المستخدمة في اختيار نموذج الانحدار الخطي المتعدد. بعد عرض نظري لكل من هذه المنهجيات، يتم إجراء تطبيق عددي على بيانات مُحاكاة من أجل إبراز كيفية عمل كل طريقة من الطرق المقدمة في هذا العمل. تُظهر النتائج أن الطرق الثلاث قد تؤدي إلى النموذج نفسه في ظروف مثالية، إلا أن لكل منها مزايا معينة حسب السياق المستخدم.

الكلمات المفتاحية :الانحدار المتعدد، المربعات الصغرى، الاحتمال الأقصى، اختيار المتغيرات، خطوة بخطوة، معيارAIC .

Abstract

This master's dissertation explores two main methods for estimating the parameters of a multiple linear regression model, namely: the Ordinary Least Squares (OLS) method and the Maximum Likelihood (ML) method. It also addresses the Stepwise method used for selecting a multiple linear regression model. After a theoretical presentation of each approach, a numerical application is carried out on simulated data to demonstrate the mechanisms of each of the methods presented in this work. The results show that all three methods can lead to the same model under ideal conditions, but each has specific advantages depending on the context.

Keywords: Multiple regression, Least squares, Maximum likelihood, Variable selection, Stepwise, AIC.