République Algérienne 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

DÉPARTEMENT DE MATHÉMATIQUES

Mémoire présenté en vue de l'obtention du Diplôme :

MASTER en Mathématiques

Option : **Statistique**

Par **Daas Mohamed EL Amine**

Titre :

Modèles Linéaires Généralisés

Membres du Comité d'Examen :

Pr.	Meraghni Djamel	UMKB	Président
Pr.	Yahia Djabrane	UMKB	Encadreur
Dr.	Zouaoui Nour Elhouda	UMKB	Examinatrice

Juin 2024

Dédicace

Je dédie cet humble travail

À mes chers parents,

À mes frères,

À mon encadreur

À toutes mes amies

À tous ceux qui m'ont soutenu et aidé.

REMERCIEMENTS

Au nom d'**ALLAH** le plus clément et le plus miséricordieux.

Avant toute chose, je remercie **ALLAH** le tout le puissant de m'avoir donné la force pour l'achèvement de ce travail et de m'avoir donné le courage et la patience durant toutes ces années d'études.

Je tiens à exprimer mes plus sincères remerciements au Professeur **Yahia Djabrane**. En tant que directeur de mon mémoire de Master, il a joué un rôle essentiel dans la réussite de ce travail. Son expertise, son expérience et sa patience ont été déterminantes pour m'aider à comprendre la méthodologie de la recherche scientifique. Ses conseils avisés m'ont poussé à fournir le meilleur de moi-même tout au long de ce travail, et pour cela, je lui suis extrêmement reconnaissant.

Je souhaite exprimer ma profonde gratitude aux membres du comité d'examen, le Professeur **Meraghni Djamel** et la Docteure **Zouaoui Nour Elhouda**, pour leur temps, leur expertise, et pour avoir accepté d'évaluer ce travail.

Je tiens également à remercier tous les professeurs du département de mathématiques qui ont accompli leur devoir avec sincérité et ont soutenu les étudiants tout au long de leur parcours universitaire. Je remercie particulièrement les Professeurs **Yahia Djabrane** et **Benatia Fatah** et **Dr.Sana Benameur** et **Dr.Fatima Ouaar** et **Dr.Hassiba Berkane** pour leurs conseils et leur soutien.

D.mohamed_el_amine.

Table des matières

Table des figures

Liste des tableaux

Introduction

La statistique a plusieurs buts dont l'exploration des données, la prise de décisions (tests et prévisions) et la modélisation. Différents modèles ont été créés pour modéliser divers problèmes statistiques comme les modèles linéaires, les modèles de comptage, les familles exponentielles, les modèles log-linéaires, etc.

Parmi ces modèles, les modèles linéaires sont très populaires. Que les variables prédictives soient numériques, catégorielles ou les deux, on peut exprimer dans une formulation unifiée la régression linéaire, l'analyse de la variance et de la covariance. Ce cadre simple est appelé Modèle Linéaire Général. Il suppose une distribution gaussienne de la variable dépendante conditionnellement aux prédicteurs, et un lien linéaire entre les variables. Cependant, ces modèles linéaires ne conviennent pas toujours à tous les problèmes statistiques, certains étant trop complexes pour être modélisés de manière aussi simple.

En 1972, de nouveaux modèles ont été formulés par John Nelder et Robert Wedderburn, appelés modèles linéaires généralisés, comme une généralisation souple de la régression linéaire. Les GLM ("General Lineair Models") généralisent la régression linéaire en permettant au modèle linéaire d'être relié à la variable réponse via une fonction lien. Cela permet d'unifier d'autres modèles statistiques, y compris la régression logistique et la régression de Poisson.

Dans ce mémoire, notre objectif est de présenter les différents modèles statistiques utilisés pour résoudre les problèmes pratiques, en mettant particulièrement l'accent sur les modèles linéaires généralisés. Le mémoire s'articule autour de trois chapitres comme suit :

Chapitre 1 : *Généralités sur les Modèles Linéaires : Dans ce chapitre, On donne un rappel des formules de la régression linéaire multiple qui constitue une généralisation naturelle de la régression linéaire simple, dont le nombre des variables explicatives supérieure ou égal 2.*

Chapitre 2 : *Modèles Linéaires Généralisés : Ce chapitre est consacré à l'étude des modèles linéaires généralisés, présentation et défintion d'un modèle linéaire généralisé et modèle de dénombrement (Odds et Odds ratio) et la régression logistique.*

Chapitre 3 : *Dans ce dernier chapitre, nous donnons une application, sous logiciel R, sur des données réelles de la maladie cardiovasculaire avec un modèle de régression logistique, une forme spéciale des modèles linéaires généralisés, dont l'explication de la variable réponse est liée à une variable explicative qui présente la classe d'âge de l'individu étudier.*

Chapitre 1

Généralités sur les Modèles Linéaires

D'une façon générale, un modèle linéaire est une expression qui relie une variable quantitative (la variable à expliquer) à des variables, quantitatives et/ou qualitatives (les variables explicatives). L'analyse des modèles linéaires porte des noms différents selon la nature des variables explicatives utilisées dans le modèle. La nomination des différentes analyses par nature des variables explicatives est comme suit :

- **Régression simple :** une seule variable quantitative.
- **Régression multiple :** plusieurs variables quantitatives.
- **Analyse de la variance :** plusieurs variable qualitatives.
- **Analyse de covariance :** une ou plusieurs variable quantitatives et plusieurs qualitatives.

Dans ce premier chapitre, nous donnons des généralités sur les modèles linéaires de la régression simple et multiple. Ces techniques de régression linéaire sont utilisées pour estimer la relation entre les variables et prédire les valeurs futures en fonction des données disponibles.

1.1 Régression Linéaire Multiple

Dans cette section, nous étendons le cas de modèle linéaire multiple, dont on cherche à expliquer une variable Y par un ensemble de p variables explicatives $X_1, X_2, ..., X_p$. Pour simplifier l'écriture du modèle, on utilisera la forme matricielle. Le modèle de régression

linéaire multiple est l'outil statistique le plus habituellement mis en œuvre pour l'étude de données multidimensionnelles.

1.1.1 Modèle de régression linéaire multiple

Le modèle de régression linéaire multiple s'écrit pour $p \geq 2$:

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{i,j} + \varepsilon_i, \qquad i = 1, ..., n. \tag{1.1}$$

où :

- x_{ij} est la $j - éme$ variable explicative pour l'individu i, $(j = \overline{1, p})$.

- β_j sont les paramètres inconnus.

- ε_i sont les termes d'erreur, il s'agit d'une variable aléatoire centrée et de variance finie, qui suit une loi normal $\varepsilon \rightsquigarrow \mathcal{N}(0, \sigma_\varepsilon^2 I_n)$.

Le modèle (1.1) s'écrit sous la forme matricielle :

$$Y = XB + \varepsilon.$$

tell que

$$Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}, \; B = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_p \end{bmatrix}, \; X = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ 1 & x_{2,1} & x_{2,2} & \cdots & x_{1,p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & x_{n,2} & \cdots & x_{n,p} \end{bmatrix}, \; \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

avec :

Y est un vecteur aléatoire de dimension n.

X est une matrice de dimension $(n, p+1) \in \mathbb{R}^n \times \mathbb{R}^{p+1}$.

ε est le vecteur d'erreur de dimension n.

Remarque 1.1.1 *La première colonne formée par la valeur* 1 *indique la constante* β_0 *et le vecteur* B *de dimension* $p+1$ *est celui des paramètres du modèle.*

Remarque 1.1.2 *1) La matrice* X *est de rang plein, c-à-d*

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

2) Les erreurs sont centrée, de même variance, et non corrélées entre elles, i.e

$$E(\varepsilon) = 0_n, Var(\varepsilon) = \sigma_\varepsilon^2 I_n, \text{avec } I_n \text{ la matrice identité dimension } n.$$

3) Les erreurs sont indépendant des X_j $(j = \overline{1,p})$.

1.1.2 Estimation des paramétres

On estime les paramètres inconnues du modèle par minimisation des moindres carrés (estimateurs des MCO) : La méthode des moindres carrés cherche la meilleure estimation des paramètres β et minimisation la quantité Ψ,

$$\hat{\beta} = \arg\min_B \sum_{i=1}^{n} \varepsilon_i^2$$

$$\Psi(B) = \Psi(\beta_0, ..., \beta_p) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{i,j} \right)^2.$$

Alors, pour calculer $\hat{B}$, nous proposons l'écriture matricielle suivante :

$$\Psi(B) = \varepsilon^t \varepsilon = \|Y - XB\|^2 = (Y - XB)^t (Y - XB) = (Y^t - B^t X^t)(Y - BX)$$
$$= Y^t Y - Y^t XB - B^t X^t Y - B^t X^t BX.$$

car $Y^t BX = B^t X^t Y$, on obtient donc

$$\Psi(B) = Y^t Y - 2B^t X^t Y - B^t X^t XB.$$

En dérivant $\Psi(B)$ par rapport à chacun des paramèters $\beta_0, \beta_1, ..., \beta_p$, on obtient le système des équations :

$$\begin{aligned}\frac{\partial \Psi}{\partial B} = 0 = -2X^tY + 2X^tX\hat{B} &\Rightarrow -X^tY + X^tX\hat{B} = 0 \\ &\Rightarrow X^tY = X^tX\hat{B} \\ &\Rightarrow (X^tX)^{-1}X^tY = (X^tX)^{-1}X^tX\hat{B} \\ &\Rightarrow \hat{B} = (X^tX)^{-1}X^tY.\end{aligned}$$

Propriété 1.1.1 *$\hat{B}$ est un estimateur sans biais de B : $E(\hat{B}) = B$.*

Preuve. Il suffit d'écrire,

$$\hat{B} = (X^tX)^{-1}X^tY = (X^tX)^{-1}X^t(XB+\varepsilon) = (X^tX)^{-1}X^tXB+(X^tX)^{-1}X^t\varepsilon = B+(X^tX)^{-1}X^t\varepsilon$$

Alors,

$$E[\hat{B}] = E[B + (X^tX)^{-1}X^t\varepsilon] = B + (X^tX)^{-1}X^tE[\varepsilon]$$

et puisque $E[\varepsilon] = 0$, alors $E[\hat{B}] = B$, donc $biais\left(\hat{B}\right) = 0$. ■

Propriété 1.1.2 *La variance de $\hat{B}$ est*:

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

Preuve. On a

$$\begin{aligned}Var(\hat{B}) &= E\left[(\hat{B} - E[\hat{B}])(\hat{B} - [\hat{B}])^t\right] = E\left[(\hat{B} - B)(\hat{B} - B)^t\right] = E\left[((X^tX)^{-1}X^t\varepsilon)((X^tX)^{-1}X^t\varepsilon)^t\right] \\ &= E\left[(X^tX)^{-1}X^t\varepsilon\varepsilon^tX(X^tX)^{-1}\right] = (X^tX)^{-1}X^tE[\varepsilon\varepsilon^t]X(X^tX)^{-1} = (X^tX)^{-1}X^t\sigma_\varepsilon^2I_nX(X^tX)^{-1} \\ &= \sigma_\varepsilon^2(X^tX)^{-1}X^tX(X^tX).\end{aligned}$$

Alors $Var(\hat{B}) = \sigma_\varepsilon^2(X^tX)^{-1}$. ■

Propriété 1.1.3 *L'estimateur de la variance résiduelle est donné par*

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

Preuve. On a

$$\sigma_{\hat{B}} = \sigma_{\varepsilon}^2 (X^t X)^{-1}$$

et on a

$$\hat{\varepsilon} = Y - \hat{Y} = (XB + \varepsilon) - X\hat{B} = (XB + \varepsilon) - X(B + (X^t X)^{-1} X^t \varepsilon) = (I - X(X^t X)^{-1} X^t)\varepsilon.$$

On pose $A = (I - X(X^t X)^{-1} X^t)$, telle que A est symétrique $(A^t = A)$ et idempotente $(A^2 = A)$, de taille (n, n). Alors,

$$\hat{\varepsilon}^t \hat{\varepsilon} = \hat{\varepsilon}^t A \varepsilon \quad \text{et} \quad E\left[\hat{\varepsilon}^t \hat{\varepsilon}\right] = \sigma_{\varepsilon}^2 tr(A)$$

car $tr(A)$ est égal $n - (p+1) = n - p - 1$. Donc,

$$\hat{\sigma}_{\varepsilon}^2 = \frac{E[\hat{\varepsilon}^t \hat{\varepsilon}]}{tr(A)} = \frac{1}{n-p-1}\sum_{i=1}^{n}\varepsilon_i^2.$$

Ce qui forme un estimateur sans biais de σ_{ε}^2. ■

1.1.3 Lois des estimateurs

Le modèle de régression linéaire multiple est :

$$Y = XB + \varepsilon \qquad \text{avec } \varepsilon \rightsquigarrow \mathcal{N}_n(0, \sigma_{\varepsilon}^2 I_n),$$

où $\mathcal{N}_n$: la loi normale dans $\mathbb{R}^n$ (la loi normale multivariée) et ε : indépendantes et identiquement distribuées (i.i.d). Donc

$$Y \rightsquigarrow \mathcal{N}_n(XB, \sigma_{\varepsilon}^2 I_n).$$

Théorème 1.1.1 *On a*

$$i)\ \hat{B} \rightsquigarrow \mathcal{N}_{p+1}\left(B, \sigma_\varepsilon^2 (X^t X)^{-1}\right) \qquad ii)\ (n-p-1)\frac{\hat{\sigma}_\varepsilon^2}{\sigma_\varepsilon^2} \rightsquigarrow \mathcal{X}^2_{n-p-1}.$$

1.1.4 Test sur les paramétres

D'après le théorème 1.1.1 : si σ_ε^2 est connue la statistique utilisée pour tester la signification des paramètres s'écrit :

$$\frac{\hat{\beta}_j - \beta_j}{\sigma_\varepsilon \sqrt{v_j}} \rightsquigarrow \mathcal{N}(0,1).$$

si σ_ε^2 est inconnue la statistique s'écrit :

$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_\varepsilon \sqrt{v_j}} \rightsquigarrow \mathcal{T}_{n-p-1}.$$

où v_j est le $j-éme$ terme diagonal de la matrice $(X^t X)^{-1}$, où $\mathcal{T}_{n-p-1}$ est la loi de student à $(n-p-1)$ degrés de liberté.

Test de student : Pour chaque paramètre β_j, nous utilisnos le test d'hypothèse pour tester β_j :

$$\begin{cases} H_0 : \beta_j = \beta, \qquad j = \overline{0.p}, \qquad \beta \in \mathbb{R} \\ H_1 : \beta_j \neq \beta \end{cases}$$

Dans ce cas, on rejet H_0.si

$$T_{\beta_j} = \frac{\left|\hat{\beta}_j - \beta\right|}{\hat{\sigma}_\varepsilon \sqrt{v_j}} > t_{1-\frac{\alpha}{2}\,(n-p-1)}.$$

avec α est le niveau de signification du test, $\alpha \in]0,1[$.

Cas particulier : si $\beta_j = 0$ pour un certain j, le test précédent est dit un test de signification.

Test sur le modèle

On note le test global comme suit :

$$\begin{cases} H_0 : B = B_0, & B_0 \in \mathbb{R}^p \\ H_1 : B \neq B_0 \end{cases}$$

Pour tester cette hypothèse, nous utilisons la statistique de Fisher :

$$F = \frac{\frac{SCE}{p}}{\frac{SCR}{n-p-1}} = \frac{MCE}{MCR},$$

où F suit une loi de Fisher avec (p)et $(n-p-1)$ degrés de liberté, les autres quantités sont définies dans le tableau suivant :

Variation	ddl	Somme des carrés (S.C)	Carré moyen (C.M)	F
Expliquée	p	SCE	$MCE = \frac{SCE}{p}$	$F = \frac{MCE}{MCR}$
Résiduelle	$n-p-1$	SCR	$MCR = \frac{SCR}{n-p-1}$	
Totale	$n-1$	SCT		

TAB. 1.1 – .Tableau d'analyse de la variance de la régression linéaire multiple.

On rejet H_0 si la statistique :

$$F = \frac{\frac{SCE}{p}}{\frac{SCR}{n-p-1}} = \frac{MCE}{MCR} > f_{1-\alpha}\left(p, n-p-1\right).$$

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

Test sur le modèle réduit

Dans la pratique, il est parfois possible de procéder au test de la nullité de certains termes seulement, le problème sera donc à tester:

$$H_0 : \beta_1 = \beta_2 = ... = \beta_q = 0, \qquad q < p$$

p le nombre exact des paramètres du modèle.

Définissons le coefficient de détermination noté R_q^2 du modèle réduit à $(p-q)$ variables. Sous l'hypothèse nulle, la statistique:

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

où SCE_q est la somme des carrées expliquée du modèle réduit.et $Q_q \rightsquigarrow \mathcal{F}(q, n-p-1)$. On rejet H_0 si $Q_q > f_{1-\alpha}(q, n-p-1)$, où f est le fractile d'ordre $(1-\alpha)$ de la loi de Fisher $\mathcal{F}(q, n-p-1)$.

1.1.5 Qualité d'ajustement

L'équation de l'analyse de la variance du modèle est donnée par les sommes aux carrés suivantes :

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

La qualité d'ajustement est jugé par le coefficient de détermination R^2, défini par:

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

L'ajustement soit parfaite dès que R^2 se rapproche de 1 ($R^2 \simeq 1$).

Il est possible de prendre de plus en considération le coefficient de détermination ajusté $\bar{R}^2$, défini par :

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

1.1.6 Prévision

La prévision dans le cas d'un modèle linéaire multiple consiste à calculer une estimation

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

avec $X_0 = (1 + X_0^1 + X_0^2 + ... + X_0^p)$ les observations qui arrivent après avoir écrire le modèle. Les intervalles de confiance des prévision de Y et $E[Y]$ au niveau de confiance $(1-\alpha)\,\%$, sont données respectivement par :

$$pré(Y) = \hat{Y}_0 \pm t\hat{\sigma}_\varepsilon \left(1 + X_0(X^tX)^{-1}X_0^t\right)^{\frac{1}{2}}$$
$$pré(E[Y]) = \hat{Y}_0 \pm t\hat{\sigma}_\varepsilon \left(X_0(X^tX)^{-1}X_0^t\right)^{\frac{1}{2}}$$

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

1.2 Régression Linéaire Simple

La régression linéaire simple est un cas particulier de la régression linéaire multiple, où seule une variable est utilisée pour prédire la variable de réponse. Alors que dans la régression linéaire multiple, plusieurs variables sont utilisées pour prédire la variable de réponse. En régression linéaire simple, la relation entre les variables est représentée par droite, tandis que les cas multiples peuvent nécessiter une représentation à plusieurs dimensions.

Définition 1.2.1 *(Régression Linéaire simple) : Un modèle de régression linéaire simple est défini par une équation de la forme :*

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad i = \overline{1.n}$$

n : *est nombre d'observation*

y_i : *est la variable expliquer et dépendante*

x_i : *est la variable explicative et indépendante*

β_0 *et* β_1 : *les paramètres inconnue du modèle*

ε_i : *l'erreur d'une variable aléatoire suit une loi normal* $\varepsilon \rightsquigarrow \mathcal{N}(0, \sigma_\varepsilon^2)$.

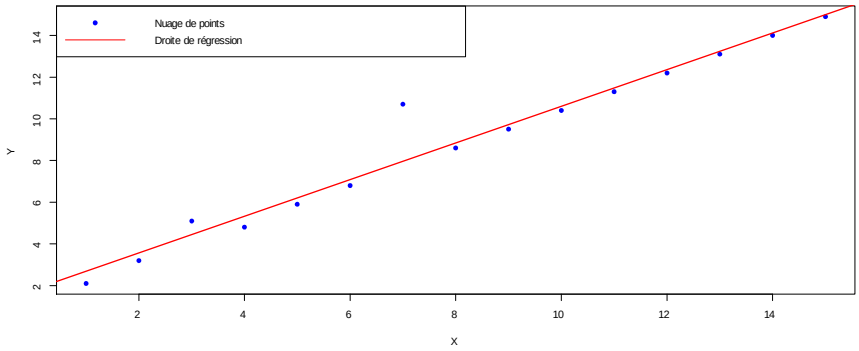

FIG. 1.1 – Ajustement du nuage de point par régression

Propriétés 1.2.1 *Dans une régression linéaire simple, on a*

- $E(\hat{\beta}_0) = \beta_0$ *et* $E(\hat{\beta}_1) = \beta_1$.
- $E(\varepsilon_i) = 0$, $E(\varepsilon_i^2) = \sigma_\varepsilon^2 < \infty$, $i = 1, ..., n$.
- $cov(\varepsilon_i, \varepsilon_j) = 0$, $\forall (i, j)$ *tel que* $i \neq j$.
- $cov(\varepsilon_i, X) = 0$, *l'erreur est indépendante du* X.

Chapitre 2

Modèles Linéaires Généralisés

Les modèles linéaires généralisés permettent d'étudier la liaison entre une variable dépendante ou réponse Y et un ensemble de variables explicatives ou prédicteurs $(X_1, X_2, ..., X_n)$. Dans de nombreuses applications, les variables à expliquer ne varient pas dans tout $\mathbb{R}$ mais dans $\mathbb{R}_+, \mathbb{N}$ ou encore un intervalle d'entiers. ll est clair que le modèle gaussien est mal adapté à cette situation.

2.1 Présentation du modèle linéaire généralisé

2.1.1 Définition du modèle

Définition 2.1.1 (Modèle linéaire généralisé) *: Le modèle linéaire généralisé (GLM) spécifie que y_i est une variable aléatoire dont la loi est paramétrée par une combinaison linéaire des régresseurs $x_i\beta$ par exemple $y_i \sim P(x_i\beta)$.*

En pratique la situation est la suivante : on dispose de données Y et X (variable explicatives) ; il faut alors spécifier une famille $(P_\theta)_{\theta\in\mathbb{R}}$ de distributions de probabilité à un paramètre réel θ ainsi qu'une fonction réelle $\eta \mapsto g(\eta)$ dont l'inverse est appelé fonction de lien, qui donne

le lien entre le paramètre θ et $X\beta$, le modèle est alors :

$$y_i \sim P_{\theta_i} \ , \ i = \overline{1.n}, \qquad E_{\theta_i}[Y] = g(X\beta). \tag{2.1}$$

Remarque 2.1.1 *Cette dernière équation (2.1) présuppose que l'application $\theta \longmapsto E_\theta[Y]$, est bijective ce qui sera toujours le cas des familles exponentielles dont θ est le paramètre naturel (le facteur de y dans l'exponentielle). On voit que le modèle linéaire gaussien rentre dans ce cadre avec la famille $N(\theta, \sigma^2)$ et $g(\eta) = \mu$.*

Choix de la famille P_θ : Les logiciels proposent typiquement les familles :

– Gamma.

– Inverse gaussienne.

– Poisson.

– Valeurs entières positives non bornées.

– Binomiale.

– Valeurs entières positives dans un intervalle...

Choix de la g : le choix par défaut proposé par les logiciels est

$$g(\eta) = E_\eta[Y],$$

ce qui conduit à $\theta_i = x_i\beta$, ce choix permet une estimation numériquement robuste, et conduit à des valeurs réalistes indépendamment de x_i (i.e par exemple comprises entre 0 et 1 si la loi est Binomiale).

Remarque 2.1.2 *Pour comprendre les implications du choix de g, précisons la paramétrisation des modèles. Ces familles ont toutes une densité de la forme :*

$$f(y_i, \theta_i, \phi) = \exp\{[Y_i\theta_i - b(\theta_i)] \diagup a(\phi) + c(Y_i, \phi)\}.$$

2.1.2 Les composantes du modèle linéaire généralisé

Les modèles linéaires généralisés sont caractérisés par trois composantes : la composante aléatoire, le prédicteurs linéaire ou composante déterministe, la fonction lien.

– **La composante aléatoire** : Elle est définie par distribution de probabilité de la variable réponse. Soit $Y_1, Y_2, ..., Y_n$ des variables d'un échantillon aléatoire de la variable réponse Y, ces variables étant supposées indépendantes admettant des distributions issues d'une famille exponentielle. Chaque observation Y_i admet une fonction de densité de la forme :

$$f(y_i, \theta_i, \phi) = \exp\{[Y_i\theta_i - b(\theta_i)]/a(\phi) + c(Y_i, \phi)\}.$$

θ_i : le paramètre naturel de la famille exponentielle.

$a(\phi)$: la fonction a la forme de $a(\phi) = \frac{\phi}{\omega_i}$ ou les poids ω_i sont connus, ϕ est appelé paramètre de dispersion.

Dans lequel ϕ est un paramètre de nuisance. Si ϕ est constante connue (en générale elle est égale 1) l'expressin précedente se met sous la forme canoniqe suivante :

$$f(y_i, \theta_i) = a(\theta_i)b(Y_i)\exp[Y_i\vartheta(\theta_i)].$$

En posant :

$$a(\theta_i) = \exp\left[\frac{-b(\theta_i)}{a(\phi)}\right], \;\; b(Y_i) = \exp[c(Y_i, \phi)], \;\; \vartheta(\theta_i) = \left[\frac{\theta_i}{a(\phi)}\right].$$

Le term $\vartheta(\theta)$ est appelé le paramètre naturel de la distribution. Tout autre paramètre de la distribution est considéré comme un paramètre de nuisance. Cette famille $f(y_i, \theta_i)$ comprend de nombreuses distributions importantes telle-que la loi de Poisson et la loi Binomial. La valeur du paramètre θ_i dépend des valeurs variables explicatives.

Remarque 2.1.3 *L'expression précédente est la forme la plus générale des modèle linéaire généralisé. Elle englobe l'ensemle des lois usuelles utilisant un ou deux paramètres, tels que la loi normale, l'inverse de la loi Normale, la loi Gamma, la loi Poisson, la loi Binomiale.*

– **La composante déterministe (prédicteurs linéaire) :** Elle est définie par la fonction linéaire des variables explicatives, utilisées comme prédicteurs dans le modèle. Dans un modèle linéaire généralisé, l'espérance mathématique de Y, notée μ, Y, notée μ, varie en fonction des valeurs des variables explicatives. Le prédicteurs linéaire est exprimé sous forme d'une combinaison linéaire.

$$\eta = \beta_0 + \beta_1 X_1 + ... + \beta_n X_n.$$

La composante déterministe du modèle se rapporte à un vecteur d'un ensemble de variable explicatives $\eta = \eta_1, ..., \eta_n$ par un modèle linéaire

$$\eta = X^t \beta,$$

X : matrice se compose de n valeurs des variables explicatives.

β : le vecteur des paramètres du modèle.

η : le vecteur prédicteurs linéaire.

– **Fonction lien :** La fonction lien spécifie comment l'espérance mathématique Y, notée μ, est liée au prédicteurs linéaire construit à partir des variables explicatives. Cette composante exprime une relation fonctionnelle entre la composante déterministe et la composante aléatoire. Soit $\mu = E[Y]$ on pose $g(\mu) = \eta$, où g appelé fonction lien, c'est une fonction différenciable et monotone, voir [1, page 33], [2, page 25].

Donc on peut modéliser l'espérance μ directement comme dans la régression linéaire, ou modéliser une fonction monotone $g(\mu)$ de l'espérance. On a alors :

$$\eta = X^t \beta, \qquad g(\mu) = \eta \Rightarrow g^{-1}[g(\mu)] = g^{-1}(\eta),$$
$$\Rightarrow \mu = g^{-1}(X^t \beta).$$

Remarque 2.1.4 *La fonction lien qui associe la moyenne μ au paramètre naturel est appelée fonction de lien canonique.*

Proposition 2.1.1 *A tout loi de probabilité de la composante aléatoire est associe une fonction spécifique de l'espérance appelée paramètre canonique. Pour la distribution normale il s'agit de l'espérance elle-même. Pour la distribution Poisson le paramètre canonique est le logarithme de l'espérance. Pour la distribution binomiale le paramètre canonique est la probabilité de succès.*

Le tableau suivant résume les différents types de modèles couverts par le modèle linéaire généralisé, avec :

– la composante a : est la composantes aléatoires.

– la composante d : est la composantes déterministes (prédicteurs linéaire).

Modèle	**composante a**	**composante d**	**fonction lien**
Régression	Normale	Quantitatives	Identité
Analyse du variance	Normale	Quantitatives	Identité
Analyse du covariance	Normale	Mixtes	Identité
Régression logistique	Binomiale	Mixtes	Logit
Modèles log-linéaires	Poisson	Mixtes	Log
Modèles multinomiales	multinomiale	Mixtes	Logit généralisé

TAB. 2.1 – Récapitulatif des principaux Modèle

2.2 Modèle de dénombrement

Contrairement aux modèles précédents, l'hypothèse de normalité est remplacée par une loi discrète. Les lois usuelles discrètes sont "loi Bernoilli, loi Binomiale, loi Poisson,..." qui appartient à la famille des modèles linéaires généralisés.

2.2.1 Odds et Odds ratio

Cas d'une variable : Soit Y une variable qualitative à j modalités. On désigne la chance ou l'**odds** de voir se réaliser la $jème$ modalités plutôt que la $kème$ par le rapport :

$$\Omega_{jk} = \frac{\pi_j}{\pi_k},$$

où π_j est la probabilité d'apparition de la $jème$ modalités. Cette quantité est estimeé par le rapport n_j/n_k des effectifs observés sur un échantillon. Ainsi,

$$\Omega_{jk} \simeq \hat{\Omega}_{jk} = \frac{n_j}{n_k}.$$

où $\pi_j = P(Y = y_j)$, $\pi \in [0, 1]$.

Exemple 2.2.1 *Lorsque la variable est binaire et suit une loi de Bernoulli de paramètre* $\pi, Y \rightsquigarrow B(\pi)$

$$\begin{cases} P(Y = 1) = \pi \\ P(Y = 0) = 1 - \pi. \end{cases}$$

Alors, il y'a

$$\Omega_{jk} = \frac{\pi}{1 - \pi}$$

‹‹chance de gain››.

Cas d'un tableau de contingence : On considère un tableau de contingence 2×2 croisant deux variables qualitatives binaires X_1 et X_2. Les paramètre de loi conjointe se mettent dans une matrice :

$$\begin{bmatrix} \pi_{11} & \pi_{12} \\ \pi_{21} & \pi_{22} \end{bmatrix}$$

où $\pi_{ij} = P[\{X_1 = i\} \text{ et } \{X_2 = j\}]$ est la probabilité d'occurrence de chaque combinaison.

– Dans la ligne 1, l'odds que la colonne 1 soit prise plutôt que la colonne 2 est :

$$\Omega_1 = \frac{\pi_{11}}{\pi_{12}}.$$

– Dans la ligne 2, l'odds que la colonne 1 soit prise plutôt que la colonne 2 est :

$$\Omega_2 = \frac{\pi_{21}}{\pi_{22}},$$

– On appelle odds ratio le rapport

$$\Theta = \frac{\Omega_1}{\Omega_2} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}.$$

– Si les variables sont indépendantes, $\Theta \geq 1$ si dans les lignes 1 on a plus de chance de prendre la première colonne, $\Theta < 1$, sinon.

Cas générale : L'odds ratio est également défini pour deux lignes (a,b) et deux colonnes (c,d) quelconques d'une table de contingence croisant deux variables à J et K modalités. L'odds ratio est le rapport

$$\Theta_{abcd} = \frac{\Omega_a}{\Omega_b} = \frac{\pi_{ac}\pi_{bd}}{\pi_{ad}\pi_{bc}}, \text{ estimé par l'odds ratio } \hat{\Theta}_{abcd} = \frac{n_{ac}n_{bd}}{n_{ad}n_{bc}}.$$

2.3 Régression logistique

La régression logistique ou **modèle logit** est un modèle de régression binomiale. Comme pour tous les modèles de régression binomiale, il s'agit de modéliser au mieux un modèle mathématique simple à des observations réelles nombreuses. En d'autres termes, d'associer à un vecteur des variables aléatoires $(x_1, ..., x_p)$ une variable aléatoire binomiale notée y.

La régression logistique constitue un cas particulier de modèle linéaire généralisé. Elle est largement utilisée dans de nombreux domaines comme par exemple :

– En médecine, pour trouver les facteurs qui caractérisent un groupe de sujets malades par rapport à des sujets sains.
– Dans le domaine bancaire, pour détecter les groupes à risque de la sous éruption d'un crédit.

Définition 2.3.1 (Modèle Logit) *: Ce modèle de régression logistique décrit la modélisation d'une variable qualitative Y à deux modalités possible* 1 *ou* 0*, sucèes ou échec.*

Les modèles de régression logistique précédents adaptés à l'explication d'une variable quantitative ne s'appliquent plus directement car la régression usuel $X\beta$ ne prend pas des valeurs simplement binaires. L'objectif est adapté à cette situation en cherchant à expliquer les probabilités :

$$P(Y = 0) = 1 - \pi \quad \text{ou} \quad P(Y = 1) = \pi,$$

ou plutôt transformation de celles-ci, par l'observation conjointe des variable explicatives. L'idée est en effet de faire intervenir une fonction réelle monotone g opérant de $[0, 1]$ dans $\mathbb{R}$ et donc chercher un modèle linéaire de la forme :

$$g(\pi_i) = x_i^t \beta, \qquad i = \overline{1.n}$$

On a trois type de cette fonction g :

$$\begin{cases} \text{probit : } g \text{ la fonction inverse de la distribution normale.} \\ \text{log-log : } g(\pi) = \ln[-\ln(1-\pi)]. \\ \text{logit : } logit(\pi) = g(\pi) = \ln\left(\frac{\pi}{1-\pi}\right), \text{ avec } g^{-1}(x) = \dfrac{\exp(x)}{1+\exp(x)}. \end{cases}$$

Exemple 2.3.1 *Soient* $y_1, y_2, ..., y_n$ $(n \geq 1)$ *observations indépendantes de* Y *telle que* $Y_i \rightsquigarrow$

$B(n,\pi)$. *La fonction de lien* $g(\mu) = logit(\pi) = \ln\left(\frac{\pi}{1-\pi}\right)$, *alors*

$$\ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1 + ... + \beta_n X_n$$
$$logit(\pi_i) = \beta X_i^t \ , \quad i = \overline{1,n}.$$

En effet, après transformation de l'équation ci-dessus, nous obtenons

$$\pi_i = \frac{\exp(\beta_0 + \beta_1 x_1 + ... + \beta_n x_n)}{1 + \exp(\beta_0 + \beta_1 x_1 + ... + \beta_n x_n)}, \ i = \overline{1,n}.$$

2.4 Famille des lois exponentielles

Considérons un échantillon de n individus $(y_1, y_2, ..., y_n)$:

$$y_i = \begin{cases} 1 \text{ où l'événement s'est réalisé pour l'individu } i \\ 0 \text{ sinon }, \quad \forall i = \overline{1,n} \end{cases}$$

La probabilité de survenue d'un événement y_i est :

$$E[y_i] = P(y_i = 1) \times 1 + P\left(y_i = 0\right) \times 0 = p_i$$

Les modèle admettant des codage $(0,1)$ sont appelés modèle **dichotomique** par convention, un modèle linéaire simple, avec un vecteur exogène $x_i = (x_i^1, ..., x_i^k)$ s'écrit comme suit

$$y_i = x_i^t \beta + \varepsilon_i \ , \quad i = \overline{1,n}$$

Donc, conditionnellement ou vecteur x_i on a

$$\varepsilon_i = \begin{cases} 1 - x_i^t\beta \text{ avec probabilité } P(y_i = 1) = p_i, \\ \\ -x_i^t\beta \text{ avec probabilité } P(y_i = 0) = 1 - p_i. \end{cases}$$

Alors, ε admet une loi discrète, et non pas normale.

Par hypothèse, $E(\varepsilon_i) = 0$, alors

$$\begin{aligned} E(\varepsilon_i) &= (1 - x_i^t\beta)P(\varepsilon_i = 1 - x_i^t\beta) + (-x_i^t\beta)P(\varepsilon_i = -x_i^t\beta) \\ &= p_i(1 - x_i^t\beta) + (1 - p_i)(-x_i^t\beta) \\ &= (p_i - x_i^t\beta) = 0 \\ &\Longrightarrow p_i = x_i^t\beta. \end{aligned}$$

de même,

$$\begin{aligned} var(\varepsilon_i) &= E(\varepsilon_i^2) - (E(\varepsilon_i))^2 = E(\varepsilon_i^2) \\ &= (1 - x_i^t\beta)^2P(\varepsilon_i = 1 - x_i^t\beta) + (-x_i^t\beta)^2P(\varepsilon_i = -x_i^t\beta) \\ &= (1 - x_i^t\beta)^2p_i + (1 - p_i)(-x_i^t\beta)^2, \end{aligned}$$

lorsque $p_i = x_i^t\beta$, alors

$$var(\varepsilon_i) = (1 - x_i^t\beta)^2x_i^t\beta + (1 - x_i^t\beta)(-x_i^t\beta)^2.$$

Définition 2.4.1 *(La famille exponentielle) : C'est la famille exponentielle des variables aléatoires y ayant la forme suivante :*

$$f(y, \mu, \phi) = \exp\{\frac{1}{\phi}[ya(y) - b(y)] + c(y, \phi)\} \tag{2.2}$$

où a, b et c sont des fonctions données, ϕ est le paramètre de dispersion et μ est le paramètre d'échelle. On pose

$$\begin{cases} Q(\theta) = \frac{\theta}{\phi}. \\ u(\theta) = \exp\{-\frac{b(\theta)}{\phi}\}. \\ v(y) = \exp\{c(y, \phi)\}. \end{cases}$$

(2.2) prendre la forme dite forme exponentielle canonique, et cela comme suit :

$$f(y_i, \phi_i) = u(\theta_i)v(y_i)\exp\{y_i, Q(\theta_i)\}. \tag{2.3}$$

La fonction qui associé la moyenne μ, au paramètre d'échelle est appelle ***fonction lien canonique.***

Exemple 2.4.1 *(Loi de Bernoulli) : Soit $Y_i \rightsquigarrow B(p_i)$:*

$$\begin{aligned} f(y_i, p_i) = P(Y_i = y_i) &= p_i^{y_i}(1-p_i)^{1-y_i},\ i = \overline{1.n} \\ &= \frac{p_i^{y_i}}{(1-p_i)^{y_i}}(1-p_i) = \left(\frac{p_i}{(1-p_i)}\right)^{y_i}(1-p_i),\ i = \overline{1.n} \\ &= (1-p_i)\exp\{y_i \log\left(\frac{p_i}{(1-p_i)}\right)\},\ i = \overline{1.n} \end{aligned}$$

donc on pose

$$\begin{cases} Q(p_i) = \log\left(\frac{p_i}{(1-p_i)}\right). \\ u(p_i) = (1-p_i). \\ v(y_i) = Id. \end{cases}$$

La paramètre d'échelle $\theta_i = \log\left(\frac{p_i}{(1-p_i)}\right)$ est la fonction lien canonique.

$$\begin{aligned} g(p_i) = E[y_i] = \theta_i &= \log\left(\frac{p_i}{(1-p_i)}\right) \\ \Longrightarrow g(x)^{-1} &= \frac{\exp(x)}{1+\exp(x)},\ c'est\ une\ fonction\ logit. \end{aligned}$$

Dans le tableau suivant nous donnons un récapitulatif des composantes des lois usuelles de la famille exponentielle :

Distribution	$Q(p_i)$	$u(p_i)$	$v(y_i)$
Loi gaussienne $N(\mu_i, \sigma^2)$	$\frac{\mu_i}{\sigma^2}$	$\exp(\frac{-1}{2}\left(\frac{\mu_i}{\sigma}\right)^2)$	$\exp(\frac{-1}{2}y_i^2/\sigma^2 - \frac{1}{2}\ln(2\pi\sigma^2))$
Binomiale $B(n, p_i)$	$\ln\left(\frac{p_i}{1-p_i}\right)$	$(1-p_i)^n$	C_n^x
Poisson $P(\lambda_i)$	$exp(-\lambda_i)$	$log(\lambda_i)$	$\frac{1}{y_i}$

TAB. 2.2 – Récapitulatif des composantes des lois de la famille exponentielle

2.4.1 Estimation des paramètres du modèle

On pose $\forall i = \overline{1, N}$:

$$y_i = \begin{cases} 1 \text{ si } p_i = F(x_i^t\beta) \\ 0 \text{ si } 1 - p_i = 1 - F(x_i^t\beta) \end{cases}$$

où

$$\begin{cases} p(y_i = 1) = p_i, \\ p(y_i = 0) = 1 - p_i, \end{cases}$$

avec $x_i^t = (x_1^1, ..., x_i^k)$, désigne un vecteur de caractéristique observable ou $\beta = (\beta_1, \beta_2, ..., \beta_k)^t \in \mathbb{R}^k$. Donc on peut considérer la variable aléatoire comme un modèle binomiale avec une probabilité $F(x_i^t\beta)$.

Définition 2.4.2 *La vraisemblance associé*

$$L = L(y_i, \beta) = p_i^{y_i}(1 - p_i)^{1-y_i}.$$

La vraisemblance associé a un échantillon $y = (y_1, y_2, ..., y_n)$ est donnée par

$$L(y_i, \beta) = \prod_{i=1}^{n} L(y_i, \beta) = \prod_{i=1}^{n} p_i^{y_i}(1 - p_i)^{1-y_i} = \prod_{i=1}^{n} [F(x_i^t\beta)]^{y_i}[1 - F(x_i^t\beta)]^{1-y_i} \qquad (2.4)$$

► *Ainsi,* $\forall x_i^t\beta \in \mathbb{R}$, *dans un modèle logit, on a*

$$F(x_i^t\beta) = \frac{\exp(x_i^t\beta)}{1+\exp(x_i^t\beta)} = \Lambda(x_i^t\beta).$$

► *Pour le modèle probit, on a*

$$F(x_i^t\beta) = \int_{-\infty}^{x_i^t\beta} \frac{1}{\sqrt{2\pi}} \exp(\frac{-z^2}{2})dz = \Phi(x_i^t\beta)$$

$$\begin{aligned}
(2.4) &\Longrightarrow \ell = \log L(y_i,\beta) \\
&= \sum_{i=1}^{n}[y_i \log(F(x_i^t\beta)) + (1-y_i)\log(1-F(x_i^t\beta))] \\
&\Longrightarrow \log L(y_i,\beta) = \sum_{i:y_i=1}^{n} \log(F(x_i^t\beta)) + \sum_{i:y_i=0}^{n} \log(1-F(x_i^t\beta))
\end{aligned}$$

On obtient une estimation du paramètre β, *en maximisant soit la fonction de vraisemblance* $L(y_i,\beta)$, *soit la fonction de log-vraisemblance* $\log L(y_i,\beta)$:

$$\begin{aligned}
G(\beta) &= \frac{\partial \log L(y_i,\beta)}{\partial\beta} \\
&= \sum_{i=1}^{n} y_i \frac{f(x_i^t\beta)}{F(x_i^t\beta)}x_i^t + (1-y_i)\frac{f(x_i^t\beta)}{1-F(x_i^t\beta)}x_i^t
\end{aligned}$$

où f *est la densité de* $F(.)$ *et* G *le vecteur du gradient*

$$\begin{aligned}
G(\beta) &= \sum_{i=1}^{n} \frac{[y_i - F(x_i^t\beta)]f(x_i^t\beta)}{F(x_i^t\beta)[1-F(x_i^t\beta)]}x_i^t \\
&= \sum_{i:y_i=1}^{n} \frac{f(x_i^t\beta)}{F(x_i^t\beta)}x_i^t - \sum_{i:y_i=0}^{n} \frac{f(x_i^t\beta)}{[1-F(x_i^t\beta)]}x_i^t.
\end{aligned} \tag{2.5}$$

Définition 2.4.3 *L'estimateur* $\hat{\beta}$ *de M.V de vecteur* $\beta \in \mathbb{R}^k$ *est défini par résolution du système d'équations en* β :

$$\hat{\beta} = \arg\max_{\{\beta\}}[\log L(y_i,\beta)] \Longleftrightarrow \frac{\partial \log L(y_i,\beta)}{\partial\beta} = G(\hat{\beta}) = 0$$

i) Dans le cas logit :

$$G_L(\hat{\beta}) = \sum_{i=1}^{n} \left[y_i - \Lambda(x_i^t \beta) x_i \right] = 0. \tag{2.6}$$

ii) Dans le cas probit :

$$G_p(\hat{\beta}) = \sum_{i=1}^{n} \frac{[y_i - \Phi(x_i^t \beta)]}{\Phi(x_i^t \beta)[1 - \Phi(x_i^t \beta)]} x_i = 0. \tag{2.7}$$

2.4.2 Matrice Hessienne et matrice d'information de Fisher

Définition 2.4.4 *La Matrice Hessienne associée à la log-vraisemblance d'un échantillon de taille n, $y = (y_1, y_2, ..., y_n)$ s'écrit*

$$\begin{aligned} H(\beta) &= \frac{\partial^2 \log(y, \beta)}{\partial \beta \partial \beta^t} \\ H(\beta) &= -\sum_{i=1}^{n} \left[\frac{y_i}{F(x_i^t \beta)^2} + \frac{(1 - y_i)}{(1 - F(x_i^t \beta))^2} \right] f(x_i^t \beta) x_i^t x_i \\ &\quad + \sum_{i=1}^{n} \left[\frac{y_i - F(x_i^t \beta)}{F(x_i^t \beta)[1 - F(x_i^t \beta)]} \right] f'(x_i^t \beta) x_i^t x_i \end{aligned}$$

dans le cas des modèles déchotomiques, on a $E[y_i] = F(x_i^t \beta)$;

$$\begin{aligned} E\left[H(\beta)\right] &= -\sum_{i=1}^{n} \left[\frac{1}{F(x_i^t \beta)} + \frac{1}{(1 - F(x_i^t \beta))} \right] f(x_i^t \beta) x_i^t x_i \\ &= -\sum_{i=1}^{n} \left[\frac{f(x_i^t \beta) x_i^t x_i}{F(x_i^t \beta)(1 - F(x_i^t \beta))} \right]. \end{aligned}$$

Définition 2.4.5 *La matrice d'information de Fisher $I(\beta)$ est donnée par $I(\beta) = -E[H(\beta)]$:*

i) *pour un modèle logit :* $I(\beta) = \sum_{i=1}^{n} \lambda(x_i^t \beta) x_i^t x_i$

ii) *pour un modèle probit :* $I(\beta) = \sum_{i=1}^{n} \frac{\phi(x_i^t \beta)}{(1 - \phi(x_i^t \beta))} x_i^t x_i$

2.4.3 Unicité du max de la log-vraisemblance

Si on admet que le maximum de $\log L(y_i, \beta)$ existe, la condition suffisante pour qu'il soit unique consiste a montrer que cette fonction est concave :

$$(2.5) \Longrightarrow \log L(y_i, \beta) = \sum_{i:y_i=1}^{n} \log\left(F(x_i^t\beta)\right) + \sum_{i:y_i=0}^{n} \log\left(1 - F(x_i^t\beta)\right)$$

Donc, il suffit de montrer que les fonctions $\log(F(x_i^t\beta))$ et $\log(1 - F(x_i^t\beta))$ sont concaves :

i) pour modèle logit :

$$\begin{aligned} F(x) &= \frac{\exp(x)}{1+\exp(x)}, \\ \frac{\partial \log \Lambda(x)}{\partial x} &= \frac{1}{\Lambda(x)} \frac{\partial \Lambda(x)}{\partial x} \\ &= \frac{1}{\Lambda(x)} \frac{\exp(x)(1+\exp(x)) - \exp(2x)}{(1+\exp(x))^2} \\ &= \frac{1}{1+\exp(x)}. \end{aligned}$$

et

$$\frac{\partial^2 \log \Lambda(x)}{\partial x^2} = \frac{\partial}{\partial x}\left(\frac{1}{1+\exp(x)}\right) = \frac{-\exp(x)}{(1+\exp(x))^2} < 0.$$

D'autre part,

$$\frac{\partial \log[1-\Lambda(x)]}{\partial x} = -\Lambda(x)$$

et

$$\frac{\partial^2 \log[1-\Lambda(x)]}{\partial x^2} = \frac{-\exp(x)}{(1+\exp(x))} < 0.$$

Alors, pour un modèle logit, l'estimateur de M.V est unique d'après la concavité. On fait la même chose pour le modèle probit.

Remarque 2.4.1 *La méthode de permettant de résoudre un système de type (2.7) et (2.6) est la méthode de Newton ou la méthode de score qui basent sur les étapes suivantes :*

i) *Des valeurs initiales* θ_0.

ii) *Une règle de passage d'un vecteur* θ_i *au suivant* θ_{i+1}.

iii) *Une règle d'arrêt* ***(convergence).***

Proposition 2.4.1 *On a la convergence en loi suivante :*

$$\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow[n\to\infty]{d} N[0, I(\beta_0)^{-1}].$$

En application, il faut remplacé $I(\beta_0)^{-1}$ *par* $I(\hat{\beta}_0)^{-1}$.

►**Test de Wald :** On désigne par β_j la j*éme* composante du vecteur $\beta = (\beta_1, ..., \beta_j, ..., \beta_k)^t \in \mathbb{R}^k$ pour tester l'hypothèse

$$\begin{cases} H_0 : \beta_j = b \\ H_1 : \beta_j \neq b \end{cases}$$

on utilise la statistique de Wald :

$$W_j = \frac{[\hat{\beta}_j - \beta]^t[\hat{\beta}_j - b]}{\hat{\nu}_{jj}},$$

ou $\hat{\nu}_{jj}$ est un estimateur de la variance de $\hat{\beta}_j$.

$$Z_j = \frac{\hat{\beta}_j - b}{\sqrt{\hat{\nu}_{jj}}} \xrightarrow[n\to\infty]{d} \mathcal{N}(0,1), \quad \text{donc} \;\; W_j \xrightarrow[n\to\infty]{d} \mathcal{X}^2_{(1)}.$$

Remarque 2.4.2 *Ce test est utilisé pour étudier la signification des paramètres* β_j *du modèle. En effet, on dit qu'un paramètre* β *est significatif à* $(1-\alpha)\,\%$*, si l'hypothèse* $H_0\,(\beta_j = 0)$ *est rejeté.*

Chapitre 3

Application

Dans ce dernier chapitre, nous donnons une application sur des données réelles (la source des données et la référence [9]) de la maladie cardiovasculaire avec un modèle de régression logistique, une forme spéciale des modèles linéaires généralisés, dont l'explication de la variable réponse (*binaire*) Y prend les valeurs :

0 : échec ou absence de la maladie,

1 : succès ou présence de la maladie,

à l'aide d'une variable explicative X qui présente la classe d'âge de l'individu étudier.

3.1 Présentation des données

Les données sur l'âge et la présence (noté 1) ou non (noté 0) d'une maladie cardiovasculaire ont été collectées pour 100 personnes et présentées dans le tableau 3.1 suivant. Dans chaque ligne du fichier, les informations suivantes sont présentées :

1. **ag :** la variable qui spécifie l'âge de l'individu,
2. **card :** qui indique la présence (1) ou l'absence (0) de maladie cardiovasculaire,
3. **Id :** qui est le numéro d'identification de l'individu,
4. **grag :** qui désigne la classe d'âge (nous avons choisi des classes d'âge de langueur 5 ans).

Les groupes sont organisés en intervalles de 5 ans et sont identifiés par des numéros (1,2,...,10). Exemple, la classe 1 est celle des individués de de 20 à 24 ans, la deuxième classe est celle des personnes de 25 à 29 ans,...

Id	grag	ag	card	Id	grag	ag	card	Id	grag	ag	card	Id	grag	ag	card
1	1	20	0	26	4	35	0	51	5	44	1	76	8	55	1
2	1	23	0	27	4	35	0	52	5	44	1	77	8	56	1
3	1	24	0	28	4	36	0	53	6	45	0	78	8	56	1
4	2	25	0	29	4	36	1	54	6	45	1	79	8	56	1
5	2	25	1	30	4	36	0	55	6	46	0	80	8	57	0
6	2	26	0	31	4	37	0	56	6	46	1	81	8	57	0
7	2	26	0	32	4	37	1	57	6	47	0	82	8	57	1
8	2	28	0	33	4	37	0	58	6	47	0	83	8	57	1
9	2	28	0	34	4	38	0	59	6	47	1	84	8	57	1
10	2	29	0	35	4	38	0	60	6	48	0	85	8	57	1
11	3	30	0	36	4	39	0	61	6	48	1	86	8	58	0
12	3	30	0	37	4	39	1	62	6	48	1	87	8	58	1
13	3	30	0	38	5	40	0	63	6	49	0	88	8	58	1
14	3	30	0	39	5	40	1	64	6	49	0	89	8	59	1
15	3	30	0	40	5	41	0	65	6	49	1	90	8	59	1
16	3	30	1	41	5	41	0	66	7	50	0	91	9	60	0
17	3	32	0	42	5	42	0	67	7	50	1	92	9	60	1
18	3	32	0	43	5	42	0	68	7	51	0	93	9	61	1
19	3	33	0	44	5	42	0	69	7	52	0	94	9	62	1
20	3	33	0	45	5	42	1	70	7	52	1	95	9	62	1
21	3	34	0	46	5	43	0	71	7	53	1	96	9	63	1
22	3	34	0	47	5	43	0	72	7	53	1	97	9	64	0
23	3	34	1	48	5	43	1	73	7	54	1	98	9	64	1
24	3	34	0	49	5	44	0	74	8	55	0	99	10	65	1
25	3	34	0	50	5	44	0	75	8	55	1	100	10	69	1

TAB. 3.1 – Tableau des données des malades

La figure 3.1, donne un aperçu sur les données du tableau précédent, de la présence ou absence de la maladie en fonction de l'âge.

3.2 Regroupement des données

Dans la suite, nous calculons la proportion de malades observée selon les classes (groupes) d'âge définies par la variable **grag**. Pour cela, on définit un vecteur noté (centre) qui donne les centres de chaque classe puis représenter le nuage de points de $p = 10$ centres en fonction

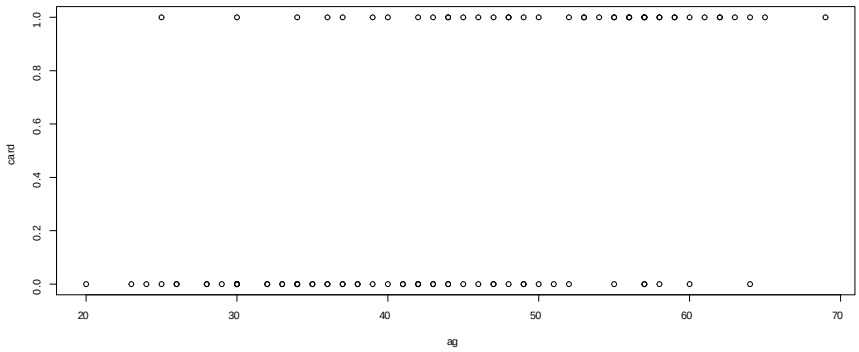

FIG. 3.1 – Maladie en fonction de l'age

de la variable **card**.

	centre	n_i	n_j	n	P
1	22.33	0	3	3	0.0000000
2	26.71	1	6	7	0.1428571
3	32.00	2	13	15	0.1333333
4	36.92	3	9	12	0.2500000
5	42.33	5	10	15	0.3333333
6	47.23	6	7	13	0.4615385
7	51.87	5	3	8	0.6250000
8	56.88	13	4	17	0.7647059
9	62.00	6	2	8	0.7500000
10	67.00	2	0	2	1.0000000

TAB. 3.2 – Tableau des proportions de malades par classe d'âge .

n_i : nombre de malades selon les classes d'âge.

n_j : nombre de personnes non malades selon les classes d'age.

n : nombre de patients selon les classes.

p : proportion de malades selon les classes d'âge.

centre : centre de chaque classe.

La figure 3.2 présente le nuage de points de la proportion de malades selon la classe d'âge.

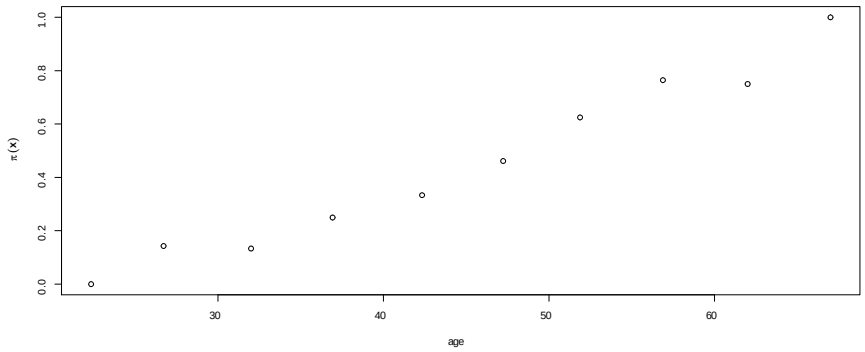

FIG. 3.2 – Proportion de malades selon la classe d'âge

3.3 Modélisation des données

Dans le langage R, la fonction **glm()** permet de faire différents types de régressions linéaires généralisées, ainsi que différents types de régressions non-linéaires. Aussi cette fonction permet d'ajuster différentes familles de modèles : logit, probit, etc. Il faut en revanche avoir recours à un argument supplémentaire : **family**, qui définit le type de modèle qu'on souhaite faire. Pour plus de détails regarder l'aide grâce au commande help(glm).

3.3.1 Modèle logistique

L'équation utilisée pour cette ajustement est celle du modèle logistique :

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 x_1)}{1 + \exp(\beta_0 + \beta_1 x_1)}.$$

Le modèle est donné sous forme linéaire, par la formule suivante :

$$\ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 X_1.$$

En utilisant la commande R suivante :

```
card.logit=glm(card~ag,family=binomial(link="logit"))
summary(card.logit)
```

Nous obtenons les résultats d'estimation suivants :

	Estimate	Std.Error	z value	$Pr(>\|z\|)$
Intercept	-5.30945	1.13365	-4.683	$2.82e-06$
ag	0.11092	0.02406	4.610	$4.0e-06$

TAB. 3.3 – Estimation et caractéristiques du modèle logistique

Alors, le modèle est

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} \quad \text{avec } \hat{\beta}_0 = -5.30945 \quad \text{et} \quad \hat{\beta}_1 = 0.11092.$$

- Std.Error : est l'écart-type des paramètres estimés : $\sigma_{\hat{\beta}_0}$ et $\sigma_{\hat{\beta}_1}$.
- z value : est la statistique de Wald.
- $Pr(>|z|)$ est la $p-value$: dans notre cas elle est $< 5\%$ donc l'hypothèse $H_0\,(B = 0)$ est rejeté et les deux paramètres sont significatifs.
- Les proportions observées et ajustées par le modèle logistique sont données par la figure 3.3 suivante, dont la ligne en bleu présente le modèle ajusté et les points en rouge sont les données.

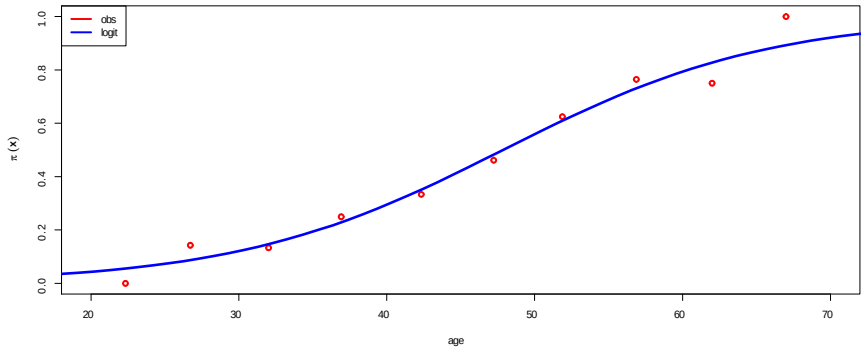

FIG. 3.3 – Proportions observées et ajustées

3.3.2 Modèle Probit

Dans la suite, nous utilisons la commande R suivante, pour l'ajustement des données par un modèle linéaire généralisé dit Probit :

```
card.probit=glm(card~ag,family=binomial(link="probit"))
summary(card.probit)
```

Nous obtenons les résultats d'estimation suivants :

	Estimate	Std.Error	z value	$Pr(>\|z\|)$
Intercept	-3.14573	0.62460	-5.036	4.74e-07
ag	0.06580	0.01335	4.930	8.20e-07

TAB. 3.4 – Estimation et caractéristiques du modèle Probit.

La signification des paramètres du modèle Probit est assuré par la $p - value << 5\%$, donc l'hypothèse $H_0\,(B = 0)$ est rejeté.

3.4 Prévision

Les proportions observées et ajustées par les modèles Probit et logistique sont données par la figure 3.4 suivante, dont la ligne en bleu présente le modèle logistique ajusté, la ligne en marron présente le modèle Probit ajusté et les points en rouge sont les données. Les deux modèles (logit et probit) que l'on voit dans la figure 3.4 sont assez semblables. En général les modèles logit et probit fournissent des valeurs très proches. Toutefois, pour des commodités de calcul, l'expression du probit, étant pas explicite, on préfère souvent le modèle logit.

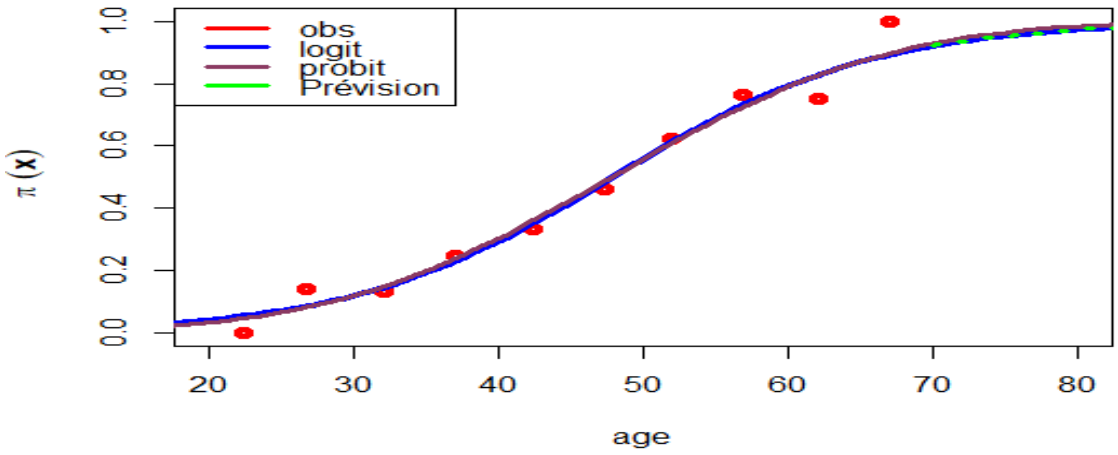

FIG. 3.4 – Proportions observées et ajustées

Comme application directe de la modélisation et l'ajustement des données par un modèle logit, nous proposons d'estimer et prédire, la cote des individus âgés de 70 à 80 ans. En utilisant le modèle ajusté :

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} \quad \text{avec } \hat{\beta}_0 = -5.30945 \quad \text{et} \quad \hat{\beta}_1 = 0.11092.$$

En remplaçant la variable x ($\hat{a}ge$) par $10, 11, ..., 20$ (personnes moin âgées) et par $70, 71, ..., 80$ (personnes plus âgées). Nous obtenons la probabilité $\pi(x)$, de la présence d'une maladie

cardiovasculaire d'une maladie cardiovasculaire pour une personne d'âge x.

Age	Prévision personnes plus âgées
70	0.9209283
71	0.9286376
72	0.9356479
73	0.9420124
74	0.9477827
75	0.9530074
76	0.9577326
77	0.9620017
78	0.9658549
79	0.9693299
80	0.9724613

Age	Prévision personnes moins âgées
10	0.01477051
11	0.01647465
12	0.01837173
13	0.02048271
14	0.02283062
15	0.02544067
16	0.02834046
17	0.03156006
18	0.03513221
19	0.03909234
20	0.04347876

TAB. 3.5 – Prévisions : personnes plus âgées et personnes moin âgées

On conclut du tableau précédent, que les personnes plus âgées sont les plus probablement touchées par l'apparition d'une maladie cardiovasculaire (un résultat semble très naturel et logique).

Remarque 3.4.1 *D'autres variables comme le poids et d'autres maladies et toute autre variable peuvent également être ajoutées à l'ensemble de données pour améliorer l'exactitude de l'analyse.*

Conclusion

Ce mémoire donne une aidé générale sur les modèles linéaires généralisés. Ce type de modèle est un outil polyvalent qui peut être utilisé dans de nombreuses situations pour analyser des variables présentant différentes distributions statistiques.

Dans ce mémoire, nous avons examiné plusieurs types de modèle linéaire généralisé avec leurs applications, notamment les cas où la variable suit une loi de Poisson ou une loi binomiale, qui sont parmi les plus courants. D'autres distributions peuvent être appliquées en utilisant à chaque fois une fonction de lien appropriée.

Nous terminons ce travail par une application sur les modèles, présentés auparavant pour montrer l'importance de ces modèles statistique dans la réalité.

Bibliographie

[1] Baccini, A., Besse, P., Dejean, S. (2008). Analyse statistique de données d'expression. Publications du laboratoire de Statistique et Probabilité. Université Toulouse.

[2] Besse, P. (2003). Pratique de la modélisation statistique. Publications du laboratoire de Statistique et Probabilité. Université Toulouse.

[3] Chikhi, M., Chavance, M. (2012). Estimation du Modèle Linéaire Généralisé et Application. Sciences & Technology. A, Exactes Sciences, 13-21.

[4] Chesneau, C. (2017). Modèles de régression. Université de Caen.

[5] Dunn, P. K., Smyth, G. K. (2018). Generalized linear models with examples in R (Vol. 53, p. 16). New York : Springer.

[6] Droesbeke, J.J., Lejeune, M., Saporta, G. (2005). Modèles statistiques pour données qualitatives. Editions Technip.

[7] Dobson, A.J. (2001). An Introduction to Generalized Linear Models. Chapman and Hall, London, second edition.

[8] Dobson, A.J., Barnett, A. G. (2018). An introduction to generalized linear models. Chapman and Hall.

[9] Fermin, A. (2015). Le Modèle linéaire généralisé (glm). https ://fermin.perso.math.cnrs.fr/Files/Modele-Logistique.pdf

[10] McCullagh, P. Nelder, J. A. (1989). Generalized Linear Models, volume 37 of Monographson Statistics and Applied Probability. Chapman and Hall, London, 2 edition.

[11] Müller, Marlene. (2004). Generalized Linear Models. http ://dx.doi.org/10.1007/978-3-642-21551-3_24.

[12] Nelder, J. A., Wedderburn, R.W.M. (1972). Generalized Linear Models. Journal of the Royal Statistical Society. Series A (General), 135(3), 370–384.

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

Annexe : Abréviations et Notations

Les différentes abréviations et notations utilisées tout au long de ce mémoire sont expliquées ci-dessous :

$E(.)$	:	Espérence mathématique.
$V(.)$	:	Variance mathématique.
$tr(A)$	:	Trace de matrice A.
GLM	:	Modèle linéaire généralisé.
SCT	:	Sommes descarrés totale.
SCA	:	Sommes descarrés Factorielles.
SCR	:	Sommes descarrés résiduelle.
$\mathcal{N}_n$	:	la loi normale dans$\mathbb{R}^n$(la loi normale multivariée).
MCO	:	Estimateurs des Moindres Carrés Ordinaires.
$i.i.d$	:	indépendant identiquement distribué.
$pré(.)$	:	Prévision.
$M.V$	:	maximum vraisemblance.
L	:	vraisemblance.
ℓ	:	log-vraisemblance.
$X^2_{(n)}$	:	Loi du chi-deux à $n \in N$ degrés de liberté.

ε	:	L'erreure.
R^2	:	Coefficient de détermination.
$H(.)$	:	Matrice Hessienne.
$I(.)$	:	Matrice d'information de Fisher.
W	:	La statistique de Wald.
X^t	:	Transposée de X.
η	:	vecteur de prédicteurs linéaire.
$t_{\left(1-\frac{\alpha}{2}\right)}$	:	Fractile d'ordre $\left(1-\frac{\alpha}{2}\right)$ de loi de Student.
T_{n-p-1}	:	Loi de student à $(n-p-1)$ degrés de liberté.
$f_{(1-\alpha)}$	:	Fractile d'ordre $(1-\alpha)$ de loi de Fisher.
$F(p, n-p-1)$	:	Loi de Fisher degrés de liberté (p) et $(n-p-1)$.

Résumé

Dans ce mémoire, nous présentons les différents modèles statistiques utilisés pour résoudre des problèmes pratiques, en mettant particulièrement l'accent sur les modèles linéaires généralisés tels que le modèle de dénombrement et la régression logistique. Nous donnons également, une application, sous logiciel statistique R, sur des données réelles de la maladie cardiovasculaire selon les classes d'âge des individus étudier.

Abstract

In this master dissertation, we present the different statistical models used to solve practical problems, with particular emphasis on generalized linear models such as the enumeration model and logistic regression. We also give an application, using R statistical software, on real data on cardiovascular disease according to the age groups of the individuals studied.

ملخـــــص

نقدم في هذه المذكرة أهم النماذج الإحصائية المختلفة المستخدمة لحل المشكلات العملية، مع التركيز بشكل خاص على النماذج الخطية المعممة مثل نموذج التعداد والانحدار اللوجستي. كما نقدم أيضًا تطبيقًا، باستخدام البرنامج الإحصائي R، على بيانات حقيقية عن أمراض القلب والأوعية الدموية وفقًا للفئات العمرية للافراد الذين خضعوا للدراسة.