République Algérienne Démocratique et Populaire
Ministère de l'enseignement Supérieur et de la Recherche Scientifique
Université Mohammed Seddik Ben Yahia - Jijel

Faculté des Sciences Exacte et Informatique
Département de Mathématique

Mémoire de fin de cycle

Présenté pour l'obtention du diplôme de

Master

Spécialité : Mathématiques.

Option : Probabilités et Statistique.

Thème

Modèles Linéaires Généralisés : Présentation et Applications

Présenté par :

Beltoum Latifa

Devant le jury composé de :

Madi Meriem	M.A.A Université de Jijel	**Président**
Guiatni Ahlam	M.A.B Université de Jijel	**Encadrant**
Cheraitia Hassen	M.C.A Université de Jijel	**Rapporteur**
Abdi Zeyneb	M.A.A Université de Jijel	**Examinateur**

Promotion **2021/2022**

Remerciements

Je tiens à exprimer toute ma reconnaissance à ma directrice de mémoire, Madmoiselle ***AHLAM GUIATNI****. Je la remercie de m'avoir encadré, orienté, aidé et conseillé.*

Je tiens à remercier sincèrement Monsieur ***HASSEN CHERAITIA****, pour son aide précieuse et pour le temps qu'il m'a consacré.*

Je remercie également toute l'équipe pédagogique de l'université de Mohammed Seddik Ben Yahya et les intervenants professionnels responsables de ma formation, pour avoir assuré la partie théorique de celle-ci.

Je remercie mes très chers parents qui ont toujours été là pour moi. Mes sœurs et mon frère, pour leurs encouragements.

Je voudrais exprimer aussi ma reconnaissance envers les amis et collègues qui m'ont apporté leur soutien moral et intellectuel tout au long de ma démarche.

Dédicace

À tous ceux qui me sont chers, Je dédie le fruit de mes 19 ans d'études.

Table des matières

Liste des tableaux

Table des figures

Notations

- $MLGs$: Les modèles linéaires généralisés.
- MLG : Les modèles linéaires générales.
- MCO : Moindre carrée ordinaire.
- E : Espérance .
- Var : La variance .
- $ANOVA$: L'analyse de la variance .
- $ANCOVA$: L'analyse de la covariance.
- SCT : Somme des carrés totale .
- SCE : Somme des carrée factorielle.
- SCR : Somme des carrés résiduelle. .
- cov : Covariance.
- ∇ : Le gradient.
- ℓ : la vraisemblance.
- $\mathcal{L}$: La log vraisemblance.
- P : Probabilité.
- $A = (a_{ij})$: La matrice de i ligne et j colonne.

Introduction Générale

Une grande partie des mathématiques appliquées consiste, d'une certaine façon, à faire de la modélisation, c'est-à-dire à définir un (ou plusieurs) modèle(s), de nature mathématique, permettant de rendre compte, d'une manière suffisamment générale, d'un phénomène donné, qu'il soit physique, biologique, économique ou autre.

De façon un peu schématique, on peut distinguer la modélisation déterministe (au sein d'un modèle déterministe, on ne prend pas en compte de variations aléatoires) et la modélisation stochastique (qui prend en compte ces variations aléatoires en essayant de leur associer une loi de probabilité).

Au sein de la modélisation stochastique, la modélisation probabiliste a surtout pour but de donner un cadre formel permettant, d'une part de décrire les variations aléatoires, d'autre part d'étudier les propriétés générales des phénomènes qui les régissent. Plus appliquée, la modélisation statistique consiste essentiellement à définir des outils appropriés pour modéliser des données observées, en tenant compte de leur nature aléatoire. Il faut noter que le terme de modélisation statistique est très général et que, à la limite, toute démarche statistique en relève. Toutefois, ce qui est traité dans ce mémoire est relativement précis et constitue une partie spécifique de la modélisation statistique.

Les méthodes de modélisation statistique sont, en fait, très nombreuses. La régression (simple/multiple), ANOVA et ANCOVA sont considérées comme un cas particulier de ces modèles statistiques qui s'appelle les modèles linéaires générales (MLG) et qui sont très utiles mais uniquement sur certaines conditions, la linéarité, l'homoscédasticité et la normalité des résidus, ces hypothèses ne sont pas toujours satisfaites et pour cela on peut dire que les MLG sont limitées.

Pour tenir compte de ces points, plusieurs solutions s'offrent à nous. La plus répandue consiste à trouver une transformation mathématique de la variable à expliquer pour la rendre normale (et son erreur avec) et pour en stabiliser les variances. On parle de transformations « normalisantes ». Ces transformations ne sont pas toutes efficaces, et leur effet normalisant est parfois difficile à quantifier. Par ailleurs, il reste évidemment préférable

d'utiliser les données d'origine plutôt que leurs valeurs transformées, ne serait-ce que pour rendre l'interprétation des résultats plus aisée. C'est dans ce cadre que se développe le modèle linéaire généralisé (MLGs).

La théorie des modèles linéaires généralisés a été formulée par John Nelder et Robert Wedderburn comme un moyen d'unifier les autres modèles statistiques y compris la régression linéaire, la régression logistique et la régression de Poisson.

L'idée reste d'utiliser une transformation mathématique sur la variable à expliquer mais en tenant compte cette fois-ci de la véritable distribution des erreurs (par exemple, une loi de Poisson dans le cas de comptages ; une loi Binomiale dans le cas de pourcentages, etc.). Ceci implique entre autre que les paramètres ne sont alors plus estimés par la simple méthode des moindres carrés – comme dans le modèle linéaire général – mais par une autre méthode d'estimation : la méthode dite du « maximum de vraisemblance ». La fonction mathématique utilisée pour transformer la variable à expliquer est appelée « fonction de lien », et plusieurs peuvent être utilisées selon la distribution réelle de la variable d'intérêt (et de son erreur). Autour de ces points, nous avons organisé notre mémoire en trois chapitres :

Le premier chapitre est intitulé : « préliminaire » qui a pour objectif de présenter les fondements théoriques des lois de probabilités, les différentes méthodes d'estimation et les tests statistiques.

Le deuxième chapitre est intitulé : « Les modèles linéaires généralisés » qui discute dans une première partie la présentation de ces modèles avec leur estimation et les différents tests utilisés dans ces modèles, en deuxième partie on présente quelques types des MLGs : logistique et poisson.

Le troisième chapitre est intitulé : « Application des modèles linéaires généralisés » qui vise essentiellement à : présenter les différents modèles (logistique et poisson) avec des données réels, estimer ces paramètres, et interpréter les résultats obtenus selon le logiciel statistique R.

1 Préliminaires

1.1 Calculs matricielles et les lois de probabilitées

1.1.1 Calculs matriciells

Définition 1.1.1. *Une matrice $A = (a_{ij})$ de type $m \times n$ est un tableau rectangulaire comprenant m lignes et n colonnes formées de nombres réels.*
L'élément situé au croisement de la i-ème ligne et de la j-ième colonne est noté a_{ij}.

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \dots & a_{mn} \end{pmatrix}$$

Définition 1.1.2.

- *Une matrice de type $n \times n$ est dite carrée d'ordre n*
- *Dans une matrice carrée, la diagonale formée par les éléments a_{ii} s'appelle la diagonale principale.*
- *Une matrice de type $1 \times n$ est appelée matrice ligne.*
- *Une matrice de type $n \times 1$ est appelée matrice colonne.*
- *Une matrice carrée de type $n \times n$ est appelée matrice identité si $a_{ij} = 0$, ij et $a_{ii} = 1$. On la note I_n*
- *Une matrice de type $m \times n$ composée uniquement de zéro est appelée matrice nulle. On la note $0_{m \times n}$.*

Définition 1.1.3. *Si $A = (a_{ij})$ et $B = (b_{ij})$ sont deux matrices de même type, leur somme $A + B$ est la matrice de même type obtenue en additionnant les tableaux élément*

par élément :

$$A + B = (c_{ij}) \ \ avec \ \ c_{ij} = a_{ij} + b_{ij}$$

Définition 1.1.4. *Le produit de deux matrices A et B est défini si le nombre de colonnes de A est égal au nombre de lignes de B.*
Si $A = (a_{ij})$ est une matrice de type $m \times n$ et $B = (b_{ij})$ est une matrice de type $n \times p$ alors le produit $AB = (c_{ij})$ est la matrice de type $m \times p$ définie par :

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \dots & a_{mn} \end{pmatrix} \times \begin{pmatrix} b_{11} & b_{12} & b_{13} & \dots & b_{1p} \\ b_{21} & b_{22} & b_{23} & \dots & b_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ b_{n1} & b_{n2} & b_{n3} & \dots & b_{np} \end{pmatrix} = \begin{pmatrix} c_{11} & c_{12} & c_{13} & \dots & c_{1n} \\ c_{21} & c_{22} & c_{32} & \dots & c_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_{m1} & c_{m2} & c_{m3} & \dots & c_{mp} \end{pmatrix}$$

avec :

$$\begin{aligned} c_{ij} &= a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots + a_{n1}b_{nj} \\ &= \sum_{i=1}^{n} a_{ik}b_{kj} \quad i = 1, \dots, m; j = 1, \dots, p \end{aligned}$$

Propriétés 1.1.1. *Le produit matriciel vérifie les propriétés suivantes :*

- $A(BC) = (AB)C$ *avec* A_{mn}, B_{np} *et* C_{pq}.
- $AI_n = I_m A = A$ *avec* $A_{m\ddot{O}n}$ *et les matrices identités* I_n *et* I_m.
- $A(B + C) = AB + AC$ *avec* A_{mn} *et* B_{np}, C_{np}.
- $(A + B)C = AC + BC$ *si* A_{mn} , B_{mn} *et* C_{np}.

Définition 1.1.5. *Soit A est une matrice carrée d'ordre n et $detA \neq 0$. S'il existe une matrice B telle que :*

$$AB = I_n = BA$$

alors B est appelée la matrice inverse de A (codée A^{-1}).

Définition 1.1.6. *La transposée d'une matrice A s'obtient en remplaçant les lignes de la matrice par ses colonnes. Si la matrice A est de dimension $m \times n$, la transposée A' , sera de dimension $n \times m$.*

$$A' = \begin{pmatrix} a_{11} & a_{21} & a_{31} & \dots & a_{n1} \\ a_{12} & a_{22} & a_{32} & \dots & a_{n2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{1m} & a_{2m} & a_{3m} & \dots & a_{mn} \end{pmatrix}' = \begin{pmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \dots & a_{mn} \end{pmatrix}$$

Propriétés 1.1.2. *Soit A et B deux matrices et k un scalaire*

- $(A+B)' = A' + B'$
- $(A')' = A$
- $(kA)' = kA'$
- $(AB)' = B'A'$

Définition 1.1.7. *Soit $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ une fonction admettant des dérivées partielles. Le gradient en $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$, noté $gradf(x)$, est le vecteur :*

$$gradf(x) = \begin{pmatrix} \frac{\partial f}{\partial x_1}(x) \\ \vdots \\ \frac{\partial f}{x_n}(x) \end{pmatrix}$$

1.1.2 Lois de probabilitées

Dans tout ce qui suit nous nous placerons dans l'espace probabilisé (Ω, P).

Variables aléatoires : généralités.

Une **variable aléatoire**, X, est une application mesurable de Ω dans $\mathbb{R}$.
E une partie de $\mathbb{R}$, $(X \in E) = X^{-1}(E) = \{\omega \in \Omega | X(\omega) \in E\}$ est **l'image réciproque** de E par l'application X.
La loi de probabilité (ou **loi**) de X est l'application P_X définie sur les parties E de $\mathbb{R}$ par :

$$P_X(E) = P(X \in E)$$

La loi image de P par l'application X est une probabilité sur $\mathbb{R}$.
La fonction de répartition F_X de la variable aléatoire X (ou de la loi P_X) est la fonction définie sur $\mathbb{R}$ par :

$$F_X(x) = P_X(]\infty, x]) = P(X \leq x)$$

Elle est à valeurs dans $[0, 1]$.
Deux variables aléatoires. X et Y sont dites **indépendantes** si pour toutes parties E et F de $\mathbb{R}$:

$$P((X \in E) \cap (Y \in F)) = P(X \in E)P(Y \in F)$$

Variables aléatoires discrètes :

On dit qu'une variable aléatoire est discrète si $X(\Omega)$ est fini ou dénombrable. On notera $X(\Omega) = \{x_1, x_2, \dots\}$.

Dans le cas de variable aléatoire discrète :

- La loi de probabilité d'une variable aléatoire discrète X est donnée par $p_i = P(X = x_i)$ pour tous les x_i de $X(\Omega)$.
- La fonction de répartition d'une variable aléatoire discrète X est une fonction en escaliers et $P_X(x_i) = F_X(x_i) - F_X(x_{i-1})$.
- Deux variables aléatoires discrètes X et Yprenant respectivement pour valeurs $X(\Omega) = \{x_1, x_2, \dots\}$ et $Y(\Omega) = \{y_1, y_2, \dots\}$ sont indépendantes ssi pour tous x_i et y_j :

$$P((X = x_i) \cap (Y = y_j)) = P(X = x_i)P(Y = y_j)$$

Loi Bernoulli :

Une épreuve de Bernoulli est une expérience aléatoire dont le résultat peut ^etre soit un succès, soit un échec, mais pas les deux simultanément.

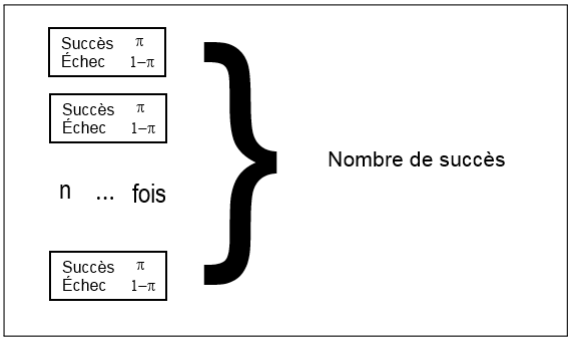
Une variable aléatoire X est dite variable aléatoire de Bernoulli de paramètre p, (pour $p \in [0, 1]$) si elle est à valeurs dans $D = \{0, 1\}$ et si :

$$P_X(1) = P(X = 1) = p;$$
$$P_X(0) = P(X = 0) = 1 - p.$$

Loi binomiale :

Considérons l'expérience qui consiste à répéter n fois une expérience aléatoire de façon indépendante telle que le résultat de chaque expérience est un succès ou un échec avec une probabilité de succès π.

On peut représenter cette expérience type par la figure suivante :



Loi de binomiale notée $B(n, \pi)$ modélise les expériences où l'on répète n fois de façons indépendantes une épreuve de Bernoulli et on compte le nombre de réussites.
Le support de cette variable aléatoire est :

$$X(\Omega) = \{0, 1, 2, \ldots, n\}$$

pour $x = \{0, 1, 2, \ldots, n\}$ la loi de probabilité est donnée par :

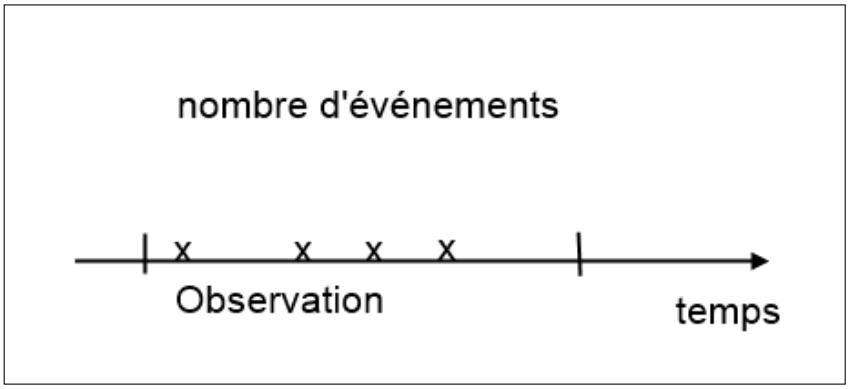
$$f(x) = \binom{n}{x} \pi^x (1-\pi)^{n-x}$$

Les principales caractéristiques numériques sont :

$$\text{Moyenne} : E(X) = n\pi.$$
$$\text{Variance} : Var(X) = n\pi(1-\pi)$$
$$\text{Écart type} : \sqrt{n\pi(1-\pi)}$$

Loi Poisson

La loi de Poisson ou modèle de Poisson permet la modélisation de l'observation d'un phénomène qui produit des événements à un rythme connu.
On s'intéresse à l'observation d'événements et on suppose :
1. un seul événement arrive à la fois
2. le nombre d'événements se produisant ne dépend que du temps de l'observation.
3. les événements sont indépendants.



La loi de Poisson notée $X \sim P(\lambda)$, où X le nombre d'événements observés dans une unité de temps, où λ représente le nombre moyen d'événements par unité de temps.
Le support de cette variable aléatoire est :

$$X(\Omega) = \{0, 1, 2, \ldots, n\} = \mathbb{N}$$

pour $x = 0, 1, 2, \ldots, \mathbb{N}$ la loi de probabilité est :

$$P_X(x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

avec :

$$\text{Moyenne} : E(X) = \lambda$$
$$\text{Variance} : Var(X) = \lambda$$
$$\text{Écart type} : \sqrt{\lambda}$$

Les lois de probabilité continues

Soit X une variable aléatoire réel . qui prend un nombre infini non dénombrable de valeurs. Si F_X est une fonction continue, on dit que X est une variable aléatoire réel continue.
Dans ce cas, la loi de X est déterminée par l'ensemble des probabilités $P(a < X < b)$, pour tout $a < b$.
La fonction de répartition d'une variable continue est :

$$F_X(t) = \int_{-\infty}^{t} f(x)\,\mathrm{d}x$$

Où f est la densité de probabilité tels que :

$$f(x) = \frac{dF_X(x)}{dx}$$

La loi exponentielle :
La loi exponentielle donne le temps d'attente avant un événement lorsque le processus est régi par une loi de Poisson.
Dans le cas de la loi de Poisson la variable aléatoire était le nombre d'événements tandis que dans la loi exponentielle c'est le temps d'attente avant le premier événement. Il est à noter que le nombre d'événements est une v.a. discrète tandis que le temps d'attente est une variable aléatoire continue. La variable aléatoire qui donne le temps d'attente avant le nouveau phénomène de Poisson est une loi exponentielle de paramètre λ=temp moyen, pour $0 \leq x \leq \infty$ sa densité est donnée par :

$$f_X(x) = \frac{e^{\frac{-x}{\lambda}}}{\lambda}$$

La probabilité $P(X \leq t)$ est donnée par : $1 - e^{\frac{-x}{\lambda}}$ La moyenne est λ et la variance est λ^2.

Loi normale :

La loi normale est très importante en statistique : plusieurs phénomènes ont une loi de probabilité très proche de la loi normale, suffisamment proche pour utiliser cette dernière pour les modéliser. De plus, elle est souvent utilisée pour faire des approximations dans le domaine de l'estimation et des tests d'hypothèses.

Plus formellement, une loi normale est une loi de probabilité absolument continue qui dépend de deux paramètres :

- l'espérance, un nombre réel noté μ.
- l'écart type, un nombre réel positif noté σ.

La densité de probabilité de la loi normale d'espérance μ, et d'écart type σ est donnée par :

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)$$

La densité a la forme suivante :

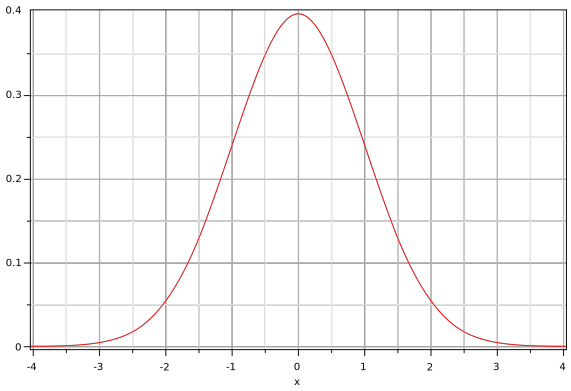


FIGURE 1.1 – La loi normale centrée réduite

1.2 Estimation des paramètres du modèle

L'estimation des paramètres est une méthode visant à attribuer une valeur au paramètre ou à l'ensemble de paramètres qui caractérisent le domaine d'étude.

1.2.1 Définition d'un estimateur

On considère un échantillon $(X_k)_{1\leq k\leq n}$ tel que chaque X_k est à valeurs dans un ensemble E (par exemple, $E = \mathbb{R}$) et de même loi P_θ qui dépend d'un paramètre inconnu θ.

Le but est d'estimer la valeur de θ à partir des valeurs des X_k.

Définition 1.2.1. *Un estimateur de θ est une variable aléatoire $\hat{\theta}_n$ telle qu'il existe une fonction $F_n : E_n \to \theta$ avec $\hat{\theta}_n = F_n(X_1, X_2, \ldots, X_n)$; c'est donc une fonction de l'echantillon.*

1.2.2 Propriétés d'un estimateur

Définition 1.2.2. *Un estimateur est **sans biais** si la moyenne de sa distribution d'échantillonnage est égale à la valeur θ du paramètre de la population à estimer, c'est-à-dire si :*

$$E(\hat{\theta}) = \theta$$

Définition 1.2.3. *Un estimateur θ est **convergent** si sa distribution tend à se concentrer autour de la valeur inconnue à estimer θ, à mesure que la taille d'échantillon augmente, c'est-à-dire :*

$$\lim_{n \to +\infty} Var(\hat{\theta}) = 0$$

Définition 1.2.4. *Un estimateur sans biais est **efficace** si sa variance est la plus faible parmi les variances des autres estimateurs sans biais.*
Ainsi, si $\hat{\theta}_1$ et $\hat{\theta}_2$ sont deux estimateurs sans biais du paramètre θ, l'estimateur $\hat{\theta}_1$ est efficace si :

$$Var(\hat{\theta}_1) \leq Var(\hat{\theta}_2) \;\; et \;\; E(\hat{\theta}_1) = E(\hat{\theta}_2) = \theta$$

1.2.3 Méthodes d'estimation

Il existe plusieurs méthodes d'estimations :

Méthode des Moindres carrés

Définition 1.2.5. *La méthode des moindres carrés consiste à estimer θ en minimisant la somme des carrés des résidus SCR telle que :*

$$S(\hat{\theta}(y)) = argmin \sum_{i=1}^{n} (\hat{\epsilon}_i^2) = argmin \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

avec :

$$\frac{\partial S(\hat{\theta}(y))}{\partial \beta_j} = 0$$

Méthode de maximum de vraisemblance

Définition 1.2.6. *La statistique* $\omega \to argmax(\theta \to \prod_{i=1}^{n} f_\theta(X_i(\omega)))$ *s'appelle l'estimateur de* ***maximum de vraisemblance*** *de* θ *.*
$L : \theta \to \prod_{i=1}^{n} f_\theta(x_i)$ *s'appelle la fonction vraisemblance du modèle.*
$\ell : \theta \to \sum_{i=1}^{n} log f_\theta(x_i)$ *s'appelle la fonction log- vraisemblance du modèle.*
pour obtenir l 'estimateur $\hat{\theta}$ *du de maximum de vraisemblance ,on maximise log- vraisemblance selon* θ *, on résolvant le système d'équation maximum de vraisemblance*

$$\frac{\partial}{\partial \theta_j} ln(_1, \ldots, \theta_k, y) = 0. \ pourj = 1, \ldots, k$$

Méthode d'intervalle de confiance

Soit Y une variable aléatoire dont la loi dépend d'un paramètre réel θ inconnu et $\alpha \in [0, 1]$ un nombre donné. On appelle « intervalle de confiance » pour le paramètre θ, de niveau de confiance $1 - \alpha$, un intervalle qui a la probabilité $1 - \alpha$ de contenir la vraie valeur du paramètre θ.

1.3 Tests d'hypothèses

Un test statistique est une procédure de décision entre deux hypothéses, l'une est dite l'hypothèse nulle notée H_0 et l'autre est dite alternative notée H_1, concernant un ou plusieurs echantillons.

1.3.1 Hypothèses et risques d'erreur

une démarche de test suit généralement les étapes suivantes :

- Choix de H_0 et de H_1. Fixer α.
- Détermination de la statistique de test.
- Allure de la région de rejet en fonction de H_1.

- Calcul de la région de rejet en fonction de α et H_0.
- Calcul de la valeur observée de la statistique de test.
- Conclusion : rejet ou acceptation de H_0 au risque α.
- Si possible, calcul de la puissance du test : $1 - \beta$.

Définition 1.3.1. *Le risque de première espèce associé à un test T est la probabilité de commettre l'erreur de première espèce. On le note traditionnellement* α *:*

$\alpha = P($ *rejeter* $H_0 | H_0$ *est vraie* $)$

Définition 1.3.2. *Le risque de seconde espèce associé à un test T est la probabilité de commettre l'erreur de seconde espèce. On le note traditionnellement* β *:*

$\beta = P($ *accepter* $H_0 | H_1$ *est vraie* $)$

Définition 1.3.3. *la puissance d'un test T est La quantité* $1 - \beta$ *. On note que par définition :* $1 - \beta = P($ *rejeter* $H_0 | H_1$ *est vraie* $)$*.*
En d'autres termes, la probabilité de rejeter H_0 *alors que* H_0 *devrait (idéalement) être rejetée est la puissance du test.*

Le tableau ci dessus illustre les cas possibles schématisés pour choisir la décision convenable :

RéalitéDécision	Accepter H_0	Rejeter H_0
H_0vraie	$1 - \alpha$	α
H_1vraie	β	$1 - \beta$

TABLE 1.1 – Différentes probabilités dans un test d'hypothèses

2 Fondements théoriques des modèles linéaires généralisés

Les modèles linéaires généralisés (MLGs) sont une classe générale des modèles statistiques. La régression (simple/multiple), ANOVA et ANCOVA sont considérées comme un cas particulier de cette classe qui s'appelle les modèles linéaires générales(MLG) et qui sont très utiles mais uniquement sur certains conditions, la linéarité, homogénéité des variances et la normalité des résidus, ces hypothèses ne sont pas toujours satisfaites et pour cela on peut dire que les MLG sont limitées.

La généralisation des MLG qui est les MLGs nous permets de modéliser notre data en utilisant des distributions pas forcément normales, mais qui appartiennent à une famille dite exponentielle. La famille exponentielle comprend un bon nombre des distributions, les plus courantes parmi elle : la distribution Normale, exponentielle, Gamma, Poisson, Bernoulli...etc.

2.1 Présentation du modèle

Les modèles linéaires généralisés sont définit par trois composantes :

2.1.1 1. Distribution

La famille exponentielle est un ensemble des distributions de probabilité peut être s'écrit sous la forme spécifie ci dessous :

$$f(y, \theta, \phi) = exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right\} \tag{2.1}$$

avec :

- θ : le paramètre naturel de la famille exponentielle.
- $b(.), c(.)$: des fonctions varient d'une famille exponentielle à une autre.
- $a(\phi)$: pour certains lois, la fonction a est de la forme :

$$a(\phi) = \frac{\phi}{\omega_i}$$

où les poids ω_i sont connus des observations, fixés à 1 pour simplifier.
ϕ : est appelé *paramètre de dispersion*, c'est un paramètre de nuisance intervenant, il égal à 1 pour les lois à un paramètre (Poisson) et peut être estimé pour les autres distributions.

Supposons maintenant que l'échantillon statistique est constitué à n variables aléatoires $\{Y_i : i = 1, ..., n\}$ indépendantes admettant des distributions issues d'une structure exponentielle, les lois de ces variables sont donc dominées par une même mesure de référence et que la famille de leurs densités par rapport à cette mesure se met sous la forme :

$$f(y_i, \theta, \phi) = exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right\} \tag{2.2}$$

La formule peut s'écrit sous la forme dite *canonique* :

$$\begin{aligned} f_c(y_i, \theta_i, \phi) &= exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right\} \\ &= exp\left\{\frac{y_i\theta_i}{a(\phi)}\right\} exp\left\{\frac{-b(\theta_i)}{a(\phi)}\right\} exp\left\{c(y_i, \phi)\right\} \\ &= exp\left\{y_iQ(\theta_i)\right\}\ \mu(\theta_i)\ b(y_i) \end{aligned} \tag{2.3}$$

où :

$$\begin{aligned} Q(\theta_i) &= \frac{\theta_i}{a(\phi)} \\ u(\theta_i) &= exp\left\{\frac{-b(\theta_i)}{a(\phi)}\right\} \\ b(y_i) &= exp\left\{c(y_i, \phi)\right\} \end{aligned}$$

2.1.2 Prédicteur linéaire

Le prédicteur linéaire est la partie déterministe de modèle, il joue un rôle similaire dans les MLGs comme dans les MLG, pour β un vecteur de p paramètres et X la matrice de planification des expériences de taille $n\times\ p$, le prédicteur linéaire est :

$$\eta = X\beta$$

avec :

$$X = \begin{pmatrix} X_1^t \\ X_2^t \\ \vdots \\ X_n^t \end{pmatrix} = \begin{pmatrix} X_{11} & X_{12} & \dots & X_{1p} \\ X_{21} & X_{22} & \dots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \dots & X_{np} \end{pmatrix}$$

tel que :

$$\mathrm{X}_i^t = \begin{pmatrix} X_{i1} \\ X_{i2} \\ \vdots \\ X_{ip} \end{pmatrix} \qquad et \qquad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}$$

2.1.3 Fonction de lien

La troisième composante des modèles linéaires généralisées exprime une relation fonctionnelle entre la moyenne naturelle de réponse (variable à expliqué) et le prédicteur linéaire.
Soit $\{\mu_i = E(Y_i), i = 1...n\}$, la fonction de lien est définie alors par :

$$\eta_i = g(\mu_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_p X_{np}$$

g est une fonction supposée monotone et différentiable, on peut donc définir l'inverse de cette fonction avec la relation :

$$g^{-1}(g(u)) = \mu$$

Dans la famille exponentielle, il y a des distributions où un certains paramètres servent de paramètre naturelle. ces paramètres sont respectivement : la moyenne, le log odds, le

logarithme de la moyenne pour les distributions : la loi normale, binomiale et la distribution de poisson.
Dans ce cas, la fonction de lien qui associe la moyenne μ_i au paramètre naturel est appelée *fonction de lien canonique* et on écrit :

$$g(\mu_i) = \theta_i \qquad i = 1, ..., n$$

où θ c'est le paramètre naturel de distribution.

Exemple 2.1.1.
Dans le cas d'un échantillon Gaussien, la famille Gaussienne se met sous la forme canonique (2.1) qui en fait une famille exponentielle de paramètre de dispersion $\phi = \sigma^2$ et de paramètre naturelle

$$\theta_i = E(Y_i) = \mu_i$$

la fonction de lien canonique donc c'est la fonction identité, l'inverse de cette fonction est simplement :

$$\mu_i = \theta_i$$

Le tableau 1.1 illustre les fonctions de liens, les supports et les moyennes de quelques distributions.

	Support de distribution	Fonction de lien	nom de fonction de lien	la fonction moyenne
normal $N(\mu, \sigma^2)$	$(-\infty; +\infty)$	μ	identité	θ
Exponentielle $Exp(\mu)$	$(0; +\infty)$	$\frac{1}{\mu}$	Inverse	$\frac{-1}{\theta}$
Gamma $G(\mu, \sigma)$	$(0; +\infty)$	$\frac{1}{\mu}$	Inverse	$\frac{-1}{\theta}$
Poisson $P(\mu)$	$\{0, 1, 2, \dots\}$	$log(\mu)$	Log	$exp(\theta)$
Bernoulli $B(\mu)$	{0,1}	$log(\frac{\mu}{1-\mu})$	Logit	$\frac{exp(\theta)}{1+exp(\theta)}$
Binomial $B(k, \mu)$	0,1,...,N	$log(\frac{\mu}{k-\mu})$	Logit	$\frac{kexp(\theta)}{1+exp(\theta)}$
Géométrique $Geo(\mu)$	{0,1,2,...}	$log(\frac{\mu}{1+\mu})$	Logit	$\frac{exp(\theta)}{1-exp(\theta)}$
Gaussien inverse $IG(\mu, \sigma^2)$	$(0, +\infty)$	$\frac{1}{\mu^2}$	Le carré inverse	$\frac{1}{\sqrt{-2\theta}}$

TABLE 2.1 – Distributions communes avec des fonctions de lien canoniques.

2.2 Estimation par Maximum de vraisemblance

2.2.1 La théorie de maximum de vraisemblance

Dans les modèles linéaires généralisées, l'estimation des paramètres β_j se fait par la méthode de maximum de vraisemblance, les estimations sont les valeurs des paramètres qui maximisent la log-vraisemblance de modèle.
Notons $\ell(\theta_i, \phi, y_i)$ la contribution de la $i$ème observation à la log de vraisemblance :

$$\begin{aligned} \ell(\theta_i, \phi, y_i) &= lnf(y_i, \theta_i, \phi) \\ &= \frac{y_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \end{aligned}$$

pour des lois issues de structure exponentielles, les conditions de régularité vérifiées permettent d'écrire :

$$\begin{cases} E\left(\frac{d\ell}{d\theta}\right) = 0 \\ E\left(\frac{d^2\ell}{d\theta^2}\right) + E\left(\frac{d\ell}{d\theta}\right)^2 = 0 \end{cases}$$

sachant que les dérivées sont :

$$\begin{cases} \frac{d\ell}{d\theta_i} = \frac{(y_i - b'(\theta_i)}{a(\phi)} \\ \frac{d^2\ell}{d\theta_i^2} = -\frac{b''(\theta_i)}{a(\phi)} \end{cases}$$

où b et b''expriment la première et la deuxième dérivées, respectivement, de b par rapport à θ.

de (1) et (3) on obtient :

$$E\left(\frac{d\ell}{d\theta_i}\right) = E\left(\frac{y_i - b'(\theta_i)}{a(\phi)}\right) = 0$$

il vient donc :

$$E(y_i) = \mu_i = b'(\theta_i)$$

et comme :

$$E\left(\frac{d^2\ell}{d\theta^2}\right) = -E\left(\frac{d\ell}{d\theta}\right)^2$$

$$E\left(-\frac{b''(\theta_i)}{a(\phi)}\right) = -E\left(\frac{y_i - b'(\theta_i)}{a(\phi)}\right)^2$$

Alors :

$$-\frac{b''(\theta_i)}{a(\phi)} = -\frac{[(y_i - b'(\theta_i)]^2}{a^2(\phi)}$$

on obtient :

$$Var(Y_i) = b''(\theta_i)a(\phi)$$

Considérons maintenant p variables explicatives et n observations dont les observations sont rangées dans la matrice de plan d'expérience X, les paramètres de modèle sont un vecteur $p \times 1$ des coefficients de régression β qui sont des fonctions de θ.

La log-vraisemblance est obtenue par la différentiation de ℓ par rapport à les éléments de β en utilisant la règle des chaînes Yields :

$$\mathcal{L}(B) = \sum_{i=1}^{n} lnf(y_i, \theta_i, \phi) = \sum_{i=1}^{n} \ell(\theta_i, \phi, y_i)$$

Calculons :

$$\frac{\partial \ell_i}{\partial \beta_j} = \frac{\partial \ell_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j}$$

On tenant compte :

- le prédicteur linéaire à n composantes s'écrit : $\eta = X\beta$.
- g la fonction lien est supposée monotone, différentielle tels que :

$$\eta_i = g(\mu_i)$$

Comme :

$$\begin{aligned}
&\frac{\partial \ell_i}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{a(\phi)} = \frac{y_i \mu_i}{a(\phi)} \\
&\frac{\partial \mu_i}{\partial \theta_i} = b''(\theta i) = \frac{Var(y_i)}{a(\phi)} \\
&\frac{\partial \eta_i}{\partial \beta_j} = x_{ij} \quad car\ \eta_i = X_i\beta = \sum_{j=1}^{n} x_{ij}\beta_j \\
&\frac{\partial \mu_i}{\partial \eta_i} \quad d\acute{e}pend\ de\ la\ fonction\ de\ lien\ \eta_i = g(\mu_i)
\end{aligned}$$

Assemblons les choses :

$$\begin{aligned}
\frac{\partial \ell_i}{\partial \beta_j} &= \frac{y_i - b''(\theta_i)}{a(\phi)} \times \frac{1}{Var(Y_i)} \times \frac{\partial \mu_i}{\partial \eta_i} \times x_{ij} \\
&= \frac{W_i(y_i - \mu_i)}{a(\phi)} \frac{\partial \eta_i}{\partial \mu_i} x_{ij}
\end{aligned}$$

tels que :

$$W_i = \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 \frac{1}{Var(Y_i)}$$

c'est la matrice de pondération.

Les équations de vraisemblance sont :

$$\sum_{j=1}^{n} \frac{W_i(y_i - \mu_i)}{a(\phi)} \frac{\partial \eta_i}{\partial \mu_i} x_{ij} = 0$$

Ce sont des équations non linéaires en β, pour la résolution on utilise soit des méthodes itératives dans lesquelles interviennent le Hessien (pour Newton-Raphson) ou la matrice d'information de Fisher, dont les variances et les covariances des paramètres estimées sont obtenues à partir de l'inverse de cette matrice, Ainsi :

$$\begin{pmatrix} Var(\hat{\beta}_0) & Cov(\hat{\beta}_0, \hat{\beta}_1) & \dots & Cov(\hat{\beta}_0, \hat{\beta}_{p-1}) \\ Cov(\hat{\beta}_1, \hat{\beta}_0) & Var(\beta_1) & \dots & Cov(\hat{\beta}_1, \hat{\beta}_{p-1}) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(\hat{\beta}_{p-1}, \hat{\beta}_0) & Cov(\hat{\beta}_1 \hat{\beta}_0) & \dots & Var(\hat{\beta}_{p-1}) \end{pmatrix} = -E \begin{pmatrix} \frac{\partial \ell}{\partial \beta_0^2} & \frac{\partial \ell}{\partial \beta_0} \frac{\partial \ell}{\partial \beta_1} & \cdots & \frac{\partial \ell}{\partial \beta_0} \frac{\partial \ell}{\partial \beta_{p-1}} \\ \frac{\partial \ell}{\partial \beta_1} \frac{\partial \ell}{\partial \beta_0} & \frac{\partial \ell}{\partial \beta_1^2} & \cdots & \frac{\partial \ell}{\partial \beta_1} \frac{\partial \ell}{\partial \beta_{p-1}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \ell}{\partial \beta_{p-1}} \frac{\partial \ell}{\partial \beta_0} & \frac{\partial \ell}{\partial \beta_{p-1}} \frac{\partial \ell}{\partial \beta_1} & \cdots & \frac{\partial \ell}{\partial \beta_{p-1}^2} \end{pmatrix}^{-1}$$

2.2.2 Qualité d'ajustement

Pour évaluer la qualité de modèle sur la base des différences entre observations et estimations, plusieurs critères sont proposées :

Déviance

Définition 2.2.1.

Pour un modèle linéaire générale avec des observations $y = (y_1, y_2, \dots, y_n)$.
On note $\mathcal{L}(\mu, y)$ la fonction de log-vraisemblance exprimé en terme des moyens $\mu = (\mu_1, \dots, \mu_n)$ et $\mathcal{L}(\hat{\mu}, \hat{y})$ le maximum de vraisemblance du modèle, cela se produit pour la plupart des modèles générales ayant un paramètre distinct pour chaque observation et un ajustement parfait $\hat{\mu} = y$, ce modèle et dit modèle saturé.

Le principe de la déviance se base sur la comparaison du modèle estimé avec le modèle saturé.

Soit $\mathcal{L}(y, \phi, y)$ le log-vraisemblance du modèle saturé et $\mathcal{L}(\hat{\mu}, \phi, y)$ le log-vraisemblance du modèle estimé.
l'expression de la déviance s'écrit :

$$D = 2\left(\mathcal{L}(y, \phi, y) - \mathcal{L}(\hat{\mu}, \phi, y)\right)$$

réécrivons cette formule avec une autre façon :

$$D = 2\sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} - 2\sum_{i=1}^{n} \frac{y_i\hat{\theta}_i - b(\hat{\theta}_i)}{a(\phi)}$$

on sait que $a(\phi) = \frac{\phi}{\omega_i}$. il résulte :

$$\begin{aligned} D &= 2\sum_{i=1}^{n} \frac{\omega_i \left[y_i(\theta_i - \hat{\theta}_i) - b(\theta) + b(\hat{\theta}_i)\right]}{\phi} \\ &= \frac{D(y, \hat{\mu})}{\phi} \end{aligned}$$

Cette dernière expression s'appelle *la déviance d'échelle*, la statistique $D(y, \hat{\mu})$ s'appelle *la déviance.*

- pour une distribution normal la déviance c'est exactement la somme des moindres carrées.
- pour certain MLGs, comme la loi binomiale et poisson, la déviance et la déviance d'échelle sont identiques.

Asypmtotiquement, lorsque n s'augmente, D suit une loi de χ^2 à $n - p$ degré de liberté, ce qui permet de l'utiliser comme un test d'adéquation du modèle selon que la déviance est jugée significativement ou non importante.

Remarque 2.2.1. *Comme $\mathcal{L}(y, \phi, y) < \mathcal{L}(\hat{\mu}, \phi, y)$, $D(y, \hat{\mu}) > 0$, par conséquence lorsque la déviance est grande l'ajustement est faible et vice-versa.*

La statistique généralisée de Pearson

L'alternative de la déviance pour comparer les valeurs observées y_i à leur prévisions par le modèle est le test χ^2 de Pearson. La statistique du est est définit par :

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\widehat{Var}(\hat{\mu}_i)}$$

On note :

- $\widehat{Var}(\hat{\mu}_i)$: la fonction de variance estimé.
- $\hat{\mu}_i$: les moyennes estimées.
- y_i : les valeurs observées.

Pour la distribution normal cette statistique est la somme des carrées résiduelles du modèle, dans ce cas la déviance et le χ^2 de Pearson coïncident.

Dans les autres cas, la déviance et χ^2 de Pearson conduisent à des résultats peu différents et dans le cas contraire c'est une mauvaise approximation de la loi asymptotique.

Sachant que l'espérance d'une loi du χ^2 est son degré de liberté en connaissant les aspects approximatifs des tests construits, l'usage est souvent de comparer les statistiques avec les degrés de liberté. le modèle peut être jugé satisfaisant pour un rapport D/ ddl plus petit que 1.

Remarque 2.2.2.
Dans les modèles linéaires généralisées, l'estimation des paramètres avec le maximum de vraisemblance cherche de minimiser la déviance et pour cela, la déviance est préférée plus que χ^2 de Pearson.

Comparaison entre deux modèles :

Pour deux modèles M_0 et M_1, la statistique de Pearson pour comparer entre eux est :

$$\chi^2(M_0/\ M_1) = \sum_{i=1}^{n} \frac{(\hat{\mu}_{0i} - \hat{\mu}_{1i})^2}{Var(\hat{\mu}_{0i})}$$

Critère d'information d'Aikaike

Ce critère a était proposée en 1973 par Aikaike, il est utilisé pour mesurer l'adéquation d'un modèle aux données et donc il nous permet de choisir un modèle parmi plusieurs autres modèles.
l'idée de ce critère est de pénaliser les fonctions de vraisemblance qui nous ramène des mesures comme :

$$D_c = D - \alpha q \phi$$

tels que :

- D :la déviance.
- q : le nombre des paramètres du modèle.
- ϕ : le paramètre de dispersion.
- si ϕ est constant $\alpha \approx 4$ au seuil 5%

le modèle ayant la valeur minimale serait alors le modèle de préférence.

2.3 Tests de validation

Nous avons vu précédemment que les modèles linéaires généralisés sont ajustés aux données par la méthode de maximum de vraisemblance, ce qui permet d'obtenir non seulement des estimations des coefficients des régressions, mais aussi des erreurs de type

asymptotique des coefficients, ce qui rend possible de faire des tests d'hypothèses des paramètres.

Les tests utilisés sont nombreux, dans cette section on va présenter deux tests : *le rapport de vraisemblance* et *le test de Wald*.

2.3.1 Test de Wald

L'estimation du maximum de vraisemblance des paramètres de certains modèles donne des estimations des paramètres et des estimations des standards erreurs des estimateurs. Les estimations des erreurs-types sont souvent des résultats asymptotiques qui sont valides pour des grands échantillons.

On va tester l'hypothèse $\{ H_0 : \beta = \beta_0 \}$ contre $\{ H_1 : \beta \neq \beta_0 \}$ par la statistique de Wald, dénotons l'erreur type asymptotique de l'estimateur $\hat{\beta}$ avec $\hat{\sigma}_{\hat{\beta}}$

$$\mathcal{Z} = \frac{\hat{\beta}}{\hat{\sigma}_{\hat{\beta}}}$$

sous l'hypothèse nulle $\mathcal{Z}$ suit une loi normale.
pour p paramètres la statistique de Wald est :

$$(K'b)'(K'(X'WX)^{-1}K)^{-1}K'b \sim \chi^2$$

avec :

- $(X'WX)^{-1}$: l'inverse de la matrice d'information observé.
- W : la matrice de pondération.
- K : la matrice de contraste définit l'ensemble H_0 des hypothèses à tester sur les paramètres $K'\beta = 0$.
- $b = (X'X)^{-1}X'Y$

2.3.2 Rapport de vraisemblance

soit le modèle 0 avec q_0 variables explicatives et $\mathcal{L}_0$ son log-vraisemblance, le modèle 1 avec q_1 variables explicatives et $\mathcal{L}_1$ son log-vraisemblance ($q_0 < q_1$).
nous testons l'hypothèse nulle H_0, selon laquelle les restrictions du modèle 1 représenté par le modèle 0 sont correctes.

Le rapport de vraisemblance est la différence des écarts résiduelles D_0 et D_1 des deux modèles emboîtés, donc la statistique du test est :

$$\begin{aligned} G_0^2 &= D_0 - D_1 \\ &= 2(\mathcal{L}_s - \mathcal{L}_0) - 2(\mathcal{L}_s - \mathcal{L}_1) \\ &= -2(\mathcal{L}_0 - \mathcal{L}_1) \end{aligned}$$

Sous l'hypothèse nulle, G_0^2 suit approximativement :

- une loi de χ^2 à $q_1 - q_0$ degrés de liberté pour les lois à 1 paramètre (Binomiale, Poisson).
- une loi de Fisher pour les loi à 2 paramètres (Gaussien, Gamma,...etc)

$$F_0 = \frac{\frac{D_0 - D_1}{q_1 - q_2}}{\hat{\phi}} \sim \mathcal{F}(q_1 - q_2)(n - k - 1)$$

$\hat{\phi}$: est le paramètre de dispersion estimé.

n,k : le nombre des paramètres des modèles 1 et 0 respectivement.

2.4 Diagnostics

De nombreux indicateurs sont proposées pour étudier les résidus qui permet de diagnostiquer la régression, pour détecter les régularités (problème de spécification), ou identifier les points isolés (atypiques ou mal modéliser).

2.4.1 Effet levier

soit la matrice de projection définit par :

$$H = W^{\frac{1}{2}}(X(X'WX)^{-1}X')W^{\frac{1}{2}}$$

L'effet de levier de l'observation i sur la valeur ajustée $\hat{\mu}_i$ est la dérivé de $\hat{\mu}_i$ par rapport à y_i, ces dérivées sont les termes diagonaux h_{ii} de cette matrice.
les termes supérieur à $\frac{3p}{n}$ (n : nombre des variables explicatives, p : nombre des paramètres) indiquent des valeurs potentiellement influentes ont besoin d'êtres examinés.

2.4.2 Résidus

Dans les modèles linéaires générales les résidus sont définies par la différence entre les valeurs observées Y et les valeurs prévu $\hat{Y}$ du modèle.

par contre dans les modèles linéaires généralisés, la variance des résidus est souvent liées au taille de $\hat{Y}$, par conséquence il faut utiliser un mécanisme de mise à l'échelle si l'on souhaite utiliser les résidus pour les graphes ou d'autres diagnostics du modèle.

Résidus de Pearson :

Les résidus de Pearson mesure la contribution de chaque observation à la significativité du test déroulant de cette statistique tels que :

$$e_{i.P} = \frac{y_i - \hat{y}_i}{\sqrt{\widehat{Var}(\hat{y}_i)}}$$

Faisant intervenir le terme diagonal de la matrice H, on peut définir les résidus de Pearson standarisés avec :

$$e_{i.Ps} = \frac{y_i - \hat{y}_i}{h_{ii}\sqrt{\widehat{Var}(\hat{y}_i)}}$$

Résidus déviance

Ces résidus mesurent la contribution de chaque observation d_i à la déviance du modèle par rapport au modèle saturé. Ces résidus s'écrit :

$$e_{i.D} = sign(y_i - \hat{y}_i)\sqrt{d_i}$$

avec :

$$D = \sum_{i=1}^{n} d_i \quad et \quad d_i = 2\omega_i[y_i(\theta_i) - \hat{\theta}_i) - b(\theta_i + b(\hat{\theta}_i)]$$

Résidus de Anscombe

Les résidus discutés précédemment ne suivent pas toujours une loi normale, par contre les résidus Anscombe suivent une loi normale, ils sont définis par :

$$e_{i.Anscombe} = \frac{A(y_i) - A(\hat{y}_i)}{\sqrt{\widehat{Var}(A(y_i) - A(\hat{y}_i))}}$$

Mais malheureusement, ce type des résidus n'est pas pratique à cause de la difficulté des calculs et surtout le choix de la fonction A qui dépend au type des données.

Remarque 2.4.1.

- *Le choix des résidus est liés aux types de test.*
- *Les résidus de déviance sont liées au la déviance comme une mesure d'ajustement de modèle et au test de rapport de vraisemblance.*
- *Les résidus de Pearson sont liées avec le test de Wald.*

3 Types des modèles linéaires généralisés

Dans ce chapitre on examine quelques types des modèles linéaires généralisées qui sont nombreux. On s'intéresse tout particulièrement à la modélisation de données quantitatives,qualitatives, binaires,et de données de comptage.

3.1 Modèles linéaires générales

La régression simple, multiple, l'ANOVA et l'ANCOVA constituent le modèle linéaire général dont ils sont tous des applications particulières des modèles linéaires généralisées.

3.1.1 Régression multiple

Le modèle de régression linéaire multiple est l'outil statistique la plus mise en oeuvre pour l'étude de données multidimensionnel d'une variable à expliqué Y quantitative en fonction des variables explicatives X qui sont qualitatives ou bien quantitatives.

On suppose que la variable Y est une variable aléatoire, les p valeurs explicatives $X = (X_1, \ldots, X_p)$ sont non aléatoires.
Dans cette situation, l'écriture du modèle est comme suit :

$$Y_i = \beta_0 + \sum_{i=1}^{p} \beta_j X_{ij} + \xi_i = \beta_0 + X_{i1}\beta_1 + \cdots +_{ip} \beta_p + \xi_i \quad \forall i = 1, \ldots, n$$

avec les hypothèses suivantes :

- $E(\xi_i) = 0$, on obtient donc $E(Y_i/X_i) = \beta_0 + X_{i1}\beta_1 + \cdots +_{ip} \beta_p$, le modèle donc est linéaire.

- $V(Y_i/X_i) = \sigma^2, \quad \forall i = 1, \ldots, n$, comme $V(\xi_i) = \sigma^2$ les variances donc sont égaux.
- la distribution de Y_i sachant X_i est gaussienne de paramètres $(\beta_0 + X_{i1}\beta_1 + \cdots + X_{ip}\beta_p, \sigma^2)$.
- pour tout $i \neq i'$, la variable Y_i/X_i est indépendante de $Y_{i'}/X_{i'}$.

Le modèle s'écrit matriciellement :

$$Y = X\beta + \xi$$

avec :

$$\text{Y}= \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad \text{X}= \begin{pmatrix} 1 & X_{11} & \ldots & X_{1p} \\ 1 & X_{21} & \ldots & X_{2p} \\ 1 & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \ldots & X_{np} \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} \quad \xi = \begin{pmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_n \end{pmatrix}$$

Y : le vecteur aléatoire à expliquer de taille $(n \times 1)$.

X : la matrice explicative de taille $n \times (p + 1)$.

β : le vecteur des paramètres inconnues de taille $(p + 1) \times 1$.

ξ : le vecteur d'erreurs de taille$(n \times 1)$.

L'estimation des coefficients du modèle se fait par l'estimation des MCO qui est dérivé de la minimisation de la somme carrées des résidus $\xi = Y - X\beta$.
c'est à dire :

$$\begin{aligned} S(B) &= Min_\beta \left\{ \sum_{i=1}^{n} (\xi_i)^2 \right\} \\ &= \xi\xi' \\ &= (Y - X\beta)'(Y - X\beta) \\ &= Y'Y - Y'X\beta - (X\beta)'X\beta \\ &= Y'Y - Y'X\beta - \beta'X'Y + \beta'X'X\beta \\ &= Y'Y - 2\beta'X'X\beta \end{aligned}$$

le minimum est attent pour :

$$\frac{S(B)}{\partial\beta|\hat{\beta}} = 0$$

c'est à dire :

$$\frac{\partial}{\partial\beta}[Y'Y - 2\beta'X'X\beta] = 0 \Leftrightarrow -2X'Y + 2X'X\beta = 0$$

dont la solution correspond bien à un minimum car la matrice $2X'X$ dite hessienne est semi-définie positive.

on obtient donc :

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

où X' désigne la matrice transposée de X.

les valeurs estimées de Y ont pour expression :

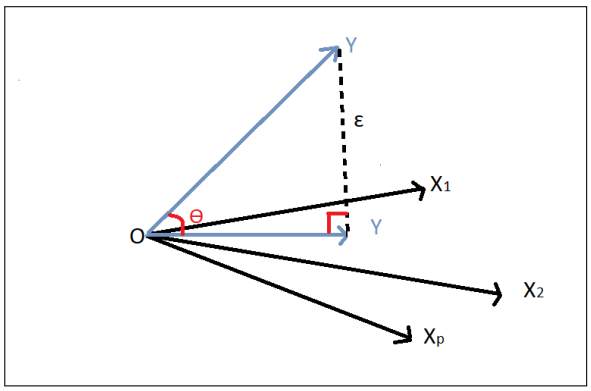
$$\hat{Y} = X\beta = X(X'X)^{-1}X'Y = HY$$

la matrice $H = X(X'X)^{-1}X'$ est appelée "hat matrix", géométriquement, c'est la matrice de projection orthogonale dans $\mathbb{R}^n$ sur le sous espace $vect(X)$ engendré par les vecteurs colonnes de X.

on note :

$$\xi = Y - \hat{Y} = Y - X\beta = (I_p - H)Y$$

Le vecteur des résidus c'est la projection de Y sur le sous-espace orthogonale de $vect(X)$ dans $\mathbb{R}^n$.



sous les hypothèses précédentes on peut montrer facilement que :

- $E(\hat{\beta}) = \beta$, l'estimateur $\hat{\beta}$ est sans biais.
- $V(\hat{\beta}) = \sigma^2(X'X)^{-1}$, donc il est aussi de variance minimale parmi tous les estimateurs linéaires par rapport à Y.

La variance des résidus noté σ^2 est définit par :

$$\sigma^2 = V(\xi_i) = V(Y_i) = E[(Y_i - E(Y_i)^2]$$

elle est estimé par :

$$\begin{aligned}\widehat{\sigma}^2 &= \frac{1}{n-(p+1)}\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2 \\ &= \frac{\sum_{i=1}^{n}(\hat{\xi}_i)^2}{n-p-1} \\ &= \frac{SCR}{n-p-1}\end{aligned}$$

Maintenant on définit R comme le coefficient de corrélation linéaire entre les Y_i et les $\hat{Y}_i$.

R^2 donne la proportion de variabilité de Y qui est expliquée par le modèle tels que :

$$\begin{aligned}R^2 &= \frac{\sum_{i=1}^{n}(\widehat{Y_i}-\overline{Y_n})}{\sum_{i=1}^{n}(Y_i-\overline{Y_n})^2} \\ &= \frac{SCE}{SCT} \\ &= 1-\frac{SCR}{SCT}\end{aligned}$$

avec :

$$\sum_{i=1}^{n}(Y_i-\overline{Y_n})^2 = \sum_{i=1}^{n}(Y_i-\overline{Y_n})^2 + \sum_{i=1}^{n}(\widehat{Y_i}-\overline{Y_n})^2$$

tels que :

SCT :somme des carrées totale.

SCE :somme factorielle.

SCR : somme des carrées résiduelle.

Si $R^2 = 1 \Rightarrow$ l'ajustement est parfait ie $\forall i : \widehat{Y_i} = Y_i$

- géométriquement : R est le cosinus de l'angle formé par$(Y-\overline{Y})$ et $(\widehat{Y}-\overline{Y})$ où $\overline{Y} = (\overline{Y_1},\ldots,\overline{Y_n})^t \in \mathbb{R}^n$.

- en statistique R^2 peut être utilisé pour tester l'ajustement de Y par $\widehat{Y}$.

les hypothèses de test : $\begin{cases} H_0 : \beta_j = 0 & \forall j : 1,\ldots,p \\ H_1 : \beta_j \neq 0 & \exists j \in \{1,\ldots,p\} \end{cases}$

La statistique du test est :

$$F_n = \frac{SCE/P}{SCR/(n-p-1)} = \frac{R^2/p}{(1-R^2)/(n-p-1)}$$

qui est distribué sous H_0 selon une loi de Fisher à p et $n-p-1$ degrés de liberté.

On rejette H_0 si $F_n \geq F_{p,n-p-1,1-\alpha}$

On désire maintenant tester la significativité d'un paramètre β_j avec les hypothèses $H_0 : \beta_j = 0$ contre $H_1 : \beta_j \neq 0$, sous l'hypothèse nulle on a :

$$T_n = \frac{\widehat{B_j}}{\widehat{\sigma_{\beta_j}}}$$

3.1.2 Analyse de la variance à 1 facteur

L'analyse de variance (ANOVA) est une technique statistique utilisée pour étudier l'effet d'un facteur (ou plusieurs facteurs) sur une variable d'intérêt de type quantitatif en utilisant un ensemble de modèles statistiques pour comparer les moyennes des différents échantillons indépendants. Les échantillons correspondent aux différentes modalités de la variable qualitative et les moyennes sont calculées sur la variable quantitative. L'application et la validité de l'analyse de variance repose sur le test de Fisher donc sur trois conditions qui sont :

- l'indépendance des échantillons : c'est à dire l'indépendance entre les différentes valeurs de la variable mesurées y_{ij}.
- la normalité des distributions : la variable quantitative étudiée suit une loi normale ie : $Y \sim \mathcal{N}(\mu, \sigma)$
- l'homogénéité des variances : les populations étudiées ont la même variance car on souhaite étudier l'effet du facteur A sur Y à travers les moyennes.
  le facteur A agit seulement sur les moyennes de Y et non sur les variances.

Sous H_0 Le modèle d'analyse de la variance à 1 facteur tels que : $H_0 : \mu_1 = \mu_2 = \cdots = \mu$ est donné par la forme suivante dite régulière :

$$y_{ij} = \mu + e_{ij}$$

où :e_{ij} : l'erreur aléatoire.

- $e_{ij} \sim \mathcal{N}(0, \sigma^2)$
  $E(e_{ij}) = 0 \qquad$ et $\qquad V(e_{ij}) = \sigma^2$.
  e_{ij} doivent être indépendants.
- $Y \sim \mathcal{N}(\mu, \sigma^2)$, en effet :

$$\begin{aligned} E(y_{ij}) &= E(\mu + e_{ij}) \\ &= E(\mu) + E(e_{ij}) \\ &= \mu \end{aligned}$$

$$\begin{aligned} V(y_{ij}) &= V(\mu + e_{ij}) \\ &= V(\mu) + V(e_{ij}) \\ &= \sigma^2 \end{aligned}$$

Sous H_1 Le modèle d'analyse de la variance à 1 facteur tels que H_1 :{au moins une moyenne est différente des autres} est donné par la forme suivante dite singulière :

$$y_{ij} = \mu + \alpha_i + e_{ij} \qquad \sim \mathcal{N}(0, \sigma^2)$$

où : α_i désigne l'effet de la modalité i sur y_{ij}.

- $e_{ij} \sim \mathcal{N}(0, \sigma^2)$ où les erreurs e_{ij} sont indépendants.
- $y_{ij} \sim \mathcal{N}(\mu + \alpha_i, \sigma^2)$, en effet :

$$\begin{aligned} E(y_{ij}) &= E(\mu + \alpha_i + e_{ij}) \\ &= \mu + \alpha_i + E(e_{ij}) \\ &= \mu + \alpha_i \end{aligned}$$

et :

$$\begin{aligned} V(y_{ij}) &= V(\mu + \alpha_i + e_{ij}) \\ &= V(\mu) + V(\alpha_i) + V(e_{ij}) \\ &= \sigma^2 \end{aligned}$$

ainsi il existe une différence entre les moyennes et il y a un effet du facteur A sur Y.
Le modèle d'analyse de la variance se met sous la forme matricielle suivante :

$$Y = X\theta + e$$

Sous la forme régulière la version matricielle du modèle est donné par :

$$\begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1n_1} \\ \vdots \\ Y_{i1} \\ \vdots \\ Y_{in_i} \end{pmatrix} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ \vdots & 0 & \dots & 0 \\ 1 & \vdots & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & 1 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_i \end{pmatrix} + \begin{pmatrix} e_{11} \\ \vdots \\ e_{1n_1} \\ \vdots \\ e_{i1} \\ \vdots \\ e_{in_i} \end{pmatrix}$$

Dans l'écriture singulière de ce modèle, la matrice X ainsi que le paramètre de moyenne θ changent :

$$\begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1n_1} \\ \vdots \\ Y_{i1} \\ \vdots \\ Y_{in_i} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & 0 & \dots & 0 \\ 1 & 1 & \vdots & \dots & 0 \\ \vdots & \vdots & \dots & \vdots & \\ 1 & 0 & \dots & 1 & \vdots \\ \vdots & \vdots & \dots & \vdots & \\ 1 & 0 & 0 & \dots & 1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \alpha_1 \\ \vdots \\ \alpha_i \end{pmatrix} + \begin{pmatrix} e_{11} \\ \vdots \\ e_{1n_1} \\ \vdots \\ e_{i1} \\ \vdots \\ e_{in_i} \end{pmatrix}$$

Les paramètres estimés du modèle sous H_0 sont :

- $\widehat{\mu} = \overline{y} = \overline{y_{..}}$ (moyenne globale).
  avec $\overline{y} = \frac{1}{N}\sum_{i=1}^{p}\sum_{j=1}^{n_i} y_{ij}$ et $N = \sum_{i=1}^{p} n_i$
- $\widehat{e_{ij}} = y_{ij} - \widehat{\mu} = y_{ij} - \overline{y_{..}}$

sous H_1 :

- $\widehat{\alpha_i} = \overline{y_i} - \overline{y_{..}}$.
- $e_{ij} = y_{ij} - \overline{y_{..}} - (\overline{y_i} - \overline{y_{..}}) = y_{ij} - \overline{y_i}$

On cherche maintenant à décomposer la variance totale car on a 2 sources de variations : variance intra-groupes et une variance inter-groupes.
Nous avons la forme singulière du modèle :

$$\begin{aligned} y_{ij} &= \mu + \alpha_i + e_{ij} \\ &= \overline{y_{..}} + (\overline{y_i} - \overline{y_{..}}) + (y_{ij} - \overline{y_i}) \end{aligned}$$

on obtient donc :

$$(y_{ij} - \overline{y_{..}}) = (\overline{y_i} - \overline{y_{..}}) + (y_{ij} - \overline{y_i})$$

$$(y_{ij} - \overline{y_{..}})^2 = (\overline{y_i} - \overline{y_{..}})^2 + (y_{ij} - \overline{y_i})^2 + 2(\overline{y_i} - \overline{y_{..}})(y_{ij} - \overline{y_i})$$

$$\sum_{i=1}^{p}\sum_{j=1}^{n_i}(y_{ij} - \overline{y_{..}})^2 = \sum_{i=1}^{p}\sum_{j=1}^{n_i}(\overline{y_i} - \overline{y_{..}})^2 + \sum_{i=1}^{p}\sum_{j=1}^{n_i}(y_{ij} - \overline{y_i})^2 + 2\sum_{i=1}^{p}\sum_{j=1}^{n_i}(\overline{y_i} - \overline{y_{..}})(y_{ij} - \overline{y_i})$$

pour une seul modalité :

$$\begin{aligned} 2n_i \sum_{j=1}^{n_i} (\overline{y}_i - \overline{y}_{..})(\overline{y}_{ij} - \overline{y}_i) &= 2n_i(\overline{y}_i - \overline{y}_{..}) \sum_{j=1}^{n_i} (\overline{y}_{ij} - \overline{y}_i) \\ &= 2n_i(\overline{y}_i - \overline{y}_{..}) \left[\sum_{j=1}^{n_i} \overline{y}_{ij} - n_i \overline{y}_i \right] \\ &= 2n_i(\overline{y}_i - \overline{y}_{..}) \left[n_i \frac{1}{n_i} \sum_{j=1}^{n_i} \overline{y}_{ij} - n_i \overline{y}_i \right] = 0 \end{aligned}$$

donc l'équation fondamentale est :

$$\sum_{i=1}^{p} \sum_{j=1}^{n_i} (y_{ij} - \overline{y_{..}})^2 = \sum_{i=1}^{p} n_i (\overline{y_i} - \overline{y_{..}})^2 + \sum_{i=1}^{p} \sum_{j=1}^{n_i} (y_{ij} - \overline{y_i})^2$$

$$SCT = SCE + SCR$$

L'estimations des variances associés un carré moyen sont :

- Variance totale : $\frac{SCT}{N-1}$
- Variance factorielle : $\frac{SCE}{p-1}$
- Variance résiduelle : $\frac{SCR}{N-p}$

Pour prendre la décision H_0 où H_1, on applique le test de Fisher.

$$F_{observée} = \frac{\frac{SCE}{p-1}}{\frac{SCR}{N-p}} \rightsquigarrow F_{théorique}(p-1, N-p)$$

si $F_{observée} \leq F_{théorique}$: on accepte H_0 au seuil α.

si $F_{observée} > F_{théorique}$: on rejette H_0 au seuil α.

Source de variation	ddl	S.C.	Carré moyenne
Factorielle	p-1	SCE	SCE/p-1
Résiduelle	N-p	SCR	SCR/N-p
Totale	N-1	SCT	SCT/N-1

TABLE 3.1 – Tableau d'ANOVA

3.2 Régression de Poisson

La régression de Poisson est un modèle de prédiction qui s'applique lorsque la variable à expliqué Y est une variable de comptage, par exemple le nombre des ciga-

rettes consommé par jour...etc, ces variables peut s'écrit sous forme des tables de contingence où des tables des fréquences.

L'objectif de cette régression est d'expliquer les variables de comptage à partir des variables explicatifs.
La distribution de la loi de Poisson est :

$$f(y,\lambda) = \frac{\lambda^y e^{-\lambda}}{y!}$$

cette distribution peut s'écrit sous la forme exponentielle.

$$f(y,\lambda) = exp\ [y\ log(\lambda) - \lambda - log(y!)]$$

en comparant avec l'expression (2.2) :

$$\theta = log(\lambda) \qquad \lambda = exp(\theta)$$

où :

$$f(y,\lambda) = exp(y\theta - exp(\theta) - log(y!))$$

avec :

$$\theta = log(\lambda), b(\theta) = exp(\theta), c(y,\phi) = log(y!), a(\phi) = 1$$

Le modèle de Poisson modélise la loi des variables à expliqués Y_i par une loi de Poisson de paramètre $\lambda_i = \lambda(x_i)$ telle que :

$$\lambda_i = X_i\beta$$

mais comme la moyenne d'une loi de poisson est forcément positive on pose alors :

$$log(\lambda_i) = X_i\beta$$

Dans ce cas la fonction de lien c'est la fonction **log** et le modèle de régression de Poisson est :

$$log(\lambda_i) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \quad x = (x_1, \ldots, x_p)$$

on peut écrire aussi :

$$\lambda_i = exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)$$

$\beta_0, \ldots, \beta_p$ sont des coefficients réels inconnus associées à les variables explicatives $x_1, \ldots, x_p$ en utilisant la méthode de maximum de vraisemblance on obtient alors

$\widehat{\beta_1}, \ldots, \widehat{\beta_p}$.

comme $Y_1, \ldots, Y_p$ sont indépendant, la fonction de vraisemblance est :

$$\ell(\beta) = \prod_{i=1}^{n} \frac{\lambda^{Y_i} e^{-\lambda_i}}{Y_i!}$$

λ_i est définit en terme de $\beta_0, \ldots, \beta_p$ et les covariables $x_{i1}, \ldots, x_{ip}$.

fixons $x_{i0} = 1$ pour tout i, le log de vraisembance est :

$$\begin{aligned}
\mathcal{L}(\beta) &= ln \prod_{i=1}^{n} \frac{\lambda^{Y_i} e^{-\lambda_i}}{Y_i!} \\
&= \sum_{i=1}^{p} Y_i ln(\lambda_i) - \lambda i - ln Y_i! \\
&= \sum_{i=1}^{p} Y_i \sum_{j=0}^{p} x_{ij} - e^{\sum_{j=0}^{p} \beta_j x_{ij}} - ln Y_i!
\end{aligned}$$

le gradient au point β est définit par :

$$\nabla \mathcal{L}(\beta) = \left(\frac{\partial \mathcal{L}}{\partial \beta_0}(\beta), \ldots, \frac{\partial \mathcal{L}}{\partial \beta_p}(\beta) \right)'$$

tels que la j-ème composante de ce vecteur est :

$$\frac{\partial \mathcal{L}}{\partial \beta_j}(\beta) = \sum_{i=1}^{n} x_{ij}(Y_i - e; \sum_{j=0}^{p} \beta_j x_{ij})$$

on obtient alors l'écriture matricielle suivante :

$$\nabla \mathcal{L}(\beta) = \sum_{i=1}^{n} x_{ij}(y_i - p(x_i)) = X'(Y - p)$$

Pour évaluer **la qualité d'ajustement** du modèle avec le critère de **la déviance** on note :

$\mathcal{L}_s$ valeur max de la log-vraisemblance modèle complet où bien dit saturé.

$\mathcal{L}_e$ valeur max de la log-vraisemblance modèle estimé.

la statistique déviance est :

$$\begin{aligned}
D &= \mathcal{L}_s - \mathcal{L}_e \\
&= 2 \left(\sum_{i=1}^{n} y_i X_i \beta - exp(X_i \beta) - ln(y_i) - \sum_{i=1}^{n} y_i X_i \hat{\beta} - exp(X_i \hat{\beta}) - ln(y_i) \right) \\
&= 2 \sum_{i=1}^{n} y_i (X_i \beta - X_i \hat{\beta}) - \left(exp(X_i \beta) - exp(X_i \hat{\beta}) \right) \\
&= 2 \sum_{i=1}^{n} y_i \left(ln(\lambda_i) - ln(\hat{\lambda}_i) \right) - (\lambda_i - \hat{\lambda}_i) \\
&= 2 \sum_{i=1}^{n} y_i \; ln \left(\frac{\lambda_i}{\hat{\lambda}_i} \right) + (\hat{\lambda}_i - \lambda_i)
\end{aligned}$$

qui suit une loi du χ^2 à $(n-p-1)$ degrés de liberté.
on peut encore utiliser **la statistique de Pearson** qui s'écrit :

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\lambda}_i)^2}{\hat{\lambda}_i} \sim \chi^2(n-p-1)$$

Dans la régression de poisson **les résidus de Pearson** sont des écarts normalisés par l'écart-type.
comme :
$E(Y) = V(Y) = \lambda$

$$e_{i.P} = \frac{y_i - \hat{y}_i}{\sqrt{\hat{\lambda}_i}}$$

et **les résidus déviance** qui sont des composantes (pour l'individu n°i) de la statistique déviance utilisée pour évaluer la modélisation, dans ce cas **les résidus déviance** sont :

$$e_{i.D} = sign(r_i)\sqrt{2y_i\ ln\left(\frac{\lambda_i}{\hat{\lambda}}\right) - r_i}$$

en effet
nous avons :

$$e_{i.D} = sign(y_i - \hat{y}_i)\sqrt{d_i} \quad avec : d_i = 2\omega_i[y_i(\theta_i) - \hat{\theta}_i) - b(\theta_i + b(\hat{\theta}_i)]$$

donc :

$$\begin{aligned} e_{i.D} &= sign(y_i - \hat{y}_i)\sqrt{2y_i(X_i\beta - X_i\hat{\beta}_i) - exp(X_i\beta) + exp(X_i\hat{\beta})} \\ &= sign(y_i - \hat{y}_i)\sqrt{2y_i(ln\lambda_i - ln\hat{\lambda}_i) - \lambda_i + \hat{\lambda}_i} \\ &= e_{i.D} = sign(r_i)\sqrt{2y_i\ ln\left(\frac{\lambda_i}{\hat{\lambda}}\right) - r_i} \end{aligned}$$

3.3 Régression Logistique

La régression logistique constitue un autre cas particulier de modèle linéaire généralisé, cette régression s'applique pour expliquer des variables dépendantes qualitative dichotomique c'est à dire de type binaire par exemple (succès/échec), (présence/absence)...etc à partir des variables explicatives.

Soit P une population considérée comme étant divisée en deux groupes G_1 et G_2 que l'on peut distinguer par des variables $X_1, X_2, \ldots, X_p$.
Soit Y la variable qualitative valant :

$$\begin{cases} 1 \textit{ si l'individu appartient à } \mathrm{G}_1 \\ 0 \textit{ sinon} \end{cases}$$

Le truc de la régression logistique consiste à modéliser la probabilité que celle-ci Y se réalisé.

Les données dont on dispose sont n observations de $(Y, X_1, \ldots, X_p)$ notées $(y_1, x_{1.1}, \ldots, x_{p.1}) \ldots (y_n, x_{1.n}, \ldots, x_{p.n})$. où pour tout $(i, j) \in \{1, \ldots, n\} \times \{1, \ldots, p\}$, $x_{i.j}$ est l'observation de la variable X_j sur le i-ème individu et y_i indique le groupe dans lequel il appartient : $y \in \{0, 1\}$.

Pour tout $i \in \{1, \ldots, n\}$:
$(x_{1.i}, \ldots, x_{p.i})$ est une réalisation du vecteur aléatoire réel $(X_1, \ldots, X_p)$ sachant que $(X_1, \ldots, X_p) = (x_{1.i}, \ldots, x_{p.i}) = x_i$ et y_i est une réalisation de $Y_i \sim \mathcal{B}(p(x_i))$
Avant de décrire les hypothèses introduites dans la régression logistique, reconsidérons la probabilité conditionnelle $P(Y = y_k|X)$.

$$\begin{aligned} P(Y = y_k|X) &= \frac{P(Y = y_k) \times P(X|Y = y_k)}{P(X)} \\ &= \frac{P(Y = y_k) \times P(X|Y = y_k)}{\sum_k P(Y = y_k) \times P(X|Y = y_k)} \end{aligned}$$

Dans le cas à deux classes, nous devons comparer simplement $P(Y = 1|X)$ et $P(Y = 0|X)$. on formant le rapport :

$$\frac{P(Y = 1|X)}{P(Y = 0|X)} = \frac{P(Y = 1)}{P(Y = 0)} \times \frac{P(X|Y = 1)}{P(X|Y = 0)}$$

Donc l'équation de la régression logistique est :

$$logit(p) = ln\left(\frac{P(Y = 1|X)}{P(Y = 0|X)}\right) = ln\left(\frac{P(x)}{1 - P(x)}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p = X\beta$$

avec :

- *logit* : est la fonction bijective dérivable de $]0, 1[$ dans $\mathbb{R}$ qui à $p(x) \to \frac{p(x)}{1-p(x)}$.
- $p(x) = p[Y = 1|\{X_1, \ldots, X_p\} = x_i]$ qui est aussi la valeur moyenne de Y quand $X_1, \ldots, X_p = x$.

Ainsi, p et $\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$ sont liés par la transformation logit, on parle de **lien logit**.

l'expression de p est :

$$p(x) = logit^{-1}(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p) = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}{1 + e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}$$

Pour mesurer l'association entre deux variables binaires on utilise l'odds ratio OR, qui est le rapport des côtes de deux valeurs x_0 et x_1 de $X = X_1, \ldots, X_p$ le réel :

$$OR(x_0, x_1) = \frac{\frac{p(x_0)}{1-p(x_0)}}{\frac{p(x_1)}{1-p(x_1)}} = \frac{p(x_0)(1-p(x_1))}{(1-p(x_0))p(x_1)} = e^{\beta}$$

Le OR est interprété comme suit :

- si $OR > 1$, l'augmentation d'une unité X_j entraîne une augmentation des chances que $\{Y = 1\}$ se réalise.
- si $OR = 1$, l'augmentation d'une unité X_jn'a pas d'impact sur Y.
- si $OR < 1$, l'augmentation d'une unité X_j entraîne une augmentation des chances que $\{Y = 0\}$ se réalise.

L'estimation des coefficients $\beta_0, \ldots, \beta_p$ se fait par le maximum de vraisemblance. la fonction vraisemblance associé à $(Y_1, \ldots, Y_n)$ tels que $Y_i \sim \mathcal{B}(p(x_i))$ est :

$$\ell(\beta) = \prod_{i=1}^{n} p(x_i)^{y_i} + (1 - p(x_i))^{(1-y_i)}$$

la log vraisemblance s'écrit donc :

$$\begin{aligned}
\mathcal{L}(\beta) &= ln(\prod_{i=1}^{n} p(x_i)^{y_i} + (1 - p(x_i))^{(1-y_i)}) \\
&= \sum_{i=1}^{n} y_i ln(p(x_i) + (1 - y_i)) ln(1 - p(x_i)) \\
&= \sum_{i=1}^{n} y_i ln(p(x_i)) + ln(1 - p(x_i)) - y_i ln(1 - p(x_i)) \\
&= \sum_{i=1}^{n} y_i ln\left(\frac{p(x_i)}{1 - p(x_i)}\right) + ln(1 - p(x_i)) \\
&= \sum_{i=1}^{n} y_i (logit(p(x)) + ln\left(1 - \frac{e^{x_i'\beta}}{1 - e^{x_i'\beta}}\right) \\
&= \sum_{i=1}^{n} y_i x_i'\beta + ln\left(\frac{1}{1 + e^{x_i'\beta}}\right) \\
&= \sum_{i=1}^{n} y_i x_i'\beta - ln(1 + e^{x_i'\beta})
\end{aligned}$$

le gradient au point β est définit par :

$$\nabla\mathcal{L}(\beta) = \left(\frac{\partial\mathcal{L}}{\partial\beta_0}(\beta), \ldots, \frac{\partial\mathcal{L}}{\partial\beta_p}(\beta)\right)'$$

tels que la j-ème composante de ce vecteur est :

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \beta_j}(\beta) &= \sum_{i=1}^{n} y_i x_{ij} - \frac{x_{ij} e^{x_i'\beta}}{1 + e^{x_i'\beta}} \\ &= \sum_{i=1}^{n} x_{ij}(y_i - p(x_i))\end{aligned}$$

on obtient alors l'écriture matricielle suivante :

$$\nabla \mathcal{L}(\beta) = \sum_{i=1}^{n} x_{ij}(y_i - p(x_i)) = X'(Y - p)$$

Maximiser la log à vraisemblance revient à résoudre le système d'équations suivant :

$$\begin{cases} \frac{\partial \mathcal{L}(\beta)}{\partial \beta_0} = \sum_{i=1}^{n}(y_i - p(x_i)) = 0 \\ \frac{\partial \mathcal{L}(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} x_{ij}(y_i - p(x_i)) = 0 \quad j = 1, \ldots, p \end{cases}$$

Donc :

$$\begin{cases} x_{11}y_1 + \cdots + x_{n1}y_n = x_{11}\frac{e^{\beta_1 x_{11} + \cdots + \beta_p x_{1p}}}{1+e^{\beta_1 x_{11} + \cdots + \beta_p x_{1p}}} + \cdots + x_{n1}\frac{e^{\beta_1 x_{n1} + \cdots + \beta_p x_{np}}}{1+e^{\beta_1 x_{n1} + \cdots + \beta_p x_{np}}} \\ \vdots \\ x_{1p}y_1 + \cdots + x_{np}y_n = x_{1p}\frac{e^{\beta_1 x_{11} + \cdots + \beta_p x_{11}}}{1+e^{\beta_1 x_{11} + \cdots + \beta_p x_{1p}}} + \cdots + x_{np}\frac{e^{\beta_1 x_{n1} + \cdots + \beta_p x_{np}}}{1+e^{\beta_1 x_{n1} + \cdots + \beta_p x_{np}}} \end{cases}$$

La résolution de ces équations se base sur les méthodes itérative en utilisant la méthode du score de Fisher.

la matrice d'information de Fisher est donnée par :

$$I_n(\beta) = E[-d^2\mathcal{L}(\beta)] = -d^2\mathcal{L}(\beta)$$

où :

$$\begin{aligned}-d^2\mathcal{L}(\beta) &= \sum_{i=1}^{n} p(x_i)(1 - p(x_i)x_i x_i^{'} \\ &= X'WX\end{aligned}$$

avec :

$W = diag(W_1, \ldots, W_n), W_i = p(x_i)(1 - p(x_i)$

les $\beta^{(t+1)}$ sont exprimés avec $\beta^{(t)}$ avec l'expression suivante :

$$\begin{aligned}\beta^{(t+1)} &= \beta^{(t)} + s[-d^2\mathcal{L}(\beta^{(t)})]^{-1}\nabla\mathcal{L}(\beta^{(t)}) \\ &= \beta^{(t)} + s(X^{(t)}W^{(t)}X)^{-1}X^{(t)}(Y - p^{(t)} \\ &= (X^{(t)}W^{(t)}X)^{-1}X^{(t)}W^{(t}[X\beta^{(t)} + s(W^{(t)})(Y - p^{(t)})] \\ &= (X^{(t)}W^{(t)}X)^{-1}X^{(t)}W^{(t)}Z^{(t)}\end{aligned}$$

et :

$$Z^{(t)} = X\beta^{(t)} + s(W^{(t)})(Y - p^{(t)})$$

Notons maintenant $\mathcal{L}_\int$ valeur max de la log-vraisemblance modèle complet où bien dit saturé et $\mathcal{L}_\urcorner$ valeur max de la log-vraisemblance modèle estimé, **la déviance** notée D est :

$$\begin{aligned} D =& \mathcal{L}_s - \mathcal{L}_e \\ &= 2\sum_{i=1}^{n}\left(Y_i ln\left(\frac{Y_i}{\hat{p}(x_i)}\right) + (1-Y_i) ln\left(\frac{1-Y_i}{1-\hat{p}(x_i)}\right)\right) \end{aligned}$$

avec : $x_i = (1, x_{1.i}, \ldots, x_{p.i})$.
Si le modèle est bien adapté au problème, la déviance suit une loi de χ^2 à $n-(p+1)$ DDL.

Pour tout $j \in \{0, \ldots, p\}$ on veut tester l'hypothèse $\{H_0 : \beta_j = 0\}$ contre $\{H_0 : \beta_j \neq 0\}$, deux tests sont proposées : **le test de Wald** et le test de rapport de vraisemblance.
pour le test de Wald, la statistique sous H_0 est :

$$\mathcal{Z} = \frac{\hat{\beta}}{\hat{\sigma}_{\hat{\beta}}} \rightsquigarrow \chi^2(1)$$

où l 'hypothèse nulle est rejetée si : $\mathcal{Z} > \chi^2(1)$.
et la statistique de **test de rapport de vraisemblance** est :

$$G_0^2 = -2\left(\mathcal{L}_0(\hat{\beta}_0, \hat{\beta}_j = 0) - \mathcal{L}_1(\hat{\beta}_0, \hat{\beta}_j \neq 0)\right) \rightsquigarrow \chi^2_\alpha(1)$$

l 'hypothèse nulle est aussi rejetée si : $G_0^2 > \chi^2_\alpha(1)$.

Pour tout $i \in \{0, \ldots, n\}$, on appelle
• La i-ème **résidus de Pearson** la réalisation de :

$$e_{i.P} = \frac{Y_i - \widehat{p}(x_i)}{\sqrt{\widehat{p}(x_i)(1-\widehat{p}(x_i))}}$$

• La i-ème **déviance résiduelle** est la réalisation de :

$$e_{i.D} = sign(Y_i - \widehat{p}(x_i)))\sqrt{2\left(Y_i ln\left(\frac{Y_i}{\hat{p}(x_i)}\right) + (1-Y_i) ln\left(\frac{1-Y_i}{1-\hat{p}(x_i)}\right)\right)}$$

4 Applications

4.1 Régression Logistique

4.1.1 Introduction

Le diabète est considéré aujourd'hui comme le mal du siècle. Cette maladie métabolique chronique, liée aux changements de mode de vie et d'habitudes alimentaires de ces 30 dernières années, voit en effet son incidence croître de manière exponentielle et touche désormais plus de 350 millions de personnes à travers le monde.

L'objectif principal de cette étude est d'expliquer l'effet de plusieurs facteurs (L'âge, le sexe, le poids, le cholestérol,HDL, LDL, Tri-glycérides, créatinine, albumine, les antécédents familiaux et HbA1c) sur la présence ou l'absence du diabète chez les individus.

Les analyses statistiques sont effectuées à l'aide du logiciel statistique R.

4.1.2 Base de données

Les données utilisées dans cette étude regroupent les caractéristiques cliniques et biochimiques des patients diabétiques et non diabétiques.
Pour les patients diabétiques, la direction du polyclinique 40 hectares-Jijel a mis à notre disposition les informations de 47 patients concernant le phénomène étudié, enregistrées en 2022.
les caractéristique des personnes non diabétiques sont obtenues par l'examen des bilans générals de 50 personnes.

Les différentes mesures effectuées sont les suivantes :

Nom des variables	**Description des variables**	**L'unité de mesure**
Sexe	Identité féminine ou masculine	(homme\femme)
Age	L'age des individus statistiques	années
Poids	Le poids	Kilogramme (kg)
Antfam	Les antécédents familiaux du diabète	(oui\non)
Chol	Le taux du cholestérol total dans le sang	gramme par litre (g\l)
HDL	Le taux cholestérol HDL (bon cholestérol)	gramme par litre (g\l)
LDL	Le taux de cholestérol LDL (mauvais cholestérol)	gramme par litre (g\l)
Trg	Le taux des triglycérides	gramme par litre (g\l)
Creat	Le taux sanguin du créatininémé	milligramme par litre (mg\l)
Glec	Glycémie à jeun qui est le taux du glucose dans le sang	gramme par litre (g\l)
Album	Le taux d'albumine	milligramme par litre (mg\l)
HbA1c	Le pourcentage d'hymoglobine glyquée	puorcentage
Diabet	La maladie du diabète	(oui\non)

TABLE 4.1 – les facteurs potentiellement explicatifs du diabète

4.1.3 Statistique descriptive des différentes variables

Le tableau 4.2 illustre la moyenne, médiane, minimum, maximum des variables quantitatives représentées dans le tableau précédent.

Les tableaux , 4.3, 4.4 et 4.5 représentent les fréquence des variables qualitatives citées dans la même table (4.1).

Le paramètre	Min	Mean	Median	Variance	Max	Statistique d'asymétrie	Statistique d'aplatissement
L'age	22.00	55.97	59.00	214.343	84.00	-0.536	-0.265
Poids	41.00	79.21	79.00	271.728	132.00	0.638	1.134
Chol	0.475	1.733	1.750	0.267	4.960	2.399	14.965
HDL	0.2800	0.4872	0.4700	0.30	1.6600	3.416	21.191
LDL	0.2300	0.9813	0.9400	0.211	4.1900	3.532	23.826
Trg	0.320	1.393	1.160	1.000	6.300	2.740	10.137
Creat	1.80	9.21	9.00	6.924	19.00	0.826	2.379
Glec	0.650	1.186	1.020	0.161	2.890	1.673	3.160
Album	5.00	43.28	45.10	948.962	200.00	2.118	9.265
HbA1c	0.0500	0.2014	0.0700	0.668	5.8000	6.791	45.332

TABLE 4.2 – Résumé statistique des données quantitative de tableau

D'après les deux statistique d'asymétrie et d'aplatissement, on peut constater que nos variables quantitative ne sont pas normalement distribuées (Skewness$\neq$ 0) et (Kurtosis$\neq$ 3)

Sexe	Le nombre	Le pourcentage
Femmmes	53	54.6%
Hommes	44	45.4%

TABLE 4.3 – Table de fréquences des modalités du sexe

Antfam	Le nombre	Le pourcentage
NON	48	49.5%
OUI	49	50.5%

TABLE 4.4 – Table de fréquences des modalités des antécédents familiaux

Diabet	Le nombre	Le pourcentage
NON	50	51.5%
OUI	47	48.5%

TABLE 4.5 – Table de fréquences des modalités du diabète

4.1.4 Présentation du modèle de régression logistique

On s'intéresse à expliquer et à prédire la probabilité d'être touché par le diabète diabète, notée P(Y = 1|X), en fonction des variables explicatives illustrées précédemment.

On considère le modèle de régression logistique multiple suivant :

$$Prob(Y_i = 1|X_i) = \frac{e^{x_i'\beta}}{1 + e^{x_i'\beta}} \qquad \forall i = 1, \ldots, 97$$

En appliquant la méthode Pas à Pas Ascendante, pour tout $i = 1, \ldots, 97$, le modèle s'écrit donc :

$$logp(Y_i = 1|X_i) = -1.18956 + 0.12310 Age_i - 4.56727 Chol_i + 8.86782 Glec_i - 0.10067 Poids_i$$

Les estimateurs β_j, $j = 1 \ldots 5$ des paramètres du modèle sont représentés dans le tableau 3.6 et les programmes correspondant se trouvent en annexe.

	Estimate	Sd.error	Z value	Pr(>|Z|)	
Intercept	-1.18956	3.83986	-0.310	0.756718	
Age	0.12310	0.03620	3.400	0.000673	***
Chol	-4.56727	1.46779	-3.112	0.001860	**
Glec	8.86782	2.65245	3.343	0.000828	***
Poids	-0.10067	0.03514	-2.865	0.000828	**

TABLE 4.6 – Estimations des paramètres du modèle

Ce tableau montre l'importance de chaque variable dans le modèle, toutes les variables sont significativement différentes de 0 (P-value< 5%), alors ces variables :l'âge, la glycémie sanguin, le cholestérol, et le poids affectent la chance d'être diabétique.

	OR	2.5%	95.5%	P	
Intercept	3.0435e-01	1.4036e-04	6.7377e+02	0.7567181	
Age	1.1310e+00	1.0645e+00	1.2313e+00	0.0006734	***
Poids	9.0423e-01	8.3341e-01	9.5960e-01	0.0041721	**
Chol	1.0386e-02	3.5407e-04	1.2800e-01	0.0018604	**
Glec	7.0998e+03	9.9393e+01	4.2660e+06	0.0008280	***

TABLE 4.7 – Odds ratio et Intervalles de confiance

• La variable Age a un OR=1.1310 cela veut dire que la probabilité d'être diabétique augmente de 13% lorsque l'age de la personne augmente d'une année.
autrement dit, le log népérien de la probabilité d'être diabétique augmente de 0.12310 (log OR) lorsque l'age de la personne augmente d'une année.

• La variable Poids a un OR=0.90423 cela veut dire que la probabilité d'être diabétique diminue de 10% (1-OR) lorsque le poids de la personne augmente d'un kilogramme.
Autrement dit, le log népérien de la probabilité d'être diabétique diminue de 0.10067 (log OR) pour chaque kilogramme supplémentaire du poids de la personne.

• La variable Chol a un OR=0.0010386 cela veut dire que la probabilité d'être diabétique diminue de 99.8% (1-OR) lorsque le cholestérol de la personne augmente d'un gramme par litre.
Autrement dit, le log népérien de la probabilité d'être diabétique diminue de 0.4.56727 (log OR) pour chaque g/l supplémentaire du cholestérol de la personne.

• La variable Glec a un OR=7.0998e+03 cela veut dire que le risque d'être diabétique multiplié par 7000 fois lorsque la glycémie de la personne varié d'une unité.
Autrement dit, le log népérien de la probabilité d'être diabétique augmente de 8.86782 (log OR) pour chaque g/l supplémentaire du glycémie sanguin de la personne.

Remarque :

Il est connu que le poids augmente le risque du diabète. Dans notre étude on a prit une population qui est déjà diagnostiqué au diabète, donc sous traitement et qui suit un régime strict, ce qui explique qu'on a pas de corrélation positive entre le poids et le diabète
(même interprétation pour le cholestérol).

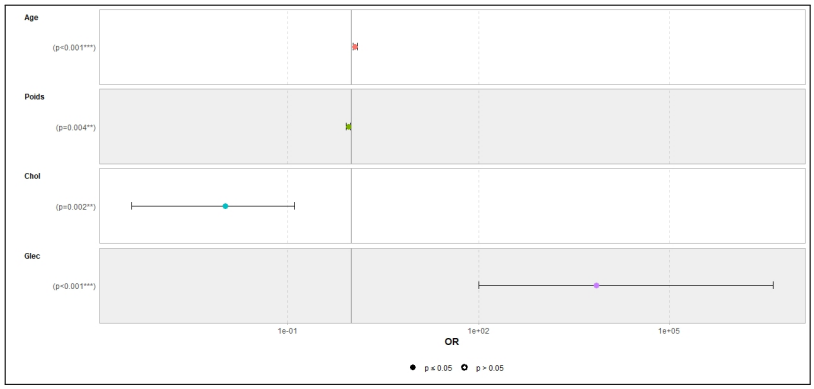


FIGURE 4.1 – Le rapport des cotes.

Variable	N	Odds ratio	p
Age	97	1.13 (1.06, 1.23)	<0.001
Poids	97	0.90 (0.83, 0.96)	0.004
Chol	97	0.01 (0.00, 0.13)	0.002
Glec	97	7099.78 (99.39, 4265991.79)	<0.001

FIGURE 4.2 – Valeurs de OR des variables significatives.

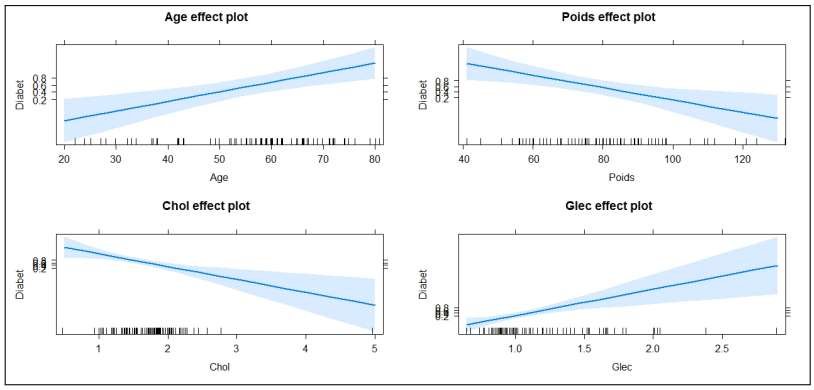


FIGURE 4.3 – L'évolution de diabète avec les variables significatives.

Les figures 1, 2 et 3 confirment l'interprétation des résultats discutées ci dessus.

4.1.5 Prévision

La matrice de confusion, est un résumé des résultats de prédiction. elle compare les données réels pour une variable cible (diabète) à celles prédites par notre modèle estimé.

	Non diabétique	Diabétique	Correct ourcentage
Non diabétique	47	3	94.0%
Diabétique	5	42	89.4%
Pourcentage global			91.8%

TABLE 4.8 – Matrice de confusion

• Sur 50 personnes non diabétiques, notre modèle a bien classé 47 personnes avec un taux de succès estimé de 94%.

• Sur 47 personnes diabétiques, notre modèle a bien classé 42 personnes avec un taux de succès estimé de 89.4%.

• En général, notre modèle a un taux très important de bon classement estimé de 91.8%, ce qui reflète le pouvoir prédictif du modèle.

4.2 Régression de poisson

4.2.1 Introduction

Les prédictions de matchs de football sont d'un grand intérêt pour les fans et la presse sportive. Au cours des dernières années, il a été le centre de plusieurs études.
On se propose de modéliser les matchs de football avec un model de régression de Poisson,Nous avons appliqué la méthodologie proposée sur la compétition de ligue algérienne professionnels de football, L'ensemble des équipes sera paramétré par une variable "Ncarton", indique le nombre des cartons jaunes et rouges obtenues durant le match, et la variable "Domicile" qui représente le lieux du match..

4.2.2 Statistique descriptive des données

Les tableau 3.9, 3.10 illustre les statistiques descriptives des données.

Jouer au domicile au non	Le nombre	Pourcentage
ND	36	50%
D	36	50%
Total	72	100%

TABLE 4.9 – Résumé statistique de la variable qualitative

Variable	Min	Mean	Median	Max	Variance
Nombre des buts marqué	0	1.069	1.00	7	2.009194
Nombre des cartes jaunes et rouges	0	1.458	1.00	7	2.617958

TABLE 4.10 – Résumé statistique des variables quantitatives

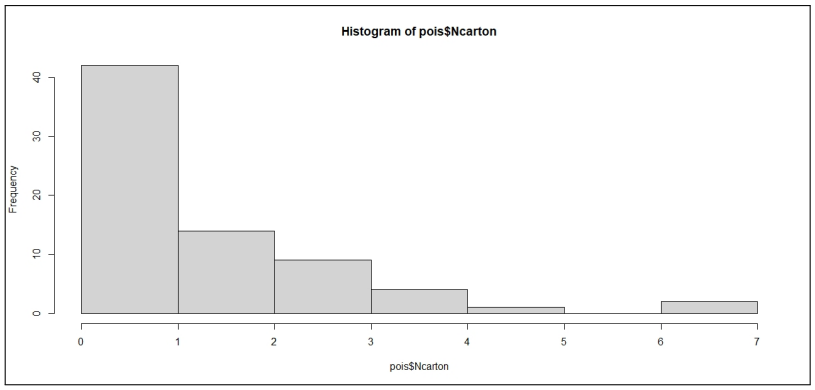


FIGURE 4.4 – Histogramme des effectifs "Nombre de but"

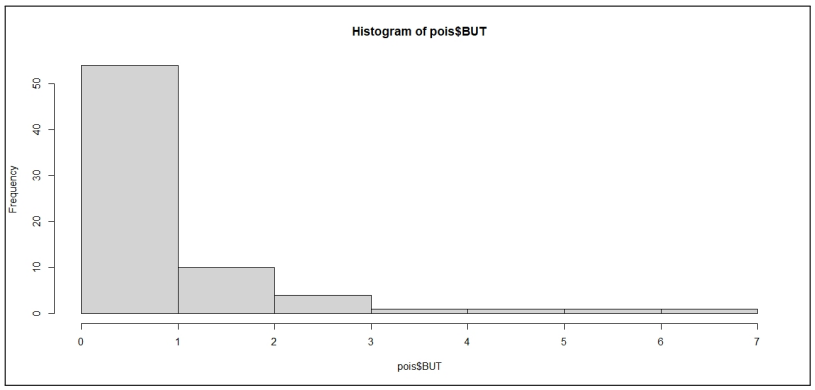


FIGURE 4.5 – Histogramme des effectifs " Nombre des cartons jaunes et rouges"

4.2.3 Estimation des paramètres

Le tableau 3.11 représente Les estimateurs des paramètres du modèle complet (modéle 1).

Le paramètre	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.52562	0.16804	3.128	0.00176 **
domicileND	-0.73860	0.24338	-3.035	0.00241 **
Ncarton	-0.11737	0.07888	-1.488	0.13675

TABLE 4.11 – Estimation des paramètres du modèle 1

La variable "domicile" est statistiquement significative, alors que la variable "Ncarton" n'est pas significative.
Pour choisir le modèle adéquate, on va utiliser la méthode à pas descendant, le modèle 2 présente les valeurs d'estimation du modèle sans la variable "Ncarton" sont illustrés dans le tableau ci dessus :

Le paramètre	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.3677	0.1387	2.652	0.00801 **
domicileND	-0.7324	0.2434	-3.009	0.00262 **

TABLE 4.12 – Estimation des paramètres du modèle 2

La sous dispersion de ces modèles (déviance résiduelle > ddl, voir tableau 3.15) nous ramène à estimer les deux modèles avec la méthode quasi Poisson, les modèles 3 et 4 présentent l'estimation quasi Poisson des modèles 1 et 2 respectivement. Les résultats d'estimations sont comme suit :

Le paramètre	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.5256	0.2215	2.373	0.0204 *
domicileND	-0.7386	0.3208	-2.303	0.0243 *
Ncarton	-0.1174	0.1040	-1.129	0.2628

TABLE 4.13 – Estimation des paramètres du modèle 3

Le paramètre	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.3677	0.1837	2.001	0.0492 *
domicileND	-0.7324	0.3225	-2.271	0.0262 *

TABLE 4.14 – Estimation des paramètres du modèle 4

Le modèle	déviance résiduelle	ddl	AIC
modp	107.60	69	211.92
modp2	110.01	70	212.33
modp3	107.60	69	
modp4	110.01	70	

TABLE 4.15 – déviance résiduelle, ddl et AIC des modèles

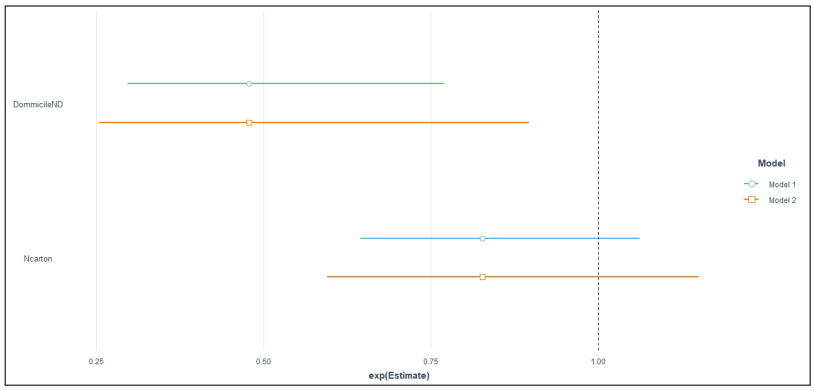


FIGURE 4.6 – Coefficients de régression des modèles 1 et 3

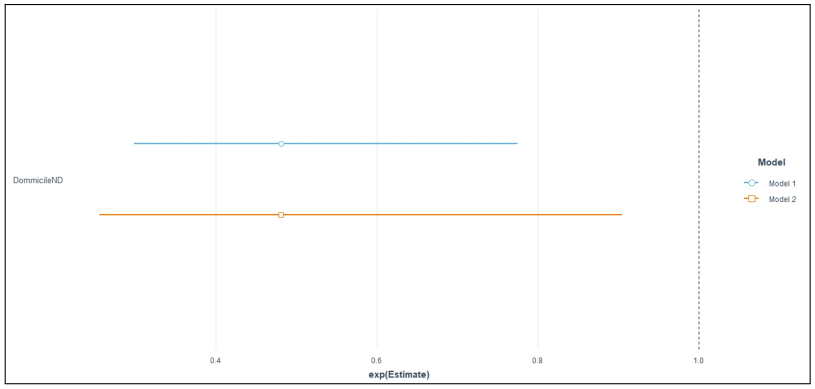


FIGURE 4.7 – Coefficients de régression des modèles 2 et 4

Le choix du modèle avec le critère de la déviance résiduelle n'est pas valable car les déviances du premier et du deuxième modèle sont égaux, ainsi on va utiliser l'écart type pour comparer entre les deux modèle.

D'aprés les tableau 3.12 et 3.14, les écarts types des paramètres du modèle 2 est inférieur à celui du modèle 4, on conclut que le modèle adéquate est le modèle 2 qui s'écrit :

$$logp(\lambda) = 0.3677 - 0.7324 dommicileND$$

Interprétation

l'estimation du paramètre "Domicile" est égal à la valeur -0.7324, donc
$exp(-0.7324) = 0.4807$, ce qui montre que le lieu du match a un effet sur la performance des joueurs cela ce traduit par une diminution de nombre des buts marquée pendant un match hors domicile avec une probabilité de 52% (1- 0.4807).

Conclusion générale

Le modèle linéaire généralisé est un outil qui peut être utilisé dans de nombreuses situations, afin d'analyser des variables qui présentent différents types de distribution statistique. Nous avons vu dans ce mémoire quelque type des MLGs avec leur application, les cas où la variable suit une loi de Poisson ou une loi Binomiale, cas qui restent les plus fréquents. D'autres lois de distribution peuvent être appliquées en utilisant à chaque fois une fonction de lien appropriée.

Dans notre cas pratique on a procéder à deux applications des MLGs.

- La modélisation du diabète par la régression logistique : Notre modèle

  $$logp(Y_i = 1|X_i) = -1.18956 + 0.12310 Age_i - 4.56727 Chol_i + 8.86782 Glec_i - 0.10067 Poids_i$$

  a un taux très important de bon classement estimé de 91.8%, ce qui reflète le pouvoir prédictif du modèle.
- La modélisation du nombre des buts dans un match par la régression de Poisson :

  $$logp(\lambda) = 0.3677 - 0.7324 dommicileND$$

  Dont on a conclut que le lieu du match a un impact sur la performance des joueurs.

Résumé

Les modèles linéaires généralisés (MLGs) sont une généralisation bien connue de modèle de régression linéaire dans les cas où la réponse est une variable discrète ou que le modèle est différent des modèles linéaires standards. Les modèles linéaires généralisés utilisés le plus souvent sont des modèles de régression logistiques pour des données binaires et des modèles log-linéaires pour des données non binaires (Poisson).
Dans ce travaille, on a présenté ces différents types des MLGs avec leurs estimations et tests statistiques, on a aussi appliqué ces tests sur des données réels à l'aide du programme R.

Mots clés : Les modèles linéaires généralisés, régression linéaire, régression logistique, régression de Poisson.

Abstract

Generalized Linear Models (GLMs) are a well-known generalization of model linear regression in cases where the response is a discrete variable or the model is different from standard linear models. The generalized linear models utilized often are logistic regression models for binary data and log-linear models for non-binary data (Poisson).
In this work, we have presented these different types of MLGs with their estimation and as well as their statistical test. Thus the application with real data is achieved with the R program.

Keywords : Generelized linear models, linear regression, logistic regression, Poisson regression.

Annexe 1

Programme de modélisation du diabète par le modèle de régression logistique sous logiciel R.

Importation de la base des données Excel vers R en csv et sa statistique descriptive

```
> diab<-read.table(file=file.choose(),header=TRUE,sep=";",dec=".")
> diab<-diab[,1:13]
> diab<-diab[1:97,]
> summary(diab)
     Sexe                Age              Poids            Antfam
 Length:97          Min.   :22.00   Min.   : 41.00   Length:97
 Class :character   1st Qu.:48.00   1st Qu.: 70.00   Class :character
 Mode  :character   Median :59.00   Median : 79.00   Mode  :character
                    Mean   :55.97   Mean   : 79.21
                    3rd Qu.:66.00   3rd Qu.: 88.00
                    Max.   :84.00   Max.   :132.00
      Chol             HDL              LDL              Trg
 Min.   :0.475   Min.   :0.2800   Min.   :0.2300   Min.   :0.320
 1st Qu.:1.410   1st Qu.:0.3800   1st Qu.:0.7200   1st Qu.:0.730
 Median :1.750   Median :0.4700   Median :0.9400   Median :1.160
 Mean   :1.733   Mean   :0.4872   Mean   :0.9813   Mean   :1.393
 3rd Qu.:2.000   3rd Qu.:0.5300   3rd Qu.:1.2000   3rd Qu.:1.660
 Max.   :4.960   Max.   :1.6600   Max.   :4.1900   Max.   :6.300
     Creat            Glec            Album            HbAlc
 Min.   : 1.80   Min.   :0.650   Min.   :  5.00   Min.   :0.0500
 1st Qu.: 7.00   1st Qu.:0.930   1st Qu.: 25.20   1st Qu.:0.0500
 Median : 9.00   Median :1.020   Median : 45.10   Median :0.0700
 Mean   : 9.21   Mean   :1.186   Mean   : 43.28   Mean   :0.2014
 3rd Qu.:11.00   3rd Qu.:1.310   3rd Qu.: 56.34   3rd Qu.:0.0800
 Max.   :19.00   Max.   :2.890   Max.   :200.00   Max.   :5.8000
    Diabet
 Length:97
 Class :character
 Mode  :character
```

```
> is.factor(diab$Diabet)
[1] FALSE
> diab$Diabet<-factor(diab$Diabet)
> diab$Sexe<-factor(diab$Sexe)
> diab$Antfam=factor(diab$Antfam)
> library(questionr)
Message d'avis :
le package 'questionr' a été compilé avec la version R 4.1.3
> freq(diab$Sexe)
   n    % val%
F 53 54.6 54.6
H 44 45.4 45.4
> freq(diab$Antfam)
     n  % val%
NON 33 34   34
OUI 64 66   66
> freq(diab$Diabet)
     n    % val%
NON 50 51.5 51.5
OUI 47 48.5 48.5
> levels(diab$Diabet)
[1] "NON" "OUI"
```

Estimation du modèle

```
> modl<-glm(Diabet~Age+Poids+Chol+Glec,data=diab,family=binomial)
> summary(modl)

Call:
glm(formula = Diabet ~ Age + Poids + Chol + Glec, family = binomial,
    data = diab)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5152  -0.3254  -0.0038   0.1807   3.2215

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.18956    3.83986  -0.310 0.756718
Age          0.12310    0.03620   3.400 0.000673 ***
Poids       -0.10067    0.03514  -2.865 0.004172 **
Chol        -4.56727    1.46779  -3.112 0.001860 **
Glec         8.86782    2.65245   3.343 0.000828 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 134.378  on 96  degrees of freedom
Residual deviance:  43.915  on 92  degrees of freedom
AIC: 53.915

Number of Fisher Scoring iterations: 7
```

L'estimation de OR

```
> odds.ratio(modl)
Waiting for profiling to be done...
                    OR      2.5 %     97.5 %         p
(Intercept) 3.0435e-01 1.4036e-04 6.7377e+02 0.7567181
Age         1.1310e+00 1.0645e+00 1.2313e+00 0.0006734 ***
Poids       9.0423e-01 8.3341e-01 9.5960e-01 0.0041721 **
Chol        1.0386e-02 3.5407e-04 1.2800e-01 0.0018604 **
Glec        7.0998e+03 9.9393e+01 4.2660e+06 0.0008280 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Prévision

```
> diabet.pred <- predict(modl, type = "response", newdata = diab)
> head(diabet.pred)
        1         2         3         4         5         6
0.9496970 0.9326276 0.9549904 0.9089076 0.9996224 0.9994180
> table(diabet.pred > 0.5, diab$Diabet) %matrice de confusion
Erreur : entrée inattendue dans "table(diabet.pred > 0.5, diab$Diabet)
> head(diabet.pred)
        1         2         3         4         5         6
0.9496970 0.9326276 0.9549904 0.9089076 0.9996224 0.9994180
> table(diabet.pred>0.5,diab$Diabet)

       NON OUI
  FALSE  47   5
  TRUE    3  42
```

Annexe 2

Programme de modélisation par le modèle de régression de Poisson sous logiciel R.

Importation de la base des données Excel vers R en csv et sa statistique descriptive

```
> pois<-read.table(file=file.choose(),header=TRUE,sep=";",dec=".")
> mean(pois$BUT)
[1] 1.069444
> var(pois$BUT)
[1] 2.009194
> mean(pois$Ncarton)
[1] 1.458333
> var(pois$Ncarton)
[1] 2.617958
> is.factor(pois$Dommicile)
[1] FALSE
> pois$Dommicile=factor(pois$Dommicile)
> is.factor(pois$Dommicile)
[1] TRUE
>
```

Estimation des modèles

```
> modp<-glm(BUT~Dommicile+Ncarton,data=pois,family=poisson)
> summary(modp)

Call:
glm(formula = BUT ~ Dommicile + Ncarton, family = poisson, data = pois)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8393  -1.1989  -0.4380   0.4416   3.5891

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.52562    0.16804   3.128  0.00176 **
DommicileND -0.73860    0.24338  -3.035  0.00241 **
Ncarton     -0.11737    0.07888  -1.488  0.13675
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 119.68  on 71  degrees of freedom
Residual deviance: 107.60  on 69  degrees of freedom
AIC: 211.92

Number of Fisher Scoring iterations: 6

    > modp2<-glm(BUT~Dommicile,data=pois,family=poisson)
    > summary(modp2)

    Call:
    glm(formula = BUT ~ Dommicile, family = poisson, data = pois)

    Deviance Residuals:
        Min       1Q   Median       3Q      Max
    -1.6997  -1.1785  -0.3917   0.3670   3.3361

    Coefficients:
                Estimate Std. Error z value Pr(>|z|)
    (Intercept)   0.3677     0.1387   2.652  0.00801 **
    DommicileND  -0.7324     0.2434  -3.009  0.00262 **
    ---
    Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

    (Dispersion parameter for poisson family taken to be 1)

        Null deviance: 119.68  on 71  degrees of freedom
    Residual deviance: 110.01  on 70  degrees of freedom
    AIC: 212.33

    Number of Fisher Scoring iterations: 6
```

Estimation quasi poisson des modèles

```
> modp3<-glm(BUT~Dommicile+Ncarton,data=pois,family=quasipoisson(link="log"))
> summary(modp3)

Call:
glm(formula = BUT ~ Dommicile + Ncarton, family = quasipoisson(link = "log"),
    data = pois)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8393  -1.1989  -0.4380   0.4416   3.5891

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.5256     0.2215   2.373   0.0204 *
DommicileND  -0.7386     0.3208  -2.303   0.0243 *
Ncarton      -0.1174     0.1040  -1.129   0.2628
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasipoisson family taken to be 1.736945)

    Null deviance: 119.68  on 71  degrees of freedom
Residual deviance: 107.60  on 69  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 6

> modp4<-glm(BUT~Dommicile,data=pois,family=quasipoisson(link="log"))
> summary(modp4)

Call:
glm(formula = BUT ~ Dommicile, family = quasipoisson(link = "log"),
    data = pois)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6997  -1.1785  -0.3917   0.3670   3.3361

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.3677     0.1837   2.001   0.0492 *
DommicileND  -0.7324     0.3225  -2.271   0.0262 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasipoisson family taken to be 1.755473)

    Null deviance: 119.68  on 71  degrees of freedom
Residual deviance: 110.01  on 70  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 6
```

Annexe 3

La table statistique de la loi normale.

Tenths Digits	Hundredths Digits									
	0.00	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

La table statistique de la loi χ^2

Degrees of freedom (*df*)	Significance level (α)							
	.99	.975	.95	.9	.1	.05	.025	.01
1	-------	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345
4	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277
5	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086
6	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812
7	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475
8	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090
9	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666
10	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209
11	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725
12	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217
13	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688
14	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141
15	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578
16	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000
17	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409
18	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805
19	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191
20	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566
21	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932
22	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289
23	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638
24	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980
25	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314
26	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642
27	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963
28	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278
29	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588
30	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892
40	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691
50	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154
60	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379
70	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425
80	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329
100	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116
1000	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807

Table statistique de Ficher

ν_2 \ ν_1	1	2	3	4	5	6	8	12	24	>25
1	161.4	199.5	215.7	224.6	230.2	234.0	238.9	243.9	249.0	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.37	19.41	19.45	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.84	8.74	8.64	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.04	5.91	5.77	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.82	4.68	4.53	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.15	4.00	3.84	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.73	3.57	3.41	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.44	3.28	3.12	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.23	3.07	2.90	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.07	2.91	2.74	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	2.95	2.79	2.61	2.40
12	4.75	3.88	3.49	3.26	3.11	3.00	2.85	2.69	2.50	2.30
13	4.67	3.80	3.41	3.18	3.02	2.92	2.77	2.60	2.42	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.70	2.53	2.35	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.64	2.48	2.29	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.59	2.42	2.24	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.55	2.38	2.19	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.51	2.34	2.15	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.48	2.31	2.11	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.45	2.28	2.08	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.42	2.25	2.05	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.40	2.23	2.03	1.78
23	4.28	3.42	3.03	2.08	2.64	2.53	2.38	2.20	2.00	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.36	2.18	1.98	1.73
25	4.24	3.38	2.99	2.76	2.60	2.49	2.34	2.16	1.96	1.71
26	4.22	3.37	2.98	2.74	2.59	2.47	2.32	2.15	1.95	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.30	2.13	1.93	1.67
28	4.20	3.34	2.95	2.71	2.56	2.44	2.29	2.12	1.91	1.65
29	4.18	3.33	2.93	2.70	2.54	2.43	2.28	2.10	1.90	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.27	2.09	1.89	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.18	2.00	1.79	1.51
60	4.00	3.15	2.76	2.52	2.37	2.25	2.10	1.92	1.70	1.39
120	3.92	3.07	2.68	2.45	2.29	2.17	2.02	1.83	1.61	1.25
>120	3.84	2.99	2.60	2.37	2.21	2.10	1.94	1.75	1.52	1.00

Bibliographie

[1] **Agresti.A** : *Foundations Of Linear and Generalized Linear Models*, WILEY, 2015.

[2] **Chavent.M** :*Régression linéaire multiple*, Université de Bordeaux.

[3] **Chesneau.C** :*Modèles de régression*, Université de Caen, 2020.

[4] **Dobson.A, Barnett.A** : *An Introduction to Generalized Linear Models*, CHAPMAN HALL/CRC, 2008.

[5] **Fox.J** : *Applied Regression Analysis and Generalized Linear models*, SAGE, 2016.

[6] **Frank.E, Harrell.Jr.** :*Regression Modeling Strategies With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis.*

[7] **Hosmer.V, Lemeshow.D** :*Applied Logistic Regression*, John Wiley Sons, INC, 2000.

[8] **Houde.L** : *Lois De Probabilité*, Université du Québec à Trois-Rivières, 2014.

[9] **Myers.R, Montgomery.D, Vining.G, Robinson.T** : *Generalized Linear Models with Applications in Engineering and the Sciences*, Wiley, 2010.

[10] **Olsson.U** : *Generalized linear models an applied approach*, Studentlitteratur, 2002.

[11] **Rakotomalala.R** : *Pratique de la Régression Logistique Régression Logistique Binaire et Polytomique*, Université Lumière Lyon 2, 2017.

[12] **Trabelsi.A, Respriget.R** :*Régression de Poisson.*