République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université A. MIRA de Béjaia

Faculté des Sciences Exactes

Département Mathématiques

Mémoire de Master en Mathématiques

Présenté par :

TOUFOUTA Noura & BOUAMAR Zineb

En vue de l'obtention du diplôme de :

Master en Mathématiques

Option : Statistique et Analyse Décisionnelle

Thème

Modèles linéaires généralisés (estimations et prédictions)

Soutenue publiquement le 20/06/2017 **devant le jury suivant :**

Présidente	M^{me} BARECHE A.	Professeur	U. de Béjaia
Promoteur	M^{r} BOURAINE M.	M.A.A	U. de Béjaia
Examinateur	M^{r}RAHMANI S.	M.C.A	U. de Béjaia

Promotion 2016/2017

Dédicaces

Je dédie ce modeste travail à mes très chers parents Djilali et Soraia et à mon mari Cherif.

à la mémoire de mes chèrs grands parents Lounes et hamou et Houaa et à la mémoire de mon oncle Farid.

à mon frère Mustapha.

à mes beaux frères Nordine, Samir et Nassim.

à mes soeurs Zahia et Farida.

à mes belles soeurs Karima, Marina, Randa, Zineb, Fozia, Sabrina et Nawel.

à mes beaux parents Mahmoud et Zara.

à mes oncles et mes tantes.

à ma grand mère laldja.

à toute la promotion SAD et à tous mes amis.

Nora

Dédicaces

Je dédie ce modeste travail à mes très chers parents Mahmoud et Zara.
à la mémoire de mes chèrs grands parents Allaoua et Said et Louiza et à la mémoire de mon oncle Zahir.
à mes frère Nordine, Samir, Nassim et cherif.
à mes soeurs Karima, Marina, Randa.
à mes belles soeurs Fozia, Sabrina, Nawel et Nora.
à mes oncles et mes tantes.
à ma grand mère Taklit.
à toute la promotion SAD et à tous mes amis.
Zineb

Remerciements

Nous remercions **Dieu tout puissant**, de nous avoir protégé et donné la force, le courage et la patience afin de terminer ce travail.

On exprime notre profonde gratitude à tout le département mathématique spécialement notre promoteur M^r BOURAINE Mohand pour sa générosité, son encouragement et son suivie afin de réaliser ce mémoire.

Nos remerciement s'adressent aussi à la présidente des jury M^{me} Bareche A. et l'examinateur M^r Rahmani S. pour l'honneur qu'ils nous font en jugeant notre soutenance.

Enfin, nous remercions toute personnes qui ont de près ou de loin participé à la réalisation de ce travail.

Table des matières

Introduction générale

La Statistique a plusieurs objets : descriptif ou exploratoire, décisionnel (tests), modélisation selon que l'on cherche à représenter des structures de donnée, confirmer ou expliciter un modèle théorique ou encore prévoir. Différents modèles ont été développés afin de pouvoir modéliser différents problèmes statistiques dont "modèles linéaires,modèles de dénombrements, familles exponentielles, modèles log linéaire etc.
Dans l'arsenal des modèles statistiques, les modèles dits linéaires sont largement dominants. Que les variables prédictives au sein de ces modèles soient numériques, catégorielles (recodées en indicatrices) ou les deux en même temps, on peut exprimer dans une formulation unifiée les méthodes bien connues que sont la régression linéaire, l'analyse de la variance et l'analyse de la covariance. Ce cadre simple, déjà assez intégrateur, est classiquement nommé Modèle Linéaire Général. Il fait l'hypothèse d'une distribution gaussienne sur la variable dépendante, conditionnellement aux prédicteurs, et d'un lien linéaire ou de proportionnalité entre variables explicatives et à expliquer. Or ces modèles ne sont pas toujours adéquats à tous problèmes statistiques, certains sont assez compliqué pour qu'ils soient modélisés par ces modèles simples.

En 1972 de nouveaux modèles ont été formulés par John Nelder et Robert Wedderburn appelés modèles linéaires généralisés comme une généralisation souple de la régression linéaire. Le GLM "General Lineair Modeles" généralise la régression linéaire en permettant au modèle linéaire d'être relié à la variable réponse via une fonction lien [2]. Aussi comme un moyen d'unifier les autres modèles statistiques y compris la régression linéaire, la régression logistique et la régression de Poisson. Ils proposent une méthode itérative dénommée méthode des moindres carrés repondérés itérativement pour l'estimation du maximum de

vraisemblance des paramètres du modèle. L'estimation du maximum de vraisemblance reste populaire et est la méthode par défaut dans de nombreux logiciels de calculs statistiques.

Notre objective dans ce mémoire est bien de présenter les différents modèles statistiques utilisés afin de modéliser différents problèmes recontrés dans la pratique. Un intérêt particulier est donné au modèles linéaires généralisés.

Ce mémoire est organisé en trois chapitres :

- Le premier chapitre sera consacré à la présentation des différents modèles statistiques :
  - modèles linéaire : régression linéaire simple et multiple ;
  - modèles de dénombrement : régression logistique et log linéaire.
- Le deuxième chapitre portera sur une formalisation mathématique d'un GLM et les outils statistiques permettant de réaliser notre étude.
- Le troisième chapitre présentera les différentes applications que nous avons réalisé sur les modèles étudiés.

Nous avons achevé notre travail par une conclusion générale.

Première partie

Partie théorique

Chapitre 1

Représentation des modèles statistiques

1.1 Introduction

Le cadre général de ce chapitre considère donc les observations d'une variable aléatoire Y dite réponse, exogène, dépendante qui doit être expliquée (modélisée) par les mesures effectuées sur p variables dites explicatives, de contrôle, endogènes, dépendantes, régresseurs. Ces variables peuvent être quantitatives ou qualitatives, ce critère détermine le type de méthode ou de modèle à mettre en oeuvre : régression linéaire, régression logistique, données de comptage et le modèle log-linéaire.

1.2 Modèles de régression linéaire

En statistique, en économétrie et en apprentissage automatique, un modèle de régression linéaire est un modèle de régression qui cherche à établir une relation linéaire entre une variable, dite expliquée, et une ou plusieurs variables, dites explicatives.

1.2.1 Régression linéaire simple

On appelle généralement modèle linéaire simple un modèle de régression linéaire avec une seule variable explicative, où cette variable Y est expliquée, modélisée par une fonction affine d'une autre variable X. La finalité d'un tel modèle est multiple et dépend donc du contexte et surtout des questions sous-jacentes. C'est une approche exploratoire ou une recherche d'une réponse à une question du type : une variable quantitative X par exemple : "la concentration d'une molécule" a-t-elle une influence sur la variable quantitative Y exemple : "une culture bactérienne". Ou enfin la recherche d'un modèle de prévision de Y en fonction de X : calibration d'un appareil de mesure d'une concentration à partir d'une mesure optique. Des concepts clefs : modèle, estimations, tests, diagnostics sont introduits et déclinés dans ce contexte élémentaire. Leur emploi et leur signification dépendent des objectifs. Ils se retrouvent dans une présentation plus générale du modèle de régression multiple et cette première partie sert donc d'introduction.
Avant tout travail de modélisation, une approche descriptive ou exploratoire est nécessaire pour dépister des difficultés dans les données : dissymétrie des distributions, valeurs atypiques, liaison non linéaire entre les variables. En fonction des résultats obtenus, une transformation préalable des variables peut s'avérer nécessaire.
Ce modèle est souvent présenté dans les manuels de statistiques à des fins pédagogiques, sous le titre d'ajustement affine.

1.2.1.1 Le modèle

On note Y la variable aléatoire réelle à expliquer et X la variable explicative (déterministe) ou effet fixe ou facteur contrôlé. Le modèle revient à supposer, qu'en moyenne, E(Y) est une fonction affine de X [9].

– Dans le cas où X est déterministe, le modèle s'écrit :

$$E(Y) = f(X) = a_0 + a_1 X. \tag{1.1}$$

– Dans le cas où X est aléatoire, le modèle s'écrit alors conditionnellement aux observations de X :

$$E(Y|X = x) = a_0 + a_1 X \tag{1.2}$$

Il conduit aux même estimations

– pour une séquence d'observations aléatoires identiquement distribuées $(y_i, x_i); i = 1, ..., n (n > 2$, et les x_i non tous égaux) le modèle s'écrit avec les observation :

$$y_i = a_0 + a_1 x_i + \epsilon_i; i = 1, ..., n \tag{1.3}$$

Les hypothèses relatives à ce modèle sont les suivantes :

– la distribution de l'erreur ϵ est indépendante de X ou X est fixe,

– l'erreur est centrée et de variance constante (homoscédasticité) :

$$\forall i = 1, ..., n; E(\epsilon_i) = 0; Var(\epsilon_i) = \sigma_\epsilon^2 \tag{1.4}$$

– a_0 et a_1 sont constants, pas de rupture du modèle.

– Hypothèse complémentaire pour les inferences :$\epsilon \sim N(0, \sigma_\epsilon^2 \, 1_p)$

– **Forme matricielle :** Le modèle s'écrit sous la forme suivantes :

$$Y = Xa + \epsilon \Leftrightarrow \begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ . & . \\ . & . \\ . & . \\ 1 & x_n \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ . \\ . \\ . \\ \epsilon_n \end{pmatrix} \tag{1.5}$$

– **Sous les hypothèses de base suivantes :**

  – $H_1 : E(\epsilon_t) = 0. \, \forall t = 1, ..., n$. Le modèle est en moyenne bien spécifié.

  – $H_2 : Var(\epsilon_t) = E(\epsilon_t^2) = \sigma_\epsilon^2, \, \forall t = 1, ..., n$. (Le risque d'amplitude de l'erreur est le même quelque soit la période).

  – H_3 : les erreurs sont non corrélés $cov(\epsilon_t, \epsilon_k) = 0, \, \forall t \neq k$.

  – H_4 : la distribution de l.erreur ϵ_t est indépendante de x.

  – H_5 : La variable explicative X est non aléatoire et prend au moins deux valeurs différentes.

1.2.1.2 Estimation

L'estimation des paramètres a_0,a_1 et σ^2 est obtenue en maximisant la vraisemblance, sous l'hypothèse que les erreurs sont gaussiennes, ou encore par minimisation de la somme des carrés des écarts entre observations et modèle (moindres carrés). Pour un jeu de données $(x_i, y_i); i = 1, ..., n$ le critère des moindres carrés s'écrit :

$$\min_{a_0,a_1} \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i)^2 \tag{1.6}$$

On pose :

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i; \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \tag{1.7}$$

$$S_x^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2; S_y^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{1.8}$$

$$S_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}); r = \frac{S_{xy}}{S_x S_y} \tag{1.9}$$

Avec S_{xy} est la covariance empirique entre x et y, S_x^2 est la variance empirique de x et S_y^2 est la variance empirique de y.

Les moindres carrés sont donnés par les formules suivantes :

$$\hat{a}_1 = \frac{S_{xy}}{S_x^2} \tag{1.10}$$

$$\hat{a}_0 = \bar{y} - \hat{a}_1 \bar{x} \tag{1.11}$$

Ces estimateurs sont des estimateurs sans biais et de variances minimum parmi les estimateurs fonctions linéaires des y_i resp. (parmi tous les estimateurs dans le cas gaussien).A chaque valeur de X correspond la valeur estimée (ou prédite, ajustée) de Y :

$$\hat{y}_i = \hat{a}_0 + \hat{a}_1 x_i \tag{1.12}$$

Les résidus calculés ou estimés sont :

$$e_i = y_i - \hat{y}_i. \tag{1.13}$$

La variance σ_ϵ^2 est estimée par la variation résiduelle :

$$S^2 = \frac{1}{n-2} \sum_{i=1}^{n} e_i^2 \tag{1.14}$$

Le coefficient de corrélation entre X et Y est donnée par :

$$\rho_{x,y} = \frac{S_{xy}}{S_x S_y} \tag{1.15}$$

Et le coefficient de détermination donné par la formule suivante :

$$R^2 = \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT}, \text{ où } 0 \leq R^2 \leq 1. \tag{1.16}$$

Avec :

- $SCT = \sum_{t=1}^{n}(y_t - \bar{y})^2$, "Somme des Carrés Totale ou Variation totale";
- $SCE = \sum_{t=1}^{n}(\hat{y}_t - \bar{y})^2 = \hat{a}_1 \sum_{t=1}^{n}(x_t - \bar{x})^2$, "Somme des Carrés Expliquée ou Variation Expliquée par la régression";
- $SCR = \sum_{t=1}^{n}(y_t - \hat{y})^2 = \sum_{t=1}^{n} \epsilon_t^2$, "Somme des Carrés Résiduelle ou Variation Résiduelle Appliquée par les Résidus".

1.2.1.3 Les tests statistiques

Même si le coefficient de détermination est grand cela ne signifie pas toujours que le modèle est bon, d'où l'utilité des tests statistiques.

- **Test de fisher :**

  Afin de construire ce test nous avons besoin d'une hypothèse supplémentaire qui est caractérisée par la normalité des erreurs ϵ_t.

  - Soit H_6 :

    $$\epsilon_t \sim N(0, \sigma_\epsilon^2) \text{ avec } t = 1...n.$$

  - Ce test repose sur l'indicateur suivant :

    $$F* = \frac{SCE}{\frac{SCR}{n-2}} = \frac{R^2}{\frac{1-R^2}{n-2}} \tag{1.17}$$

  - Le test de Ficher consiste à vérifier si le modèle est globalement significatif. Cela revient à tester dans le modèle si la variable explicative x est significative . On considère le test suivant :

    $$H_0 : a_1 = 0 \quad \text{contre} \quad H_1 : a_1 \neq 0$$

    .

- Sous H_0, $F^* \sim Fisher(1, n-2)$. Pour un seuil de signification α "généralement égal à 0.05", $f_\alpha(1; n-2)$ est la valeur lue sur la table de Fisher à 1 et n-2 d.d.l.
  - Si $F^* > f_\alpha(1; n-2)$, alors on rejette H_0. Par conséquent, on accepte H_1. D'où la variable explicative est significative dans le modèle. Le modèle est alors globalement significatif.
  - Si $F^* \leq f_\alpha(1; n-2)$ alors on accepte H_0. D'où, la variable explicative du modèle est non significative au seuil α : Le modèle n'est alors pas globalement significatif.

- **Test de Student :**

  Le test de Student, consiste à mesurer la signification statistique de la variable explicative x, sur la variable dépendante Y.
  - On note :

$$t^*_{\hat{a_1}} = \frac{\hat{a}_1}{\hat{\sigma}_{\hat{a}_1}}$$

  - Considérons le test suivant :

$$H_0 : a_1 = 0 \text{ contre } H_1 : a_1 \neq 0$$

  - Le test est basé sur la statistique suivante :

$$t^*_{\hat{a}_1} = \frac{\hat{a}_1}{\hat{\sigma}_{\hat{a}_1}} \sim t(n-2)$$

  - Sous H_0 : $t^*_{\hat{a_1}} \sim t(n-2)$ et on note $t_{(\frac{\alpha}{2}, n-2)}$la valeur lue sur la table de Student pour un seuil de signification fixé α et $(n-2)$ d.d.l. On a la règle de décision suivante :
    - Si $t_{\hat{a}_1} =\mid \frac{\hat{a}_i}{\hat{\sigma}_{\hat{a}_1}} \mid > t_{(\frac{\alpha}{2}, n-2)}$, on rejette H_0, d'où on accepte H_1 : Le paramètre a_1 est alors significativement différent de 0. La variable explicatif est alors significativement dans le modèle.
    - Si $t_{\hat{a}_1} =\mid \frac{\hat{a}_i}{\hat{\sigma}_{\hat{a}_1}} \mid \leq t_{(\frac{\alpha}{2}, n-2)}$, on accepte H_0. D'où le paramètre $a_1 = 0$ la variable X est alors non significative de Y :

1.2.1.4 Intervalle de confiance

Pour obtenir un intervalle de confiance pour le paramètre a_1 au seuil de signification α (ou bien au niveau de confiance $1-\alpha$), on utilise la statistique :

$$t^*_{\hat{a_1}} = \frac{\hat{a}_1 - a_1}{\sigma_{a1}} \sim Student(n-2).$$

D'où,

$$P[-t_{\frac{\alpha}{2}} < \frac{\hat{a}_1 - a_1}{\sigma_{\hat{a}_1}} < t_{\frac{\alpha}{2}}] = 1-\alpha \Leftrightarrow P[\hat{a}_1 - t_{\frac{\alpha}{2}}\sigma_{\hat{a}_1} < a_1 < \hat{a}_1 + t_{\frac{\alpha}{2}}\sigma_{\hat{a}_1}] = 1-\alpha.$$

D'où, l'intervalle de confiance pour au seuil α est donnée par :

$$IC_\alpha(a_1) = [\hat{a}_1 - t_{\frac{\alpha}{2}}\sigma_{\hat{a}_1}, \hat{a}_1 + t_{\frac{\alpha}{2}}\sigma_{\hat{a}_1}].$$

1.2.1.5 Exemple de regression linéaire simple

L'analyse de la température de fonctionnement d'un procédé chimique sur le rendement de produit a donné les valeurs suivantes :
Pour la température x_i et le rendement correspondant y_i.

Température C°	**Rendement %**
100	45
110	51
120	54
130	61
140	66
150	70
160	74
170	78
180	85
190	89

TABLE 1.1: *Tableau des données de l'analyse de la température*

Le graphe ci-dessous représente les points $(x_i; y_i)$ pour ces données et suggère une relation linéaire entre X et Y.

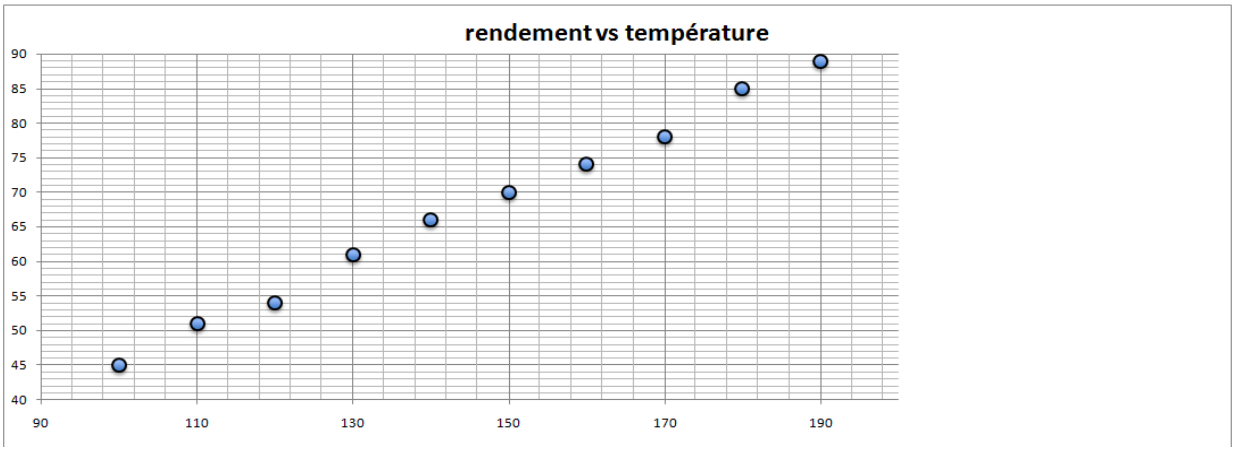

FIGURE 1.1: *rendement vs température.*

La droite de régression estimée est :

$$\hat{Y} = \hat{a}_0 + \hat{a}_1 X$$

Les variables aléatoires $\hat{a}_0$ et $\hat{a}_1$ sont des estimateurs de a_0 et de la pente a_1.

Les valeurs estimées en chaque point de X_i pour $i = 1, ..., 10$ sont données dans les tableaux suivants :

Température C°	**Rendement %**	$(X_i - \bar{x})^2$	$(Y_i - \bar{Y})^2$	$(X_i - \bar{x})(Y_i - \bar{Y})$
100	45	2025	497.29	1003.5
110	51	1225	265.69	570.5
120	54	625	176.89	332.5
130	61	225	39.69	94.5
140	66	25	1.69	6.5
150	70	25	7.29	13.5
160	74	225	44.89	100.5
170	78	625	114.49	267.5
180	85	1225	313.29	619.5
190	89	2025	470.89	976.5

TABLE 1.2: *tableau1 des estimations*

$\hat{Y}$	$(Y_i - \hat{Y}_i)^2$	$(\hat{Y}_i - \bar{Y})^2$
45.5636	0.3176	472.4695
50.3939	0.3673	285.8148
55.2242	1.4987	145.8239
60.0545	0.8938	52.4966
64.8848	1.2435	5.8329
69.7151	0.0811	5.8329
74.5454	0.2975	52.4966
79.3755	1.8927	145.8239
84.2060	0.6303	285.8148
89.0363	0.0013	472.4695

TABLE 1.3: *Tableau 2 des estimations*

La droite de régression obtenue est donnée dans la figure suivantes :

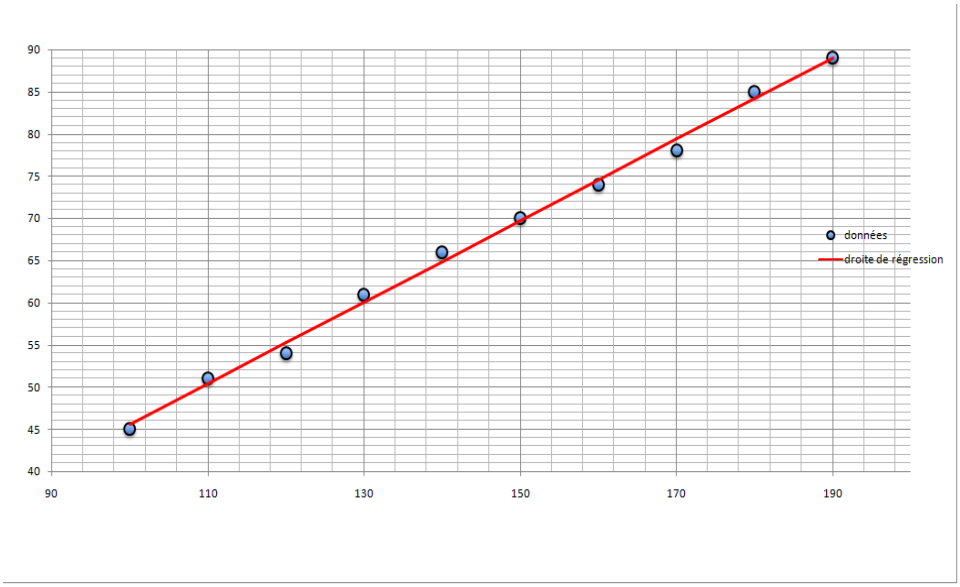

FIGURE 1.2: *rendement vs température.*

1.2.2 Régression linéaire multiple

Le modèle de régression linéaire multiple est l'outil statistique le plus habituellement mis en oeuvre pour l'étude de données multidimensionnelles. Cas particulier de modèle linéaire, il constitue la généralisation naturelle de la régression simple dans lequel une variable quantitative Y est expliquée, modélisée, par plusieurs variables quantitatives X_j pour $j = 1, ..., p$.

1.2.2.1 Le modèle

Une variable quantitative Y dite à expliquer (ou encore, réponse, exogène, dépendante) est mise en relation avec p variables quantitatives $X^1, ..., X^p$ dites explicatives (ou encore de contrôle, endogènes, indépendantes, regresseurs)[10].
Les données sont supposées provenir de l'observation d'un échantillon statistique de taille n $(n > p + 1)$ de $\mathbb{R}^{(p+1)}$:

$$(x_i^1, ..., x_i^j, ..., x_i^p, y_i); i = 1, ..., n. \tag{1.18}$$

L'écriture du modèle linéaire dans cette situation conduit à supposer que l'espérance de Y appartient au sous-espace de $\mathbb{R}^n$ engendré par $\mathbf{1}, X^1, ..., X^p$ où $\mathbf{1}$ désigne le vecteur de $\mathbb{R}^n$ constitué de 1. C'est-à-dire que les $(p+1)$ variables aléatoires verifient :

$$y_i = a_0 + a_1 x_i^1 + a_2 x_i^2 + ... + a_p x_i^p + \epsilon_i; i = 1, 2, ..., n. \tag{1.19}$$

sous les hypothèses suivantes :

- Les ϵ_i sont des termes d'erreurs , d'une variable U, non observés, indépendants et identiquement distribuées ;

$$E(\epsilon_i) = 0, Var(U) = \sigma_\epsilon^2 \mathbf{I}. \tag{1.20}$$

- Les termes x^j sont supposés déterministes (facteurs contrôles) ou bien l'erreur U est indépendante de la distribution conjointe de $X^1, ..., X^p$ .On écrit dans ce dernier cas que :

$$E(Y|X^1, ..., X^p) = a_0 + a_1 x_i^1 + a_2 x_i^2 + ... + a_p x_i^p \tag{1.21}$$

$$Var(Y|X^1, ..., X^p) = \sigma_\epsilon^2 \tag{1.22}$$

- Les paramètres inconnus $a_0, ..., a_p$ sont supposés constants.
- En option, pour l'étude spécifique des lois des estimateurs, une quatrième hypothèse considère la normalité de la variable d'erreur $U \sim N(0, \sigma_\epsilon^2 \mathbf{I})$. Les ϵ_i sont alors i.i.d de loi $N(0, \sigma_\epsilon^2)$.

  Les données sont rangées dans une matrice $\mathbf{X}(n \times (p+1))$ de terme général x_i^j, dont la première colonne contient le vecteur 1 ($x_0^i = 1$) et dans un vecteur Y de terme général y_i. En notant les vecteurs :

  $\epsilon = [\epsilon_1, ..., \epsilon_p]'$ et $a = [a_0, a_1 ..., a_p]'$,le modèle s'écrit matriciellement :

$$y = \mathbf{X}a + \epsilon. \tag{1.23}$$

- **Forme matricielle**

  L'écriture précédente du modèle est d'un maniement peu pratique. Afin d'en alléger l'écriture et de faciliter l'expression de certains résultats, on a habituellement recours aux notations matricielles. En écrivant le modèle, observation par observation, on

aura :

$$\begin{cases} y_1 = a_0 + a_1 x_{1,1} + ... + a_p x_{1,p} + \epsilon_1 \\ y_2 = a_0 + a_1 x_{2,1} + ... + a_p x_{2,p} + \epsilon_2 \\ \quad . \\ \quad . \\ \quad . \\ y_n = a_0 + a_1 x_{n,1} + ... + a_p x_{n,p} + \epsilon_n \end{cases} \tag{1.24}$$

Soit, sous forme matricielle :

$$\underset{(n,1)}{Y} = \underset{(n,p+1)}{\mathbf{X}} \underset{(p+1,1)}{a} + \underset{(n,1)}{\epsilon}$$

Avec :

$$\begin{pmatrix} y_1 \\ . \\ . \\ . \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & . & . & . & x_{1p} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ 1 & x_{n1} & . & . & . & x_{np} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ . \\ . \\ . \\ a_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ . \\ . \\ . \\ \epsilon_n \end{pmatrix}. \tag{1.25}$$

Nous remarquons la première colonne de la matrice $\mathbf{X}$, composée de 1, qui correspond au coefficient a_0 (coefficient du terme constant).

La dimension de la matrice $\mathbf{X}$ est donc de n lignes et $p+1$ colonnes (p étant le nombre de variables explicatives réelles, c'est-à-dire constante exclue).

– **Sous les hypothèses de base suivantes :**

Hypothèses stochastiques :

- H_1 : Les X_j sont non aléatoires c-à-d. les x_{ij} sont observés sans erreur ;
- H_2 : $E[\epsilon_i] = 0$, l'espérance de l'erreur est nulle. En moyenne, le modèle est bien spécifié ;
- H_3 : $E[\epsilon_i^2] = \sigma_\epsilon^2$, la variance de l'erreur est constante, c'est l'hypothèse de homoscédasticité ;
- H_4 : $cov(\epsilon_i, \epsilon_j) = 0$ pour $i \neq j$, les erreurs sont indépendantes, c'est l'hypothèse de non auto-corrélation des résidus ;
- H_5 :$cov(x_{ij}, \epsilon_i) = 0$, l'erreur est indépendante des variables exogènes ;

– H_6 : $\epsilon_i \curvearrowright N(0, \sigma_\epsilon^2)$, les erreurs sont distribuées selon une loi normale.

Hypothèses structurelles :

– H_7 : La matrice (**X'X**) est régulière c-à-d. $det((\mathbf{X'X})^{-1}) \neq 0$ et $(\mathbf{X'X})^{-1}$ existe. Elle indique l'absence de colinéarité entre les exogènes ;

– H_8 : $\frac{(\mathbf{X'X})}{n}$ tend vers une matrice finie non singulière lorsque $n \curvearrowright +\infty$;

– H_9 : $n > p + 1$, le nombre d'observations est supérieur au nombre de paramètres à estimer. Dans le cas où $n = p + 1$, nous avons une interpolation, la droite passe exactement par tous les points. Lorsque $n < p + 1$, la matrice (**X'X**) n'est plus inversible.

Remarque :

On appel **X'** la matrice transposé de **X**.

1.2.2.2 Estimation des paramètres

La méthode des moindres carrés cherche la meilleure estimation des paramètres "a_i" en minimisant la quantité S :

$$S = \sum_i \epsilon_i^2 \tag{1.26}$$

avec :

$$\epsilon = Y - X\hat{a} \tag{1.27}$$

"ϵ", l 'erreur observée (le résidu).

– **L'estimateur des moindres carrès ordinaires**

– a) principe de calcul : Pour trouver les paramètres qui minimise S :

$$S = \sum \epsilon\epsilon' = \sum_i \epsilon_i^2 \tag{1.28}$$

$$= \sum_i [y_i - (a_0 + a_{i,1}x_1 + ... + a_{i,p}x_p)]^2 \tag{1.29}$$

On doit resoudre :

$$\frac{\partial S}{\partial a} = 0 \tag{1.30}$$

Il ya (p + 1) équation dite "équation normales" à résoudre

$$\begin{aligned} S = \epsilon'\epsilon &= (Y - \mathbf{X}a)'(Y - \mathbf{X}a) \\ &= (Y' - (\mathbf{X}a)')(Y - \mathbf{X}a) \\ &= (Y' - a'\mathbf{X}')(Y - \mathbf{X}a) \\ &= Y'Y - a'\mathbf{X}'Y - Y'Xa + a'\mathbf{X}'\mathbf{X}a \\ &= Y'Y - a'\mathbf{X}'Y - a'\mathbf{X}'Y + a'\mathbf{X}'\mathbf{X}a \\ &= Y'Y - 2a'\mathbf{X}'Y + a'\mathbf{X}'\mathbf{X}a \\ \Longrightarrow \frac{\partial S}{\partial a} &= -2(\mathbf{X}'Y) + 2(\mathbf{X}'\mathbf{X})a = 0 \\ \Longrightarrow \hat{a} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y \end{aligned}$$

Ce qui donne :

$$\hat{a} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y, \tag{1.31}$$

avec :

$$(\mathbf{X}'\mathbf{X}) = \begin{pmatrix} n & \sum_i x_{i,1} & . & . & . & \sum_i x_{i,p} \\ \sum_i x_{i,1} & \sum_i x_{i,1}^2 & & & & \sum_i x_{i,1}x_{i,p} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & \sum_i x_{i,p}^2 \end{pmatrix}$$

Matrice des sommes des produits croisés entre les variables exogènes-symétrique (son inverse et aussi symétrique).

Si les variables sont centrées alors $\frac{1}{n}(\mathbf{X}'\mathbf{X})$ = matrice de variance covariance.

Si les variables sont centrées et rèduites alors $\frac{1}{n}(\mathbf{X}'\mathbf{X})$ = matrice de corrélation.

$$(\mathbf{X}'Y) = \begin{pmatrix} \sum_i Y_i \\ \sum_i Y_i x_{i,1} \\ . \\ . \\ \sum_i Y_i x_{i,p} \end{pmatrix}$$

vecteur des sommes des produits croisés entre l'endogène et les variables exogènes.
Si les variables sont centrées alors $\frac{1}{n}(\mathbf{X}'Y)$ = vecteur de covariance Y et $\mathbf{X}$.
Si les variables sont centrées et réduite alors $\frac{1}{n}(\mathbf{X}'Y)$ = vecteur des corrélations entre Y et $\mathbf{X}$.

– b) Biais de $\hat{a}$: Etape1 : expimer $\hat{a}$ en fonction de a

$$\begin{aligned}\hat{a} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'[\mathbf{X}a + \epsilon] \\ \hat{a} &= a + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\epsilon;\end{aligned}$$

Etape2 : voir sous quelle conditions $E[\hat{a}] = a$

$$\begin{aligned}E[\hat{a}] &= a + E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\epsilon] \\ &= a + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E[\epsilon] \\ &= a\end{aligned}$$

Parce que $\mathbf{X}$ non aléatoire, et $E[\epsilon] = 0$ (d'aprés les hypothèses H_1 et H_2)

1.2.2.3 Matrice de variance-covariance

La matrice de variance-covariance des coefficients est importante, car elle renseigne sur la variance de chaque coefficient estimé, et permet de faire des tests d'hypothèse, notamment de voir si chaque coefficient est significativement différent de zéro. Elle est définie par :

$$\Omega_{\hat{a}} = \begin{pmatrix} V(\hat{a}_0) & cov(\hat{a}_0, \hat{a}_1) & . & . & . & cov(\hat{a}_0, \hat{a}_p) \\ cov(\hat{a}_1, \hat{a}_{0)} & V(\hat{a}_0) & . & . & . & cov(\hat{a}_1, \hat{a}_p) \\ . & & & & & \\ . & & & & & \\ cov(\hat{a}_p, \hat{a}_{0)} & . & . & . & . & V(\hat{a}_p) \end{pmatrix}$$

La matrice est symétrique, sur la diagonale principale nous observons les variances des coefficients estimés.

Cette matrice est définie de la manière suivante :

$$\Omega_{\hat{a}} = E[(\hat{a} - a)(\hat{a} - a)'] \tag{1.32}$$

Or :

$$\begin{aligned} \hat{a} - a &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\epsilon \\ (\hat{a} - a)' &= \epsilon'\mathbf{X}[(\mathbf{X}'\mathbf{X})^{-1}]' \\ &= \epsilon'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \qquad \text{car } (\mathbf{X}'\mathbf{X})^{-1} \text{ est symétrique} \end{aligned}$$

Ainsi :

$$(\hat{a} - a)(\hat{a} - a)' = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\epsilon\epsilon'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \tag{1.33}$$

En passant à l'espérance mathématique, et sachant que les $\mathbf{X}$ sont non-stochastiques d'après l'hypothèse (H_1) :

$$E[(\hat{a} - a)(\hat{a} - a)'] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E[\epsilon\epsilon']\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \tag{1.34}$$

La quantité $E[\epsilon\epsilon']$, de dimension (n, n), représente la matrice de variance covariance des erreurs, en voici le détail :

$$E[\epsilon\epsilon'] = \begin{pmatrix} E(\epsilon_1^2) & E(\epsilon_1\epsilon_2) & . & . & . & E(\epsilon_1\epsilon_n) \\ . & & & . & . & \\ . & & & & & \\ . & & & & & \\ . & . & & . & . & E(\epsilon_n^2) \end{pmatrix}$$

Nous observons les variances des erreurs sur la diagonale principale, et les covariances sur les autres cases.

Or, par hypothèse (H_1) la variance de l'erreur est constante $V(\epsilon_i) = E(\epsilon_i^2) = \sigma_i^2$ et d'après l'hypothese (H_4) leurs covariances sont nulles $cov(\epsilon_i, \epsilon_i') = 0$.

De ce fait :

$$E[\epsilon\epsilon'] = \sigma_\epsilon^2\mathbf{I} \tag{1.35}$$

Où **I** est la matrice unité de dimension (n, n).

La matrice de variance covariance des estimateurs s'en retrouve grandement simplifiée. En effet :

$$\begin{aligned} E[(\hat{a} - a)(\hat{a} - a)'] &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E[\epsilon\epsilon']\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma_\epsilon^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'I\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma_\epsilon^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma_\epsilon^2(\mathbf{X}'\mathbf{X})^{-1} \end{aligned}$$

Nous trouvons ainsi la matrice de variance covariance des coefficients estimés :

$$\Omega_{\hat{a}} = \sigma_\epsilon^2(\mathbf{X}'\mathbf{X})^{-1} \tag{1.36}$$

1.2.2.4 Estimation de la variance du résidu

Pour la variance du résidu ($\sigma^2 = var[\epsilon]$), on peut utiliser l'estimateur sans biais construit à partir de la variance des résidus observés :

$$S^2 = \hat{\sigma}^2 = \frac{1}{n - p - 1}\sum_{i=1}^{n}\hat{\epsilon_i}^2 \tag{1.37}$$

Les $\hat{\epsilon}$ correspondent aux résidus observés :

$$\hat{\epsilon} = Y - \hat{Y} \tag{1.38}$$

On remarque deux choses par rapport à l'estimateur classique de :

$$S_{n-1}^2 = \hat{\sigma}^2 = \frac{1}{n - 1}\sum_{i=1}^{n}(y_i - \bar{y})^2 \tag{1.39}$$

- on n'inclut pas l'espérance des résidus, car celle-ci est supposée être zéro selon l'yposthese (H_2). Surtout, les résidus du modèle ont exactement une moyenne de zéro lorsqu'une constante est introduite dans le modèle.
- La somme des carrés est divisée par $(n - p - 1) = n - (p + 1)$ et non par $n - 1$. En fait,$n - p - 1$ correspond aux degrés de liberté du modèle (le nombre d'observations moins le nombre de coefficients à estimer). On remarque effectivement que :

$$E[\hat{\epsilon}'\hat{\epsilon}] = \sigma^2(n - p - 1) \tag{1.40}$$

1.2.2.5 Estimation de la matrice de variance-covariance

Il suffit de remplacer la variance théorique des résidus σ^2 par son estimateur sans biais des moindres carrés :

$$S^2 = \hat{\sigma}^2 = \frac{1}{n-p-1}\sum_{i=1}^{n}\hat{\epsilon_i}^2 \tag{1.41}$$

L'estimateur de la matrice de variance-covariance des résidus devient :

$$\widehat{var}[\hat{a}] = \hat{\sigma}^2(\mathbf{X}^t\mathbf{X})^{-1} \tag{1.42}$$

La variance estimée $\hat{\sigma}^2_{\hat{a}_j}$ de l'estimation du paramètre $\hat{a}_j$ est lue sur la diagonale principale de cette matrice.

1.2.2.6 Équation d'analyse de la variance et qualité d'un ajustement

Soient les deux relations suivantes :

$$a) \sum_{i=1}^{n} y_i = \sum_{i=1}^{n}\hat{y}_i \longrightarrow \bar{y} = \bar{\hat{y}} \tag{1.43}$$

$$b) \sum_{i=1}^{n}\epsilon_i = 0 \tag{1.44}$$

De a et b , nous en déduisons l'équation fondamentale d'analyse de la variance :

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2 + \sum_{i=1}^{n}\epsilon_i^2 \tag{1.45}$$

D'où :

$$SCT = SCE + SCR$$

tels que :

- **SCT :** représente la somme des carrés totale, elle signifie la quantité de variabilité dans les données originales ;
- **SCE :** représente la somme des carrés expliquée, elle signifie la quantité de variabilité dans le modéle ajusté aux données originales ;
- **SCR :** représente la somme des carrés résiduelles, elle signfie la quantité de variabilité non-expliquée par l'ajustement.

La variabilité totale (SCT) est égale à la variabilité expliquée (SCE) + la variabilité des résidus (SCR). Cette équation va nous permettre de juger la qualité de l'ajustement d'un modèle ; en effet, plus la variance expliquée est " proche " de la variance totale, meilleur est l'ajustement global du modèle. C'est pourquoi nous calculons le rapport SCE sur SCT :

$$R^2 = \frac{\sum_i (\hat{y}_i - \bar{\hat{y}})^2}{\sum_i (y_i - \bar{y})^2} = 1 - \frac{\sum_i \epsilon_i^2}{\sum_i (y_i - \bar{y})^2} \tag{1.46}$$

ou bien :

$$R^2 = \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT} \tag{1.47}$$

R^2 est appelé le coefficient de détermination, et R le coefficient de corrélation multiple. R^2 mesure la proportion de la variance de Y expliquée par la régression de Y sur X. Dans le cas de données centrées (moyenne nulle) et seulement dans ce cas, le coefficient de détermination est égal à :

$$R^2 = \frac{\hat{Y}'\hat{Y}}{Y'Y} = 1 - \frac{\epsilon'\epsilon}{Y'Y} \tag{1.48}$$

Cette qualité de l'ajustement et l'appréciation que l'on a de R^2 doivent être tempérées par le degré de liberté de l'estimation. En effet, lorsque le degré de liberté est faible, il convient de corriger le R^2 afin de tenir compte du relativement faible nombre d'observations comparé au nombre de facteurs explicatifs par le calcul d'un R^2 " corrigé " noté $\bar{R}^2$:

$$\bar{R}^2 = 1 - \frac{n-1}{n-p-1}(1 - R^2) \tag{1.49}$$

On a $\bar{R}^2 < R^2$ et si n est grand $\bar{R}^2 \simeq R^2$.

1.2.2.7 Les tests statistiques

– **Test de fisher**

Dans ce test, nous allons nous interroger sur la signification globale du modèle de régression, c'est-à-dire si l'ensemble des variables explicatives a une influence sur la variable à expliquer. Ce test peut être formulé de la manière suivante : existe-t-il au moins une variable explicative significative

Soit le test d'hypothèses :

$$H_0 : "a_1 = a_2 = ... = a_p = 0"$$

contre

$$H_1 : \text{" il existe au moins un des coefficients non null "}$$

Le cas où l'hypothèse H_0 est acceptée signifie qu'il n'existe aucune relation linéaire significative entre la variable à expliquer et les variables explicatives (ou encore que la Somme des Carrés Expliqués n'est pas significativement différente de 0).

Nous reprenons l'équation fondamentale d'analyse de la variance :

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2 + \sum_{i=1}^{n}\epsilon_i^2 \tag{1.50}$$

La régression est jugée significative si la variabilité expliquée est significativement différente de 0. Le tableau d'analyse de la variance permet d'effectuer le test de Fisher.

$$F^* = \frac{\frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}{p}}{\frac{\sum_{i=1}^{n}\epsilon_i^2}{(n-p-1)}} \tag{1.51}$$

Démonstration :

$$\begin{aligned}
\frac{\frac{SCE}{P}}{\frac{SCR}{n-p-1}} &= \frac{\frac{\frac{SCT}{SCT} - \frac{SCR}{SCT}}{P}}{\frac{\frac{SCT}{SCT} - \frac{SCE}{SCT}}{n-p-1}} \\
&= \frac{\frac{1-(1-R^2)}{p}}{\frac{1-R^2}{n-p-1}} \\
&= \frac{\frac{R^2}{P}}{\frac{1-R^2}{n-p-1}}
\end{aligned}$$

Alors :

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

Le tableau de l'analyse de la variance :

Source de variation	Somme des carrés	Degré de liberté	carrés moyens
$x_1, ..., x_p$	$SCE = \sum_i (\hat{y}_i - \bar{\hat{y}})^2$	p	$\frac{SCE}{p}$
Résidu	$SCR = \sum_t \epsilon_i^2$	$n-p-1$	$\frac{SCR}{n-p-1}$
Total	$SCT = \sum_t (y_i - \bar{y})^2$	$n-1$	$\frac{SCT}{n-1}$

L'hypothèse de normalité des erreurs implique que sous l'hypothèse H_0, F^* suit une loi de Fisher (rapport de deux chi-deux). Nous comparons donc ce F^* calculé au F théorique à p et $(n-p-1)$ degrés de liberté : si $F^* > F$ nous rejetons l'hypothèse H_0, le modèle est globalement explicatif.

– **Test de student**

Pour savoir la signification individuelle de chacune des variables explicatives (constante incluse),on utilise le test de Student sous l'hypothèse de normalité des erreurs, qui nous permettra de savoir si chaque paramètre est significativement différent de zéro ou non.

On test l'hypothése :

$$H_0 : "a_i = 0" \text{ contre } H_1 : "a_i \neq 0" \ (i = 0, ..., p)$$

Sous H_0 on calcule le ratio de student :

$$t^*_{\hat{a}_i} = |\frac{\hat{a}_i}{\hat{\sigma}_{\hat{a}_i}}| \longrightarrow t_{(n-(p+1))}$$

Où $\hat{\sigma}_{\hat{a}_i}$ est l'écart type estimé, obtenu à partir de la matrice des variances-covariances estimée :

$$\hat{\Omega}_{\hat{a}} = \hat{\sigma}^2_{\epsilon}(\mathbf{X}'\mathbf{X})^{-1}$$

Avec :

$$\hat{\sigma}^2_{\epsilon} = \frac{\hat{\epsilon}'\hat{\epsilon}}{n-p-1} \quad \text{et} \quad \hat{\epsilon}_i = y_i - \hat{y}_i.$$

On compare $t_{\hat{a}_i}$ à $t_{(\frac{\alpha}{2}, n-p-1)}$, lu sur la table de student.

– Si $t_{\hat{a}_i} > t_{(\frac{\alpha}{2}, n-p-1)}$, alors on rejette H_0.

– Si $t_{\hat{a}_i} < t_{(\frac{\alpha}{2}, n-p-1)}$, alors on accepte H_0.

1.2.2.8 Intervalles de confiances

L'intervalle de confiance permet de savoir s'il ya une relation entre la variable explicative X_{pi} et la variable expliquée Y , en tenant compte des autres variables explicatives qui sont dans le modèle.

Au niveau de confiance α, l'intervalle de confiance pour chaque coefficient de régression

a_i tel que $(i = 0, ..., p)$ est donnée par :

$$IC_{a_i} = [\hat{a}_i - \hat{\sigma}_{\hat{a}_i} t_{(\frac{\alpha}{2}, n-p-1)}; \hat{a}_i + \hat{\sigma}_{\hat{a}_i} t_{(\frac{\alpha}{2}, n-p-1)}] \tag{1.52}$$

Avec :

- $t_{\frac{\alpha}{2}}$ est le quantile de la distribution t.
- $\hat{\sigma}_{\hat{a}_i}$ est l'écart type estimé.

– Si $l'IC_{a_i}$ exclut la valeur zéro, alors on déduit qu'il ya une association linéaire significative entre Y et X_{pi}.

– Si $l'IC_{a_i}$ inclut la valeur zéro, alors on déduit que la variable X_i n'apporte aucune information supplémentaire pour la prédiction de Y.

1.2.2.9 Exemple de régression linéaire multiple

– **Les données :**

Prenant l'exemple suivant à deux variables :

x_{i1}	x_{i2}	y_i
9	52	38
13	70	57
22	78	43
11	83	84
18	68	17

TABLE 1.4: *Tableau des données*

– **Estimation :**

$\bar{x_1} = 14.6$, $\bar{x_2} = 70.2$ et $\bar{y} = 47.8$

$SCE_1 = 113.2$, $SCE_2 = 560.8$

$SPE_{12} = 106.4$, $SPE_{1y} = -230.4$,$SPE_{2y} = 670.2$

$$b_0 = \bar{y} - b_1\bar{x}_1 - b_2\bar{x}_2 = -31.2$$

;

$$b_1 = \frac{SCE_2 SPE1y - SPE_{12} SPE_{2y}}{SCE_1 SCE_2 - SPE_{12}^2} = -3.84$$

$$b_2 = \frac{SCE_1 SPE2y - SPE_{12} SPE_{1y}}{SCE_1 SCE_2 - SPE_{12}^2} = 1.92$$

– **Le modèle estimé est :**

$$\hat{y} = -31.2 - 3.84x_1 + 1.92x_2$$

1.3 Modèles de dénombrement

Les modèles décrits dans cette partie s'intéressent plus particulièrement à la description ou l'explication d'observations constituées d'effectifs comme, par exemple, le nombre de succès d'une variable de Bernouilli lors d'une séquence d'essais ou encore le nombre d'individus qui prennent une combinaison donnée de modalités de variables qualitatives ou niveaux de facteurs. Contrairement aux modèles précédents "modèle linéaire" basés sur l'hypothèse de normalité des observations, les lois concernées sont maintenant discrètes et associées à des dénombrements : loi de Poisson, binomiale, multinomiale. Néanmoins, tous les modèles considérés dans cette partie appartiennent à la famille des modèles linéaires généralisés.

1.3.1 Modèle de Régression logistique

Historiquement, la régression logistique ou régression binomiale fut la première méthode utilisée, notamment en marketing pour le scoring et en épidémiologie, pour aborder la modélisation d'une variable binaire binomiale (nombre de succès pour n_i essais) ou de Bernoulli (avec $n_i = 1$) : possession on non d'un produit, bon ou mauvais client, décès ou survie d'un patient, absence ou présence d'une pathologie... . Bien connue dans ces types d'applications et largement répandue, la régression logistique conduit à des interprétations pouvant être complexes mais rentrées dans les usages pour quantifier, par exemple, des facteurs de risque liés à une pathologie, une faillite... . Cette méthode reste donc la plus utilisée même si, en terme de qualité prévisionnelle, d'autres approches sont susceptibles, en fonction des données étudiées, d'apporter de bien meilleurs résultats. Il est donc important de bien maîtriser les différents aspects de la régression logistique dont l'interprétation des paramètres, la sélection de modèle par sélection de variables ou par régularisation. Cas particulier de modèle linéaire général, la régression logistique reprend

la plupart des usages des méthodes de cette famille : estimation par maximisation de la vraisemblance, statistiques de test suivant asymptotiquement des lois du chi-deux, calcul des résidus, observations influentes, critère pénalisé (AIC) d'Akaïke pour la sélection de modèle. Néanmoins, certaines spécificités méritent d'être soulignées pour un meilleur usage de même qu'il est important de rappeler que d'autres méthodes peuvent conduire à de meilleure prévision, donc de meilleurs scores et que c'est souvent un bon investissement que de faire évouer ses habitudes[11].

1.3.1.1 Représentation du modèle

Ce modèle de regression logistique décrit la modélisation d'une variable qualitative Z à 2 modalités : 1 ou 0, succès ou échec. Les modèles de régression précédents adaptés à l'explication d'une variable quantitative ne s'appliquent plus directement car le régresseur linéaire usuel Xβ ne prend pas des valeurs simplement binaires. L'objectif est adapté à cette situation en cherchant à expliquer les probabilités[20] :

$$\pi = p(Z = 1) \text{ou} 1 - \pi = p(Z = 0) \tag{1.53}$$

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

$$g(\pi_i) = x_i'\beta \tag{1.54}$$

Il existe de nombreuses fonctions, dont le graphe présente une forme sigmoïdale et qui sont candidates pour remplir ce rôle, trois sont pratiquement disponibles dans les logiciels :

- probit : g est alors la fonction inverse de la fonction de répartition d'une loi normale, mais son expression n'est pas explicite.
- log-log : avec g définie par :

  $$g(\pi) = \ln[-\ln(1 - \pi)] \tag{1.55}$$

  mais cette fonction est dissymétrique.
- logit est définie par :

  $$g(\pi) = logit(\pi) = \ln \frac{\pi}{1 - \pi}; avec : g^{-1}(x) = \frac{e^x}{1 - e^x} \tag{1.56}$$

Plusieurs raisons, tant théoriques que pratiques, font préférer cette dernière solution. Le rapport $\pi/(1-\pi)$ qui exprime une "cote", appelé l'odds et la régression logistique s'interprète donc comme la recherche d'une modélisation linéaire du "log odds" tandis que les coefficients de certains modèles expriment des "odds ratio" c'est-'a-dire l'influence d'un facteur qualitatif sur le risque (ou la chance) d'un échec (d'un succès) de Z.
On se limite à la description de l'usage élémentaire de la régression logistique. Des compléments concernant l'explication d'une variable qualitative ordinale (plusieurs modalités), l'intervention de variables explicatives avec effet aléatoire.

1.3.1.2 Formulation du modèle

- Modèle logistique à une variable explicative X :

$$Y = \pi(X) + \epsilon = \frac{\exp^{\beta_0+\beta_1 X}}{1+\exp^{\beta_0+\beta_1 X}} + \epsilon \tag{1.57}$$

  Remarque :
  - On a

$$\epsilon = 1 - \pi(X) \text{ Si } Y = 1$$

  - Ou

$$\epsilon = -\pi(X) \text{ Si } Y = 0$$

- Modèle logistique multivarié :

$$Y = P(Y = 1 \mid X_j) + \epsilon = \pi(X_j) + \epsilon = \frac{\exp^{\beta_0+\sum_{j=1}^{p}\beta_j X_j}}{1+\exp^{\beta_0+\sum_{j=1}^{p}\beta_j X_j}} + \epsilon \tag{1.58}$$

- Transformation LOGIT :

$$Logit[\pi(X)] = \ln(\frac{\pi(X)}{1-\pi(X)}) = \beta_0 + \beta_1 X \tag{1.59}$$

- Intérêts :
  - Permet de revenir à un modèle linéaire classique
  - Interprétation des coefficients du modèle comme une mesure d'association de X par rapport à Y (Notion d'odds-ratio)

1.3.1.3 Estimation des coefficients

En régression logistique, la MMC "Méthode des moindres carrés ordinaires" ne permet pas d'obtenir une estimation des coefficients. On utilise alors la Méthode du maximum de vraisemblance (Maximum likelihood),qui est une Méthode classique permettant d'estimer les paramètres d'une loi, d'un modèle.

- **Maximum de vraisemblance** :
  - Estimateurs des paramètres sans biais et de faible variance.
  - n variables aléatoires Y_i i.i.d qui suivent une loi de Bernouli de paramètre β B(β).
  - La vraisemblance d'un n-échantillon $y_1, y_2, \ldots, y_n$ est définie comme la probabilité d'observer cet échantillon :

$$P(Y_i = y_i) = \beta_i^{y_i}.(1-\beta_i)^{1-y_i} \tag{1.60}$$

  - Les variables Y_i étant indépendantes on aura :

$$L(\beta, y_1..., y_n) = \prod_{i=1}^{n} \beta^{y_i}.(1-\beta_i)^{1-y_i} \tag{1.61}$$

    On aura donc selon le modèle logistique pour une seule variable explicative X quantitative :

$$L = \prod_{i=1}^{n} (\frac{\exp^{\beta_0+\beta_1 X}}{1+\exp^{\beta_0+\beta_1 X}})^{y_i} (1 - \frac{\exp^{\beta_0+\beta_1 X}}{1+\exp^{\beta_0+\beta_1 X}})^{1-y_i} \tag{1.62}$$

  - trouver les paramètres $\hat{\beta}_0$ et$\hat{\beta}_1$ qui maximisent la probabilité d'observer l'échantillon "i.e. maximisation de la vraisemblance"

$$\max_{\beta_0,\beta_1}(L) = \max_{\beta_0,\beta_1}(\prod_{i=1}^{n} (\frac{\exp^{\beta_0+\beta_1 X}}{1+\exp^{\beta_0+\beta_1 X}})^{y_i} (1 - \frac{\exp^{\beta_0+\beta_1 X}}{1+\exp^{\beta_0+\beta_1 X}})^{1-y_i}) \tag{1.63}$$

    Pour des raisons de simplicité de calcul, on passe généralement par le log :

$$\max_{\beta_0,\beta_1}(L) = \max_{\beta_0,\beta_1}(\log L) \tag{1.64}$$

  - En utilisant la Méthode de Newton-Raphson (analyse numérique) pour calculer $\hat{\beta}_0, \hat{\beta}_1$ on a recours aux dérivées partielles :

$$\frac{\partial L}{\partial \beta_0} = 0 \text{et} \frac{\partial L}{\partial \beta_1} = 0$$

– l'intérêt est de pouvoir calculer pour tout i :

$$P(y_i = 1 \mid_{X=x_i}) = \frac{\exp^{\hat{\beta}_0+\hat{\beta}_1 X_i}}{1+\exp^{\hat{\beta}_0+\hat{\beta}_1 X_i}} \tag{1.65}$$

– **Variance et écart-type** :

En posant $\hat{\beta} = (\hat{\beta}_0, ..., \hat{\beta}_1)^t$, On a :

$$\hat{V}(\hat{\beta}) = \left(-\frac{\partial^2}{\partial\beta^2} lnL(\beta, Y)\right)^{-1} = (X^t W X)^{-1}, \tag{1.66}$$

Où X=

$$\begin{pmatrix} 1 & x_{1,1} & .. & x_{p,1} \\ 1 & x_{1,2} & .. & x_{p,2} \\ : & : & : & : \\ 1 & x_{1,n} & .. & x_{p,n} \end{pmatrix}$$

Et

$$W = diag(\hat{p}(x_1)(1-\hat{p}(x_1)), ..., \hat{p}(x_n)(1-\hat{p}(x_n))). \tag{1.67}$$

$$\sigma(\hat{\beta}_j) = \sqrt{\hat{V}(\hat{\beta}_j)}, \text{racine carrée de la j + 1-ème composante du vecteur} \hat{V}(\hat{\beta}_j) \tag{1.68}$$

1.3.1.4 Modèle binomial

On considère, pour $i = 1, ...I$ différentes valeurs fixées $x_i^1, ..., x_i^p$ des variables explicatives $X^1, ..., X^p$. Ces dernières pouvant être des variables quantitatives ou encore des variables qualitatives, c'est-à-dire des facteurs issus d'une planification expérimentale. Pour chaque groupe, c'est-à-dire pour chacune des combinaisons de valeurs ou facteurs, on réalise n_i observations telle que :

$$n = \sum_i^I n_i \tag{1.69}$$

de la variable Z qui se mettent sous la forme :

$$\frac{y_1}{n_1}, ..., \frac{y_I}{n_I} \tag{1.70}$$

où y_i désigne le nombre de "succès" observés lors des n_i essais. On suppose que toutes les observations sont indépendantes et qu'à l'intérieur d'un même groupe, la probabilité π_i

de succès est constante. Alors, la variable y_i sachant n_i et d'espérance $E(y_i) = n_i\pi_i$ suit une loi binomiale $B(n_i, \pi_i)$ dont la fonction de densité s'écrit :

$$P(Y = y_i) = \begin{pmatrix} n_i \\ y_i \end{pmatrix} \pi_i^{y_i}(1 - \pi_i)^{(n_i - y_i)} \tag{1.71}$$

On suppose que le vecteur des fonctions logit des probabilités π_i appartient au sous-espace vect $\{X^1, ..., X^q\}$ engendré par les variables explicatives :

$$logit(\pi_i) = x_i'\beta, i = 1, ..., I \tag{1.72}$$

ce qui s'écrit encore :

$$\pi_i = \frac{\exp^{x_i'\beta}}{1 + \exp^{x_i'\beta}}, i = 1, ..., I \tag{1.73}$$

Le vecteur des paramètres est estimé par maximisation de la log-vraisemblance. Il n'y a pas de solution analytique, celle-ci est obtenue par des méthodes numériques itératives (par exemple Newton Raphson) dont certaines reviennent à itérer des estimations de modèles de régression par moindres carrés généralisés avec des poids et des métriques adaptés à chaque itération.

L'optimisation fournit une estimation b de β,il est alors facile d'en déduire les estimations ou prévisions des probabilités π_i :

$$\hat{\pi}_i = \frac{\exp^{x_i'b}}{1 + \exp^{x_i'b}}, i = 1, ..., I \tag{1.74}$$

et ainsi celles des effectifs :

$$\hat{y}_i = n_i\hat{\pi}_i. \tag{1.75}$$

1.3.1.5 Significativité de la régression

– **Test de Wald :**

Soit $j\epsilon\{0, ..., p\}$.Le test de Wald permet d'évaluer l'influence de X_j sur Y.

On considère les hypothèses :

$$H_0 : \beta_j = 0 \text{ contre } H_1 : \beta_j \neq 0.$$

On calcule la réalisation z_{obs} de

$$Z_* = \frac{\hat{\beta}_j}{\hat{\sigma}(\hat{\beta}_j)}. \tag{1.76}$$

On considère une v.a $Z \sim N(0,1)$.

Alors la p-valeurs associée est :

$$p - valeur = p(\mid Z \mid \geq \mid Z_{obs} \mid).$$

Dans ces tests de significativité de la régression un code est attribué à chaque résultat pour évaluer le degré de significativité :

Le code *, signifie que l'influence de X_j sur Y est significative, le code **, signifie qu'elle est très significative et le code *** signifie qu'elle est hautement significative.

– **Déviance :**

On appelle déviance la réalisation D de :

$$\hat{D} = 2\sum_{i=1}^{n}(Y_i \ln(\frac{Y_i}{\hat{p}(x_i)}) + (1-y_i)\ln(\frac{1-y_i}{1-\hat{p}(x_i)})) \tag{1.77}$$

Avec $x_i = (1, x_{1,i}, ..., x_{p,i})$

C'est 2 fois la différence entre les log-vraisemblances évaluées en Y_i et $\hat{p}(x_i)$.

– **Loi de $\hat{D}$:**

Si le modèle est bien adapté au problème, la loi limit de $\hat{D}$ est $X^2(n-(p+1))$

– **Test de la déviance :**

Soit $j\epsilon\{0, ..., p\}$. Le test de la déviance vise à évaluer l'influence (ou la contribution) de X_j sur Y .

La p-valeur associée utilise la loi du Chi-deux :

Le code *, signifie que l'influence de X_j sur Y est significative, le code **, signifie qu'elle est très significative et le code *** signifie qu'elle est hautement significative.

– **Test lr :**

On considère les hypothèses :

$$H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0$$

Contre

$$H_1 : \text{il y a au moins un coefficient non nul}$$

.

Pour ce faire, on utilise le test du rapport de vraisemblance (lr pour likelihood ratio) (asymptotique). La p-valeur associée utilise la loi du Chi-deux :

Le code *, signifie que l'influence de X_j sur Y est significative, le code **, signifie qu'elle est très significative et le code *** signifie qu'elle est hautement significative.

1.3.1.6 Intervalle de confiance

Intervalle de confiance pour β_j : soit $j\epsilon\{0, ..., p\}$ Un intervalle de confiance pour β_j au seuil α ,$\alpha\epsilon[0, 1]$, est la réalisation i_{β_j} de :

$$I_{\beta_j} = [\hat{\beta}_j - z_\alpha \tilde{\hat{\sigma}}(\hat{\beta}_j), \hat{\beta}_j + z_\alpha \tilde{\hat{\sigma}}(\hat{\beta}_j)]$$

Où z_α est le réel vérifiant $P(\mid Z \mid \geq z_\alpha) = \alpha$, avec $Z \sim N(0, 1)$.

Intervalle de confiance pour p(x) : Un intervalle de confiance pour p(x), avec $x = (x_1, ..., x_p)$, au niveau $100(1 - \alpha)\%$,$\alpha\epsilon]0, 1[$, est la réalisation $i_{p(x)}$ de :

$$I_{p(x)} = \left[logit^{-1}(logit\hat{p}(x) - z_\alpha \hat{\sigma}(logit\hat{p}(x))), logit^{-1}(logit\hat{p}(x) + z_\alpha \hat{\sigma}(logit\hat{p}(x)))\right],$$

où z_α est le réel vérifiant$P(\mid Z \mid \geq z_\alpha) = \alpha$, avec $Z \sim N(0, 1)$.

1.3.1.7 Exemple de régression logistique

– **les données :**

l'étude a était réalisée sur un échantillon de 65 malade vérifiant si un patient est atteint d'une maladie cardiaque **y** La variable à expliquer binaire : "Présence/absence de la maladie cardiaque" ;

X La variable explicative quantitative : "L'âge du patient".

Les résultats obtenus sont données dans le graphe suivant :

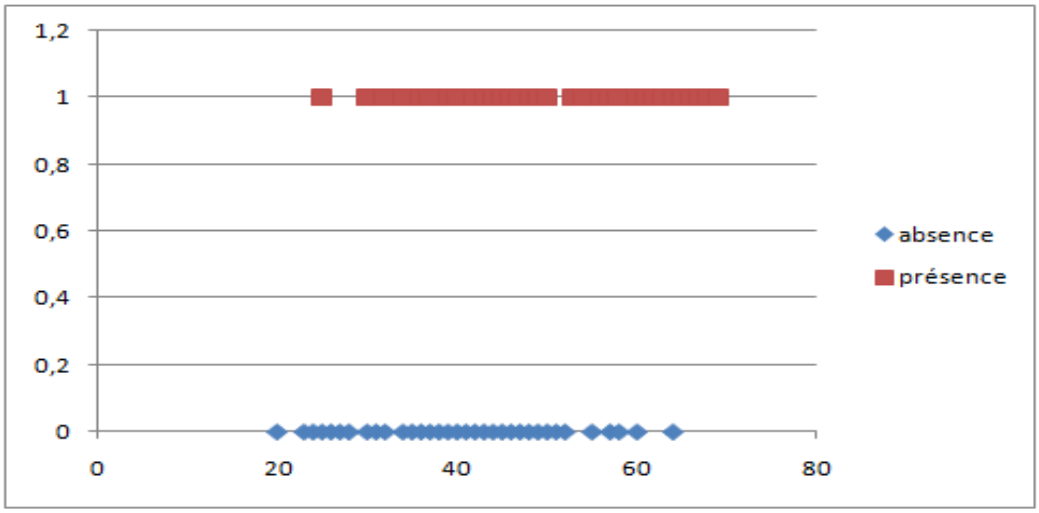

FIGURE 1.3: *1.Présence ou absence de la maladie cardiaque*

Nous constatons que le graphe donné ci-dessous ne montre pas clairement l'existence d'une liaison entre Y et X.

Par contre si l'on utilise la variable âge découpée en classes et la proportion de malades par classe, la liaison entre Y et X apparaît plus clairement sous la forme d'une courbe en S :

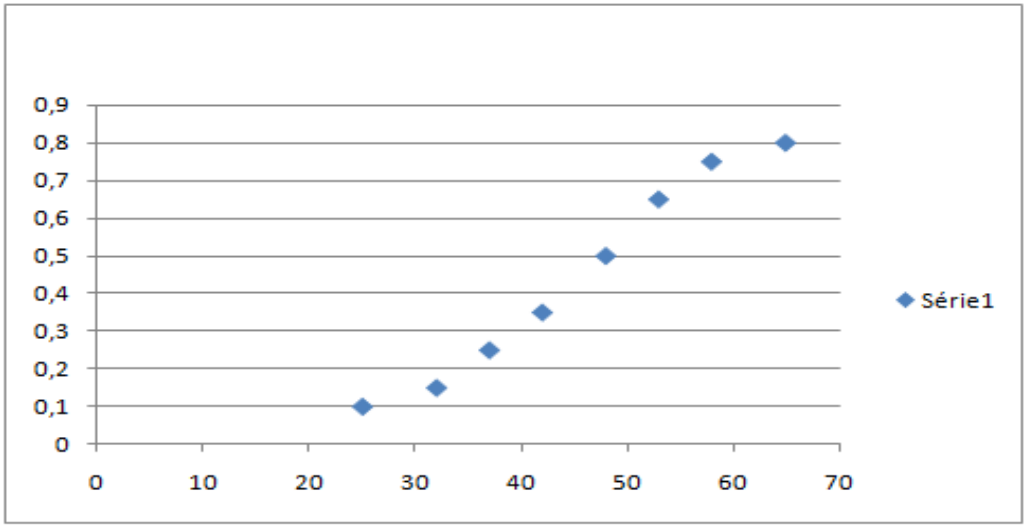

FIGURE 1.4: *2.La liaison entre X et Y*

– **Application du modèle logistique**

L'objectif est de modéliser $\Pi(x) = prob(Y = 1 \mid X = x)$

On a :

$$\Pi(x) = \frac{\exp^{\beta_0+\beta_1 x}}{1+\exp^{\beta_0+\beta_1 x}}$$

$$\Pi^{'}(x) = \beta_1 \Pi(x)(1-\Pi(x))$$

Nous avons choisit la fonction logistique car elle est adaptée à la modélisation de probabilités en prenant ses valeurs entre 0 et 1 selon une courbe en S.

Son utilisation est par exemple indiquée lors de la modélisation du risque individuel de développer une maladie dans les études épidémiologiques.

En effet, en considérant que la variable x représente un indice résultant de la combinaison de plusieurs facteurs de risque, on peut interpréter F(x) comme le risque d'être atteint de cette maladie.

Dans ce contexte le risque est minimal pour de faibles valeurs de x.

Il augmente pour les valeurs intermédiaires de x.

Il apparaît proche de 1 pour des valeurs plus élevées de x.

– **Estimation des paramètres du modèle logistique :**

$$\Pi(x_i) = \Pi_i = p(Y = 1|X = x_i) = \frac{\exp^{\beta_0+\beta_1 x_i}}{1+\exp^{\beta_0+\beta_1 x_i}}$$

Vraisemblance

$$L(\beta) = \prod_{i=1}^{n} \Pi(x_i)^{y_i}(1-\Pi(x_i))^{1-y_i}$$

Log de la vraisemblance

$$LogL(\beta) = \sum_{i=1}^{n} y_i Log\Pi_i(x) + (1-y_i)Log(1-\Pi_i(x))$$

Maximum de vraisemblance

On obtient $\hat{\beta}$ en annulant $\frac{\partial LogL(\beta)}{\partial \beta}$

On aura :

$$\hat{\beta}_0 = -5.3095$$

$$\hat{\beta}_1 = 0.1109$$

Et

$$-2LogL = 107.35\text{est minimum}$$

– **Le modèle estimé :**

$$\hat{\Pi}(x) = \frac{\exp^{-5.3095+0.1109x}}{1+\exp^{-5.3095+0.1109x}}$$

On conclu que la probabilité d'être atteint de la maladie augmente avec l'âge.

1.3.2 Modèle log-linéaire ou poissonien

Le modèle log-linéaire ou modèle poissonien s'intéresse plus particulièrement à la description ou l'explication d'observations constituées d'effectifs ; nombre de succès d'une variable de Bernouilli lors d'une séquence d'essais dans la cas précédent de la régression logistique, nombre d'individus qui prennent une combinaison donnée de modalités de variables qualitatives ou niveaux de facteurs, dans le cas présent. Ce modèle fait également partie de la famille du modèle linéaire général en étant associé à une loi de Poisson. Il est également appelé aussi modèle log-linéaire et s'applique principalement à la modélisation d'une table de contingence complète[12].

1.3.2.1 Représentation du modèle

Le modèle est associé à une loi de poisson décrit comme suit :

– Soit λ ,un réel et Y une variable aléatoire réelle, $Y \sim P(\lambda)$ si et seulement si$\forall\ k\epsilon N$:

$$P(Y=k) = \exp^{-\lambda}\frac{\lambda^k}{k!} \tag{1.78}$$

En conséquence, on aura :

$$E(Y) = Var(Y) = \lambda \tag{1.79}$$

Le modèle de régression de Poisson est donné par :

$$\ln(y) = \beta_o + \beta_1 x_1 + ... + \beta_i x_i + ... + \beta_k x_k \tag{1.80}$$

– y est la réalisation de Y variable endogène suivant une loi de poisson ;

– β_o ordonnée à l'origine ;

– β_i coefficient associé à la i^{me} variable explicative x_i

Les données se présentent généralement sous la forme d'une table de contingence obtenue par le croisement de plusieurs variables qualitatives et dont chaque cellule contient un effectif ou une fréquence à modéliser. Nous nous limiterons à l'étude d'une table élémentaire

en laissant de côté des structures plus complexes, par exemple lorsque des zéros structurels, des indépendances conditionnelles, des propriétés de symétrie ou quasi-symétrie, une table creuse, sont à prendre en compte. D'autre part, sous sa forme la plus générale, le modèle peut intégrer également des variables quantitatives. Ce type de situation se retrouve en analyse des correspondances simple ou multiple mais ici, l'objectif est d'expliquer ou de modéliser les effectifs en fonction des modalités prises par les variables qualitatives.

1.3.2.2 Distributions

On considère la table de contingence complète constituée à partir de l'observation des variables qualitatives $X^1, X^2, ..., X^p$ sur un échantillon de n individus. Les effectifs $\{y_{jk...l}$;j=1...J ;k=1...K ;l=1...L$\}$ de chaque cellule sont rangés dans un vecteur y à I composantes (I=J*K*L).

L'hypothèses sur la distribution est considérées en fonction du contexte expérimental :

- **Poisson** :

  Le modèle le plus simple consiste à supposer que les variables observées Y_i suivent des lois de Poisson indépendantes de paramètre $\lambda_i = E(Y_i)$.La distribution conjointe admet alors pour densité :

$$f(y, \lambda) = \prod_{i=1}^{I} \frac{\lambda_i^{y_i} \exp^{-\lambda_i}}{y_i!} \tag{1.81}$$

1.3.2.3 Estimation du modèle

Nous avons :

$$\ln[E(y)] = \ln(\lambda) = \beta_o + \beta_1 x_1 + ... + \beta_i x_i + ... + \beta_k x_k \tag{1.82}$$

Le but est d'estimer β_0 et le vecteur β des coefficients β_i en utilisant la méthode du maximum de vraisemblance :

- Fonction de vraisemblance :

$$L = \prod_{i=1}^{n} P(Y_i = k_i) = \prod_{i=1}^{n} \exp^{-\lambda_i} \frac{\lambda_i^{k_i}}{k_i!} \tag{1.83}$$

avec n le nombre d'observations ;

$\lambda_i = \exp^{\beta_0+\beta x_i}$;

x_i=$\{x_{i1}, .., x_{ij}\}$ et β=$\{\beta_1, ..., \beta_j\}$

– Logarithme de la Vraisemblance :

$$\ln(L) = \sum_{i=1}^{n} [k_i \ln(k_i) - \lambda_i] - cte \tag{1.84}$$

– Maximisation grâce à la dérivée :

$$\begin{aligned} s(\beta_0, \beta) &= (\frac{\partial \ln(L)}{\partial \beta_0}, \frac{\partial \ln(L)}{\partial \beta}) \\ &= (\sum_{i=1}^{n} (y_i - \lambda_i), \sum_{i=1}^{n} x_i (y_i - \lambda_i) \end{aligned}$$

– Variance et écart-type :En posant $\hat{\beta} = (\hat{\beta}_0, ..., \hat{\beta}_p)^t$, On a :

$$\hat{V}(\hat{\beta}) = ((-\frac{\partial^2}{\partial \beta^2} lnL(\beta, Y))^{-1} = (X^t W X)^{-1},$$

Où X=

$$\begin{pmatrix} 1 & x_{1,1} & .. & x_{p,1} \\ 1 & x_{1,2} & .. & x_{p,2} \\ : & : & : & : \\ 1 & x_{1,n} & .. & x_{p,n} \end{pmatrix}$$

Et $W = diag(\hat{\lambda}(x_1), ..., \hat{\lambda}(x_n))$.

$$\hat{\sigma}^2 = \sqrt{\hat{V}(\hat{\beta})}$$

1.3.2.4 Significativité de la régression

– **Test de Wald :** Soit $j \epsilon \{0, ..., p\}$. L'objectif du test de Wald est d'évaluer l'influence (ou la contribution) de X_j sur Y .

On considère les hypothèses :

$$H_0 : \beta_j = 0 \text{ contre } H_1 : \beta_j \neq 0.$$

On calcule la réalisation z_{obs} de

$$Z_* = \frac{\hat{\beta}_j}{\hat{\sigma}(\hat{\beta}_j)}. \tag{1.85}$$

On considère une v.a $Z \sim N(0,1)$.

Alors la p-valeurs associée est :

$$p - valeur = p(\mid Z \mid \geq \mid Z_{obs} \mid).$$

Dans ces tests de significativité de la régression un code est attribué à chaque résultat pour évaluer le degré de significativité :

Le code *, signifie que l'influence de X_j sur Y est significative, le code **, signifie qu'elle est très significative et le code *** signifie qu'elle est hautement significative.

– **Déviance :**

On appelle déviance la réalisation D de :

$$\hat{D} = 2\sum_{i=1}^{n}(Y_i \ln(\frac{Y_i}{\hat{\lambda}(x_i)}) + (1-y_i)\ln(\frac{1-y_i}{1-\hat{\lambda}(x_i)})) \tag{1.86}$$

Avec $x_i = (1, x_{1,i}, ..., x_{p,i})$

C'est 2 fois la différence entre les log-vraisemblances évaluées en Y_i et $\hat{\lambda}(x_i)$.

– **Loi de $\hat{D}$:**

Si le modèle est bien adapté au problème (ou exact), la loi limite (ou exacte) de $\hat{D}$ est $X^2(n-(p+1))$

– **Test de la déviance :**

Soit $j \epsilon \{0, ..., p\}$. L'objectif du test de la déviance est d'évaluer l'influence (ou la contribution) de X_j sur Y.

La p-valeur associée utilise la loi du Chi-deux :

Le code *, signifie que l'influence de X_j sur Y est significative, le code **, signifie qu'elle est très significative et le code *** signifie qu'elle est hautement significative.

– **Test lr :**

On considère les hypothèses :

$$H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0$$

Contre

$$H_1 : \text{il y a au moins un coefficient non nul}$$

.

Pour ce faire, on utilise le test du rapport de vraisemblance (lr pour likelihood ratio) (asymptotique). La p-valeur associée utilise la loi du Chi-deux :

Le code *, signifie que l'influence de X_j sur Y est significative, le code **, signifie qu'elle est très significative et le code *** signifie qu'elle est hautement significative.

1.3.2.5 Intervalle de confiance

Intervalle de confiance pour β_j : soit $j\epsilon 0, ..., p$ Un intervalle de confiance pour β_j au seuil α ,$\alpha\epsilon[0, 1]$, est la réalisation i_{β_j} de :

$$I_{\beta_j} = [\hat{\beta}_j - z_\alpha \tilde{\hat{\sigma}}(\hat{\beta}_j), \hat{\beta}_j + z_\alpha \tilde{\hat{\sigma}}(\hat{\beta}_j)]$$

Où z_α est le réel vérifiant $P(\mid Z \mid \geq z_\alpha) = \alpha$, avec $Z \sim N(0, 1)$.

Intervalle de confiance pour $\lambda(x)$: Un intervalle de confiance pour p(x), avec $x = (x_1, ..., x_p)$, au seuil α ,$\alpha\epsilon]0, 1[$, est la réalisation $i_{\lambda(x)}$ de :

$$I_{\lambda(x)} = \left[\exp(\ln(\hat{\lambda}(x)) - z_\alpha \hat{\sigma}(\ln(\hat{\lambda}(x)))), \exp(\ln(\hat{\lambda}(x)) + z_\alpha \hat{\sigma}(\ln(\hat{\lambda}(x))))\right],$$

où z_α est le réel vérifiant$P(\mid Z \mid \geq z_\alpha) = \alpha$, avec $Z \sim N(0, 1)$.

1.3.2.6 Exemple de régression de poisson

La régression de Poisson permet de modéliser des comptages distribués selon une loi de Poisson en fonction de variables explicatives quantitatives ou qualitatives.
Nous représentons un exemple sur le risque de Mélanome présenté dans Tenenhaus (1993)

- **Les données :**

  Y :Comptage ; $X_1 ... X_k$ Variables explicatives.

  Le tableau suivants résume les données : où :

  N : La région du nord ; S : La région du sud ; Population : Nombre estimé de personnes soumises au risque.

- **Le modèle**

$$Y_i \sim Poisson(\lambda_i)$$

Age	**Région**	**Y=mélanome**	**Population**
<35	N	61	2880262
35-44	N	76	564535
45-54	N	98	592983
55-64	N	104	450740
65-74	N	63	270908
>74	N	80	161850
<35	S	64	1074246
35-44	S	75	220407
45-54	S	68	198119
55-64	S	63	134084
65-74	S	47	70708
>74	S	27	34233

TABLE 1.5: *Tableau des données*

$$Log(\lambda_i) = \beta_0 + \beta_1 X_{1i} + ... + \beta_p X_{pi}$$

– **Vraissemblance**

$$L = \prod_{i=1}^{12} \frac{exp^{\lambda_i} \lambda_i^{y_i}}{y_i!}$$

Où i est l'indice de la i^{ime} population.

– **Estimation**

On estime les β_j en maximisant la vraissemblance.

Le modèle estimé est donné par :

$$\begin{aligned} \ln(\lambda_i) &= -6.89 - 2.94(age < 35) - 1.15(age35 - 44) - 1.03(age45 - 54) \\ &- 0.70(age55 - 64) - 0.58(age65 - 74) - 0.82Nord \end{aligned}$$

Chapitre 2

Modèles linéaires généralisés (estimations et prédictions)

2.1 Introduction

En statistique, le modèle linéaire généralisé (GLM) est une généralisation de la regression linéaire. Il permet d'étudier la liaison entre une variable réponse et un ensemble de variables explicatives ou prédicateurs $(x_1, x_2, ...x_n)$.
Les modèles linéaires généralisés ont été formules par Nelder et Wedderburn en 1972 et ont été exposés de façon complète par Mc Cullagh et Nelder en 1989 comme un moyen d'unifier les modèles statistiques y compris la regression linéaire, la regression logistique , la Regression log linéaire. Ils proposent une méthode iterative dénommée méthode des moindres carres ré-pondérés itérativement pour l'estimation du maximum de vraisemblance des paramètres du modèle. L'estimation du maximum de vraisemblance reste populaire et est la méthode par défaut dans de nombreux logiciels de calculs statistiques [2].

2.2 Famille exponentielle

En probabilités et statistique, une famille exponentielle est une classe de distributions de probabilité qui ont des propriétés algébriques intéressantes. Souvent, elles sont les seules

à présenter ces propriétés que nous présenterons dans ce qui suit. Les familles exponentielles apparaissent de façon naturelle dans la recherche de distributions lors d'applications statistiques, en particulier dans les méthodes bayesiennes. La famille exponentielle comprend quantité de distributions parmi les plus courantes : Normale, exponentielle, gamma, chi-carré, bêta, Dirichlet, Bernoulli, Bernoulli multinomiale, Poisson, Wishart, Wishart Inverse, etc. D'autres distributions courantes ne forment une famille exponentielle que si certains paramètres sont fixes et de valeur connue, telles la binomiale (nombre de tirages fixés), multinomiale (idem) et binomiale négative (nombre d'échecs fixés). Parmi les distributions d'usage courant qui ne sont pas de famille exponentielle, on peut citer la t de Student, la plupart des mixtures, ainsi que la famille des distributions uniformes de bornes non fixées. La famille exponentielle est à la base des fonctions de distribution utilisées dans le Modèle linéaire généralisé, qui comprend la plupart des modèles de régression en statistique et en économétrie[**?**].

2.2.1 Définition

Un modèle statistique sur un espace des observations E est dit famille exponentielle générale s'il existe un entier p, des fonctions η, T, C et h tels que les densités puisse s'écrire, pour tout θ de ϑ, sous la forme :

$$f_\theta(x) = \exp^{(\eta(\theta),T(x))} C(\theta)h(x) \tag{2.1}$$

Avec les contraintes que :

- T soit une fonction mesurable à valeurs dans R^p ;
- η soit une fonction à valeurs dans R^p ;
- C soit une fonction réelle positive qui ne dépend pas de x ;
- h soit une fonction borélienne positive qui ne dépend pas de θ

Le vecteur aléatoire T(X) est appelé statistique canonique du modèle. Si la fonction T est l'identité, la famille exponentielle est dite naturelle.

On parle de forme canonique d'une famille exponentielle générale quand les densités de probabilités ont la forme :

$$f_\theta(x) = \exp^{(\theta,T(x))} C(\theta)h(x) \tag{2.2}$$

pour tout θ de ϑ, ce qu'il est toujours possible d'obtenir quitte à reparamétriser la famille par $\theta' = \eta(\theta)$. Dans ce cas le paramètre θ de la famille exponentielle est appelé paramètre canonique.

2.2.2 propriétés algébriques

- La caractérisation d'une distribution en famille exponentielle permet de reformuler la distribution à l'aide de ce qu'on appelle des paramètres naturels .
- Statistique fréquentiste : elles permettent d'obtenir facilement des statistiques d'échantillonnage, à savoir les statistiques suffisantes naturelles de la famille, qui résument un échantillon de données à l'aide d'un nombre réduit de valeurs.
- Statistique bayésienne : elles possèdent des prieures conjuguées qui facilitent la mise à jour des distributions dites "subjectives".

De plus, la distribution prédictive a posteriori d'une variable aléatoire de famille exponentielle (à prieure conjuguée) peut toujours s'écrire en forme close (pour autant que le facteur de normalisation de la famille exponentielle puisse lui-même s'écrire en forme close). Il est à noter toutefois que souvent ces distributions ne sont pas elles-mêmes de famille exponentielle. Exemples courants : la t de Student, la bêta-binomiale ou la Dirichlet-multinomiale.

2.3 Présentation du modèle linéaire généralisé :

2.3.1 Définition du modèle

Dans bien des applications, les variables à expliquer ne varient pas dans tout R mais dans R_+,N ou encore un intervalle d'entiers. Il est clair que le modèle gaussien est mal adapté à cette situation. Le modèle linéaire généralisé (GLM) spécifie que y_i est une variable aléatoire dont la loi est paramétrée par une combinaison linéaire des régresseurs $x_i\beta$, par exemple $y_i \sim P(x_i\beta)$.

En pratique la situation est la suivante : on dispose de données y et X (réponses et variables explicatives) ; il faut alors spécifier une famille $(P_\theta)_{\theta \epsilon R}$ de distributions de probabilité à un paramètre réel θ ainsi qu'une fonction réelle $\eta \mapsto r(\eta)$ dont l'inverse est appelé fonction

de lien, qui fait le lien entre le paramètre θ et $x\beta$. Le modèle est alors[2] :

$$y_i \sim P_{\theta_i}, i = 1...n$$

$$E_{\theta_i}[Y] = r(x_i\beta)$$

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

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

- Valeurs réelles positives
  - Gamma
  - Inverse gaussienne
- Valeurs entières positives non bornées
  - Poisson
- Valeurs entières positives dans un intervalle
  - Binomiale

    La distribution binomiale négative est également proposée mais semble peu utilisée

Choix de r. Le choix par défaut proposé par les logiciels est $r(\eta) = E_\eta[Y]$, ce qui conduit à $\theta_i = x_i\beta$; ce choix permet une estimation numériquement robuste, et conduit à des valeurs réalistes indépendamment de x_i (i.e par exemple comprises entre 0 et 1 si la loi est binomiale).Pour comprendre les implications du choix de r, précisons la paramétrisation des modèles. Ces familles ont toutes une densité de la forme :

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

2.3.2 Les composantes du modèle linéaire généralisé :

Les modèles linéaires généralisés sont caractérisés par trois composantes : la composante aléatoire, le prédicteurs linéaire ou composante déterministe et la fonction lien[21].

- **La composante aléatoire :** identifie la distribution de probabilités de la variable à expliquer, qui est définie par la distribution de probabilité de La variable réponse y. Elle peut être choisie dans la famille exponentielle à laquelle appartiennent les

lois normales, binomiales, de Poisson, gamma, etc... Une propriété de ces lois est que pour chacune d'elle, il existe une relation spécifique entre l'espérance $E(Y) = \mu$ et la variance $Var(Y) = a(\phi)Var(\mu)$, souvent $a(\phi) = \phi$.

- **Prédicteur linéaire :** ou la composante déterministe est définie par la fonction linéaire des variables explicatives, utilisées comme prédicteurs dans le modèle. Dans un modèle linéaire généralisé, l'espérance mathématique de Y, notée μ, varie en fonction des valeurs des variables explicatives. Le prédicteur linéaire est exprime sous forme d'une combinaison linéaire.

$$\eta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n$$

- **la fonction lien :** exprime une relation fonctionnelle entre la composante déterministe et la composante aléatoire. Elle spécifie le fait que l'espérance mathématique de Y , notée μ, est liée au prédicteur linéaire et construit à partir des variables explicatives.

2.3.2.1 La composante aléatoire :

Elle est définie par la distribution de probabilité de la variable réponse.

Soit $Y_1, Y_2, ..., Y_n$ des variables d'un échantillon aléatoire de la variable réponse Y, ces variables étant supposées indépendantes admettant des distributions issues d'une famille exponentielle.

Chaque observation Y_i admet une fonction de densité de la forme :

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

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

Remarque : l'expression précedente représente la forme la plus générale des modèles linéaires généralisés. Elle englobe l'ensemble des lois usuelles utilisant un ou deux paramètres, tels que la loi normale, l'inverse de la loi normale, la loi gamma, la loi Poisson et la loi binomiale. Dans lequel ϕ est un paramètre de nuisance. Si ϕ est une constante connue (en générale elle est égale a 1) l'expression précédante se met sous la forme canonique

suivante :

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

En posant :

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

Le terme $\vartheta(\theta)$ est appelé le paramètre naturel de la distribution. Tout autre paramètre de la distribution est considéré comme un paramètre de nuisance.

Cette famille $f(Y_i, \theta)$ comprend de nombreuses distributions importantes telles-que la loi de Poisson et la loi binomiale. La valeur du paramètre θ_i dépend des valeurs des variables explicatives.

2.3.2.2 La composante déterministe :

La composante déterministe du modèle se rapporte à un vecteur d'un ensemble de variables explicatives $\eta = \eta_1, \eta_2, ..., \eta_n$ par un modèle linéaire :

$$\eta = X\beta$$

La matrice X se compose de n valeurs des variables explicatives, β est le vecteur des paramètres du modèle, le vecteur η est appelé le prédicteur linéaire La fonction lien spécifie comment l'espérance mathématique de Y, notée μ, est liée au prédicteur linéaire construit à partir des variables explicatives.

2.3.2.3 La fonction lien :

Cette composante exprime une relation fonctionnelle entre la composante déterministe et la composante aléatoire.

Soit $\mu = E(Y)$ on pose $\eta = g(\mu)$.

où g est appelé fonction lien, c'est une fonction différenciable et monotone.

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

$$g(\mu) = X'\beta$$

La fonction lien qui associe la moyenne μ au paramètre naturel est appelée fonction de lien canonique. A toute loi de probabilité de la composante aléatoire est associée une

fonction spécifique de l'espérance appelée paramètre canonique. Pour la distribution normale il s'agit de l'espérance elle-même. Pour la distribution Poisson le paramètre canonique est le logarithme de l'espérance . Pour la distribution binomiale le paramètre canonique est la probabilité de succès. Le tableau suivant résume les différents types de modèles couverts par le modèle linéaire généralisé[4].

Composante a	**Lien**	**Composante d**	**Modèle**
Normale	Identité	Quantitatives	Régression
Normale	Identité	Quantitatives	Analyse du var
Normale	Identité	Mixtes	Analyse du cov
Binomiale	Logit	Mixtes	Régression logistique
Poisson	log	Mixtes	Modèles log-linéaires
Multinomiale	Logit généralisé	Mixtes	Modèles à réponses multinomiales

TABLE 2.1: *Récapitulatif des principaux modèles*

Avec :

Composante a : les composantes aléatoires ;

Composante d : les composantes déterministes.

Les fonctions de lien typiquement utilisées sont données dans le tableau suivants :

Lien	$\mu = r(\eta)$
identité	η
logarithme	e^{η}
logit	$\frac{1}{1+e^{-\eta}}$
loglog complémentaire	$1 - \exp(-e^{\eta})$
probit	$\Phi(\eta)$
puissance	$\eta^{1/\alpha}$

TABLE 2.2: *Tableau des fonctions de lien*

2.3.3 Les distributions du modèle

2.3.3.1 La distribution de Poisson :

Si y_i désigne l'effectif de la i^{me} cellule distribuée selon une loi de Poisson de paramètres $E(y_i) = \mu_i$ et $Var(y_i) = \mu_i$. Sa distribution de probabilité est définie par :

$$f(y_i, \mu_i) = e^{-\mu_i} \frac{\mu_i^{y_i}}{y_i!}$$

$$\log f(y_i, \mu_i) = (y_i \log \mu_i - \mu_i - \log y_i!)$$

La 2^{me} forme exponentielle de la distribution de Poisson est définie par :

$$f(y_i, \mu_i) = \exp(y_i \log \mu_i - \mu_i - \log y_i!)$$

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

$$f(y_i, \theta) = \exp(-\mu_i) \frac{1}{y_i!} \exp[y_i \log \mu_i]$$

Les composantes du modèle sont données par :
$\vartheta(\theta_i) = log\mu_i$, $a(\theta_i) = \exp(-\mu_i)$, $b(y_i) = \frac{1}{y_i!}$.
La fonction de lien canonique est $g(\mu_i) = log\mu_i$, elle permet de modéliser le logarithme de l'espérance, le modèle utilisant ce lien est le modèle *loglinaire*.

2.3.3.2 La distribution binomiale :

Notons $Y_1, Y_2, ..., Y_n$ des variables d'un échantillon aléatoire de taille n de la variable de réponse Y, ces variables étant supposées indépendantes. Chaque variable Y_i est binaire (succès-échec), ou de façon plus générale Y_i peut être le nombre de succès au cours d'un certain nombre d'essais. On supposera alors que la composante aléatoire est distribuée selon une loi binomiale de paramètres $E(y_i) = \Pi_i$ et $Var(y_i) = \Pi_i(1-\Pi_i)$. Sa distribution de probabilité est définie par :

$$f(y_i, \Pi_i) = C_n^{y_i} \Pi_i^{y_i} (1 - \Pi_i)^{n - y_i}$$

$$\log f(y_i, \Pi_i) = \log C_n^{y_i} + y_i \log \Pi_i + (n - y_i) \log(1 - \Pi_i)$$

$$\log f(y_i, \Pi_i) = \{y_i \log \frac{\Pi_i}{1 - \Pi_i} + n \log(1 - \Pi_i) + \log C_n^{y_i}\}$$

La 2^{me} forme exponentielle de la distribution de binomiale est définie par :

$$f(y_i, \Pi_i) = (1-\Pi_i)^n C_n^{y_i} \exp[y_i \log \frac{\Pi_i}{1-\Pi_i}]$$

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

Les paramètres sont :

$a(\theta_i) = (1-\Pi_i)^n$, $b(y_i) = C_n^{y_i}$, $\vartheta(\theta_i) = \log \frac{\Pi_i}{1-\Pi_i}$

La fonction lien canonique est $g(\Pi_i) = \log \frac{\Pi_i}{1-\Pi_i}$ elle est appelée fonction *logit* de Π.

2.3.3.3 La distribution normale :

Soit $Y_1, Y_2, ..., Y_n$ des variables d'un échantillon aléatoire gaussien de la variable$Y$, ces variables sont supposées indépendantes. Alors la composante aléatoire est distribuée selon une distribution normale de paramètres (μ, σ^2), les fonctions de densités de ces variables s'écrivent.

$$f(y_i, \mu_i, \sigma_i^2) = \frac{1}{\sigma_i^2 \sqrt{2\Pi}} e^{\frac{-1}{2\sigma_i^2}(y_i - \mu_i)^2}$$

$$\log f(y_i, \mu_i, \sigma_i^2) = \log \frac{1}{\sigma_i^2 \sqrt{2\Pi}} - \frac{1}{2\sigma_i^2}(y_i - \mu_i)^2$$

$$\log f(y_i, \mu_i, \sigma_i^2) = \exp[\frac{-1}{2\sigma_i^2} y_i^2 + \frac{2}{2\sigma_i^2} y_i \mu_i - \frac{-1}{2\sigma_i^2} \mu_i^2 - log\sigma_i^2 \sqrt{2\Pi}]$$

La 2^{me} forme exponentielle de la loi Normale selon la forme du GLM est :

$$\log f(y_i, \mu_i, \sigma_i^2) = \exp\{[\frac{y_i \mu_i}{\sigma_i^2}] + [-\frac{\mu_i^2}{2\sigma_i^2}] + [-\frac{y_i^2}{2\sigma_i^2} - \frac{1}{2}\log[2\Pi\sigma_i^2]]\}$$

Les composantes du modèles sont :

$\vartheta(\theta_i) = \mu_i$, $a(\theta_i) = \exp[-\frac{\mu^2}{2}]$, $b(y_i) = \exp\{-\frac{y_i^2}{2} - \frac{1}{2}\log[2\Pi\sigma_i^2]\}$ $\phi = \sigma_i^2$.

La famille gaussienne se met sous cette forme canonique. C'est une famille exponentielle de paramètre de dispersion $\phi = \sigma_i^2$ et de paramètre naturel $\theta_i = E(y_i) = \mu_i$, Donc la fonction de lien canonique est la fonction identité $g(\mu_i) = \mu_i$.

Le tableau suivant indique les composantes de la famille exponentielle pour des lois usuelles :

Distribution	$\theta(\mu)$	$b(\theta)$	$a(\phi)$
Normale $N(\mu, \sigma^2)$	μ	$\frac{\theta^2}{2}$	σ^2
Binomiale$\beta(1, \mu)$	$\log \frac{\mu}{1-\mu}$	$\log(1 + e^{\theta})$	1
Poisson $P(\mu)$	$\log \mu$	e^{θ}	1
Gamma $G(\mu, \nu)$	$\frac{-1}{\mu}$	$-\log -\theta$	1
Gauss inverse $IG(\mu, \sigma^2)$	$\frac{-1}{2\mu^2}$	$-\sqrt{-2\theta}$	σ^2

TABLE 2.3: *Tableau récapitulatif des composantes des lois de la famille exponentielle*

2.3.4 Principe d'estimation d'un modèle linéaire généralisé :

2.3.4.1 La méthode des moindres carrées :

Les paramètres du modèle linéaire sont classiquement estimés par la méthode des moindres carrées qui vise à minimiser la somme des carrés des écarts entre les valeurs observées Y_i et les valeurs prédites $\mu_i = \sum_{j=0}^{p} \beta_j X_{ij}$.
Donc la somme des carrées des écarts entre les réponses observées et prédites est une fonction quadratique des paramètres inconnus :

$$\sum_{i=1}^{n} [Y_i - \sum_{j=0}^{p} \beta_j X_{ij}]^2$$

2.3.4.2 La méthode du maximum de vraisemblance

L'estimation des paramètres β_j est calculée par la maximisation du log de vraisemblance du modèle linéaire généralisé. Cette estimation s'applique à toutes les lois de distributions appartenant à la famille exponentielle de la 2^{me} forme.

– **Estimation des coefficients par la méthode du maximum de vraisemblance :**

Soit Y une variable qui obéit à une loi de distribution $f(y_i, \theta_i, \phi)$. A partir d'un certain nombre d'observations de Y,$(Y_1, Y_2, .., Y_n)$ on détermine les valeurs inconnues des paramètres θ_i. La méthode du maximum de vraisemblance postule que cette valeur de θ_i devrait être celle qui maximise la probabilité et d'obtenir les valeurs

observées de Y. La procedure d'estimation par la méthode du maximum de vraisemblance de la première expression du modèle linéaire généralisé est définie par :
La Fonction de vraisemblance du GLM :

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

En général Il est plus pratique d'opérer sur la transformation logarithmique de $f(y_i,\theta_i,\phi)$ qui s'exprime comme une somme de fonctions de θ plutôt qu'un produit de fonctions comme c'est le cas pour $f(y_i;\theta_i,\phi)$. Que l'on désigne par $\log f(y_i;\theta_i,\phi)$ et on la note $l(y_i;\theta_i,\phi)$.
Le log de vraisemblance de la i^{me} observation est exprimée par :

$$l(y_i;\theta_i,\phi) = [y_i\theta_i - b(\theta_i)]/a(\phi) + c(y_i,\phi)$$

La Maximisation du log de vraisemblance est en fonction des premières et deuxièmes dérivées. alors on l'applique aux résultats de la vraisemblance :

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

$$\frac{\partial^2 l}{\partial \theta_i^2} = -b^{"}(\theta_i)/a(\phi)$$

Les conditions de régularités des lois appartenant aux familles exponentielles sont vérifiées et permettent d'écrire :
$E(\frac{\partial l}{\partial \theta_i}) = 0$ et $-E(\frac{\partial^2 l}{\partial \theta_i^2}) = E(\frac{\partial^2 l}{\partial \theta_i^2})$
Alors :

$$E(Y_i) = \mu_i = b^{'}(\theta_i)$$

Et comme

$$E\{b^{"}(\theta_i)/a(\phi)\} = E\{[Y_i - b^{'}(\theta_i)]/a(\phi)^2]\} = Var(Y_i)/a^2(\phi)$$

Donc :

$$Var(Y_i) = b^{"}(\theta_i)a(\phi)$$

Ainsi ϕ est appelé paramètre de dispersion lorsque, μ est la fonction d'identité.

– **Équations de vraisemblance**

Soit X La matrice qui se compose de p observations des variables explicatives, β un vecteur de p paramètres du modelé et η le prédicateur linéaire à n composantes défini par .

$$\eta = X\beta$$

La fonction lien est supposée être monotone et différenciable telle que :

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

La fonction lien canonique est donnée par :

$$g(\mu_i) = \theta_i$$

Soit n observations indépendantes et θ dépend de β ,le log de vraisemblance est défini par :

$$L(\beta) = \sum_{i=1}^{n} \log f(y_i; \theta_i, \phi) = \sum_{i=1}^{n} \log l(y_i; \theta_i, \phi)$$

Pour obtenir les équations de vraisemblance, on Calcule :

$$\frac{\partial l_i}{\partial \beta_j} = \frac{\partial l_i}{\partial \theta_i}, \frac{\partial \theta_i}{\partial \mu_i}, \frac{\partial \mu_i}{\partial \eta_i}, \frac{\partial \eta_i}{\partial \beta_j}$$

Comme :

$$\frac{\partial l_i}{\partial \theta_i} = [y_i - b^{'}(\theta_i)]/a(\phi)\} = \{[y_i - \mu_i]/a(\phi)$$

$$\frac{\partial \mu_i}{\partial \theta_i} = \frac{Var(Y_i)}{a(\phi)} = b^{"}(\theta_i)$$

Avec $\frac{\partial \eta_i}{\partial \beta_j} = x_{ij}$, car $\eta_i = \sum_{j=1}^{p} \beta_j x_{ij}$

Puisque,$\frac{\partial \mu_i}{\partial \eta_i}$ dépends de la fonction lien $\eta_i = g(\mu_i)$ du modèle alors :

$$\frac{\partial l_i}{\partial \beta_j} = \frac{(y_i - \mu_i)}{a(\phi)} \times \frac{a(\phi)}{Var(Y_i)} \times \frac{\partial \mu_i}{\partial \eta_i} \times x_{ij}$$

Les équations de vraisemblance sont données par :

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ij}}{Var(Y_i)} \times \frac{\partial \mu_i}{\partial \eta_i} = 0$$

Ces équations sont non linéaires en β.pour les résoudre on utilise des méthodes itératives telles que la méthode de newton Raphson (ou l'on utilise le Hessien) ou la

méthode des scores de fisher (on utilise la matrice d'information). Les éléments de la matrice d'information sont données par :

$$E(\frac{\partial^2 L(\beta)}{\partial\beta_j\partial\beta_k})$$

Elle est notée par F et égale à :

$$F = X^{'}WX$$

De terme général

$$F = E(\frac{\partial^2 L(\beta)}{\partial\beta_j\partial\beta_k}) = -\sum_{i=1}^{n}\frac{x_{ik}x_{ij}}{Var(Y_i)} \times (\frac{\partial\mu_i}{\partial\eta_i})^2$$

Ou W est la matrice diagonale, ces éléments sont définies par :

$$w_i = \frac{1}{Var(Y_i)} \times (\frac{\partial\mu_i}{\partial\eta_i})^2$$

La procédure d'estimation par la méthode du maximum de vraisemblance de la deuxième expression du modèle linéaire généralisé est, (plusieurs simplifications interviennent) :

$$\eta_i = \frac{\partial\mu_i}{\partial\theta_i} = \sum_{j=1}^{p}\beta_j x_{ij}$$

$$\frac{\partial\mu_i}{\partial\eta_i} = \frac{\partial\mu_i}{\partial\theta_i} = \frac{\partial b^{'}(\theta_i)}{\partial\theta_i} = b^{"}(\theta_i)$$

Ainsi

$$\frac{\partial l_i}{\partial\beta_j} = \frac{(y_i-\mu_i)}{Var(Y_i)} \times b^{"}(\theta_i) \times x_{ij} = \frac{(y_i-\mu_i)}{a(\phi)} \times x_{ij}$$

Les termes $\frac{\partial^2 L(\beta)}{\partial\beta_j\partial\beta_k}$ ne dépendent plus de y_i alors on montre que le Hessien est égale à la matrice d'information et donc les méthodes de newton Raphson coïncident avec les méthodes de résolutions des scores de Fisher.

Si, en plus $a(\phi)$ est constante pour les observations, alors les équations de vraisemblance s'écrivent :

$$X^{'}y = X^{'}\mu$$

La valeur de la fonction en son maximum joue un rôle central dans la définition du test statistique du rapport de vraisemblance.

2.3.5 Construction pratique d'un modèle linéaire généralisé

La Construction pratique d'un modèle linéaire généralisé est une interprétation des résultats qui sont basée sur :

- Le choix du modèle (ie.la loi de probabilité de la fonction de réponse) ;
- Les statistiques permettant d'apprécier l'adéquation du modèle aux données ;
- Les tests d'hypothèse concernant les coefficients du modèle ;
- La construction d'intervalles de confiance pour les coefficients du modèle ;
- L'analyse des déviances et des résidus.

2.3.5.1 Le choix du modèle :

Le plus souvent le choix de la loi de probabilité de la fonction de réponse découle naturellement de la nature du problème étudié. On peut alors choisir comme fonction de lien, la fonction de lien canonique associée à la loi de probabilité de la fonction de réponse étudiée.
Il est toujours possible d'utiliser d'autre fonction de lien par exemple : l'identité, le Logit, Probit, Puissance et Logarithme.

2.3.5.2 Adéquation du modèle :

Deux statistiques sont utiles pour juger de l'adéquation du modèle aux données :

- La déviance normalisée.
- La statistique du Khi-deux de Pearson.

Un modèle linéaire généralisé est défini par la loi de probabilité $f(y;\theta)$ de la réponse Y et la nature de la fonction de lien g reliant l'espérance μ et Y aux variables explicatives $X_1, X_2, ..., X_k$, $g(\mu_i) = x_i'\beta$.
On note b l'estimation du maximum de vraisemblance de β. Pour mesurer l'adéquation du modèle étudié aux données, on construit tout d'abord un modèle saturé, modèle basé sur la même loi de probabilité et la même fonction de lien, mais contenant autant de variables explicatives indépendantes que de données : ce modèle permet de reconstruire

parfaitement les données. On note b_{max} l'estimation du vecteur des paramètres β pour ce modelé.

– **La déviance :**

Dans le modèle linéaire, la décomposition de la somme des carrées en somme des carrées expliquée par le modèle, SCL, et somme des carrées résiduelle, SCR, fournit une mesure d'adéquation classique, le coefficient de détermination

$$R^2 = \frac{SCL}{SCL + SCR}$$

Il fournit une mesure pratique d'adéquation parce qu'il est compris entre 0 et 1 et que sa borne supérieure peut être atteinte à condition d'utiliser un nombre suffisamment élevé de coefficients (modèle dit saturé).

Ce coefficient, dont le calcul ne fait intervenir que les valeurs observées et prédites, pourrait être calculé pour n'importe quel modèle linéaire généralisé, cependant il ne bénéficierait pas des mêmes avantages. Il suffit d'imaginer le cas de données binaires comme dans une enquête épidémiologique ou tous les sujets de l'échantillon ont des profils de risque différents, par exemple parce que certains risques sont mesurés sur une échelle continue. Les observations ne peuvent prendre n'importe quelle valeur de l'intervalle $[0, 1]$.

On préfère utiliser un coefficient appelé déviance $D = 2(L_{max} - L_0)$ qui représente le double du logarithme d'un rapport de vraisemblance. Il est toujours possible de construire un modèle qui comprend autant de paramètres que d'observations. Ce modèle comprend autant de paramètres que d'observations distinctes (modèle dit saturé) et puisqu'il ne permet pas de les résumer. Cependant on peut le considérer comme une référence : aucun modèle ne s'ajuste mieux et sa log-vraisemblance L_{max} est la plus élevée parmi tous ceux de la famille considérée. La log-vraisemblance L_0 du modèle dont on veut mesurer l'adequation est inférieure, mais si la difference n'est pas trop élevée on pourra affirmer qu'il s'ajuste bien aux données. Sous l'hypothèse que le modèle considère est correct, on montre que $2(L_{max} - L_0)$ suit asymptotiquement une distribution de $chi - 2$ dont le nombre de degrés de liberté est égal à la différence entre le nombre de paramètres des 2 modèles. On peut remarquer qu'il est strictement équivalent de comparer deux modèles emboîtés en calculant la différence

de leurs déviances ou le double de la différence de leur log-vraisemblances.

Dans le modèle linéaire qui suppose des observations normales de variance constante, la deviance $\frac{\sum(y_i-\hat{y})^2}{s^2}$ est apparentée à la somme des carrés résiduelles $SCR = \sum(y_i - \hat{y})^2$.

Dans un modèle logistique, qui suppose des observations y_i distribuées selon des lois binomiales $B(n_i, \Pi_i)$, la vraisemblance du modèle saturé qui prédit les probabilités Π_i est $\upsilon_0 = \prod \Pi_i^{y_i}(1-\Pi)^{n_i-y_i}$, celle du modèle saturé qui prédit les fréquences observées $f_i = y_i/n_i$ et $V_{max} = \prod f_i^{y_i}(1-f_i)^{n_i-y_i}$ et la déviance vaut donc :

$$D = 2\sum[y_i \log \frac{f_i}{n_i} + (n_i - y_i)\log\frac{1-f_i}{1-\Pi_i}]$$

$$D = 2\sum o_{ij}\log\frac{O_{ij}}{E_{ij}}; j = 0, 1$$

indice les deux modalités de réponse possibles. De façon générale, le tableau suivant donne la forme de la déviance pour quatre des principales distributions de la famille exponentielle. La déviance des différentes distributions de la famille exponentielle sont données dans le tableau suivant :

Distribution	**Déviance**
Normale	$\sum(y-\hat{\mu})^2$
Poisson	$\sum[y\log\frac{y}{\hat{\mu}} - (y-\hat{\mu})]$
Binomiale	$2[y\log\frac{y}{\hat{\mu}} + (n-y)\log(\frac{n-y}{n-\hat{\mu}})]$
Gamma	$2\sum(-\log\frac{y}{\hat{\mu}} + \frac{y-\hat{\mu}}{\hat{\mu}})$

TABLE 2.4: *Déviance des différentes distributions de la famille exponentielle*

– **Les résidus :** L'observation des résidus permet d'identifier les observations qui s'ajustent mal au modèle. Les résidus bruts représentent les différences $Y_i - \hat{Y}_i$entre réponse observée et prédite. Leur variance vaut $\sigma^2(1-h_{ii})$. On peut leur préférer les résidus de Pearson $\frac{Y_i-\hat{Y}_i}{s}$, de déviance $(1-h_{ii})$, dont les carrées représente la contribution de l'observation au X^2 de Pearson, ou les résidus studentisés $\frac{Y_i-\hat{Y}_i}{s\sqrt{1-h_{ii}}}$ de variance unité.

Ces trois types de residus existent sous une forme dite de prediction qui utilise la

regression effectuée sur l'ensemble de l'échantillon moins le point considère. En appelant X_i la matrice des variables explicatives sans la ligne x_i, β_i et s_i les estimations correspondantes des coefficients β et l'écart type résiduel σ, $\hat{Y}_i$ la valeur prédite par x_i et β_i, On obtient les résidus de prédiction $Y_i - \hat{Y}_i$ et $\frac{Y_i - \hat{Y}_i}{s\sqrt{1-h_{ii}}}$, noter que dans cette formule, le bras de levier h_{ii}, qui ne depend que des covariables X, et calculé sur la matrice complete. Il est plus pertinent de comparer a une distribution de Student (ou pratiquement a une distribution normale) la valeur des residus studentisés de prediction que celle des résidus de Student simples car une observation suffisamment excentrée a la fois du point de vue des Co variables et du point de vue de la variable réponse Y,a un résidus nul.

Deuxième partie

Application

Chapitre 3

Application

Ce chapitre sera consacré à l'application des différents modèles statistiques déjà présentés sur des données réelles.

3.1 Application de la régression linéaire simple

3.1.1 Introduction

Notre première application portera sur la régression linéaire simple.
Pour se faire nous allons procéder comme suit :

- Récolte de données ;
- Représentation graphique des données ;
- Estimation des paramètres du modèle ;
- Tests statistiques ;
- Intervalle de confiance.

3.1.2 Récolte de données

Notre étude portera sur le trafics aérien dans l'aéroport de béjaia, où nous allons étudier le nombre de passagers en fonction du temps ; Les données couvrent la période 2005 jusqu'à 2013 (Voir Table 3.1) .

On note par :

NP : nombre de passagers (la variable à expliquer) ;

AN : année (la variable explicative).

Nombre de passagers	**Année**
103455	2005
104043	2006
115864	2007
205312	2008
229333	2009
217453	2010
233782	2011
249156	2012
295020	2013

TABLE 3.1: *Nombre de passagers par année*

3.1.3 Représentation graphique des données

Dans notre étude nous avons utilisé le logiciel R, nous avons introduit les données du nombre de passagers sous forme d'un vecteur que nous avons nommé NP en utilisant la formule suivante :

$> NP$=c(103455,104043,115864,205312,229333,217453,233782,249156,295020)

$> NP$

[1]103455 104043 115864 205312 229333 217453 233782 249156 295020

Et nous avons introduit les années sous forme d'un vecteur que nous avons nommé AN en utilisant la formule suivante :

$> AN$=c(2005,2006,2007,2008,2009,2010,2011,2012,2013)

$> AN$

[1] 2005 2006 2007 2008 2009 2010 2011 2012 2013

Puis nous avons utilisé la formule suivante pour tracer le nuage de points :

$> plot(AN, NP)$

Le graphe ci-dessous représente le résultat obtenu :

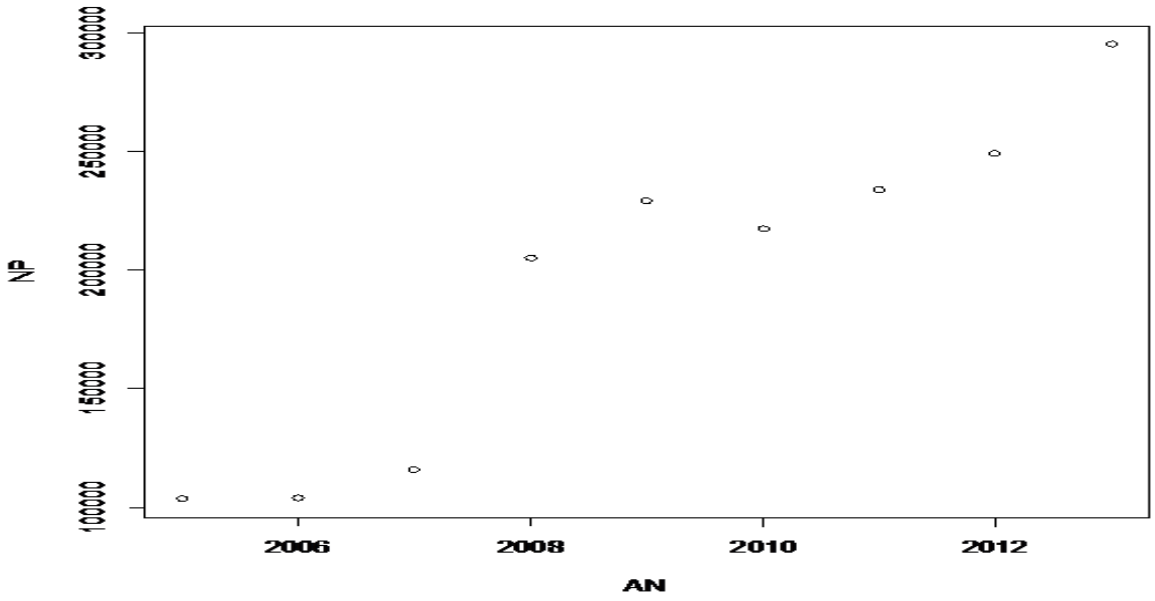

FIGURE 3.1: *nombre de passager en fonction du temps*

On remarque une régularité dans le nuage, presque un alignement. Cet élément suggère qu'une droite serait un bon résumé de ces points. Donc le modèle adéquat pour ce problème est le modèle de regression linéaire simple.

3.1.4 Estimation des paramètres du modèle

Le modèle est donné par la formule suivante :

$$y = ax + b$$

Nous avons appliqué la méthode des moindres carrées ordinaires (MCO) pour estimer les paramètres du modèle.

Pour se faire nous avons utilisé la commande suivante :

$> lm(NP \sim AN)$

Call :

$> lm(formula = NP \sim AN)$

Coefficients :

(Intercept) AN

-48341812 24160

Les paramètres estimées sont :

$\hat{\alpha} = 24160$ et $\hat{\beta} = -48341812$

Le modèle estimé est donné par :

$$\hat{y} = \hat{\alpha}x + \hat{\beta} + \epsilon_t$$

$$\hat{y} = 24160x - 48341812 + \epsilon_t$$

avec $t = 1, ..., 9$.

La droite d'ajustement représentée sur le même plan que le nuage de points (par la commande "abline" sous R) est donnée par :

> abline(48341812, 24159.6, col ='red')

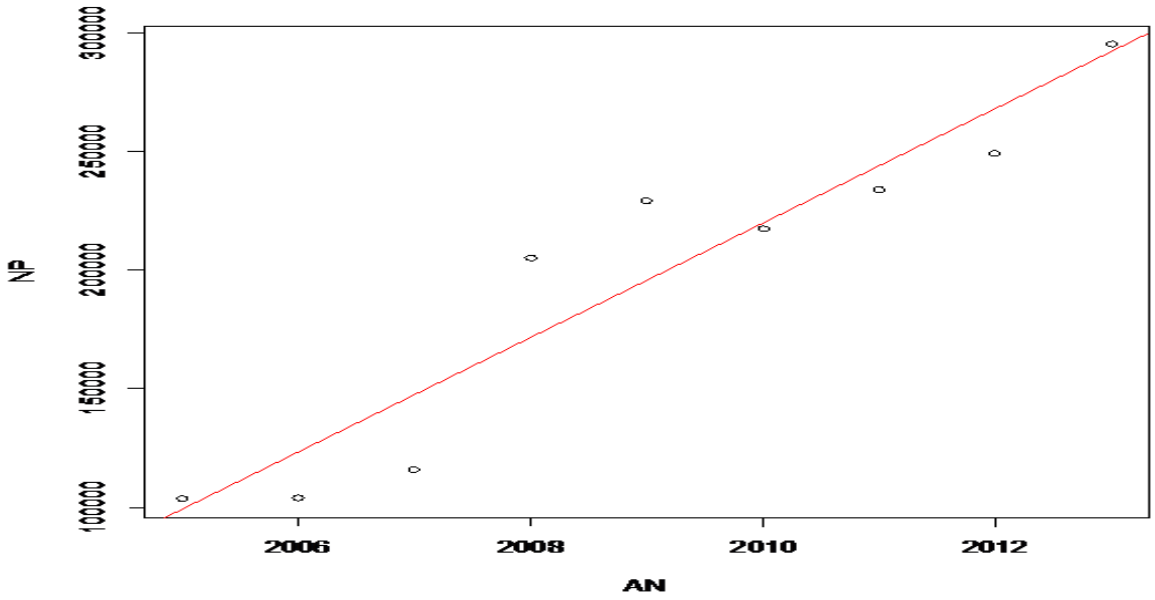

FIGURE 3.2: *nombre de passager en fonction du temps*

Les valeurs estimée $\hat{y}_i$ sont obtenues par :

$\hat{y}_i = \hat{\alpha}x_i + \hat{\beta}$

La commande suivante nous donne les estimations :

$> (yChap < -alphaChap * AN + betaChap)$

[1] 98185.82 122345.42 146505.02 170664.62 194824.22 218983.82 243143.42 267303.02 291462.62

Les résidus sont obtenus par :

$e_i = y_i - \hat{y}_i$.

$> (residus < -NP - yChap)$

[1] 5269.178 -18302.422 -30641.022 34647.378 34508.778 -1530.822 -9361.422 -18147.022 3557.378

La somme des carrés des résidus (SCR) :

$> (SCR < -sum(residus^2))$

[1] 4124860662

La somme des carrés totale (SCT) :

$> (SCT < -sum((NP - mean(NP))^2))$

[1] 39146036992

Ces deux valeurs permettent d'obtenir la somme des carrés expliquée (SCE=SCT-SCR) :

La somme des carrés expliquée SCE :

$> (SCE < -SCT - SCR)$

[1] 35021176330

Le coefficient de corrélation linéaire :

La commande cor() permet d'afficher directement le coefficient de corrélation linéaire, et on obtient un résultat très proche de 1

$> cor(AN, NP)$

[1] 0.9458482

Le coefficient de détermination (R^2) :

$> (R2 < -1 - SCR/SCT)$

[1] 0.8946289

La variance des résidus :

$> (hatSigma2_u < -SCR/(n-2))$

[1] 589265809

La variance estimée des estimateurs de α et β sont données par :

La variance estimée de α : $> (hatSigma2_alphaChap < -hatSigma2_u/(sum(AN^2) - n * mean(AN)^2))$

[1] 9821097

La variance estimée de β :

$> (hatSigma2_betaChap < -mean(AN)^2 * hatSigma2_alphaChap + hatSigma2_u/n)$

[1] 3.963881e+13

3.1.5 Tests statistiques

Le modèle étant un modèle de regression simple, alors les tests utilisés sont Fisher et Student qui permettront de tester la signification de la variable " temps " dans le modèle.

Test de student :

Test :"$H_0 : a_1 = 0$" contre "$H_1 : a_1 \neq 0$"

Pour n=9 taille de l'échantillon, k=1 nombre de variable explicative.

En utilisant la commande suivante :

$> (tObs_alpha < -(alphaChap - 0)/hatSigma_alphaChap)$

On trouve :

[1] 7.709208

On compare à la valeur théorique :

$> qt(p = 1 - 0.05/2, df = n - 2)$

[1] 2.364624

On trouve : $t^*_{\hat{a_1}} = 7.709$. Au seuil de signification $\alpha = 5\%$,la valeur théorique lue sur la table de student est $t_{(\alpha/2,n-2)} = t_{(0.025,7)} = 2.3646$.

On a $\mid t^*_{\hat{a_1}} \mid = 7.709 >= t_{(0.025,7)} = 2.3646$.

Donc La variable "temps" est une variable explicative significative pour le modèle.

Le test de Fisher :

Test :"$H_0 : a_1 = 0$" contre "$H_1 : a_1 \neq 0$"

Pour n=9 taille de l'échantillon, k=1 nombre de variable explicative.

En utilisant la commande suivante :

$> (Fobs < -(R2/1)/((1-R2)/(n-2)))$

On trouve :

[1] 59.43188

On compare à la valeur théorique :

$> qf(p = 1 - 0.05, df1 = 1, df2 = n - 2)$

[1] 5.591448

On trouve : $F^* = 59.43$. Au seuil de signification $\alpha = 5\%$,la valeur théorique lue sur la table de Fisher pour n-2 degrés de liberté est $F_{5\%} = 5.59$.

On a $F^* = 59.43 > F_{\alpha(1,7)} = 5.59$

D'où on rejette largement l'hypothèse H_0 et on accepte l'hypothèse H_1 d'où la variable "temps" est une variable explicative significative pour le modèle.

3.1.6 Intervalle de confiance

Pour calculer l'intervalle de confiance pour a_1, on utilise la fonction prédifinie "confint" comme suit au niveau 95% :

$> confint(lm(NP \sim AN))$

On aura : $IC_{95\%}(a_1) = [16749.19 \pm 31570.01]$

3.2 Application de la régression linéaire multiple

3.2.1 Introduction

Dans cette partie notre application sera réalisé sur la régression linéaire multiple. Pour se faire nous allons procédé comme suit :

- Récolte de données ;
- Estimation des paramètres du modèle ;
- Tests statistiques ;
- Intervalle de confiance.

3.2.2 Récolte de données

Notre étude portera sur les sorties de marchandises dans le port de béjaai durant l'année 2014 que nous souhaitons expliquer en fonctions de certains facteurs (entrées de marchandises et le temps donnée par mois).

Le tableau suivants regroupe nos données :

Sorties de marchandises	**Entrées de marchandises**	**Mois**
1412117	1441120.5	1
1184007	1102467.2	2
1600881	1751045.8	3
1528210	1396056.9	4
1481342	1543081.9	5
1479066	1664679.3	6
1626816	1474142.2	7
1560075	152765.3	8
1598639	1698379.1	9
1533378	1488074.2	10
1840133	1830762.3	11
1862422	1819132.5	12

TABLE 3.2: *les entrées et sorties de marchandises pendant une année*

3.2.3 Estimation des paramètres du modèle

Le modèle est donné par la formule suivante :

$Sorties = a_0 + a_1 entres + a_2 Temps + \epsilon$;

D'où notre modèles à deux facteurs est données par :

$Y = a_0 + a_1 X^1 + a_2 X^2 + \epsilon$

Nous avons appliqué la méthode des moindres carrées qui cherche la meilleure estimation des paramètres "a_i" pour estimer les paramètres du modèle.

Pour se faire nous allons tout d'abord introduire nos données sous forme de vecteurs dans

le logiciel R :

>Sorties=c(1412116.79,1184007.47,1600881.35,1528209.79,1481342.06,1479066.43,1626816.03,
1560075.15,1598639.11,1533378.01,1840132.73,1862421.65)

> Sorties

[1] 1412117 1184007 1600881 1528210 1481342 1479066 1626816 1560075 1598639 1533378
1840133 1862422.

>Entrées=c(1441120.49,1102467.17,1751045.82,1396056.94,1543081.89,1664679.29,
1474142.18,152765.32,1698379.09,1488074.23,1830762.34,1819132.54)

>Entrées

[1] 1441120.5 1102467.2 1751045.8 1396056.9 1543081.9 1664679.3 1474142.2 152765.3
1698379.1 1488074.2 1830762.3 1819132.5

>Temps=c(1 :12)

>Temps

[1] 1 2 3 4 5 6 7 8 9 10 11 12.

Les commandes suivantes nous donne les estimations :

$> Tab = data.frame(Sortie, Entres, Temps)$

$> Step(lm(Sortie \tilde{} Entres + Temps, Data = tab)$

> Call

$lm(Formula = Sortie \tilde{} Entres + Temps, Data = tab)$

Coeffecients :

(Intercept) Entrées Temps

5.54×10^5 5.81×10^{-1} 1.724×10^4

Les estimateurs sont donnés par :

$\hat{a}_0 = 5.54 \times 10^5$

$\hat{a}_1 = 5.81 \times 10^{-1}$

$\hat{a}_2 = 1.724 \times 10^4$

D'où le modèle estimé est donné par :

$\hat{y} = 5.54 \times 10^5 + 5.81 \times 10^{-1} X^1 + 1.724 \times 10^4 X^2 + \epsilon$

3.2.4 Tests statistiques

Test individuel de Student

– Test sur le coefficient a_1 :

Soient les hypothèses suivantes :

H_0 : "$a_1 = 0$" contre H_1 : "$a_1 \neq 0$"

ce test est basé sur la statistique :

$t^*_{\hat{\alpha}_1} = \mid \frac{\hat{\alpha_1}}{\hat{\sigma}_{\hat{\alpha_1}}} \mid \longrightarrow t_{(n-3)}$

En utilisant la commande suivante sous le logiciel R on aura :

> (tObs <-hatalpha1/ hatSigmaalpha)

> [1] 3.57 *

> p-value=0.0002768

On a donc :

$t^*_{\hat{\alpha}_1} = 3.57$

On compare à la valeur théorique :

> (T-tab <- qt(p = 1-0.05/2, df = n-3))

[1] 2.055529

On trouve $\mid t^*_{\hat{a_1}} \mid = 3.57 > t_{tab} = 2.3646$.

Et on a :

p-value=0.0002768 < * "0.05"

D'où rejette l'hypothèse H_0 et on accepte H_1.

le coefficient a_1 est significativement différent de zéro.

– Test sur le coefficient a_2 :

Soient les hypothèses suivantes :

H_0 : "$a_2 = 0$" contre H_1 : "$a_2 \neq 0$"

ce test est basé sur la statistique :

$t^*_{\hat{\alpha}_2} = \mid \frac{\hat{\alpha_2}}{\hat{\sigma}_{\hat{\alpha_2}}} \mid \longrightarrow t_{(n-3)}$

En utilisant la commande suivante sous le logiciel R on aura :

> (tObs <-hatalpha2/ hatSigmaalpha)

> [1] 1.91 *

> p-value=0.0002768

On a donc :

$t^*_{\hat{a}_1} = 1.91$

On compare à la valeur théorique :

> (T-tab <- qt(p = 1-0.05/2, df = n-3))

[1] 1.765

On trouve $\mid t^*_{\hat{a_1}} \mid = 1.91 > t_{tab} = 1.765$.

Et on a :

p-value=0.0002768 < * "0.05"

D'où on rejette l'hypothèse h_0 et on accepte h_1.

le coefficient a_2 est significativement différent de zéro.

Test globale de Fisher :

On test les hypothèses suivantes :

"$H_0 : a_1 = a_2 = 0$" contre "$H_1 : \exists a_i \neq 0, i = 1, 2$"

Ce test est basé sur la statistique de fisher suivante :

$F^* = \frac{\frac{R^2}{P}}{\frac{1-R^2}{n-p-1}} = \frac{\frac{R^2}{1}}{\frac{1-R^2}{n-2}} \longrightarrow F_{(2,n-3)}$

En utilisant la commande suivante :

> (F-obs <- (R2 / p) / ((1-R2) / (n-p-1)))

[1] 23.29

On aura :

$F^* = 23.29$

La valeur lue sur la table de fisher est :

> (F-tab <- qf(p = 1-0.05, df1 = p, df2 = n-p-1))

$F_{tab} = 4.2$

Donc :

$F^* > F_{tab}$ d'où on rejette H_0 ; donc le modèle est globalement significatif.

3.2.5 Intervalle de confiance

Pour obtenir les intervalles de confiance pour a_0,a_1 et a_2 on utilise la commande suivante :

$> Confit(lm(Sortie \tilde{} Entre + Temps))$

On aura :

intercept 64470.6868 1.043450e+06

Entrée 0.21287166 9.494191e-01

Temps -3181.9632094 3.765826e+04

D'où :

$IC_{a_0} = [64470.6868; 1.043450 \times 10^{06}]$

$IC_{a_1} = [0.21287166; 9.494191 \times 10^{-01}]$

$IC_{a_2} = [-3181.9632094; 3.765826 \times 10^{04}]$

Remarquons bien que :

$\hat{a_0} \in IC_{a_0}$

$\hat{a_1} \in IC_{a_1}$

$\hat{a_2} \in IC_{a_2}$

D'ou le modèle est validé.

3.3 Application de la régression logistique

3.3.1 Introduction

Dans ce qui suit notre application portera sur la régression logistique.

Pour se faire nous allons procédé comme suit :

- Récolte de données ;
- Représentation graphique des données ;
- Estimation des paramètres du modèle ;
- Tests statistiques ;
- Intervalle de confiance.

3.3.2 Récolte des données

On souhaite étudier la présence ou l'absence d'une maladie cardiovasculaire chez 40 personnes où x représente l'âge et y la variable indiquant s'ils sont atteint ou non de la maladie.

Voici le tableau qui regroupe les données :

Age	**P/A**	**Age**	**P/A**	**Age**	**P/A**	**Age**	**P/A**
47	1	66	0	67	0	55	0
35	1	35	1	32	1	56	0
22	1	52	1	38	0	31	1
39	1	61	0	51	1	48	0
30	1	28	1	51	0	50	0
46	0	67	0	27	1	66	0
45	1	25	1	38	0	21	1
45	1	31	1	48	1	44	1
34	1	68	1	52	1	43	0
52	1	47	1	50	1	59	0

TABLE 3.3: *Tableau des données des malades*

Avec :

Age : représente l'age du malade ;

P/A : représente la présence ou l'absence de la maladie(ie 0 :absence 1 :présence).

3.3.3 Représentation graphique des données

Nous avons introduit les données de l'âge des personnes dans le logiciel R sous forme d'un vecteur que nous avons nommé X en utilisant la formule suivante :

>x=c(47,35,22,39,30,46,45,45,34,52,66,35,52,61,28,67,25,31,68,47,67,32,38,51,51,27,
38,48,52,50,55,56,31,48,50,66,21,44,43,59)

$> x$

[1] 47 35 22 39 30 46 45 45 34 52 66 35 52 61 28 67 25 31 68 47 67 32 38 51 51 27 38 48

52 50 55 56 31 48 50 66 21 44 43 59

Et nous avons introduit les données de présence ou absence de la maladie sous forme d'un vecteur y :

>y=c(1,1,1,1,1,0,1,1,1,1,0,1,1,0,1,0,1,1,1,1,0,1,0,1,0,1,0,1,1,1,0,0,1,0,0,0,1,1,0,0)

> y

[1] 1 1 1 1 1 0 1 1 1 1 0 1 1 0 1 0 1 1 1 1 0 1 0 1 0 1 0 1 1 1 0 0 1 0 0 0 1 1 0 0

Puis nous avons utilisé la formule suivante pour la représentation graphique :

>Plot(x,y)

Le graphe ci-dessous représente le résultat obtenu :

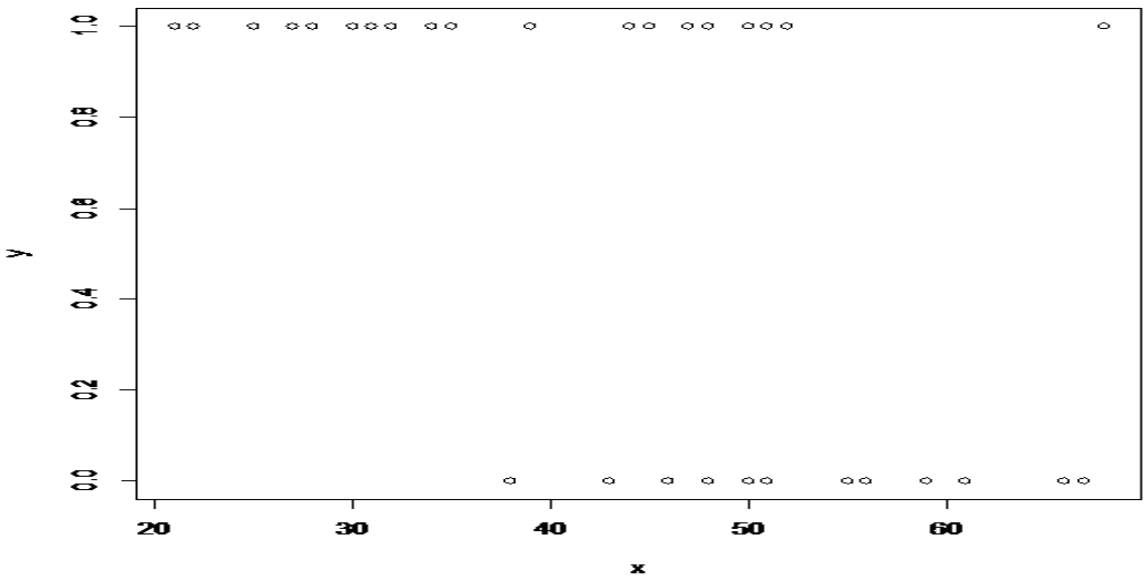

FIGURE 3.3: *présence ou absence de la maladie*

Si l'on représente la relation entre logit(E(Y)) et X, on retrouve bien une relation linéaire. En revanche, l'échelle des ordonnées n'est pas aisée à interpréter. On procède donc à une transformation inverse de la relation :

En utilisant les formules suivantes sous R :

> logit-ypredit=-0.12×x+5.95

> ypredit=exp(logit-ypredit)/(1+ exp(logit-ypredit)) transfo inverse de logit

> plot(x,y)

> o=order(x)

> points(x,ypredit, col="red",type="l", lwd=2)

On obtient le graphe suivant :

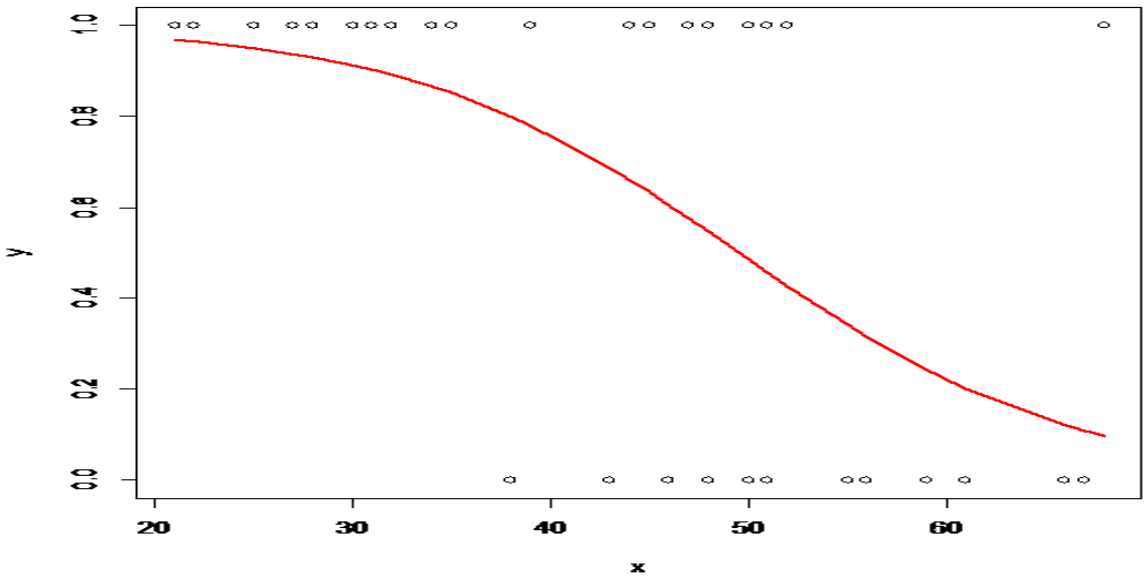

FIGURE 3.4: *présence ou absence de la maladie*

3.3.4 Estimation des paramètres du modèle

Le modèle est donné par la formule suivante :

$logit(E(y)) = \beta_0 x + \beta_1$

En utilisant les formules suivantes sur R :

$> reg = glm(y\tilde{}x, family = binomial(link = logit))$

$> reg$

$Call : glm(formula = y\tilde{}x, family = binomial(link = logit))$

Coefficients :

(Intercept) x

5.9462 -0.1156

On obtient donc le modèle suivant :

$Logit(E(y)) = -0,12x + 5.94$

3.3.5 Tests statistiques

Test de Wald :

Le test de Wald permet d'évaluer l'influence de X_j sur Y.

En utilisant les formules suivantes :

> Summary(req)

> Estimate Std. Error z value Pr(>|z|)

on obtient :

Coefficients :

(Intercept) 5.9462 1.9599 3.034 0.00241 **

x -0.1156 0.0397 -2.912 0.00360 **

Remarquons bien que le code "**" est attribué à ce test ce qui signifie que l'influence de X_j sur Y est très significative.

Test de la déviance :

Le test de la déviance vise à évaluer la contribution de X_j sur Y .

Dans ce test la p-valeur associée utilise la loi du Chi-deux.

En utilisant la formule suivante sous "R" :

> anova(reg, test = "Chisq" ,Model : binomial, link : logit)

On aura :

Df Deviance Resid. Df Resid. Dev Pr(>Chi)

NULL 39 52.925

x 1 13.308 38 39.617 0.0002643 ***

Dans ce test le code attribué est "***", d'où l'influence de X_j sur Y est hautement significative.

Test lr :

Pour ce test, on utilise le test du rapport de vraisemblance (lr pour likelihood ratio)

En utilisant les formules suivantes :

> reg=glm(y~x, family = binomial)

> reg0=glm(y~1, family = binomial)

> anova(reg0, reg, test = "Chisq")

On obtient le résultat suivant :

Resid. Df Resid. Dev Df Deviance Pr(>Chi)

1 39 52.925

2 38 39.617 1 13.308 0.0002643 ***

Aussi le code attribué à ce test est "***", d'où l'influence de X_j sur Y est hautement significative.

On conclut que le modèle est significatif.

3.3.6 Intervalle de confiance

Intervalle de confiance pour β_j :

Pour obtenir l'intervalles de confiance pour β_j on utilise la formule suivante :

> confint.default(reg, level = 0.95) On obtient le résultat suivant :

(Results) 2.5 % 97.5 %

(Intercept) 2.1049627 9.7874537

x -0.1933855 -0.0377741

Intervalle de confiance pour p(x) :

Pour obtenir l'intervalles de confiance pour $p(x)$ on utilise la formule suivante :

> logitp = predict.glm(reg, se.fit = TRUE)

> iclogit = c(logitp fit - 1.96 × logitp se.fit,logitp fit + 1.96 × logitp se.fit)

> ic = (exp(iclogit) / (1 + exp(iclogit))

On obtient le résultat suivant :

(Results) 2.5 % 97.5 %

Px 0.43559408 0.70198207

3.4 Application du modèle linéaire généralisé

Dans cette dernière partie notre étude portera sur les modèles linéaires généralisés. Pour pouvoir procéder a cette application nous avons suivis les étapes suivantes :

- Analyser les données "le choix de la loi de probabilité de la fonction réponse" ;
- Estimation des paramètres du modèle ;
- Tests statistiques d'hypothèse sur les paramètres du modèle ;

3.4.1 Les données

Dans notre étude nous souhaitons étudier l'influence de la dose d'un poison (disulfide de carbone) sur la mortalité de cafards.

Les données sont rangés dans le tableau suivant :

Epreuve	**la dose**	**total**	**les morts**
1	1.691	59	6
2	1.724	60	16
3	1.755	62	18
4	1.784	56	28
5	1.811	63	52
6	1.837	59	53
7	1.861	62	61
8	1.884	60	60

TABLE 3.4: *Tableau des données*

3.4.2 Modélisation du problème

On note :

i=0,...,8 les groupes ;

n_i : taille du i^{me} groupe ;

Y_i : nombre de morts dans le groupe i ;

x_i : la dose de poison ;

Π_i : la probabilité de mourir dans le groupe i.

Le modèle est alors :

$$y_i \sim B(n_i, \Pi_i)$$

$B(n_i, \Pi_i)$: la loi Binomiale de paramètre n_i et Π_i.

Avec :

$$\Pi_i = a + bx_i$$

On veut garder la simplicité d'interprétation du modèle linéaire.

On ne modélise pas directement Π_i mais $g(\Pi_i)$:

g :fonction lien

$g(\Pi_i) = a + bx_i$

$g : [0, 1] \longrightarrow \mathbb{R}$

La fonction de lien adéquate pour le modèle est logit :

$g(\Pi) = log(\frac{\Pi}{1-\Pi})$

3.4.3 Estimation des paramètres du modèle

Pour estimer les paramètres du modèle sous R on utilise les formules suivantes :

>cafards< −read.table("cafards.dat", header=TRUE)

>attach(cafards)

> cafards

ldose total morts

1.691 59 6

1.724 60 13

[...]

> y< −cbind(morts,total-morts)

> model< −glm(y~ldose, family=binomial(link="logit"))

> y.prop< −morts/total

> model.prop< −glm($y.prop$~ldose, weights=total, family=binomial(link="logit"))

Les résultats obtenus sont :

Call :

glm(formula = y~ ldose, family = binomial(link = "logit"))

Deviance Residuals :

Min $1Q$ Median $3Q$ Max

-1.5878 -0.4085 0.8442 1.2455 1.5860

Coefficients :

Estimate Std. Error z value Pr(> $|z|$)

(Intercept) -60.740 5.182 -11.72 <2e-16 ***

ldose 34.286 2.913 11.77 <2e-16 ***

Interprétation des paramètres

$g(\Pi) = log(\frac{\Pi}{1-\Pi}) = a + bx_i = -11.72 + 5.182x_i$

$\Pi_i = \frac{\exp(a+bx_i)}{1+\exp(a+bx_i)}$

On a :

$\Pi(.) = \frac{\exp(a)}{1+\exp(a)}$: la probabilité de décès quand on ne met pas de poison.

$\Pi_i(x) = 0.5$ Dose l'étale à 50% :

Pour $x_i = \frac{-a}{b}$

$\frac{\Pi_2/(1-\Pi_2)}{\Pi_1/(1-\Pi_1)} = \exp(b(x_2 - x_1))$

Calcule de la vraisemblance :

En utilisant la formule suivante sous R :

> LVsat <− sum(log(dbinom(morts,total,morts/total)))

On aura :

$v(y_1, y_2, ..., y_n)$=-13.09902.

Calcule $E(y_i)$:

$E(Y_i)$ est estimé comme la moyenne p_0 par max de vraisemblance ;

En utilisant les formules suivantes :

> p_0 <− sum(morts)/sum(total)

> LV_0 <− sum(log(dbinom(morts,total,p_0)))

On aura :

E(y)= -155.2002 **Calcule de la deviance :**

$D = 2(l_{max} - L_0)$

En utilisant la formule suivante :

> dev0 = 2*(LVsat-LV0)

D= 284.2024

Deviance residuelle :

En utilisant la formule suivante :

> devx = 2*(LVsat-LVx)

On aura :

Null deviance : 284.202 Chi_2 on 7 degrees of freedom

Residual deviance : 11.116 Chi_2 on 6 degrees of freedom

Dx=11.11558

Le modèle estimé :

Nous avons tracé la courbe du modèle estimé en utilisant les formules suivantes sous R :

> y< −cbind(morts,total-morts)

> model< −glm(y~ldose, family=binomial(link="logit"))

> curve y

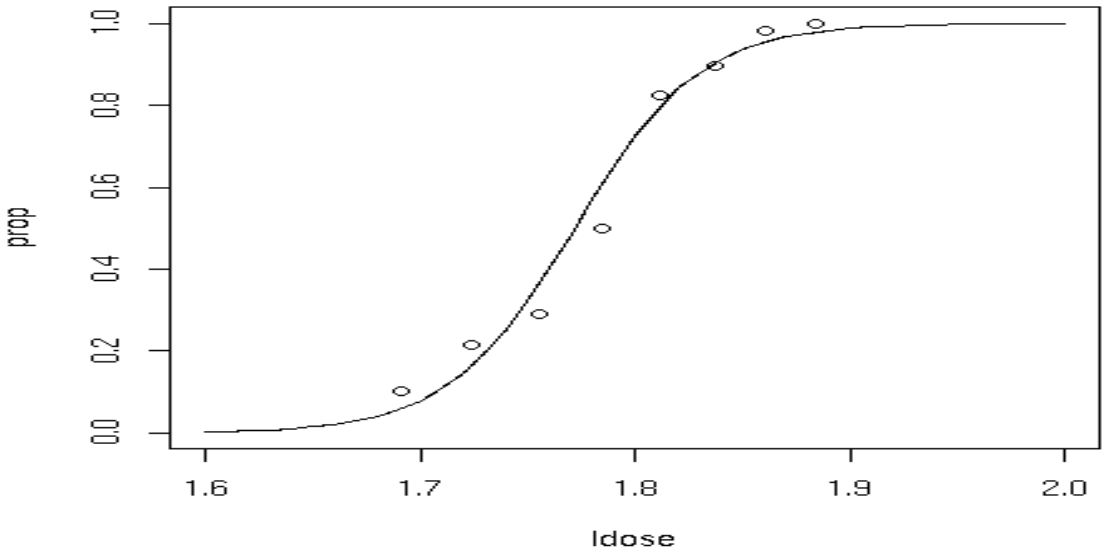

FIGURE 3.5: *courbe du modèle estimé*

Interprétation de la courbe :

- Le poison n'est utile qu'à une dose supérieure à 1.6 ;
- la mortalité atteint son maximum à une dose supérieure ou égale à 1.9.

Conclusion générale

En dépit du fait que les modèles linéaires généralisés sont désormais des outils classiques en statistiques, ils ne sont encore que trop rarement intégrés. Leurs extensions récentes, avec inclusion d'effets aléatoires ou de classes latentes, rendent ces outils d'une importance considérable pour une étude statistique. Leur maîtrise suppose d'avoir clairement à l'esprit les hypothèses inhérentes à tout modèle de régression, en particulier les concepts de distribution, de moyenne et de variance conditionnelles.

Nous avons tenus dans ce mémoire à mettre en évidence ce type de modèles statistiques ainsi que la démarche à suivre pour leurs réalisation : la récoltes de données, proposition du modèle adéquat, estimations des paramètres etc.

Nous terminons ce travail par une application sur les différents modèles, présentés auparavant, sur des données réelles.

Enfin, nous souhaitons que notre travail a permit de montrer l'importance de ces modèles statistiques et qu'il servira de référence pour d'éventuelles études statistiques théoriques ou appliquées.

Bibliographie

[1] D.Bernard, Cours de deuxième année de master, Université Rennes I, 28 avril 2017.

[2] P.Besse, Pratique de la modélisation Statistique, université paul sabatier, janvier 2003.

[3] A.Charpentier, Modèles linéaires généralisés, ACT2040 ,université de Québec Montréal 2013.

[4] M.Chikhi et M.Chavance, Estimation du modèle linéaire généralisé et application, A – N°35, 21/02/2012.

[5] Claude et Michel, Modèle linéaire généralisé application dans la prédiction de la qualité du café rwandais, Mémoire licencié en Mathématiques Appliquées,University Avenue, Novembre 2008

[6] Cullagh et Nelder, Modèles linéaires généralisés, cours université toulouse 3, 1989.

[7] Cullagh et Nelder, Introduction au modèle linéaire, université de toulouse ,1990.

[8] Cullagh et Nelder, Introduction au modèle linéaire général, université de toulouse, 1992.

[9] Cullagh et Nelder, Introduction au modèle de régression linéaire simple, université de toulouse, 1992.

[10] Cullagh et Nelder, Introduction au modèle de régression linéaire multiple, université de toulouse, 1992.

[11] Cullagh et Nelder , Introduction au modèle de régression logistique, université de toulouse, 1992.

[12] Cullagh et Nelder, Introduction au modèle de régression log linéaire, université de toulouse, 1992.

[13] P.Givord, Introduction à l'econométrie, école Centrale de Paris,Année 2006-2007.

[14] F.Lale, Formation statistique au modèle linéaire général, support de cours,28 avril 1995.

[15] J.Lenoir, Modèles linéaire généralisé GLMS, université de picardie.

[16] S.Mazouz et A.Acherchour, Etude statistique du trafics aérien et du nombre de passager de l'aéroporrt de béjaia, mémoire licence, université de béjaia.

[17] J.Nelder And R.Wedderburn, Cours sur les modèle linéaire généralisé, 27/3/2016.

[18] F.Picard, Première notions de statistique régression linéaire, UMR CNRS-5558, université lyon1.

[19] J.S. Pierre, Modèle linéaire généralisé, UMR6552, 14 novembre 2004.

[20] R.Rakotomalala, Pratique de la régression logistique, Université Lumière Lyon 2, 12/02/2014.

[21] Sophia, Introduction au modèle linéaire généralisé, Université de Nice, Octobre 2011.

[22] A.Trabelsi et R.Resprige, Cours régression de Poisson, université de lille.