République Algérienne Démocratique et Populaire
Ministére de l'Enseignement et de la Recherche Scientifique

Université A.MIRA de Béjaïa
Faculté des Sciences Exactes
Département de Mathématiques

Mémoire de fin de cycle Master

En vue de l'obtention du diplôme de Master

Option : Probabilité Statistique et Application

THÈME

La régression logistique bayésienne

---

présenté par :

M^{elle} **ZITOUT** Yasmina

Soutenu, le 15 octobre 2020 devant le jury composé de :

Mme	TABTI.H	MCB	Présidente	U.A.Mira Béjaia
Mme	LAGHA.K	MCA	Rapporteur	U.A.Mira Béjaia
Mr	BOURAINE.M	MAA	Examinateur	U.A.Mira Béjaia
Mme	TIMRIDJINE.K	MCA	Examinatrice	U.A.Mira Béjaia

Année universitaire 2019/2020

Remerciements

Éloge à Allah, le tout-puissant de m'avoir donné la force, le courage et la volonté pour réaliser ce modeste travail.

Au terme de ce travail, j'adresse mes sincères remerciements à Mme LAGHA.K qui a assuré l'encadrement de ce mémoire. Je la remercie pour sa grande disponibilité, sa patience, ses précieux conseils et son optimisme contagieux. Je la remercie aussi pour les sujets passionnants vers lesquels elle m'a dirigé et toute l'aide qu'elle m'a fourni et grâce à laquelle j'ai pu mener à terme ce travail. Je lui en suis infiniment reconnaissante.

J'adresse mes vifs remerciements à Mme TABTI.H pour l'honneur qu'elle nous a fait en acceptant de présider le jury et d'évaluer ce mémoire à sa juste valeur.

Mes remerciements chaleureux s'adressent également à Mr BOURAINE.M et Mme TIMRIDJINE.K pour avoir accepté d'examiner ce travail et de l'enrichir avec leurs critiques.

Je remercie aussi toute personne de prés ou de loin qui a contribué dans l'élaboration de ce travail.

Dédicaces

Je dédie ce modeste travail

À mes chèrs et estimables parents qui n'ont cessé de me guider vers le droit chemin ainsi que pour leurs sacrifices et leur amour permanent.

À mes chers frères Mohammed, Djamel et Rafik, leurs épouses, ainsi que leurs enfants.

À mes chers frères Zahir et Kamel.

À mes chères soeurs Naima et Samira, leurs époux, ainsi que leurs enfants.

À ma copine : Kahina, merci pour ton amour, amitié. Tu été toujours là pour me soutenir, et m'écouter.

À toute la promotion de Master 2 PSA (2019/2020).

Z.Yasmina

Table des matières

Table des figures

Liste des tableaux

Notation & Abréviation

v.a	variable aléatoire.
$\mathbb{E}$	Espérance mathématique.
$\mathbb{V}$	Variance mathématique.
$\mathbb{P}$	Probabilité.
$\mathbb{N}$	Ensemble des entiers naturels.
$\mathbb{R}$	Ensemble des réels.
ddl	degré de liberté.
SCT	Somme des Carrés Totale.
SCE	Somme des Carrés Expliquée.
SCR	Somme des Carrés Résiduelle.
MCO	Moindres Carrés Ordinaires.
IC	Intervalle de Confiance.
diag	La diagonale d'une matrice.
GLM	Generalized Linear Models.
CM	Chaine de Markov.
MCMC	Markov Chain Monte Carlo.
iid	indépendant et identiquement distribué.
fdr	fonction de répartition.
ϵ	Erreur aléatoire.
MLE	Maximum Likelihood Estimator.
MV	Maximum de Vraisemblance.
det	Le déterminant d'une matrice.
log	Logarithme népérien.
ML	Maximum Likelihood.
ROC	Receiver Operating Characteristic.

Introduction

L Homme est curieux et c'est sans doute ce qui explique le mieux son cheminement depuis le début de l'humanité jusqu'à nos jour. Ce besoin de comprendre les phénomènes observés et le désir de les anticiper est au coeur de ses préoccupations, c'est ce qui explique l'émergence et le succès de la statistique qui est une discipline scientifique, et quant à elle une méthode et un ensemble de techniques qui consiste à recueillir, traiter et interpréter des données issues d'expériences et d'études de un ou plusieurs phénomènes. Son objectif principal est de mener grâce à l'observation d'un phénomène aléatoire une inférence sur la distribution de probabilité qui est à l'origine de ce phénoméne, elle est utilisée dans presque tous les domaines de l'activité humaine : ingénierie management, économie, biologie, informatique, etc.

Le terme "régression" a une origine curieuse, il remonte à l'étude du physiologiste et anthropologue Francis Galton (vers la fin du XIX éme siècle) sur la relation entre la taille des parents et celle des enfants. Galton a introduit le mot régression pour désigner dans des problèmes d'hérédité la diminution progressive (régression) des écarts par rapport à la moyenne, d'une génération à la suivante [17]. L'analyse de régression est une méthodologie statistique qui est utilisée pour établir une relation entre une variable quantitative et une ou plusieurs autres variables qui peuvent être quantitatives, qualitatives ou un mélange des deux sous forme d'un modèle.

Les modèles de régression sont devenus une composante intégrale de toute analyse statistique visant à décrire la relation entre une variable à expliquée et une ou plusieurs variables explicatives, ils sont largement utilisés dans de nombreuses disciplines et applications. Plusieurs modèles peuvent être distingués, tels que les modèles linéaires, linéaire généralisé, non linéaires, logistique stochastiques et bien d'autres [9].

Le modèle linéaire consiste à mettre en relation une variable à expliquer (quantitative et continue) et une ou plusieurs variables explicatives (qualitatives et ou quantitatives). Lorsque cette variable réponse est qualitative dans ce cas on parle du modèle logistique [32].

La régression logistique, en tant que modèle, a l'avantage d'expliquer une variable dépendante qualitative et une ou plusieurs variables indépendantes qualitatives ou quantitatives qui offrent plusieurs variantes en fonction du nombre et la nature de la variable dépendante. Parmi ces variantes on peut citer la régression logistique binaire dont la variable dépendante est qualitative binaire, elle peut égalemment s'appeler variable dichotomique, c'est à dire elle ne peut prendre que deux valeurs possibles ou polytomique nominale dans ce cas la variable dépendante doit être qualitative nominale, c'est à dire elle admet plus de deux modalités non ordonnées. Et enfin polytomique ordinale lorsque la variable dépendante est qualitative ordinale, c'est à dire elle prend égalemment plus de deux modalités ordonnées. Elle est considéré comme l'une des procédures statistiques les plus utilisés en recherche et est devenue de plus en plus populaire grâce à la disponibilité aisée des outils informatiques appropriés et aussi plus importante dans la plupart des domaines d'études tels que : médecine, économie et d'autres domaines similaires.

La régression logistique est également utilisée par les chercheurs et les analystes tel que Cox et Snell (1970) qui discutait de la régression logistique [8], Anderson (1972) qui traitait la discrimination logistique d'échantillons séparés [1] et bien d'autres à trois fins :

1. Pour décrire la probabilité que le résultat ou la variable réponse soit égal à 1.

2. Pour catégoriser les résultats ou les prévisions.

3. Pour accéder aux côtes ou aux risques associés au prédicteurs du modèle.

Le modèle logistique est unique en ce qu'il peut répondre à ces trois objectifs [28]. L'accent principal est d'aider à guider l'analyste dans l'utilisation des capacités du modèle logistique et ainsi aider les analystes à mieux comprendre leurs données et à faire des prédictions.

Avant de commencer une étude approfondie du modèle de régression logistique il est important de comprendre que l'objectif d'une analyse utilisant ce modèle est le même que celui de tout autre modèle de régression utilisé en statistique, c'est à dire de trouver le meilleur ajustement et le plus parcimonieux. En 1972 de nouveaux autres modèles ont été formulés par John Nelder et Robert Wedderburn appelé les Modèles Linéaires Généralisés (*Generalized Linear Models*(GLM)) comme une généralisation souple de la régression linéaire [41]. Les GLM généralise la régression linéaire en permettant au modèle linéaire d'être relié à la variable réponse via une fonction lien aussi comme un moyen d'unifier les autres modèles statistique y compris la régression linéaire et la régression logistique.

Les applications réussies des modèles cités précédemment nécessitent une bonne compréhension à la fois de la théorie sous jacente et des problèmes pratiques rencontrés lors de l'utilisation des modèles dans des situations réelles.

L'approche bayésienne est de plus en plus utilisée dans l'analyse statistique pour l'ajustement et l'étude de plusieurs modèles de régression y compris le modèle logistique connu sous le nom de régression logistique bayésienne. Pour obtenir les distributions a posteriori des paramètres de régression logistique, les méthodes d'approximation de Monté Carlo par Chaine de Markov (MCMC) sont très puissantes et indispensables et ont donc été mise au point afin d'approximer la loi a posteriori lorsque l'on ne sait pas le faire analytiquement.

Certains travaux sur l'application des méthodes bayésiennes aux modèles logistiques sont contenues dans Genkin et al [19], comme par exemple la régression logistique bayésienne appliquée à la catégorisation des textes linguistiques et aussi la comparaison des deux méthodes d'estimation classiques et bayésiennes. Mila et Michailides (2006) ont étudié la prédiction de la gravité de la panicule et de la brulure des pousses de pistachio en Californie en utilisant une régression logistique bayésienne [35], ils ont noté que les méthodes bayésiennes donnaient des résultats plus cohérents. Gordovil, Guardia, Pero et Fuente (2010) ont présenté l'estimation bayésienne comme une alternative aux procédures classiques d'estimation par la régression logistique dans l'étude du Trouble Déficitaire de l'Attention avec Hyperactivité (TDAH) dans un échantillon méxicain [25]. Ces études montrent clairement que l'application de la régression logistique tout comme d'autres méthodes statistiques couvre un large éventail de domaine de recherche par les statisticiens.

Le but de ce mémoire est de montrer l'interêt et l'impact de l'estimation bayésienne dans les modèles logistiques.

Ce mémoire est structuré en trois chapitres essentiels agencés comme suit :

☞ Le premier chapitre introduit les concepts de base de l'inférence classique et bayésienne ainsi que les notations et la terminologie appropriée sur lesquels se fonde l'analyse des deux écoles (classique et bayésienne) dont nous aurons besoin pour établir les prochains chapitres de ce mémoire.

☞ Le deuxième chapitre consiste en une présentation des trois modèles classiques de régression (linéaire, logistique et GLM) avec leurs propriétés. Ce chapitre sera défini en trois parties.

— La première partie couvre les concepts fondamentaux de la régression linéaire, la présentation des modèles linéaire, les limites ainsi que les extensions.

— La deuxième partie concerne le modèle de régression linéaire généralisé.

— La troisième partie sera consacrée aux différents types de modèles logistiques, les méthodes d'estimation ainsi leurs propriétés.

☞ Le troisième chapitre introduit l'approche bayésienne pour l'estimation des paramètres des modèles logistique, en posant les bases théoriques qui mènent aux méthodes MCMC, nous présentons quelques notions sur les MCMC, des diagnostics de convergence y sont aussi discutés.

Nous terminons par une conclusion où on citera quelques perspectives de recherche.

Chapitre1

Inférence classique et bayésienne

Un aspect important de l'inférence statistique consiste à obtenir des estimations fiables des caractéristiques d'une population. L'école classique (fréquentiste) et l'école bayésienne traitent le problème différemment. Dans l'école classique les données sont des observations provenant d'une population, dont la loi dépend d'un paramètre inconnu θ qu'il faut estimer et dans l'école bayésienne, on tient compte d'une information supplémentaire qui nous aide à donner plus de précision dans l'estimation, avec le paramètre θ considéré comme une variable aléatoire.

Dans ce chapitre, nous allons présenter l'inférence faite par les deux écoles.

1.1 Inférence classique

On considère un modèle statistique $(\mathcal{X}, \mathcal{A}, \{P_\theta, \theta \in \Theta\})$ où $\mathcal{X}$ est l'espace des observations $(= \mathbb{R}^n)$, $\mathcal{A}$ est la sigma algèbre, Θ c'est l'espace des paramètres et P_θ c'est la loi de probabilité dépendant du paramètre $\theta \in \Theta$, définie sur cet espace. Considérons une variable aléatoire (v.a) X distribuée suivant la loi de probabilité P_θ ayant pour densité f_θ.

Soit $\underline{X} = (X_1, X_2, ...X_n)$ un échantillon de taille n issu de la v.a X, et on note $\underline{x} = (x_1, x_2, ...x_n)$ l'observation de $\underline{X}$.

1.1.1 Notions préliminaires

Definition 1.1.1. Un **estimateur** est une statistique $\hat{\theta}$ permettant de proposer une valeur pour le paramètre inconnu à estimer θ.

1.1.2 Propriétés de l'estimateur

- **Estimateur sans biais**

$\hat{\theta}$ est un estimateur sans biais de θ si :

$$\mathbb{E}(\hat{\theta}_n - \theta) = \mathbb{E}(\hat{\theta}_n) - \theta = 0.$$

- **Convergence**

Il existe plusieurs modes de convergence dont on définit quelques-uns :

- *Convergence en probabilité*

$\hat{\theta}$ converge en probabilité vers θ si :

$$\forall \ \ \epsilon > 0, \quad \lim_{n \to +\infty} \mathbb{P} \ [\mid \hat{\theta}_n - \theta \mid > \epsilon] = 0.$$

-*Convergence presque-sûre*

$\hat{\theta}$ converge presque surement vers θ si :

$$\mathbb{P} \ (\lim_{n \to +\infty} \hat{\theta}_n = \theta) = 1.$$

-*Convergence en moyenne d'ordre p*

$\hat{\theta}$ converge vers θ en moyenne d'ordre p si :

$$\lim_{n \to +\infty} \mathbb{E}[\mid \hat{\theta}_n - \theta \mid^p] = 0.$$

-*Convergence en loi*

$\hat{\theta}$ converge en loi vers θ, si et seulement si pour toute fonction φ continue bornée on a :

$$\lim_{n \to +\infty} \mathbb{E}[\varphi(\hat{\theta}_n)] = \mathbb{E}[\varphi(\theta)].$$

Pour plus de détails sur d'autres différents modes de convergence voir [18].

- **Variance minimale**

L'estimateur $\hat{\theta}$ est de variance minimale si et seulement si pour tout autre estimateur $\hat{\theta}^*$ sans biais pour θ, on a :

$$\mathbb{V}(\hat{\theta}) \leq \mathbb{V}(\hat{\theta}^*).$$

- **Dominance**

$\hat{\theta}$ est dominant si pour tout autre estimateur $\hat{\theta}^*$, on a :

$$MSE(\hat{\theta}) \leq MSE(\hat{\theta}^*),$$

avec $MSE(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^2] = \mathbb{V}(\hat{\theta}) + B(\hat{\theta}, \theta)^2$, appelé l'erreur quadratique moyenne de $\hat{\theta}$ par rapport à θ et $B(\hat{\theta}, \theta)^2$ est le biais de $\hat{\theta}$.

- **Admissibilité**

L'estimateur $\hat{\theta}$ est admissible s'il n'existe pas un autre estimateur $\hat{\theta}^*$ strictement meilleur au sens où, pour tout $\theta \in \Theta$

$$MSE(\hat{\theta}^*) \leq MSE(\hat{\theta})$$

et pour un certain θ_0 on a :

$$MSE_{\theta_0}(\hat{\theta}^*) < MSE_{\theta_0}(\hat{\theta})$$

- **Efficacité**

L'estimateur $\hat{\theta}$ est efficace de θ si

$$\mathbb{V}(\hat{\theta}) = \frac{1}{I_n(\theta)},$$

où $\frac{1}{I_n(\theta)}$ est la borne de *Fréchet-Darmois-Cramer-Rao (FDCR)* et $I_n(\theta)$ est l'information de Fisher apportée par l'échantillon de taille n sur le paramètre θ, définie par :

$$I_n(\theta) = nI_X(\theta) = n\mathbb{E}\left[\left(\frac{\partial ln(f(X,\theta))}{\partial \theta}\right)^2\right].$$

1.1.3 Méthodes classiques d'estimation

(a) La méthode du maximum de vraisemblance (MV)(en anglais Maximum Likelihood ML)

Soit $\underline{X} = (X_1, ..., X_n)$ un échantillon de taille n issu d'une variable aléatoire X de loi de probabilité $f_\theta(x) = f(x, \theta), \theta \in \Theta$ dont la fonction de vraisemblance est définie par :

$$\mathcal{L}(\theta, \underline{x}) = \prod_{i=1}^{n} f(\underline{x}, \theta),$$

où $\underline{x} = (x_1, ..., x_n)$.

Le principe de la méthode MV consiste à trouver l'estimateur $\hat{\theta}_{MV}$ qui maximise $\mathcal{L}(\theta, \underline{x})$ c'est-à-dire :

$$\hat{\theta}_{MV} = \underset{\theta \in \Theta}{argmax}\,(\mathcal{L}(\theta, \underline{x})) = \underset{\theta \in \Theta}{argmax}\; ln(\mathcal{L}(\theta, \underline{x})).$$

L'estimateur du MV est obtenu en résolvant le système d'équations suivant :

$$\begin{cases} \frac{\partial}{\partial\theta} ln(\mathcal{L}(\theta, \underline{x})) & = 0 \\ \frac{\partial^2}{\partial\theta^2} ln(\mathcal{L}(\theta, \underline{x})) & < 0. \end{cases}$$

Propriétés des estimateurs du maximum de vraisemblance

Les estimateurs obtenus par la méthode du maximum de vraisemblance sont [4] :

— Convergents.

— Asymptotiquement sans biais, cependant $\hat{\theta}_{MV}$ n'est pas nécessairement sans biais.

— Asymptotiquement efficaces.

— Asymptotiquement gaussiens.

(b) Estimation par la méthode des moments

Soit $\underline{X} = (X_1, ..., X_n)$ un échantillon de taille n issu d'une variable aléatoire X de densité $f(x, \theta)$ où $\theta = (\theta_1, \theta_2, ..., \theta_k)$ est un paramètre inconnu. La méthode consiste à estimer les paramètres inconnus $\theta_1, \theta_2, ..., \theta_k$ en égalisant certains moments théoriques (qui dépendent de ces paramètres) $\mu_r = \mathbb{E}(X^r)$, $r = 1, ..., k$ avec les moments empiriques $M_r = \frac{1}{n}\sum_{i=1}^{n} X_i^r$. La solution du système $M_r = \mu_r$, $r = 1, ..., k$ nous donne les estimateurs de $\theta_1, \theta_2, ..., \theta_k$ [51].

$$\begin{cases} M_1 = & \mu_1 \\ M_2 = & \mu_2 \\ \quad \vdots & \\ M_k = & \mu_k \end{cases}$$

Dans la plupart des cas, les estimateurs obtenus par la méthode des moments sont consistants, convergents, asymptotiquement normaux mais en général ne sont pas efficaces.

1.1.4 Tests d'hypothèse

Les tests d'hypothèses constituent un autre aspect important de l'inférence statistique dont on distinguera deux classes de tests : `les tests paramétriques` et `non paramétriques`.

Leurs principes consiste à décider entre deux hypothèses : l'hypothèse nulle H_0 et l'hypothèse alternative H_1, au vu des résultats d'un échantillon.

Il existe de nombreux tests statistiques classiques parmi lesquels on peut citer :

⋆ **Test paramétrique de Student**

⋆ **Test paramétrique de Fisher**

⋆ **Test non paramétrique du khi-deux de Pearson généralisé**

⋆ **Test paramétrique de rapport de vraisemblance**

⋆ **Test paramétrique de Wald**

Qui seront définis par la suite dans le chapitre 2 (Modèles de régression). Pour d'autres types de tests statistiques nous renvoyons au [34].

1.2 Inférence bayésienne

La démarche bayésienne consiste à traiter le paramètre inconnu θ comme une v.a, en lui associant une loi de probabilité sur l'espace Θ dite *loi a priori*, notée $\pi(\theta)$ cette loi reflète la connaissance a priori (éventuelle) du paramètre.

On appelle le modèle statistique bayésien, le modèle $(\mathcal{X}, \mathcal{A}, P_\theta)$ muni de la loi a priori $\pi(\theta)$ et il est noté par $(\mathcal{X}, \mathcal{A}, P_\theta, \pi(\theta))$. Soit la v.a X de loi de probabilité P_θ de densité f_θ. Dans le modèle bayésien on interprète la densité f_θ comme la loi conditionnelle par rapport à θ.

Principales lois de probabilités utiles dans l'approche bayésienne

★ **Loi a priori**

La loi a priori qui est notée par $\pi(\theta)$ est une loi de probabilité modélisant l'information disponible sur le paramètre d'intérêt θ, avant le recueil des données, sa détermination est l'essence de la statistique bayésienne.

★ **La loi a posteriori**

Sa densité est donnée par :

$$\pi(\theta \mid x) = \frac{f(x \mid \theta)\pi(\theta)}{\int_\Theta f(x \mid \theta)\pi(\theta)d\theta}.$$

On peut utiliser la notion de proportionnalité, c'est-à-dire :

$$\pi(\theta \mid x) \propto f(x \mid \theta)\pi(\theta).$$

pour plus de détails [voir Annexe p 86].

★ **La loi du couple** (θ, x)

La densité de la loi du couple est définie comme suit :

$$g(\theta, x) = f(x \mid \theta)\pi(\theta).$$

★ **La loi marginale de** x

On la note par m(x), elle est calculée comme suit :

$$m(x) = \int_\Theta f(x \mid \theta)\pi(\theta)d\theta.$$

Remarque 1.2.1. Lorsqu'on manipule un échantillon $\underline{X} = (X_1, ...X_n)$, on remplace dans les expressions précédentes X par $\underline{X}$ et x par $\underline{x}$ et $f_\theta(x)$ par $f_\theta(\underline{x}) = \mathcal{L}(\theta, \underline{x})$.

1.2.1 Bases décisionnelles de l'analyse bayésienne

En pratique, l'inférence statistique conduit à une décision finale prise par le "décideur" et il est important de pouvoir comparer les différentes décisions au moyen d'un critère d'évaluation, qui va apparaitre sous forme de fonction de **coût** (perte), qui est une fonction mesurable de $(\Theta \times \mathcal{D})$ à valeures dans $\mathbb{R}_+$ notée par $\ell(\theta, \delta(x))$.

Soit $\mathcal{D}$ l'ensemble des règles de décisions δ, qui sont des applications de $\mathcal{X}$ dans $\mathcal{A}$. Le but est de trouver une règle de décision $\delta \in \mathcal{D}$, c'est à dire essayer de choisir un estimateur.

Parmi les fonctions de coût usuelles, on peut citer [50] :

⋆ **Le coût quadratique**

La fonction de coût quadratique est la fonction définie par :

$$\ell(\theta, \delta) = (\theta - \delta(x))^2.$$

Une variante de cette fonction de coût est une fonction de coût quadratique pondérée [voir Annexe p 87].

⋆ **Le coût absolu**

La fonction de coût absolu est la fonction définie par :

$$\ell(\theta, \delta(x)) = |\theta - \delta(x)|.$$

Une généralisation du coût absolu est connue sous le nom de fonction de coût linéaire par morceaux [voir Annexe p 87].

⋆ **Le coût 0-1**

La fonction de perte 0-1 est l'application ℓ définie par :

$$\ell(\theta, \delta(x)) = \begin{cases} 0, & \theta \in \Theta_0 \\ 1, & \theta \in \Theta_1. \end{cases}$$

• **Risque fréquentiste**

Pour une fonction de perte donnée $\ell(\delta, \theta)$, le risque fréquentiste est le coût moyen (l'espérance mathématique) du coût d'une règle de décision qui est défini par :

$$R(\theta, \delta) = \mathbb{E}_\theta[L(\theta, \delta(x))] = \int_{\mathcal{X}} \ell(\theta, \delta(x)) f(x \mid \theta) dx.$$

• **Risque a posteriori**

Le risque a posteriori noté par $\rho(\pi, \delta \mid x)$ est défini comme étant la moyenne du coût par rapport à la loi a posteriori :

$$\rho(\pi, \delta \mid x) = \mathbb{E}^\pi[\ell(\theta, \delta(x))] = \int_{\Theta} \ell(\theta, \delta(x)) \pi(\theta \mid x) d\theta.$$

• **Risque intégré**

Pour une fonction de perte donnée, on définit le risque intégré comme étant le risque fréquentiste moyenné sur les valeurs de θ selon leurs distributions a priori π noté $r(\pi, \delta)$, par :

$$r(\pi, \delta) = \mathbb{E}[R(\theta, \delta)] = \int_{\Theta} R(\theta, \delta)\pi(\theta)d\theta.$$

• **Risque de bayes**

Le risque de bayes est la quantité donnée par :

$$r(\pi) = r(\pi, \delta^{\pi}) = \underset{\delta \in \mathcal{D}}{inf}\ r(\pi, \delta).$$

Dans le cas où $r(\pi, \delta) < \infty$, $\delta^{\pi} = \underset{\delta \in \mathcal{D}}{argmin}(\rho(\pi, \delta(x)))$, où $\rho(\pi, \delta \mid x)$ est le risque a posteriori.

Remarque 1.2.2. Cette décision δ^{π} est appelée estimateur bayésien.

•**Risque minimax**

On appelle risque minimax (minimum du risque maximum) associé à la fonction de coût ℓ, la valeur :

$$\overline{R} = \underset{\delta \in \mathcal{D}}{inf}\ \underset{\theta \in \Theta}{sup} R(\theta, \delta) = \underset{\delta \in \mathcal{D}}{inf}\ \underset{\theta \in \Theta}{sup} \mathbb{E}_{\theta}[\ell(\theta, \delta(x))].$$

Quelques théorèmes et Lemmes importants sur la minimaxité seront donnés [voir Annexe p 86-87].

1.2.2 Estimation bayésienne

On appelle estimateur de bayes associé à une fonction de coût $\ell(\theta, \delta(x))$ et à une loi a priori π, la décision δ^{π} qui minimise le risque a posteriori c'est-à-dire :

$$\delta^{\pi} = \underset{\delta \in \mathcal{D}}{argmin}(\rho(\pi, \delta(x))). \tag{1.1}$$

En particulier on peut citer :

• **Estimateur MMSE (Minimum Mean Square Error)**

L'estimateur MMSE de θ, noté $\hat{\theta}_{MMSE}$, associé à la fonction de coût quadratique, relativement à la loi a prioti π, est la moyenne a posteriori de θ, c'est-à-dire :

$$\hat{\theta}_{MMSE}(x) = \mathbb{E}(\theta \mid x), \quad x \in \mathbb{R}.$$

• **La médiane a posteriori**

L'estimateur Bayésien associé à la loi a priori π et à la fonction de coût absolu est le fractile d'ordre $\frac{1}{2}$ (médiane) de la loi a posteriori. C'est alors la médiane a posteriori qui est donné par :

$$\mathbb{P}(\theta \mid \delta) = \mathbb{P}(\theta < \delta \mid x) = \frac{1}{2}.$$

• **Estimateur MAP (Maximum A Posteriori)**

L'estimateur MAP de θ est obtenu par maximisation de la loi a posteriori, c'est à dire :

$$\hat{\theta}_{MAP}(x) = \underset{\theta}{argmax}(\pi(\theta \mid x)).$$

Il a le grand avantage de ne pas dépendre d'une fonction de perte, et est utile pour les approches théoriques.

Remarque 1.2.3. L'estimateur MAP n'est pas un estimateur bayésien, car il ne vérifie pas la relation (1.1).

Propriétés des estimateurs de bayes [10]

P1. Les estimateurs de Bayes sont admissibles.

P2. Les estimateurs de Bayes sont biaisés.

Sous certaines hypothéses de régularité le plus souvent satisfaites en pratique, on a les deux propriétés :

P3. Les estimateurs de Bayes sont convergents en probabilité (quand la taille de l'échantillon $n \longrightarrow +\infty$).

P4. La loi a posteriori peut être asymptotiquement (i.e pour de grandes valeurs de n) approximée par une loi normale $\mathcal{N}(\mathbb{E}(\theta \mid x), \mathbb{V}ar(\theta \mid x))$.

où $\mathbb{V}ar(\theta \mid x) = \mathbb{E}[(\theta - \mathbb{E}(\theta \mid x))^2 \mid x]$ est la variance a posteriori de θ.

1.2.3 Intervalles de confiance bayésiens

L'approche bayésienne présente l'avantage de permettre une construction directe d'une région de confiance. Deux intervalles seront définis par [31] :

• **Intervalle de confiance a priori**

Un intervalle de confiance a priori J de niveau $1 - \alpha$ est l'intervalle vérifiant :

$$\mathbb{P}(\theta \in J) = \int_J \pi(\theta) = 1 - \alpha.$$

• **Intervalle de confiance a posteriori**

Un intervalle de confiance a posteriori I, de niveau $1 - \alpha$ est l'intervalle vérifiant :

$$\mathbb{P}(\theta \in I \mid X) = \int_I \pi(\theta \mid X) = 1 - \alpha.$$

1.2.4 Approche bayésienne des tests

On veut tester : $H_0 : \theta \in \Theta_0$ contre $H_1 : \theta \in \Theta_1$, où $\Theta_0 \cup \Theta_1 = \Theta$ et $\Theta_0 \cap \Theta_1 = \emptyset$.

Par définition, les décisions bayésiennes sont celles qui minimisent le coût a posteriori $\rho(\pi, \delta \mid x)$.

On a deux décisions possibles : $\begin{cases} d_0 : & \text{on ne rejette pas } H_0, \\ d_1 : & \text{on rejette } H_0. \end{cases}$

En pratique, on accepte l'hypothèse H_0 ou H_1 dès que sa probabilité a posteriori $\alpha_0 = \mathbb{P}(H_0 \mid x)$ ou $\alpha_1 = \mathbb{P}(H_1 \mid x)$, respectivement est suffisamment forte (≥ 0.9 ou 0.95).

Le facteur de bayes

On appelle facteur de bayes le rapport de probabilité a posteriori des hypothèses nulle et alternative $\frac{\alpha_0}{\alpha_1}$ (en anglais "posterior odds ratio") sur le rapport des probabilités a priori de ces mêmes hypothèses $\frac{\pi_0}{\pi_1}$ (en anglais "prior odds ratio") qui permet de comparer la pertinence du choix d'un modèle par rapport à un autre, soit [31] :

$$B(x) = \frac{\text{posterior odds ratio}}{\text{prior odds ratio}} = \frac{\frac{\alpha_0}{\alpha_1}}{\frac{\pi_0}{\pi_1}} = \frac{\alpha_0 \pi_1}{\alpha_1 \pi_0}.$$

Ce rapport évalue la modification de la vraisemblance sous l'ensemble Θ_0 par rapport à celle sous l'ensemble Θ_1, dûe à l'observation.

Cas particulier

Si $\Theta_0 = \{\theta_0\}$ et $\Theta_1 = \{\theta_1\}$, le facteur de bayes n'est que le rapport de vraisemblance classique qui est défini par :

$$B = \frac{f(x \mid \theta_0)}{f(x \mid \theta_1)} = \frac{\int_{\Theta_0} f(x \mid \theta_0)\pi_0(\theta)d\theta}{\int_{\Theta_1} f(x \mid \theta_1)\pi_1(\theta)d\theta}.$$

1.2.5 Modélisation de l'information a priori

Sans contexte, le point le plus critique de l'analyse bayésienne est le choix de la loi a priori, il est donc nécessaire le plus souvent de faire un choix (partiellement) arbitraire de la loi a priori ce qui peut avoir un impact considérable sur l'inférence qui en découle.

Le plus souvent on ne dispose pas de suffisamment d'informations a priori sur le paramètre inconnu θ pour construire la loi a priori. Dans la pratique, on a recours à des lois usuelles (lois normales, lois gamma, etc) ou à des lois dites conjuguées, lois subjectives et lois a priori maximale (lois informatives).

En l'absence d'informations a priori, on introduira la notion de loi a priori non informative qui permet de rester dans un cadre bayésien, alors même que l'on ne dispose pas d'information a priori, leurs choix sont motivés par des a priori qui donnent des a posteriori correspondant à des estimations fréquentistes, des a priori qui ont une interprétation attractive ou des a priori permettant une forme analytique pour un a posteriori et parmi les techniques les plus populaire dans la construction de ses lois on peut citer : la loi de Laplace, Jeffrey, ...etc.

Bien souvent, les connaissances subjectives sont relativement vagues et il est difficile de spécifier une loi statistique a priori précise pour les représenter. On dispose, dans ce cas, d'une certaine latitude dans le choix de cette loi a priori, certaines propriétés sont associées à cette loi :

— Le calcul de la densité a posteriori, à partir de la distribution a priori et la distribution de l'échantillon d'observations, doit être aussi simple que possible.

— Les règles de cohérence et de bon sens doivent être respectées.

— La distribution a posteriori doit être du même type que la distribution a priori, afin de permettre un calcul d'actualisation itératif.

Lorsque les paramètres recherchés sont constants, on choisira préférentiellement :

• Une densité a priori uniforme entre deux valeurs extrêmes du paramètre θ_1 et θ_2 pour θ, ou entre 0 et 1. Si les connaissances initiales sont faibles, l'a priori uniforme ne favorise aucune valeur potentielle du paramètre et donne donc le maximum de poids aux observations du retour d'expérience. Par contre, elle varie si l'on fait un changement de variable, ce qui est regrettable.

• Une distribution non informative de Jeffrey, conjuguée à la loi du phénomène observé, lorsque les connaissances a priori sont très faibles et qu'il faut favoriser les données du retour d'expérience plutôt que les connaissances a priori.

Et lorsque les paramètres recherchés sont variables, ils seront modélisés par des densités appropriées : lois de Weibull, Gamma, Beta, dont les paramètres de forme et d'échelle a priori suivront des lois Normale, log-normale ou uniforme, en fonction des connaissances initiales disponibles sur ces paramètres [52].

Chapitre2

Modèles de régression

L'origine du mot régression fut introduit par Sir Francis Galton (1885), un scientifique Britannique du 19éme siècle qui travaillait sur l'hérédité, cet auteur fut un des premiers qui décrivit les liens de dépendance entre certaines variables physiologiques, il chercha à expliquer la taille des fils en fonction de celle des pères. Il constata que lorsque le père était plus grand que la moyenne "taller than medocrity", son fils avait tendance à être plus petit que lui et au contraire que lorsque le père était plus petit que la moyenne "shorter than mediocrity" son fils avait tendance à être plus grand que lui. Les résultats obtenus l'on conduit à sa théorie dite théorie de "régression toward mediocrity".

La régression est l'une des méthodes les plus connues et les plus appliquées en statistiques pour l'analyse des données. Elle est utilisée pour décrire la relation existante entre une variable à expliquer et une ou plusieurs variables explicatives.

Le modèle n'est pas une fin en soi, il doit plutôt aider la prise de décision. Dans l'éminence de la prise de décision on effectue généralement quelques prévisions conditionnelles et on se demande comment évolueront certaines grandeurs controlées si tel événement se produit ou si telle décision doit être prise. C'est à ce moment que les modèles de régression servent puisqu'ils ont pour but d'expliquer la variabilité d'un phénoméne mesurable par d'autres facteurs également mesurables.

Ces modèles sont construits aussi dans le but d'expliquer la variance d'un phénomène (variable dépendante) à l'aide d'une combinaison de facteurs explicatifs (variables indépendantes).

Il existe plusieurs types de régressions paramétriques et non paramétriques, chaque type ayant son importance et ses conditions d'application. Au sein de ce chapitre nous présenterons en détail deux types de modèles paramétriques : la régression linéaire et la régression logistique et enfin nous terminerons par les modèles linéaires généralisés qui présentent l'aspect général des deux types de modèles linéaires et logistiques.

2.1 Cadre général

En introduisant dans cette partie les modèles de régression sous leurs formes générales. Nous nous plaçons dans le cadre de l'estimation d'une fonction de régression notée par m. La variable aléatoire à expliquer sera notée Y, tandis que la variable aléatoire explicative sera notée X. Les différents modèles de régression que nous aborderons peuvent tous être écrit sous la forme :

$$Y = m(X) + \epsilon, \quad m \in \mathcal{C} \tag{2.1}$$

où $\mathcal{C}$ est une classe de fonctions et ϵ est une variable aléatoire indépendante de X.

On se limitera au cas où Y est une variable réelle.

Les modèles de régression se distingueront selon la nature de la variable X (réelle, vectorielle ou fonctionnelle) et selon la nature de la relation m linéaire ou non linéaire.

Definition 2.1.1. Nous dirons que le modèle définit par (2.1) et $m \in \mathcal{C}$ est un modèle paramétrique pour variable réelle lorsque la classe $\mathcal{C}$ est indexable par un nombre fini de paramètres réels. Par opposition nous parlerons de modèles non paramétrique pour variable réelle lorsque $\mathcal{C}$ n'est pas indexable par un nombre fini de paramètres réels.

— Dans le cas où la variable X est réelle ($X \in \mathbb{R}$), le modèle linéaire s'écrit :

$$Y = aX + b + \epsilon,$$

avec a,b $\in \mathbb{R}$ et $\epsilon \sim \mathcal{N}(0, \sigma^2)$

— Dans le cas où la variable X est vectorielle ($X \in \mathbb{R}^p$), le modèle linéaire s'écrit :

$$Y = A^t X + b + \epsilon, \quad A \in \mathbb{R}^p \text{ et } b \in \mathbb{R}$$

Ces deux modèles entrent dans la catégorie des modèles paramétriques, les modèles non paramétriques sont obtenus lorsque nous prenant des hypothèses de régularité sur la fonction m du type :

$\mathcal{C}$=\{fonctions k-fois continument différentiables \}

ou encore l'exemple du modèle additif pour lequel :

$$\mathcal{C} = \{m, m(x_1, ..., x_p) = \mu + m_1(x_1) + ... + m_p(x_p), \mu \in \mathbb{R}, m_i \in \mathcal{C}^0\}$$

où $\mathcal{C}^0 = \{\phi, \int \phi(t)dt = 0\}$ (voir [13]).

Remarque 2.1.1. Dans la régression non paramétrique le modèle s'écrit :

$$m(x) = \mathbb{E}(Y \mid X = x)$$

Où Y est une variable aléatoire réelle et X est une variable aléatoire réelle, avec la condition de régularité sur m, à savoir m est k fois continûment dérivable ($k \in \mathbb{N}$) et peut être estimé par la méthode du noyau [56].

2.2 Modèle linéaire classique simple

On note Y la variable aléatoire réelle à expliquer (variable endogène, dépendante ou réponse) et X la variable explicative ou effet fixe (variable exogène).

Le modèle linéaire simple s'écrit :

$$Y = b_0 + b_1 x + \epsilon.$$

L'écriture matricielle de ce modèle est donnée par :

$$\underset{(n,1)}{Y} = \underset{(n,2)}{X}\ \underset{(2,1)}{b} + \underset{(n,1)}{\epsilon}$$

Avec :

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \quad b = \begin{pmatrix} b_0 \\ b_1 \end{pmatrix} \text{ et } \epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Où (X_i, Y_i), $i = \overline{1.n}$ sont n observations de (X, Y).

ϵ est l'erreur aléatoire.

$b = (b_0, b_1)^t$ est le vecteur des paramètres inconnus (à estimer).

Hypothèses du modèle

$H_1 : \mathbb{E}(\epsilon_i) = 0 \ \ \forall i = \overline{1.n}$ (erreurs centrées).

$H_2 : \mathbb{C}ov(\epsilon_i, \epsilon_j) = 0, \forall (i \neq j)$ (erreurs non corrélées).

$H_3 : \mathbb{E}(\epsilon_i^2) = \sigma_\epsilon^2$, la variance de l'erreur est constante (hypothése d'homoscédasticité).

$H_4 : \epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2 I_n), \quad i = \overline{1.n}.$

✠ Démarche de modélisation

Le processus de modélisation d'un modèle linéaire classique simple passe par plusieurs étapes qui sont définis comme suit :

1. **Estimation des paramètres du modèle**

   Les estimateurs des coefficients obtenus par la méthode MCO qui sont donnés par [47] :

$$\begin{cases} \hat{b}_0 = & \overline{y} - \hat{b}_1 \overline{x} \\ \hat{b}_1 = & \frac{\sum_{i=1}^n x_i y_i - n\overline{xy}}{\sum_{i=1}^n x_i^2 - n\overline{x}^2} \end{cases}$$

   La variance résiduelle estimée est [12] : $\hat{\sigma}_\epsilon^2 = \frac{\sum_{i=1}^n \epsilon_i^2}{n-2}$.

2. **Etablir la qualité d'ajustement**

   La qualité d'ajustement du modèle sera jugée en calculant le coefficient de détermination qui est défini par :

$$R^2 = \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT} = 1 - \frac{\sum_{i=1}^{n} \epsilon_i^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \quad \text{avec} \quad 0 < R^2 < 1.$$

   Avec :

   • SCE $= \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2$ est la variabilité expliquée par la variable X.

   • SCR $= \sum_{i=1}^{n} \epsilon_i^2$ représente la variabilité non expliquée par le modèle ou c 'est la variabilité expliquée par les résidus.

   • SCT $= \sum_{i=1}^{n}(y_i - \overline{y})^2$ représente l'information disponible dans les données.

3. **Tester la réalité de la relation entre Y et les variables exogénes**

   (a) Test de Fisher il permet de vérifier la pertinence globale du modèle. On considére le test :

$$\begin{cases} H_0 : & "b_1 = 0" \\ H_1 : & "b_1 \neq 0" \end{cases}$$

   On utilise la statistique de Fisher suivante :

$$F^* = \frac{\frac{SCE}{1}}{\frac{SCR}{(n-2)}} = \frac{\frac{R^2}{1}}{\frac{(1-R^2)}{(n-2)}} \sim F_{(1,n-2,\alpha)}.$$

   Si $F^* \leq F_{(1,n-2,\alpha)}$, alors on accepte H_0 et on dira que le modèle est globalement mauvais.

   Si $F^* > F_{(1,n-2,\alpha)}$, on rejette H_0 et on dira que le modèle est globalement bon.

   (b) Test de student il permet de vérifier la pertinence des variables prises individuellement. On considére le test :

$$H_0 : "b_i = 0" \quad \text{contre} \quad H_1 : "b_i \neq 0" \quad \text{pour} \quad i \in \{0, 1\}$$

   On compare la valeur calculée $t^*_{b_i}$ à la valeur tabulée $t_{\frac{\alpha}{2}}$, tel que :

   Si $t^*_{\hat{b}_0} = |\frac{\hat{b}_0}{\hat{\sigma}_{\hat{b}_0}}| > t_{(\frac{\alpha}{2},n-2)}$, on rejette H_0, la variable x_i contribue dans l'explication de Y.

   Si $t^*_{\hat{b}_1} = |\frac{\hat{b}_1}{\hat{\sigma}_{\hat{b}_1}}| \leq t_{(\frac{\alpha}{2},n-2)}$, on accepte H_0, la variable x_i ne contribue pas dans l'explication de Y.

4. **Intervalle de confiance** Les intervalles de confiances de b_0 et b_1 respectivement seront donnés par :

$$IC_{b_0} = [\hat{b}_0 - t_{\frac{\alpha}{2}}(n-2)\hat{\sigma}_{\hat{b}_0}; \hat{b}_0 + t_{\frac{\alpha}{2}}(n-2)\hat{\sigma}_{\hat{b}_0}]$$
$$IC_{b_1} = [\hat{b}_1 - t_{\frac{\alpha}{2}}(n-2)\hat{\sigma}_{\hat{b}_1}; \hat{b}_1 + t_{\frac{\alpha}{2}}(n-2)\hat{\sigma}_{\hat{b}_1}]$$

★ **Exemple** [voir Annexe p 88-89]

2.3 Modèle linéaire classique multiple

On considére un couple de v.a $(\underline{X}, Y)$ où Y est une v.a à valeurs réelles et $\underline{X}$ est un vecteur de variables à valeurs dans $\mathbb{R}^p$. Nous considérons un échantillon $(Y_i, X_{i1}, ..., X_{ip})$ de taille n de valeurs $(y_i, x_{i1}, ..., x_{ip})$.

Le modèle linéaire multiple est une généralisation du modèle linéaire simple qui est défini par :

$$Y = b_0 + b_1 x_1 + b_2 x_2 + ... + b_p x_p + \epsilon.$$

L'écriture matricielle de ce modèle est donnée par :

$$\underset{(n,1)}{Y} = \underset{(n,p+1)}{X} \ \underset{(p+1,1)}{b} + \underset{(n,1)}{\epsilon} \tag{2.2}$$

Avec :

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} ; X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n2} & \cdots & x_{np} \end{pmatrix} \quad b = \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_p \end{pmatrix} \text{ et } \epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Hypothèses du modèle :

H_1 : Les variables explicatives sont observées sans erreurs (non aléatoires).

H_2 : $\mathbb{E}(\epsilon_i) = 0$, pour $i = \overline{1.n}$, erreurs centrées.

H_3 : $\mathbb{E}(\epsilon_i^2) = \sigma_\epsilon^2$, la variance de l'erreur est constante (hypothèse d'homoscédasticité).

H_4 : $\mathbb{E}(\epsilon_i \epsilon_j) = 0 \ \ \forall i \neq j$, les erreurs sont non corrélées.

H_5 : $\mathbb{C}ov(\epsilon_i, x) = 0$, pour $i = \overline{1.n}$, l'erreur est indépendante des variables explicatives.

H_6 : $\epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2 I_n)$, $i = \overline{1.n}$.

Si l'une ou plusieurs de ces hypothèses ne sont pas vérifiées (de même pour le cas simple) donc on ne peut pas appliquer le modèle.

✠ Démarche de modélisation

1. **Estimation des paramètres du modèle**

   L'estimateur des MCO des coefficients du modèle s'écrit [47] :

   $$\hat{b} = (X^t X)^{-1} X^t Y$$

   L'estimateur sans biais des moindres carrés de la variance des résidus σ_ϵ^2 est défi ni comme suit [12] :

   $$\hat{\sigma}_\epsilon^2 = \frac{||Y - X\hat{b}||}{n - p - 1} = \frac{\sum_{i=1}^n \epsilon_i^2}{n - p - 1}.$$

   Par conséquent, l'estimateur de la matrice de variance-covariance des $\hat{b}$ est donné par [47] :

   $$\hat{\Omega}_{\hat{b}} = \hat{\sigma}_\epsilon^2 (X^t X)^{-1}.$$

2. **Etablir la qualité d'ajustement**

   La qualité d'ajustement du modèle sera jugée en calculant le coefficient de détermination qui est défini par :

$$R^2 = \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT} = 1 - \frac{\sum_{i=1}^{n} \epsilon_i^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \quad \text{avec} \quad 0 < R^2 < 1.$$

3. **Tester la réalité de la relation entre Y et les variables exogénes**

   (a) Test de Fisher il permet de vérifier la pertinence globale du modèle. On considére le test :

$$\begin{cases} H_0 : "b_1 = b_2 = ... = b_p = 0" \\ \text{contre} \\ H_1 : \exists\ j = 1, ..., p, \quad b_j \neq 0 \end{cases}$$

   Ce test repose sur la statistique de décision suivante :

$$F^* = \frac{\frac{SCE}{p}}{\frac{SCR}{(n-p-1)}} = \frac{\frac{R^2}{p}}{\frac{(1-R^2)}{(n-p-1)}} \sim F_{(p,n-(p+1))}.$$

   Si $F^* > F_{(p,n-p-1,\alpha)}$, alors on rejette l'hypothése nulle H_0 donc on accepte l'hypothése alternative H_1, le modèle est globalement significatif. D'où il y a au moins une variable explicative significative de Y.

   Si $F^* \leq F_{(p,n-p-1,\alpha)}$, on accepte H_0, alors le modèle est rejeté. D'où, il n' y a aucune variable explicative significative de Y.

   (b) Test de student il permet de vérifier la pertinence des variables prises individuellement. On considére le test :

$$H_0 : "b_j = 0" \quad \text{contre} \quad H_1 : "b_j \neq 0", \quad \text{pour} \quad j = 0, ..., p$$

   On compare la statistique de décision $t^* = |\frac{\hat{b}_j}{\hat{\sigma}_{\hat{b}_j}}|$, à $t_{(\frac{\alpha}{2},n-p-1)}$ lue sur la table de student à (n-p-1) ddl, tel que :

   Si $t^* > t_{(\frac{\alpha}{2},n-p-1)}$, alors on rejette H_0, la variable x_i contribue dans l'explication de Y.

   Si $t^* \leq t_{(\frac{\alpha}{2},n-p-1)}$, alors on accepte H_0, la variable x_i ne contribue pas dans l'explication de Y.

4. **Intervalle de confiance** l'intervalle de confiance pour chaque coefficient de régression est donné par :

$$IC_{b_j} = [\hat{b}_j - \hat{\sigma}_{\hat{b}_j} t_{(\frac{\alpha}{2},n-p-1)}; \hat{b}_j + \hat{\sigma}_{\hat{b}_j} t_{(\frac{\alpha}{2},n-p-1)}]$$

★ **Exemple** [voir Annexe p 89-90]

2.4 Limites du modèle linéaire

Le modèle linéaire n'est pas adapté lorsque l'une ou plusieurs hypothèses de base ($H_1 - H_6$) ne sont pas vérifiées, il devient limité aussi dans les cas suivants :

— Le modèle linéaire ne permet pas de modéliser des variables de réponse discrètes (exp : y=0 ou y=1, comptage, ...)[36].

En effet, l'hypothèse $Y \sim \mathcal{N}(Xb, \sigma^2)$ est très contraignante car en pratique on a souvent recours à d'autres lois.

— Le modèle linéaire ne permet pas aussi de tenir compte d'une relation non linéaire entre la variable de réponse Y et les variables explicatives. Cette contrainte de linéarité entraine en particulier qu'on ne peut pas imposer de borne à l'espérance de la variable à expliquer [36].

— Les hypothèses de base sur le terme d'erreur sont par ailleurs des hypothéses fortes, pas toujours réalistes ainsi dans de nombreuses situations il n'est pas réaliste de définir une valeur unique σ^2 pour les erreurs de toutes les observations, par exemple lorsque certaines observations sont moins précises que d'autres, les erreurs peuvent également être corrélées plutôt qu'indépendantes. Il est alors nécessaire de faire d'autres hypothèses et de définir une loi de probabilité plus complexe incluant un nombre plus élevé de paramètres pour décrire la loi du terme d'erreur [37].

Il est donc nécessaire de s'affranchir de ces trois contraintes afin d'élargir le champ d'application de la régression, ainsi dans cette optique trois extensions principales seront une bonne alternative à savoir *les modèles non linéaires* , *à effets mixtes* et *le modèles linéaire généralisé.*

2.5 Extensions du modèle linéaire

✠ Les Modèles non linéaires

Le modèle non linéaire survient dans le cas où une discipline scientifique spécifie la forme que les données doivent suivre et cette forme est non linéaire, il est généralement utilisé lorsqu'on ne peut pas modéliser de manière adéquate la relation avec des paramètres linéaires [55].

La différence fondamentale entre le modèle non linéaire et linéaire tient aux formes fonctionnelles acceptables du modèle. De façon plus spécifique, le modèle linéaire requiert des paramètres linéaires ce qui n'est pas le cas du modèle non linéaire.

Les modèles non linéaires ont été appliqué à un large éventail de situations grace à leurs principaux avantages qui sont : la parcimonie, l'interprétabilité et la prédiction. Plusieurs formes de modèles non linéaires sont explorées dans la littérature.

• **Exemples de modèles non linéaires** :

Il existe de très nombreux exemples de modèles définis par une combinaison non linéaire des paramètres, combinaison qui peut néanmoins être linéarisée. Quelques fonctions classiques sont rassemblées dans le tableau ci-dessous, qui ne donne qu'un très petit aperçu de la variété des situations rencontrées en pratique :

Modèle	**Transformation**	**Modèle linéarisé**
$Y = ae^{bX}$	$Z = ln(Y)$	$Z = ln(a) + bX$
$Y = a + b\ ln(X)$	$H = ln(X)$	$Y = a + bH$
$Y = b(X)^a$	$ln(Y)$	$Y = a\ ln(X) + ln(b)$
$Y = \frac{X}{-a+bX}$	$H = \frac{1}{X}$, $Z = \frac{1}{Y}$	$Z = b - aH$

TABLE 2.1 – Exemples de modèles non linéaires

✠ **Le modèle à effets mixtes**

Le modèle linéaire à effets mixtes est une extension du modèle linéaire qui prend en compte la variabilité liée aux individus. Pour plus de détails voir [22].

✠ **Le modèle linéaire généralisé**

Ce modèle sera abordé de façon détaillé dans la section (2.6).

Pour un traitement plus complet sur la régression linéaire nous renvoyons aux livres [12] et [30].

2.6 Les modèles linéaires généralisés (GLM)

Les modèles linéaires généralisés, communément appelés Generalized Linear Model (GLM) qui est une terminologie introduite par Nelder et Wedderburn (1972), est une classe de modèles permettant la modélisation d'une variable réponse Y dont la loi appartient à la famille exponentielle naturelle (Normale, Poisson, Bernoulli, ...etc), qui a une place importante dans la modélisation statistique. Leurs champs d'applications sont considérablement plus grands que celui des modèles linéaires classiques.

Les GLM sont une extension du modèle linéaire permettant de traiter des observations dont la loi de probabilité appartient à une famille de lois élargie.

Partant du même principe de la régression linéaire, le modèle linéaire classique s'écrit (voir (2.2)) :

$$Y = Xb + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2 I_n),$$

avec $Y \sim \mathcal{N}(Xb, \sigma^2 I_n)$ ce qui conduit à $\mathbb{E}(Y) = Xb$.

Le modèle linéaire généralisé est la donnée d'une loi de probabilité pour Y et d'une fonction g appelée fonction de lien qui sert à généraliser la relation entre $\mathbb{E}(Y)$ et le prédicteur linéaire Xb pour former la relation suivante :

$$g(\mathbb{E}(Y)) = Xb$$

Cela permet d'établir une relation non linéaire entre l'espérance de la variable réponse Y et les variables explicatives et d'envisager des observations de nature variée.

Le cas linéaire est obtenu pour une fonction de lien identité.

2.6.1 Modélisation

La démarche d'écriture d'un modèle linéaire généralisé est constituée de deux étapes :

• Étape 1

Choix d'une loi de probabilité pour la variable Y, cette loi doit appartenir à la famille exponentielle naturelle, qui s'écrit sous la forme suivante :

$$f(y, \omega, \phi) = exp\left\{\frac{y\omega - b(\omega)}{a(\phi)} + c(y, \phi)\right\}, \tag{2.3}$$

- b est une fonction trois fois dérivable et sa dérivée première $b^{'}$ est inversible.
- $a(.)$ et $c(.)$ sont des fonctions dérivables.
- ω est le paramètre naturel lié aux deux premiers moments de la loi.
- ϕ est un paramètre de dispersion ou de nuisance.

Propriétés

Deux propriétés concernant le paramètre naturel ω découlent de l'écriture de la densité pour les lois appartenant à la famille exponentielle [38] :

$$\mathbb{E}(Y) = \mu = b^{'}(\omega) \tag{2.4}$$
$$\mathbb{V}(Y) = b^{''}(\omega)a(\phi) \tag{2.5}$$

Preuve :

f_Y étant la densité de probabilité de Y appartenant à la famille exponentielle naturelle et vérifie que $\int f_Y(y, \omega, \phi) = 1$. Alors, en dérivant par rapport a ω, on obtient :

$$\frac{\partial}{\partial\omega}\int f_y(y, \omega, \phi)dy = \int \frac{\partial}{\partial\omega} f_y(y, \omega, \phi)dy = 0.$$

Donc

$$\begin{aligned}\int \frac{1}{f_Y(y,\omega,\phi)}\frac{\partial}{\partial\omega}f_y(y,\omega,\phi)f_y(y,\omega,\phi)dy=0 &\implies \int \frac{\partial}{\partial\omega}logf_Y(y,\omega,\phi)f_Y(y,\omega,\phi)dy=0\\ &\implies \int \frac{1}{a(\phi)}[y-b^{'}(\omega)]f_Y(y,\omega,\phi)dy=0\\ &\implies \frac{1}{a(\phi)}(\mathbb{E}(Y)-b^{'}(\omega))=0\end{aligned}$$

D'où, $\mathbb{E}(Y)=b^{'}(\omega)$.

En dérivant une seconde fois, on obtient :

$$\frac{\partial}{\partial\omega^2}\int f_Y(y,\omega,\phi)dy= \quad 0$$

Donc

$$\frac{\partial}{\partial\omega}\int\frac{\partial}{\partial\omega}f_Y(y,\omega,\phi)dy=\frac{\partial}{\partial\omega}\int\frac{\partial}{\partial\omega}log\ f_Y(y,\omega,\phi)f_Y(y,\omega,\phi)dy= \quad 0$$

Par conséquent

$$\int\frac{\partial^2}{\partial\omega^2}\left(log\ f_Y(y,\omega,\phi)\right)f_Y(y,\omega,\phi)dy+\int\frac{\partial}{\partial\omega}log\ f_Y(y,\omega,\phi)\frac{\partial}{\partial\omega}f_Y(y,\omega,\phi)dy= \quad A+B=0$$

où,

$$\begin{aligned}A &= \int\frac{\partial^2}{\partial\omega^2}(logf_Y(y,\omega,\phi))f_Y(y,\omega,\phi)dy\\ &= \int\frac{\partial}{\partial\omega}[\frac{1}{a(\phi)}(y-b^{'}(\omega))]f_Y(y,\omega,\phi)dy\\ &= \int\frac{1}{a(\phi)}(-b^{''}(\omega))f_Y(y,\omega,\phi)dy=-\frac{1}{a(\phi)}b^{''}(\omega)\int f_Y(y,\omega,\phi)dy=-\frac{b^{''}(\omega)}{a(\phi)}.\end{aligned}$$

et

$$\begin{aligned}B &= \int\frac{\partial}{\partial\omega}logf_Y(y,\omega,\phi)\frac{\partial}{\partial\omega}f_Y(y,\omega,\phi)dy\\ &= \int\frac{\partial}{\partial\omega}logf_Y(y,\omega,\phi)\frac{\partial}{\partial\omega}log(f_Y(y,\omega,\phi))f_Y(y,\omega,\phi)dy\\ &= \int\left(\frac{\partial}{\partial\omega}logf_Y(y,\omega,\phi)\right)^2f_Y(y,\omega,\phi)dy\\ &= \int(\frac{1}{a(\phi)}(y-b^{'}(\omega))^2f_Y(y,\omega,\phi)dy=\frac{1}{(a(\phi))^2}\int(y-b^{'}(\omega))^2f_Y(y,\omega,\phi)dy=\frac{\mathbb{V}(Y)}{(a(\phi))^2}\end{aligned}$$

D'où, B+A $=0 \Leftrightarrow \frac{\mathbb{V}(Y)}{a(\phi)^2}-\frac{b^{''}(\omega)}{a(\phi)}=0 \Rightarrow \mathbb{V}(Y)=\frac{b^{''}(\omega)a(\phi)^2}{a(\phi)}=b^{''}(\omega)a(\phi)$. Ce qui donne $\mathbb{V}(Y)=b^{''}(\omega)a(\phi)$.

Exemple de la loi gaussienne :

Soit $Y \sim \mathcal{N}(\mu, \sigma^2)$, sa fonction densité est donnée par :

$$\begin{aligned} f(y,\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} &= exp\left\{ log\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(y-\mu)^2}{2\sigma^2}} \right) \right\} \\ &= exp\left\{ log\left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) - \frac{(y-\mu)^2}{2\sigma^2} \right\} \\ &= exp\left[\frac{-y^2}{2\sigma^2} + \frac{2y\mu}{2\sigma^2} + \frac{\mu^2}{2\sigma^2} - log\sqrt{2\pi\sigma^2} \right] \\ &= exp\left[\frac{y\mu - \frac{\mu^2}{2}}{\sigma^2} - \frac{1}{2} log(2\pi\sigma^2) - \frac{y^2}{2\sigma^2} \right]. \end{aligned}$$

Par identification avec la formule (2.3), donc les composantes de la loi seront données par :

$$\omega = \mu, \quad b(\omega) = \frac{\mu^2}{2} = \frac{\omega^2}{2}, \quad a(\phi) = \phi = \sigma^2, \quad c(y,\phi) = \frac{-1}{2} log(2\pi\sigma^2) - \frac{y^2}{2\sigma^2} = \frac{-1}{2} log(2\pi\phi) - \frac{y^2}{2\phi}$$

Le tableau suivant indique les composantes de certaines lois de la famille exponentielle naturelle :

Distribution de Y	ω	ϕ	$a(\phi)$	$b(\omega)$	$c(y,\phi)$
$\mathcal{N}ormale(\mu,\sigma^2)$	μ	σ^2	ϕ	$\frac{\omega^2}{2}$	$-\frac{1}{2}\{\frac{y^2}{\sigma^2} + log(2\pi\sigma^2)\}$
$\mathcal{P}$oisson(μ)	$log(\mu)$	1	ϕ	$exp(\omega)$	$-log\ y!$
$\mathcal{B}$ernouli(y,μ)	$log(\frac{\mu}{1-\mu})$	1	ϕ	$log(1+exp(\omega))$	0
$\mathcal{B}$inomiale(y,n,μ)	$log(\frac{\mu}{1-\mu})$	1	ϕ	$nlog(1+exp(\omega))$	$log\binom{n}{y}$
$\mathcal{G}$amma(μ,α)	$\frac{\alpha}{\mu}$	$\frac{-1}{\mu}$	ϕ	$log(\mu\ \omega)$	$(\mu-1)logy - log(\Gamma(\mu))$

TABLE 2.2 – Tableau des composantes des lois de la famille exponentielle

• Étape 2

Modélisation du lien entre l'espérance de Y et les variables explicatives au travers d'une fonction de lien g inversible, différentiable et monotone :

$$g(\mathbb{E}(Y)) = g(\mu) = Xb.$$

Si Y est une v.a admettant une densité appartenant à la famille exponentielle telle que $\mathbb{E}(Y) = b^{'}(\omega) = \mu$, alors le lien est dit "canonique".

Exemples de fonctions de lien

Les fonctions de lien les plus fréquemment utilisées en médecine, épidémiologie, économie,... etc. sont indiquées dans le tableau suivant [54] :

log	$g(x) = log\ x$
logit	$g(x) = log(\frac{x}{1-x})$
probit	$g(x) = \Phi^{-1}(x)$ où $\Phi(.)$ est la fdr $\mathcal{N}(0,1)$.
complementary log-log	$g(x) = log(-log(1-x))$
log-log	$g(x) = log(-log(x))$

TABLE 2.3 – Exemples de fonctions de lien

Elles sont représentées dans la figure suivante :

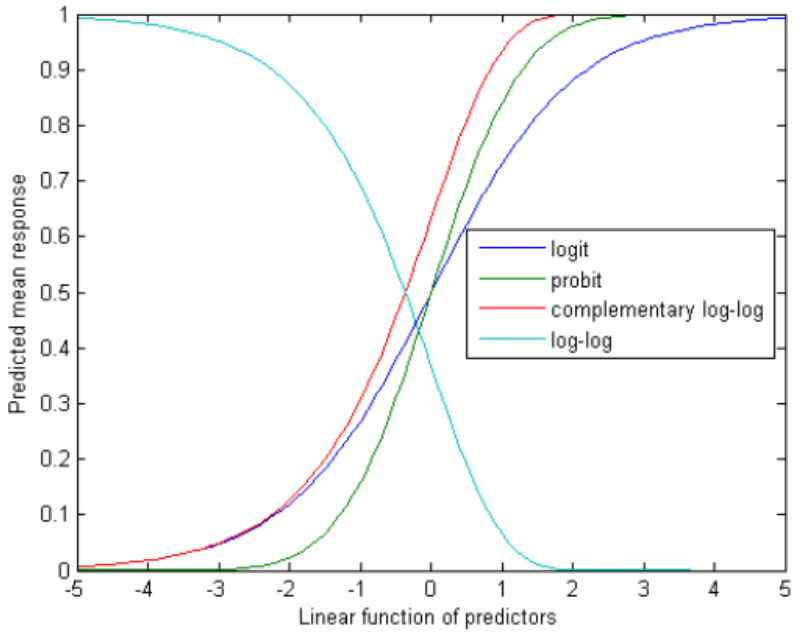


FIGURE 2.1 – Représentation graphique des fonctions de lien

✠ Choix de la fonction de lien

Le choix de la fonction de lien est important car cette fonction permet de s'assurer que les valeurs prédites par le modéle vont restés dans des limites raisonnables, c'est à dire il faut la choisir de sorte qu'elle conduise à la meilleure adéquation entre le modèle et les observations. Le choix est lié directement à la loi de la variable réponse.

(a) Cas de lois de poisson

Si les observations sont des comptages (prennent des valeurs entières non négatives) modélisée par une loi de poisson, il faudra s'assurer que les prédictions du modèle donnent des valeurs positives ou nulles car les valeurs négatives seraient aberrantes dans ce cas. La fonction de lien la plus appropriée sera la fonction "log" car les valeurs prédites par le prédicteur linéaire donnent par transformation inverse des valeurs toutes positives ou nulles qui respecte bien la nature de Y.

(b) Cas de loi bernoulli

Si les observations sont données binaires suivant une loi de Bernoulli. La fonction de lien appropriée sera la fonction "logit" car elle permet de supprimer le risque d'obtenir un modèle qui prédit des probabilités négatives ou plus grandes que 1. Cette fonction de lien est trés utilisée pour ses propriétés plus simples mais surtout pour la simplicité de son interprétation.

En pratique, si aucune raison de choisir une fonction de lien spécifique ne s'impose, le choix par défaut consiste à choisir la fonction de lien naturel, c'est à dire choisir g telle que $g(\mu) = (b^{'})^{-1}(\mu)$, permettant d'assurer la convergence des algorithmes d'estimation vers le MLE.

2.6.2 Estimation des paramètres

Considérons un échantillon $(Y_1, Y_2, ..., Y_n)$ de n variables aléatoires indépendantes et identiquement distribués de réalisations $(y_1, y_2, ..., y_n)$ de densité de probabilité f_{Y_i}, où Y_i est la réponse au point $\underline{x}_i = (x_{i1}, ..., x_{ip}), \quad i = \overline{1.n}$

Si la fonction de lien utilisée est celle du lien naturel, alors :

$$\omega_i = g(\mathbb{E}(Y_i) = x_i\theta.$$

La fonction de vraisemblance de Y s'écrit alors :

$$\mathcal{L}(y, \theta, \phi) = \prod_{i=1}^{n} f(y_i, \omega_i, \phi) = \prod_{i=1}^{n} \left[exp\left\{ \frac{y_i\omega_i - b(\omega_i)}{a(\phi)} + c(y_i, \phi) \right\} \right].$$

La log-vraisemblance du modèle est alors définie comme suit :

$$\begin{aligned} \ell(y, x_i\theta, \phi) &= log(\mathcal{L}(y, x_i\theta, \phi)) \\ &= log\left(\prod_{i=1}^{n} \left[exp\left\{ \frac{y_i x_i\theta - b(x_i\theta)}{a(\phi)} + c(y, \phi) \right\} \right] \right) \\ &= \sum_{i=1}^{n} \frac{y_i x_i\theta - b(x_i\theta)}{a(\phi)} + c(y_i, \phi) \end{aligned}$$

Les paramètres à estimer sont θ et de ϕ qui rendent maximale cette fonction de log-vraisemblance sont solutions du système d'équations aux dérivées partielles suivant :

$$\begin{cases} \frac{\partial}{\partial \theta_j}\ell(y,\theta,\phi) = 0 & \text{pour } \ j = 1,...p \\ \frac{\partial}{\partial \phi}\ell(y,\theta,\phi) = 0 \end{cases}$$

Pour simplifier l'écriture, on pose $a(\phi) = 1$ dans cette partie, ce qui ne change pas les résultats obtenus. On a alors :

$$\frac{\partial \ell(y,\theta,\phi)}{\partial \theta_j} = \sum_{i=1}^{n} x_i^j [y_i - b^{'}(x_i\theta)], \quad j = 1,...,p.$$

Et donc $\sum_{i=1}^{n} x_i[y_i - b^{'}(x_i\theta)] = 0$.

Pour tous les autres modèles linéaires généralisés, ce système à p équations est un systéme non linéaire en θ et il n'y a pas d' expression explicite pour les estimateurs. Pour obtenir les estimations du maximum de vraisemblance, on a recours à des algorithmes d'optimisation itératifs. Les deux algorithmes les plus utilisés sont l'algorithme de Newton-Raphson et l'algorithme du Fisher-scoring.

2.6.3 Propriétés de l'estimateur MV

Notons $\hat{\theta}_n$ l'estimateur du maximum de vraisemblance (EMV). Cet estimateur vérifie les propriétés suivantes [2] :

Théorème 2.6.1. *Sous certaines conditions de régularité de la densité de probabilité, l'EMV possède les propriétés suivantes :*

1. *$\hat{\theta}_n$ converge en probabilité vers θ (ce qui implique que $\hat{\theta}_n$ est asymptotiquement sans biais).*

2. *$\hat{\theta}_n$ converge en loi vers une loi gaussienne*

$$\sqrt{I_n(\theta,\phi)}(\hat{\theta}_n - \theta) \sim \mathcal{N}(0, Id),$$

où $I_n(\theta,\phi) = -\mathbb{E}\left[\frac{\partial^2 \ell(y,\theta,\phi)}{\partial^2 \theta}\right]$ est la matrice d'information de Fisher évaluée en θ et ϕ (vraie valeur des paramètres) sur un échantillon de taille n quand le domaine ne dépend pas de θ

Preuve voir [2].

2.6.4 Tests d'hypothèses

Dans cette section nous présentons l'idée générale des tests d'hypothéses dans le cas des GLM qui vont permettre de déterminer si les variables explicatives présentent dans le modèle sont pertinentes ou non. Pour cela les tests de modèles emboités sont les plus généraux et permettent de répondre à la majorité des questions qui se posent. Ces tests permettent de déterminer si un sous-ensemble de variables explicatives suffit à expliquer la variable Y .

On dira que les modèles M_0 et M_1 sont emboitès s'ils ont la même distribution de probabilité et la même fonction de lien mais la composante linéaire du modèle M_0 est un cas particulier de la composante linéaire du modèle plus général M_1. Deux tests basés sur le principe des modèles emboités sont : le test de maximum de vraisemblance et celui de Wald [2].

2.6.5 Qualité d'ajustement

Dans les modèles linéaires classiques une mesure globale d'ajustement est fournie par le coefficient de détermination R^2. Pour les modèles linéaires généralisés deux mesures d'ajustement sont disponibles : *La déviance* et *la statistique du khi-deux de Pearson.*

(a) La déviance

Le modèle estimé (null) ℓ, qui est le modèle à un paramètre est comparé avec le modèle dit saturé ℓ_{sat}, c'est-à-dire le modèle à n paramètre ou chaque moyenne de Y_i est remplacée par Y_i. Cette comparaison est basée sur l'expression de la déviance D des logs-vraisemblances ℓ et ℓ_{sat} des deux modèles null et saturé (respectivement) :

$$D = -2\{\ell - \ell_{sat}\},$$

Rappelant que $\omega_i = g(\mathbb{E}(Y_i))$, et par définition du modèle saturé on a chaque moyenne de Y_i est remplcé par Y_i, c 'est à dire $\mathbb{E}(Y_i)$ sera remplacé par Y_i.

On aura donc $\ell_{sat} = \sum_{i=1}^{n} \frac{y_i g(Y_i) - b(g(Y_i))}{a(\phi)} + c(y_i, \phi)$.

Approximativement D suit une khi-deux à $(n-p)$ degré de liberté ($D \sim \chi^2_{(n-p)}$)

- Si $D > \chi^2_{(1-\alpha,n-p)}$, alors le modèle est globalement mauvais.
- Si $D \leq \chi^2_{(1-\alpha,n-p)}$, alors le modèle est globalement bon.

L'objectif, lors de l'ajustement d'un GLM, sera de minimiser D.

(b) La statistique du khi-deux de Pearson généralisé

Une deuxième mesure qui a été proposée pour juger de la qualité de l'ajustement est la statistique du khi-deux de Pearson, défini par :

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\mathbb{V}(\hat{\mu}_i)}, \quad \hat{\mu}_i = g^{-1}(x_i\hat{\theta})$$

où Approximativement χ^2 suit une khi-deux à $(n-p)$ degré de liberté ($\chi^2 \sim \chi^2_{(n-p)}$)

- Si $\chi^2 > \chi^2_{(1-\alpha,n-p)}$ alors le modèle est globalement mauvais, il est donc rejeté.
- Si $\chi^2 \leq \chi^2_{(1-\alpha,n-p)}$, alors le modèle est globalement bon.

Remarque 2.6.1. Le test de Pearson possède des propriétés asymptotiques différentes de la déviance et peut conduire à des résultats différents.

2.7 Régression logistique

La régression logistique est une technique de modélisation mathématique qui peut-être utilisée pour décrire la relation de plusieurs variables dépendantes qualitatives ou quantitatives. Elle constitue un cas particulier du modèle linéaire généralisé (GLM).

La régression élaborée en 1944 par Joseph Berkson permet de discriminer une variable réponse Y binaire à partir d'une matrice de p variables explicatives [3], ensuite elle a été développée par le statisticien David Cox en 1958, elle est devenue très populaire. Selon la science citation Index, plus de 2500 articles ont été publié en 1999 dans lesquels les mots "régression logistique" sont apparues. En 2004 Pohar et al soutiennent que la régression logistique est une méthode robuste, flexible et facile à utiliser, le but est de trouver le meilleur modèle et le plus parcimonieu pour décrire la relation entre la variable dépendante et les variables indépendantes [43]. Une année après, Desjardin l'a défini comme étant une technique permettant d'ajuster une surface de régression à des données lorsque la variable dépendante est dichotomique [11]. Hasti et al (2009) pour eux la régression logistique est un modèle plus robuste que la régression linéaire peu importe le type de distribution des données [26].

Le succès de cette approche repose surtout sur les nombreux outils qui permettent d'interpréter de manière approfondie les résultats obtenus.

✠ Cadre d'application

La régression logistique est largement répandue dans des domaines très variés. On peut citer de façon non exhaustive :

— En médecine, elle permet par exemple de trouver les facteurs qui caractérisent un groupe de sujets malades par rapport à des sujets sains.

— En assurances, elle permet de cibler une fraction de la clientèle qui sera sensible à une police d'assurance sur tel ou tel risque particulier.

— Dans le domaine bancaire, pour détecter les groupes à risque lors de la souscription d'un crédit.

— En économétrie, pour expliquer une variable discrète. Par exemple, les intentions de vote aux élections.

De nombreux autres types d'applications ont été décrits dans la littérature, par exemple White, Pearson et Wilson (1999) ont examiné la mise en oeuvre du Just à temps pratiques de fabrication utilisant des modèles de régression logistique [57], Palma, Beja et Rodrigues (1999) ont modélisé des observations de Lynx [42] et dans une application particulièrement contemporaine de Hu et Heisey (1991) qui ont utilisé la régression logistique pour prédire " la valeur de cache" des objets sur le World Wide Web [15].

Il est évident que la régression logistique est le domaine des praticiens plutôt que des statisticiens. Soit moins de 1% du grand nombre d'articles sur la régression logistique apparaissent dans les revues statistiques [53].

✠ Types de régression logistique

Différents types de régression logistique existent, possédant chacun leur procédé statistique et conduisant à l'élaboration de différents modèles théoriques, on va se limiter à trois types les plus utilisés en pratique qui sont [20] :

• ***La régression logistique binaire***

Ce type correspond au cas où la variable réponse Y est du type binaire.

• ***La régression logistique polytomique nominale***

Elle est utilisée pour modéliser la relation entre une variable dépendante nominale qui présente plus de deux modalités et il n'y a pas de relation d'ordre naturel entre ces modalités. Par exemple, un biologiste peut être intéressé par les choix alimentaires des alligators (une famille de crocodiliens). Les alligators adultes peuvent avoir des préférences différentes des jeunes, la variable de résultat ici sera les types d'aliments et les variables prédictives pourraient être la taille des alligators et d'autres variables environnementales.

• ***La régression logistique polytomique ordinale***

Ce type de régression concerne les situations où la variable Y présente plus de deux modalités ordonnées. Un exemple typique est la description de l'intensité de l'attaque d'individus par un parasite, cette description étant réalisée par exemple sur la base d'une échelle à quatre niveaux notés A, B, C et D tel que A représente l'absence d'attaque, B attaque faible, C attaque modérée et D attaque forte.

✠ Les limites de la régression logistique

1. La régression logistique nécessite des échantillons de grande taille pour pouvoir atteindre un bon niveau de stabilité.

2. Les catégories auxquelles appartiennent les variables indépendantes doivent être mutuellement exclusives, car il s'agit d'une variable dichotomique.

3. Examiner les corrélations entre les prédicteurs avant de procéder à l'élaboration du modèle. Lorsque certains prédicteurs sont fortement corrélés entre eux, il est préférable d'en éliminer quelques-uns puisqu'il s'agit des variables redondantes.

4. La régression logistique présente la limite de ne présumer que les réponses non reliées.

Remarque 2.7.1. Le modèle de régression logistique traité dans ce mémoire est adapté au cas binaire.

2.7.1 Modèle logistique binaire

Soit Y une variable à expliquer à valeurs dans $\{0,1\}$ et $\underline{X} = (X_1, ..., X_p)^t \in \mathbb{R}^p$ le vecteur des variables explicatives de valeur $\underline{x} = (x_1, ..., x_p)^t$ concourant à l'explication de Y.

Comme Y est une variable binaire, alors elle suit une loi de Bernoulli de probabilité $p(\underline{x}) = \mathbb{E}(Y \mid \underline{X} = \underline{x})$. On a dans le cas des modèles GLM la relation suivante :

$$g(\mathbb{E}(Y \mid \underline{X} = \underline{x})) = \underline{x}^t \beta,$$

où $\beta = (\beta_0, \beta_1, ..., \beta_p)^t$ est le vecteur des paramètres.

Dans le cas d'une variable binaire quatre fonctions de lien sont considérées dans la littérature pour tenir compte de ce système (voir tableau 2.3).

On utilise pour la fonction de lien g, la fonction "logit" pour ses propriétés théoriques plus simples mais surtout pour la simplicité de son interprétation [7]. La fonction de lien log-log est adaptée au cas où l'on considère qu'il y a une asymétrie entre les probabilités de succès et d'échec, et du coté de faciliter la fonction logit et plus simple à manipuler que la fonction probit.

D'où, le modèle logistique binaire (modèle logit) est donné par :

$$g(\mathbb{E}(Y \mid \underline{X} = \underline{x})) = logit(p(\underline{x}))$$

tel que :

$$logit(p(\underline{x})) = log\left(\frac{p(\underline{x})}{1 - p(\underline{x})}\right) = \beta_0 + \sum_{j=1}^{p} \beta_j x_j = \underline{x}^t \beta, \tag{2.6}$$

où $\beta_0, \beta_1, ..., \beta_p$ sont les paramètres réels inconnus à estimer.

Avec OR $= \frac{p(\underline{x})}{(1-p(\underline{x}))}$ est appelé le **rapport de chance (Odds Ratio)** qui est le rapport entre la probabilité de succès $p(\underline{x}) = \mathbb{P}(Y = 1 \mid \underline{X} = \underline{x})$ et la probabilité d'échec $1 - p(\underline{x}) = \mathbb{P}(Y = 0 \mid \underline{X} = \underline{x})$.

• OR< 1, indique une influence négative de X sur Y, c'est à dire il correspond à une diminution du phénomène étudié.

• OR $= 1$, signifie l'absence d'effet entre X et Y.

• OR>1, indique une influence positive de X sur Y, c'est à dire il correspond à une augmentation du phénomène étudié.

Puisque la fonction de lien (logit) est inversible, alors la fonction inverse est :

$$logit^{-1}\left(log\left(\frac{p(\underline{x})}{1 - p(\underline{x})}\right)\right) = p(\underline{x}) = \mathbb{P}(Y = 1 \mid \underline{X} = \underline{x}) = \mathbb{E}(Y \mid \underline{X} = \underline{x}).$$

D'un autre coté, à partir de (2.6) :

$$\begin{aligned} log\left(\frac{p(\underline{x})}{1-p(\underline{x})}\right) &= \underline{x}^t\beta, \quad \underline{x} = (1, x_1, ..., x_p) \\ \Longrightarrow \frac{p(\underline{x})}{1-p(\underline{x})} &= e^{\underline{x}^t\beta} \\ \Longrightarrow p(\underline{x}) &= \frac{e^{\underline{x}^t\beta}}{1+e^{\underline{x}^t\beta}} = \frac{exp[logit(p(\underline{x}))]}{1+exp[logit(p(\underline{x}))]} \end{aligned}$$

D'où, $logit^{-1}\left(log\left(\frac{p(\underline{x})}{1-p(\underline{x})}\right)\right) = \frac{e^{\underline{x}^t\beta}}{1+e^{\underline{x}^t\beta}} = p(\underline{x}), \quad 0 < p(\underline{x}) < 1.$

où La probabilité $p(\underline{x})$ est appelée la fonction logistique (ou sigmoide).

La figure suivante représente la fonction logistique :

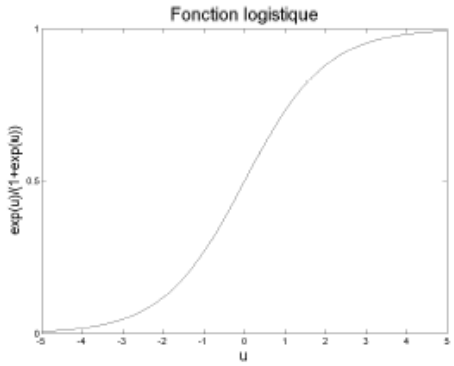


FIGURE 2.2 – La Fonction logistique

Hypothèses

• **Résidus non normaux**

Comme la variable réponse Y est du type binaire donc, chaque $\epsilon = Y - x^t\beta$ (erreur aléatoire) ne peut prendre que deux valeurs possibles, c'est-à-dire [30] :

$$\epsilon = \begin{cases} 1 - \underline{x}^t\beta & si \quad Y = 1 \\ -\underline{x}^t\beta & si \quad Y = 0. \end{cases}$$

Ainsi ϵ admet nécessairement une loi discrète ce qui exclut en particulier l'hypothèse de normalité des résidus.

• **Variance non constante**

Étant donné que la variable Y est de Bernoulli :

$$\begin{cases} \mathbb{E}(Y=1 \mid \underline{X}=\underline{x}) = & \mathbb{P}(Y=1 \mid \underline{X}=\underline{x}) \\ \mathbb{V}(Y=1 \mid \underline{X}=\underline{x}) = & \mathbb{P}(Y=1 \mid \underline{X}=\underline{x})(1-\mathbb{P}(Y=1 \mid \underline{X}=\underline{x})) \end{cases}$$

Ce qui implique que la variance n'est pas une constante et varie selon x.

2.7.2 Estimation des paramètres

La méthode d'estimation du maximum de vraisemblance, est la méthode standard utilisée par les statisticiens pour estimer les paramètres d'un modèle logistique. Cependant deux autres méthodes peuvent être utilisés qui sont la méthode des moindres carrés pondérés non itératifs et l'analyse de fonction discriminante.

On dispose de n observations $\underline{x}_i = (x_{i1}, ..., x_{ip})^t$, $\quad y \in \{0,1\}$, avec $Y_i \mid \underline{X}_i = \underline{x}_i \sim \mathcal{B}(p(\underline{x}_i))$ où Y_i est une variable endogène, pour $i = \overline{1.n}$.

La fonction de masse conditionnelle de $Y \mid \underline{X} = \underline{x}$ s'écrit :

$$\mathbb{P}(y_i \mid \underline{X}_i = \underline{x}_i) = p(\underline{x}_i)^{y_i}(1-p(\underline{x}_i))^{1-y_i}, \quad \text{avec} \quad p(\underline{x}_i) = \frac{e^{logit(p(\underline{x}_i))}}{1+e^{logit(p(\underline{x}_i))}}, \quad i = \overline{1.n}$$

D'où, la fonction de vraisemblance sera donnée par :

$$\begin{aligned} \mathcal{L}(\beta, \underline{y}) &= \prod_{i=1}^{n} \mathbb{P}(Y_i = y_i \mid \underline{X_i} = \underline{x_i}), \quad \underline{y} = (y_1, ..., y_n)^t \\ &= \prod_{i=1}^{n} p(\underline{x}_i)^{y_i}(1-p(\underline{x}_i))^{1-y_i} \\ &= \prod_{i=1}^{n} \left(\frac{e^{logit(p(\underline{x}_i))}}{1+e^{logit(p(\underline{x}_i))}}\right)^{y_i} \left(1 - \frac{e^{logit(p(\underline{x}_i))}}{1+e^{logit(p(\underline{x}_i))}}\right)^{1-y_i} \\ &= \prod_{i=1}^{n} \left(\frac{e^{logit(p(\underline{x}_i))}}{1+e^{logit(p(\underline{x}_i))}}\right)^{y_i} \left(\frac{1}{1+e^{logit(p(\underline{x}_i))}}\right)^{1-y_i} \\ &= \prod_{i=1}^{n} \frac{e^{logit(p(\underline{x}_i))y_i}}{(1+e^{logit(p(\underline{x}_i))})^{y_i+1-y_i}} \\ \Rightarrow \mathcal{L}(\beta, \underline{y}) = \prod_{i=1}^{n} \left(\frac{e^{logit(p(\underline{x}_i))^{y_i}}}{1+e^{logit(p(\underline{x}_i))}}\right) &= \prod_{i=1}^{n} \left(\frac{exp(\underline{x}_i^t \beta y_i)}{1+exp(\underline{x}_i^t \beta)}\right). \end{aligned}$$

Pour faciliter les manipulations, on préfére travailler sur la log vraisemblance qui est donnée par :

$$\begin{aligned}\ell(\beta, \underline{y}) &= log\ \mathcal{L}(\beta, \underline{y}) \\ &= \sum_{i=1}^{n} log\left(\frac{e^{\underline{x}_i^t \beta y_i}}{1+e^{\underline{x}_i^t \beta}}\right) \\ &= \sum_{i=1}^{n}\left[\underline{x}_i^t \beta y_i - log(1+e^{\underline{x}_i^t \beta})\right].\end{aligned}$$

D'où,

$$\begin{aligned}\frac{\partial \ell}{\partial \beta_r} &= \sum_{i=1}^{n}\left(x_{ir}y_i - \frac{x_{ir}e^{\underline{x}_i^t \beta}}{1+e^{\underline{x}_i^t \beta}}\right), \quad r = 0, ...p \\ &= \sum_{i=1}^{n} x_{ir}\left[y_i - \frac{e^{\underline{x}_i^t \beta}}{1+e^{\underline{x}_i^t \beta}}\right] \\ &= \sum_{i=1}^{n} x_{ir}\left[y_i - p(\underline{x}_i)\right].\end{aligned}$$

L'estimateur du maximum de vraisemblance $\hat{\beta} = \underset{\beta}{\text{argmax}}\ \ell(\beta, \underline{y})$ est obtenu en résolvant le système suivant :

$$\begin{cases} \frac{\partial \ell}{\partial \beta_r} = & \sum_{i=1}^{n} x_{ir}(y_i - p(\underline{x}_i)) = 0, \quad r = 1, .., p \\ \frac{\partial^2 \ell}{\partial \beta_r \partial \beta_k} = & -\sum_{i=1}^{n} x_{ir}x_{ik}\frac{e^{\underline{x}_i^t \beta}}{(1+e^{\underline{x}_i^t \beta})^2} < 0, \quad r, k = 1, .., p. \end{cases}$$

La résolution de ce système n'est pas possible analytiquement. Elle exige alors l'utilisation des méthodes d'optimisation numériques (itératives) notamment l'algorithme de Newton-Raphson et Mak (1993) [39] qui permettent de résoudre ce type de problème.

Algorithme de Newton-Raphson

Nous considérons le vecteur gradient $\nabla \ell = \frac{\partial \ell}{\partial \beta}$ et la matrice hessienne $\nabla^2 \ell = \left(\frac{\partial^2 \ell}{\partial \beta_i \partial \beta_j}\right)_{i,j}$

On peut montrer que [27] :

$$\nabla \ell = \underline{X}^t\left[\underline{Y} - \mathbb{E}(Y \mid \underline{X}, \beta)\right] \tag{2.7}$$

$$\nabla^2 \ell = -\underline{X}^t W X \tag{2.8}$$

où $W = diag\left\{\frac{e^{\underline{X}_i^t \beta}}{(1+e^{\underline{X}_i^t \beta})^2}\right\}_i$.

On présente la solution $\beta^{(k)}$, donnée par :

$$\beta^{(k)} = \beta^{(k-1)} - (\nabla^2 \ell_{k-1})^{-1} \nabla \ell_{k-1},$$

avec $\beta^{(k-1)}$ est la valeur de β à l'itération $(k-1)$ et $\nabla \ell_{k-1} = \underline{X}^t[\underline{Y} - \mathbb{E}(Y \mid X, \beta^{(k-1)})]$ le vecteur gradient évalué en $\beta^{(k-1)}$.

Lorsque nous combinons les équations (2.7) et (2.8) nous obtenons :

$$\beta^{(k)} = \beta^{(k-1)} + (\underline{X}^t W^{(k-1)} X)^{-1} \underline{X}^t [\underline{Y} - \mathbb{E}(Y \mid \underline{X}, \beta^{(k-1)})],$$

$W^{(k-1)}$ et $\mathbb{E}(Y \mid X, \beta^{(k-1)})$ sont respectivement la matrice diagonale W et le vecteur des espérance $\mathbb{E}(Y \mid \underline{X}, \beta)$ évalué en $\beta^{(k-1)}$.

On peut résumer l'algorithme en 2 étapes :

1. Le choix d'un point de départ $\beta^0 = (0, ..., 0)$.

2. Tant que le critére de convergence n'est pas vérifié ($||\beta^{(k)} - \beta^{(k-1)}|| < \epsilon$) :

   — Evaluer le vecteur gradient $\nabla \ell$ et la matrice hessienne $\nabla^2 \ell$ en $\beta^{(k-1)}$.

   — Mettre à jour : $\beta^{(k)} = \beta^{(k-1)} + (\underline{X}^t W^{(k-1)} X)^{-1} \underline{X}^t [\underline{Y} - \mathbb{E}(Y \mid \underline{X}, \beta^{(k-1)})]$.

Nous avons une divergence de l'algorithme de Newton Raphson lorsque la matrice hessienne $\nabla^2 \ell$ est impossible à calculer.

★ Propriétés asymptotiques de l'estimateur $\hat{\beta}$:

Pour présenter les propriétés asymptotiques de l'estimateur de maximum de vraisemblance $\hat{\beta}$ du paramètre β dans le modèle de régression logistique, nous supposons que les hypothèses suivantes sont vérifiées [21] :

1. H_1 : Les variables exogènes sont uniformément bornées , c-à-d $\exists\ M < \infty : |X| \leq M$

2. H_2 : Soit λ_{1n} et λ_{pn} les valeurs propres respectivement minimale et maximale de la matrice $\underline{X}^t W X$. Alors il existe une constante $K < \infty$, telle que $\frac{\lambda_{pn}}{\lambda_{1n}} \leq K$ pour tout n.

Théorème 2.7.1. (*Existence et consistance*) *Sous les hypothèses H_1 et H_2 l'estimateur du maximum de vraisemblance noté $\hat{\beta}$ de β existe presque sûrement quand n tend vers $+\infty$, et $\hat{\beta}$ converge presque sûrement quand n tend vers $+\infty$ vers la vraie valeur β_* si et seulement si $\lim\limits_{n \to +\infty} \lambda_{1n} = +\infty$.*

Démonstration. Voir [21] □

Théorème 2.7.2. (*Normalité asymptotique*) *Sous les hypothèses H_1 et H_2 et si l'estimateur de vraisemblance $\hat{\beta}$ converge asymptotiquement vers β_* alors*

$$\sqrt{n}(\hat{\beta}_n - \beta_*) \to \mathcal{N}(0, \phi(\beta_*)^{-1}), \ \textit{quand} \ \ n \to +\infty$$

où $\phi(\beta_*) = -\mathbb{E}\left(\nabla^2 \ell(\beta_*, \underline{y})\right)$ est la matrice de l'information de Fisher.

Démonstration. Voir [21] □

2.7.3 Les mesures d'ajustement

★ Les pseudos-R^2

Dans un modèle linéaire, le coefficient de détermination R^2 mesure la qualité d'ajustement du modèle et quantifie le pourcentage de variation totale expliquée par le modèle. Contrairement au modèle linéaire, le coefficient de détermination dans le modèle logistique ne mesure pas la proportion de variance expliquée mais plutôt une amélioration du modèle complet (contenant tous les prédicteurs) par rapport au modèle nul (ne contenant que la constante du modèle).

$\mathcal{L}$ et ℓ représente la fonction de vraisemblance et log vraisemblance pour le modèle contenant tous les prédicteurs respectivement.

$\mathcal{L}_0$ et ℓ_0 représente la fonction de vraisemblance et log vraisemblance pour le modèle contenant uniquement la constante respectivement.

En effet, plusieurs formes de R^2 ont été proposées dans la littérature, nous en distinguons quelques-unes, présentés comme suit [48] :

• **McFadden's R^2**

C'est l'un des premiers indicateurs que l'on retrouve dans la litérature le plus simple et le plus adapté, qui est défini par :

$$R^2_{MF} = 1 - \frac{\ell(\beta, \underline{y})}{\ell_0(\beta, \underline{y})},$$

où Min $R^2_{MF} = 0$ si $\ell(\beta, \underline{y}) = \ell_0(\beta, \underline{y})$ et Max $R^2_{MF}= 1$ si $\mathcal{L}(\beta, \underline{y}) = 1$ c-à-d $\ell(\beta, \underline{y}) = 0$.

• **Cox et Snell's R^2**

Cette mesure populaire, peut être calculée à partir de n'importe quel modèle estimé par la méthode du maximum de vraisemblance, qui sera défini par :

$$R^2_{CS} = 1 - \left(\frac{\mathcal{L}_0(\beta, \underline{y})}{\mathcal{L}(\beta, \underline{y})}\right)^{\frac{2}{n}},$$

Min $R^2_{CS} = 0$ et Max R^2_{CS} si $\mathcal{L}(\beta, \underline{y}) = 1 \Longrightarrow max[R^2_{CS}] = 1 - \mathcal{L}_0(\beta, \underline{y})^{\frac{2}{n}}$.

• **Nagelkerke's R^2**

Ce pseudo est une version ajusté du R^2 de Cox et Snell, donné par :

$$R^2_N = \frac{R^2_{CS}}{max[R^2_{CS}]}$$

Min $R^2_N = 0$ et Max $R^2_N= 1$.

Lorsque ces coefficients sont proches de 1 cela signifie que le modèle est globalement significatif, toutefois ces R^2 sont souvent petits et difficiles à interpréter, ils sont généralement considérés comme correcte si $R^2 > 0.2$.

★ Matrice de confusion

La matrice de confusion est une autre procédure d'évaluation de la régression logistique elle confronte (évalue) les valeurs observées de la variable dépendante Y avec celles qui sont prédites puis comptabilise les bonnes et les mauvaises prédictions. Comme on est dans le cas binaire $y \in \{0, 1\}$ où la modalité 1 de la variable à prédire correspond à la classe "positive" et l'autre à la classe "négative", la matrice est définie par :

	$\hat{y} = 1$	$\hat{y} = 0$
$y = 1$	**VP**	**FN**
$y = 0$	**FP**	**VN**

VN est le nombre de vrais négatifs, c'est-à-dire les observations qui ont été classés négatives et qui le sont réellement.

FN représente le nombre de faux négatifs, c'est-à-dire les individus qui ont été classés négatifs.

FP représente le nombre de faux positifs, c'est-à-dire les individus qui ont été classés positifs.

VP est le nombre de vrais positifs, c'est-à-dire les observations qui ont été classées positives et qui le sont réellement.

D'autres indicateurs peuvent être déduits pour rendre compte de la corrélation entre les valeurs observées et les valeurs prédites de la variable Y, qui sont donnés par :

⋆ La sensibilité (Taux de Vrai Positif (TVP))

Elle indique la capacité du modèle à retrouver les positifs, elle se calcule comme suit :

$$\textbf{TVP} = \frac{VP}{VP+FN}.$$

⋆ La spécificité (Taux de Vrai Négatif (TVN))

À l'inverse de la sensibilité, elle indique la proportion de négatifs détectés :

$$\textbf{TVN} = \frac{VN}{FP+VN}.$$

⋆ La précision

Elle indique la proportion de vrais positifs parmi les individus qui ont été classés positifs. Elle estime la probabilité d'un individu d'être réellement positif lorsque le modèle le classe comme tel elle est définie par :

$$\textbf{Précision} = \frac{VP}{VP+FP}.$$

⋆ **Taux de bonne détection global**

Il correspond à la probabilité de bon classement du modèle, il est défini par

$$\textbf{Taux de bonne détection}= \frac{VP+VN}{VP+FN+FP+VN}.$$

On déduit que la sensibilité et la spécificité jouent un rôle particulier dans l'évaluation du modèle. un bon modèle doit présenter des valeurs assez fortes de taux de bonne détection, des valeurs élevées de sensibilité, précision et spécificité.

Remarque 2.7.2. En plus des indicateurs de performances cités précédemment, il existe deux autres indicateurs simples qui sont **le taux d'erreur** : il représente le rapport entre le nombre de mauvaises prédictions et la taille de l'échantillon, (c'est a dire, la somme des deux valeurs non diagonales de la matrice de confusion divisé par n, avec n est la taille de l'échantillon). Plus ce taux est proche de 0 meilleur est la quantité prédictive du modèle on convient que la qualité prédictive du modèle est mauvaise dès que ce taux dépasse la valeur 0.5 et **la courbe ROC (Receiver Operating Characteristic)** [voir Annexe p 93].

2.7.4 Les tests statistiques

Test du rapport de vraisemblance

On définit le test du rapport de vraisemblance (en anglais Likelihood Ratio Test) :

$$\begin{cases} H_0 : & "\beta_1 = \beta_2 = ... = \beta_p = 0" \\ & \text{contre} \\ H_1 : & \exists\ j = 1, ..., p,\ \beta_j \neq 0 \end{cases}$$

Ce teste repose sur la statistique de décision suivante :

$$MV = -2log\left(\frac{\mathcal{L}_0(\beta)}{\mathcal{L}(\hat{\beta})}\right) \sim \chi^2_p,$$

avec :

$\mathcal{L}_0(\beta, \underline{y})$ correspond à la vraisemblance sans les variables explicatives du modèle.

$\mathcal{L}(\hat{\beta}, \underline{y})$ correspond à la vraisemblance avec les variables explicatives du modèle.

Règle de décision

Nous comparons la valeur calculée MV au quantile d'ordre $(1-\alpha)$ de la loi du χ^2 à p degrés de liberté, notée par $\chi^2_{(1-\alpha,p)}$.

• Si MV$> \chi^2_{(1-\alpha,p)}$, alors on rejette H_0 et on dira que le modèle est globalement bon. D'où il ya au moins une variable explicative significative de y.

• Si MV$\leq \chi^2_{(1-\alpha,p)}$, alors on accepte H_0 et on dira que le modèle est mauvais. D'où il n'y a aucune variable explicative significative de y.

Test individuel de Wald

On s'intéresse ici à la contribution de chaque variable individuellement. En effectuant les tests définis par :

$$H_0 : "\beta_j = 0" \text{ contre } H_1 : "\beta_j \neq 0", \text{ pour } j = 1, ..., p$$

En raison de la normalité asymptotique de l'estimateur du maximum de vraisemblance. Ce test repose sur la statistique $\frac{\hat{\beta}_j}{\hat{\sigma}_{\hat{\beta}_j}}$ qui suit approximativement une loi normale centrée réduite telle que :

$$W = \frac{\hat{\beta}_j}{\hat{\sigma}_{\hat{\beta}_j}} \sim \mathcal{N}(0,1), \quad j = 1, ..., p$$

Règle de décision

Nous comparons la valeur calcuée W au quantile d'ordre $(1 - \frac{\alpha}{2})$ de la loi normale centrée réduite noté par $z_{1-\frac{\alpha}{2}}$.

- Si $W > z_{1-\frac{\alpha}{2}}$ alors on rejette H_0, donc la variable X_j est significative de y.
- Si $W \leq z_{1-\frac{\alpha}{2}}$ alors on accepte H_0, donc la variable X_j n'est pas significative de y.

2.7.5 Intervalle de confiance

Un complément important aux tests de signification du modèle logistique est le calcul et l'interprétation des intervalles de confiance pour les paramètre β_j, $j = 0, ..., p$ au seuil de signification α, (ou bien au niveau de confiance $(1 - \alpha)$), on considére la statistique :

$$W = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}} \sim \mathcal{N}(0,1)$$

À partir de la symétrie de la loi normale, on considère un intervalle symétrique tel que :

$$\mathbb{P}[-z_{1-\frac{\alpha}{2}} < \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}} < z_{1-\frac{\alpha}{2}}] = 1 - \alpha$$

$$\Rightarrow \mathbb{P}[\hat{\beta}_j - z_{1-\frac{\alpha}{2}}\hat{\sigma}_{\hat{\beta}_j} < \beta_j < \hat{\beta}_j + z_{1-\frac{\alpha}{2}}\hat{\sigma}_{\hat{\beta}_j}]$$

On obtient l'intervalle de confiance de β_j, j=0,...,p au seuil α, donné par :

$$IC_{\beta_j} = [\hat{\beta}_j - z_{1-\frac{\alpha}{2}}\hat{\sigma}_{\hat{\beta}_j}, \hat{\beta}_j + z_{1-\frac{\alpha}{2}}\hat{\sigma}_{\hat{\beta}_j}]$$

Où $z_{1-\frac{\alpha}{2}}$ est le quantile d'ordre $1 - \frac{\alpha}{2}$ de la loi normale centrée réduite.

✠ **Différence entre la régression linéaire et la régression logistique binaire**

Régression linéaire	Régression logistique binaire
Une approche linéaire qui modélise la relation entre une variable dépendante et une ou plusieurs variables indépendantes.	Un modèle statistique qui prédit la probabilité d'un résultat qui ne peut avoir que deux valeurs possibles.
Il peut y'avoir une colinéarité entre les variables indépendantes.	Il ne devrait pas y avoir de colinéarité entre les variables indépendantes
Les résidus sont supposés être normalement distribués.	Les résidus doivent être indépendants mais pas forcément normalement distribués.
La méthode d'estimation est la méthode des moindres carrés.	La méthode d'estimation utilisée est celle du maximum de vraisemblance.
Le résultat (variable dépendante) est continue il peut y avoir un nombre infini de valeurs possibles.	Dans la régression logistique on a que 2 possibilités de réponse.
Le lien géométrique (la représentation graphique) de la solution est linéaire (droite de régression).	Utilise une fonction courbe sigmoïde (S-shaped).

TABLE 2.4 – Régression linéaire versus Régression logistique binaire

Commentaire

— Pour que le modèle s'adapte bien aux données, on suppose que les variables indépendantes ne sont pas corrélées et qu'ils sont significativement liées à la réponse Y.

— Sous les hypothèses habituelles de régression linéaire, la méthode des moindres carrés donne des estimateurs avec un certain nombre de propriétés statistiques souhaitables. Malheureusement, lorsque la méthode des MCO est appliquée à un modèle logistique, les estimateurs n'ont plus ces mêmes propriétés. Elle peut être appliquée dans le cas particulier où les données sont groupées [29].

De plus le caractère (binaire) de la variable réponse rend la méthode MCO impossible à mettre en oeuvre, car lorsqu' on éléve la quantité $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ qui est la somme des erreurs au carrés, ce qui n'a pas de sens dans la régression logistique.

2.7.6 Autres types de régression logistique

✠ **Modèle de régression logistique multinomiale (polytomique nominale)**

La régression logistique multinomiale est une généralisation de la régression logistique binaire. Ici la variable dépendante Y admet plus de 2 modalités (non ordonnée). On souhaite expliquer une variable réponse Y à K modalités $y_1, ...y_K$ en fonction de p variables explicatives $X_1, X_2, ..., X_p$.

Considérons n observation $(\underline{Y}, X_{i1}, X_{i2}, ..., X_{ip})$ de valeurs $(\underline{y}, x_{i1}, x_{i2}, ..., x_{ip})$, $i = \overline{1.n}$.

⋆ **La stratégie de modélisation :**

On va modéliser (K-1) rapports de probabilités c-à-d. Prendre une modalité comme référence (ex. la derniére), et exprimer (K-1) logit par rapport à cette référence (ex. les "non-malades" à opposer à différentes catégories de maladies).

La derniére probabilité, appartenance à la K-éme catégorie, est déduite des autres :

$$p_K(\underline{x}_i) = 1 - \sum_{k=1}^{K-1} p_k(\underline{x}_i).$$

On écrit (K-1) équations logit :

$$logit(p_k(\underline{x}_i)) = log\left(\frac{p_k(\underline{x}_i)}{p_K(\underline{x}_i)}\right) = \beta_{0k} + \beta_{1k}x_{i1} + ... + \beta_{pk}x_{ip}, \quad i = \overline{1.n}, \quad k = 1, ...K-1,$$

et on déduit les (K-1) probabilités d'affectation :

$$p_k(\underline{x}_i) = \frac{e^{logit(p_k(\underline{x}_i))}}{1 + \sum_{k=1}^{K-1} e^{logit(p_k(\underline{x}_i))}}, \qquad k = 1, ...K-1,$$

et la derniére $p_K(\underline{x}_i) = 1 - \sum_{k=1}^{K-1} p_k(\underline{x}_i)$, on a donc $\sum_{k=1}^{K} p_k(\underline{x}_i) = 1$.

La régle d'affectation est : $Y = y_k \Leftrightarrow y_k = argmax(p_k(\underline{x}_i))$.

⋆ Estimation des paramètres

La méthodes d'estimation des paramètres est celle du maximum de vraisemblance La fonction de vraisemblance est définie par :

$$\mathcal{L} = \prod_{i=1}^{n} [p_1(\underline{x}_i)^{y_1} \times .... \times (p_K(\underline{x}_i)^{y_K}]$$

D'où la log-vraisemblance est donnée par :

$$\ell = \sum y_1 ln\ p_1(\underline{x}_i) + ... + y_K ln\ p_K(\underline{x}_i)$$

Il y a (K-1)(p+1) paramètres à estimer. On peut s'appuyer de nouveau sur la méthode de Newton-Raphson.

Avec $G = \begin{pmatrix} G \\ \vdots \\ G_{K-1} \end{pmatrix}$ est le vecteur gradient de dimension (K-1)(p+1)1

G_k est de dimension $(p + 1 \times 1)$, avec pour chaque case on a :

$$[g_{k,j}] = \sum_i x_j[y_k - p_k(\underline{x}_i)].$$

La matrice hessienne, de dimension (K-1)(p+1)× (K-1)(p+1), qui sera donnée par :

$$H = \begin{pmatrix} H_{11} & \cdots & H_{1,K-1} \\ \vdots & \ddots & \vdots \\ & & H_{K-1,K-1} \end{pmatrix}$$

$H_{i,j}$ est de dimension $(p+1 \times p+1)$, définie par :

$$H_{ij} = \sum p_i(\underline{x}_i)\left[\delta_{ij} - p_j(\underline{x}_i)\right] xx^t$$

avec $x = (1, X1, ...X_p(\underline{x}_i)$ et $\delta_{ij} = \begin{cases} 1 & si \quad i = j \\ 0 & si \quad i \neq j \end{cases}$

Exemple voir [48]

✠ Modèle de régression logistique ordinale

La régression logistique ordinale est utilisée pour modéliser la relation entre une variable dépendante Y qui prend plus de 2 modalités ordonnées et les variables indépendantes.

Considérons la variable réponse Y avec k modalités et $\underline{x} = (x_1, ..., x_p)^t$ le vecteur des variables explicatives (covariables).

(a) Cas des logits adjacents

Son principe est de calculer le logit du passage d'une catégorie à l'autre. Même idée que le modèle multinomial, sauf que la catégorie de référence change à chaque étape. On évalue le passage de la modalité (k) à (k-1).

Les (K-1) équations logit sont définies par :

$$\begin{cases} logit_1(p(\underline{x})) = & ln\left(\frac{p_1(\underline{x})}{p_2(x)}\right) = \beta_{01} + \beta_{11}x_1 + ... + \beta_{p1}x_p \\ \quad ... & \\ logit_{K-1}(p(\underline{x})) = & ln\left(\frac{p_{K-1}(\underline{x})}{p_K(\underline{x})}\right) = \beta_{0,K-1} + \beta_{1,K-1}x_1 + ... + \beta_{p,K-1}x_p \end{cases}$$

Cette écriture peut être vue comme une ré-interprétation du modèle multinomial.

$$\begin{cases} ln\left(\frac{p_2(\underline{x})}{p_1(\underline{x})}\right) & = & -logit_1(\underline{x}) \\ ln\left(\frac{p_3(\underline{x})}{p_1(\underline{x})}\right) & = & -logit_2(\underline{x}) - logit_1(\underline{x}) \\ \quad ... & & \\ ln\left(\frac{p_K(\underline{x})}{p_1(\underline{x})}\right) & = & -logit_{K-1}(\underline{x}) - ... - logit_2(\underline{x}) - logit_1(\underline{x}) \end{cases}$$

Remarque 2.7.3. On peut utiliser les résultats du modèle multinomial pour estimer les paramètres. Les évaluations et les tests de significativité sont les mêmes.

(b) Cas des ODDS-RATIO cumulatifs

Voyons dans ce cas ce qu'il en est pour l'interprétation des coefficients, pour cela on va modéliser comme suit :

La probabilité cumulée est définie comme suit :

$$\mathbb{P}(Y \leq k \mid X) = p_1 + ... + p_k$$

Les logits cumulatifs sont donnés par :

$$logit_k = ln\left(\frac{\mathbb{P}(Y \leq k \mid X)}{\mathbb{P}(Y > k \mid X)}\right) = ln\left(\frac{\mathbb{P}(Y \leq k \mid X)}{1 - \mathbb{P}(Y \leq k \mid X)}\right) = ln\left(\frac{p_1 + ... + p_k}{p_{k+1} + ... + p_K}\right)$$

Les (K-1) équations logit sont définies (dans un premier temps) comme suit :

$$\begin{cases} logit_1 & = \beta_{0,1} + \beta_{1,1}x_1 + ... + \beta_{p1}x_p \\ ... & \\ logit_{K-1} & = \beta_{0,K-1} + \beta_{1,K-1}x_1 + ... + \beta_{p,K-1}x_p \end{cases}$$

Introduisons de nouveau l'hypothése : le rôle d'une variable ne dépend pas du niveau de Y.

$$logit_k = \beta_{0,k} + \beta_1 x_1 + ... + \beta_p x_p.$$

Exemple voir [48]

Nous renvoyons aux livres de [33] et [27], pour plus de détails et exemples d'applications de la régression logistique.

Chapitre3

La régression logistique bayésienne

3.1 La régression logistique bayésienne

La différence entre l'approche classique et l'approche bayésienne est que l'approche classique suppose que les paramètres sont des valeurs inconnues à estimer et l'approche bayésienne suppose que les coefficients ne sont plus fixes mais plutôt des variables aléatoires suivant une certaine loi de probabilité connue appelée loi a priori de densité $\pi(\beta)$. Dans le cas de la régression logistique l'introduction de cette loi donne lieu à la régression logistique bayésienne. Son principe est d'actualiser les informations données par les observation y en introduisant d'autres informations a priori. La loi résultante de cette actualisation est la loi a posteriori $\pi(\beta \mid y)$.

En résumé l'inférence bayésienne pour les modèles logistiques suit les étapes suivantes :

— Écrire la fonction de vraisemblance des données.

— Introduire une distribution a priori pour les paramètres inconnus du modèle.

— Trouver la distribution a posteriori des paramètres.

Au sein de ce chapitre, nous allons développer l'approche bayésienne pour traiter la régression logistique.

3.1.1 Modèle logistique bayésien dans le cas binaire

La méthodologie décrit l'inférence bayésienne en mettant l'accent sur trois composantes clés : *la distribution a priori*, *la fonction de vraisemblance* et *la distribution a posteriori*.

On dispose de n observations $\underline{x}_i = (x_{i1}, ..., x_{ip})^t$, y$\in \{0, 1\}$ où $Y_i \mid \underline{X}_i = \underline{x}_i \sim \mathcal{B}(p(\underline{x}_i))$ avec Y_i est une variables aléatoire endogéne, $i = \overline{1.n}$.

Considérons le modèle logistique binaire, qui est défini par :

$$logit(p(\underline{x})) = log\left(\frac{p(\underline{x}_i)}{1 - p(\underline{x}_i)}\right) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} = \underline{x}^t \beta$$

où $p(\underline{x}_i) = \frac{exp[logit(p(\underline{x}_i))]}{1+exp[logit(p(\underline{x}_i))]} = \mathbb{P}(Y = 1 \mid \underline{X} = \underline{x})$

La densité a posteriori est donnée par :

$$\pi(\beta \mid y) \propto f(y \mid \beta)\pi(\beta). \tag{3.1}$$

Rappelons que la fonction de vraisemblance est définie par :

$$f(y \mid \beta) = \prod_{i=1}^{n} \left(\frac{e^{\underline{x}_i^t \beta y_i}}{1 + e^{\underline{x}_i^t \beta}}\right).$$

Supposons que la densité a priori est une normale multivariée telle que :

$$\pi(\beta) = \frac{1}{2\pi|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}\beta^t \Sigma^{-1} \beta}. \tag{3.2}$$

où $\Sigma = \sigma^2 I$ est la matrice de covariance, avec I comme étant la matrice identité.

D'où la formule (3.2) peut également s'écrire :

$$\pi(\beta) = \prod_{i=1}^{n} \left(\frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} e^{\frac{-1}{2\sigma^2 \beta_i^2}}\right). \tag{3.3}$$

En excluant tous les termes qui ne dépendent pas de β, la formule (3.1) sera :

$$\pi(\beta \mid y) \propto \prod_{i=1}^{n} \left(\frac{e^{\underline{x}_i^t \beta y_i}}{1 + e^{\underline{x}_i^t \beta}}\right) e^{\frac{-1}{2\sigma^2}\beta^t \beta}. \tag{3.4}$$

L'expression (3.4) n'a pas de forme explicite car c'est une fonction complexe des paramètres. Dans cette situation des méthodes de simulation sont souvent nécessaires afin d'estimer la loi a posteriori pour chacun des paramètres du modèle, on peut citer les méthodes MCMC qu'on définera dans la section (3.2).

3.1.2 Choix des lois a priori et les facteurs influents

La 2éme étape du modèle consiste à définir des distributions a priori pour les coefficients β. Cette spécification est l'aspect majeure de l'analyse bayésienne. De plus il n'existe pas toujours un choix unique, toutes les distributions a priori peuvent être utilisées en fonction de l'information a priori disponible, le choix peut inclure deux catégories : informatives et non informatives.

Dans l'analyse bayésienne du modèle de régression logistique les lois a priori conjuguées des coefficients n'exsite pas alors faire des inférences devient très compliqué.

Très souvent la régression logistique bayésienne nécessite des loi a priori non informatives d'une part car elles aboutissent à des estimations a posteriori trés proche de celles obtenues à partir d'une analyse fréquentiste et d'autre part, elles tendent à imiter une approche du maximum de vraisemblance mais il faut observer que cette approche non informative sur les paramàtres n'est pas informative sur les paramàtres des variables d'origine.

Dans ce contexte bayésien, les distributions a priori sont normalement placées sur les coefficients de régression généralement sous la forme de distribution gaussiennes. Une autre partie indispensable serait donc les facteurs influents par rapport a la distribution a priori et par rapport à l'échantillon.

Parmi les facteurs qui agissent potentiellement sur l'estimation des coefficients sont : l'épaisseur des queues, le positionnement et l'échelle. En plus de la distribution, rappelons que la densité a posteriori est aussi influencée par la fonction de vraisemblance.

3.2 Méthodes MCMC

Les méthodes de Monté Carlo par Chaine de Markov sont apparues en 1950 pour la physique statistique, elles ont des applications presque illimités. Même si leurs performances varient largement selon la complexité du problème, leurs principe consiste à générer une chaine de markov ergodique qui converge vers sa distribution stationnaire qui est exactement la distribution a posteriori.

Dans cette section nous nous intéressons à quelques techniques d'approximation et les résultats principaux relatifs aux méthodes MCMC. Nous commençons d'abord par présenter quelques notions relatives aux chaines de Markov ainsi que leurs diagnostics de convergence, nous développerons en particulier trois méthodes d'estimation numériques tels que la méthode d'échantillonnage par tranche, l'approximation de Laplace et celle de Genkin et al que nous appliquons par la suite pour notre modèle de régression logistique.

3.2.1 Notions de base des méthodes MCMC

- **Chaine de Markov**

Une chaine de markov est un processus stochastique $(X_n)_{n\in\mathbb{N}}$ qui peut-être soit discret soit continu satisfaisant la propriété de markov suivante [45] :

$$\begin{aligned}\mathbb{P}_{ij}^{n} &= \mathbb{P}(X_{n+1}=j \mid X_n=i, X_{n-1}=i_{n-1}, ..., X_0=i_0)\\ &= \mathbb{P}(X_{n+1}=j \mid X_n=i), \quad \forall n\in\mathbb{N},\end{aligned}$$

où $i_0, i_1, ...i_{n-1}$ et j sont les états de la chaine $(X_n)_{n\in\mathbb{N}}$.

Le processus markovien est sans mémoire car l'état futur du processus ne dépend que de l'état présent de ce processus et non pas de ceux du passé.

Les définitions suivantes précisent les propriétés nécessaires pour la convergence des chaines de Markov produites par les algorithmes MCMC.

- **Irréductibilité**

Une chaine de Markov est dite irréductible si elle ne possède qu'une seule classe d'équivalence c-à-d tous ses états communiquent entre eux [45].

- **Récurrence et transiance**

Un état i est dit réccurent si partant de i on y revient presque surement en temps fini [23] :

$$\mathbb{P}(T_i < +\infty \mid X_0 = i) = 1,$$

avec $T_i = inf\{n \geq 1, X_n = i\}$ est le temps de retour en i.

Dans le cas $\mathbb{P}(T_i = +\infty \mid X_0 = i) = 1$, l'état i est dit transitoire, c-à-d c'est avec une probabilité positive qu'on le quitte pour ne jamais y revenir.

On dit qu'un état est récurrent positif lorsque le temps moyen de retour en i est fini :

$$\mu_i = \mathbb{E}(T_i \mid X_o = i) < +\infty$$

Dans le cas où $\mu_i = \mathbb{E}(T_i \mid X_o = i) = +\infty$, i est dit récurent nul.

- **Périodicité**

On dit que l'état i est apériodique si $d(i) = 1$ avec $d(i) = PGCD\{n \geq 1, \mathbb{P}_{ii}^{(n)} > 0\}$ représente la période de l'état i. Une chaine de Markov est dite apériodique si tous ses états sont apériodiques.

- **Ergodicité**

Une chaine de Markov irréductible, récurrente positive et apériodique est dite ergodique.

Une importante littérature est également consacré aux méthodes MCMC, voir ainsi les livres de Gamerman [23], Robert et Casella [46].

3.2.2 Diagnostics de convergence

Nous concluons notre présentation sur les méthodes MCMC par une discussion de leur convergence. La vérification de cette convergence est une étape essentielle dans toutes simulations MCMC il est important de vérifier la convergence pour tous les paramètres du modèle et pas seulement pour un sous-ensemble de paramètres. Nous présentons également trois types de convergence pour lesquels une évaluation est nécessaire.

⋆ **Convergence vers la loi stationnaire**

La convergence vers la distribution stationnaire est nécessaire car nous cherchons à approximer la densité a posteriori, il faut donc s'assurer que notre chaine atteigne sa distribution stationnaire.

Le principal outil d'évaluation de la convergence vers la stationnarité consiste à lancer plusieurs chaines en parallèle afin de comparer leurs performances. Évidemment, cela signifie que la chaine la plus lente du groupe détermine le diagnostic de convergence et que le choix de la distribution initiale est extrêmement important pour garantir que les différentes chaines sont bien réparties (par rapport à la distribution cible).

⋆ **Convergence des moyennes**

Une fois approximativement réglé le problème de convergence vers la distribution stationnaire nous nous trouvons à nouveau dans le cadre classique de Monte Carlo à savoir la convergence de la moyenne empirique $\frac{1}{l}\sum_{l=1}^{L} h(\beta^{(l)})$ vers $\mathbb{E}[h(\beta)]$ pour une fonction quelconque h telle que $\mathbb{E}[h(\beta)] < \infty$ nous devons nous assurer que la chaine ait exploré l'ensemble du support de la densité a posteriori afin d'inférer adéquatement sur celle-ci.

⋆ **Convergence vers un échantillon iid**

La convergence vers un échantillonnage i.i.d. est une autre forme de convergence. L'idée générale est de produire un échantillon quasi-indépendant par sous-échantillonnage (ou échantillonnage par lots) pour réduire la corrélation entre les itérations de la chaine de Markov.

Plusieurs diagnostics ont été proposés dans la littérature pour vérifier la convergence des méthodes MCMC, les diagnostics de Gelman et Rubin (1992) et de Raftery et Lewis (1992) sont actuellement les plus populaires dans la communauté statistique du moins en partie parce que les programmes informatiques pour leur mise en oeuvre sont disponibles auprès de leurs créateurs.

• **Méthode de Gelman et Rubin (1992)**

Gelman et Rubin ont développé une méthode permettant de générer m chaines indépendantes chacune de longueur $2n$ chaque chaine possède des valeurs initiales différentes échantillonées selon une distribution plus dispersée que la distribution cible. Afin de diminuer l'influence de la distribution des valeurs initiales. Gelman et Rubin proposent de rejeter les n premières itérations de chaque chaine afin de garder les n dernières. L'étape qui va suivre est une analyse de la variance qui consiste à calculer une mesure de convergence pour chaque coefficient β qui est le facteur de réduction d'échelle (scale reduction factor) qui consiste à comparer la variation des valeurs des paramètres échantillonnés à l'interieur des chaines (variance interchaine) et entre les chaines (variance intra-chaine), noté par $\sqrt{\hat{G}}$ qui sera défini comme suit [6] :

$$\sqrt{\hat{G}} = \sqrt{\left(\frac{n-1}{n} + \frac{m+1}{mn}\frac{B}{W}\right)\frac{df}{df-2}},$$

avec $B = \frac{n}{m-1}\sum_{i=1}^{m} \frac{n\sum_{i=1}^{m}(\bar{\beta}_{i.}-\bar{\beta}_{..})}{m-1}$ est la variance inter chaine.

$W = \frac{\sum_{i=1}^{m} s_i^2}{m}$ ou $s_i^2 = \frac{\sum_{j=1}^{n}(\beta_{ij}-\bar{\beta}_{i.})}{n-1}$.

$\bar{\beta}_{i.} = \frac{\sum_{j=1}^{n}\beta_{ij}}{n}$ et $\bar{\beta}_{..} = \frac{\sum_{i=1}^{m}\bar{\beta}_{i.}}{m}$.

où β_{ij} est le j'éme vecteur de la iéme chaine et df est le nombre de ddl d'une loi de student qui approxime la densité a posteriori.

Lorsque $\sqrt{\hat{G}}$ est très proche de 1 indique que les chaines sont très susceptibles d'avoir convergé en leur distribution stationnaire et que le tirage s'effectue sur la densité a posteriori.

En complément de cette méthode, il est souvent commun d'utiliser en plus un diagnostic graphique pour vérifier la convergence en superposant chacun des m chaines exécutées en paralléle sur un graphique, nous examinons visuellement si les m chaines semblant échantillonnées selon une distribution commune (distribution stationnaire) ou non. La première méthode graphique vient de Gelfand et Smith [24].

• **Raftery et Lewis (1992)**

Raftery et Lewis (1992) cherchent à savoir le temps nécessaire pour qu'un algorithme MCMC obtient une estimation précise d'un quantile extrême de la densité a posteriori. Pour plus de détails voir [49].

Pour une revue complète et détaillée sur d'autres diagnostics de convergence nous renvoyons à [6].

3.2.3 Algorithmes et méthodes d'approximation

(a) Algorithme d'échantillonnage par tranche (slice sampling)

C'est un algorithme de Markov Chain Monte Carlo (MCMC) proposé par Neal (2003) à partir d'une fonction de densité de probabilité non normalisée, généralement inconnue qu'on note par $h(\beta)$. Il ne génére pas d'échantillons indépendants, mais plutot une séquence markovienne dont la distribution stationnaire est la distribution cible. Cependant, il différe des autres algorithmes MCMC bien connus car seule la loi à posteriori est mise à l'échelle et doit être spécifiée. Aucune proposition ou distribution marginale n'est nécessaire [40].

Posons $h(\beta \mid Y, X)$ telle que :

$$h(\beta \mid Y, X) = f(Y \mid \beta, X)\pi(\beta \mid X) \propto \pi(\beta \mid Y, X).$$

Introduisons une variable aléatoire auxiliaire, notée U. Posons la loi conditionnelle :

$$U \mid \beta \sim \mathcal{U}[0, h(\beta \mid Y, X)].$$

Puisque la fonction de densité conditionnelle est :

$$f(u \mid \beta) = \frac{1}{h(\beta \mid y, x)} \mathbf{1}_{\{u<h(\beta|y,x)\}}.$$

Nous avons

$$f(u, \beta) = f(u \mid \beta)\pi(\beta \mid y, x) \propto \frac{1}{h(\beta \mid y, x)} \mathbf{1}_{\{u<h(\beta|y,x)\}} h(\beta \mid y, x).$$

De plus nous avons :

$$f(\beta \mid u) = \frac{f(u, \beta)}{f(u)} = f(u, \beta) \propto \mathbf{1}_{\{\beta|u<(\beta|Y,X)\}}$$

La méthode d'échantillonnage consiste à choisir itérativement des valeurs de u et β à partir d'une valeur de départ $\beta^{(0)}$, nous avons les étapes suivantes à l'itération k :

(1) Tirer $u^{(k)}$ uniformément sur l'intervalle $[0, h(\beta^{(k-1)})]$

(2) Tirer $\beta^{(k)}$ uniformément sur l'ensemble $S = \{\beta \mid h(\beta \mid Y) > u^{(k)}\}$

L'étape (1) est triviale, en effet après avoir calculé $h(\beta^{(k-1)} \mid Y)$ il est facile d'échantillonner $u^{(k)}$ uniformément sur l'intervalle $[0, h(\beta^{(k-1)} \mid Y)]$ il est toutefois pus difficile d'éffectuer l'étape (2) puisque $h(\beta \mid Y, X)$ n'est pas nécessairement inversible, l'ensemble $S = \{\beta \mid h(\beta \mid Y) > u^{(k)}\}$ n'est pas toujours simple à déterminer.

Pour échantillonner l'ensemble S, nous appliquons une procédure pas à pas [voir Annexe p 94]

(b) L'approximation de Laplace

Approximer la distribution a posteriori $\pi(\beta \mid y) = \frac{f(y|\beta)\pi(\beta)}{\int f(y|\beta)\pi(\beta)d\beta}$ par une loi normale multivariée $\pi(\beta \mid y) \sim \mathcal{N}(\mu, \Sigma)$ avec μ est le vecteur des espérances et Σ est la matrice de variance-covariance [44].

En utilisant une notation condensée, on aura

$$\pi(\beta \mid y) = \frac{e^{ln(\pi(\beta|y))}}{\int e^{ln(\pi(\beta|y))}d\beta}$$

Nous approchons $ln(\pi(\beta \mid y))$ au numérateur et au dénominateur.

Supposons que :

$$ln(\pi(\beta \mid y)) = g(\beta) \tag{3.5}$$

Approximons $g(\beta)$ en utilisant le développement de Taylor à l'ordre 2.

$$g(\beta) \approx g(z) + (\beta - z)^t \nabla g(z) + \frac{1}{2}(\beta - z)^t \nabla^2 g(z)(\beta - z), \tag{3.6}$$

où z est un point choisi arbitrairement dans le domaine de g, choisissons $z = \beta_{MAP}$, avec β_{MAP} est l'estimateur de β obtenu par maximisation de la loi a posteriori.

A partir de (3.6), l'approximation de Laplace est :

$$\begin{aligned} \pi(\beta \mid y) &= \frac{e^{g(\beta)}}{\int e^{g(\beta)}d\beta} \\ &\approx \frac{e^{g(z)+(\beta-z)^t\nabla g(z)+\frac{1}{2}(\beta-z)^t(\nabla^2 g(z)(\beta-z))}}{\int e^{g(z)+(\beta-z)^t\nabla g(z)+\frac{1}{2}(\beta-z)^t(\nabla^2 g(z)(\beta-z))}} \end{aligned}$$

Cela peut être simplifiée en deux étapes :

1- Le terme $e^{g(\beta_{MAP})}$ dans le numérateur et le dénominateur peut être considéré comme une constante car il ne varie pas en β . Il est donc simplifié.

2- Par définition de β_{MAP}, le vecteur $\nabla ln(\pi(\beta \mid y)) = 0$, approximation est alors :

$$\pi(\beta \mid y) = \frac{e^{\frac{1}{2}(\beta-\beta_{MAP})^t(-\nabla^2 ln\pi(\beta|y))(\beta-\beta_{MAP})}}{\int e^{\frac{1}{2}(\beta-\beta_{MAP})^t(-\nabla^2 ln\pi(\beta|y))(\beta-\beta_{MAP})}d\beta} \tag{3.7}$$

Donc l'approximation de Laplace de $\pi(\beta \mid y)$ est une gaussienne et est donnée par :

$$\pi(\beta \mid y) \sim \mathcal{N}(\mu, \Sigma) \tag{3.8}$$

où $\mu = \underset{\beta}{argmax}(ln(\pi(\beta \mid y))) = \hat{\beta}_{MAP} = \hat{\beta}_{MMSE}$ et $\Sigma = [-\nabla^2 ln\pi(\beta \mid y)]^{-1}$.

• **Propriétés de l'approximation de Laplace** [44] :

1. Il est généralement simple si les dérivées (premières et secondes) peuvent être calculées facilement.

2. Il est cher si le nombre de paramètres est très grand (en raison de calcul et de l'inversion de hessien).

3. Il peut mal faire si le (vrai) a posteriori est multimodal.

4. Il peut réellement s'appliquer lorsqu' on travaille avec n'importe quelle fonction de perte régularisée pour obtenir une distribution a posteriori gaussienne sur les paramètres.

(c) La méthode de Genkin et al (2007)

Genkin et al proposent un algorithme d'estimation ponctuelle des coefficients β. L'idée est d'introduire une méthode d'estimation rapide et efficace. Ils proposent de passer par l'estimateur du maximum a posteriori (MAP). L'approche bayésienne proposée évite le sur-ajustement des données et est efficace pour la prédiction, en utilisant une densité a priori de Laplace de moyenne 0 et de variance $\frac{2}{\lambda_j^2}$, définie par [19] :

$$\pi(\beta_j) = \frac{\lambda_j}{2} e^{-\lambda_j|\beta_j|}. \tag{3.9}$$

En appliquant cette méthode au modèle logistique binaire, pour lequel la fonction vraisemblance est :

$$f(y \mid \beta) = \prod_{i=1}^{n} \left[\frac{e^{y_i(\beta_0+\sum_{j=1}^{p}\beta_j x_{ij})}}{1+e^{\beta_0+\sum_{j=1}^{p}\beta_j x_{ij}}} \right]. \tag{3.10}$$

Et en utilisant la loi a priori définie dans la formule (3.9), la densité a posteriori est :

$$\pi(\beta \mid y) = \prod_{i=1}^{n} \left[\frac{e^{y_i(\beta_0+\sum_{j=1}^{p}\beta_j x_{ij})}}{1+e^{\beta_0+\sum_{j=1}^{p}\beta_j x_{ij}}} \right] \frac{\lambda_j}{2} e^{-\lambda_j|\beta_j|}.$$

Le log de la loi a posteriori sera donnée par :

$$\begin{aligned}
\ell(\beta) &= log\left(\prod_{i=1}^{n}\left[\frac{e^{y_i(\beta_0+\sum_{j=1}^{p}\beta_j x_{ij})}}{1+e^{\beta_0+\sum_{j=1}^{p}\beta_j x_{ij}}}\right]\frac{\lambda_j}{2}e^{-\lambda_j|\beta_j|}\right)\\
&= \sum_{i=1}^{n} y_i\left(\beta_0+\Sigma_{j=1}^{p}\beta_j x_{ij}\right) - log\left(1+e^{\beta_0+\sum_{j==1}^{p}\beta_j x_{ij}}\right) + log\left(\frac{\lambda_j}{2}e^{-\lambda_j|\beta_j|}\right)\\
&= \sum_{i=1}^{n} y_i\left(\beta_0+\Sigma_{j=1}^{p}\beta_j x_{ij}\right) - log\left(1+e^{\beta_0+\sum_{j=1}^{p}\beta_j x_{ij}}\right) + log\left(\frac{\lambda_j}{2}-\lambda_j|\beta_j|\right)\\
&= \sum_{i=1}^{n} y_i\left(\beta_0+\sum_{j=1}^{p}\beta_j x_{ij}\right) - log\left(1+e^{\beta_0+\sum_{i=1}^{n}\beta_j x_{ij}}\right) + log(\lambda_j) - log(2) - \lambda_j|\beta_j|.
\end{aligned}$$

Alors l'estimation MAP est alors le β qui maximise $\ell(\beta)$.

3.3 Cas de loi a priori non informative de Jeffreys

La loi a priori de Jeffreys est peut être la loi a priori non informative la plus largement utilisée dans l'analyse bayésienne. Pour le modèle logistique la loi a priori de Jeffreys est attrayante car elle ne céssecite aucune élicitation d'hyperparamètres, il y a eu une énorme littérature sur cette loi et ses propriétés pour une grande variété d'applications.

Cette littérature est trop vaste pour être énuméré dans son intégralité ici mais on peut citer quelques travaux tels que Firth (1993) a suggéré l'utilisation de la loi de Jeffreys comme solution au problème du biais dans les estimateurs du maximum de vraisemblance [14], Chen et al (2008) ont étudié les propriétés et l'implémentation de la loi a priori de Jeffreys pour les modèles logistiques [5]. Bien que la littérature sur la loi a priori de Jeffreys soit vaste mais il y a trés peu de discussion sur les propriétés théoriques sur cette loi pour les modèles de régression logistique.

3.3.1 Modèle logistique

La densité a posteriori est donnée par :

$$\pi(\beta \mid y) \propto f(y \mid \beta)\pi(\beta).$$

La fonction de vraisemblance est :

$$f(y \mid \beta) = \prod_{i=1}^{n}\left(\frac{e^{\underline{x}_i\beta y_i}}{1+e^{\underline{x}_i\beta}}\right).$$

Comme β est un vecteur $\beta = (\beta_1, ...\beta_p)$ dans ce cas la loi a priori de Jeffreys sera donnée par :

$$\pi(\beta) = \sqrt{det(I(\beta))},$$

où $I(\beta)$ est la matrice d'information de fisher dont les éléments sont donnés par :

$$I(\beta)_{rk} = \mathbb{E}\left[-\frac{\partial^2}{\partial\beta_r\partial\beta_k} log f(y \mid \beta)\right].$$

On a :

$$\frac{\partial^2}{\partial\beta_r\partial\beta_k} log f(y \mid \beta) = -\sum_{i=1}^{n}\left[x_{ir}x_{ik}\left(\frac{e^{\underline{x}_i\beta}}{(1+e^{\underline{x}_i\beta})^2}\right)\right].$$

On aura donc :

$$I(\beta)_{rk} = \mathbb{E}\left[\sum_{i=1}^{n} x_{ir}x_{ik}\left(\frac{e^{\underline{x}_i\beta}}{(1+e^{\underline{x}_i\beta})^2}\right)\right] = \sum_{i=1}^{n} x_{ir}x_{ik}\left(\frac{e^{\underline{x}_i\beta}}{(1+e^{\underline{x}_i\beta})^2}\right).$$

D'où la loi a priori de Jefferys est :

$$\pi(\beta) = \sqrt{det\left[(I(\beta))_{rk}\right]}.$$

La loi a posteriori sera donnée comme suit :

$$\pi(\beta \mid y) = \prod_{i=1}^{n}\left(\frac{e^{\underline{x}_i\beta y_i}}{1+e^{\underline{x}_i\beta}}\right)\sqrt{det\left[(I(\beta))_{rk}\right]}. \tag{3.11}$$

Comme l'expression de la loi a posteriori qui est donnée pour une loi a priori non informative de Jeffreys, définie dans la formule (3.11) est complexe, donc elle ne peut pas être calculée manuellement cela nécessite donc des méthodes de simulations.

Application

Dans ce chapitre nous présentons deux applications, la premiére est dans le domaine de santé et la deuxiéme dans le domaine économique.

1. Problème cardiaque

La présence ou l'absence d'une maladie cardiaque (CHD) peut être appliquée par l'age de l'individu (voir [27]).

Le modèle logistique permettant de décrire cette liaison est présenté par :

$$logit(\mathbb{P}(Y = 1 \mid X = x)) = log\left(\frac{p(x)}{1 - p(x)}\right) = \beta_0 + \beta_1 x$$

où Y = CHD est la variable binaire qui explique la "présence" ou "non" de la maladie cardiaque.

$$\text{Y} = \begin{cases} 1 & \text{si l'individu est atteint de la maladie} \\ 0 & \text{sinon} \end{cases}$$

et X = "Âge" de l'individu. La probabilité que l'individu soit atteint de la maladie CHD est :

$$p(x) = \frac{e^{logit(p(x))}}{1 + e^{logit(p(x))}} = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

où β_0 et β_1 sont les paramètres à estimé.

A partir des données [27] présentées en Annexe p 95, nous avons tracé la figure suivante :

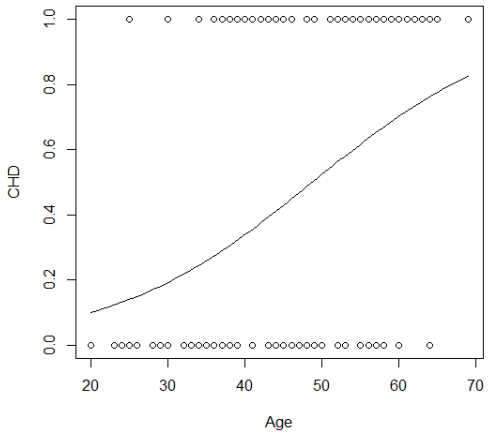


FIGURE 3.1 – Nuage de points de CHD par Age

La figure (3.1) montre la nature dichotomique de la relation entre le résultat (CHD) et la variable indépendante (Age).

✠ Estimation des paramètres

Les estimateurs de maximum de vraisemblance de β_0 et β_1 sont $\hat{\beta}_0 = -5.309$ et $\hat{\beta}_1 = 0.111$

La probabilité que l'individu soit atteint de la maladie (CHD) est :

$$\widehat{p(x)} = \frac{e^{-5.309+0.111x}}{1+e^{-5.309+0.111x}}$$

D'où le modèle estimé est : $logit(p(x)) = -5.309 + 0.111x$.

✠ Tests statistiques

(a) Test de rapport de vraisemblance

```
> reg00=glm(CHD~1,family=binomial)
> anova(reg00,regg,test="Chisq")
Analysis of Deviance Table

Model 1: CHD ~ 1
Model 2: CHD ~ Age
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1        99     136.66
2        98     107.35  1    29.31 6.168e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

D'aprés les résultats du test, on a $-2\ell_0(\beta) = 136.66$ et $-2\ell(\hat{\beta}) = 107.35$, on aura donc :

MV$= -2log\left(\frac{\mathcal{L}_0(\beta)}{\mathcal{L}(\hat{\beta})}\right) = (-2\ell_0(\beta))-(-2\ell(\hat{\beta})) = 136.66\text{-}107.35 = 29.31$. et $\chi^2_{(0.95,1)}$=3,8415.

Comme on a MV >3,8415, alors on dira que le modèle est globalement bon. D'où la variable explicative x(Âge) significative de y (CHD).

(b) Test de Wald

```
> summary(regg)

Call:
glm(formula = CHD ~ Age, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9718  -0.8456  -0.4576   0.8253   2.2859

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.30945    1.13365  -4.683 2.82e-06 ***
Age          0.11092    0.02406   4.610 4.02e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 136.66  on 99  degrees of freedom
Residual deviance: 107.35  on 98  degrees of freedom
AIC: 111.35

Number of Fisher Scoring iterations: 4
```

On a $W = \frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}} = \frac{0.11092}{0.02406} = 4.61014$, où $z_{1-\frac{\alpha}{2}}$ est le quantile d'ordre $1-\frac{\alpha}{2}$ de la loi normale centrée réduite qui est égale à 1.96.

Comme W>1.96, donc la variable Âge est significative de la maladie (CHD).

✠ Intervalle de confiance

```
> confint.default(regg,level=0.95)
                  2.5 %      97.5 %
(Intercept) -7.53137370 -3.0875331
Age          0.06376477  0.1580775
```

IC_{β_0}=[-7.53137370 ; -3.0875331] IC_{β_1}=[0.06376477 ; 0.1580774]

On remarque que $\hat{\beta}_1 \in IC_{\beta_1}$, ce qui signifie que nous sommes convaincus à 95% que la corrélation entre la maladie CHD et l'age de l'individu est comprise entre [0.06376477 ; 0.1580774].

✠ La qualité d'ajustement

(a) Les pseudos R^2

On a $-2\ell_0(\beta) = 136.66 \Longrightarrow \ell_0(\beta) = \text{-}68.33 \Longrightarrow \mathcal{L}_0(\beta) = 2.1118256 \times 10^{-30}$.

$-2\ell(\beta) = 107.35 \Longrightarrow \ell(\beta) = \text{-}53.675 \Longrightarrow \mathcal{L}(\beta) = 4.8892662 \times 10^{-24}$.

$max[R^2_{CS}] = 1 - (\mathcal{L}_0(\beta))^{\frac{2}{n}} = 1 - (2.1118256 \times 10^{-30})^{\frac{2}{100}}$.

Les pseudos de **McFadden's**, **CoxSnell** et **Nagelkerke** seront présentés dans le tableau suivant :

Pseudos	Formule	Formule calculée
R^2 de McFadden	R^2_{MF}=$1 - \left(\frac{\ell(\beta)}{\ell(\beta_0)}\right)$	$R^2_{MF} = 1 - \frac{-53.675}{-68.33} = 0.2144$
R^2 de CoxSnell	$R^2_{CS} = 1 - \left(\frac{\mathcal{L}_0(\beta)}{\mathcal{L}(\beta)}\right)^{\frac{2}{n}}$	$R^2_{CS} = 1 - \left(\frac{2.1118256\times10^{-30}}{4.8892662\times10^{-24}}\right)^{\frac{2}{100}} = 0.25408$
R^2 de Nagelkerke	$R^2_N = \frac{R^2_{CS}}{max[R^2_{CS}]}$	$R^2_N = \frac{0.254084649}{1-(2.1118256\times10^{-30})^{\frac{2}{100}}} = 0.341040$

Comme ses pseudos $R^2 > 0.2$, on peut dire que le modèle est globalement validé.

(b) La matrice de confusion

```
> prev1=predict(regg,type="response")
> prev11=factor(ifelse(prev1>0.5,"1","0"))
> mc=table(CHD,prev11)
> mc
   prev11
CHD  0  1
  0 45 12
  1 14 29
```

VP = 45, ils ont projeté du positif et cela s'est avéré vrai. Dans cet exemple ils ont prédit que la maladie CHD serait présente chez l'individu et elle l'est.

FN = 12, la prédiction négative est fausse, c'est à dire la prédiction que la maladie est présente chez l'individu est fausse.

VN = 29, ils ont prédit négatif et c'est vrai, c'est à dire ils ont prédit que la maladie est absente et elle l'est.

FP = 14, la prédiction positive est fausse, c'est à dire la prédiction que la maladie est absente est fausse.

Nous en déduisons les principaux indicateurs d'évaluation tirésà partir de la matrice de confusion :

(1) La précision

$$\textbf{Précision} = \frac{45}{45+14} = 76.27\%$$

La précision signifie que seulement 76.27% des patients appartiennent à la classe réelle "présence de la maladie" parmi tous les patients qui devrait être atteint de la maladie.

(2) La sensibilité

$$\textbf{Sensibilité} = \frac{45}{45+12} = 78.95\%$$

La sensibilité signifie que 78.95% des patients qui sont atteint de la maladie ont été correctement classés comme ils ont une maladie.

(3) La spécificité

$$\textbf{Spécificité} = \frac{29}{14+29} = 67.44\%$$

La spécificité signifie qu'il y a 67.44% de négatifs détectés.
.

✠ La prédiction

Afin de prédire la probabilité que l'individu atteint la maladie à un age donné, on utilise la commande suivante :

```
> prev5=predict(regg,type="response")
> prev5
```

L'estimation de la probabilité d'avoir une maladie pour un individu de 30 ans et 60 ans est :

$$\hat{p}(x=30) = \frac{e^{-5.309+0.111(30)}}{1+e^{-5.309+0.111(30)}} = 0.1214254 \quad \text{et} \quad \hat{p}(x=60) = \frac{e^{-5.309+0.111(60)}}{1+e^{-5.309+0.111(60)}} = 0.7942.$$

Comme $0.7942 > 0.1214$, ce qui confirme que la maladie cardiaque CHD est en relation avec l'age de l'individu.

2. Problème du défaut bancaire

L'exemple présenté ici pour appliquer le modèle logistique consiste à traiter un problème économique du défaut bancaire qui est extrait de [16]. Nous cherchons à déterminer quels clients seront en défaut sur leur dette de carte de crédit (c'est-à-dire, le risque de non-remboursement) en fonction du statut du client, le montant mensuel moyen d'utilisation de la carte et le revenu du client. On dispose d'un échantillon de 50 clients. L'objectif est d'expliquer la présence ou l'absence du défaut bancaire à partir de son montant, le revenu et le statut du modèle logistique :

$$logit(\mathbb{P}(Y = 1 \mid \underline{X} = x)) = log\left(\frac{p(\underline{x})}{1 - p(\underline{x})}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \tag{3.12}$$

où Y="defaut" est la variable binaire qui explique la "présence" ou "non" du défaut bancaire.

$$\text{Y} = \begin{cases} 1 & \text{si le client fait defaut} \\ 0 & \text{sinon} \end{cases}$$

et $\underline{X} = (X_1, X_2, X_3)$ de valeur $\underline{x} = (x_1, x_2, x_3)$ avec :

X_1="client" est la variable "statut" binaire qui vaut 1 si le client est un étudiant et 0 sinon.

X_2="balanc" est le montant mensuel.

X_3="revenu" est le revenu du client.

Et la probabilité que le client fasse défaut sur sa dette est

$$p(\underline{x}) = \frac{e^{logit(p(\underline{x}))}}{1 + e^{logit(p(\underline{x}))}} = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3}}{1 + e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3}}$$

β_0, β_1, β_2 et β_3 sont les paramètres à estimer par le ML.

Les données sont présentées dans le tableau (voir Annexe p 96).

• Représentation graphique des données

Les diagrammes de dispersion ainsi que les courbes logistiques des données du tableau 3.1 sont présentés dans les figures suivantes :

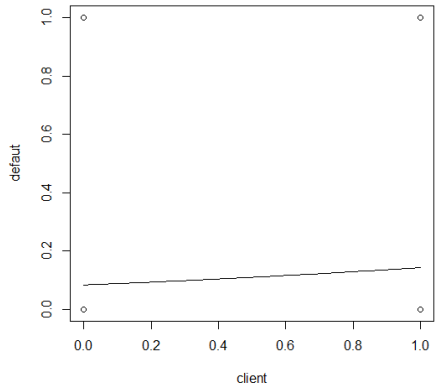


FIGURE 3.2 – Nuage de points de defaut par client

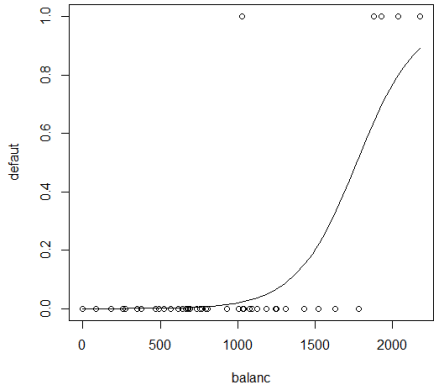


FIGURE 3.3 – Nuage de points de defaut par balanc

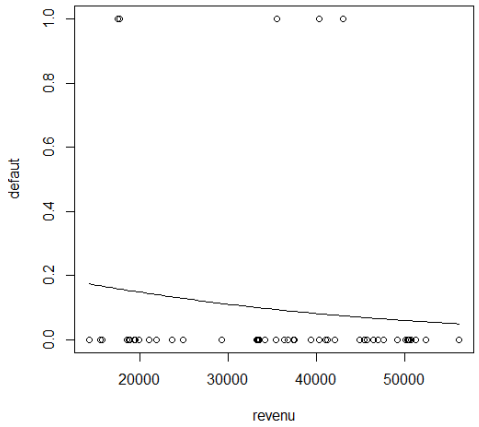


FIGURE 3.4 – Nuage de points de defaut par revenu

• Interprétation

Les figures (3.2), (3.3) et (3.4) montrent des diagrammes de dispersion de (defaut versus client), (defaut versus balanc) et (defaut versus revenu) respectivement. Elles montrent la nature dichotomique de toute relation entre le résultat et la variable indépendante. On remarque que tous les points tombent sur l'une des deux droites parallèles représentant l'absence du défaut (y=0) et la présence du défaut (y=1) ces graphiques dépeint assez clairement la nature dichotomique de la variable réponse.

✠ Estimation (MV) des paramètres

Les estimateurs des paramètres β_0, β_1, β_2 et β_3 du modèle logistique (3.12) sont donnés grace à la fonction prédéfinie "glm" suivante :

```
> reg=glm(defaut~client+balanc+revenu,data=Default,family=binomial)
> reg

Call:  glm(formula = defaut ~ client + balanc + revenu, family = binomial,
    data = Default)

Coefficients:
(Intercept)       client       balanc       revenu
 -9.831e+00   -2.333e+00    6.473e-03   -3.714e-06

Degrees of Freedom: 49 Total (i.e. Null);  46 Residual
Null Deviance:      32.51
Residual Deviance: 12.72        AIC: 20.72
```

FIGURE 3.5 – Résultats de l'estimation classique

On aura donc :

$$\hat{\beta}_0 = -9.8311e+00, \quad \hat{\beta}_1 = -2.333e+00, \quad \hat{\beta}_2 = 6.473e-03, \quad \hat{\beta}_3 = -3.714e-06$$

Ces coefficients sont saisis dans l'équation de régression logistique pour estimer la probabilité, qu'un individu fasse défaut sur sa dette est alors :

$$\hat{p}(\underline{x})) = \frac{e^{-9.8311e+00-2.333e+00x_1+6.473e-03x_2-3.714e-06x_3}}{1+e^{-9.8311e+00-2.333e+00x_1+6.473e-03x_2-3.714e-06x_3}}$$

D'où le modèle estimé est :

$$logit(p(\underline{x})) = -9.8311e+00 - 2.333e+00x_1 + 6.473e-03x_2 - 3.714e-06x_3$$

- **Interprétation des résultats**

Pour le modèle de régression logistique les valeurs des coefficients ne sont pas directement interprétables comme dans le cas de la régression linéaire, seuls les signes des coefficients indiquent si la variable agit positivement ou négativement sur la probabilité $p(\underline{x})$. C'est à dire, si le coefficient est positif cela implique une augmentation du logit et s'il est négatif implique une diminution du logit.

Dans notre cas, la variable client contribue négativement sur la probabilité que le client fasse défaut sur sa dette car le signe du coefficient associé est négatif. De même pour la variable revenu. Par contre pour la variable explicative balanc, elle contribue positivement sur cette probabilité.

• **Odds-Ratio**

Dans le cadre d'un modèle logistique généralement on ne présente pas les coefficients du modèle mais les valeurs exponentielles. Cette dernière correspondant en effet à des Odds-Ratio également appelés rapport des côtes, dont la formule est donnée par :

$$OR = \frac{p(\underline{x})}{1 - p(\underline{x})}$$

On obtient les Odds-Ratio donnés par :

```
> exp(coef(reg))
 (Intercept)       client       balanc       revenu
5.377458e-05 9.695603e-02 1.006494e+00 9.999963e-01
```

FIGURE 3.6 – Résultats des Odds-Ratio

L'odds-ratio de la variable balanc est de 1.006494 > 1, cela signifie qu'il y a une augmentation du risque de défaut sur la dette du client, c'est à dire lorsque le montant moyen mensuel augmente d'une unité, le risque de défaut sur sa dette augmente, donc il y a une association entre le montant mensuel d'utilisation de la carte et le risque que le client fasse défaut sur sa dette.

✠ Les tests statistiques

(a) Test de rapport de vraisemblance

On utilise la commande suivante sous R, on obtient :

```
> reg0=glm(defaut~1,family=binomial)
> anova(reg0,reg,test="Chisq")
Analysis of Deviance Table

Model 1: defaut ~ 1
Model 2: defaut ~ client + balanc + revenu
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1        49     32.508
2        46     12.724  3   19.784 0.0001881 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

FIGURE 3.7 – Résultats du test de rapport de vraisemblance

Où MV = 32.508-12.724 = 19.784, et $\chi^2_{(0.95,3)} = 7.8147$.

Comme on a MV> 7.8147, alors on dira que le modèle est globalement bon. D'où, il y a au moins une variable explicative significative de y.

(b) Test de Wald

En utilisant la fonction "summary" on obtient les résultats suivants :

```
> summary(reg)

Call:
glm(formula = defaut ~ client + balanc + revenu, family = binomial,
    data = Default)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.12477  -0.20786  -0.07572  -0.01804   2.59116

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -9.831e+00  7.020e+00  -1.400   0.1614
client      -2.333e+00  3.534e+00  -0.660   0.5090
balanc       6.473e-03  2.704e-03   2.393   0.0167 *
revenu      -3.714e-06  1.304e-04  -0.028   0.9773
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 32.508  on 49  degrees of freedom
Residual deviance: 12.724  on 46  degrees of freedom
AIC: 20.724

Number of Fisher Scoring iterations: 8
```

FIGURE 3.8 – Résultats du test de Wald

On a $W_1 = \frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}} = \frac{-2.333e+00}{3.534e+00} = -0.660158$.

Où $z_{1-\frac{\alpha}{2}}$ est le quantile d'ordre $1-\frac{\alpha}{2}$ de la loi normale centrée réduite qui est égale à 1.96.

Comme on a $W_1 < 1.96$, donc la variable "statut" n'est pas significative de y.

Et $W_2 = \frac{\hat{\beta}_2}{\hat{\sigma}_{\hat{\beta}_2}} = \frac{6.473e-03}{2.704e-03} = 2.393860 > 1.96$, donc la variable "balanc" est significative de y.

Et $W_3 = \frac{\hat{\beta}_3}{\hat{\sigma}_{\hat{\beta}_3}} = \frac{-3.71e-06}{1.304e-04} = -0.385041319 < 1.96$ donc la variable "revenu" n'est pas significative de y.

D'aprés le test de Wald la seule variable significative du modèle est la variable "balanc", on constate que cette variable est statistiquement associée au défaut bancaire.

✠ Intervalles de confiance

Afin d'obtenir les intervalles de confiance des coefficients β_0, β_1, β_2 et β_3 on utilise la formule suivante :

```
> confint.default(reg,level=0.95)
                    2.5 %       97.5 %
(Intercept) -2.358998e+01 3.9285641629
client      -9.259393e+00 4.5923981133
balanc       1.172138e-03 0.0117733971
revenu      -2.593807e-04 0.0002519517
```

FIGURE 3.9 – Intervalles de confiance

$IC_{\beta_0} = [-2.358998e+01; 3.9285641629]$, $IC_{\beta_1} = [-9.25393e+00; 4.5923981133]$

$IC_{\beta_2} = [1.172138e-0.3; 0.0117733971]$, $IC_{\beta_3} = [-2.593807e-04; 0.0002519517]$

On a $\hat{\beta}_0 \in IC_{\beta_0}$.

$\hat{\beta}_3 \notin IC_{\beta_3}$ cela signifie que nous sommes pas convaincus à 95% que la corrélation entre le défaut bancaire et le revenu des clients est comprise entre [-2.593807e-04 ;0.0002519517].

$\hat{\beta}_2 \in IC_{\beta_2}$, ce qui signifie que nous sommes convaincus à 95% que la corrélation entre le défaut bancaire et le montant mensuel de l'utilisation de la carte de crédit pour les clients est comprise entre [1.172138e-03 ;0.0117733971].

$\hat{\beta}_1 \in IC_{\beta_1}$, ce qui signifie que nous sommes convaincus à 95% que la corrélation entre le défaut bancaire et le statut des clients est comprise entre [1.172138e-03 ;0.0117733971].

✠ La qualité d'ajustement

(a) Les pseudos R^2

Les pseudos R^2 de CoxSnell, Nagelkerke, McFadden, Tjur et sqPearson sont donnés par la commande suivante :

```
> library(modEvA)
> RsqGLM(model=reg)
$CoxSnell
[1] 0.3267815

$Nagelkerke
[1] 0.683585

$McFadden
[1] 0.6085914

$Tjur
[1] 0.6152032

$sqPearson
[1] 0.6362105
```

FIGURE 3.10 – Les pseudos R^2

Comme ses pseudos R^2 sont souvent petits et difficiles à interpréter, on dit donc que le modèle est globalement validé dés que ses R^2 dépasse la valeur 0.2. C'est le cas dans cet exemple.

(b) La matrice de confusion

```
> prev=predict(reg,type="response")
> prev1=factor(ifelse(prev>0.5,"1","0"))
> mc=table(defaut,prev1)
> mc
      prev1
defaut  0  1
     0 45  0
     1  2  3
```

FIGURE 3.11 – La matrice de confusion

Nous avons donc 2 prédictions incorrectes sur un total de 50.

VP=45, on a projeté du positif et cela s'est avéré vrai. Dans notre exemple on a prédit que le défaut bancaire serait présent chez le client et l'est.

FN=0, notre prédiction négative est fausse, c'est à dire la prédiction que le défaut bancaire est présent chez le client est fausse.

VN=3, on a prédit négatif et c'est vrai, c'est à dire on a prédit que le défaut bancaire est absent et il l'est.

FP =2, notre prédiction positive est fausse, c'est à dire la prédiction que le défaut bancaire est absente est fausse.

(1) La précision

La précision= $\frac{VP}{VP+FP} = \frac{45}{45+2} = 95\%$.

On a 95% des clients qui appartiennent à la classe réelle "présence de défaut bancaire" parmi tous les clients qui devraient avoir un défaut bancaire.

(2) La sensibilité

$TVP = \frac{VP}{VP+FN} = \frac{45}{45+0} = 100\%$

Les 100% des clients qui ont un défaut bancaire ont été correctement classés.

(3) La spécificité

$TVN = \frac{VN}{FP+VN} = \frac{3}{2+3} = 60\%$

il y a 60% de négatifs détectés.

(c) Le taux d'erreur

La valeur du taux d'erreur est donnée par :

```
> t=(mc[1,2]+mc[2,1])/sum(mc)
> t
[1] 0.04
```

On a le taux d'erreur est de 4% qui est proche de 0, donc la qualité prédictive du modèle est meilleur.

(d) La courbe ROC

La courbe ROC est donnée par la commande suivante :

```
> library(Epi)
> ROC(form=defaut~client+balanc+revenu,plot="ROC")
```

Elle est représentée comme suit :

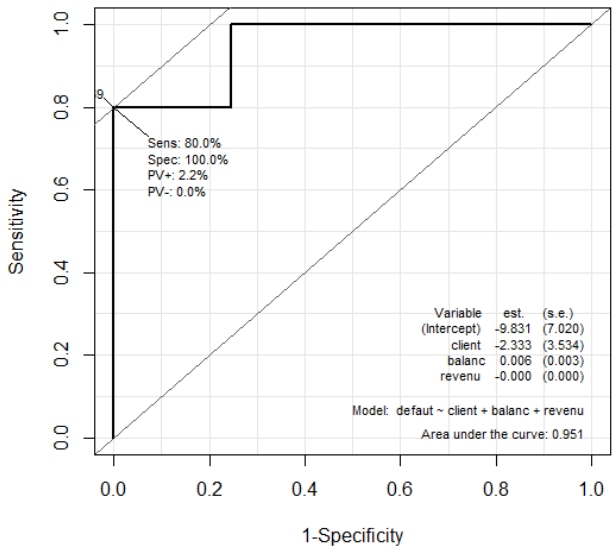


FIGURE 3.12 – La courbe ROC

D'aprés la figure (3.12) on remarque que la courbe ROC est proche du coin gauche vers le haut et la valeur de l'air sous la courbe (AUC)= 0.951 qui est proche de 1, donc le modèle est précis.

On a trouvé que la variable "balanc" est la seule variable qui contribue dans l'explication du défaut bancaire, le modèle retenu est :

$$logit(\mathbb{P}(Y = 1 \mid X = x)) = log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_2 x_2$$

Aprés avoir refaire l'estimation pour le modèle avec la variable "balanc", on obtient :

```
> reg22=glm(defaut~balanc,family=binomial)
> reg22

Call:  glm(formula = defaut ~ balanc, family = binomial)

Coefficients:
(Intercept)       balanc
  -9.059491     0.005128

Degrees of Freedom: 49 Total (i.e. Null);  48 Residual
Null Deviance:      32.51
Residual Deviance: 14.53        AIC: 18.53
```

$\hat{\beta}_0$ = -9.059491 et $\hat{\beta}_2$ = 0.005128.

D'où le modèle retenu est : $logit(p(x)) = -9.059491 + 0.005128x_2$.

✠ La prédiction

Après avoir construit le modèle de régression logistique qui est défini dans la formule (3.12) pour relier le défaut bancaire et les variables client, le montant moyen mensuel d'utilisation de la carte de crédit et le revenu de client. Nous appliquons ce modèle pour faire des prédictions.

Afin de prédire la probabilité que le client possède un défaut à un montant mensuel donné nous utilisons la fonction "predict" qui renvoie pour chaque individu la probabilité qu'il possède un défaut "defaut", sous R comme suit :

```
> prev=predict(reg22,type="response")
> prev
```

On peut ainsi avoir par exemple une estimation de la probabilité d'avoir un défaut pour un client qui possède un montant mensuel d'utilisation de la carte de crédit de 758.1342851.
$\hat{p}(x_2 = 758.1342851) = \frac{e^{-9.059491+0.005128(758.1342851)}}{1+e^{-9.059491+0.005128(758.1342851)}} = 0.00564245 = 5.64245 \times 10^{-3}$

3. Estimation bayésienne

Dans cette section nous allons analyser les données précédentes traitées par le modèle logistique binaire classique en utilisant l'approche bayésienne, et cela via le logiciel WinBUGS qui prendra en entrée un modèle composé de lois a priori et de vecteurs de valeurs initiales et retournes les sorties de l'algorithme de Gibbs pour cette loi a posteriori, afin de configurer un modéle bayésien avec des distributions a priori normales informatives (la plus utilisée

pour les paramètres du modèle logistique). Nous déterminons les distribution a posteriori des paramètres de régression et nous appliquons l'approche MCMC.

On utilise la commande suivante pour insérer les données :

```
list(defaut=c(0,0,1,0,0,0,0,0,0,0,0,0,0,1,0,0,0,1,0,0,0,0,0,0,0
,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0), ..., n=50)
```

✠ spécification du modèle

On a la loi a posteriori pour β est définie par :

$$\pi(\beta \mid y) = \prod_{i=1}^{n} \left(\frac{e^{\underline{x}_i \beta y_i}}{1 + e^{\underline{x}_i \beta}} \right) \prod_{j=0}^{p} e^{\frac{-1}{2\sigma_j^2}(\beta_j - \mu_j)^2}.$$

En prenant pour les lois a priori des β_j, j=0,...3, la loi normale de paramètres $\mu_j = 0$ et $\sigma_j^2 = 1000$, j=0,...3, $\mathcal{N}(0, 1000)$.

Remarque

L'utilisation de la loi normale avec une grande variance ($\sigma_j^2 = 1000$), a permis de ne pas ajouter d'informations supplémentaires et de ne prendre en compte que celle apportée par la vraisemblance. Cette loi aurait pu être remplacée par d'autres lois, en particulier par la loi uniforme et la loi non informative de Jeffreys. La loi normale a été choisie pour sa particularité de conjugaison : on s'attend ainsi à observer des densités proches de gaussiennes pour les lois a posteriori.

Sous WinBUGS, on obtient :

```
model
{
for (i in 1:n) {
logit(p[i]) <- alpha + b.client*client[i] + b.balanc*balanc[i]+b.revenu*revenu[i]
defaut[i] ~ dbern(p[i])
}
alpha ~ dnorm(0.0,1.0E-4)
b.client ~ dnorm(0.0,1.0E-4)
b.balanc ~ dnorm(0.0,1.0E-4)
b.revenu ~ dnorm(0.0,1.0E-4)
}
```

Après avoir défini le modèle, on a donc besoin des valeurs initiales pour démarrer la chaine de Makov qui sont données par la commande "list", comme suit :

```
list(alpha=0, b.client=0, b.balanc=0,b.revenu=0)
```

Dans notre cas on a choisi 20000 comme le nombre d'itérations.

• **Résultats de l'estimation**

WinBUGS calcule également des statistiques relatives a la loi a posteriori, dans le quel on a généré une chaine de Markov avec 20000 échantillons, qui sont données dans le tableau suivant :

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
alpha	-15.79	8.399	0.5405	-34.06	-14.87	-1.569	1	20000
b.balanc	0.01004	0.003744	1.964E-4	0.004099	0.009548	0.01881	1	20000
b.client	-3.176	4.119	0.1946	-11.83	-3.01	4.661	1	20000
b.revenu	1.097E-5	1.46E-4	8.497E-6	-2.935E-4	1.304E-5	2.946E-4	1	20000

TABLE 3.1 – Résultats de l'estimation bayésienne

D'où le modèle est : $logit(p(\underline{x})) = -15.79 + 0.01004\beta_1 - 3.176\beta_2 + 1.097 \times 10^{-5}\beta_3$.

node représente le nom des paramètres inconnus.

mean est la moyenne a posteriori des paramètres, on a l'estimation a posteriori de β_2(b.balanc) et β_3(b.revenu) est positive et cela indique une augmentation du défaut du client et une estimation négative de β_1(b.client), cela indique que le statut du client croissant entraine une diminution de la valeur du défaut.

sd est l'écart-type a posteriori et le fait d'avoir un écart type faible pour chaque paramètre engendre un intervalle de confiance plus restreint pour chacun des paramètres ce qui signifie que l'estimation est plus précise et c'est le cas pour les paramètres associés aux variables "balanc" et "revenu" qui possèdent des écart-types très petits.

MC error est l'erreur standard de Monte Carlo, est calculée en multipliant la variance des valeurs échantillonnées par l'inverse de la taille effective de l'échantillon pour tenir compte de la corrélation. Une régle de base est que le MC error doit être inférieur à 5% de l'écart type du paramètre estimé (sd).

En effet, pour la variable balanc on a **sd** = 0.003744 et **MC error** = 0.0001964 et le rapport des deux est de $\frac{0.0001964}{0.003744}$=0.0524.

et pour la variable revenu, son **sd**=0.000146 et **MC error** = 0.000008497 et le rapport des deux est $\frac{0.000008497}{0.000146} = 0.0582$.

faisons de même pour la variable client avec un **sd** = 4.119 et **MC error** = 0.1946 et rapport des deux est de $\frac{0.1946}{4.119}$=0.047 ou 4.7% ce qui est inférieur à la régle empirique de 5% suggérant que l'erreur dûe à la simulation n'est pas assez importante pour justifier l'incertitude.

median est la médiane a posteriori.

start représente l'indice de départ des simulations (aprés rodage).

sample est le nombre de simulation utilisés pour approximer la distribution a posteriori (20000).

2.5% et **97.5**% sont les bornes inférieures et supérieures des intervalles de crédibilité. On constate que β_1(b.client) qui est le coefficient de la variable client dans le modèle n'est pas significativement différent de 0 puisque l'intervalle de confiance bayésien [-11.83 ;4.661] contient le zéro, de même pour la variable revenu. La variable balanc en revanche semble contribuer de maniére significative à la variation du défaut bancaire du client puisque β_2 est significativement différent de zéro car l'intervalle de crédibilité [0.004099 ;0.01881] ne contient pas zéro.

Remarque

On retrouve le même résultat que l'estimation classique concernant l'acceptation de la variable "balanc".

Une nouvelle estimation bayésienne de $\hat{\beta}_2$ est calculée en utilisant uniquement la variable "balanc" dans le modèle, les résultats obtenus sont :

$\hat{\beta}_0$ = -11.21 et $\hat{\beta}_2 = 0.006382$.

Le modèle est : $logit(p(x)) = -11.21 + 0.006382x_2$.

et la probabilité estimé est donnée par :

$$\hat{p}(x) = \frac{e^{-11.21+0.006382x_2}}{1+e^{-11.21+0.006382x_2}}.$$

La prédiction pour $x_2 = 758.1342851$ est donnée par :

$$\hat{p}(x) = \frac{e^{-11.21+0.006382(758.1342851)}}{1+e^{-11.21+0.006382(758.1342851)}} = 0.001706.$$

Comaparaison

	Prédiction de la prababilité
Méthode classique MV	5.64245×10^{-3}
Méthode bayésienne	1.706×10^{-3}

Comme $1.706\times10^{-3} < 5.64245\times10^{-3}$, donc $p(\underline{x})_{classique} > p(\underline{x})_{bayesienne}$. D'où, le client a moins de possibilités de faire défaut avec l'estimation bayésienne. On rappelle que l'estimation bayésienne est plus précise que l'estimation classique

✠ Diagnostics de convergence de la distribution a posteriori

Maintenant à la convergence, un MCMC crée un échantillon à partir de la distribution a posteriori et nous voulons généralement savoir si cet échantillon est suffisamment proche de la loi a posteriori pour être utilisé pour l'analyse, pour cela il existe plusieurs diagnostics qui seront présentés comme suit :

(a) Les séries chronologiques

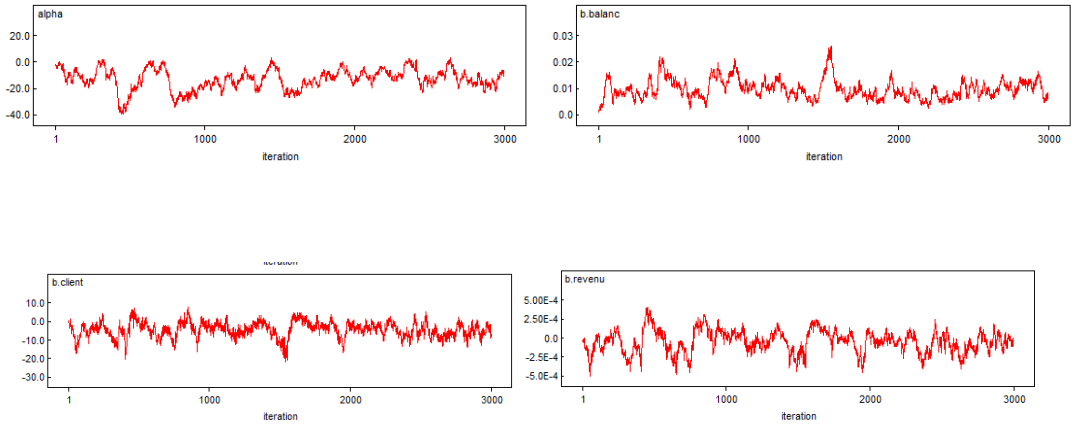
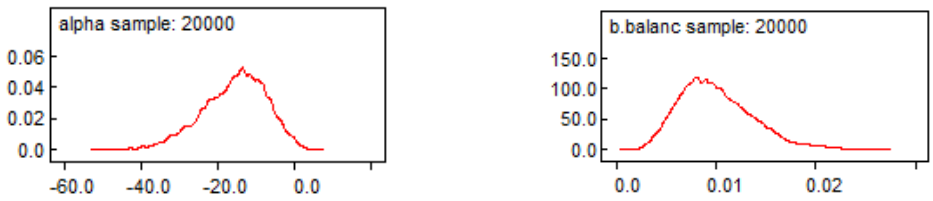


FIGURE 3.13 – Les graphes des séries chronologiques

D'aprés la figure (3.13) on déduit que la chaine de Markov est relativement stationnaire ceci implique que la chaine a atteint ou est proche de sa distribution stationnaire, on peut donc raisonnablement supposer que nos simulations sont tirés de la distribution a posteriori souhaitée $\pi(\beta \mid y)$ qui est une normale (voir tableau 3.1).

(b) Les densités a posteriori

Les représentations graphiques des densités a posteriori sous WinBUGS sont les suivantes :



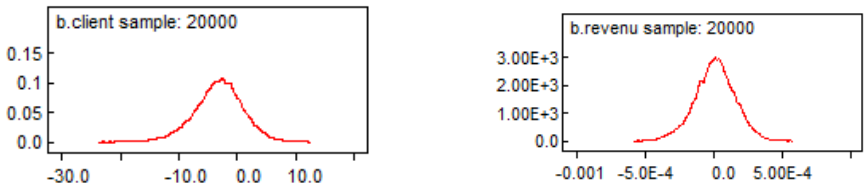


FIGURE 3.14 – Densités a postériori des paramètres du modèle

La qualité des graphes de la figure (3.14) est satisfaisante. Car les densités des lois a posteriori sont proches de la densité gaussiennes.

(c) Le diagnostic de Gelman-Rubin

Ce diagnostic évalue la convergence MCMC en analysant la différence entre plusieurs chaines en comparant les variances intra et inter chaine pour chaque paramètre du modèle et on dit que la convergence est aquise si la variance intra est similaire à la variance inter chaine et une grande différence entre les deux chaines indique une non convergence, les graphes de convergence sont donnés comme suit :

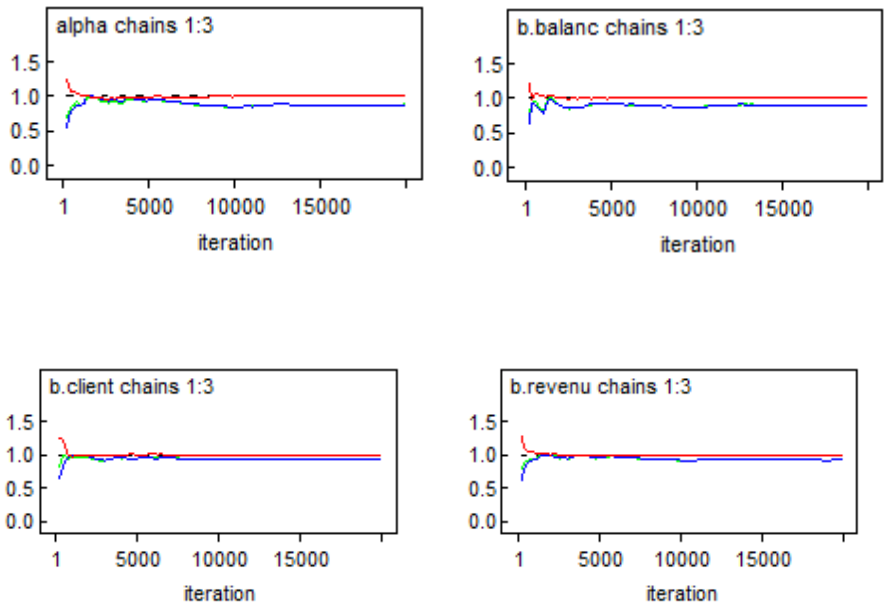


FIGURE 3.15 – Diagnostics de convergence de Gelman Rubin

La figure (3.15) montre trois lignes :

• La ligne rouge représente le graphe du facteur de réduction d'échelle $\sqrt{\hat{G}}$ pour chaque paramètre, ce facteur doit être proche de 1 pour dire qu'il y a convergence, on remarque que c'est le cas pour les variables client, balanc et revenu.

• La ligne verte représente le graphe de la variance inter chaine (B).

- La ligne bleu représente le graphe de la variance intra chaine (W).

On voit bien que les deux lignes (verte et rouge) correspondantes aux variances inter et intra chaine respectivement se rapprochent, ce qui signifie que les deux chaines ont convergés. De cela on conclut que la chaine converge vers sa distribution cible souhaitée.

✠ L'approximation de Laplace

Une autre approche bayésienne peut être utilisée à savoir l'approximation de Laplace qui consiste à déterminer l'estimateur $\hat{\beta}_{MAP}$ par le maximum a posteriori. On utilise la fonction "*lapl_aprx*" qui génére un mode μ et une matrice de variance-covariance σ pour une distribution a posteriori normale multivariée pour le logit bayesien. Le but de l'approximation de Laplace est d'estimer le mode a posteriori et la variance de chaque paramètre, qui seront donnés par :

```
> library(bayesdistreg)
> D1=data.frame(x11=client,x2=balanc,x3=revenu)
> lapl_aprx ( defaut,D1)
$mode
  (Intercept)           x11            x2            x3
-9.830710e+00 -2.333498e+00  6.472768e-03 -3.714481e-06

$var
              (Intercept)           x11            x2            x3
(Intercept)  4.928274e+01 -1.454767e+01 -1.215441e-02 -7.815501e-04
x11         -1.454767e+01  1.248693e+01 -1.861620e-03  3.870049e-04
x2          -1.215441e-02 -1.861620e-03  7.314063e-06  6.097219e-08
x3          -7.815501e-04  3.870049e-04  6.097219e-08  1.701573e-08
```

$$\mu = \beta_{MAP} = (-9.830710e+00\ , -2.333498e+00\ , 6.472768e-03\ , -3.714481e-06)$$

$$\Sigma = \begin{pmatrix} 4.928274e+01 & -1.454767e+01 & -1.215441e-02 & -7.815501e-04 \\ -1.454767e+01 & 1.248693e+01 & -1.861620e-03 & 3.870049e-04 \\ -1.215441e-02 & -1.861620e-03 & 7.314063e-06 & 6.097219e-08 \\ -7.815501e-04 & 3.870049e-04 & 6.097219e-08 & 1.701573e-08 \end{pmatrix}$$

Logiciels utilisés

1. **Logiciel R**

FIGURE 3.16 – Le logo du logiciel R

R est un langage de programmation et un environnement mathématique utilisé pour le traitement de données et l'analyse statistique. C'est un projet GXU (Général Public Licence) fondé sur le langage S et sur l'environnement développé dans les laboratoires Bell par John.Chambers et ses collègues. Le logiciel R a été développé dans les années 90 par Robert Gentleman et Ross Ihaka (Département de Statistique, Université d'Auckland, Nouvelle-Zélande). C'est un logiciel libre (avec code source) et peut être distribué librement. Le logiciel R comporte des moyens qui rendent possibles la manipulation des données et les calculs. Il a aussi la possibilité d'exécuter des programmes stockés dans des fichiers textes et excel. Ce logiciel trés puissant possède :

- Un système efficace de manipulation et de stockage des données,
- Un grand nombre d'outils pour l'analyse des données et les méthodes statistiques.
- Des moyens graphiques pour visualiser les analyses.
- Un langage de programmation simple et performant comportant : conditions, boucles.

Enfin, R est un logiciel performant en terme de calcul et de représentation graphique. C'est pourquoi, dans le cadre de notre étude statistique, nous l'avons utilisé afin d'estimer le modèle statistique et de calculer les prévisions.

2. **Logiciel WinBUGS**

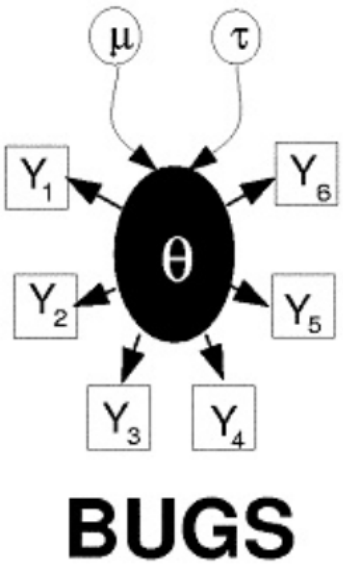


FIGURE 3.17 – Le logo du logiciel WinBUGS

WinBUGS est un logiciel statistique d'analyse bayésienne utilisant les méthodes MCMC. Il est basé sur le projet BUGS (Bayesian inference Using Gibbs Sampling) lancé en 1989, il peut-être intégré au logiciel R à l'aide du package R2WinBUGS dans R. Il a été développé par une équipe de chercheurs britanniques de la RMC biostatistics Unit Cambridge et de l'Imperial collège school of Medecine de londres.

WinBUGS nécesite une connaissance approfondie des statistiques bayésiennes pour créer et évaluer les modéles de manière apropriée. Notons qu'une des particularité lorsque l'on code dans un programme de type WinBUGS est que les lois normales ne prennent pas comme paramètres la moyenne et la variance mais moyenne et précision où la précision est définie comme l'inverse de la variance.

3. **Logiciel latex**

FIGURE 3.18 – Le logo du logiciel Latex

Latex, prononcé /la.tek/, est un langage informatique performant créé par Leslie-Lamport en 1983, conçu pour la rédaction des documents scientifiques. Il est notamment utilisé par les mathématiciens pour la qualité du rendu et génération des formules mathématiques. A la différence des traitements de texte usuels (tel MS Word), il ne s'agit pas d'un logiciel dit "WYSIWYG" (what you see is what you get), c'est à dire lors de la frappe, on ne voie pas de document tel qu'il sera imprimé, il faut le compiler avec LATEX puis utiliser un programme de visualisation. LATEX permet de produire des documents sans soucier de leur exacte mise en page. Ce sont des commandes préfixées qui serviront à la mise en page automatique à l'aide d'un environnement LATEX, et peut être facilité par l'utilisation d'un éditeur graphique. Sous Windows, l'environnement le plus utilisé est Miktex.

Conclusion

Historiquement l'étude des modèles décrivant les modalités prises par des variables qualitatives date de plusieurs années, les travaux les plus marquants de cette époque sont ceux de Berkson consacrés aux modèle logistique.

Nous avons étudié dans ce mémoire la classe des modèles GLM qui contient les modéles linéaires classiques et aussi les modèles logistique pour lesquels nous avons abordé deux méthodes d'estimation pour les paramètres : la méthode du maximum de vraisemblance et la méthode bayésienne renforcée par les techniques d'approximation MCMC.

La qualité de l'estimation est analysée par des quantités tels que le coefficient de détermination dans le cas des modèles classiques et les pseudos R^2 dans le cas des modèles logistique, la matrice de confusion et les diagrammes ROC, la courbe logistique ainsi que les représentations graphiques des fonction de lien et cela afin d'évaluer la qualité du modèle logistique.

Un exemple et une application sont présentés afin d'illustrer la modélisation logistique binaire en utilisant les logiciels R et WinBUGS qui permettent d'implémenter l'approximation de Laplace ainsi que l'algorithme de Gibbs, de cela on voit bien que les performances des ordinateurs ont rendu faisable des procédés de simulation efficaces et la disponibilité des programmes informatique a facilité le calcul des probabilité a posteriori qui était jusque la d'une complexité décourageante, et nous avons obtenu la même conclusion par les deux méthodes d'estimation.

Faute du covid-19 qui nous a privé de faire un stage pour avoir des données réelles, nous nous sommes contenté de présenter deux applications dont les données de la premiére application sont tirées d'un livre et les données de la deuxiéme sont tirées d'un cours. Nous avons établi nous-même toutes les estimations présentés et nous pouvons conclure que même si les deux méthodes d'estimations présentées ont permis d'accepter la même variable mais la valeur du coefficient de cette variable est plus grande dans l'estimation bayésienne, ce qui nous a permis d'avoir dans les résultats de prévision, moins de chances de faire défaut que le résultat obtenu avec l'estimation classique.

Terminons sur ce constat que les méthodes de régression sont des méthodes très puissantes mais qui doivent être utilisées avec beaucoup de discernement et de prudence. Évidemment la régression logistique polytomique reste des modèles très riches à développer et qu'on n'a pas pu faire dans le cadre de ce mémoire .

La bibliographie jointe démontre par son volume que les modèles logit ont ces dernières années su séduire un nombre grandissant de statisticiens et de chercheurs par la simplicité de ses applications et interprétation.

Annexe

1. Le raisonnement proportionnel

Il est parfois possible d'éviter le calcul de la densité marginale de x :

$$m(x) = \int_{\Theta} f(x \mid \theta)\pi(\theta)d\theta$$

On dit que deux fonctions f et g définies sur le même espace $\mathcal{Y}$ sont proportionnelles et on note $f \propto g$ s'il existe une constante a telle que :

$$f(y) = ag(y)$$

pour tout $y \in \mathcal{Y}$

Propriété 3.3.1. *La relation $\propto$ est une relation d'équivalence, en particulier :*

$$f \propto g \quad et \quad g \propto h \Longrightarrow f \propto h$$

Dans le contexte bayésien on a :

$$\pi(\theta \mid x) \propto f(x \mid \theta)\pi(\theta)$$

En tant que fonction de θ, $\pi(\theta \mid x)$ et $\pi(\theta)$ sont proportionnelles, la quantité $\frac{1}{m(x)}$ étant une constante qui ne dépend pas de θ, elle joue le rôle de a qui apparait dans la définition au sens où elle ne dépend pas de θ. L'écriture $\pi(\theta \mid x) \propto f(x \mid \theta)\pi(\theta)$ est souvent reformulée de la façon suivante :

$$\pi(\theta \mid x) \propto \mathcal{L}(x \mid \theta)\pi(\theta)$$

où $\mathcal{L}(x \mid \theta)$ désigne la vraisemblance.

2. Fonction de perte quadratique pondérée

On appelle fonction de perte quadratique pondérée tout fonction de forme :

$$\ell(\theta, \delta(x)) = \omega(\theta)(\theta - \delta)^2$$

où $\omega(\theta)$ est une fonction positive.

Cette perte présente la caractéristique intéressante de permettre à l'erreur quadratique $(\theta-\delta)^2$ d'être pondérée par une fonction de θ.
ainsi l'estimateur bayésien associé à la loi a priori π et au coût quadratique pondéré est donné par :

$$\delta^\pi = \frac{\mathbb{E}^\pi[\omega(\theta)\theta \mid x]}{\mathbb{E}^\pi[\omega(\theta) \mid x]}$$

3. Fonction de coût linéaire par morceaux

Definition 3.3.1. La fonction de coût linéaire par morceaux sera définie comme suit :

$$\ell(\theta, \delta(x)) = \begin{cases} k_2(\theta - \delta(x)) & \theta > \delta(x) \\ k_1(\delta(x) - \theta) \quad \theta \leq \delta(x), k_1, k_2 & \text{sont des constantes positives.} \end{cases}$$

L'estimateur de bayes associé est le fractile $\frac{k_2}{k_1+k_2}$ de $\pi(\theta \mid x)$ on donne ainsi L'estimateur de Bayes associé à la loi a priori π et a la fonction de coût linéaire par morceaux qui est le fractile $\frac{k_2}{k_1+k_2}$ de la distribution a posteriori $\pi(\theta \mid x)$.

En particulier, si $k_1 = k_2$, l'estimateur de Bayes est la mediane de $\pi(\theta \mid x)$.

4. Minimaxité

Théorème 3.3.2. ***(Judith Rousseau 2009)***

Le risque de Bayes est toujours plus petit que le risque minimax

$$\underline{R} = \sup_\pi r(\pi) = \sup_\pi \inf_{\delta \in \mathcal{D}} r(\pi, \delta) \leq \overline{R} = \inf_{\delta \in \mathcal{D}} \sup_\theta R(\theta, \delta)$$

$\underline{R}$ désigne le risque maxmin, et $\overline{R}$ désigne le risque minimax.

Lemme 3.3.3. (Christian P. Robert 2006)

Si δ^π est un estimateur de bayes pour π et si $R(\theta, \delta^\pi) \leq r(\pi)$ pour tout θ dans le support de π, l'estimateur δ^π est minimax et π est la distribution la moins favorable.

Les distributions les moins favorables sont celles qui ont le risque de bayes le plus grand.

Lemme 3.3.4. Soit π_n une suite de lois a priori propres. S'il existe un estimateur de bayes δ^*, tel que $\forall\theta \in \Theta$:

$$R(\theta, \delta^*) \leq \lim_{n \longrightarrow +\infty} r(\pi_n) < \infty$$

alors δ^* est minimax.

Théorème 3.3.5. ***(Robert(2001))***

Si $\ell(\theta, \delta)$ est une fonction de perte continue et convexe en θ $\forall\theta \in \Theta$, alors il existe un estimateur minimax.

5. Modèle linéaire classique simple

Les ventes mensuelles d'huile au sein de l'entreprise Cevital (Béjaia), durant la période 2013-2016, peuvent être appliqué par le temps.

Le modèle est donné par :

$$ventes_i = b_0 + b_1 temps_i + \epsilon_i, \quad i = \overline{1.48}$$

Les données sont présentées dans le tableau suivant :

temps	ventes	temps	ventes
1	470167	25	642949
2	355946	26	603559
3	556447	27	685384
4	321028	28	699363
5	449438	29	662647
6	503373	30	809186
7	471372	31	714734
8	567199	32	718835
9	466799	33	786479
10	429591	34	790237
11	580271	35	680734
12	537257	36	638097
13	531577	37	540038
14	432456	38	652281
15	489200	39	648593
16	613084	40	702104
17	671969	41	741572
18	689404	42	845232
19	505472	43	928942
20	627782	44	738119
21	356946	45	757607
22	444503	46	700911
23	530801	47	845128
24	568279	48	862128

TABLE 3.2 – Les ventes mensuelles d'huile [2013-2016]

Le nuage des points est représenté dans la figure suivante :

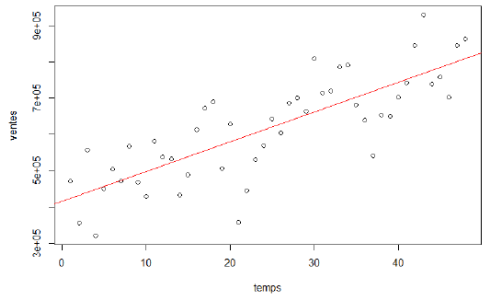


FIGURE 3.19 – Nuage des points et la droite de régression

Le modéle estimé est donné par : $\widehat{ventes} = 86810+8194\text{temps}$.

Les résultats de l'estimation sont résumés dans le tableau suivant :

Paramétres $\hat{b}_i$	**Ecart-type**	**Ratios de student**	**Intervalle de confiance**
$\hat{b}_0$=86810	$\hat{\sigma}_{\hat{b}_0}$=25456.4	$t^*_{\hat{b}_0}$=16.310	IC_{b_0}=[363952.674 ;466434.74]
$\hat{b}_1$=8194	$\hat{\sigma}_{\hat{b}_1}$=904.5	$t^*_{\hat{b}_1}$=9.059	IC_{b_1}=[6373.232 ;10014.39]

En utilisant les valeur de : $\sigma^2_\epsilon = 86810$ $F^* = 82.07$ $R^2 = 0.6408$ $F_{(1,46,0.95)} = 4.05$ $t_{(0.025,46)} = 2.013$.

En appliquant les tests de Fisher et de Student on conclut :

- Le modèle est globalement significatif.
- Le temps est une variable significative des ventes mensuelle.

6. Modèle linéaire classique multiple

Dans cette partie on veut expliquer la variable **ventes** (variable à expliquer) par les ventes des différents article d'huiles (variables explicatives) : **Huileelio5L** et **Huilefridor5L**, pour l'année 2014.

Le modèle est donné par :

$$\text{ventes} = b_0+b_1(\text{Huileelio5L})+b_2(\text{Huilefridor5L})$$

Les données des ventes mensuelles (en litre) de l'année 2014, sont présentées dans le tableau suivant :

ventes	Huileelio5L	Huilefridor5L
14824	4776	4878
13756	4469	4474
15225	4939	4921
16555	5301	5346
15377	5161	4889
12784	4059	4059
9932	3336	3201
13496	4362	4250
15000	4721	4927
16451	5152	5373
16699	5131	5547
15222	4961	5022

Le modèle estimé est donné par :

$$\widehat{ventes} = 88.558+1.080(\text{Huileelio5L})+1.993(\text{Huilefridor5L})$$

Les résultats de l'estimation sont résumés dans le tableau suivant :

Paramètres $\hat{b}_i$	**Ecart-type**	**Ratios de student**	**Intervalle de confiance**
$\hat{b}_0$=88.558	$\hat{\sigma}_{\hat{b}_0}$=417.4735	$t^*_{\hat{b}_0}$=0.212	IC_{b_0}=[-855.8323729 ;1032.948777]
$\hat{b}_1$=1.080	$\hat{\sigma}_{\hat{b}_1}$=0.3230	$t^*_{\hat{b}_1}$=3.344	IC_{b_1}=[0.3495133 ;1.810750]
$\hat{b}_1$=1.993	$\hat{\sigma}_{\hat{b}_1}$=0.2786	$t^*_{\hat{b}_1}$=7.154	IC_{b_1}=[1.3628013 ;2.623116]

En utilisant les valeur de : $\sigma^2_\epsilon = 155.5 \quad F^* = 828.1 \quad R^2 = 0.994 \quad F_{(2,9,0.95)} = 4.256 \quad t_{(0.025,9)} = 2.262$.

En appliquant les tests de Fisher et de Student on conclut :

- Le modèle est globalement significatif.
- Les ventes (elio5L et fridor5L) sont significatives de la vente totale des huiles.

7. Les composantes des lois de la famille exponentielle

(a) Distribution de poisson

Soit $Y \sim \mathcal{P}(\mu)$, sa fonction de masse de est donnée par :

$$\begin{aligned} f(y,\mu) &= \frac{e^{-\mu}\mu^y}{y!} \\ &= exp\left\{log\left(\frac{e^{-\mu}\mu^y}{y!}\right)\right\} \\ &= exp\left\{log(e^{-\mu}) + log\left(\frac{\mu^y}{y!}\right)\right\} \\ &= exp\left\{-\mu + log(\mu^y) - log(y!)\right\} \\ &= exp\left\{-\mu + ylog(\mu) - log(y!)\right\} \end{aligned}$$

Par identification avec la formule (2.3), Les composantes du modèle sont données par :

$\omega = log(\mu)$, $\quad b(\omega) = \mu = exp(\omega)$, $a(\phi) = 1$, et $c(y,\phi) = -log(y!)$

La fonction de lien associée est $g(\mu) = log(\mu)$.

(b) Distribution Bernoulli

Soit $Y \sim \mathcal{B}(y,\mu)$, sa distribution de probabilité est défini par :

$$\begin{aligned} f(y,\mu) &= \mu^y(1-\mu)^{1-y} \\ &= exp\left\{log(\mu^y(1-\mu)^{1-y})\right\} \\ &= exp\left\{log(\mu^y) + log(1-\mu)^{1-y}\right\} \\ &= exp\left\{ylog(\mu) + (1-y)log(1-\mu)\right\} \\ &= exp\left\{ylog(\mu) + log(1-\mu) - ylog(1-\mu)\right\} \\ &= exp\left\{ylog\left(\frac{\mu}{1-\mu}\right) + log(1-\mu)\right\} \end{aligned}$$

Cette equation a une forme de la famille exponentielle naturelle dont on donne les paramètres :

$\omega = log\left(\frac{\mu}{1-\mu}\right)$, $\quad b(\omega) = -log(1-\mu) = log(1+exp(\omega))$, $\quad a(\phi) = 1$

La fonction lien canonique est $g(\mu) = log\left(\frac{\mu}{1-\mu}\right)$ elle est appelée fonction logit.

(c) Distribution Binomiale

Soit $Y \sim \mathcal{B}(y, n, \mu)$, sa distribution est donnée par :

$$\begin{aligned} f(y,n,\mu) &= \binom{n}{y}\mu^y(1-\mu)^{n-y} \\ &= exp\left\{log\binom{n}{y}\mu^y(1-\mu)^{n-y})\right\} \\ &= exp\left\{log\binom{n}{y} + log(\mu^y) + log(1-\mu)^{n-y}\right\} \\ &= exp\left\{log\binom{n}{y} + ylog(\mu) + (n-y)log(1-\mu)\right\} \\ &= exp\left\{log(ny) + ylog(\mu) + nlog(1-\mu) - ylog(1-\mu)\right\} \\ &= exp\left\{ylog\left(\frac{\mu}{1-\mu}\right) + nlog(1-\mu) + log\binom{n}{y}\right\} \end{aligned}$$

$\omega = log(\frac{\mu}{1-\mu}), \quad b(\omega) = -nlog(1-\mu) = nlog(1+exp(\omega)), \quad a(\phi) = 1, \quad c(y,\phi) = log\binom{n}{y}$
la fonction de lien canonique est $g(\mu) = log(\frac{\mu}{1-\mu})$.

(d) distribution gamma

Soit $Y \sim \mathcal{G}(\mu, \alpha)$, sa densité sera donnée par :

$$\begin{aligned} f(y,\mu,\alpha) &= \frac{\alpha^\mu y^{\mu-1}e^{-\alpha y}}{\Gamma(\mu)} \\ &= exp\left\{log\left(\frac{\alpha^\mu y^{\mu-1}e^{-\alpha y}}{\Gamma(\mu)}\right)\right\} \\ &= exp\left\{log(\alpha^\mu + log(y^{\mu-1}) + log(e^{-\alpha y}) - log(\Gamma(\mu))\right\} \\ &= exp\left\{\mu log(\alpha) + (\mu-1)log(y) - \alpha y - log(\Gamma(\mu))\right\} \\ &= exp\{-\alpha y + \mu log\alpha + (\mu-1)log\ y - log(\Gamma(\mu))\} \\ &= exp\left\{\frac{-\frac{\alpha}{\mu}y + \frac{\mu}{\mu}log\alpha}{\frac{1}{\mu}} + (\mu-1)log\ y - log(\Gamma(\mu))\right\} \\ &= exp\left\{\frac{\frac{\alpha}{\mu}y - \frac{\mu}{\mu}log\alpha}{\frac{-1}{\mu}} + (\mu-1)logy - log(\Gamma(\mu))\right\} \end{aligned}$$

Cette expression a une forme de la famille de dispersion exponentielle donne les paramètres sont :
$\omega = \frac{\alpha}{\mu}$, $b(\omega) = log(\alpha) = log(\mu\omega)$ $a(\phi) = \phi = \frac{-1}{\mu}$,$c(y,\phi) = (\mu-1)logy - log(\Gamma(\mu))$.

D'aprés la propriété (2.4) on a $\mathbb{E}(y) = b^{'}(\theta) = \frac{\mu}{\mu\omega} = \frac{1}{\omega} = \theta^{-1}$.

Et d'aprés (2.5) on a $\mathbb{V}(y) = b^{''}(\omega)a(\phi) = \frac{\mu}{\alpha^2}$.

La fonction de lien associée est $g(\mu) = (\frac{\mu}{\alpha})^{-1}$, qui est appelée la fonction de lien inverse.

8. La courbe ROC (Receiver Operating Characteristic)

La courbe ROC est une représentation graphique de la relation existante entre la sensibilité et la spécificité pour toutes les valeurs seuils possibles. L'ordonnée représente la sensibilité et l'abscisse correspond à la spécificité. Sa construction nécessite l'emploi d'un logiciel de calcul spécialisé.

La figure ci-dessous représente une courbe ROC :

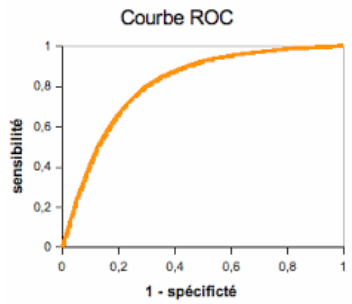


FIGURE 3.20 – La courbe ROC

Il est possible de caractériser numériquement la courbe ROC en calculant la surface située sous la courbe. C'est le critère AUC (Area Under Curve) (qui peut être calculer grace à un algorithme de tri), il exprime la probabilité de placer un individu positif devant un négatif. qui est définie par :

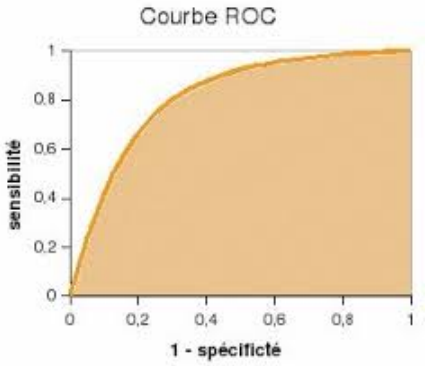


FIGURE 3.21 – Représentation graphique de l'aire sous la courbe ROC : AUC

C'est une mesure de performance du modèle dans la prédiction, un modèle parfait aura une mesure AUC de 1. Ainsi, plus le modèle est précis, plus la courbe ROC est proche du coin gauche du graphique vers le haut et la mesure AUC est proche de 1.

9. La procédure pas à pas

Les étapes seront résumé comme suit :

1. Définir l'intervalle $I = (L, U)$

2. Échantillonner uniformément $\beta^{(k)}$ sur I. Choisir la valeur échantillonnée si elle est incluse sur l'ensemble S, sinon passer à l'étape suivante.

3. Réduire l'intervalle I, rejeter la valeur $\beta^{(k)}$ et recommencer l'étape précédente.

à l'étape (1) Pour définir l'intervalle I dans l'étape 1 nous devons tout d'abord choisir un intervalle de longueur arbitraire w autour de $\beta^{(k-1)}$ puis éloigner la borne inférieure d'une distance w jusqu'à ce que celle ci soit exclue de l'ensemble S. La valeur obtenue composera la borne in inférieure L de l'intervalle I. Nous effectuons la même procédure pour définir la borne supérieure U afin d'obtenir l'intervalle $I = (L, U)$ qui, idéalement, seraient tel que $I \backslash S = \emptyset$ Or, cette condition n'est pas nécessaire. Évidemment, il est parfois même impossible d'obtenir $I \backslash S = \emptyset$ lorsque $h(\beta \mid Y, X)$ n'est pas unimodale.

Ensuite à l'étape (2), nous échantillonnons uniformément une valeur de $\beta^{(k)}$ sur l'intervalle I. Nous évaluons la valeur de $\beta^{(k)}$. Si cette valeur ne fait pas partie de l'ensemble S nous rejetons $\beta^{(k)}$ si par contre $\beta^{(k)}$ est incluse dans S, nous gardons cette valeur.

Si nous rejetons la valeur $\beta^{(k)}$ à l'étape (2), nous réduisons l'intervalle I à l'étape (3). Si la valeur $\beta^{(k)}$ rejetée est plus petite que $\beta^{(k-1)}$, L devient $L = \beta^{(k)}$, si toutefois la valeur $\beta^{(k)}$ rejetée est plus grande que $\beta^{(k-1)}$, U devient $U = \beta^{(k)}$.

Nous remarquons qu'en tout temps $\beta^{(k-1)} \in$ I. Ainsi, la procédure pas à pas nous garantit qu'elle prendra fin à un certain moment et que nous obtiendrons une valeur $\beta^{(k)} \in$ S (Neal (2003)). De plus, la procédure devrait s'effectuer sur un nombre plutôt restreint d'itérations puisque l'intervalle I se refermera de plus en plus sur l'ensemble S non vide. Aussi, nous pouvons montrer que chaque itération laisse la distribution stationnaire invariante. Ainsi pour chaque itération aprés convergence de la chaine en sa distribution stationnaire, nous échantillonnons sur la densité a posteriori.

Y	X	Y	X
CHD	**Age**	**CHD**	**Age**
0	20	0	38
0	23	0	38
0	24	0	39
0	25	1	39
1	25	0	40
0	26	1	40
0	26	0	41
0	28	0	41
0	28	0	42
0	29	0	42
0	30	0	42
0	30	1	42
0	30	0	43
0	30	0	43
0	30	1	43
1	30	0	44
0	32	0	44
0	32	1	44
0	33	1	44
0	33	0	45
0	34	1	45
0	34	0	46
1	34	1	46
0	34	0	47
0	34	0	47
0	35	1	47
0	35	0	48
0	36	1	48
1	36	1	48
0	36	0	49
0	37	0	49
1	37	1	49
0	37	0	50
1	50	1	57
0	51	1	57
0	52	0	58
1	52	1	58
1	53	1	58
1	53	1	59
1	54	1	59
0	55	0	6 0
1	55	1	60
1	55	1	61
1	56	1	62
1	56	1	62
1	56	1	63
0	57	0	64
0	57	1	64
1	57	1	65
1	57	1	69

Y	X_1	X_2	X_3
defaut	**client**	**balance**	**revenu**
0	0	758.1342851	33220.5755
0	0	930.7169386	46501.2757
1	1	2177.1508689	17659.7478
0	0	0.0000000	24892.9157
0	1	565.8300588	21042.2277
0	0	1076.1265844	23632.5203
0	1	1031.8699298	18668.4835
0	1	768.4037418	15417.8415
0	1	690.4210492	19273.7324
0	0	469.8442413	51308.3185
0	0	617.8600254	50177.7734
0	1	1779.0496987	15689.7766
0	0	0.0000000	47669.7041
1	0	1928.2802834	35492.1282
0	0	1123.7192597	56217.6849
0	0	794.6461084	41033.5877
0	1	1428.0668825	19818.2917
1	0	1028.7672066	40346.8333
0	0	1006.2029775	50501.7615
0	0	691.7517135	45420.9691
0	0	524.8381501	41268.4236
0	0	493.9141608	37409.1839
0	0	1520.9804782	37510.5394
0	0	665.0397566	47062.2531
0	1	1520.4421011	18462.4307
0	0	1094.7804726	34190.8765
0	0	735.0910408	44933.3232
0	0	681.6935764	33327.1130
0	0	376.0345440	50748.2717
1	0	2037.9433537	43016.0722
0	1	1630.1995895	14232.6615
0	1	0.0000000	21881.7059
0	0	184.4275319	36731.6270
0	0	349.6606656	39391.3223
0	0	673.5552703	49169.7285
0	0	82.7245249	42048.4448
0	0	1247.1206052	50539.9074
0	1	640.6395430	29236.6302
0	0	273.4469724	52492.7498
0	0	0.0000000	35377.1407
0	0	803.8311868	33417.7724
0	0	1248.3757490	37469.8646
0	0	0.0000000	40348.3142
0	1	1307.2049729	19381.5413
0	0	0.0000000	45793.3930
0	0	1186.0987051	50353.9254
0	1	1037.5730180	18769.5790
1	1	1878.0011459	17473.1840
0	0	736.2348369	36313.6336
0	0	260.1621754	33551.7153

Bibliographie

[1] ANDERSON, J.A. *Separate sample logistic discrimination.* Biometrika, 1972, 59, 19-35.

[2] BEL, L et al. *Le Modèle Linéaire et ses Extensions : Modèle linéaire général, modèle linéaire généralisé, modèle mixte, plans d'expériences.* Paris, Ellipses, 2016, 328p.

[3] BERKSON, J. *Application of the logistic function to bio-assay.* Journal of the American Statistical Association, 1944, V(39), 357-365.

[4] CHESNEAU, C. *Sur l'Estimateur du Maximum de Vraisemblance (EMV).* France : Université de Caen, cours de Licence, 2017, 68p.

[5] CHEN, M.H et al. *Properties and Implementation of Jeffreys's Prior in Binomial Regression Models.* Journal of the American Statistical Association, 2008, V(108), 1659-1664.

[6] COWLES, M.K et CARLIN, B.P. *Markov chain monte carlo convergence diagnostics* : A comparative review. Journal of the American Statistical Association, 1996, 91, 883-904.

[7] COLLETT, D. *Modelling binary data.* 2^e éd. London, Chapman et Hall, 2014, 369p.

[8] COX, D.R et SNELL, E.J. *The Analysis of Binary Data.* 2^e éd. London, Chapman and Hall-CRC, 1970, 240p.

[9] DAVISON, A.C. *Statistical Models.* 1^e éd. New York, Cambridge University Press, 2008, 738p.

[10] DUPUIS, J. *Statistique bayésienne et algorithmes MCMC*, Cours de Master 1 (IMAT), 2007, 16p.

[11] DESJARDINS, J. *L'analyse de régression logistique. Tutorial in quantitative Methods for Psychology*, 2005, V(1), 35-41.

[12] DODGE, Y et ROUSSON, V. *Analyse de régression appliquée.* 2^e éd. Paris, Dunond, 2004, 280p.

[13] FERRATY, F et VIEU, P. *Statistique fonctionnelle : Modèles non paramétriques de régression.* Toulouse. France : Université Paul Sabatier, Notes de cours de DEA, 2002.

[14] FIRTH, D. *Bias Reduction of Maximum Likelihood Estimates*. Biometrika, 1993, V(80), 27-38.

[15] FOONG, A.P, HU, Y.H et HEISEY, D.M. *Logistic regression in an adaptive Web cache*. IEEE Internet Computing, 1999, V(3), 27-36.

[16] FERMIN, A. *Le Modèle linéaire généralisé (glm)*, cours, 2015.

[17] GALTON, F. *Regression Towards Mediocrity in Hereditary Stature*. The Journal of the Anthropological Institute of Great Britain and Ireland, 1886, V(15), 246-263.

[18] GUAYADER, A. *Statistique*. France : Université Sorbonne, cours de Master mathématiques et applications, 2020, 115p.

[19] GENKIN, A et al. *Large-scale bayesian logistic regression for text categorization*. Technometrics, 2007, V(49), 291-304.

[20] GILLET, A et al. *Principaux modéles utilisés en régression logistique*. Biotechnol. Agron. Soc. Environ, 2011, V(15), 425-433.

[21] GOURIEROUX, M. *Asymptotic Properties of the Maximum Likelihood Estimator in Dichotomous Logit Models*. J Econom, 1981, V(17), 83-97.

[22] GAUDART, J et al. *Modèles linéaires à effets mixtes*, 2010, 12p.

[23] GAMERMAN, D. *Markov Chain Monte Carlo : Stochastic Simulation for Bayesian Inference*. 2^e éd. London, Chapman et Hall CRC Press, 2006, 342p.

[24] GELFAND, A.E et ADRIAN, F.M. SMITH. *Sampling-based approaches to calculating marginal densities*. Journal of the American Statistical Association, 1990, V(85), 398-409.

[25] GORDOVIL, M, GUARDIA, O, PERO, C et FUENTE, S. *Classical and Bayesian Estimation in the Logistic Regression Model Applied To Diagnosis of Child Attention Deficit Hyperactivity Disorder*. Psychological Reports, 2010, V(106), 1-15.

[26] HASTIE, T, TIBSHIRANI, R et FRIEDMAN, J. *The Elements of Statistical Learning*. 2^e éd. Springer-Verlag, 2009, 809p.

[27] HOSMER, D.W, LEMESHOW, S et STURDIVANT, R.X. *Applied logistic regression*. 3^e éd. John Wiley et Sons, 2013, 508p.

[28] HILBE, J.M. *Practical Guide to Logistic Regression*. Taylor and Francis Group, 2016, 162p.

[29] HOSMER, D.W et LEMESHOW, S. *Applied logistic regression*. 2^e éd. John Wiley et Sons, 2000, 397p.

[30] KUTNER, M.H et al. *Applied Linear Statistical Models*. 5^e éd. 2004, 1415p.

[31] LAGHA, K. *Cours Statistique Bayésienne*, master 1 Statistique et Analyse Décisionnelle. Université de A.MIRA, Bejaia, 2014, 50p.

[32] LALANNE, C, GEORGES, S et PALLIER, C. *Statistique appliquées à l'expérimentation en sciences humaines*, 133p.

[33] MENARD, S. *Applied logistic regression analysis : Quantitative Applications in the Social Sciences*, 2001, 120p.

[34] MILLOT, G. *Comprendre et réaliser les tests statistiques à l'aide de R*. 4^e éd. 2018, 960p.

[35] MILA, A et MICHAILIDES, T.J. *Use of Bayesian Methods to Improve Prediction of Panicle and Shoot Blight Severity of Pistachio in California*. Phytopathology, 2006, V(96), 1142-1147.

[36] MAKOWSKI, D. *Une extension du modèle linéaire : le modèle linéaire généralisé*. Paris, cours de Master en Agronomie, 2016, 12p.

[37] MAKOWSKI, D et MONOD, H. *Analyse statistique des risques agro-environnementaux Etudes de cas*. France, Springer-Verlag, 2011, 171p.

[38] MCCULLAGH, P et NELDER, J.A. *Generalized Linear Models*. 2^e éd. London, chapman and hall, 1989, 526p.

[39] MAK, T. *Solving Non-Linear Estimation Equations*. J. Roy. Stat. Soc. Ser. B, 1993, V(55), 945-955.

[40] NEAL, R.M. *Slice sampling*. Annals of statistics,2003, V(31), 705-74.

[41] NELDER, J.A et WEDDERBURN, R.W.M. *Generalized Linear Models*. Journal of the Royal Statistical Society, 1972, V(135), 370-384.

[42] PALMA, L, BEJA, P et RODRIGUES, M. *The use of sighting data to analyse Iberian lynx habitat and distribution*. Journal of Applied Ecology, 1999, V(36), 812-824.

[43] POHAR, M et al. *Comparaison of logistic regression and linear discriminant analysis : A simulation study*, 2004, V(1), 143-161.

[44] PIYUSH, R. *Bayesian Logistic Regression : Bayesian Generative Classi cation Topics in Probabilistic Modeling and Inference* , 2019, 28p.

[45] RUCH, J.J et CHABANOL, M.L. *Chaine de markov*. Bordeaux 1, Cours de préparation à l'agrégation, Université, 2013, 27p.

[46] ROBERT, C.P et CASELLA, G. *Méthodes de Monte-Carlo avec R*. 1^e éd. Paris, Springer, 2011, 273p.

[47] RAKOTOMALALA R. *Econométrie La régression linéaire simple et multiple*. Lyon 2, Cours de Licence, 2018, 183p.

[48] RAKOTOMALALA R. *Pratique de la Régression Logistique*. Lyon 2, Cours de Licence, 2015, 272p.

[49] RAFTERY, A.E. et LEWIS, S.M. *One Long Run with Diagnostics : Implementation Strategies for Markov chain Monte Carlo*. Statistical Science, 1992, V(7), 493-497.

[50] ROBERT, C.P. *Le choix bayésien principe et pratique*. 1^e éd. Paris, Springer-Verlag, 2006, 654p.

[51] SAADI, N. *Statistique inférentielle*, cours de 2éme année stid, Université A.MIRA, Bejaia, 2020, 49p.

[52] SUHNER M C et PROCACCIA, H. *Démarche bayésienne et application à la sureté de fonctionnement*, Ed Lavoisier, France, 2003, 410p.

[53] THOMAS, P. Ryan, *Some issues in logistic regression*. Communications in Statistics-Theory and Methods, 2000, V(29), 9-10.

[54] TOMAS, J. *Modeles lineaires et GLMS Analyse logit et Regression De Poisson Analyse d'un Portfeuille d'assurance Algorithme IRWLS avec R*, Institut de Science Financiére et d'Assurances-Université Claude Bernard Lyon 1 France.

[55] WOOD, S.N, *Generalized Additive Models : An Introduction with R*, 1^{st} éd, 2006, 397p.

[56] WEGLARCZYK, S et al. *Kernel density estimation and its application.* ITM Web of Conferences, 2018, V(23), 8p.

[57] WHITE, R, PEARSON, J, WILSON, J. *JIT manufacturing : A survey of implementations in small and large U.S. manufacturers.* Management Science, 1999, V(45), 1-15.