RÉPUBLIQUE ALGÉRIENNE DÉMOCRATIQUE ET POPULAIRE
MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITÉ MENTOURI CONSTANTINE
FACULTÉ DES SCIENCES EXACTES

DÉPARTEMENT DE MATHÉMATIQUES

N° d'ordre :.....
N° de série :.....

MÉMOIRE DE MAGISTÈRE EN MATHÉMATIQUES

Option :**Probabilités Statistiques**

THEME

COMPARAISON DES MODÈLES PROBIT ET LOGIT ET APPLICATION EPIDEMIOLOGIQUE

Présenté par :**MAHIDEB SAADIA**

Devant le jury composé de :

Président : RAHMANI Fouad.L , Maitre de Conférences, Université de Constantine

Rapporteur : CHIKHI Malika ,Maitre de Conférences, Université de Constantine

Examinateur :MESSACI Fatiha Professeur, Université de Constantine

Examinateur : NEMOUCHI Naima , Maitre de Conférences, Université de Constantine

.

Soutenu le :20/06/2012.

DEDICACES.

A mes parents

A mes chers frères

A ma chere soeur Fatima.Z et ses enfants

A Halima Mahideb et toute ma famille Mahideb

A tous mes collègues

Et à tous mes enseignants.

Remerciements.

Ce travail intitulé « **Comparaison des Modèles Probit et Logit et Application Epidémiologique** » a été réalisé à l'Université Mentouri Constantine sous la direction de Madame **Malika CHIKHI**, Maître de Conférences à Université Mentouri Constantine.

Ma profonde gratitude et ma reconnaissance la plus distinguée vont particulièrement à Monsieur **F. Lazhar. RAHMANI**, Maître de Conférences à l'Université Mentouri Constantine, me fait un grand honneur de présider le Jury de ce Mémoire. Qu'il veuille bien trouver ici l'expression distinguée de ma haute considération, de mes profonds respects et de mes sincères remerciements.

Le suivi permanent, sous la direction éclairée et fructueuse, de Madame **M. CHIKHI**, Maître de Conférences à Université Mentouri Constantine a permis à ce travail de se concrétiser grâce aux précieux conseils toujours encourageants, à l'entière disponibilité et à la facilité de consultation qu'elle as toujours prodiguée, qu'elle veuille bien trouver ici l'expression profonde de mon estime considérable et de mes sincères remerciements pour avoir accepter de diriger ce travail et d'en être le Rapporteur.

Ma profonde gratitude et ma reconnaissance la plus distinguée vont particulièrement à Madame le Professeur **F. MESSACI**, de l'Université Mentouri Constantine, qui me fait un grand honneur d'avoir bien voulu accepter d'examiner ce travail. Qu'elle veuille bien trouver ici l'expression sincère de ma haute et respectueuse considération et de mes vifs remerciements.

Ma considération, ma profonde gratitude et mes remerciements les plus vifs et les plus distingués vont également à Madame **N. NEMOUCHI**, Maître de Conférences à Université Mentouri Constantine, qui m'honore considérablement de sa présence au Jury en qualité d'Examinateur.

Je ne saurai oublier d'exprimer cordialement à l'ensemble des collègues, des chercheurs et du Personnel du Département de Mathématiques de **l'Université Mentouri Constantine** ma haute et sincère gratitude pour leur soutien moral et leur constante sympathie.

Table des matières

INTRODUCTION

Ce mémoire de magister intitulé : « **Comparaison des Modèles Probit et Logit et Application Epidémiologique** » est structuré en quatre chapitres essentiels agencés comme suit :

Le premier chapitre dénommé " Le modèle linéaire généralisé " consiste en une synthèse bibliographique et une présentation des définitions générales des différents modèles exponentielles avec leurs propriétés et leur utilisation

Le deuxième chapitre présente les modèles logistiques et le modèle probit pour une variable expliquée binaires en fonction de plusieurs variables explicatives (qualitatives ou quantitatives), les fonctions qui les définissent sont d'abord rappelés puis les modèles sont interprétés ainsi que l'estimation des paramètres par la méthode du Maximum de Vraisemblance.

Le troisième chapitre présente une « Comparaison des modèles probit et logit ».

Dans le quatrième chapitre les modèles vus ci-dessus sont appliqués aux données d'une étude épidémiologique visant à mettre en évidence, dans un échantillon 200 patients, les facteurs de risques de la maladie coronarienne et les résultats sont comparés et interprétés. L'intérêt de ce travail provient du fait que ces méthodes, ont été introduites dans certains logiciels récents spécialisés, utilisées, notamment dans le domaine épidémiologique.

Mots clés : modèle linéaire généralisé, modèle Probit et Logit, Maximum de Vraisemblance, épidémiologie, maladie coronarienne.

Chapitre 1

Le modèle linéaire généralise

1.1 qu'est ce qu'un modèle statistique ?

un modèle statistique ([1]) est une représentation simplifiée, donc fausse, de la réalité. Il représente le comportement d'une variable d'intérêt, la variable réponse, en fonction des valeurs des variables explicatives. Dans les études expérimentales celles-ci sont contrôlées alors que dans les études d'observations elles sont simplement observées. statistique (on dit aussi stochastique), il ne cherche pas à prédire exactement la valeur de la réponse, mais seulement à décrire la distribution de toutes les réponses possibles pour une combinaison donnée des variables explicatives. des modèles relativement frustes peuvent se contenter de résumer ces distributions par leur moyenne et leur variance. Les modèles les plus complets les définissent entièrement en spécifiant le type de lois à laquelle appartiennent les distributions (lois normales, lois binomiales,- ...) et les paramètres qui permettent d'identifier une loi unique parmi toutes celles possibles.

1.2 Les composantes du modèle linéaire généralisé

Le modèle logistique ([2]) étend le modèle linéaire dans deux directions :
- la modélisation porte sur des réponses binomiales et non normales ;
- le prédicteur linéaire n'est pas supposé égal aux fréquences attendues mais à une fonction de ces valeurs attendues (leur logit).

Le modèle linéaire généralisé([3]), dont le modèle linéaire([4]) et le modèle logistique([3]) sont des cas particuliers explore systématiquement ces deux directions. Il comprend trois composantes

– le prédicteur linéaire

$$\eta = \beta_0 + \beta_1 \mathrm{X}_1 + ... + \beta_p \mathrm{X}_p \tag{1.1}$$

– la fonction de lien qui relie l'espérance de la réponse, μ ,au predicteur linéaire.

$$g(\mu) = \eta \tag{1.2}$$

– la composante aléatoire qui spécifie comment les réponses fluctuant autour de l'espérance.

La composante aléatoire :

peut être choisie dans une famille relativement large , la famille exponentielle, à laquelle appartiennent les lois normales , binomiales , de poisson , gamma , etc. ... Une propriété de ces lois est que pour chacune , il existe une relation spécifique entre l'espérance

$$\mathrm{E}(\mathrm{Y}) = \mu$$

et la variance

$$\mathrm{V}ar(\mathrm{Y}) = a(\Phi)\mathrm{V}(\mu)$$

Souvent

$$a(\Phi) = \Phi$$

. Si la ième réponse est la moyenne de n_i observation, comme dans le cas d'une fréquence binomiale ,

$$a(\Phi) = \frac{\Phi}{n_i} \tag{1.3}$$

où les poids n_i sont des effectifs connus . La fonction V ,appelée fonction de variance est une caractéristique de chacune des lois de la famille exponentielle : variance constante pour la loi normale .

$$\mathrm{V}(\mu) = 1$$

égale à l'espérance pour la loi de poisson

$$\mathrm{V}(\mu) = \mu$$

au carrée de l'espérance pour la loi gamma

$$\mathrm{V}(\mu) = \mu^2$$

et

$$\mathrm{V}(\mu) = \mu(1-\mu)$$

pour la loi binomiale.

Le choix de la composante aléatoire dépend donc de la relation observée entre moyenne et variance des observations pour différentes combinaisons de variables explicatives . Le coefficient Φ , appelé paramètre de dispersion ,est parfois une caractéristique fixe de la loi (égal à 1 pour Poisson et binomiale), mais c'est souvent un paramètre qui doit être estimé ou spécifié (σ^2pour la loi normale).

Le prédicteur :

comme dans la régression , est linéaire par rapport aux coefficients β .Les variables explicatives peuvent comprendre des fonctions non linéaires de données observées ou contrôlées, comme le carré du temps , le log d'une dose , le rapport du poids sur le carré de la taille.

La fonction de lien :

doit être monotone et différentiable , conditions qui autorisent des formes très diverses . On la choisit généralement croissante pour des raisons d'interprétation.

Dans le modèle linéaire le lien est la fonction identité ([2]).

$$\mu = \eta$$

dans le modéle logistique ([5]).

$$logit(\pi) = log\frac{\pi}{1-\pi} = \eta$$

D'autres liens couramment employés sont les fonctions

$$log(\mu)$$

$$log[-log(1-\pi)]$$

ou

$$probit(\pi)$$

Pour chaque loi ,il existe une fonction de lien , dite canonique ,qui présente certains avantages théoriques et que les logiciels choisissent lorsqu'ils ne reçoivent pas d'instruction contraire .

C'est le lien identité pour la loi normale , le logit pour la binomiale ,le log pour poisson .Cela ne signifie pas que ce soit toujours le meilleur choix . la meilleure fonction de lien à utiliser est celle qui conduit à la meilleure adéquation entre le modèle et les observations mais souvent plusieurs fonctions conduisent à des résultats voisins et le choix peut être effectué sur d'autres critères (facilité de communication , usage, ...).

1.2.1 Variance constante : modélisation d'observation normales

La loi normale ,de variance σ^2 fixée , correspond à la situation où la variance des réponses ne dépend pas de leur moyenne . C'est l'hypothèse sur laquelle reposent la régression linéaire , l'analyse de variance et l'analyse de covariance qui sont trois versions du modèle linéaire ne différant que par la nature des variables explicatives (respectivement quantitatives , qualitatives, ou mixtes) .

L'analyse de variance à un facteur suppose que la variable réponse suit une distribution normale dont la moyenne dépend du niveau d'un facteur explicatif et de variance σ^2 , indépendante du niveau du facteur . Ce facteur est une variable qualitative , ici à trois modalités . Le modèle peut s'écrire

$$\begin{aligned} \mathrm{Y}_{ij} &= \mu_j + \varepsilon_{ij} \\ \varepsilon_{ij} &\sim \mathcal{N}(0, \sigma^2) \end{aligned} \tag{1.4}$$

où Y_{ij} est la ième réponse observée au niveau j du facteur . Dans les notations du modèle linéaire généralisé ([1]), il peut s'écrire

$$\begin{aligned} \eta_{ij} &= \beta_0 + \sum_{j=1}^{p-1} \beta_j \mathrm{X}_{ij} \\ &= \sum_{j=0}^{p-1} \beta_j \mathrm{X}_{ij} \end{aligned} \tag{1.5}$$

$$\mu_{ij} = \eta_{ij} \tag{1.6}$$

$$\mathrm{Y}_{ij} \sim \mathcal{N}(\mu_{ij}, \sigma^2)$$

où les $\mathrm{X}_{ij} (j \geqslant 1)$ sont des variables qui permettent d'identifier les niveaux du facteur et où l'on a introduit $\mathrm{X}_{i0} = 1$,associé à l'ordonnée à l'origine β_0

1.2.2 Réponse binomiales :

pour supprimer le risque d'obtenir un modèle qui prédise des probabilités négatives ou plus grandes que 1, il suffit d'utiliser comme fonction de lien l'inverse de la fonction de répartition d'une variable aléatoire définie sur $\mathbb{R}$. C'est le cas du **logit** qui doit son nom à la distribution logistique sa densité ressemble à celle de la distribution (une courbe en cloche symétrique) mais un peu plus aplatie .

Le lien **probit**([6]) correspond à l'inverse de la fonction de répartition d'une loi normale centrée réduite . Il est particulièrement utilisé en toxicologie , sans que cela signifie qu'il conduise à de "meilleurs" résultats dans ce domaine que le lien logit. De fait il est généralement difficile de choisir entre ces deux liens sur d'autres critères que l'interprétation des paramètres . Les liens logit et probit sont des fonctions symétriques par rapport au point $\frac{1}{2}$ au sens où elles vérifient

$$g(\pi) = -g(1-\pi)$$

ce qui implique

$$logit(\frac{1}{2}) = probit(\frac{1}{2}) = 0$$

En conséquence il est indifférent de modéliser la survenue d'un événement ou celle de son complémentaire : seuls changent le signes des paramètres

Une autre fonction de lien est utilisée dans le cadre d'une méthode de titrage biologique applicable aux suspensions où l'on ne sait pas dénombrer directement les micro-organismes ou cellules d'intérêt tout en étant capable de reconnaître leur présence ou leur absence .
Elle consiste à diluer successivement le prélèvement initial où les cibles sont en concentration θ . A chaque niveau i de dilution , on constitue plusieurs échantillons de volume identique où l'on recherche la présence de la cible . Sous l'hypothèse d'une répartition uniforme des cibles à chaque niveau de dilution , leur nombre Z dans un échantillon donné suit une distribution de poisson de paramètre $\frac{\theta}{x^i}$ où x est la facteur de dilution . La probabilité qu'un échantillon ne contienne aucune cible et donne une réponse négative est donc

$$\mathrm{P}[\mathrm{Y}=0] = \exp(\frac{-\theta}{x^i})$$

et le nombre de réponses positives au niveau i suit une distribution binomiale de probabilité

$$\pi_i = 1 - \exp(\frac{-\theta}{x^i})$$

La fréquence des réponses positives a donc une espérance π_i qui vérifie

$$\log[-\log(1-\pi i)] = \log\theta - i\log x \tag{1.7}$$

le log est donc l'ordonnée à l'origine d'un modèle linéaire généralisé de lien appelé log **-** log **du complément**([3]) . Les termes $-i\log(x)$ ne sont pas estimés , ce sont des termes de compensation , connus , qui ne dépendent que du niveau et du facteur de dilution . Contrairement au logit et au probit , le lien $\log(-\log)$ n'est pas symétrique par rapport à $\frac{1}{2}$.

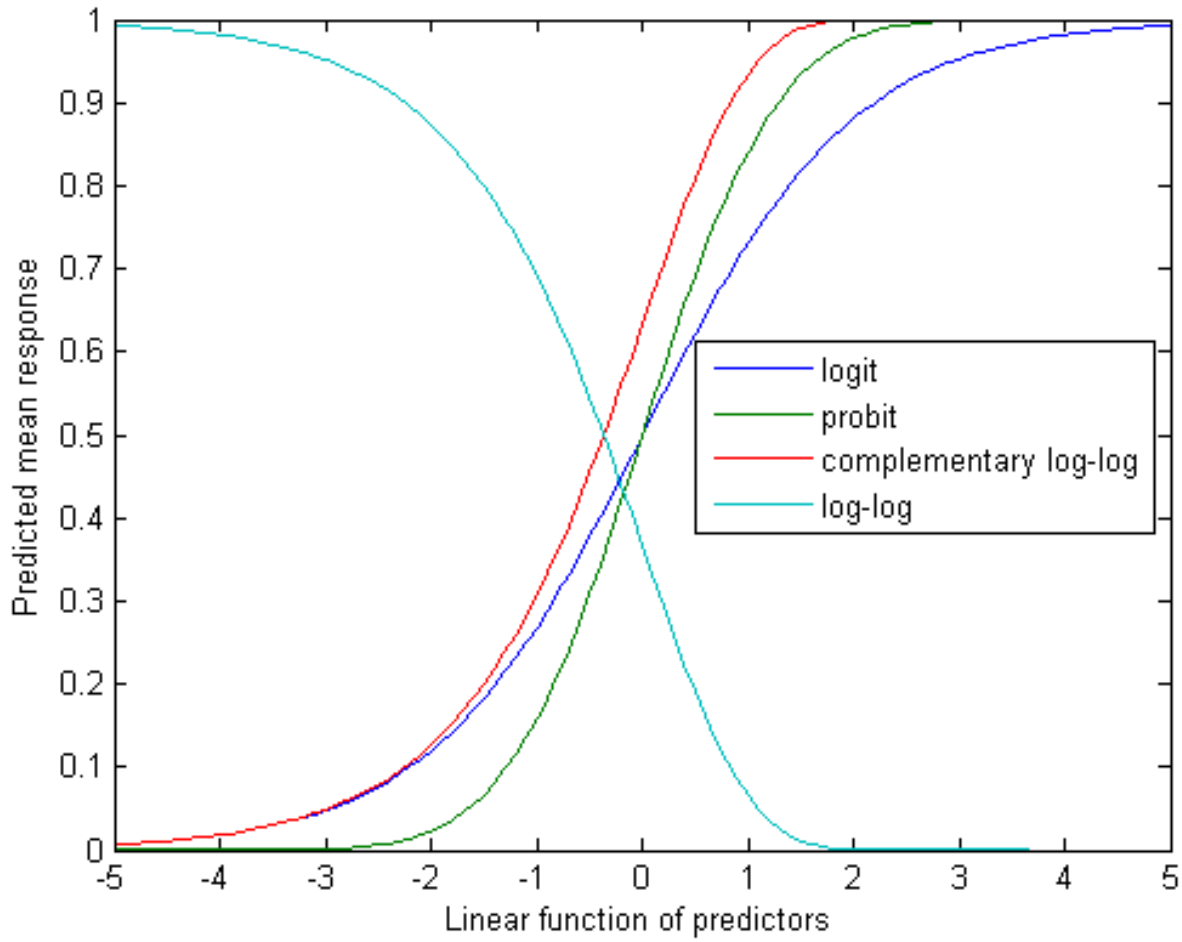


FIGURE 1.1 – FONCTION DE LIENS

.

Chapitre 2

Modèle Logit Et Modèle Probit

2.1 Modèle Dichotomiques Univariés

2.1.1 Introduction

Par modèle dichotomique([7]) , on entend un modèle statistique dans lequel la variable expliquée ne peut prendre que deux modalités (variable dichotomique) . Il s'agit alors généralement d'expliquer la survenue ou la non survenue d'un événement.

Hypotèse On considère un échantillon de N individus indicés $i = 1, .., \mathrm{N}$. pour chaque individu . on un certain évenement s'est réalisé et on not y_i la variable codée associée à l'évenement . on pos $\forall i \in [1, \mathrm{N}]$:

$$y_i = \begin{cases} 1 & \text{si l'événement s'est réalisé pour l'individu } i \\ 0 & \text{si l'événement ne s'est pas réalisé pour l'individu } i \end{cases} \tag{2.1}$$

On remarque ici le choix du codage (0,1) qui est traditionnellement retenu pour les modèles dichotomique . En effet , celui-ci permet définir la probabilité de survenue de l'événement comme l'espérance de la variable codée y_i , puisque :

$$\begin{aligned} \mathrm{E}(y_i) &= \mathrm{P}rob(y_i = 1) \times 1 + \mathrm{P}rob(y_i = 0) \times 0 \\ &= \mathrm{P}rob(y_i = 1) \\ &= pi \end{aligned}$$

L'objectif des modèle dichotomiques([8]) consiste alors à expliquer la survenue de l'événement considéré en fonction d'un certain nombre de caractéristiques observées pour les individus de l'échantillon . Comme nous le verrons par la suite , on cherche dans ces modèles , à spécifier la probabilité d'appartition de cet événement .

2.1.2 Spécification linéaire des variables à expliquer dichotomiques

Supposons que l'on dispose de N observations y_i , $\forall i = 1, .., \mathrm{N}$ d'une variable à expliquer dichotomique codée $y_i = 1$ ou $y_i = 0$ par convention , lorsque parallèlement les observations de K variables explicatives sont $x_i = (x_i^1 ... x_i^{\mathrm{K}})$, $\forall i = 1, ..., \mathrm{N}$.Dans ce cas , le modèle linéaire simple s'écrit :

$$y_i = x_i\beta + \varepsilon_i \qquad \forall i = 1, ..., \mathrm{N} \tag{2.2}$$

où $\beta = (\beta_1 ... \beta_{\mathrm{K}})' \in \mathbb{R}^{\mathrm{K}}$ désigne un vecteur de K paramètres inconnus et où les εi sont supposées être indépendamment distribuées. On peut alors mettre en évidence plusieurs problèmes liés à l'utilisation de cette spécification linéaire simple pour modéliser notre variable dichotomique.

Premièrement , les termes de gauche et de droite de l'équation (2.2) sont de nature différentes. La variable y_i est de type qualitative tandis que la somme $x_i\beta + \varepsilon_i$ est une variable quantitative. On peut répondre à ceci que le membre de gauche correspond en fait au codage (ici 0 ou 1) associé à la variable qualitative ; dès lors , il n'y aurait plus de problème . Mais il est évident que ce codage est lui même par nature arbitraire , et que les valeurs de β obtenues pour ce codage sont nécessairement différentes de celles obtenues pour tout autre codage. Elles seraient par exemple de $\alpha\beta$ si le codage était de type (0,α). Ainsi, le premier problème de l'application du modèle linéaire simple à une variable dichotomique , est que le paramètre β du modèle (2.2) n'est pas interprétable.

Deuxièmement, une étude graphique montre l'approximation linéaire est peu adaptée au problème posé. considérons pour cela le modèle linéaire avec une seule

variable explicative (K = 1), notée x_i^1, et une constante. On pose $\beta = (\beta_0 \beta_1)'$ et l'on considère le modèle linéaire suivant :

$$y_i = \beta_0 + x_i \beta_1 + \varepsilon_i \qquad \forall i = 1, .., \mathrm{N} \tag{2.3}$$

voir la figure (2.1) .

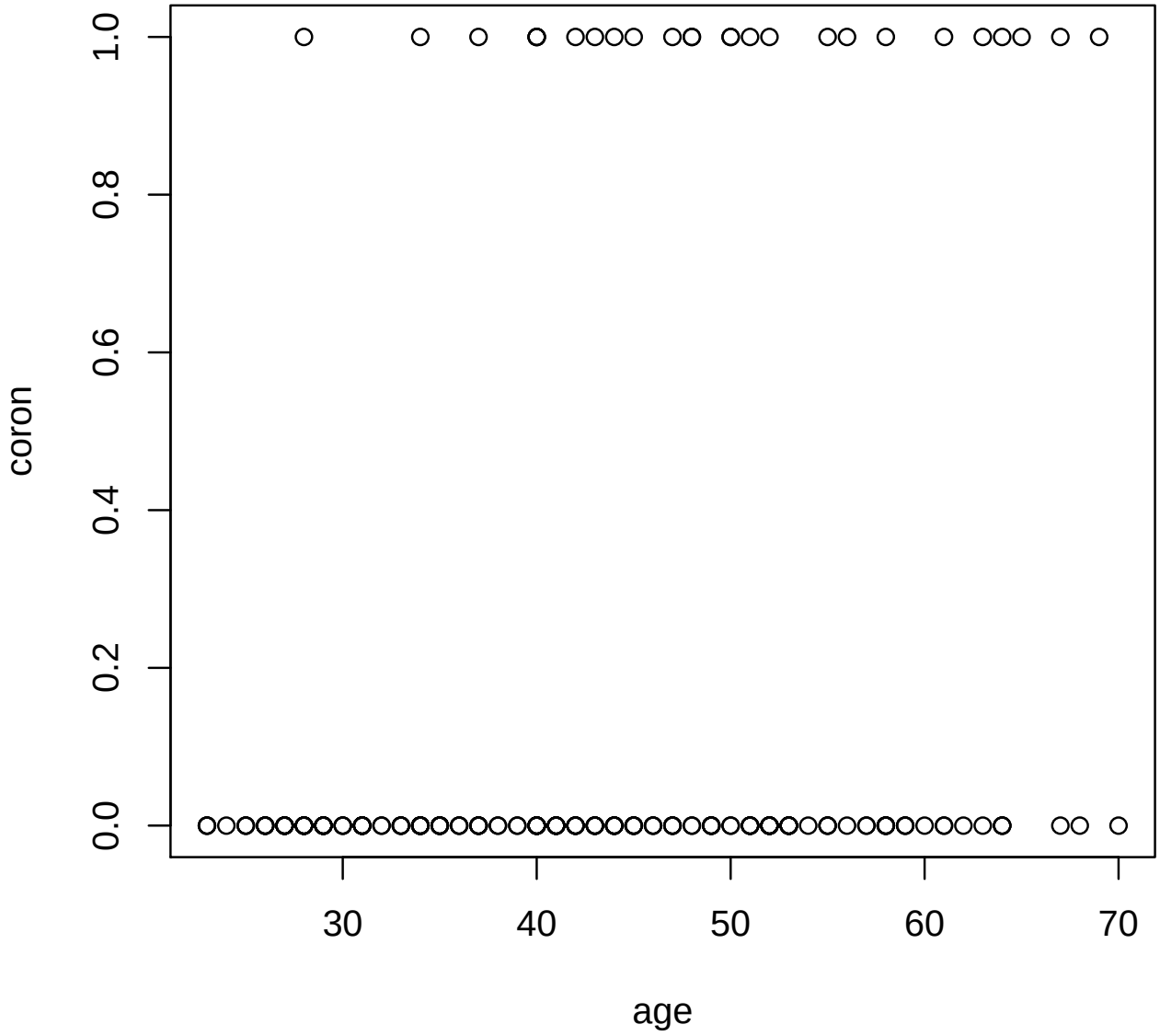


FIGURE 2.1 – Représentation directe de "coron" en fonction de l'âge

Pour constater l'inadéquation de ce modèle à reproduire correctement la variable à expliquer dichotomique y_i , il suffit de se placer dans un repère (x^1, y) et de reproduire les N différents couples (x_i^1, y_i), $\forall i = 1, .., \mathrm{N}$. Naturellement , du fait du statut dichotomique de la variable à expliquer, le nuage de points ainsi obtenu soit sur la droite $y = 0$, soit sur la parallèle $y = 1$.Ainsi , il est impossible d'ajuster de façon satisfaisante, par une seule droite, le nuage de points, associé à une variable dichotomique qui , par nature, est réparti sur deux droites parallèles.(2.1)

Troisièmement, la spécification linéaire standard ne convient pas aux variables dichotomiques, et plus généralement aux variables qualitatives, car elle pose un certain nombre de problèmes mathématiques.

1. Sachant que dans le cas d'une variable à expliquer y_i dichotomique , celle-ci ne peut prendre que les valeurs 0 ou 1 , la spécification linéaire (2.2) implique que ε_i ne peut prendre, elle aussi, que 2 valeurs, conditionnellement au vecteur x_i :

$$\varepsilon_i = 1 - x_i\beta$$

avec une probabilité de

$$p_i = \mathrm{P}rob(y_i = 1)$$

$$\varepsilon_i = -x_i\beta$$

avec une probabilité de

$$1 - p_i$$

Ainsi , la perturbation ε_i du modèle (2.2) admet nécessairement une loi discrète, ce qui exclut en particulier l'hypothèse de normalité des résidus.

2. Lorsque l'on suppose que les résidus ε_i sont de moyenne nulle, la probabilité p_i associée à l'événement $y_i = 1$ est alors déterminée de façon unique. En effet , écrivons l'espérance des résidus :

$$\mathrm{E}(\varepsilon_i) = p_i(1 - x_i\beta) - (1 - p_i)x_i\beta$$

$$= p_i - x_i\beta = 0$$

On en déduit immédiatement que :

$$p_i = x_i\beta = \mathrm{P}rob(y_i = 1) \tag{2.4}$$

Ainsi la quantité $x_i\beta$ correspond à une probabilité et doit par conséquent satisfaire un certain nombre de propriétés et en particulier appartenir à l'intervalle fermé [0,1].

$$0 \leqslant x_i\beta \leqslant 1 \qquad \forall i = 1, .., \mathrm{N} \tag{2.5}$$

Or rien n'assure que de telles conditions soient staisfaites par l'estimateur des Moindres Carrés utilisé dans le modèle linéaire (2.2) . Si de tels cantraintes ne sont pas assurées, le modèle

$$y_i = \beta_0 + x_i\beta_1 + \varepsilon_i \qquad \mathrm{E}(\varepsilon_i) = 0 \qquad \forall i = 1, .., \mathrm{N}$$

n'a pas de sens.

3. Enfin, même si l'on parvenait à assurer le fait que les contraintes (2.5) soient satisfaites par l'estimateur des Moindres Carrés des paramètres du modèle linéaire, il n'en demeurerait pas moins une difficulté liée à la présence d'hétéroscedasticité. En effet, on constate immédiatement que, dans le modèle (2.2), la matrice de variance covariance des résidus varie entre les individus en fonction de leur caractéristiques associées aux expliqatives x_i puisque :

$$\mathrm{V}(\varepsilon_i) = x_i\beta(1 - x_i\beta) \qquad \forall i = 1, .., \mathrm{N} \tag{2.6}$$

Pour démontrer ce résultat il suffit de considérer la loi discrète des résidus et de calculer la variance :

$$\mathrm{V}(\varepsilon_i) = \mathrm{E}(\varepsilon_i^2) = (1 - x_i\beta)^2\mathrm{P}rob(y_i = 1) + (-x_i\beta)^2\mathrm{P}rob(y_i = 0)$$

$$= (1 - x_i\beta)^2 p_i + (-x_i\beta)^2(1 - p_i)$$

Sachant que d'après la relation (2.4) on a $p_i = x_i\beta$, on en déduit que :

$$\begin{aligned} \mathrm{V}(\varepsilon_i) &= (1 - x_i\beta)^2 x_i\beta + (-x_i\beta)^2(1 - x_i\beta) \\ &= (1 - x_i\beta)x_i\beta[(1 - x_i\beta) + x_i\beta \\ &= (1 - x_i\beta)x_i \end{aligned}$$

Or, de plus ce problème d'hétéroscédascticité ne peut pas être résolu par une méthode d'estimation des Moindres Carrés Généralisés tenant compte de la contrainte d'inégalité (2.4), puisque la matrice de variance covariance des perturbations (2.5) dépend du vecteur β des paramètres à estimer dans la spécification linéaire, qui est par nature supposé inconnu.

Pour toutes ces différentes raison, la spécification linéaire des variables explicatives qualitatives, et plus spécialement dichotomique, n'est jamais utilisée et l'on recourt à des modèles logit ou probit, que nous allons à présent étudier, pour représenter ces variables.

2.2 La régression logistique

Nous nous plaçons tout d'abord dans un contexte de classification binaire, c'est-à-dire que nous supposons qu'il existe seulement deux groupes à discriminer.

2.2.1 Régression logistique binaire :

L'éxpression mathématique du modèle logistique dans le cas d'une seule variable X est la suivante :

$$\pi(\mathrm{X}) = p(\mathrm{Y} = 1/\mathrm{X}) = \frac{exp(\alpha + \beta \mathrm{X})}{1 + exp(\alpha + \beta \mathrm{X})}$$

Description graphique de la fonction logistique :

En dehors de raisons mathématiques liées à la théorie du modèle linéaire deux raisons principales conduisent au choix de la fonction logistique.
Cette fonction a une forme sigmoide qui correspond à une forme de relation souvent observée entre "une dose d'exposition" $g(\mathrm{X})$ et la fréquence $\mathrm{Y} = \pi(\mathrm{X})$ d'une maladie à cette dose, comme le montre la figure suivante .

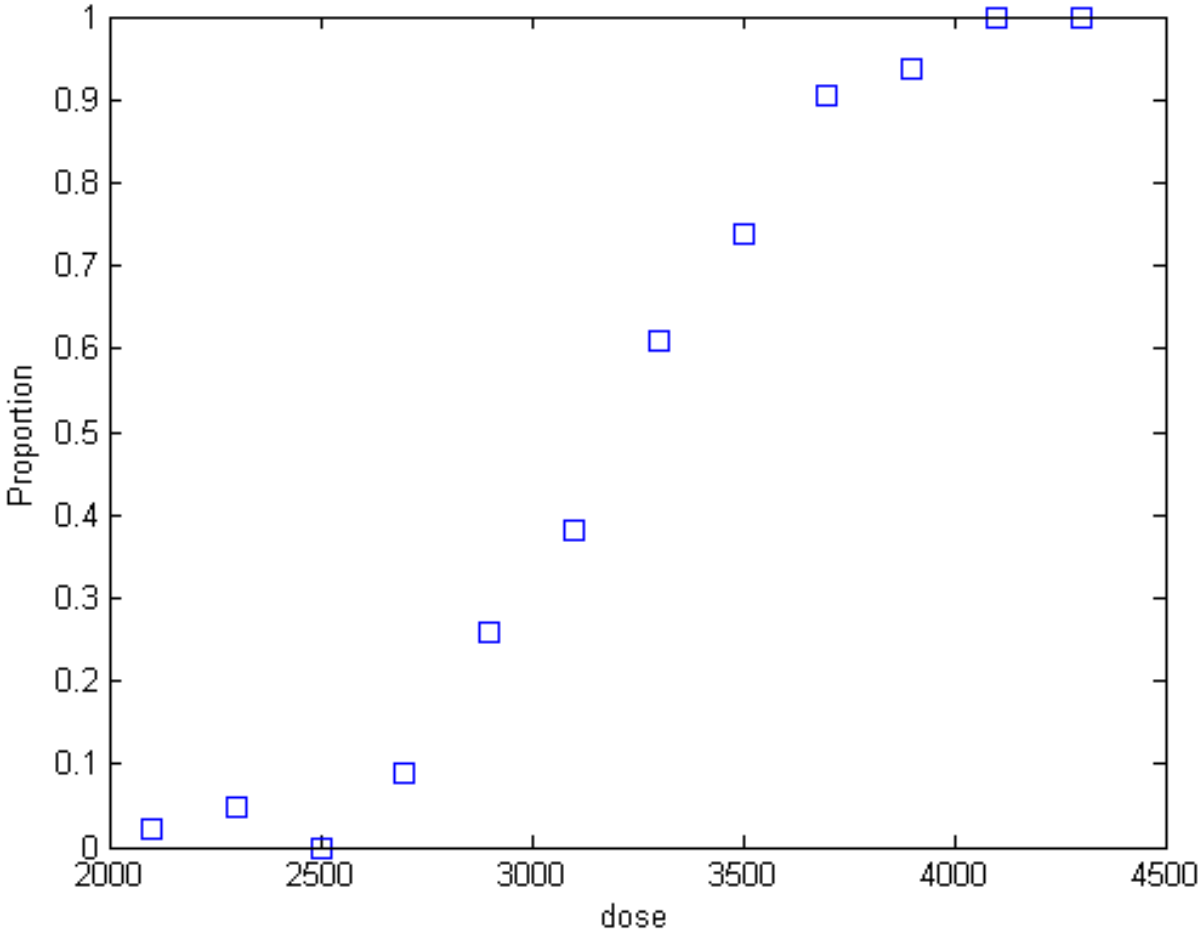

Proportion
dose
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
2000
2500
3000
3500
4000
4500

FIGURE 2.2 –

Elle exprime la relation entre la maladie M et une exposition X à partir de paramètres β directement liés à l'odds ratio (rapport de cote) qui est une mesure d'association trés fréquemment utilisé en épidémiologie.

Si $\pi(\mathrm{X}) = \mathrm{P}(\mathrm{Y} = 1 \setminus \mathrm{X})$, décrit la probabilité de Y=1 pour une valeur X donnée le modèle logistique s'écrit sous une autre forme à partir de logit de π.

$$g(\mathrm{X}) = logit\pi(\mathrm{X})$$

$$= \log[\frac{\pi(\mathrm{X})}{1-\pi(\mathrm{X})}]$$

avec

$$g(\mathrm{X}) = \alpha + \beta\mathrm{X}$$

par définition

Ce modèle est dit **saturé** si les valeurs prédites sont égales aux valeurs observées. Cela suppose que le modèle a autant de paramètres qu'il y a deux modalités dans les données.

Interprétation générale des coefficients :

- α représente l'ordonnée à l'origine et correspond à :

$$log\frac{\pi(0)}{1-\pi(0)}$$

$$\alpha = g(0)$$

$*$ par définition. l'odds ratio mesurant l'association entre la maladie et l'exposition , il est noté **OR** ou **RC** et est égale à :

$$\mathrm{OR} = \mathrm{RC} = \frac{\pi(1)/1-\pi(1)}{\pi(0)/1-\pi(0)}$$

• β est le logarithme d'odds ratio associé à une augmentation de X d'une unité.

• **Mesure de l'effet de X**

$$\mathrm{OR}(\mathrm{X}'vc\mathrm{X}") = e^{\beta(\mathrm{X}'-\mathrm{X}'')}$$

En particulier , pour X dichotomique (X=1,0), OR = RC = e^{β} est issue de la comparaison de la catégorie X=1 à la catégorie X=0.
Sous ce modèle les probabilités sont données par :

$$\pi(\mathrm{X}) = \frac{e^{\alpha+\mathrm{X}\beta}}{1+e^{\alpha+\mathrm{X}\beta}} \qquad et \qquad 1-\pi(\mathrm{X}) = \frac{1}{1+e^{\alpha+\mathrm{X}\beta}}$$

Propriétés du modèle logistique : Les bonnes propriétés du modèle logistique peuvent être illustrées par un exemple. Considérons le cas où Y désigne l'appartition d'une maladie M : M désigne la présence de la maladie, soit Y = 1, et M^c son absence, soit Y = 0. La covariable est le facteur d'exposition, traditionnellement appelé E. On a alors le tableau d'effectifs suivant, qui peut avoir été obtenu de diverses manières, soit prospective, soit rétrospective.

	M	M^c	Total
E	a	c	n_1
E^c	b	d	n_0
Total	m_1	m_0	n

Le modèle logistique peut être défini comme suit :

$$logit\pi(x) = log(\frac{\pi(x)}{1-\pi(x)}) = \alpha + \beta x$$

tq $\pi(x) = \mathrm{P}(\mathrm{M}/x)$

on note E s'il y a exposition, soit $x = 1$. E^c s'il n'ya pas exposition, soit $x = 0$.

- Pour les sujets exposés :

$$\pi(1) = \mathrm{P(M/E)} = \frac{e^{\alpha+\beta}}{1+e^{\alpha+\beta}}$$

$$1-\pi(1) = \mathrm{P(M^c/E)} = \frac{1}{1+e^{\alpha+\beta}}$$

- Pour les sujets non exposés :

$$\pi(0) = \mathrm{P(M/E^c)} = \frac{e^{\alpha}}{1+e^{\alpha}}$$

$$1-\pi(0) = \mathrm{P(M^c/E^c)} = \frac{1}{1+e^{\alpha}}$$

2.2.2 Variable explicative qualitative :

Il y a deux façons d'avoir des variables explicatives qualitatives. Tout d'abord, la variable peut être qualitative par nature (sexe). La deuxième manière consiste à regrouper une variable continue en classes. Soit X une variable qualitative admettant m modalités, le modèle logistique permettant d'expliquer une variable dichotomique Y par X s'écrit([9]).

$$logitp(x) = \beta_0 + \beta_1 1_1(x) + ... + \beta_m 1_m(x)$$

où $1_j(x)$ désignent les indicatrices

$$1_j(x) = \begin{cases} 1 & \text{si x correspond à la } j^{me} \text{ modalité de x} \\ 0 & \text{sinon} \end{cases}$$

Avec un léger abus de notation, on écrira

$$logitp(x) = x\prime\beta$$

avec

$$\beta = (\beta_0, \beta_1, ..., \beta_m)$$

et

$$x = (1, 1_1(x), ..., 1_m(x))$$

Toutes les variables qualitatives sont découpées en variables indicatrices dans le modèle, à chaque modalité d'une variable correspond un coefficient. Nous sommes maintenant à même de définir le modèle logistique dans un cadre général.

définition 1 *(Régression logistique)*

Soit Y une variable binaire à expliquer et $X = (X1, ..., Xp) \in \mathbb{R}^p$ p variables explicatives. Le modèle logistique propose une modélisation de la loi de $Y|X = x$ par une loi de Bernoulli de paramètre $\pi(x) = P(Y = 1|X = x)$ telle que :

$$log\frac{\pi(x)}{1-\pi(x)} = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p = x\prime\beta \tag{2.7}$$

ou encore

$$logitp(x) = x\prime\beta$$

logit désignant la fonction bijective et dérivable de $]0, 1[$ dans $\mathbb{R}$:

$$\pi \longmapsto log(\pi/(1-\pi)).$$

On déduit de (1.1)

$$\pi(x) = P(Y = 1|X = x) = \frac{exp(x\prime\beta)}{1 + exp(x\prime\beta)}.$$

La regression logistique est une manière de modéliser la relation entre une variable à expliquer Y qualitative a deux classes et des variables explicatives X_i qui peuvent être quantitatives ou qualitatives. Ce modèle est utilise en biologie humaine et épidemiologie pour étudier les relations entre une maladie Y et des facteurs de risque X_i , Il permet de calculer la probabilite de survenue de la maladie quand la valeur des variables X_i est connue.

$$p(Y = 1/X_1, X_2, ..., X_p) = \frac{exp(\alpha + \sum \beta_i x_i)}{1 + exp(\alpha + \sum \beta_i x_i)}$$

Cette méthode est utilisable chaque fois que le paramètre de santé au quel on s'intéresse peut être représenté par une variable à deux modalités (présence ou absence d'un signe, malade ou non malade)

• on notera les deux modalités de Y par 1 en cas de malade , 0 pour les non malade

2.2.3 Variable explicative continue

Exemple 1 *:*

Nous souhaitons expliquer la variable Y présence (1)/ absence (0) d'une maladie coronariennes (coron) par l'âge des patients. Les données sont représentées sur la figure (2.1).

Cette figure montre qu'il est difficile de modéliser les données brutes, la variabilité de la variable Coron est élevée pour tout âge. Une méthode permettant de réduire cette variabilité consiste à regrouper les patients par classe d'âge. Nous obtenons le tableau suivant :

Age	n	Absent	Present	Moyenne
[20 ;30[	39	38	1	0.0256
[30,40[	53	48	5	0.0943
[40,50[	50	42	8	0.16
[50,60[	42	28	14	0.33333
[60,70[	15	10	5	0.375
[70,80[	1	1	0	0

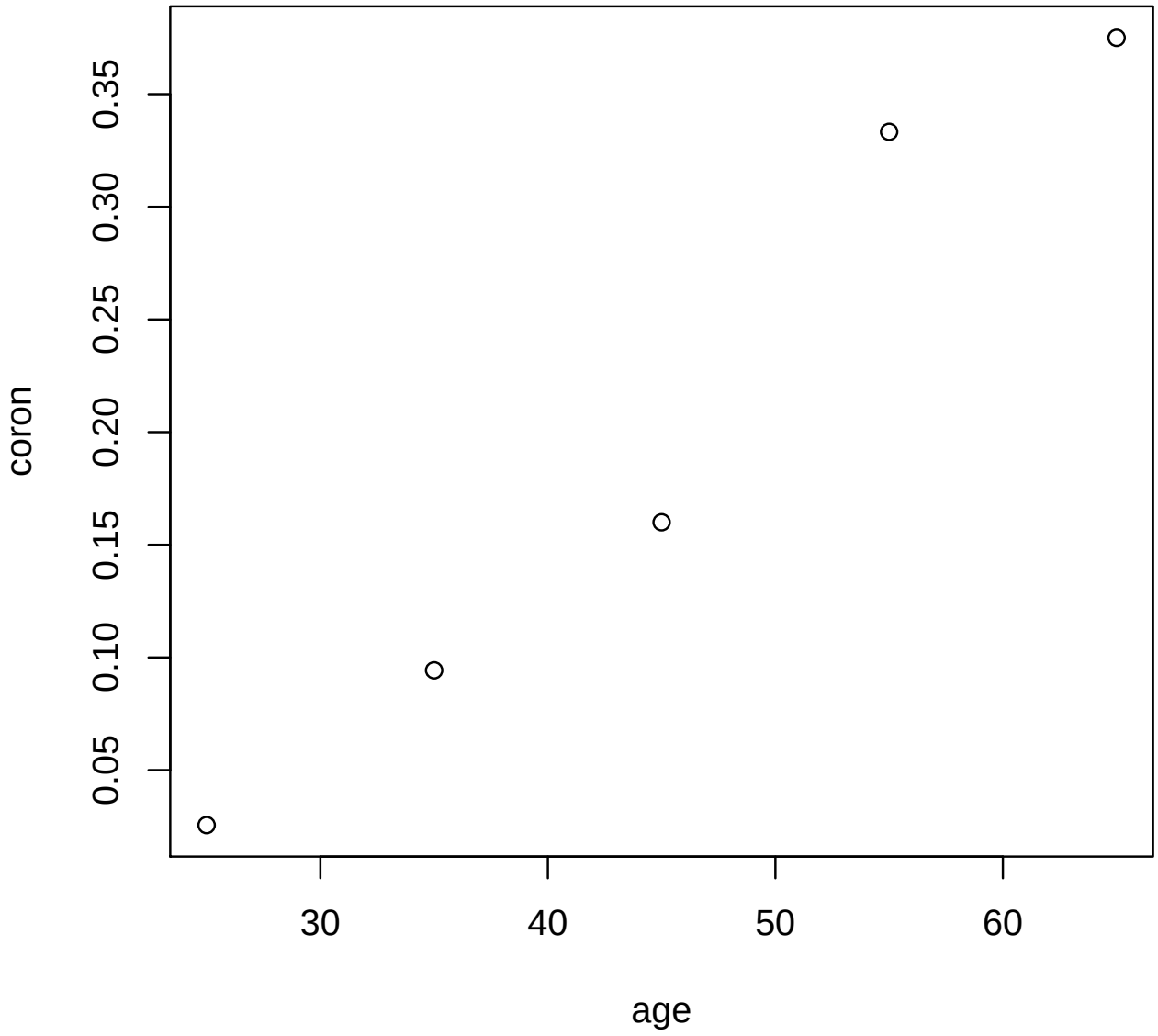


FIGURE 2.3 – Fréquence de "coron" par classe d'âge en fonction de l"âge

2.3 Présentation des modèles dichotomiques en termes de variable latente :

Bien que cela ne soit pas nécessaire on présente les modèles dichotomiques en termes de variables latentes([7]) ou inobservée $y_i{}^*$, la variable observée y_i étant alors un indicateur des valeurs prises par $y_i{}^*$. Cette référence à une variable latente permet de mieux comprendre l'émergence des modèles dichotomiques à partir de certains problèmes .

Exemple :

Tiré d'une étude biologique, concerne la probabilité pour un mineur de contracter une maladie des poumons (événement codé $y_i = 1$) lorsque sa tolérance inobservable, notée $y_i{}^*$, aux conditions de travail et en particulier aux poussières de charbon est inférieure à certain seuil δ inconnue. On suppose que la tolérance est liée à l'âge du mineur noté x_i. ce modèle peut s'écrire sous la forme :

$$y_i = \begin{cases} 1 & \text{si } y_i{}^* = \beta 1 + \beta_2 x_i + \varepsilon_i < \delta \\ 0 & \text{sinon} \end{cases}$$

où ε_i a une distribution logistique. Ici l'événement y_i (maladie) apparait quand la variable latent $y_i{}^*$ est inferieure àun seuil δ. Une autre manière aurait consister à coder l'événement "maladie" en 0. Par suite, nous considérons un modèle où l'on a $y_i = 1$ lorsque $y_i{}^* > \delta$.

Proposition 1

Tout modèle dichotomique univarié peut s'écrire sous la forme d'une équation de mesure de la forme :

$$y_i = \begin{cases} 1 & \text{si } y_i{}^* > \delta \\ 0 & \text{sinon} \end{cases}$$

où $\delta \in \mathbb{R}$ et où la variable latente $y_i{}^*$ inobservable est définie en fonction de caractéristiques observables x_i et d'une pertubation ε_i $iid(0, \sigma_\varepsilon^2)$:

$$y_i{}^* = x_i\beta + \varepsilon_i$$

ce modèle peut également s'exprimer sous la forme :

$$p_i = prob(y_i = 1) = \mathrm{F}(x_i\beta - \delta)$$

F(.)désigne la fonction de répartition associéée à la loi des εi .

Ainsi , si F(.) égale à la fonction de répartition de la loi logistique on retrouve le modèle logistique et si F(.) égale à la fonction de répartition de la loi normale centrée réduite on retrouve le cas du modèle probit.

2.4 Modèle Probit

Un modèle probit est un modèle à réponse binaire([10]) qui emploie une fonction de lien probits.

définition 2 *:*

Le modèle dichotomique suivant :

$$p_i = \mathrm{P}rob(y_i = 1 \backslash x_i) = \mathrm{F}(x_i\beta) \qquad \forall i = 1, .., \mathrm{N} \tag{2.8}$$

définit le modèle probit,dont la fonction de répartition F(.) correspond à la fonction de répartition de la loi normale centrée réduite $\forall x \in \mathbb{R}$:

$$\mathrm{F}(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} dz \tag{2.9}$$

Ainsi, pour une valeur donnée du vecteur des variables explicatives et du vecteur des paramètres β , on peut définir le modèle d'une façon équivalente :

définition 3

Dans le modèle probit on définit la probabilité associé à l'évenement y_i, comme la valeur de la fonction de répartition de la loi normale centrée réduite $\mathcal{N}(0,\ 1)$ considèrée au point $x_i\beta$:

$$\mathrm{F}(x_i\beta) = p_i = \int_{-\infty}^{x_i\beta} \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} dz \qquad \forall i = 1, .., \mathrm{N} \tag{2.10}$$

Remarque1 : les même propriété du modèle logit s'applique pour le modèle probit

Remarque2 : Le modèle dichotomique probit admet pour variable expliquée, non pas un codage quantitatif associé à la réalisation d'un événement , mais la probabilité d'apparition de cet événement, conditionnellement aux covariables .

Remarque3 : Ce modèle est estimée par la méthode du maximum de vraisemblance , une telle estimation étant appelé une régression des probits.

2.5 Estimation des Paramètres par la Méthode du Maximum de Vraisemblance

Nous allons utiliser l'échantillon $(x_1, y_1), ..., (x_n, y_n)$ pour estimer les paramètres β par la méthode du maximum de vraisemblance([5]). Cette méthode consiste à chercher

$$\beta = (\beta_0, \beta_1, ..., \beta_p)$$

qui maximise la vraisemblance

$$\prod_{i=1}^{n} \mathrm{P}(\mathrm{Y} = yi|\mathrm{X} = xi).$$

Rappelons que par définition du modèle logistique $\mathrm{Y}|\mathrm{X} = x \sim \mathrm{B}er(p(x))$. Par conséquent :

$$\prod_{i=1}^{n} \mathrm{P}(\mathrm{Y} = yi|\mathrm{X} = xi) = \prod_{i=1}^{n} \pi_i^{y_i}(1 - \pi_i)^{1-y_i}$$

avec $\pi_i = \mathrm{P}(\mathrm{Y} = 1|\mathrm{X} = x_i)$. En passant au log nous avons alors

$$\mathcal{L}(\beta) = \sum_{i=1}^{n} [y_i log(\pi_i) + (1 - y_i) log(1 - \pi_i)]$$

$$= \sum_{i=1}^{n} [y_i log(\frac{\pi_i}{1-\pi_i}) + log(1-\pi_i)].$$

D'après(2.7) nous obtenons

$$\mathcal{L}(\beta) = \sum_{i=1}^{n} [y_i x_i{}'\beta - log(1 + exp(x_i{}'\beta))]. \tag{2.11}$$

En dérivant par rapport au paramètre β nous avons que

$$\frac{\partial \mathcal{L}}{\partial \beta}(\beta) = [\frac{\partial \mathcal{L}}{\partial \beta_0}(\beta), ..., \frac{\partial \mathcal{L}}{\partial \beta_p}(\beta)]'$$

avec

$$\frac{\partial \mathcal{L}}{\partial \beta_j}(\beta) = \sum_{i=1}^{n} [y_i x_{ij} - \frac{x_{ij} exp(x_i{}'\beta)}{1 + exp(x_i\prime\beta)} = \sum_{i=1}^{n} [x_{ij}(y_i - p_i)].$$

Ce qui donne en écriture matricielle

$$\frac{\partial \mathcal{L}}{\partial \beta}(\beta) = \sum_{i=1}^{n} [x_i(y_i - p_i)].$$

Une condition nécessaire d'optimum (sur $\mathbb{R}$) est l'annulation des dérivées à l'optimum, nous obtenons l'équation suivante (appelée équation du score) :

$$\mathrm{S}(\beta) = \frac{\partial \mathcal{L}}{\partial \beta}(\beta) = \sum_{i=1}^{n} x_i{}'\{y_i - \mathrm{P}(\mathrm{Y} = 1|\mathrm{X} = x_i)\} = \mathrm{X}'(\mathrm{Y} - \mathrm{P}) = 0. \tag{2.12}$$

P est le vecteur de dimension n des $\mathrm{P}(\mathrm{Y} = 1|\mathrm{X} = x_i)$ qui dépend de β. On note $\hat{\beta}$ une solution de $\mathrm{S}(\beta) = 0$. Trouver explicitement $\hat{\beta}$ n'est pas possible. En effet, l'équation (1.8) s'écrit :

$$\begin{cases} x_{11}y_1 + ... + x_{n1}y_n = x_{11}\frac{exp(\beta_1 x_{11}+...+\beta_p x_{1p})}{1+exp(\beta_1 x_{11}+...+\beta_p x_{1p})} + ... \\ \qquad\qquad +x_{n1}\frac{exp(\beta_1 x_{n1}+...+\beta_p x_{np})}{1+exp(\beta_1 x_{n1}+...+\beta_p x_{np})} \\ ..... \\ x_{1p}y_1 + ... + x_{np}y_n = x_{1p}\frac{exp(\beta_1 x_{11}+...+\beta_p x_{1p})}{1+exp(\beta_1 x_{11}+...+\beta_p x_{1p})} + ... \\ \qquad\qquad +x_{np}\frac{exp(\beta_1 x_{n1}+...+\beta_p x_{np})}{1+exp(\beta_1 x_{n1}+...+\beta_p x_{np})} \end{cases}$$

Remarque : pour le modèle Probit : la fonction p_i est la fonction de répartition gaussienne

Dans le cas de la régression logistique (binaire), il existe une procédure spécifique dite IRLS (Iterative Reweighted Least Squares) issue de la procédure de Newton-Raphson.

2.5.1 Méthode de Newton-Raphson

La méthode de Newton-Raphson([11]) permet une résolution numérique des équations du score. On part tout d'abord d'une valeur initiale arbitraire de β, notée β^0 . On note

$$\beta^1 = \beta^0 + h$$

une valeur candidate pour être solution de $\mathrm{S}(\beta) = 0$, c'est-à-dire

$$\mathrm{S}(\beta^0 + h) = 0$$

Par un développement limité à l'ordre un de la fonction S , on obtient l'approximation suivante :

$$\mathrm{S}(\beta^0 + h) \simeq \mathrm{S}(\beta^0) + h\mathrm{S}'(\beta^0). \tag{2.13}$$

Comme

$$\mathrm{S}(\beta^0 + h) = 0$$

on obtient pour h la valeur suivante :

$$h = -[\mathrm{S}'(\beta^0)]^{-1}\mathrm{S}(\beta^0).$$

Il vient

$$\beta^1 = \beta^0 - [\frac{\partial^2 \mathcal{L}}{\partial\beta\partial\beta'}(\beta^0)]^{-1}\frac{\partial \mathcal{L}}{\partial\beta}(\beta^0) \tag{2.14}$$

On itère le processus. La procédure se résume de la manière suivante :
1. choix d'un point de départ β^0 ;
2. On construit β^{k+1} à partir de β^k

$$\beta^{k+1} = \beta^k + \mathrm{A}^k \nabla\mathcal{L}|_{\beta^k}$$

où $\nabla\mathcal{L}|_{\beta^k}$ est le gradient au point β^k et $\mathrm{A}^k = -(\nabla^2\mathcal{L}|_{\beta^k})^{-1}$ est la matrice de "pas" de l'algorithme (l'inverse du hessien de $\mathcal{L}$ au point β^k)
Algorithme maximisation de la vraisemblance
Require : β^0
$k \longleftarrow 1$
repeat
$\beta^{k+1} \longleftarrow \beta^k + \mathrm{A}^k\nabla\mathcal{L}_k$
$k \longleftarrow k + 1$
until
$\beta^{k+1} \approx \beta^k$ et / ou $\mathcal{L}(\beta^{k+1}) \approx \mathcal{L}(\beta^k)$

2.5.2 Calcul des estimateurs

Calculons la matrice hessienne

$$\nabla^2\mathcal{L} = \{\frac{\partial^2 \mathcal{L}}{\partial\beta_r\partial\beta_s}\}_{1\leqslant r,s\leqslant p}$$

tq

$$\frac{\partial^2 \mathcal{L}}{\partial \beta_r \partial \beta_s} = -\sum_{i=1}^{n} x_i^r x_i^s \frac{exp(x_i^{'}\beta)}{(1+exp(x_i^{'}\beta))^2}$$

$$= -\sum_{i=1}^{n} x_i^r x_i^s p_i(1-p_i)$$

par conséquent

$$\nabla^2 \mathcal{L} = \frac{\partial^2 \mathcal{L}}{\partial \beta_2}$$

$$= -\sum_{i=1}^{n} x_i x_i^{'} \mathrm{P}(\mathrm{Y}=1|\mathrm{X}=x_i)(1-\mathrm{P}(\mathrm{Y}=1|\mathrm{X}=x_i))$$

On note

• p_i^k la probabilité $\mathrm{P}(\mathrm{Y}=1|\mathrm{X}=x_i)$ estimée à l'étape k de l'algorithme ;

• P^k le vecteur colonne de dimension n dont le i^{me} élément est p_i^k ;

• W_k la matrice diagonale $\mathrm{diag}(p_i^k(1-p_i^k))$.

Il vient

$$-(\nabla^2 \mathcal{L}|_{\beta^k})^{-1} = (\mathrm{X}^{'}\mathrm{W}^k\mathrm{X})^{-1} \tag{2.15}$$

Nous sommes maintenant à même de calculer β^{k+1}.

$$\beta^{k+1} = \beta^k + (\mathrm{X}^{'}\mathrm{W}^k\mathrm{X})^{-1}\mathrm{X}^{'}(\mathrm{Y}-\mathrm{P}^k)$$

$$= (\mathrm{X}^{'}\mathrm{W}^k\mathrm{X})^{-1}\mathrm{X}^{'}\mathrm{W}^k(\mathrm{X}\beta^k + \mathrm{W}^{k-1}(\mathrm{Y}-\mathrm{P}^k))$$

$$= (\mathrm{X}^{'}\mathrm{W}^k\mathrm{X})^{-1}\mathrm{X}^{'}\mathrm{W}^k\mathrm{Z}^k,$$

où

$$\mathrm{Z}^k = \beta^k + \mathrm{W}^{k-1}(\mathrm{Y}-\mathrm{P}^k).$$

Cette équation est simplement une régression pondérée où les poids W^k dépendent de X et β^k. Les poids sont donc réévalués à chaque étape de l'algorithme, une étape étant une simple régression pondérée. A la dernière étape $k^\star$ de l'algorithme, on note $\mathrm{W}^{k^\star} = \mathrm{W}^\star$. On obtient l'estimateur du maximum de vraisemblance :

$$\hat{\beta} = (\mathrm{X}^{'}\mathrm{W}^{k^\star}\mathrm{X})^{-1}\mathrm{X}^{'}\mathrm{W}^{k^\star}\mathrm{Z}^{k^\star}$$

$$= (\mathrm{X}^{'}\mathrm{W}^\star\mathrm{X})^{-1}\mathrm{X}^{'}\mathrm{W}^\star\mathrm{Z}^{k^\star}$$

2.6 Tests

On peut obtenir des statistiques pivotales, c'est-à-dire des statistiques dont on connaît la loi asymptotique, qui permettent de tester des contraintes sur les coefficients, en particulier leur nullité. On obtient à chaque fois une statistique asymptotiquement χ^2, on compare donc les valeurs obtenues aux quantiles du χ^2.

2.6.1 Rapport de vraisemblance

Dans le cadre de l'estimation par maximum de vraisemblance, le test le plus naturel consiste à construire un rapport de vraisemblance([14]). Pour tester une contrainte de rang p - r sur β de dimension p, on utilise le résultat suivant :

$$\mathrm{LR} = -2(log\mathrm{L}(\hat{\beta}) - log\mathrm{L}(\hat{\beta}^c)) \xrightarrow{\mathcal{L}} \chi^2_r$$

pour N $\longrightarrow \infty$
où $\hat{\beta}^c$ est l'estimateur du maximum de vraisemblance sous la contrainte.

2.6.2 Score

On peut aussi utiliser la nullité du score (aussi appelé test du multiplicateur de Lagrange), en mesurant la norme $||.||_2$ du score évalué en $\hat{\beta}^c$:

$$\frac{\partial log\mathrm{L}(\beta)}{\partial \beta} \mid_{\beta=\hat{\beta}^c} \mathrm{I}(\hat{\beta}^c)^{-1} \frac{\partial log\mathrm{L}(\beta)}{\partial \beta} \mid_{\beta=\hat{\beta}^c} \xrightarrow{\mathcal{L}} \chi^2_r$$

pour N $\longrightarrow \infty$

2.6.3 test de Wald

Le test de Wald, proche du test de score, sert spécifiquement à tester la nullité des coefficients :

$$\sum_{j=1}^{p} \frac{\hat{\beta}_j^2}{\mathrm{I}_j(\hat{\beta})} \xrightarrow{\mathcal{L}} \chi^2_p$$

pour N $\longrightarrow \infty$

Chapitre 3

Comparaison des modèles probit et logit

La question que l'on se pose immédiatement est de savoir quelles sont les différences fondamentales entre les modèles probit et logit ? Quand doit on utiliser l'un plutôt que l'autre ? Quelles sont les propriétés particulières de ces deux modèles ? Bien entendu , ces deux modèles ne différent que par la forme de la fonction de répartition F(.) . Ainsi, il faut donc se rappeler quelles sont les propriétés respectives des lois logistiques et normales, pour comprendre quelles peuvent être les différences et les similitudes entre les modèles logit et probit.

3.1 Similitudes

Les modèles logit ont été introduits comme des approximations de modèles probit permettant des calculs plus simples. Dés lors, il n'existe que peu de différences entre ces deux modèles dichotomiques. Ceci s'explique par la proximité des familles de lois logistiques et normales. Les deux fonctions de répartition sont en effet sensibblement proches, comme on peut le constater à partir du tableau (3.1) où sont reportées les valeurs de ses fonctions pour différentes valeurs de x . Mais cette similitude est encore grande si l'on considère une loi logistique transformée de sorte à ce que la variance soit identique à celle de la loi normale réduite. En effet, nous avons vu que la loi logistique usuelle admet pour fonction de répartition

$$\mathrm{L}(x) = \frac{1}{1+\exp(-x)} = \frac{\exp(x)}{1+\exp(x)}$$

Cette loi a une espérance nulle et une variance égale à $\pi^2/3$. C'est pourquoi, il convient de normaliser la loi logistique de sorte à obtenir une distribution de variance unitaire, comparable à celle de la loi normale réduite. On définit pour cela une loi logistique transformée.

définition 4

La loi logistique transformée de paramètre λ admet pour fonction de répartition, notée $\mathrm{L}_\lambda(x)$ $\forall x \in \mathbb{R}$

$$\mathrm{L}_\lambda(x) = \frac{\exp(\lambda x)}{1+\exp(\lambda x)} = \frac{1}{1+\exp(-\lambda x)} \tag{3.1}$$

A cette fonction de répartition correspond une variance de $\frac{\pi^2}{(3\lambda^2)}$. Ainsi, il convient de comparer la loi normale centrée réduite à la loi logistique transformée, de paramètre $\lambda = \pi/\sqrt{3}$, dont la fonction de répartition est définie comme suit :

$$\widetilde{\mathrm{L}}(x) = \mathrm{L}_{\pi/\sqrt{3}}(x) = \frac{1}{1+e^{-\frac{\pi x}{\sqrt{3}}}} \tag{3.2}$$

Cette loi admet par construction une variance unitaire. On observe ainsi à partir du tableau (3.1), que les réalisations de cette fonction $\mathrm{L}_{\pi/\sqrt{3}}(.)$ sont très proches de celles de la fonction F(.) associée à la loi normale réduite et ce notamment pour des valeurs de x proche de 0, c'est à dire des valeurs dites centrales, car proches de la moyenne de la distribution.

Certains auteurs proposent d'utiliser d'autres paramètres λ afin de mieux reproduire encore la fonction de répartition de la loi normale pour des valeurs centrales. En particulier ([11]) propose d'utiliser un paramètre $\lambda = 1.6$ et donc de retenir la loi logistique transformée $\mathrm{L}_{1.6}(.)$. Comme on peut l'observer sur le tableau (3.1) , la fonction de paramètre 1.6 est encore plus proche de F(.) que la fonction de paramètre $\pi/\sqrt{3}$. pour les valeurs centrales proches de 0.

x	0	0.15	0.3	0.45	0.6	0.75	1	1.5
$\mathrm{F}(x)$	0.5000	0.5596	0.6179	0.6736	0.7257	0.7734	0.8413	0.9332
$\mathrm{L}(x)$	0.5	0.5374	0.5744	0.6106	0.6457	0.6792	0.7311	0.8176
$\mathrm{L}_{\pi/\sqrt{3}}(x)$	0.5	0.5676	0.6328	0.6934	0.7481	0.7958	0.8598	0.9382
$\mathrm{L}_{1.6}(x)$	0.5	0.5597	0.6177	0.6726	0.7231	0.7685	0.8320	0.9168

TABLE 3.1 – Comparaison des Fonctions de Répartition $\mathrm{L}_{\pi/\sqrt{3}}(x)$ et $\mathrm{F}(x)$

Quoiqu'il en soit, il apparaît ainsi que les fonctions de répartition des lois normales centrées réduites et des lois logistiques simples ou transformées sont extêmement proches. Par conséquent, les modèles probit et logit donnent généralement des résultats relativement similaires.

En effet, les valeurs estimées des paramètres dans les modèles probit et logit ne sont pas directement comparables puisque les variances des lois logistiques et normale réduite ne sont pas identiques. Cette différence de variance implique que la normalisation des coefficients β n'est pas identiques et que par conséquent les estimateurs de ces paramètres obtenus dans les deux modèles ne fournissent pas des réalisations identiques.

Proposition 2

Supposons que l'on note respectivement $\hat{\beta}_{\mathrm{P}}$ et $\hat{\beta}_{\mathrm{L}}$ les estimateurs des paramètres β obtenus dans les modèles probit et logit ([11]) propose en première approximation d'utiliser la relation suivante entres les estimations probit et logit :

$$\hat{\beta}_{\mathrm{L}} \simeq 1.6\hat{\beta}_{\mathrm{P}} \tag{3.3}$$

Toutefois, si ces approximations sont relativement sur certains échantillons comportant peu de valeurs **extrêmes** (c'est à dire lorsque la moyenne des valeurs $x_i\beta$ est proche de zéro), elles seront moins précises en présence de nombreuses valeurs $x_i\beta$ éloignées de zéro. Une façon équivalente de vérifier l'adéquation de cette approximation consiste à observer si la valeur moyenne des probabilités p_i est proche de 0.5 (Davidson et Mackinnon 1984). Si tel est le cas, les estimateurs des coefficients du modèle logit seront environ 1.6 fois superieurs à ceux du modèle probit.(voir Annex)

Considérons par exemple le coefficient de la variable age. Le modèle logit nous donne une estimation de 0.06333 pour ce paramètre alors que le modèle probit donne une estimation de 0.03496. On vérifie alors que, pour cet échantillon, les approximations (3.3) sont satisfaisantes puisque selon cette formule, on devrait obtenir une estimation logit de paramètre de l'ordre de $0.03496 \times 1.6 = 0.055936$ ou 0.06337 si l'on considère léapproximation 0.03496 $\times \frac{\pi}{\sqrt{3}}$. Ces approximations sont en effet très proches de la vraie estimation du paramètre dans le modèle logit.

Proposition 3

On note $\hat{\beta}_{\mathrm{P}}$ l'estimateur obtenu dans le modèle probit, $\hat{\beta}_{\mathrm{L}}$ l'estimateur obtenu dans le modèle logit, et $\hat{\beta_{\mathrm{LP}}}$ l'estimateur obtenu dans le modèle linéaire. (Amemiya [11]) propose les approximations suivantes pour les modèles probit et linéaire :

$$\hat{\beta_{\mathrm{LP}}} \simeq 0.4\hat{\beta}_{\mathrm{P}} \tag{3.4}$$

pour tous les paramètres à l'exception de la constante

$$\hat{\beta_{\mathrm{LP}}} \simeq 0.4\hat{\beta}_{\mathrm{P}} + 0.5 \tag{3.5}$$

pour la constante

et les approximations suivantes pour les modèles logit et linéaire :

$$\hat{\beta_{\mathrm{LP}}} \simeq 0.25\hat{\beta}_{\mathrm{L}} \tag{3.6}$$

pour tous les paramètres à l'exception de la constante

$$\hat{\beta_{\mathrm{LP}}} \simeq 0.25\hat{\beta}_{\mathrm{L}} + 0.5 \text{ } \textbf{pour la constante} \tag{3.7}$$

Ainsi si l'on considère l'exemple des données de chapitre 4 (Application) , les estimations de la constante et des paramètres des variables age et wt dans le modèle linéaires sont respectivement égales à -0.458952, 0.006916 et 0.001791. Or, si l'on compare ces résultats à ceux obtenus à partir des modèles logit et probit (voir chap 4), on obtient les résultats relativement proches. Ainsi, dans le cas du modèle logit pour la variable age l'approximation donnerait

$0.25 \times 0.06358 = 0.015895$ et $0.25 \times 0.01593 = 0.003982$ pour la variable wt. Pour la constante l'approximation donne une valeur approchée égale à $0.25 \times (-7.49929) + 0.5 = -1.37$. Cers approximations seront d'autant plus proches des valeurs estimées qu'il y a aura un grand nombre d'observations $x_i\beta$ proches de 0, car en effet les fonctions de répartition des lois logistiques et normales ne se démarquent pas d'une droite dans cette zone

En conclusion, il apparaît que les résultats des modèles probit et logit sont généralement similaires que ce soit en termes de probabilité ou en termes d'estimation des coefficients β si l'on tient compte des problèmes de normalisation. C'est le sens de cette conclusion d'Amemiya ([11]).

Toutefois, comme le note (Amemiya (1981)), il convient d'être prudent dans l'utilisation des approximations pour comparer les modèles probit et logit. Il est toujours préférable de raisonner en termes de probabilités $p_i = \mathrm{F}(x_i\beta)$ et non en termes d'estimation des paramètres β pour comparer ces résultats

3.2 Différences

Si les deux modèles sont sensiblement identiques, il existe cependant certaines différences entre les modèles probit et logit, comme le souligne d'ailleurs (Amemiya). Nous évoquerons ici deux principales différences :
1. La loi logistique tend à attribuer aux événements "extrêmes" une probabilité plus forte que la distribution normale.
2. Le modèle logit facilite l'interprétation des paramètres β associées au variables explicatives x_i

Nous allons à présent étudier successivement ces deux propriétés. Premièrement, la fonction de densité associée à la loi logistique possède en effet des queues de distribution plus épaisses que celles de la fonction de densité de la loi normale (distribution à queues "plates"). La loi logistique présente donc un excès de Kurtosis : il s'agit d'une distribution leptokurtique. En d'autres termes, nous avons vu que les lois logistique et normale appartiennent à la même famille des lois exponentielles et sont par nature très proches, notamment pour les valeurs proches de

la moyenne de la distribution. Toutefois, le profil de ces deux distributions diffère aux extrémités du support : pour la loi normale, les valeurs extrêmes sont moins pondérées, la fonction de répartition tendant plus vite vers 0 à gauche du support et vers 1 à droite.

Economiquement, cela implique que le choix d'une fonction logistique (modèle logit) suppose une plus grande probabilité attribuée aux évenements "extrêmes", comparativement au choix d'une loi normale (modèle probit), que ce soit à droite ou gauche de la moyenne de la distribution, les lois normales et logistiques étant symétriques. Pour visualiser ce phénomène, il convient de comparer la fonction de répartition associée à la loi normale centrée réduite avec la fonction de répartition associée à la loi logistique possédant les deux premiers moments identiques à la loi $\mathcal{N}(0,1)$.

Sur le graphique (3.2) est reportée la différence $\widetilde{\mathrm{L}}(x) - \mathrm{F}(x)$ en fonction de x : On constate qu'à droite du support, pour des valeurs élevées de x (x $>$ 1.5 environ), on a $\widetilde{\mathrm{L}}(x) < \mathrm{F}(x)$. La fonction de répartition de la loi normale est au dessus de celle de la loi logistique. Etant donnée la définition de la fonction de répartition,$\mathrm{F}(x) = \mathrm{P}rob(\mathrm{X} \leqslant x)$, cela signifie que la probabilité que la réalisation de la variable X soit inférieure au seuil x est plus grande dans le cas de la loi normale que dans le cas de la loi logistique. Inversement, pour un seuil w donnée, la probabilité d'obtenir des valeurs supérieures à ce seuil (des valeurs "extrêmes") est plus grande dans le cas de la loi logistique que dans le cas de la loi normale. On vérifie ainsi la propriété de la loi logistique qui sur-pondère les valeurs extrêmes en comparaison de la loi normale. Naturellement, puisque les distributions sont symétriques, on obtient le même résultat à gauche du support pour des valeurs très faibles de x (x $<$ -1.5 environ).

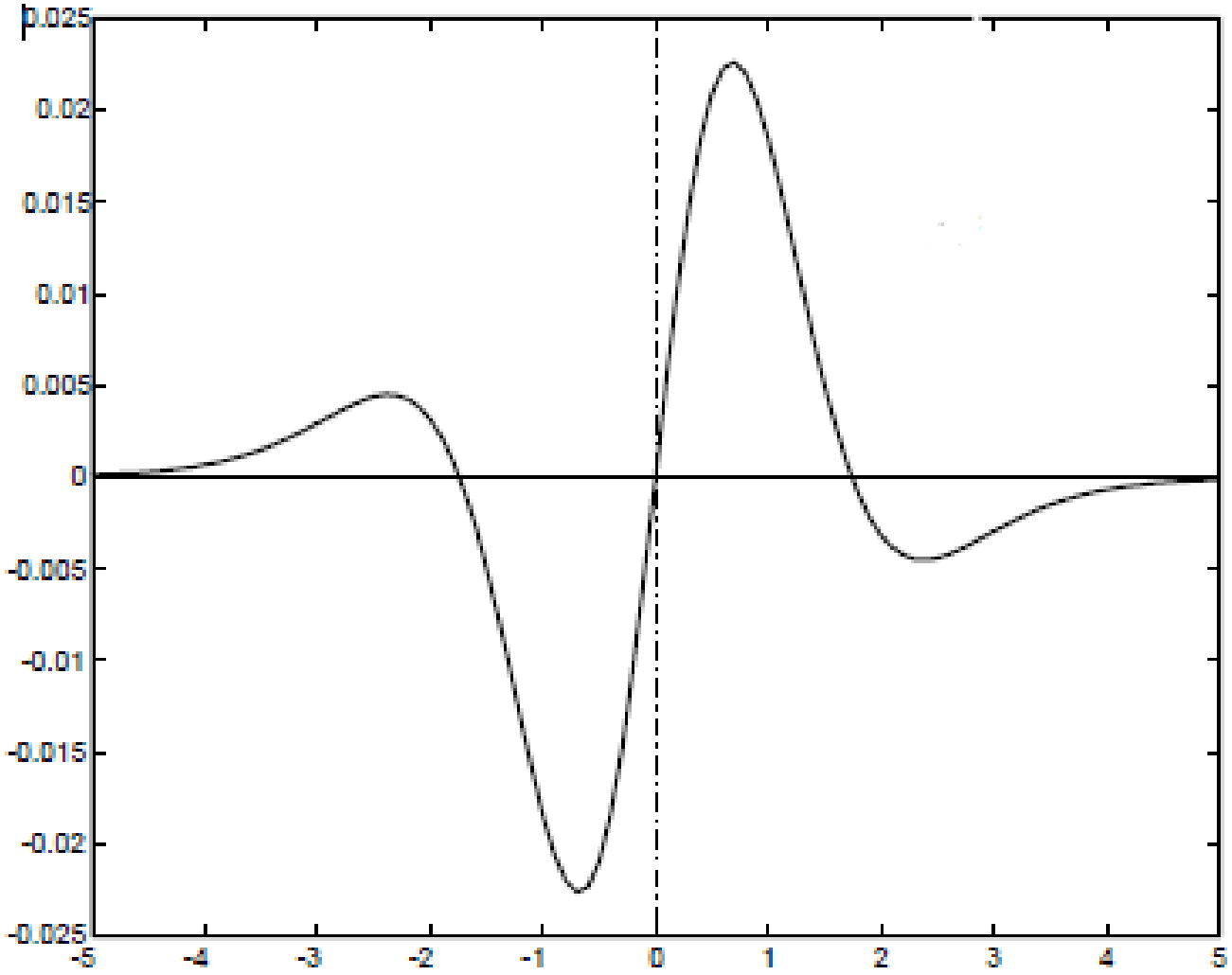


FIGURE 3.1 – Différence des Fonctions de Répartition $\widetilde{\mathrm{L}}(x) - \mathrm{F}(x)$

Proposition 4

De façon générale, la quantité $c_i = \frac{p_i}{1-p_i}$ représente le rapport de la probabilité associée à l'événement $y_i = 1$ à la probabilité de non survenue de cet événement : il s'agit de la cote ("odds"). Dans un modèle logit, cette cote correspond simplement à la quantité $e^{x_i\beta}$:

$$c_i = \frac{p_i}{1-p_i} = e^{x_i\beta} \tag{3.8}$$

Si ce rapport est égal à c_i pour l'individu i, cela signifie qu'il y a c_i fois plus de chance que l'événement associé au code $y_i = 1$ se réalise, qu'il ne se réalise pas ("c_i contre 1" dans le langage usuel).

Exemple : Si on maintenant fixés le poids et le cholesterol la côte estimée (odds : prob(coron)/prob(non coron)) d'un évenement coronariennes est multiplies par $exp(10 \times 0.053003641) = 1.699$ pour tout accroissement de l'age d'une dizaine d'années

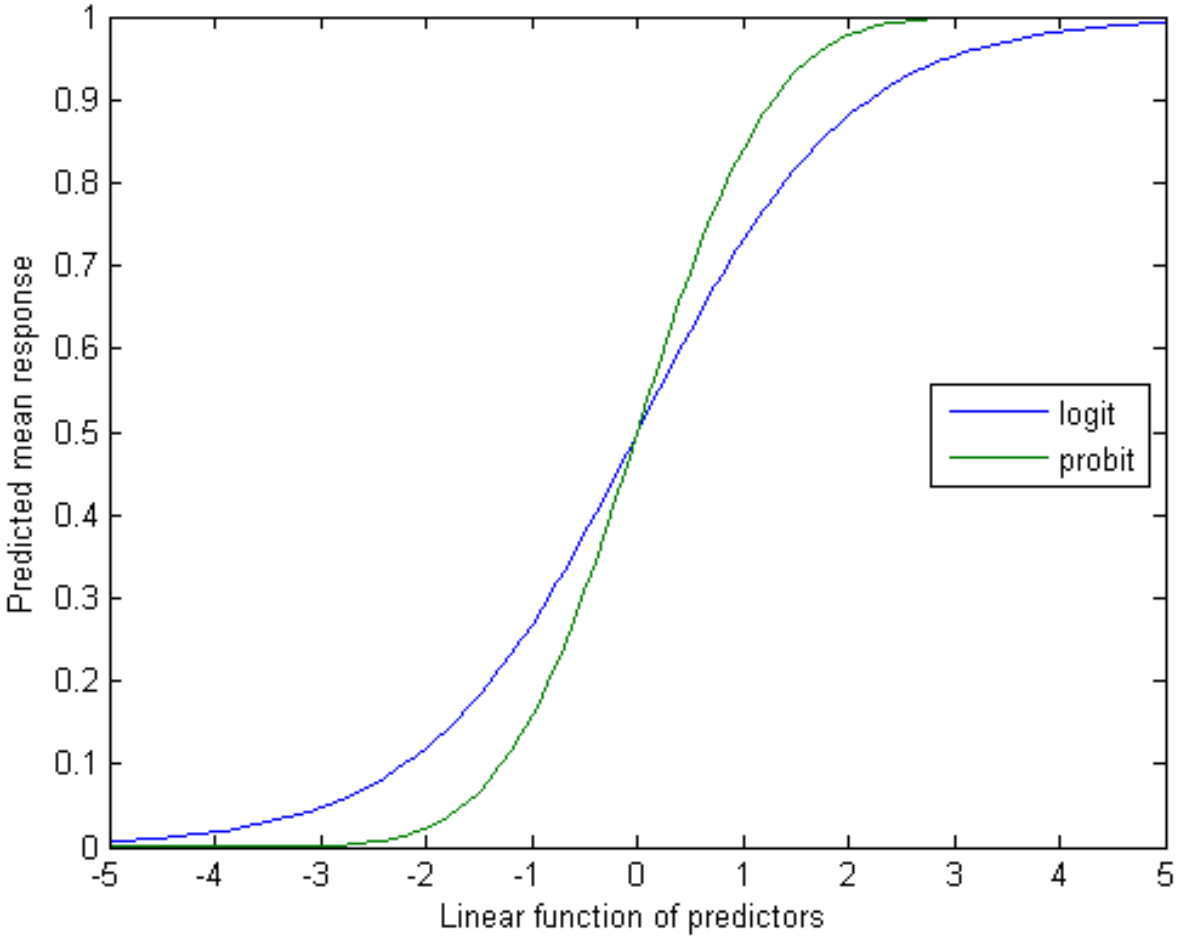


FIGURE 3.2 – FONCTIONS LOGIT ET PROBIT

Chapitre 4

Application

4.1 Introduction

l'étude épidémiologique des données ci - dessous a pour but d'identifier et d'analyser les facteurs de risques de la maladie coronarienne chez les patients.

On dispose d'un échantillon de 200 patients de sexe masculin qui sont attients ou non de maladies coronariennes (coron)(Y=0 ont été ; Y=1 non atteints) . Le but est d'expliquer Y par les variablles suivantes : l'âge (age) , le poids(wt), la taille (height), la tension (sbp et dbp, systolique et diastolique) et le taux de cholestérol (chol) .

les modèles explicatifs appliqués à la maladie coronarienne sont le modèle logit et le modèle probit.

L'intérêt de ce travail provient du fait que ces méthodes, ont été introduites dans certains logiciels récents spécialisés, utilisés, notamment dans le domaine épidémiologique.

4.2 Matériels et Méthodes

Les modèles logistiques à logit généralisés et modèles probit ont été appliqués avec comme variable expliquée maladie coronarienne, et variables explica-

tives l'âge, le poids, la taille , la tension(systolique et diastolique) et le taux de cholestérol.

l'échantillon considéré comprend 200 hommes . a partir des renseignements de l'étude 6 variables sont constitués.

Le programme utilisé pour traiter ces données est les programmes du logiciel R([15]).

4.3 Resultats et Commentaires.

Modèle 1 :

Modèle logistique saturé

Le tableau (4.1) montre les relations entre la variable coronn et les variables explicatives l'âge (age) , le poids(wt), la taille (height), la tension (sbp et dbp, systolique et diastolique) et le taux de cholestérol (chol) pour le modèle logistique.

	Estimate	Std.Errror	Z value	Pr(>|z|)
(Intercept)	-4.538765	7.479006	-0.607	0.5439
age	0.045908	0.023528	1.951	0.0510.
sbp	0.006826	0.020194	0.338	0.7353
dbp	-0.006784	0.038351	-0.177	0.8596
chol	0.006296	0.003631	1.734	0.0829 .
height	-0.073488	0.106198	-0.692	0.4889
wt	0.020028	0.009885	2.026	0.0427 *

TABLE 4.1 – Estimation d'un Modèle saturé logit

Les sorties du logiciel R sont :
Null deviance : 154.55 on 199 degrees of freedom
Residual deviance : : 134.91 on 193 degrees of freedom AIC : 148.91
Number of Fisher Scoring iterations : 5
Ce modèle possède 7 paramètres et il s'écrit alors :

$$y = -4.5387 + 0.0459age + 0.0068sbp - 0.0067dbp + 0.0062chol - 0.0734height + 0.020wt$$

Les coifficients qui correspondent à P<0.05 peuvent être concidère comme significatif ; l'age et le poids ont un effet sur la maladie .

si on juge d'apres le degré de signification $\mathrm{P}r(> |z|)$ aucun paramètre n'est significatif le log de vraisemblance $\mathrm{L}_1 = 134.91$, le critère AIC(Akaike Information Criterion définie comme AIC=-2Log Vraissemblance +2. nbre des paramètres) AIC=148.91
On peut donc rejeter avec un d'effet de la tention et la taille sur la maladie .

Modèle probit saturé

Le tableau (4.2) montre les relations entre la variable coronn et les variables explicatives l'âge (age) , le poids(wt), la taille (height), la tension (sbp et dbp, systolique et diastolique) et le taux de cholestérol (chol) pour le modèle probit.

	Estimate	Std.Errror	Z value	Pr(>|z|)
(Intercept)	-2.482580	3.993068	-0.622	0.5341
age	0.026419	0.012620	2.093	0.0363 *
sbp	0.003821	0.011362	0.336	0.7366
dbp	-0.004554	0.020863	-0.218	0.8272
chol	0.003144	0.001989	1.580	0.1141
height	-0.039455	0.056858	-0.694	0.4877
wt	0.010861	0.005455	1.991	0.0465 *

TABLE 4.2 – Estimation d'un Modèle saturé probit

Les sorties du logiciel R sont :

Null deviance : 154.55 on 199 degrees of freedom
Residual deviance : : 134.65 on 193 degrees of freedom AIC : 148.65
Ces resultats sont similaire à ceux du tableau (4.1)

D'apres les valeurs de $Pr(> |z|)$, aucune des valeurs explicatives ne paraît être déterminante, et voire on peut essayer d'enlever les variables sbp , dbp et height . La valeur de la log de vraisemblance $L_1 = 134.65$, le critère AIC(Akaike Information Criterion) AIC=148.65

Modèle 2 :

Modèle logistique

Le tableau (4.3) montre les relations entre la variable coronn et les variables explicatives l'âge (age) , le poids(wt) et le taux de cholestérol (chol) pour le modèle logistique.

	Estimate	Std.Errror	Z value	Pr(>|z|)
(Intercept)	-9.237982	2.071162	-4.460	8.18 e^{-06} ***
age	0.053010	0.020822	2.546	0.0109*
chol	0.006509	0.003588	1.814	0.0696 .
wt	0.017451	0.008279	2.108	0.0350 *

TABLE 4.3 – Estimation d'un Modèle logit

Les sorties du logiciel R sont :

Null deviance : 154.55 on 199 degrees of freedom
Residual deviance : : 135.58 on 193 degrees of freedom AIC : 143.58
Number of Fisher Scoring iterations : 5
ce modèle s'écrit ainsi :

$$y = -9.237982 + 0.0530age + 0.0065chol - 0.0174wt$$

Modèle probit

Le tableau (4.4)montre les relations entre la variable coronn et les variables explicatives l'âge (age), le poids(wt) et le taux de cholestérol (chol) pour le modèle probit.

	Estimate	Std.Errror	Z value	Pr(>|z|)
(Intercept)	-5.029043	1.078867	-4.661	3.14 e^{-06}***
age	0.029916	0.011231	2.664	0.00773 **
chol	0.003226	0.001974	1.634	0.10227
wt	0.009425	0.004545	2.074	0.03810 *

TABLE 4.4 – Estimation d'un Modèle probit

Les sorties du logiciel R sont :
Null deviance : 154.55 on 199 degrees of freedom
Residual deviance : : 135.25 on 196 degrees of freedom AIC : 143.25
Ces resultats sont similaires à ceux du tableau (4.3)

si on regarde à nouveaux les valeurs de $\Pr(> |z|)$ l'age et le poids sont des facteurs significatifs pour le survenue d'incidents coroniens .
On peut eliminer le cholesterol , qui est correlé avec l'age

Modèle 3 :

Modèle logistique

Le tableau (4.5) montre les relations entre la variable coronn et les variables explicatives l'âge (age) et le poids(wt) pour le modèle logistique. Ce modèle possède 3 paramètres.

	Estimate	Std.Errror	Z value	Pr(>|z|)
(Intercept)	-7.499292	1.709867	-4.386	1.16 e^{-05} ***
age	0.063579	0.019674	3.232	0.00123**
wt	0.015926	0.007963	2.000	0.04549 *

TABLE 4.5 – Estimation d'un Modèle logit

Les sorties du logiciel R sont :

Null deviance : 154.55 on 199 degrees of freedom
Residual deviance : : 138.81 on 197 degrees of freedom AIC : 144.81
cet modèle s'écrit comme :

$$y = -7.499292 + 0.063579age + 0.015926wt$$

On voit que l'age et le poids sont des facteur significatifs pour la survenue de la maladie coronarienne .
cependant le modèle ne parait pas très performant pour ce qui est de diminuer la deviance par à rapport au modèle nulle sous covariables.
En fait on calcul R^2

$$R^2 = \frac{\mathbf{154.55 - 138.8}}{\mathbf{154.55}}$$

$$R^2 = \mathbf{0.10} = \mathbf{10\ \%}$$

Ce la veut dire qu'il ya seulement 10% de la deviance du modèle saturé sont expliques par notre modèle .

Si on utille toutes les variables , R^2=13%
($R^2 = \frac{\mathbf{154.6 - 134.9}}{\mathbf{154.6}} = 0.13 = 13\%$) donc il ya que 13% de la deviance du modèle saturé.

Mais cela n'est pas inhabituel dans le regression logistique .Le modèle complet à un R^2 plus grand, mais on lui préferé un modèle plus parcimonieux (plus economique) qui a l'avantage d'être plus stable.

Modèle Probit

Le tableau (4.6) montre les relations entre la variable coronn et les variables explicatives l'âge (age) et le poids(wt) pour le modèle probit. Ce modèle possède 3 paramètres.

	Estimate	Std.Errror	Z value	Pr(>|z|)
(Intercept)	-4.245736	0.914358	-4.643	3.43 e^{-06} ***
age	0.034972	0.010516	3.236	0.000883***
wt	0.009139	0.004411	2.072	0.038251*

TABLE 4.6 – Estimation d'un Modèle Probit

Les sorties du logiciel R sont :

Null deviance : 154.55 on 199 degrees of freedom
Residual deviance : : 138.14 on 197 degrees of freedom AIC : 144.14
cet modèle s'écrit comme :

$$y = -4.245736 + 0.034972age + 0.009139wt$$

même interprétation que précèdent.

Modèle 4 : Le meilleur modèle par la fonction step

Modèle logistique

Nous pouvons maintenant essayer d'obtenir un modèle de maniere automatique en utilisant la fonction **step** de **R**([42]).

.

	Df	Deviance	AIC
dbp	1	134.74	146.74
sbp	1	134.81	146.81
height	1	135.17	147.17
<none>		134.72	148.72
chol	1	137.66	149.66
age	1	138.68	150.68
wt	1	139.00	151.00

TABLE 4.7 – resultat step

Start : AIC=148.72
coronn ∼ age + sbp + dpb + chol + height + wt

	Df	Deviance	AIC
sbp	1	134.83	144.83
height	1	135.17	145.17
<none>		134.74	146.74
chol	1	137.66	147.66
age	1	138.71	148.71
wt	1	139.29	149.29

TABLE 4.8 – resultat step

Step : AIC=146.74
coronn ∼ age + sbp + chol + height + wt

	Df	Deviance	AIC
height	1	135.32	143.32
<none>		134.83	144.83
chol	1	137.86	145.86
wt	1	139.91	147.91
age	1	140.23	148.23

TABLE 4.9 – resultat step

Step : AIC=144.83
coronn ∼ age + chol + height + wt

	Df	Deviance	AIC
<none>		135.32	143.32
chol	1	138.52	144.52
wt	1	139.93	145.93
age	1	142.16	148.16

TABLE 4.10 – resultat step

Step : AIC=143.32
coronn ∼ agee +chol + wt

Le meilleur modèle obtenu par la regression logistique pas à pas qui chercher à trouver le modèle qui a la valeur minimal pour cp (qui est le critère AIC)

$$y{=}\alpha + \beta_1 age + \beta_2 cho + \beta_3 wt$$

dont le fit est obtenu par modèle (1)

Le meilleur modèle que nous avions trouves est :

$$coron \sim age + wt$$

Il est difficile de choisir entre ces deux modèles. si on regarde les dessin des residus partiels pour chacun des deux modèle .

Aucun des deux modèle ne paraît préférable à l'autre . Pour le modèle qui utilise le poids, l'age et le cholesterol . Le graphe des residus partiel suggérer qu'il est necessaire de tenir compte du cholesterol.

Les residus partiels ceux que l'on obtient en utilisant seulement le poids et l'age n'apperaient pas lies linéairement au cholesterol. En l'absence d'autre critere de selection. On decide choisir le modèle qui ce le plus petit AIC ce sera :

$$coron \sim age + chol + wt$$

Modèle probit

Nous pouvons maintenant essayer d'obtenir un modèle de maniere automatique en utilisant la fonction **step** de **R**([42]) pour le modèle probit.

.

	Df	Deviance	AIC
dbp	1	134.70	146.70
sbp	1	134.76	146.76
height	1	135.12	147.12
<none>		134.65	148.65
chol	1	137.36	149.36
wt	1	138.77	150.77
age	1	139.01	151.01

TABLE 4.11 – resultat step

Start : AIC=148.65
coronn ∼ age + sbp + dpb + chol + height + wt

	Df	Deviance	AIC
sbp	1	134.76	144.76
height	1	135.13	145.13
<none>		134.70	146.70
chol	1	137.36	147.36
wt	1	138.98	148.98
age	1	139.07	149.07

TABLE 4.12 – resultat step

Step : AIC=146.7
coronn ∼ age + sbp + chol + height + wt

	Df	Deviance	AIC
height	1	135.25	143.25
<none>		134.76	144.76
chol	1	137.50	145.50
wt	1	139.58	147.58
age	1	140.62	148.62

TABLE 4.13 – resultat step

Step : AIC=144.76
coronn ∼ age + chol + height + wt

	Df	Deviance	AIC
<none>		135.25	143.25
chol	1	138.14	144.14
wt	1	139.93	145.93
age	1	142.55	148.55

TABLE 4.14 – resultat step

Step : AIC=143.25
coronn ∼ agee +chol + wt

Le meilleur modèle obtenu par la regression probit pas à pas qui à chercher trouver le modèle qui à la valeur minimal pour CP (qui est le critère AIC)

$$y{=}\alpha + \beta_1 age + \beta_2 cho + \beta_3 wt$$

Les residus partiels ceux que l'on obtient en utilisant seulement le poids et l'age n'apperaient pas lies linéairement au cholesterol. En l'absence d'autre critere de selection. On decide choisir le modèle qui ce le plus petit AIC ce sera :

$$coron \sim age + chol + wt$$

4.4 Interprétation des paramètre

Si on maintenant fixés le poids et le cholesterol la côte estimée (odds : prob(coron)/prob(non coron)) d'un évenement coronariennes est multiplies par $exp(10 \times 0.053003641) = 1.699$ pour tout accroissement de l'age d'une dizaine d'années .

Comme les estimateurs des beta sont approximativement normaux , un intervalle de confiance approximatif à 95 % pour β (age) est égal à $0.05300364 + / - 2 \times 0.020821917$, Soit [0.01135981 ;0.09464748] . est un intervale de confiance approximatif à 95% pour un accroissement de la cote lorsque l'âge augment de 10 ans est (exp(10 × 0.01135891)) , coronarien augment d'un facteur situé entre 1.120 et 2.58 par tranche de 10 ans d'âge , à poids et taux de cholestérol constants .
On peut bien sûr faire des calculs analogues poids et pour le taux de cholestérol.

Si on fixés le poids et l'age la côte estimée d'un évenement coronariennes est multiplies par $exp(54 \times 0.006509) = 1.421$ pour tout accroissement de cholestérol d'une 54 mg/dl .

Comme les estimateurs des beta sont approximativement normaux , un intervalle de confiance approximatif à 95 % pour β (chol) est égal à $0.006509 + / - 2 \times 0.003588$, Soit [-0.000667 ;0.013685] . est un intervale de confiance approximatif à 95% pour un accroissement de la cote lorsque le cholestérol augment de 54 mg/dl est (exp(54 × −0.000667)) , coronarien augment d'un facteur situé entre 0.965 et 2.093 par tranche de 54 mg/dl de cholestérol , à poids et age constants .

Si on fixés le cholestérol et l'age la côte estimée d'un évenement coronariennes est multiplies par $exp(10 \times 0.017451) = 1.191$ pour tout accroissement de poids d'une 10Kg .

Comme les estimateurs des beta sont approximativement normaux , un intervalle de confiance approximatif à 95 % pour β (poids) est égal à $0.017451 +/- 2 \times 0.008279$, Soit [0.000893 ;0.034009] . est un intervale de confiance approximatif à 95% pour un accroissement de la cote lorsque le poids augment de 10 Kg est $(\exp(10 \times 0.000893))$, coronarien augment d'un facteur situé entre 1.010 et 1.405 par tranche de 10 Kg de poids , à age et taux de cholestérol constants .

4.5 Discussion

4.5.1 Comparaison des modèles à partir de résultats précédents pour Les modèles saturé M1 ,M2 :

Les résultats obtenus aux tableaux 1 et 2 sont pratiquement les même .

On voit que l'effet du poids , l'age et le taux du cholestérol sont significatifs pour les deux modèles (logit et probit) l'effet de la tension systolique et diastolique et la taille n'est pas significatif pour l'existence des signes de la maladie coronarienne.

Le modèle logistique ne peut pas être rejetté en favori du modèle probit puisque la statistique du log des vraisemblances veut U=D_1 - D_2=19.9
Modèle logit U=19.65
Modèle probit U=19.9
U=19.9 - 19.65 = 0.25 , valeur non significative pour une distribution khi2 , à 5 ddL.

4.5.2 Comparaison des modèles à partir de résultats précédents pour les modèles M3,M4 :

Les résultats des tableaux 3 et 4 montrent que les effets des variables âge , chole , et le poids , sont des facteurs significatif pour la survenue d'accident coronarien et équivalents dans les deux modèles logit et probit
La variable cholestérol à un effet significatif moins marquant mais identique dans les deux modèles .

Par contre , on remarque que l'effet de l'age et le poids sont très significatifs pour la survenue de la maladie.

La comparaison des modèles 1 et 3 ,2 et 4 à partir de la statistique U et du test de Wald montre que la valeur pour une distribution est significative pour une distribution de khideux à cinq degrés de liberté .Le modèle 3 peut donc être choisi plutôt que le modèle 3 même chose pour le modèle 2 et 4

Enfin la comparaison des résultats obtenus avec les modèles 1 et 3 , 2 et 4 par le critére AIC est le plus petit permet de choisir 3,4 plutôt qu'un autre.

Chapitre 5

Conclusion générale

Le but de ce travail est de comparer les différents modèles logit et probit du chapitre (3) . Nous avons comparé et interprété les différents modèles en terme de probabilité et en terme d'estimation des coefficients β

Ces modèles ne différent que par la forme de la fonction de répartition, ceci s'explique par la proximité des familles de lois logistique et normales ces des fonction sont sensiblement proche , mais cette similitude est encore grande si l'on considère la loi logistique transformé .

Les valeurs estimées des paramètres dans les modèles probit et logit ne sont pas directement comparables (voir application) puisque les variances des lois logistique est normale réduite ne sont pas identique .

Cette différence de variance implique que les estimateurs de ces paramètres obtenus dans les deux modèles ne sont pas identiques .

Différente approximations permettant d'approches les estimations des modèles logit et probit à partir des estimations obtenues sont proposes et pressentes et présentes "dans notre application"

Si les deux modèles sont sensiblement identiques , il existe cependant certaines différence entre les modèles probit et logit (voir Application) .

La loi logistique tend à attribuer aux évènements une probabilité plus forte que la distribution normale .

Le modèle logit faciliter l'interprétation des paramètres β associées aux variables explicatives (voir Application)

Dans notre application chapitre 4 . Il n'est pas apparu de différences notables entre les deux modèles logit et probit . Similitude et différence entre les deux modèles ont été démontré et présenter dans l'application

On conclusion, il apparait que les résultats des modèles probit et logit sont généralement similaires que ce soit en termes de probabilité ou en termes d'estimation des coefficient β si l'on tient compte des problèmes de normalisations . Mais il convient d'être prudent dans l'utilisation des approximations pour comparer les modèles probit et logit . Il est préférable de raisonner en termes de probabilité et non termes d'estimations des paramètres β pour comparer les résultats.

Le choix de l'un des deux modèles peut être effectué au vu des distributions de variable expliquée quand celle-ci comprend un nombre suffisant de modalités et les covariables sont en nombre modéré.

Le choix peut aussi eventuellement être effectué on fonction des connaissances a priori sur la distribution de la variable expliquée

Ces deux modèles on été compares car ils existe dans le logiciel R..

Chapitre 6

ANNEXES

6.1 PROGRAMMES INFORMATIQUES

Exemple N° 1 :

pour le modèle logit
variable expliqué : coron
variable explicative : age , wt(poids) , height (la taill), sbp (la tension systolique) , dbp (la tension diastolique) et chol (le taux de cholestérol).
Eexemple de programme R([42]) :

> reg.logit<-glm (coron ∼ age+sbp+dbp+chol+height+wt,
family = binomial (link=logit), data =donnees)
> summary(reg.logit)
> step(reg.logit)
> plot(reg.logit)

Exemple N° 2 :

pour le modèle probit
variable expliqué : coron
variable explicative : age , wt(poids) , height (la taill), sbp (la tension systolique) , dbp (la tension diastolique) et chol (le taux de cholestérol).
Eexemple de programme R([42]) :
>reg.probit<- glm (coron ∼ age+sbp+dbp+chol+height+wt,
family = binomial (link=probit), data =donnees)
> summary(reg.probit)
> step(reg.probit)
> plot(reg.probit)

Exemple N° 3 :

pour le modèle linéaire
variable expliqué : coron
variable explicative : age , wt(poids) , height (la taill), sbp (la tension systolique) , dbp (la tension diastolique) et chol (le taux de cholestérol).
Eexemple de programme R([42]).

> reg.lin<-lm (coron ∼ age+sbp+dbp+chol+height+wt, data =donnees)
> summary(reg.lin)
> step(reg.lin)
> plot(reg.lin)

6.2 GRAPHIQUES

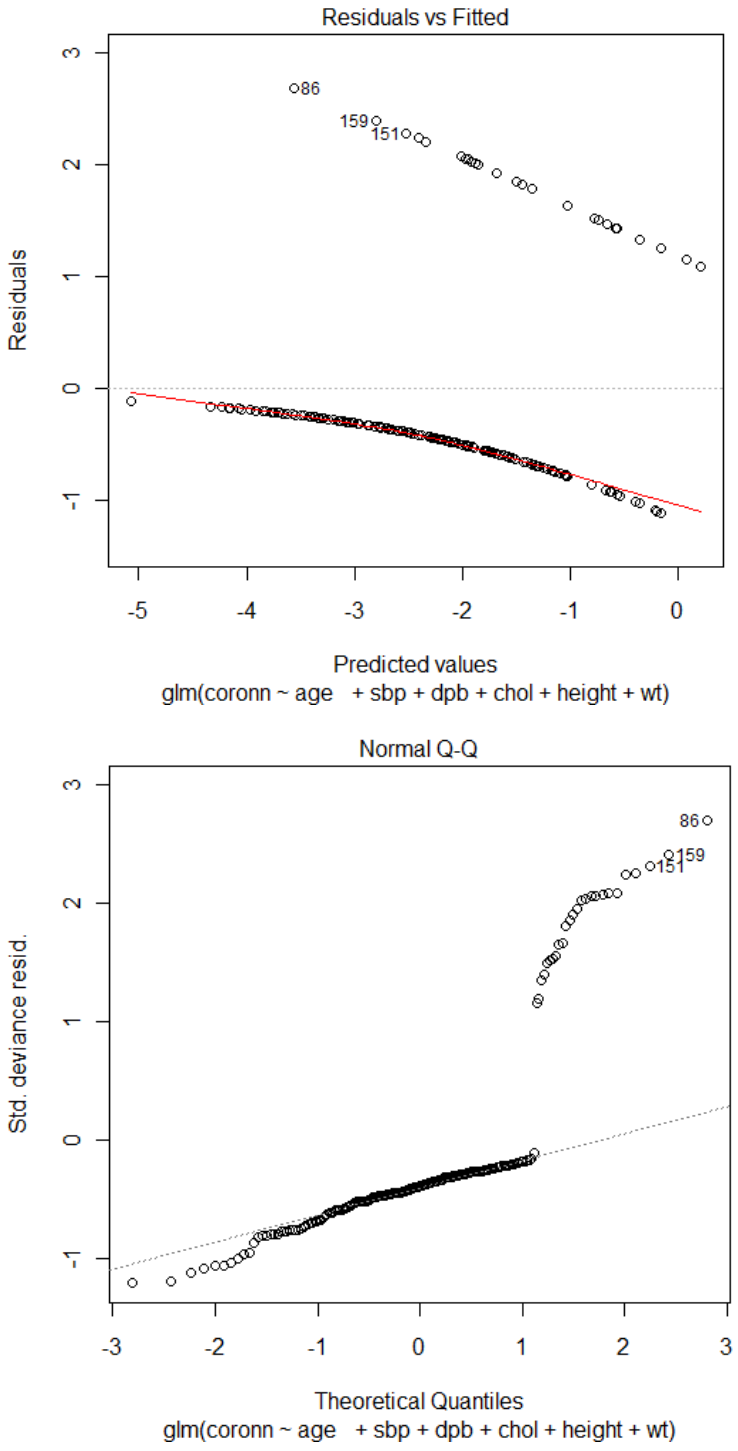


FIGURE 6.1 – Modèle logistique saturé

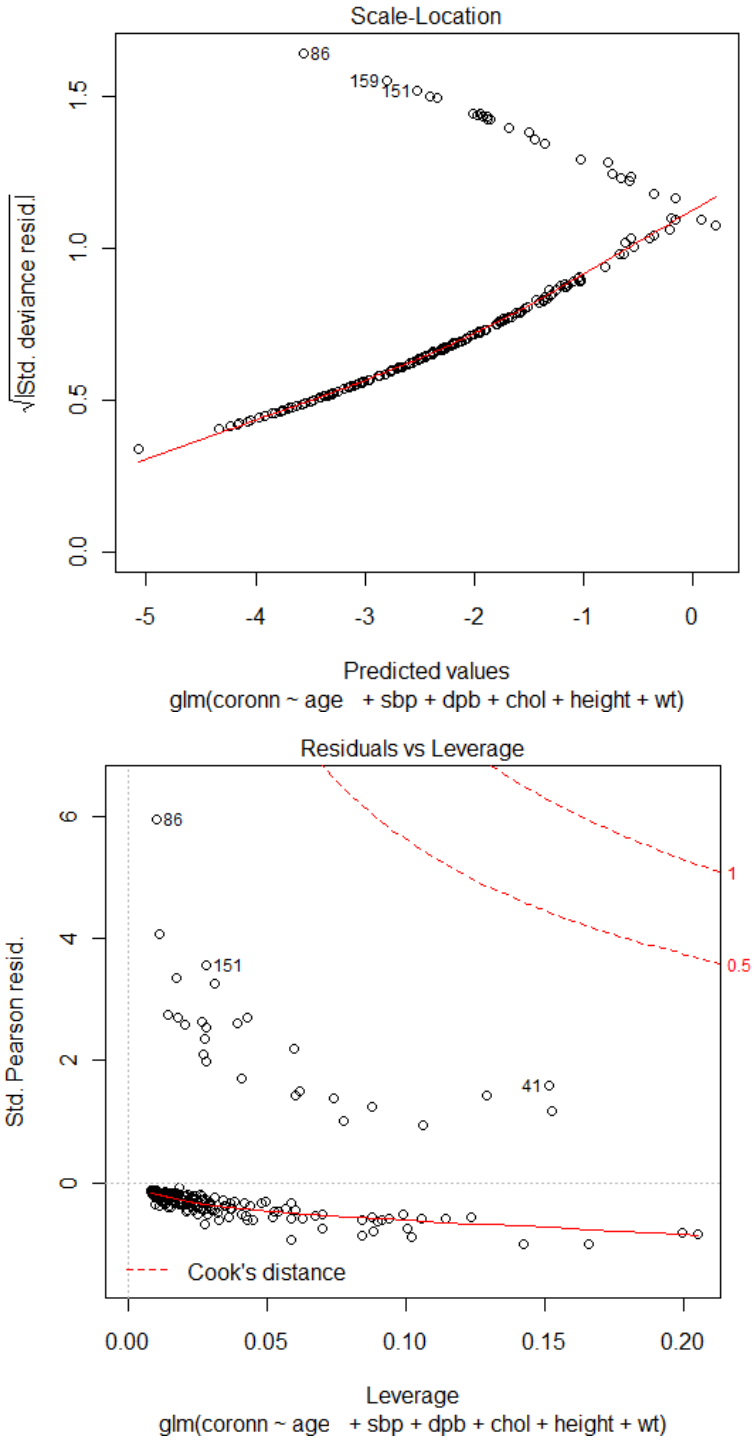


FIGURE 6.2 – Modèle logistique saturé

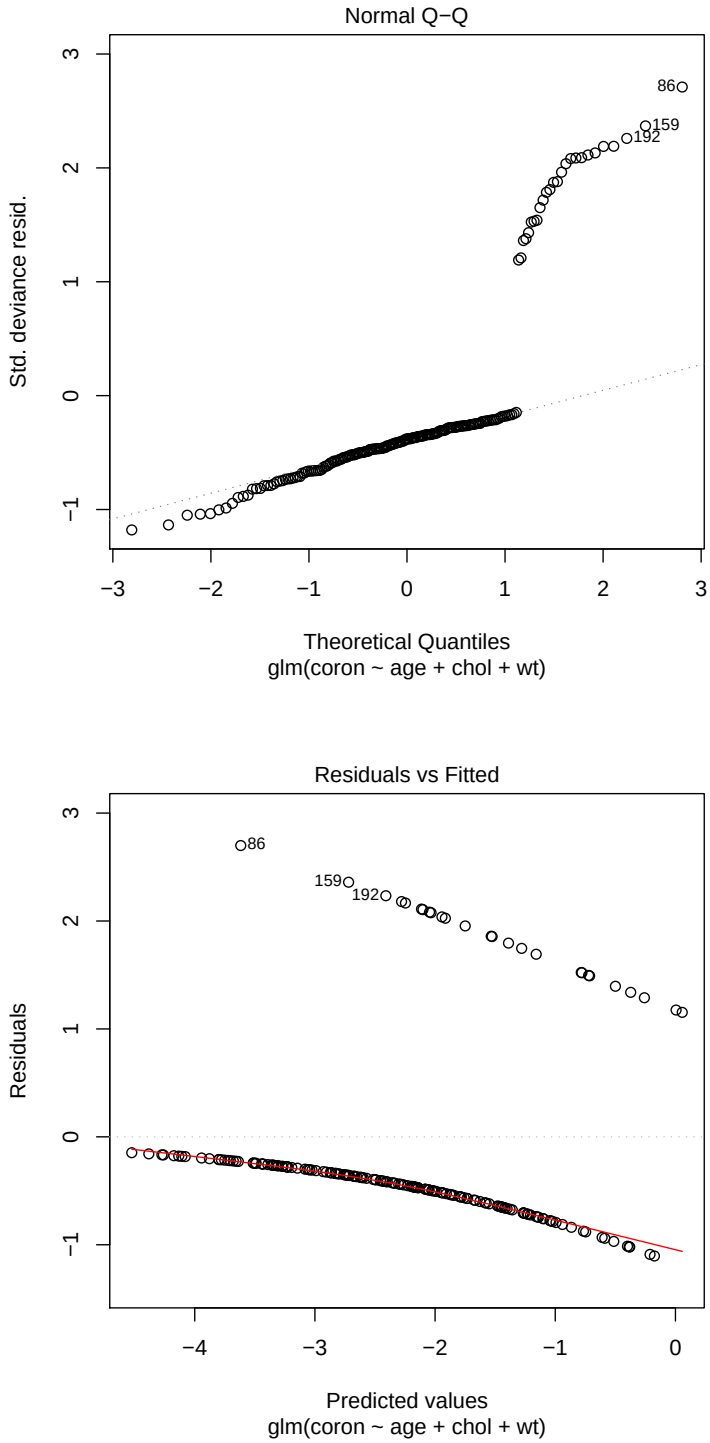


FIGURE 6.3 – Modèle logistique

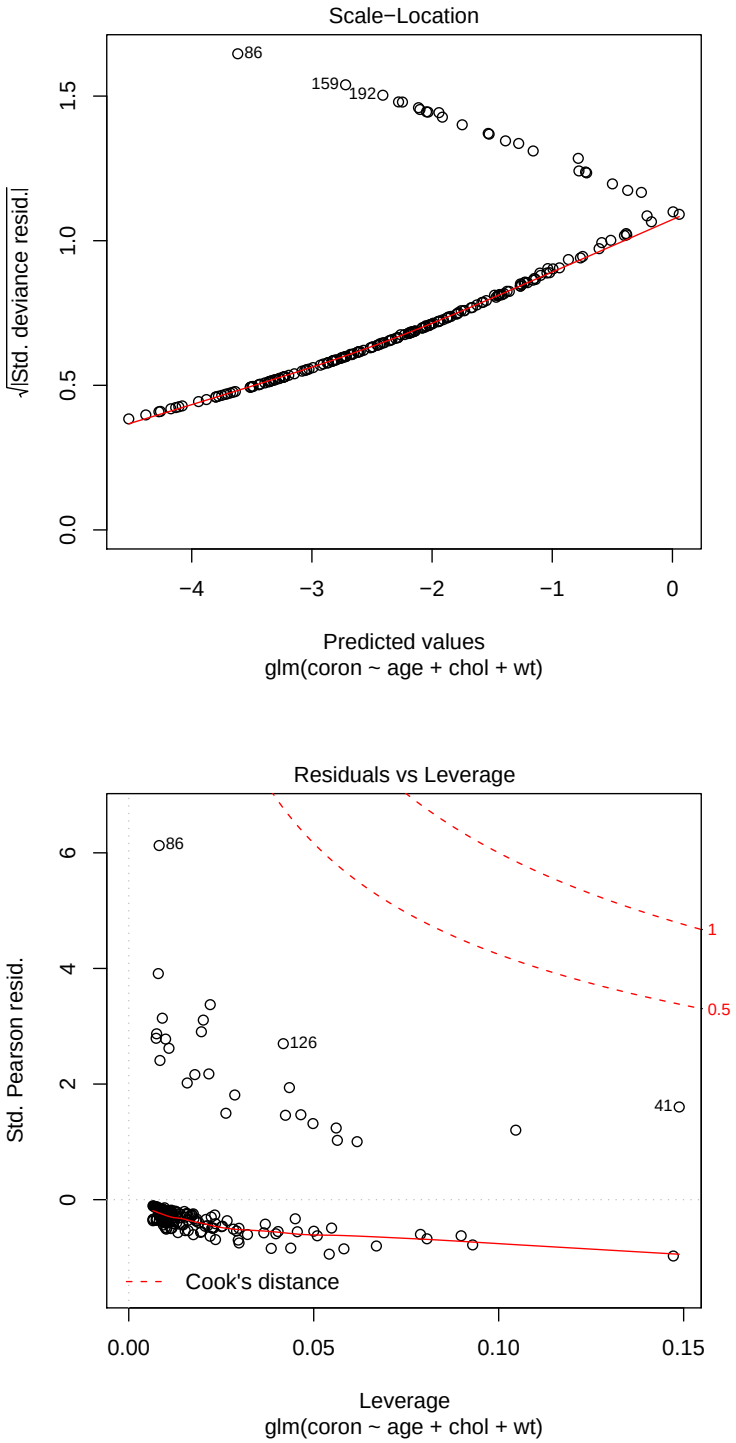
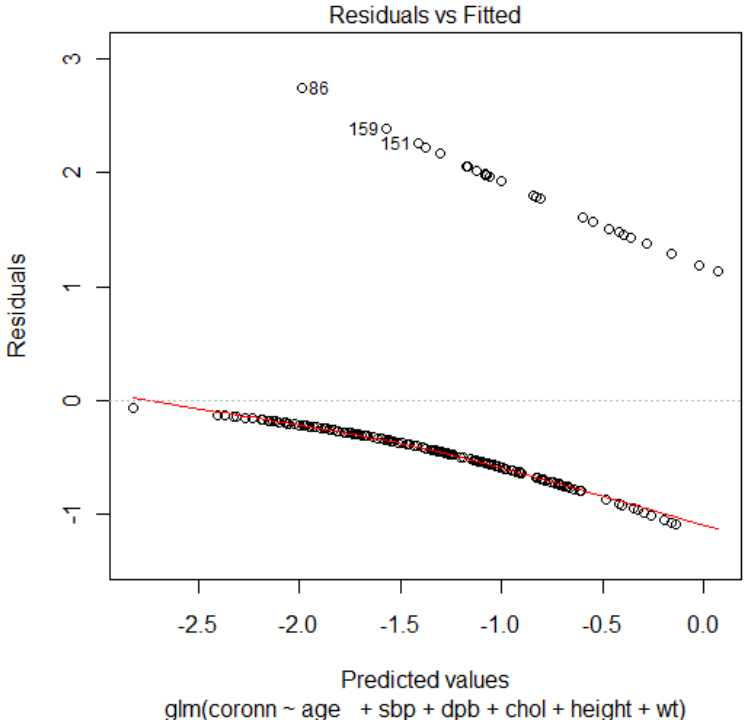


FIGURE 6.4 – Modèle logistique



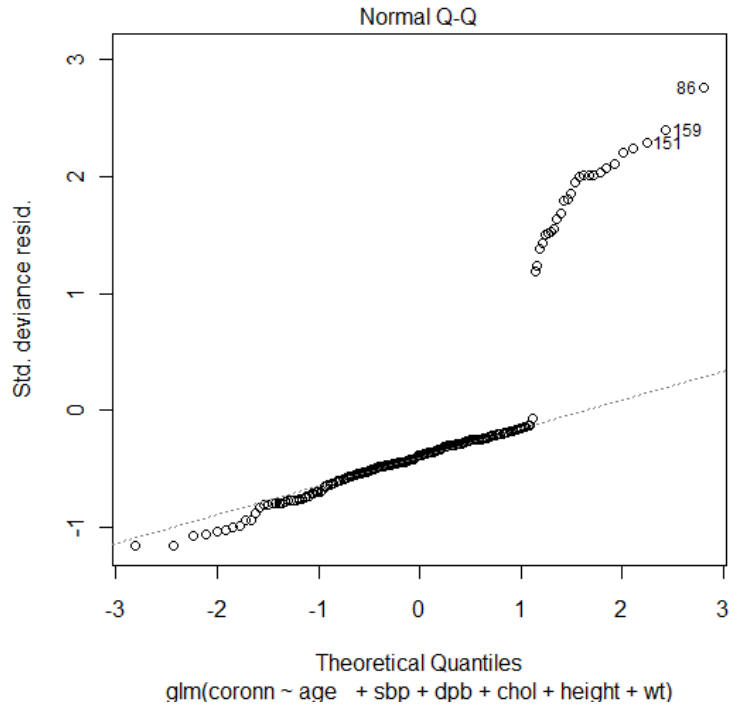


FIGURE 6.5 – Modèle probit saturé

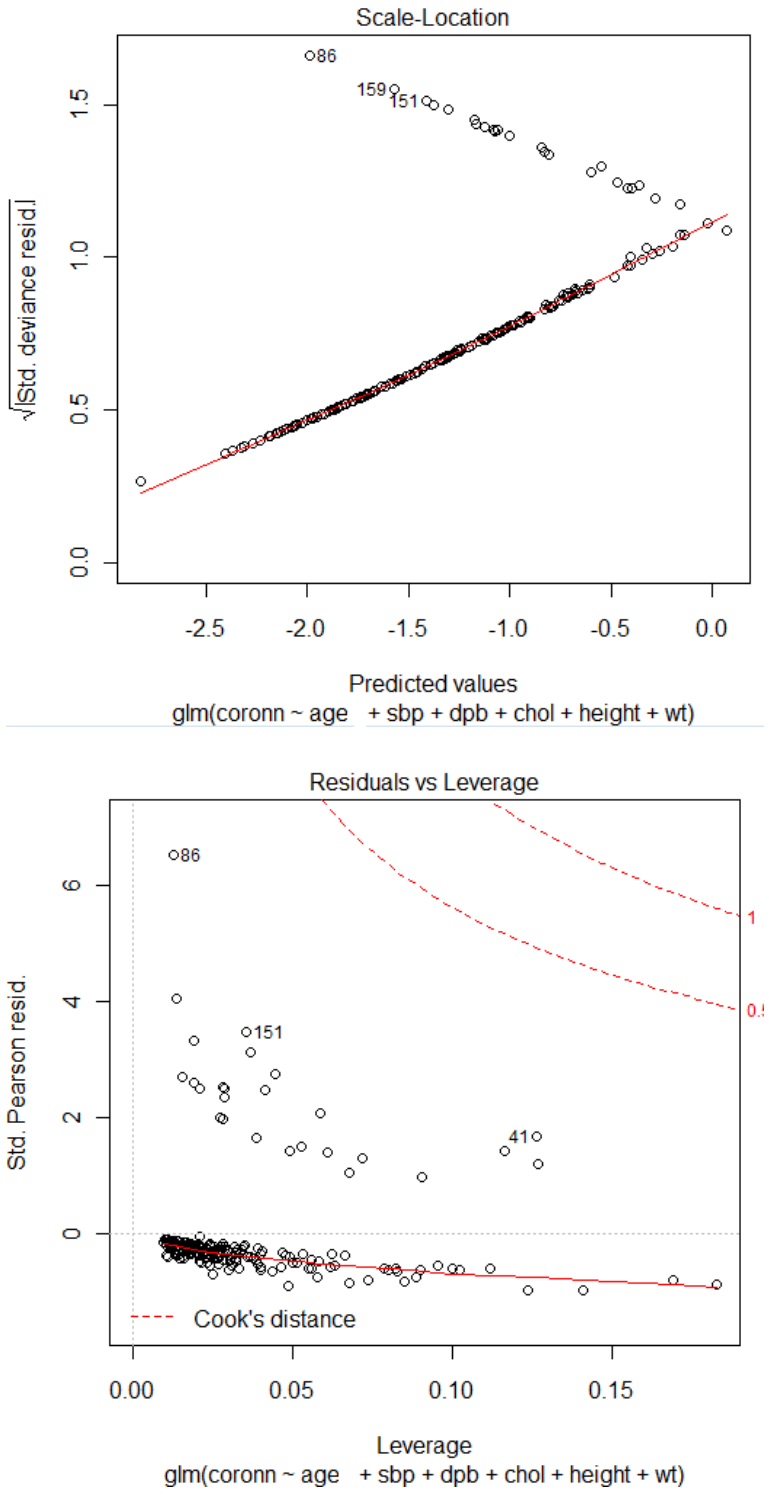


FIGURE 6.6 – Modèle probit saturé

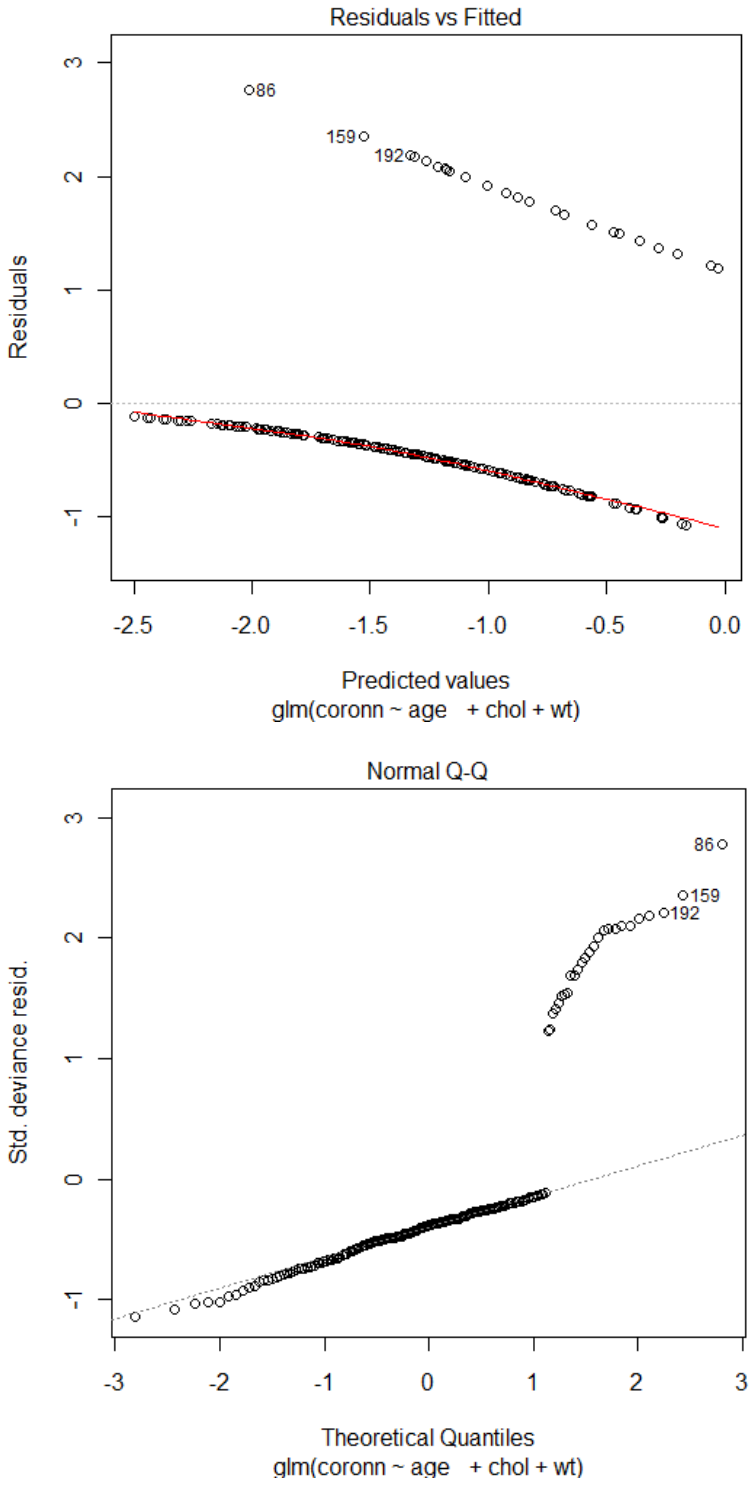


FIGURE 6.7 – Modèle probit

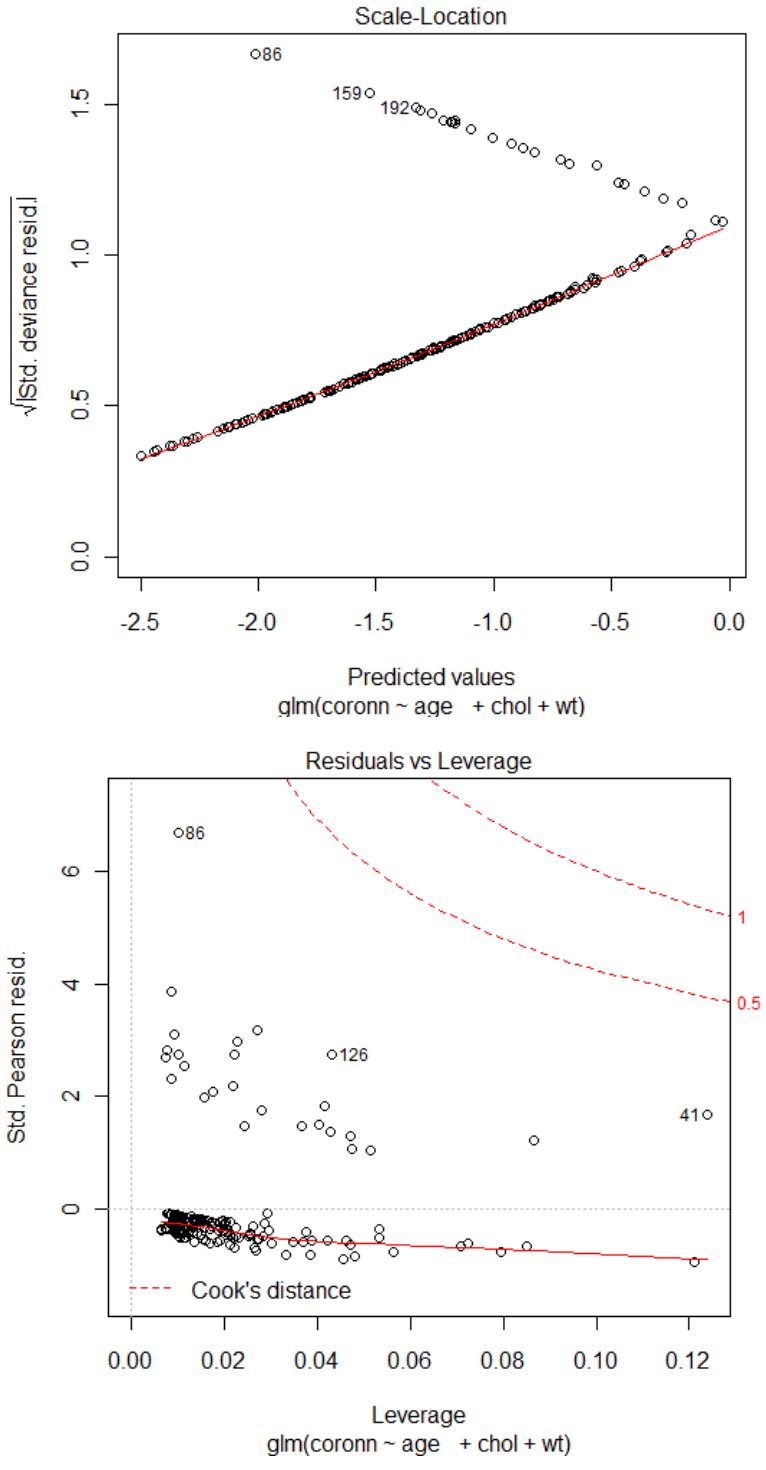


FIGURE 6.8 – Modèle probit

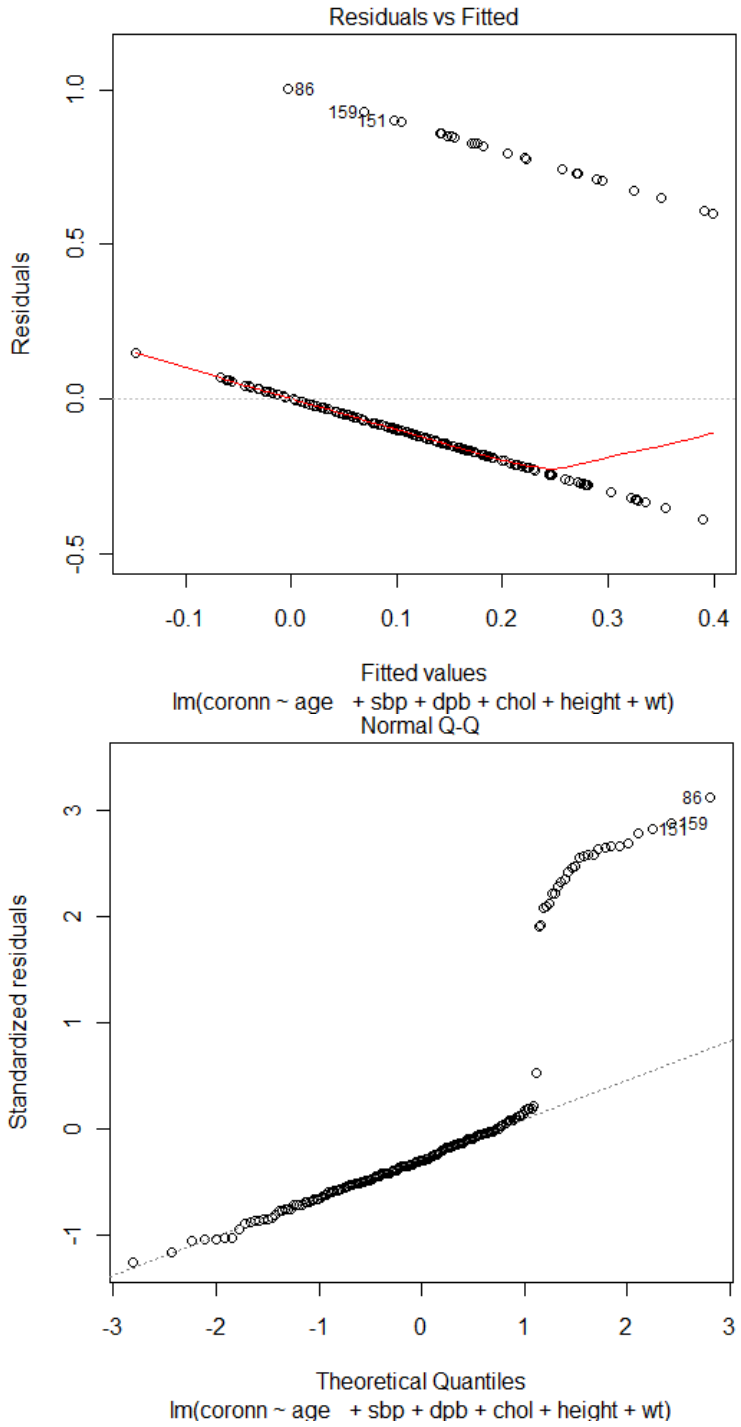


FIGURE 6.9 – Modèle linéaire saturé

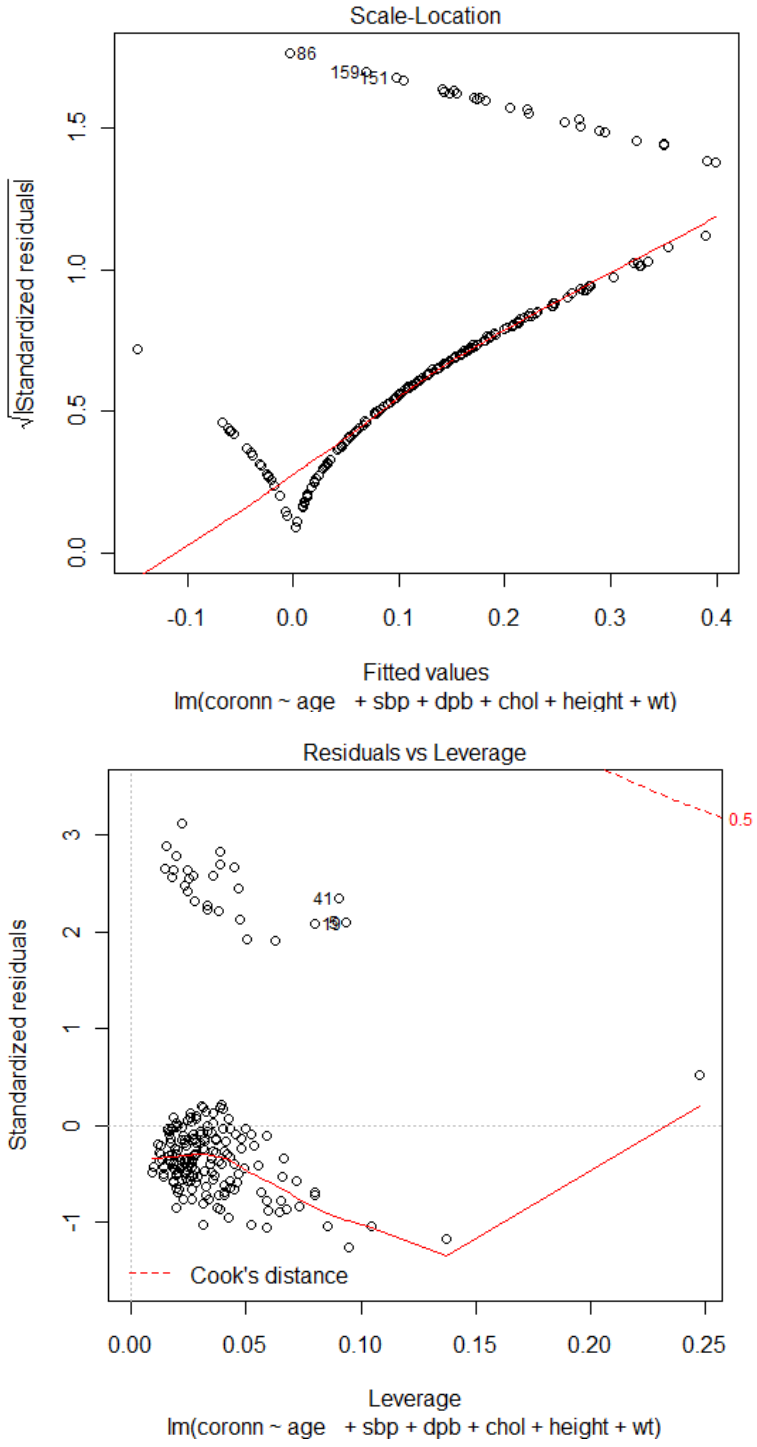
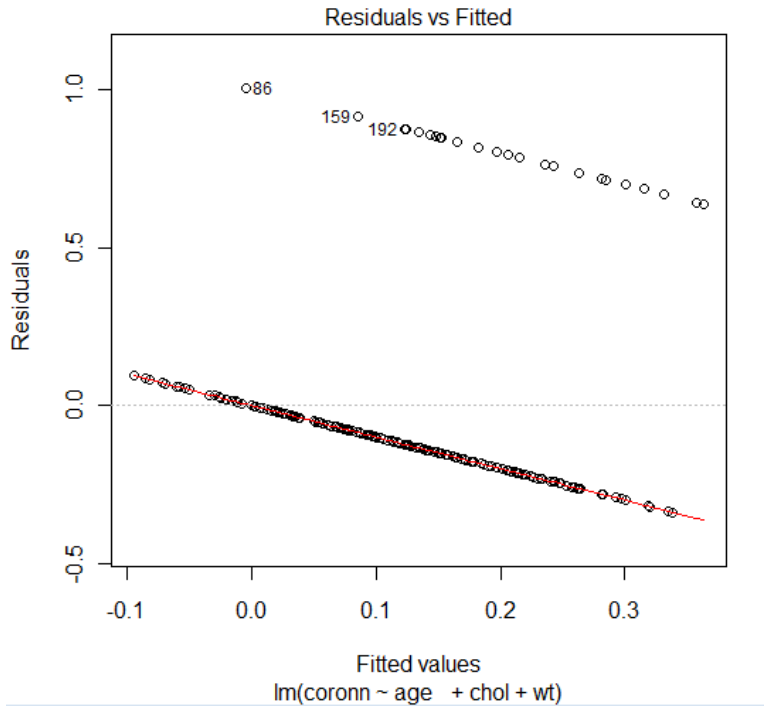


FIGURE 6.10 – Modèle linéaire saturé



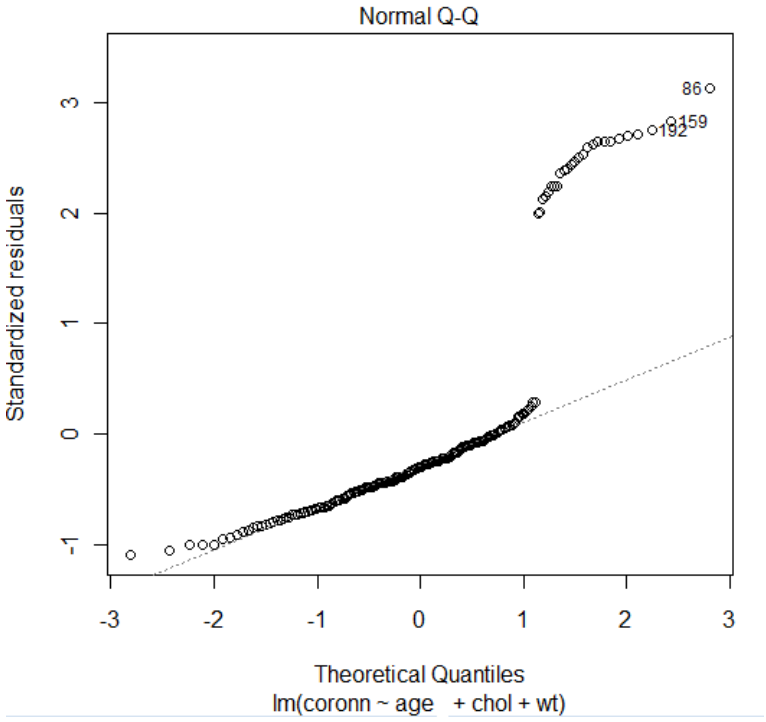


FIGURE 6.11 – Modèle linéaire

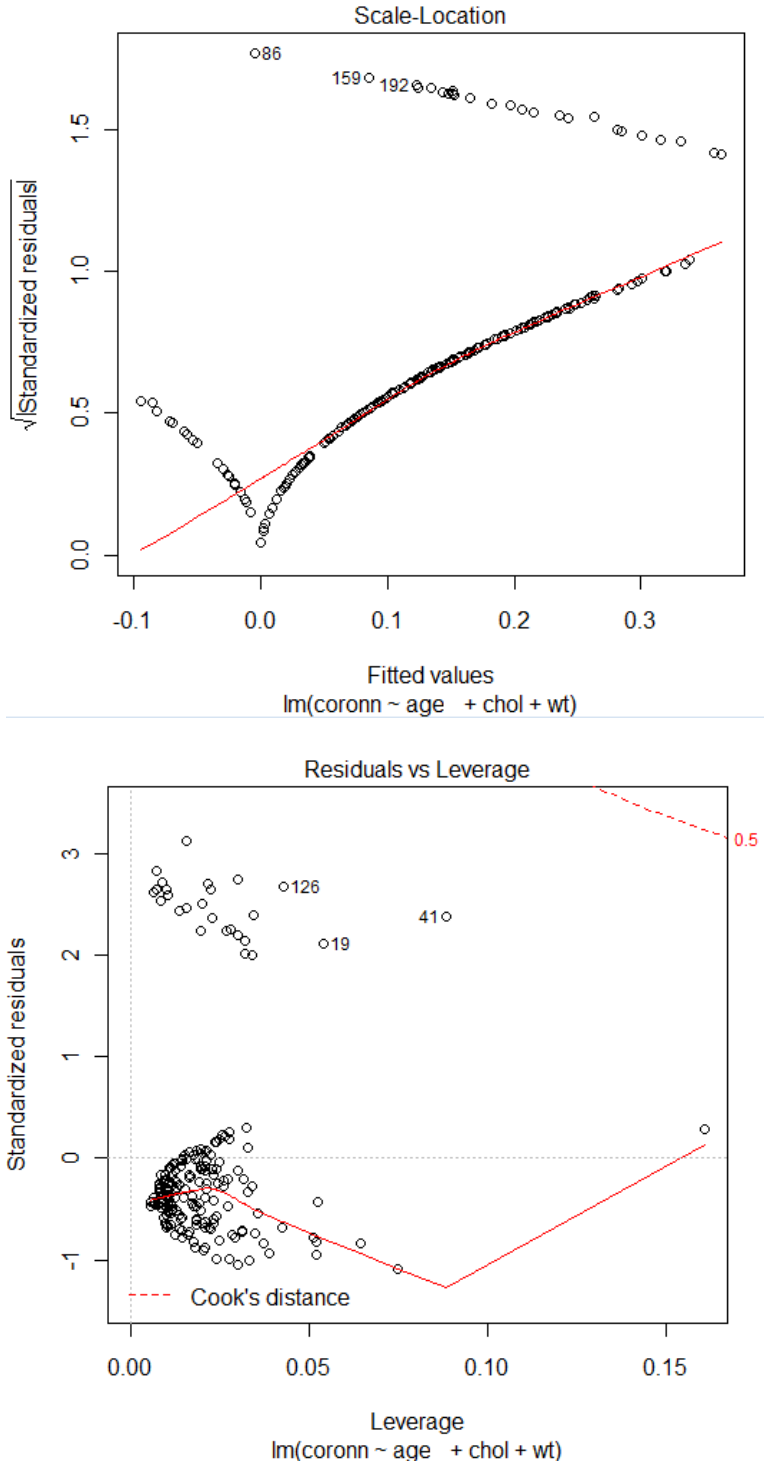


FIGURE 6.12 – Modèle linéaire

Bibliographie

[1] Lee, Youngjo ; Nelder, John Pawitan, Yudi (**2006**). "Modèles linéaires généralisés avec des effets aléatoires" : Analyse unifiée par l'intermédiaire de H-probabilité. Boca Raton : Colporteur et Hall/CRC. ISBN 1-584-88631-5.

[2] McCullagh, Peter ; Nelder, John (**1989**). "Modèles linéaires généralisés". Londres : Colporteur et Hall. ISBN 0-412-31760-5.

[3] CULLAGH M.C. and J. NELDER M.C., (**1983**) : "Generalised Linear Models ", London Chapman and Hall.

[4] Dobson, A.J. ; Barnett, A.G. (**2008**). "Introduction aux modèles linéaires généralisés", troisième édition. Londres : Colporteur et Hall/CRC.

[5] Laurent Rouvière .**(2008-2009)** Régression sur variables catégorielles.

[6] P.L. Gonzales, (**2005**)."Modèles à réponses dichotomiques, in Modèles statistiques pour données qualitatives", Droesbeke, Lejeune et Saporta Editeurs, Chapitre 6, pages 99-136, Technip.

[7] C. Hurlin Modèles Dichotomiques Univariés .Poly . De cours

[8] R. Rakotomalala, Régression logistique - Une approche pour rendre calculable P(Y/X).

[9] D. Garson, "Logistic Regression".

[10] M.C.Cullagh (**1980**)"Regression models for ordinal data"(With discussion), J.Roy.Statist.Soc. ,B42,pp .109-142.

[11] Amemiya, T. (**1985**). "Économétrie avançée". Pression d'université de Harvard. ISBN 0-674-00560-0.

[12] R version 2.8.1 (**2008-12-22**).

[13] Agresti (**1991**). " categorial data Analysis ", John Wiley and Sons ,Inc.

[14] Agresti, Alan. (**2002**). "Analyse de données catégorique". New York : Wiley-Interscience. ISBN 0-471-36093-7.

[15] M Bouchoul-Chikhi , T . Moreau, M Chavance et B Bru .(**1999**) ."Modèle logistique cumulatif pour une variable Réponse Ordinale à c Catégories.", Rev. Sc et Tech , Université Mentouri Constantine, N° 11, pp.17-19,

[16] O.D . Williams , and J. E. Grizzle. (**1972**)."analysis of contingency tables having orderd response cateegories " , J-Amer.Statist.Assoc , 67 , pp. 55-63 .

[17] D.G. Clayton (**1974**)"Some odds-ratio Statistics for the analysis of ordered categorical data ", Biometrics, 61, pp.525-531.

[18] Hardin, James ; Hilbe, Joseph (**2001, 2007**). "Modèles et prolongements linéaires généralisés". Station d'université : Pression de Stata.

[19] Bois, Simon (**2006**). "Modèles additifs généralisés " : Une introduction avec R. Colporteur et Hall/CRC. ISBN 1-584-88474-6.

[20] Bonheur, C.I. (**1935**). "Le calcul de la courbe de dosage-mortalité". Annales de la biologie appliquée (22) 134-167.

[21] Bonheur, C.I. (**1938**)." La détermination de la courbe de dosage-mortalité de petits nombres". Journal trimestriel de la pharmacologie (11) 192-216.

[22] Balakrishnan, N. (**1991**). "Manuel de la distribution logistique". Marcel Dekker, Inc. ISBN 978-0824785871.

[23] Vert, William H. (**2003**). "Analyse économétrique", cinquième édition. Apprenti Hall. ISBN 0-13-066189-9.

[24] Hosmer, David W. ; Stanley Lemeshow (**2000**). "Régression logistique appliquée", 22eme E-D.. New York ; Chichester, Wiley. ISBN 0-471-35632-8.

[25] COX D.R., (**1972**) : " Regression Models and Life Tables (With discussion) ", J. Roy. Statist. Soc., B.34, pp.187-220.

[26] FADEN D. MC, (**1987**) : " Regression Based Specification Tests for the Multinomial Logit Models ", Journal of Econometrics, 34, pp.63-82.

[27] HUBERT C. and GUIHENNEUC.C., (**1999**) : "Certificat de Maîtrise de Biostatistiques et Modélisation", UFR Biomédicale, Paris V.

[28] Mc GULLAGH P and NELDERJ.A., (**1989**) , " Generalized Linear Models ", Chapman and hall, New York.

[29] NIKULIN M.S., (**1973**) :"Chi-square tes for continution distributions with shift and scale parametres ,Theoriy of probability and its applications" , 18, 3, 559-568.

[30] OLIVER F.R.,(**1964**) : "Methods of estimating the logistic growth function, Appli.Statist"., 13,57-66.

[31] REED L.J and BERKSON J, (**1929**) :" The application of the logistic function to expirimental data, J. Physical Chemistry", 33,760-779.

[32] J. Jaccard,(**2001**). "Intercation E _ects in Logistic Regression, Series : Quantitative Applications in the Social Sciences", n0135, Sage Publications.

[33] D.W. Hosmer, S.(**2000**). "Lemeshow, Applied Logistic Regression", Second Edition, Wiley.

[34] S. Menard,(**2002**) "Applied Logistic Regression Analysis (Second Edition)", Series : "Quantitative Applications in the Social Sciences", n0106, Sage Publications.

[35] J.P. Nakache, J.(**2003**) "Confais, Statistique Explicative Appliquée", Partie 2, "Modèle Logistique", pages 77-168, Technip.

[36] A.A. O'Connell(**2006**), "Logistic Regression Models for Ordinal Response Variables", Series : Quantitative Applications in the Social Sciences, n0146, Sage Publications,.

[37] G. Saporta,(**2006**) "Probabilités, Analyse de données et Statistique", Section 18.6, "Régresion logistique binaire (deux groupes)", pages 475-480, Technip.

[38] A. Slavkovic, "STAT 504 - Analysis of discrete data"".

[39] M. Tenenhaus(**2007**), Statistique - Méthodes pour décrire, expliquer et prévoir, Chapitre 11, "La régression logistique binaire", pages 387-460 ; Chapitre 12, "Régression logistique multinomiale : réponses polytomique et ordinale", pages 461-499, Dunod.

[40] Wikipedia, Régression Logistique.

[41] M.Chikhi. (**2010**). Régression logistique binaire. Poly. De cours.

[42] M.Chikhi. (**2011**). Régression logistique multinomiale. Poly. De cours .

RESUME

Le travail présenté dans ce **Mémoire de Magister** s'articule et met en relief principalement une synthèse bibliographique des méthodes de données binaires et porte plus particulièrement sur la **comparaison des modèles Logit et Probit** dans le cas d'une réponse dichotomique.

La première partie consiste en une présentation générale des modèles linéaires généralisés suivi du développement des modèles logistiques et du modèle probit pour une variable réponse binaire en fonction de plusieurs variables explicatives (qualitatives ou quantitatives), les fonctions qui les définissent sont rappelées ainsi que l'interprétation des paramètres par la méthode du Maximum de Vraisemblance.

La deuxième partie porte sur la présentation et l'interprétation de l'étude comparative des modèles Logit et Probit.

Enfin les modèles cités ci-dessus sont appliquées aux données d'une **étude épidémiologique** visant à mettre en évidence, dans un échantillon 200 patients, les facteurs de risques de la maladie coronarienne, les résultats sont comparés discutés et interprétés.

Mots clés : modèle linéaire généralisé, modèle Probit et Logit, Maximum de Vraisemblance, épidémiologie, maladie coronarienne.

Abstract

The work presented in this Thesis Magister articulates and highlights primarily a literature review of methods of binary data and focuses on the comparison of Logit and Probit models in the case of a dichotomous response
The first part consists of an overview of generalized linear models followed by the development of logistic models and the probit model for a binary response variable depending on several variables (qualitative or quantitative), the functions that define them are recalled and the interpretation of parameters by Maximum Likelihood
The second part focuses on the presentation and interpretation of the comparative study of Logit and Probit models.
Finally the models mentioned above are applied to data of an epidemiological study to highlight, in a sample of 200 patients, the risk factors of coronary artery disease ; the results are compared and discussed interpreted.

Keywords : generalized linear model, Probit and Logit, Maximum Likelihood, epidemiology, coronary artery disease.