PEOPLE'S DEMOCRATIC OF ALGERIA
MINISTERY OF HIGHER EDUCATION AND SCIENTIFIC RSEARCH
UNIVERSITY MOULOUD MAMMERI
TIZI-OUZOU

FACULTY OFSCIENCES

DEPARTEMENT OF MATHEMATICS

MASTER'S DISSERTATION
With the a view to obtaning the master degree

Option: Probability-Statistic.

Theme :

Logistic regression by the bayesian approach and application

realized by

Ms **ADGHAR fatma**

Defended in front of the examination board composed by:

Mrs.*ATIL Lynda Senior Lecturer Class A President*
Mr.*FELLAG Hocine Professeur Supervisor*
Mrs*BELKACEM Cherifa Senior Lecturer Class B Examinor*

2020/2021

Thanks

First of all, I thank the almite **Allah** for giving me the strength and courage to achieve this modest work.

I address my sincere thanks to **Mr FELLAG Hocine** who supervised this thesis.

I thank him for his great availability, her patience , his precious advice and contagious optimism and all the help he gave me to complete this work.

It's a great pleasure that I express my gratitude to the members of the jury for agreeing to honor my work with their judgmants.

ALL my gratitude is addressed to all teachers who have fllowed me tirelessly throughout my university studies.

Dedications

I dedicate this modest work To my dear and esteemed parents who have never ceased to guide me to the right path as well as for their sacrifices and their permanent love.

To my dear brothers and sisters .

To my friends: thank you for your love and friendship. You were always there to support me and listen to me.

To the entire promotion of Master 2 PS (2020/2021).

ADGHAR Fatma

Abstract

This dissertation is devoted to study the logistic regression using the Bayesian approach. To obtain the posterior distributions of the parameters regression, the Monte Carlo approximation methods by Markov chain(MCMC) are very powerful and indispensable and have there fore been developed in order to approximate the posteriori distribution when one does not know how to do it analytically.
Keywords Logistic regression ,Bayesian logistic regression,Markov chain(MCMC), approximate,posteriori ditribution.

Contents

List of Tables

List of Figures

Notation and abbreviation

rv	random variable
E	Expectation.
V	Mathematical variance
$\mathbb{R}$	Set of reals
$\mathbb{N}$	Set of natural numbers.
ddl	degree of freedom
CI	Confidence Interval
$MMSE$	Minimum Mean Square Error
MAP	Maximum Posterior
GLM	Generalized Linear Models
$MCMC$	Markov Chain Monte Carlo
iid	independent and identically distributed
ε	independent and identically distributed sample
OR	Odds Ratio
L	Likelihood
Ll	Log Likelihood
D_M	Log Residul deviance
CM	Confusion Matrix
ER	Error Rate
ER	Sucess Rate
Se	Sensibility
TPR	True Positive Rate
Sp	Specifity
ROC	Receiver Operating Characteristic
AUC	Area under the curve
	Receiver Operating Characteristic
ML	Maximum Likelihood

Introduction

The man is curious and this is probably the best explanation for his development from the beginning of mankind to the present day. This need to understand observed phenomena and the desire to anticipate them is at the heart of his preoccupations. This explains the emergence and success .

Is an interdisciplinary art of quantification under uncertainty used by physicists, economists, engineers, geographers, biologists, insurers, psychologists, meteorologists, business managers, etc.
All practitioners concerned with building a bridge between theory and experimental data on solid foundations. Its main objective is to carry out, through the observation of a random phenomenon, an inference on the probability distribution which is at the origin of this phenomenon.

Sometimes the studied phenomenon can be simply described by a few graphical representations of basic data analysis. Often the problem is much more complicated because multiple influenced factors have to be taken into account.
In such situation, the classical statistician uses both deterministic reasoning by the absurd, with the aim to propose acceptable values for the parameters describing the effets of the explanatory factors, and probabilistic reasoning,to reflect the variability of the observed results due to noise.In contrast, the Bayesian statistician uses the same framework of thought to deal with the interaction of these two levels of uncertainty: ignorance of the possible values of the parameters and the randomness of the noise affecting the experimental results.

In statistical data analysis, one is interested in the estimation of several unknown parameters in models such as regression models. Regression is concerned with the mean change in a dependent variable to be explained

conditional on explanatory variables. Usually, the unknown parameters or regression coefficients of the model are estimated by the least squares method or by the maximum likelihood method, however, one might want to know more about the conditional distribution of the variable to be explained.
A statistical model can also be perturbed when the dependent variable is qualitative which is represented by the logistic model. In this case, we are interested in logistic regression.

Logistic regression is a mathematical modeling technique that can be used to describe the relationship of several qualitative or quantitative dependent variables. It is a special case of the generalized linear model (GLM)which been formulated by John Nelder and Robert in 1972 as a flexible generalization of linear regression [17].

The Bayesian approach is increasingly used in statistical analysis for the adjustmentand the study of several regression models including the logistic model known as Bayesian logistic regression. To obtain the posterior distributions of the parameters regression, the Monté Carlo approximation methods by Markov chain(MCMC) are very powerful and indispensable and have there fore been developed in order to approximate the posteriori ditribution when one does not know how to do it analytically.

Some works on the application of Bayesian methods to logistics models are continued in Genkin and al [9], as for example the Bayesian logistic regression applied to the categorization of linguistic texts and also the comparison of the two estimation methods classical and Bayesian. Mila and Michailides (2006) studied the prediction of the severity ofpanicle and scorch of pistachio shoots in California using lo-Bayesian gistics [15], they noted that Bayesian methods gave moreconsistent. Gordovil, Guardia, Pero and Fuente (2010) presented the Bayesian estimate asan alternative to the classical estimation procedures by logistic regression in the study of theAttention Deficit Hyperactivity Disorder (ADHD) in a Mexican sample [10].
Logistic regression is widely used in a wide variety of fields. We can citein a non-exhaustive way:

- In medicine, it allows for example to find the factors that characterize a groupof sick subjects compared to healthy subjects.

- In insurance, it makes it possible to target a fraction of the clientele who will be sensitive to a policyinsurance on a particular risk.

- In econometrics, to explain a discrete variable. For example, voting intentionsin the elections.

- In the banking sector, to detect groups at risk when subscribing to acredit.

Many other types of applications have been described in the literature, for example White,Pearson and Wilson (1999) examined the implementation of Just in Time manufacturing practices using logistic regression models [24], Palma, Beja and Rodrigues (1999) modeledobservations of Lynx [18] and in a particularly contemporary application of Hu and Heisey (1991) who used logistic regression to predict the "cache value" of objects on the World Wide Web [6].Obviously, logistic regression is the domain of practitioners rather than statisticians.That is, less than 1% of the large number of articles on logistic regression appear in the statistical reviews [23].

These studies clearly show that the application of logistic regression like other statistical methods covers a wide range of research area by statisticians.

This dissertation is structured in three essential chapters arranged as follows:

The first chapter introduces the basic concepts of Bayesian inference as well as the notations and the appropriate terminology .

second chapter presents the logistic regression model. This chapter will be defined in three parts:

- ▶ The first part covers the fundamental concepts of the logistic regression in binary case.
- ▶ The second part concerns the generalized the logistic regression model to the polytomous model.
- ▶ The third part will introduces the Bayesian approach for the estimation of parameters logistic models, by laying the theoretical bases which lead to MCMC methods.

end with an application followed with conclusion where we will cite some research perspectives.

Chapter 1

The Bayesian tools

Introduction

The Bayesian analysis,created by Pierre-Simon de Laplace and Thomas Bayes (1774), has for first step the study of situation and identify an uncertainly carried on en unknown parametere θ then to qunantify this one throughth a probalistic partition by using the element of calculation.

The Bayesian approach consists in treating the unknown parameter θ as a randam variable, in association with it ciating a probability distribution on the space Θ called prior distribution, denoted $\pi(\theta)$, this distribution reflects the knowledge prior sence of the parameter θ.
This prior distribution is updated by extracting information contained in the observations of X , to obtain another distribution, master of Bayesian, called posterior distribution.

1.1 Bayes' theorem

Theorem 1.1.1 *Let $(\Omega, \mathcal{A}, \mathcal{P}_0)$ be a probability space and A,B two events such that $P[B] \neq 0$. The theory of Bayes gives the expression defined by:*

$$P[A|B] = \frac{P[A \bigcap B]}{P[B]}$$

Bayes' theorem is a discounting principle, because it describes the updating of the observation from A , from $P[A]$ to $P[A|B]$ once B is observed.

Thomas Bayes (1764) actually proved a continuous version of this result, namely, that given two random variables x and y, with conditional distribution $f(x|y)$ and marginal distribution $g(y)$, the conditional distribution of y given x is

$$g(y|x) = \frac{f(x|y)g(y)}{\int f(x|y)g(y)dy}$$

.

1.2 Presentation of the Bayesian model

Let us define the spaces involved in a Bayesian model:

- **Observation space**
  Noted $\mathcal{X}$, it represents all the results following a study of a phenomenon.

- **Actions space**
  Noted $\mathcal{A}$, it represents all the actions or decisions to be taken after obtaining information.

- **Space of states of nature**
  Noted Θ, it represents the space of unknown parameters θ .

Let us define the distribution of probabilities involved in the Bayesian analysis:

1.2.1 The prior distribution

The prior distribution which is denoted by $\pi(\theta)$ is a probability distribution modeling the information available on the parameter of interest θ.

1.2.2 The Likelihood

It is the distribution of observations or even the conditional distribution of$X$ given θ , denoted by $f(x|\theta)$, it is given by:

$$f(x|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

.

1.2.3 Joint distribution of the couple (θ, x)

Usually denoted by $f(\theta, x)$, its formula is given by:

$$f(\theta, x) = f(x|\theta)\pi(\theta)$$

.

1.2.4 Marginal distribution

We denote it by $f(x)$, it is calculated as follows:

$$f(x) = \int_\Theta f(x|\theta)\pi(\theta)d\theta = \int_\Theta f(\theta, x)d\theta$$

1.2.5 The posterior distribution

Its density is given by:

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

We can use the notion of proportionality, that is to say:

$$\pi(x|\theta) \propto f(x|\theta)\pi(\theta)$$

.

1.3 The basic decision and Bayesian approach

Definition 1.3.1 *A Bayesian statistical model is made of a parametric ,the likelihood distribution $f(x|\theta)$, and prior distribution on the parameters,$\pi(\theta)$.*

In practice, statistical inference leads to a final decision taken by the decision maker and it is important to be able to compare the different decisions by means of an evaluation criterion , which will appear as a loss function.

Definition 1.3.2 *A loss function is any measurable function* $\mathcal{L} : \Theta \times \mathcal{D} \longrightarrow [0, +\infty[$.

This loss function is supposed to evaluate the penalty (or error) $\mathcal{L}(\theta, d)$ associated with the decision d when the parameter takes the value θ. However, in practice, this function is often difficult to determine, for example the difficulty is present when the spaces Θ and $\mathcal{D}$ are large (infinite dimensions), it is then impossible to determine each action $a = \delta(x)$ for each value of θ .

Definition 1.3.3 *Let* $\mathcal{L}(\theta, \delta)$ *be a loss function and let* π *be a prior distribution for* θ *. We call bayes estimator, the decision rule* $\delta_\pi(x)$ *which satisfies:*

$$\delta^\pi(x) = arg \min_{\delta \in \mathcal{D}} E^\pi[\mathcal{L}(\theta, \delta)|x]$$

This estimator will be determined analytically or numerically depending on the complexity of the loss function $\mathcal{L}$ and the posterior distribution $\pi(\theta|x)$.

It has the great advantage of not depending on a loss function, and is useful for theoretical approaches.

The interest of the decision-making approach is to find the best possible evaluation for the function of θ, which can lead to zero loss at best when this parameter is known. In otherwise, the cost function would lose its continuity, which could prevent the choice of a decision-making procedure.

To obtain a comparison criterion from a loss function, we consider the average of it, this is called frequentist risk.

Definition 1.3.4 *We call frequentist risk the average loss of a decision rule denoted* $R(\theta, \delta(x))$ *and defined through :*

$$R(\theta, \delta(x)) = E_\theta[\mathcal{L}(\theta, \delta(x))] = \int_{\mathcal{X}} \mathcal{L}(\theta, \delta(x)) f(x|\theta) dx$$

The frequentist paradigm relies on this criterion to compare estimators and, if possible, to select the best estimator, the reasoning being that estimators are evaluated on their long-run performance for all possible values of the parameter θ. Notice, however, that there are several difficulties associated with this approach.

The Bayesian approach defines another risk that integrates the parameter space to remedy to the difficulties present in the first approach.

Definition 1.3.5 *We call a posterior risk the average of the loss compared to the posterior distribution, commonly denoted by* $\rho(\pi, \delta(x))$ *, it is defined by:*

$$\rho(\pi, \delta(x)) = E^{\pi(.|x)}[\mathcal{L}(\theta, \delta(x))] = \int_{\Theta} \mathcal{L}(\theta, \delta(x))\pi(\theta|x)d\theta$$

It is also possible to define the integrated risk by definig a prior distribution π*.*

Definition 1.3.6 *We define the integrated risk as being the frequentist risk averaged over the value of* θ *according to its prior distribution, it is denoted by* $r(\pi, \delta(x))$ *:*

$$r(\pi, \delta(x)) = E_{\pi}[R(\theta, \delta(x))] = \int_{\Theta}\int_{\mathcal{X}} \mathcal{L}(\theta, \delta(x))f(x|\theta)\pi(\theta)dxd\theta$$

The interest of this risk is that it associates a real number with each estimator, which allows a direct comparison between these estimators.
The value $r(\pi, \delta^{\pi}) = \inf_{\delta \in \mathcal{D}} r(\pi, \delta) < \infty$ is called Bayesian risk.

Another formulation of the Bayesian estimator is proposed in the following.

Definition 1.3.7 *A Bayes estimator associated with a prior distribution* π *and a loss function* $\mathcal{L}$ *is a estimator* δ^{π} *verifying, for each* $x \in \mathcal{X}$ *:*

$$\delta^{\pi}(x) = arg\min_{\delta \in \mathcal{D}} \rho(\pi, \delta(x))$$

- ⋆ Properties of bayes estimators
- Bayes estimators are eligible.
- Bayes estimators are biased.
  Under certain regularity assumptions most often satisfied in practice, we have the two properties:

- The Bayes estimators are convergent in probability (when the sample size $n \longrightarrow +\infty$).

- The posterior distribution can be asymptotically approximate by a normal distribution(when the sample size $n \longrightarrow +\infty$)

$$\mathcal{N}(E(\theta|x), Var(\theta|x))$$

where $Var(\theta|x) = E[(\theta - E(\theta|x))^2|x]$ is the posterior variance of θ.

Definition 1.3.8 *We call minimax risk associated with the cost function $\mathcal{L}$, the value $\overline{R}$:*

$$\overline{R} = \inf_{\delta \in \mathcal{D}} \sup_{\theta \in \Theta} R(\theta, \delta(x))$$

The minimaxity criterion aims to minimize the average cost in the least favorable case, that is a kind of insurance against the worst.

Proposition 1.3.1 *Bayes risk is always smaller than minimax risk*

$$\underline{R} = \sup r(\pi, \delta^\pi) = \sup_{\pi} \inf_{\delta \in \mathcal{D}} r(\pi, \delta) \leq \inf_{\delta \in \mathcal{D}} \sup_{\theta \in \theta} R(\theta, \delta(x)) = \overline{R}$$

$\underline{R}$ is called the maximin risk, and $\overline{R}$ is the minimax risk.

Definition 1.3.9 *An estimator δ_0 is said to be inadmissible if there exists an estimator δ_1 such that for all $\theta \in \Theta$:*

$$R(\theta, \delta_1) \leq R(\theta, \delta_0)$$

In other words, δ_0 is admissible if there is no estimator δ such that:

$$R(\theta, \delta) \leq R(\theta, \delta_0)$$

Proposition 1.3.2 *If the Bayes estimator associated with a prior distribution π is unique, it is admissible.*

Proposition 1.3.3 *δ_0 is admissible and of constant risk, then δ_0 is the unique minimax estimator.*

1.3.1 Usual loss functions

1.The quadratic loss

Definition 1.3.10 *The quadratic loss function is the function defined by:*

$$\mathcal{L}(\theta, \delta(x)) = (\theta - \delta(x))^2$$

Proposition 1.3.4 *The Bayesian estimator $\delta^\pi(x)$ associated with the prior distribution π and the quadratic loss function is the mean of the posterior distribution of θ :*

$$\delta\pi(x) = E^\pi[\theta|x]$$

Corollary 1.3.1 *The Bayes estimator δ^π associated with π and with the weighted quadratic loss function $\mathcal{L}(\theta, \delta) = \omega(\theta)(\theta - \delta)^2$, where $\omega(\theta)$ is a non negative function, is*

$$\delta^\pi(x) = \frac{E[\omega(\theta)\theta|x]}{E[\omega(\theta)|x]}$$

2.The absolute loss function

Definition 1.3.11 *The absolute loss function is defined as follows:*

$$\mathcal{L}(\theta, \delta(x)) = |\theta - \delta(x)|$$

Proposition 1.3.5 *The Bayesian estimator associated with the prior distribution $\pi$and with the absolute loss function is the median of the posterior distribution $\pi(\theta|x)$.*

In a more general case, this function is weighted, which gives the linear cost function by pieces:

$$\mathcal{L}(\theta, \delta(x)) = \begin{cases} k_2(\theta - \delta(x)) & \theta > \delta(x) \\ k_1(\delta(x) - \theta) & \theta < \delta(x) \end{cases}$$

Proposition 1.3.6 *The Bayesian estimator associated with the prior distribution π and the linear cost perpiece is the fractile of order $\frac{k_2}{k_1+k_2}$ of the posterior distribution $\pi(\theta|x)$.*

3.The $0-1$ loss function

This loss is used for hypothesis testing, it is an example of non-quantitative loss.

Definition 1.3.12 *The loss function $\mathcal{L}$ is defined by:*

$$\mathcal{L}(\theta, \delta(x)) = \begin{cases} 0 & : (\theta - \delta(x)) < \varepsilon \\ 1 & : (\theta - \delta(x)) > \varepsilon \end{cases}$$

where ε is very small.

Proposition 1.3.7 *The Bayesian estimator associated with the prior distribution π and the $0-1$ cost function is the mode of the posterior distribution $\pi(\theta|x)$.*

4.The Linex loss function

Definition 1.3.13 *We define the Linex loss function as follows:*

$$L(\theta, \delta(x)) = expc(\delta(x) - \theta) - c(\delta(x) - \theta) - 1$$

. with $c \in \mathbb{R}$.

Proposition 1.3.8 *The Bayesian estimator associated with the prior distribution π and the loss function Linex is:*

$$\delta^{\pi}(x) = -\frac{1}{c} \ln[E_{\theta} \exp(-c\theta)]$$

where E_{θ} is the posterior expectation.

1.3.2 The choice of prior distribution

In practice, the prior information is generally insufficient, which makes the choice of the distribution priori difficult and not precise. Due to lack of resources or time, the researcher can not construct an exact prior, he must then rely on partial information on the data model.
Most often, it is then necessary to make a arbitrary choice of the prior distribution,in particular, the systematic use of parametrized distributions (like

the normal, gamma, beta,...,etc) and the further reduction to conjugate distributions (defined below) can not be justified at all times.

In the absence of prior information, we will introduce the notion of non-informative prior distribution which makes it possible to remain in a Bayesian framework, even when there is no information prior, their choices are motivated by prior distribution which give a posterior corresponding to frequentist estimates. A prior distribution which have an attractive interpretation or prior distribution allowing an analytical form for posterior distribution and among the most popular techniques in the construction of its laws one can quote: the distribution of Laplace, Jeffrey, ... etc.

Partially informative approach

1.Maximum entropy

If some characteristics of the prior distribution (moments, quantiles,...,etc.) are known, assuming that they can be written as prior expectations,

$$E_\pi[g_k(\theta)] = \omega_k$$

$(k = 1, ..., K)$, a way to select a prior π satisfying these constraints is the maximum entropy method, developed in Jaynes (1980, 1983). In a finite setting, the entropy is defined as

$$\varepsilon(\pi) = -\sum_i \pi(\theta_i) \log(\pi(\theta_i))$$

This quantity has been introduced by Shannon (1948) as a measure of uncertainty in information theory and signal processing.
The maximization of entropy under constraints makes it possible to search for the least informational distribution, the principle of the method is to calculate $arg\min \varepsilon(\pi)$, with constraint $E_\pi[g_k(\theta)] = \omega_k$. The solution of this gives:

$$\pi^* \propto \frac{\exp\{\sum_{i=1}^n \lambda_k g_k(\theta_i)\}}{\sum_j \exp\{\sum_{i=1}^n \lambda_k g_k(\theta_j)\}}$$

where the λ_k are the Lagrange multipliers.
The extension to the continuous case is quite delicate, it becomes more complicated to use the maximum method entropy. A first difficulty being that there is no formal definition of entropy, Jaynes (1968) then proposed the

following formula:

$$\varepsilon(\pi) = -E[\ln \frac{\pi(\theta)}{\pi_0(\theta)}] = \int_\theta \ln \frac{\pi(\theta)}{\pi_0(\theta)} d\theta$$

In this case, the maximum entropy distribution is given by the density

$$\pi^* \propto \frac{\exp\{\sum_{i=1}^n \lambda_k g_k(\theta)\pi_0(\theta)\}}{\int \exp\{\sum_{i=1}^n \lambda_k g_k(\eta)\pi_0 d\eta\}}$$

2.Conjugate priors distributions

The conjugate prior approach, which originated in Raiffa and Schlaifer (1961),it describes a particular link that connects the model that follows the data and the prior distrubutions of the unknown parameters . It's can be considered as a point of starting point for the development of prior distributions based on limited prior information.

Definition 1.3.14 *A family $\mathcal{F}$ of probability distributions on Θ is said to be conjugate with respect to the likelihood function $f(x|\theta)$ if for all $\pi \in \mathcal{F}$, the posterior distribution $\pi(\theta|x)$ also belongs to $\mathcal{F}$*

The conjugate prior distributions are generally associated with a particular type of sampling distribution that always allows for their derivation wich is a characteristic of conjugate priors, as we will see below. These distributions constitute what are called exponential families, studied in detail in Brown (1986).

Definition 1.3.15 *Let μ be a σ finite measure on $\mathcal{X}$, and let Θ be the parameter space. Let C and h be functions, respectively, from $\mathcal{X}$ and Θ to $\mathbb{R}_+$, and let R and T be functions from Θ and X to $\mathbb{R}^k$.*

$$f(x|\theta) = C(\theta)h(x)\exp R(\theta)T(x)$$

is called an exponential family of dimension k. In the particular case when $\Theta \subset \mathbb{R}^k, \mathcal{X} \subset \mathbb{R}^k$ and

$$f(x|\theta) = C(\theta)h(x)\exp\{\theta.x\}$$

the family is said to be natural.

Proposition 1.3.9 *A conjugate family for $f(x|\theta)$ is given by*

$$\pi(\theta|\mu,\lambda) = K(\mu,\lambda)\exp\{\theta\mu - \lambda\psi(\theta)\}$$

where $K(\mu,\lambda)$ is the normalizing constant of the density. The corresponding posterior distribution is $\pi(\theta|\mu+x,\lambda+1)$

$f(x\|\theta)$	$\pi(\theta)$	$\pi(\theta\|x)$
Normal $\mathcal{N}(\theta,\sigma^2)$	Normal $\mathcal{N}(\mu,\tau^2)$	Normal $\mathcal{N}(\varrho(\sigma^2\mu+\tau^2),\varrho\sigma^2\tau^2)$
Poisson $\mathcal{P}(\theta)$	Gamma $\mathcal{G}(\alpha,\beta)$	$\mathcal{G}(\alpha+x,\beta+1)$
Gamma $\mathcal{G}(\alpha,\beta)$	Gamma $\mathcal{G}(\nu,\theta)$	$\mathcal{G}(\alpha+\nu,\beta+x)$
Binomial $\mathcal{B}(n,\theta)$	Beta $\mathcal{B}e(\alpha,\beta)$	$\mathcal{B}e(\alpha+x,\beta+n-x)$
Negative Binomial $\mathcal{N}eg(m,\theta)$	Beta $\mathcal{B}e(\alpha,\beta)$	$\mathcal{B}e(\alpha+m,\beta+x)$
Normal $\mathcal{N}(\mu,\frac{1}{\theta})$	Gamma $\mathcal{G}(\alpha,\beta)$	$\mathcal{G}(\alpha+\frac{1}{2},\frac{(\mu-x)^2}{2})$

Table 1.1: Natural conjugate priors for some common exponential families

1.3.3 Noninformative prior distributions

The bayesian approach is also applied even when we have no prior information which adds a new point in the critiques against this approach.But this involves a particular type of particular prior distributions must be derived from the sample distribution, since this is the only available information. For obvious reasons, they are called noninformative priors.
We describe below some of the most important techniques in the derivation of noninformative priors.

Definition 1.3.16 *A noninformative distribution is a distribution which does not carry any information on the parameters to be estimate, it does not give more weight to a particular value of the parameter.*

Definition 1.3.17 *A prior distribution $\pi(\theta)$ is said to be improper if it is a σ-finite measure and which verifies:*

$$\int_\Theta \pi(\theta)d\theta = \infty$$

▶ Laplace prior's distribution

Historically, Laplace was the first to use noninformative techniques since, although he had no information,he used a uniform prior. His reasoning, later

called the Principle of Insufficient Reason, was based on the equiprobability of elementary events and therefore appeared to be sound enough.
Suppose Θ is a set of size k then:

$$\pi(\theta) = \frac{1}{k}$$

The resulting distributions are improper when the parameter space is not compact.

The Jeffreys' prior distribution

The Jeffreys noninformative prior distributions are based on Fisher information, given by:

$$I(\theta) = E_\theta \left[\frac{\partial \log f(X|\theta)^2}{\partial \theta^2} \right]$$

in the one-dimensional case. Under some regularity assumptions, this information can also be written as

$$I(\theta) = -E_\theta \left[\frac{\partial \log f(X|\theta)}{\partial^2 \theta} \right]^2$$

The Jeffreys prior distribution is

$$\pi^*(\theta) \propto I^{\frac{1}{2}}(\theta)$$

when π^* is proper When θ is a multidimensional parameter, the Fisher information matrix is defined as
For $\theta \in \mathbb{R}^k$, $I(\theta)$ has the following elements,

$$I_{ij}(\theta) = -E_\theta \left[\frac{\partial^2 \log f(x|\theta)}{\partial \theta_i \partial \theta_j} \right] (i, j = 1, \cdots, k)$$

and the Jeffreys noninformative prior is then defined by

$$\pi^*(\theta) \propto \left[\det I^{\frac{1}{2}}(\theta) \right]$$

1.3.4 Credibility interval

The Bayesian analogue of a classical confidence interval is called confidence region or credibility interval defined as follows:

Definition 1.3.18 *An α-credible region of $100(1-\alpha)\%$ for $\theta (0 < \alpha < 1)$ is a subset c of Θ such that*

$$1 - \alpha \leq P[\theta \in c|x] = \begin{cases} \int_C \pi(\theta|x) d\theta \\ \sum_{\theta \in C} \pi(\theta|x) \end{cases}$$

There exists an infinity of α-credible regions for θ , to choose the right set, we are interested to the region which has the minimum volume, the volume

V being defined by:

$$V(C) = \int_C d\nu(\theta)$$

if $\pi(\theta|x)$ is absolutely continuous with respect to a measure ν .

For this, we will introduce the notion of HPD region ie "Highest Posterior Density".

Definition 1.3.19 *An HPD region α-credible for θ "Highest Posterior Density" , the subset C_x of Θ of the form $C_x = \{\theta \in \Theta : \pi(\theta|x) \geq k(\alpha)\}$*

where $k(\alpha)$ is the largest bound such that

$$P^{\pi}[\theta \in C_x|x] \geq 1 - \alpha$$

.

1.3.5 Bayesian approach to testing

Consider a statistical model $f(x|\theta)$ with $\theta \in \Theta$. Given a subset of interest of Θ, Θ_0, which sometimes is a single point Θ_0, the question to be answered is whether the true value of the parameter θ belongs to Θ_0, i.e
We want to test:
$H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$, where $\Theta_0 \cup \Theta_1 = \Theta$ and $\Theta_0 \bigcap \Theta_1 = \varnothing$.

By definition, Bayesian decisions are those which minimize the posterior loss $\updownarrow(\pi, \delta|x)$.

We have two possible decisions:
d_0 : we do not reject H_0
d_1 :we reject H_0

In practice, we accept the hypothesis H_0 or H_1 as soon as its posterior probability $\alpha_0 = P(H_0|x)$ or $\alpha_1 = P(H_1|x)$.

1.3.6 The bayes factor

The bayes factor is ratio of the posterior probabilities of the null and alternative hypotheses $\frac{\alpha_0}{\alpha_0}$ (posterior odds ratio) to the prior probabilities of these same hypotheses $\frac{\pi_0}{\pi_1}$ (prior odds ratio):

$$B = \frac{posterior \;\; odds \;\; ratio}{priori \;\; odds \;\; ratio} = \frac{\frac{\alpha_0}{\alpha_1}}{\frac{\pi_0}{\pi_1}} = \frac{\alpha_0 \pi_1}{\alpha_1 \pi_0}$$

This ratio evaluates the modification of the likelihood under the set Θ_0 compared to that under the set Θ_1 , due to observation.

Particular case

If $\Theta_0 = \theta_0$ and $\Theta_1 = \theta_1$, the bayes factor is only the classical likelihood ratio which is defined by:

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

Conclusion

Without context, the most critical point of Bayesian analysis is the choice of the prior distribution, it is therefore most often necessary to make a (partially) arbitrary choice of the a priori distribution. which can have a considerable impact on the resulting inference.

Most often we do not have enough prior information about the parameter unknown θ to construct the prior distribution. In practice, we have recourse to usual distributions (gaussian distributions, gamma distribution,...,etc.) or to conjugate distribution.

In the absence of prior information, we will introduce the notion of noninformative a priori distribution which makes it possible to remain in a Bayesian framework, their choices are motivated by prior which give posterior corresponding to frequentist estimates, prior which have an attractive interpretation or prior allowing an analytical form for posterior and among the most popular techniques in the construction of its distribution one can quote: Laplace's distribution, Jeffrey's distribution, ... etc.

Chapter 2

The logistic regression

introduction

The term "regression" has a curious origin, it goes back to the study of the physiologist and anthropologist Francis Galton (towards the end of the 19th century) on the relationship between the size of the parents and that of the children. The results obtained led him to his theory known as "regression towards mediocrity".

Regression models have become an integral component of any statistical analysis aimed at writing the relationship between a variable to be explained and one or more explanatory variables, and are widely used in many disciplines and applications. Several models can be distinguished, such as linear, generalized linear, logistic and many others.

In this chapter we will introduce the logistic model which has become an integral component in data analysis, a modelling technique which, in its most popular version, aims to predict and explain the values of a qualifying variable Y which is more often binary from a collection of variables X.

2.1 Binary logistic regression

Definition 2.1.1 *Let Y be a variable with values in $\{0,1\}$ and $X = (x_1, \cdots, x_p)$ designate p explanatory variables. The logistic model proposes a modelling of the distribution of $Y|X = x$ by a Bernoulli distribution of parameter $p_\beta(x) = P_\beta(Y = 1|X = x)$ such that :*

$$p_\beta(x) = \frac{\exp(\beta_0 + x_1\beta_1 + x_2\beta_2 + ... + x_p\beta_p)}{1 + \exp(\beta_0 + x_1\beta_1 + x_2\beta_2 + ... + x_p\beta_p)} \tag{2.1}$$

$$= \frac{\exp(X^t\beta)}{1 + \exp(X^t\beta)} \tag{2.2}$$

This probability is also called a logistic function.

(2,1) can also be written by

$$log\frac{p_\beta(x)}{1 - p_\beta(x)} = \beta_0 + x_1\beta_1 + x_2\beta_2 + \cdots + x_p\beta_p \tag{2.3}$$

or even

$$logitp_\beta(x) = \beta_0 + x_1\beta_1 + x_2\beta_2 + \cdots + x_p\beta_p \tag{2.4}$$

where $\beta_0, \beta_0, \cdots, \beta_p$ are the unknown real parameters to be estimated. *logit* denoting (2,3)and (2,4) the bijective and derivable function in $]0, 1[$ to $\mathbb{R}$:

$$p_\beta(x) \mapsto \frac{p_\beta(x)}{1 - p_\beta(x)}$$

.
The ratio between the probability of success $p_\beta(x)$ and the probability of failure $(1 - p_\beta(x))$ is called odds "odds ratio" .

Remark 2.1.1

The logistic function is bounded between 0 *and* 1.*It is called the link function*

2.2 Interpretation of β coefficients

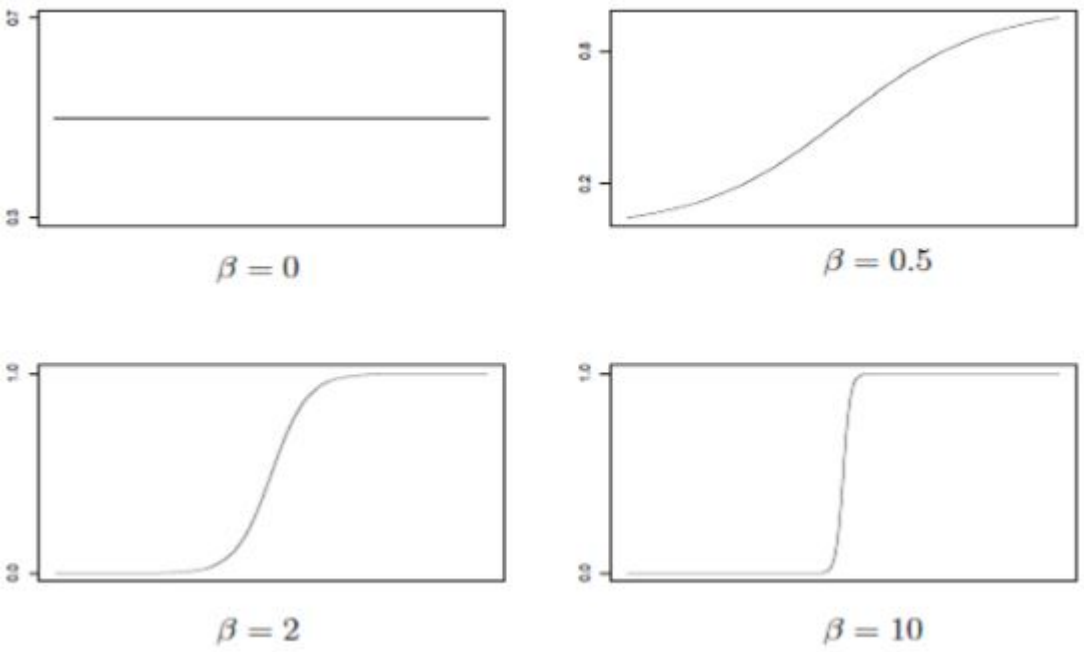


Figure 2.1: Representation of the logit function for a different values of β coefficients

We notice that:

- If β is small, we have a wide range of values of x for which $\pi_\beta(x)$ is around 0.5.
- $\pi_\beta(x) = 0.5$ in the extreme case $\beta = 0$.
- As β increases, the area where $\pi_\beta(x)$ is close to 0.5 decreases. $\pi_\beta(x)$ becomes close to 0 or 1 for a large number of values of x. This can be interpreted as follows :the larger β is, the better we discriminate.

This interpretation depends on the values of x . This is why the interpretation of β coefficients is effected in terms of odds ratios.

2.3 Estimation of the parameters β

Now the model is adequately specified, we seek to estimate the effect of the β coefficients. To do this, we seek the maximum likelihood. Since maximizing likelihood becomes maximizing log-likelihood, we have for $j = \{1, 2, ...p\}$

$$\widehat{\beta_p} = \arg\max \ell(\beta|Y, X) = \arg\max_{\beta_p} \sum_{i=1}^{n} \ln[f(Y_i|X_i, \beta)]$$

ℓ represents the log-likelihood and$f(Y_i|X_i, \beta) = \pi(X)$ the density function associated with Y_i. We have

$$f(Y_i|X_i, \beta) = \left(\frac{\exp(X^t\beta)}{1+\exp(X^t\beta)}\right)^{y_i} \left(1 - \frac{\exp(X^t\beta)}{1+\exp(X^T\beta)}\right)^{1-y_i} = \frac{\exp(X^t\beta y_i)}{1+\exp(X^t\beta y_i)}$$

for $i = \{1, 2, ...n\}$Thus, the log-likelihood is written as

$$\ell(\beta|Y, X) = \sum_{i=1}^{n} \ln\left[\frac{\exp(X^t\beta y_i)}{1+\exp(X^t\beta y_i)}\right] = \sum_{i=1}^{n} \left[X_i^t\beta y_i - \ln\left(1 + e^{X_i^T\beta y_i}\right)\right]$$

Let us now derive the log-likelihood from the coefficients β. We have

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{n} X_i \left[y_i - \frac{\exp(X^t\beta y_i)}{1+\exp(X^t\beta y_i)}\right]$$

It then requires the use of numerical optimization methods (iterative) in particular the algorithm of Newton-Raphson and Mak (1993) which allow to solve this type of problem.

2.3.1 Asymptotic properties of the estimator $\widehat{\beta}$

To present the asymptotic properties of the maximum likelihood estimator $\widehat{\beta}$ of the parameter β in the logistic regression model, we assume that the following assumptions are verified cite:

1 H_1: The exogenous variables are uniformly bounded, i.e. $\exists M < \infty$: $|X| \leq M$

2 H_2: Let λ_{1n} and λ_{pn} be the minimum and maximum eigenvalues of the matrix X^tWX , respectively. Then there exists a constant $K < \infty$, such that $\frac{\lambda_{pn}}{\lambda_{1n}} \leqslant K$

Theorem 2.3.1 *(Existence and consistency) Under hypotheses H_1 and H_2 the maximum likelihood estimator noted $\widehat{\beta}$ of β almost surely exists when n tends to $+\infty$, and $\widehat{\beta}$ almost surly converges when n tends to $+\infty$ to the true value β_* if and only if n tends to $+\infty$, $\lambda_{1n} \to +\infty$.*

proof see [11]

Theorem 2.3.2 *(Asymptotic normality) Under the hypotheses H_1 and H_2 and if the estimator of likelihood $\widehat{\beta}$ converges asymptotically to β_*then $\sqrt{n}(\widehat{\beta}_n - \beta_*) \longrightarrow \mathcal{N}(0, \phi(\beta_*))$ when $n \longrightarrow +\infty$ where $\phi(\beta_*) = -E(\bigtriangledown^2 \ell(\beta_*, y))$ is is the Fisher information matrix.*

proof see [11]

2.3.2 The variance-covariance matrix

In this section we will use the asymptotic properties for the approximation of the estimates.To facilitate the calculations, we assume that $\beta = (\beta_0, \beta_1)$ (logistic regression with two modalities).
The variance of $\widehat{\beta}_0$ and $\widehat{\beta}_1$ is obtained by considering the second derivatives of the log-likelihood function, thus forming the Fisher information matrix.

$$\mathbb{I} = \begin{pmatrix} \sum_{i=1}^{n}\{p(x_i)(1-p(x_1))\} & \sum_{i=1}^{n}\{x_i p(x_i)(y_i - p(x_i)\} \\ \sum_{i=1}^{n}\{p(x_i)(1-p(x_i)\} & \sum_{i=1}^{n}\{x_i^2(y_i - p(x_i)\} \end{pmatrix}$$

The variance-covariance matrix $V(\widehat{\beta})$ is estimated by the inverse matrix of Fisher information.

2.3.3 Statistical tests

Likelihood ratio test

We de fine the likelihood ratio test as :

$$\begin{cases} \mathrm{H}_0 : \beta_0 = \beta_1 = \cdots = \beta_p = 0 \\ \mathrm{H}_1 : \exists j = 1, \cdots P \qquad\qquad \beta_j \neq 0 \end{cases}$$

This test is based on the following decision statistics:

$$ML = -2\log\left(\frac{\mathcal{L}_0(\beta_0)}{\mathcal{L}(\widehat{\beta})}\right) \sim \mathcal{X}_P^2$$

With: $\mathcal{L}_0(\beta_0, y)$ corresponds to the likelihood without the explanatory variables of the model.

$\mathcal{L}(\widehat{\beta}, y)$ corresponds to the likelihood with the explanatory variables of the model.

Decision rule

We compare the calculated value ML to the quantile of order $(1-\alpha)$ of the law of $\mathcal{X}^2$ at p degrees of freedom, denoted by $\mathcal{X}^2_{(1-\alpha,p)}$

- If $ML > X^2_{(1-\alpha,p)}$ then we reject H_0 and say that the model is globally good. Hence there is at least one significant explanatory variable of y.

- If $ML < X^2_{(1-\alpha,P)}$ then we accept H_0 and we will say that the model is bad. Hence there is no significant explanatory variable of y.

Individual Wald test

We are interested here in the contribution of each variable individually. By carrying out the tests defined by:

$$\begin{cases} \mathrm{H}_0 : \beta_j = 0 \\ \mathrm{H}_1 : \beta_j \neq 0 \end{cases} \quad (\forall j = 1, \cdots p)$$

Due to the asymptotic normality of the maximum likelihood estimator,this test is based on the statistic $\dfrac{\widehat{\beta}_j}{\sigma_{\widehat{\beta}_j}}$ which approximately follows a reduced central normal distribution such that:

$$W = \frac{\widehat{\beta}_j}{\sigma_{\widehat{\beta}_j}} \sim \mathcal{N}(0,1)\ , j = 1, \cdots, p$$

Decision rule

We compare the calculated value W to the quantile of order $(1-\frac{\alpha}{2})$ of the reduced central normal distribution denoted by $Z_{1-\frac{\alpha}{2}}$.

- $W > Z_{1-\frac{\alpha}{2}}$then we reject H_0, so the variable X_j is significant of y.

- $W < Z_{1-\frac{\alpha}{2}}$ then we accept H_0, so the variable X_j isn't significant of y.

2.3.4 Confidence interval

An important complement to the logistic model significance tests is the computation and interpretation of confidence intervals for the parameters β_j, $j = 0, \cdots, p$ at the significance level α , (or at the level of confidence $(1-\alpha)$, we consider the statistic:

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

From the symmetry of the normal distribution, we consider a symmetrical interval such that:

$$\mathrm{P}\left[-Z_{1-\frac{\alpha}{2}} < \frac{\widehat{\beta}_j - \beta_j}{\widehat{\sigma}_{\widehat{\beta}_j}} < Z_{1-\frac{\alpha}{2}}\right] = 1-\alpha$$

$$\Rightarrow \mathrm{P}\left[\widehat{\beta}_j - Z_{1-\frac{\alpha}{2}}\widehat{\sigma}_{\widehat{\beta}_j} < \beta_j < \widehat{\beta}_j + Z_{1-\frac{\alpha}{2}}\widehat{\sigma}_{\widehat{\beta}_j}\right] = 1-\alpha$$

We obtain the confidence interval of β_j, $j = 0, \cdots, p$ at the sizeα, given by:

$$\mathrm{IC}_{B_j} = \left[\widehat{\beta}_j - Z_{1-\frac{\alpha}{2}}\widehat{\sigma}_{\widehat{\beta}_j}, \widehat{\beta}_j + Z_{1-\frac{\alpha}{2}}\widehat{\sigma}_{\widehat{\beta}_j}\right]$$

where $Z_{1-\frac{\alpha}{2}}$ is the quantile of order $1-\frac{\alpha}{2}$ of the reduced central normal distribution.

Remark 2.3.1 *1- Logistic regression requires large sample sizes to be able to achieve a good level of stability.*

2- The categories to which the independent variables belong must be mutualy exclusive, because it is a dichotomous variable.

3- Chek the correlations between the predictors before proceeding to the development of the model. When certain predictors are strongly correlated

with each other, it is preferable to eliminate them a few since they are redundant variables.

4- Logistic regression has the limitation of assuming only unrelated responses.

2.4 Polytomous logistic regression

In the previous section we introduced the logistic regression model in the univariate context.As in the case of linear regression,the strenght of modeling is techniques lies in its ability to model many variables,somes of wich may be on different measurement scales.This be refered to as *polytomous case"*[12].

- The modalities of Y are ordered: there is a natural hierarchy between them. For example in biostatistics, it can be a diagonasis of the state of health (very good, good, average, poor health), on the stage of development of a disease, or on the size or nature a tumor (absent, benign, or malignant tumor). We speak in this case of model ordered polytomous.

- There is no order relation on the modalities of Y, the variable to be explained is purely nominal: agreement for a loan (yes, no, examination of the file). We speak in this case of model nominal polytomous or multinomial polytomous model.
  As in the binary case, here we are trying to model the law of $Y|X = x$.

2.4.1 Multinomial logistic regression

Multinomial logistic regression is used to model nominal outcome variables, in which the log odds of the outcomes are modeled as a linear combination of the predictor variables.

Multinomial logistic regression is a generalization of binary logistic regression. Here the dependent variable Y admits more than 2(unordered) modalities. We want to explain a response variable Y to k modalities $y_1, \cdots y_k$ as a function of p explanatory variables $X_1, X_2, \cdots, X_p$.

2.4.2 modelization

We will model $K-1$ probability ratios ie. Take a modality as a reference (ex the last one), and express $K-1$ logit with compared to this reference (ex the "non-patients" to oppose to various categories of diseases).
The last probability, belonging to the K^{th} category, is deduced from the others:

$$\mathrm{p}_k(x_i) = 1 - \sum_{k=1}^{K-1} \mathrm{p}_k(x_i)$$

We write $K-1$ logit equations:

$$logit(\mathrm{p}_k(x_i)) = \log\left(\frac{\mathrm{p}_k(x_i)}{\mathrm{p}_K(x_i)}\right) = \beta_{0k} + \beta_{1k}x_{i1} + \cdots + \beta_{pk}x_{ip}$$

$$i = \overline{1,n} \quad k = 1, \cdots, K-1$$

and we deduce the $K-1$ probabilities of assignment:

$$\mathrm{p}_k(x_i) = \frac{e^{logit(\mathrm{p}_k(x_i))}}{1 + \sum_{k=1}^{K-1} \mathrm{p}_k(x_i)} \; k = 1, \cdots, K-1$$

and the last one $\mathrm{p}_K(x_i) = 1 - \sum_{k=1}^{K-1} \mathrm{p}_k(x_i)$ and we have $\sum_{k=1}^{K} \mathrm{p}_k(x_i) = 1$

The assignment rule is:

$$Y = y_k \Leftrightarrow arg\max \mathrm{p}_k(x_i)$$

2.4.3 Parameter estimation

The parameter estimation method is that of maximum likelihood.The likelihood function is defined by:

$$\mathrm{L} = \prod_{i=1}^{n} \mathrm{p}_k(x_i)^{y_1} \times \cdots \times \mathrm{p}_k(x_i)^{y_k}$$

Hence the log-likelihood is given by:

$$\ell = \sum_{i=1}^{n} y_1 \ln \mathrm{p}_k(x_i) + \cdots + y_k \ln \mathrm{p}_k(x_i)$$

There are $(K-1)(p+1)$ parameters to be estimated. We can again rely on the NewtonRaphson method:

With $G = \begin{pmatrix} G \\ \vdots \\ G_{K-1} \end{pmatrix}$ is the gradient vector of dimension $(K-1)(p+1) \times 1$

G_k is of dimension $(p+1) \times 1$, for each case we have:

$$g_{k,j} = \sum_{j=1} x_i(y_k - \mathrm{p}_k(x_i))$$

The Hessian matrix, of dimension $(K-1)(p+1) \times (K-1)(p+1)$, which will be given by:

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

$H_{i,j}$ is of dimension $(p+1) \times (p+1)$, defined by:

$$H_{i,j} = \sum \mathrm{p}_i(x_i)[\delta_{ij} - \mathrm{p}_j(x_i)]XX^t$$

With $X = (1, X1, ...Xp(xi))$ and

$$\delta_{ij} = \begin{cases} 1 : i = j \\ 0 : i \neq j \end{cases}$$

2.5 Ordinal logistic regression

Ordinal logistic regression is used to model the relationship between a dependent variable Y which takes more than 2 ordered modalities and the independent variables.
Consider the response variable Y with k modalities and $x = (x_1, ..., x_P)^t$ the vector of explanatory variables (covariates).

2.5.1 Case of adjacent logits

Its principle is to calculate the logit of the passage from one category to another. Same idea as the multinomial model, except that the reference category changes at each step. We evaluate the passage from the modality (k) to $(k-1)$.

The $(k-1)$ logit equations are de fined by:

$$\begin{cases} logit_1(p(x)) = \ln(\frac{p_1(x)}{p_2(x)}) = \beta_{01} + \beta_{11}x_1 + \cdots + \beta_{p,1}x_p \\ \cdots \\ logit_{k-1}(p(x)) = \ln(\frac{p_{1-K}(x)}{p_k(x)}) = \beta_{0,K-1} + \beta_{1,K-1}x_1 + \cdots + \beta_{p,K-1}x_p \end{cases}$$

This writing can be seen as a reinterpretation of the multinomial model.

$$\begin{cases} \ln(\frac{p_2(x)}{p_1(x)}) = -logit_1(x) \\ \ln(\frac{p_3(x)}{p_1(x)}) = -logit_2(x) - logit_1(x) \\ \cdots \\ \ln(\frac{p_{1-K}(x)}{p_k(x)}) = logit_{k-1}(x) - \cdots - logit_2(x) - logit_1(x) \end{cases}$$

Remark 2.5.1 *We can use the results of the multinomial model to estimate the parameters. Significance assessments and tests are the same*

2.5.2 Case of cumulative odds-ratio

Let us see in this case what it is for the interpretation of the coefficients, for that we will model as follows:
The cumulative probability is de fi ned as follows:

$$\mathbb{P}(Y < k|X) = p_1 + \cdots + p_k$$

The cumulative logits are given by:

$$logit_k = \ln \frac{\mathbb{P}(Y \preceq k|X)}{\mathbb{P}(Y \succ k|X)} = \ln \frac{\mathbb{P}(Y \preceq k|X)}{1 - \mathbb{P}(Y \preceq k|X)} = \ln(\frac{p_1 + \cdots + p_k}{p_{k+1} + \cdots + p_K})$$

The $(K-1)$ logit equations are defined (first) as follows:

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

Let us re-introduce the hypothesis: the role of a variable does not depend on the level of Y.

$$logit_k = \beta_{0,k} + \beta_{1,k}x_1 + \cdots + \beta_{p,p}x_p$$

2.6 Bayesian logistic regression

In this section,we will apply the Bayesian approach to regression models.The difference between the two approachs is that the classical approach assumes that the parameters are unknow values to be estimated and the Bayesian approach assumes that the coefficients are no longer correctives but rather random variables following a certain know probabiliy distribution called a priori density distributions$\pi(\beta)$.

In the case of logistic regression, the introduction of this distribution rise to Bayesian logistic regression. Its principle is to update the information given by the observations and by introducing other information prior. The distrubtion resulting from this update is the posterior law $\pi(\beta|y)$.
In summary, Bayesian inference for logistic models follows the following steps:

- Write the likelihood function of the data.
- Introduce an prior distribution for the unknown parameters of the model.

- Find the posterior distribution of the parameters.

  In this part, we will develop the Bayesian approach to deal with regression.

2.7 Bayesian logistic model (the binary case)

The methodology describes Bayesian inference with an emphasis on three key components: the prior distribution, the likelihood function and the posterior distribution.

2.7.1 Presentation of the model

We have n observations $X = (X_1, X_2, \cdots, X_2)^t$ is an explanatory variable and Y a variable with value in $0, 1$ or $Y_i|X_i = x_i \sim \ beta(p(x_i))$ with Y_i is an endogenous random variable, $\beta = (\beta_0, \cdots \beta_p)^t$ are unknown regression parameters. The binary logistic model, which defined by:

$$logit(p_\beta(x_i)) = \log\left(\frac{p_\beta(x_i)}{1 - p_\beta(x_i)}\right) = x^t\beta \tag{2.5}$$

(2,5) It refers to the logistic probability model for a'success',that is,a certain event happening. where x is some given value of some predictor, Then the probability of a success is :

$$p_\beta(x_i) = P_\beta(Y = 1|X = x) = \frac{\exp\left(logit(p_\beta(x_i))\right)}{1 + (logit(p_\beta(x_i)))} \tag{2.6}$$

The posterior density is given by:

$$\pi(\beta|y) \propto f(y|\beta)\pi(\beta) \tag{2.7}$$

Recall that the likelihood function is defined by:

$$f(y|x, \beta) = \prod_{i=1}^{n} f(y_i|x_i, \beta) = \prod_{i=1}^{n} \frac{e^{X_i^t\beta}y_i}{1 + e^{X_i^t\beta}}$$

suppose that β follows an prior density of multivariate normal distribution

such that

$$\pi(\beta) = \frac{1}{2\pi\Sigma^{\frac{1}{2}}} e^{-\frac{1}{2}\beta^t \Sigma^{-1}} \tag{2.8}$$

where Σ is the covariance matrix with σ^2 on its diagonal. We suppose that the coefficients β are independent. Thus, the covariances of the coefficients β will be zero and $\Sigma = \sigma^2\mathbf{I}$ with $\mathbf{I}$ being the identity matrix.
Hence, formula(2,9) can also be written by :

$$\pi(\beta) = \prod_{i=1}^{n} \left(\frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} e^{-\frac{1}{2\sigma^2\beta_i^2}} \right) \tag{2.9}$$

Excluding all terms that do not depend on β, the formula (2,7) will be :

$$\pi(\beta|y) \propto \prod_{i=1}^{n} \left(\frac{e^{x_i^t \beta y_i}}{1 + e^{x_i^t \beta}} \right) e^{-\frac{1}{2\sigma^2}\beta^t\beta} \tag{2.10}$$

The expression (2,10) does not have an explicit form because it is a complex function of the parameters. In this situation, simulation methods are often necessary to estimate the posterior distribution for each of the parameters of the model, such as the MCMC methods that will be described in the next section.

2.7.2 MCMC methods

Markov Chain Monte Carlo methods appeared in 1950 for statistical physics and have almost unlimited applications. Although their performance varies widely depending on the complexity of the problem, their principle is to generate an ergodic Markov chain that converges to its stationary distribution which is exactly the posterior distribution.

Basics notions of MCMC methods

1 **Markov chain**

Markov chains were introduced by Andreï iAndreïevitch Markov (1856-1922) who was a student of Chebyshev. While reading the novel Eugene Onegin by the Russian Alexander Pushkin, he examined the first $20,000$ letters of the text and considered them as $20,000$ random experiments whose result is a consonant or a vowel.He then showed that

the occurrence of a vowel immediately after a consonant is a Markov process. In other words, if we draw a letter at random, it is a vowel or a consonant according to the probabilities given by the statistics , but the status of a letter depends on the status of the previous one (non-independent events).

Definition 2.7.1 *Consider a stochastic process $(X_t)_t$ with discrete time $\mathbb{T}$ and discrete state space E.*
Markov chain $(X_t)_t$ is said to be cha^ıne de Markov if

$$\begin{aligned} P(X_{n+1} = j | X_0 = i_0, \cdots, X_{n-1} = i_{n-1}, X_{n-1} = i_n) \\ = P(X_{n+1} = j | X_n = i) = P_{ij}(n) \end{aligned}$$

$i_0, \cdots, i_{n-1}, i, j \in E, n \in \mathbb{T}$

This means that the future evolution of the process depends only on its most recent past, i.e., that the present contains all the past information of the process. In other words, the process tends to forget its past; this is called a memoryless process.$P_{ij}(n)$ is the probability that the emarrant system from the initial state i_0 reaches the state j after $n+1$ transitions.

The following definitions describe the properties necessary for the convergence of Markov chains produced by MCMC algorithms.

1.1 Irreducibility

Definition 2.7.2 *A Markov chain is said to be irrecoverable if it has only one equivalence class, i.e. all its states communicate with each other*

1.2 Recurrence

Definition 2.7.3 *A state i is said to be recurrent if starting from i one almost surely returns to it in in finite time:*
$P(T < +\infty | X_0 = i) = 1$, *where* $\mathbb{T} = \inf\{n \geq 1, X_n = i\}$ *is the return time to i [7].*

1.3 transitivity

Definition 2.7.4 *A state is said to be transient i if it is with a positive probability that one leaves it never to return to it i.e*
$P(\mathbb{T}_i = +\infty | X_0 = i) = 1$

A state is said to be recurrent positive when the average time of return to i is finite:
$\mu_i = E(\mathbb{T}_i | X_0 = i) < +\infty$ *In the case where*$\mu_i = E(\mathbb{T}_i | X_0 = i) = +\infty$, *i is said to be recurrent zero.*

1.4 **Periodicity**

Definition 2.7.5 *A state i is said to be aperiodic if* $d(i) = 1$ *with* $d(i) = PGCDn \geqslant 1, P(n)_{ii} > 0$ *representing the period of state i. A Markov chain is said to be aperiodic if all its states are aperiodic.*

1.5 **Ergodicity** An irrecoverable, positively recurrent and aperiodic Markov chain is said to be ergodic.[7]

2.Convergence diagnostics

We conclude our presentation on MCMC methods with a discussion of their convergence. The verification of this convergence is an essential step in all MCMC simulations : it is important to verify the convergence for all the parameters of the model and not only for a subset of parameters. We also present three types of convergence for which an evaluation is necessary.

- **2.1 Convergence to the stationary distribution**
  Convergence to the stationary distribution Convergence to the stationary distribution is necessary because we are trying to approximate the posterior density , so we must ensure that our chain reaches its stationary distribution. The main tool for evaluating convergence to the stationary distribution is to run several chains in parallel to compare their performance. Obviously, this means that the slowest chain in the group determines the convergence diagnosis and that the choice of the initial distribution is extremely important to ensure that the diff'erentes chains are well spread out .

- **Convergence of means**
  Once we have approximately solved the problem of convergence to the

stationary distribution we are again in the classical Monte Carlo framework namely the convergence of the empirical mean $\frac{1}{k}\sum_{k=1}^{k} h(\beta^k)$ to $E[h(\beta)]$ for some function$h$ such that $E[h(\beta)] < \infty$ we need to ensure that the chain has explored the whole support of the a posteriori density in order to infer adequately on it.

- **Convergence to an *iid* sample**
  Convergence to an $i.i.d$ sample is another form of convergence. The general idea is to produce a quasi-independent sample by subsampling to reduce the correlation between the iterations of the Markov chain. Several diagnostics have been proposed in the literature to verify the convergence of MCMC methods, the diagnostics of Gelman and Rubin (1992) and Raftery and Lewis (1992) are currently the most popular in the statistical community at least in part because the computer programs for their implementation are available from their creators.

2.7.3 Algorithms and approximation methods

1. Metropolis-Hastings with the random walk

The Metropolis-Hastings algorithm is an MCMC method whose purpose is to obtain a random sample of a probability distribution when direct sampling is difficult. In particular, it is not necessary to calculate the π partition function, which is often a difficult task. For this reason, this method is widely used in statistical physics [1].

In our case we will focus on a very particular case, the Metropolis-Hastings with the random walk algorithm which is a very simple approach that can be used when π is very poorly known , its principle is to generate a random walk with a correction in order to cover all the possibilities, hence the efficiency. In addition, the probability of acceptance does not depend on g anymore. The chaıne depends on it via the propositions [16].

Algorithm generate

$$y_n \backsim g(y - x^{(t)})$$

take

$$x^{t+1} = \begin{cases} y_t \ , \ with\ \rho(x^t, y_t) \\ x_t^t \ , with\ 1 - \rho(x^t, y_t) \end{cases}$$

where

$$\rho(x^t, y_t) = \{1, \frac{\pi(y_t)}{\pi(x^t)}\}$$

2.Laplace's approximation

Laplace's approximation is an analytic—although asymptotic—alternative to Monte Carlo simulations. This method was introduced by Laplace and is thus called Laplace approximation. It consists to approximate the posterior distribution

$$\pi(\beta|y) = \frac{f(y|\beta)\pi(\beta)}{\int f(y|\beta)\pi(\beta)d\beta}$$

by a multivariate $\pi(\beta|y) \backsim \mathcal{N}(\mu, \Sigma)$

with μ is the vector of expectations and Σ is the matrix of variance-covariance .

Using a condensed notation, we will have

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

We approach $\ln(\pi(\beta|y))$ at the numerator and denominator. Suppose that :

$$\ln(\pi(\beta|y)) = g(\beta) \tag{2.11}$$

Approximate $g(\beta)$ using the 2-order Taylor expansion.

$$g(\beta) \approx g(z) + (\beta - z)^t \bigtriangledown g(z) + \frac{1}{2}(\beta - z)^t \bigtriangledown^2 g(z)(\beta - z) \tag{2.12}$$

where z is an arbitrarily chosen point in the domain of g, choose $z = \beta_{MAP}$ (see Annex) , with β_{MAP} is the estimator of β obtained by maximizing the

posterior distribution. From (2,12), the Laplace approximation is:

$$\pi(\beta|y) = \frac{e^{g(\beta)}}{\int e^{\beta} d\beta} \tag{2.13}$$

$$\approx \frac{e^{g(z)+(\beta-z)^t \bigtriangledown g(z)+\frac{1}{2}(\beta-z)^t \bigtriangledown^2 g(z)(\beta-z)}}{\int e^{g(z)+(\beta-z)^t \bigtriangledown g(z)+\frac{1}{2}(\beta-z)^t \bigtriangledown^2 g(z)(\beta-z} d\beta} \tag{2.14}$$

This can be simplified in two steps:

1 The term $e^{g(\beta_{MAP})}$ in the numerator and the denominator can be considered as a constant because it does not vary in β. It is therefore simplified.

2 By definition of β_{MAP} , the vector $\bigtriangledown \ln \pi(\beta|y) = 0$, approximation is then:

$$\pi(\beta|y) = \frac{e^{\frac{1}{2}(\beta-\beta_{MAP})^t - \bigtriangledown^2 g(z)(\beta-\beta_{MAP})}}{\int e^{\frac{1}{2}(\beta-\beta_{MAP})^t - \bigtriangledown^2 g(z)(\beta-\beta_{MAP}} d\beta} \tag{2.15}$$

So the Laplace approximation of $\pi(\beta|y)$ is a Gaussian and is given by:
where $\mu = arg\max_{\beta}(\ln(\pi(\beta|y))) = \widehat{\beta}_{MAP} = \widehat{\beta}_{MMSE}$
and $\Sigma = [-2 \ln \bigtriangledown \pi(\beta|y)] - 1$. (see Annex)

3.Case of a non-informative prior Jeffreys' distribution

Jeffreys' prior distribution is perhaps the most widely used non-informative prior distribution in Bayesian analysis. For the logistic model, Jeffreys' prior distribution is attractive because it does not cease any elicitation of hyperparameters, there has been an enormous literature on this law and its properties for a wide variety of applications.

This literature is too vast to be listed in its entirety here but we can cite some works such as Firth (1993) suggested the use of Jeffreys' distribution as a solution to problem of bias in maximum likelihood estimators [5], Chen et al (2008) have studied the properties and implementation of Jeffreys' prior law for logistics models [3].
Although the literature on Jeffreys' prior distribution is extensive but there is very little discussion on the theoretical properties on this law for logistic regression models.

presentation of the model

The posterior density is given by:

$$\pi(\beta|y) \propto f(y|\beta)\pi(\beta)$$

As β is a vector $\beta = (\beta_1, \cdots, \beta_p)$ in this case the a priori distribution of Jeffreys will be given by:

$$f(y|\beta) = \prod_{i=1}^{n} \left(\frac{e^{x_i \beta y_i}}{1 + e^{x_i \beta}} \right)$$

where $I(\beta)$ is the information matrix of fisher whose elements are given by:

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

We have :

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

We will therefore have:

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

Hence Jefferys' prior distribution is:

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

The posterior law will be given as follows:

$$\pi(\beta|y) = \prod_{i=1}^{n} \left(\frac{e^{x_i \beta y_i}}{1 + e^{x_i \beta}} \right) \sqrt{det[(I(\beta))_{rk}]}$$

This expression is complex, so it cannot be calculated manually therefore it requires simulation methods.

Chapter 3

Application of logistic regression with R

In this chapter we will present an application in the medical field using softwar R.

The objective is to predicte the value taken by the binary random variable Y taking two modalities $\{0, 1\}$.

We have a sample ε of size n. The value taken by Y for an individual ω is denoted by $Y(\omega)$.

The file has p descriptors $\{X_1, X_2, \cdots, X_p\}$. the vector of values for individu ω is written $X_1(\omega), X_2(\omega), \cdots, X_p(\omega)$.
For a given individual, his prior probability of being positive is written by $P(y(\omega) = 1) = P(\omega)$ we will denote it P.

When the sample is drawn from a random selection from the population, without distinction membership classes, if n^+ is the number of positive observations in ε, p can be estimated by

$$\widehat{P} = \frac{n^+}{n}$$

The posterior probability of an individual to be positive i.e. knowing the values taken by the descriptors are noted $P(Y(\omega) = +|X(\omega)) = \pi(\omega)$

This is the probability that we are trying to model in supervised study.

3.1 Data

Consider the fictitious dataset comprising 20 observations and 3 predictor variables to illustrate binary logistic regression The goal is to predict the presence or absence of a heart problem y: heart, with "presence" = " + " and "absence" = " − " from

- X_1:age is quantative **r.v**.
- X_2: maxrate is quantitative **r.v** (blood pressure).
- X_3: angina is binary **r.v** .

3.2 The logit model

The Logit of an individual ω is written:
$X = (x_1, \cdots, x_p)$ and $\beta = (\beta_0, \cdots, \beta_p)$ are the parameters that we want to estimate from the data.

$$C(x) = \log \frac{\pi(\omega)}{1 - \pi(\omega)} = \beta_0 + \beta_1 x_1(\omega) + \cdots, \beta_p x_p(\omega) = X^t \beta$$

with $X(\omega) = (1, X_1(\omega), \cdots)$ and $\beta = (\beta_0, \cdots, \beta_p)$ is the vector of the parameters.

The quantity $\frac{\pi}{1-\pi} = \frac{P(y=+|x)}{P(y=-|x)}$expresses an odds .

Let $C(X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$, we can find π by means of the function logistics

$$\pi = \frac{\exp(C(X))}{1 + \exp(C(X))} = \frac{1}{1 + e^{-C(X)}}$$

Assignment rule

The assignment rule can be based on π in different ways:

- If $\frac{\pi}{1-1\pi} > 1$ then $Y = +$.
- If $\pi > 0.5$ then $Y = +$.
- If $C(X) > 0$then $Y = +$.

3.3 Interpretation of coefficients β

Relative risk

Definition 3.3.1 *We called the relative risk the increased chance of being positive in the exposed group compared to the control group.*

$$RR = \frac{P(+|1)}{P(+|0)}$$

According to the data table we have the following matrix :

```
       angina1
heart1  0  1
     0 12  2
     1  3  3
```

we obtain the following contingency table:

$Y\|X$	1	0	full
1	a=2	b=12	a+b=14
0	c=3	d=3	c+d =5
full	a+c=5	b+d=15	n=20

Table 3.1: contingency table

So we have:

$$RR = \frac{P(+|1)}{P(+|0)} = \frac{\frac{a}{a+c}}{\frac{b}{b+d}} = 3$$

- People who have angina are 3 times more likely than those who do not have angina to develop heart disease.
- RR characterises a relationship between heart disease and the occurrence of angina.
- When $RR = 1$, it means that angina has no impact on the disease.

Definition 3.3.2 *The odds ratio is defined as a ratio of probabilities in a group.*

Remark 3.3.1 *The odds for an individual x to obtain the answer $Y = 1$ can be written:odds$(x) = \frac{\pi(x)}{1-\pi(x)}$, where $\pi(x) = P(Y = 1|X = x)$.*

- *If odds* > 1 *then* $Y = 1$.
- *If odds* < 1 *then* $Y = 0$.

Exemple 3.3.1 *According to contingency table , the odds (in the exposed group) are written :* $odds(1) = \frac{P(+|1)}{P(-|1)} = \frac{\frac{a}{(a+c)}}{\frac{c}{(a+c)}} = 1.5$

Interpretation

In the group of people with angina pectoris, one has : 1.5 times more likely to have heart disease than not.

Definition 3.3.3 *The odds ratio between two individuals* x *and* x' *is:*

$$OR(x, x') = \frac{odds(x)}{odds(x')} = \frac{\frac{\pi(x)}{1-\pi(x)}}{\frac{\pi(x')}{1-\pi(x')}}$$

This is the ratio between the odds of the exposed group and the odds of the control group.

$OR = OR(1, 0) = \frac{odds(1)}{odds(0)} = \frac{ad}{bc} = 6$

The OR indicates same thing as the relative risk, i.e. in the exposed group you are 6 times more likely to have the disease than in the control group.

- $OR = 1$, the disease is independent of the symptom. no impact on Y.
- $OR > 1$, the disease is more frequent for individuals who have the symptom *Odds* increase ($odds(x) > odds(x')$).
- $OR < 1$, disease is more common for individuals who do not have the symptom Odds decrease ($odds(x) < odds(x')$).

3.4 Parameter estimation

$Y \in \{+, -\}$ (or $\{1.0\}$). For an individual ω, we model the probability using the law binomial ,$\mathcal{B}(1, \pi)$, with

$$p(Y(\omega)|X(\omega)) = \pi(\omega)^{Y(\omega)} \times [1 - \pi(\omega)]$$

- If $Y(\omega) = 1 \Rightarrow p(Y(\omega) = 1|X(\omega)) = \pi$.
- If $Y(\omega) = 0 \Rightarrow p(Y(\omega) = 1|X(\omega)) = 1 - \pi$.

3.4.1 Likelihood

The likelihood of a sample ε is written as

$$L = \prod_{\omega} p(Y(\omega)|X(\omega)) = \pi(\omega)^{y(\omega)} \times [1 - \pi(\omega)]^{1-y(\omega)}$$

It corresponds to the probability of obtaining the sample E from a selection from the population.
The maximum likelihood method consists in determining the vector of the parameters ,$\beta = (\beta_0, \cdots, \beta_p)$ which maximize the probability of observing this sample.

3.4.2 Log-likelihood

Log-likelihood of a sample ε is written as

$$LL = \sum_{\omega} (Y(\omega) \times \ln \pi(\omega) + (1 - Y(\omega)) \times \ln(1 - \pi(\omega)))$$

3.4.3 Deviance

Deviance is a measure of the gap between the data and the model. It is based on the likelihood function L.

Definition 3.4.1 *The quantity:*

$$D_M = -2LL$$

is called deviance (or residual deviance). Opposite to the log-likelihood, it is positive.

We can compare it to the sum of the squares of the residuals of the multiple linear regression, but the latter always reflects the difference between the data and the model.

- It makes it possible to quantify the goodness of fit of the model, and the comparison of nested models which make it possible to evaluate the contribution of one or more predictors in relation to a basic model.

- The null deviance D_0 calculated on the model only composed of the finding would then correspond to the sum of the total squares.

In some works, we define deviance D more generically:

$$\begin{aligned} D &= 2\ln\left[\frac{L \;\; Saturated \;\; model}{L \;\; Studied \;\; model}\right] \\ &= 2\ln LL \;\; (Saturated \;\; model) - 2LL \;\; (Studiedmodel) \\ &= D_M - 2LL(Studied \;\; model) \\ &= -2\sum_{\omega}\left[y\ln\left(\frac{\widehat{\pi}}{y}\right) + (1-y)\ln\left(\frac{1-\widehat{\pi}}{1-y}\right)\right] \end{aligned}$$

3.5 Evaluation of the logistic regression

A saturated model for individual data is a model that perfectly reconstructs the values of the dependent variable, i.e.$\widehat{\pi}(\omega) = y(\omega)$.

- Its likelihood is equal to 1.
- Its loglikelihood is equal to 0. In this context, $D = D_M$.

The objective is to minimize this deviance .
Application to the file $Heart = f(age, maxrate, angina)$. We will use R software to determine:

1 The deviance value D_M

2 The coefficient of the parameters $\widehat{\beta} = (\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2)$

3 Determine an estimate of the function $C(X)$.

```
Call:
glm(formula = heart ~ age + maxrate + angine, family = "binomial",
    data = Heart)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9773  -0.5437  -0.3876   0.5093   1.7577

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 14.49379    7.95464   1.822   0.0684 .
age         -0.12563    0.09380  -1.339   0.1805
maxrate     -0.06356    0.04045  -1.572   0.1161
angine       1.77901    1.50449   1.182   0.2370
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 24.435  on 19  degrees of freedom
Residual deviance: 16.618  on 16  degrees of freedom
  (1 observation deleted due to missingness)
AIC: 24.618

Number of Fisher Scoring iterations: 5
```

Figure 3.1: Processing of the heart file

1 The deviance $D_M = 16.618$

2 The values of the parameters which made it possible to obtain it are: $\widehat{\beta} = (14493.7, -0.1256, -0.0636, 1.7790)$

3 The estimated Logit predicting the occurrence of heart disease from age, max rate and angina, is written: $C(X) = (14.4937 - 0.1256X_1 - 0.0636X_2 + 1.7790X_3)$

3.5.1 The fit measures

The pseudo R^2's in the logistic model do not measure the proportion of variance explained but rather an improvement in the full model (containing all predictors) over the null model (containing only the model constant).
ℓ and $\mathcal{L}$ represent the likelihood function and log likelihood for the model containing all predictors respectively.
ℓ_0 and $\mathcal{L}_0$ represent the likelihood function and log likelihood for the model

containing only the constant respectively.

In addition, several forms of R^2 have been proposed in the literature, we distinguish some of them, presented as follows:

McFadden's R^2

This is one of the simplest and most suitable early indicators found in the literature, which is defined by :

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

Where $\min R^2_{MF} = 0$ if $\ell(\beta, y) = \ell_0(\beta, y)$ and $\max R^2_{MF} = 1$ if $\mathcal{L}(\beta, y) = 1$ i.e $\ell(\beta, y) = 0$

Cox and Snell's R^2

This popular measure, can be computed from any model estimated by the maximum likelihood method, which will be define by :

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

$\min R^2_{CS} = 0$ if $\mathcal{L}(\beta, y) = 1$ with $\max R^2_{CS} = 1 - \mathcal{L}(\beta, y)$

Nagelkerke's R^2_N

This pseudo is an adjusted version of Cox and Snell's R^2_N, given by :

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

With $\min R^2_N = 0$ and $\max R^2_N = 1$.
This is a simple normalization of the Cox and Snell R^2.

When these coefficients are close to 1 it means that the model is globally significant, however these R^2 are often small and difficult to interpret, they are generally considered correct if $R^2_N > 0.2$.
The R^2 of Cox and Snell , Nagelkerke and McFadden are given by the following command:

```
> library(modEvA)
> RsqGLM(model=reglo)
$CoxSnell
[1] 0.3235144

$Nagelkerke
[1] 0.4587037

$McFadden
[1] 0.3199108

$Tjur
[1] 0.3620991

$sqPearson
[1] 0.3472096
```

Figure 3.2: The diffrents pseudo R^2

As its pseudo R^2's are often small and difficult to interpret, the model is said to be globally valid as soon as its R^2's exceed the value 0.2. This is the case in this example.

3.5.2 The confusion matrix

The confusion matrix is another procedure for evaluating logistic regression. It compares (evaluates) the observed values of the dependent variable Y with those that are predicted and then counts the good and bad predictions.

Definition 3.5.1 *The confusion matrix is defined by:*

$$CM = \begin{pmatrix} Number\ of\ 0\ predicted; 0\ in\ reality & Number\ of\ 1\ predicted; 0\ in\ reality \\ Number\ of\ 0\ predicted; 1\ in\ reality & Number\ of\ 1 predicted; 1\ in\ reality \end{pmatrix}$$

Generic form of the Confusion Matrix

$Y\|\widehat{y}$	1	0	full
1	a=2	b=12	a+b
0	c=3	d=3	c+d
full	a+c=5	b+d	n

Table 3.2: contingency table of confusion matrix

- a are true positives: observations that have been classified as positive and are actually positive.
- c are false positives: individuals classified as positive who are in fact negatives.
- b are false negatives.
- d are true negatives.

Several indicators can be derived to account for the agreement between observed and predicted values.

Error rate, Success rate, Sensitivity

- **The error rate** is given by:

$$ER = \frac{b+c}{n} = 1 - \frac{a+d}{n}$$

It estimates the probability of misclassification of the model.

- **The success rate** corresponds to the probability of the model being correctly classified

$$SR = \frac{a+d}{n}$$

- **The sensitivity** (or true positive rate (TPR)) indicates the ability of the model to find the positives

$$Se = Sensitivity = TPR = \frac{a}{a+b}$$

- **The precision** indicates the proportion of true positives among the individuals who have been classified as positive

$$precision = \frac{a}{a+c}$$

  It estimates the probability of an individual being truly positive when the model classifies him as such (it is also called the positive predictive value (PPV)).

- **The specificity**, in contrast to the sensitivity, indicates the proportion of negatives detected.

$$Sp = Specificity = \frac{d}{c+d}$$

```
> prev=predict(reglo,newdata=Heart,type = "response")
> print(prev)
        1         2         3         4         5         6         7
0.87894733 0.58154537 0.39220275 0.37820752 0.21335852 0.87655486 0.01640958
        8         9        10        11        12        13        14
0.07103688 0.37750865 0.03624840 0.85841939 0.10575388 0.10366373 0.40566043
       15        16        17        18        19        20        21
0.12437705 0.05836647 0.17271990 0.13818549 0.13712678 0.07370700         NA
> prev1=factor(ifelse (prev >0.5,1,0))
> mc<-table(Heart$heart,prev1)
> print(mc)
   prev1
     0  1
  0 13  1
  1  3  3
```

Figure 3.3: confusion matrix

The confusion matrix is formed by comparing the Heart and Prediction columns. The main indicators for evaluating the classifiers:

- Error rate $ER = \frac{1+3}{20} = 0.20$
- Success rate $SR = Sp\frac{3+13}{20} = 0.80$
- Sensitivity $TPR = \frac{3}{6} = 0.50$
- $Accuracy = \frac{3}{4} = 0.75$
- Specificity $Sp = \frac{13}{14} = 0.93$

3.5.3 The Roc curve

The *ROC* curve is a graphical representation of the relationship between the sensitivity and the specificity for all possible threshold values. The ordinate represents the sensitivity and the abscissa corresponds to the specificity.
It is possible to digitally characterise the *ROC* curve by calculating the area under the curve *AUC*.
This is the *AUC* criterion which expresses the probability of placing a positive individual in front of a negative one. *AUC* is a measure of the performance of the model in prediction, a perfect model will have an *AUC* measure of 1. Thus, the more accurate the model, the closer the *ROC* curve is to the left hand corner of the graph and the *AUC* measure is close to 1.

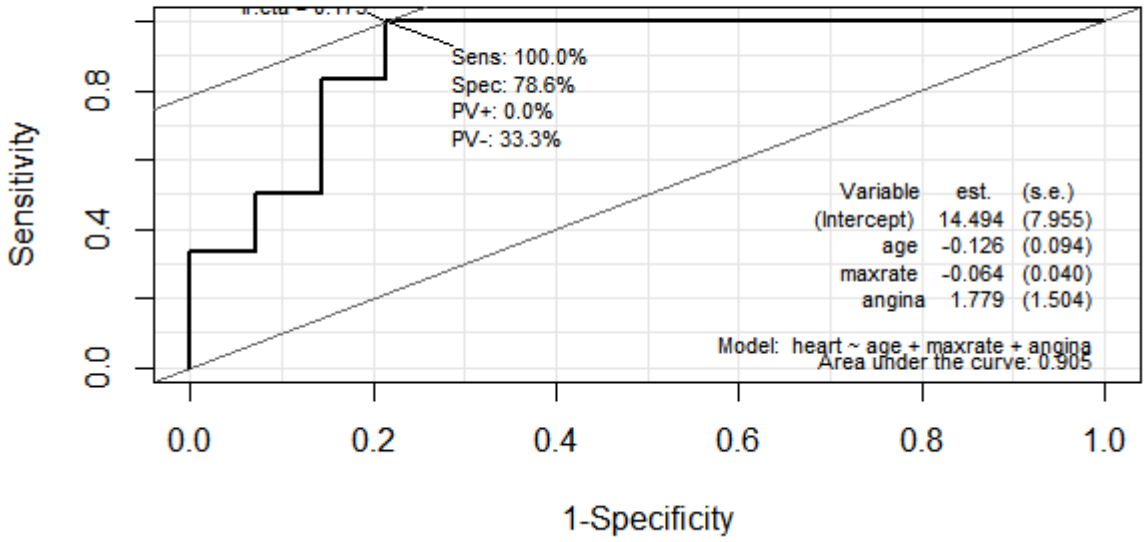


Figure 3.4: The ROC curve

From the figure below we can see that the *ROC* curve is close to the left corner upwards and the value of the air under the curve $AUC = 0.905$ which is close to 1, so the model is accurate

3.6 Statistical Evaluation of Regression

We have two strategies for implementing these tests:

1 **The likelihood ratio principle**:

- This approach detects the alternative hypothesis better when it is true.
- The disadvantage is that it is more demanding in terms of machine resources: each hypothesis to be evaluated gives rise to a new estimation of the parameters, thus to an optimisation process.

2 **The Wald test**:

- The main advantage is that the information that one wishes to exploit is all available at the end of the estimation of the complete model.
- The disadvantage is that the Wald test tends to favour the null hypothesis.

3.6.1 Likelihood ratio test

```
> reglo1<-glm(heart~1,family = "binomial",data =Heart )
> anova(reglo1,reglo,test = "Chisq")
Analysis of Deviance Table

Model 1: heart ~ 1
Model 2: heart ~ age + maxrate + angina
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        19     24.435
2        16     16.618  3   7.8169  0.04995 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 3.5: Results of the likelihood ratio test

Or $ML = 24.435 - 16.618 = 7.817$, and $\mathcal{X}^2_{0.95,3} = 7.8147$.

Since $ML > 7.8147$, then the model is globally good. Hence, there is at least one significant explanatory variable of y.

3.6.2 Wald test

```
Call:
glm(formula = heart ~ age + maxrate + angine, family = "binomial",
    data = Heart)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9773  -0.5437  -0.3876   0.5093   1.7577

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 14.49379    7.95464   1.822   0.0684 .
age         -0.12563    0.09380  -1.339   0.1805
maxrate     -0.06356    0.04045  -1.572   0.1161
angine       1.77901    1.50449   1.182   0.2370
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 24.435  on 19  degrees of freedom
Residual deviance: 16.618  on 16  degrees of freedom
  (1 observation deleted due to missingness)
AIC: 24.618

Number of Fisher Scoring iterations: 5
```

Figure 3.6: Wald test results

We have $W_1 = \frac{\widehat{\beta^1}}{\widehat{\sigma}_{\widehat{\beta^1}}} = \frac{-0.12563}{0.09380} = -1.3393$.
Where $z_{1-\frac{\alpha}{2}}$ is the quantile of order $1 - \frac{\alpha}{2}$ of the reduced central normal distribution which is equal to 1.96.
Since we have $W_1 < 1.96$, therefore the variable 'age' is not significative of y.

$$W_2 = \frac{\widehat{\beta^2}}{\widehat{\sigma}_{\widehat{\beta^2}}} = \frac{-0.06356}{0.04045} = -1.5713$$

.

We have $W_2 < 1.96$, therefore the variable 'maxrate' is not significative of y.

$$W_3 = \frac{\widehat{\beta^3}}{\widehat{\sigma}_{\widehat{\beta^3}}} = \frac{1.77901}{1.50444} =$$

.

We have $W_3 < 1.96$, therefore the variable 'angina' is not significative of y. Regarding the Wald test: Less powerful, more conservative. It favours the null hypothesis H_0. For the fichier "Heart" data H_0 was never rejected regardless of the test implemented.

When the value of the coefficient is high, the estimate of the standard deviation is exaggerated. Again H_0 is favoured in individual tests, this leads us to wrongly remove important variables from the model.

3.7 Confidence intervals

Relying on the Wald test ,we can construct the confidence interval at confidence level $1 - \alpha$ for any individual coefficient.

```
> confint.default(reglo,level=0.9)
                  5 %         95 %
(Intercept)  1.4095662 27.578014811
age         -0.2799259  0.028657638
maxrate     -0.1300874  0.002966815
angina      -0.6956545  4.253680345
```

Figure 3.7: Confidence intervals for different coefficient

3.8 prediction

To obtain a "logit" prediction for a new individual ω' to be classified, we need to apply the estimated coefficients from logistic regression

$$\widehat{\pi(x)} = C(x(\omega')) = \widehat{\beta}_0 + \widehat{\beta}_1 x_1(\omega') + \cdots + \widehat{\beta}_p x_p(\omega')$$

Where $\widehat{\beta} = (\widehat{\beta}_0, \cdots, \widehat{\beta}_p)$ the vector of estimated parameters.

From the Logit, we can derive an estimate of the posterior probability of the individual being positive, i.e. $\widehat{\pi}(\omega') = \frac{1}{1+e^{\widehat{C}}}$ and applying the standard affectation rule, we obtain $\widehat{y}$ If $\widehat{\pi} > 0.5$ then $y = +$ else $y = -$

```
> prev=predict(reglo,newdata=Heart,type = "response")
> print(prev)
         1          2          3          4          5          6          7
0.87894733 0.58154537 0.39220275 0.37820752 0.21335852 0.87655486 0.01640958
         8          9         10         11         12         13         14
0.07103688 0.37750865 0.03624840 0.85841939 0.10575388 0.10366373 0.40566043
        15         16         17         18         19         20         21
0.12437705 0.05836647 0.17271990 0.13818549 0.13712678 0.07370700         NA
> prev1=factor(ifelse (prev >0.5,1,0))
> print(prev1)
   1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16   17
   1    1    0    0    0    1    0    0    0    0    1    0    0    0    0    0    0
  18   19   20   21
   0    0    0 <NA>
Levels: 0 1
```

Figure 3.8: Prediction results

3.9 Bayesian estimation

In this section we will analyze the previous data processed by the classic binary model using the Bayesian approach, and this via R software .

3.9.1 Metropolis Hasting with the random walk

We use the MCMCpack package, which has the MCMClogit function for estimating Bayesian logistic models. The algorithms in MCMCpack employ a random walk version of the Metropolis-Hastings algorithm when estimating a logistic model [13]

For our example ,we shall employ by default multivariate normal priors on all of the parameters. It is used because we have more than one parameter. The priors are noninformative, and therefore do not appreciatively influence the model. That is, the data, or rather likelihood, is the prime influence on the parameter estimates, not the priors.
The output is given as usual:

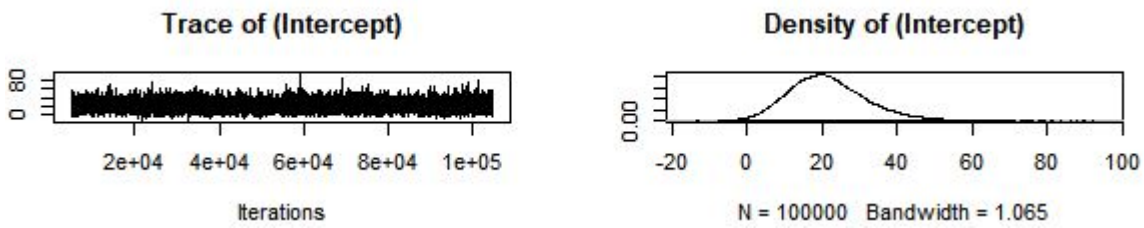
```
Iterations = 5001:105000
Thinning interval = 1
Number of chains = 1
Sample size per chain = 1e+05

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

                Mean       SD  Naive SE Time-series SE
(Intercept) 21.56319 10.52635 0.0332872      0.1546486
age         -0.19550  0.12506 0.0003955      0.0017740
maxrate     -0.09046  0.05145 0.0001627      0.0006925
angina       2.60393  1.91840 0.0060665      0.0261470

2. Quantiles for each variable:

               2.5%     25%     50%      75%     97.5%
(Intercept)  3.7020 14.2978 20.6337 27.75890 45.027449
age         -0.4828 -0.2661 -0.1808 -0.10825  0.005593
maxrate     -0.2069 -0.1203 -0.0855 -0.05466 -0.002428
angina      -0.8426  1.3102  2.4723  3.76600  6.763332
```

Figure 3.9: Estimation result using MCMC

Although interpretations differ, the posterior mean values are larger then the maximum likelihood coefficients.

3.9.2 Convergence diagnostics of the posterior distribution

For the convergence,we want to know if this sample is close enough to the posterior law to be used for the analysis, for this there are several diagnoses which will be presented as follows:



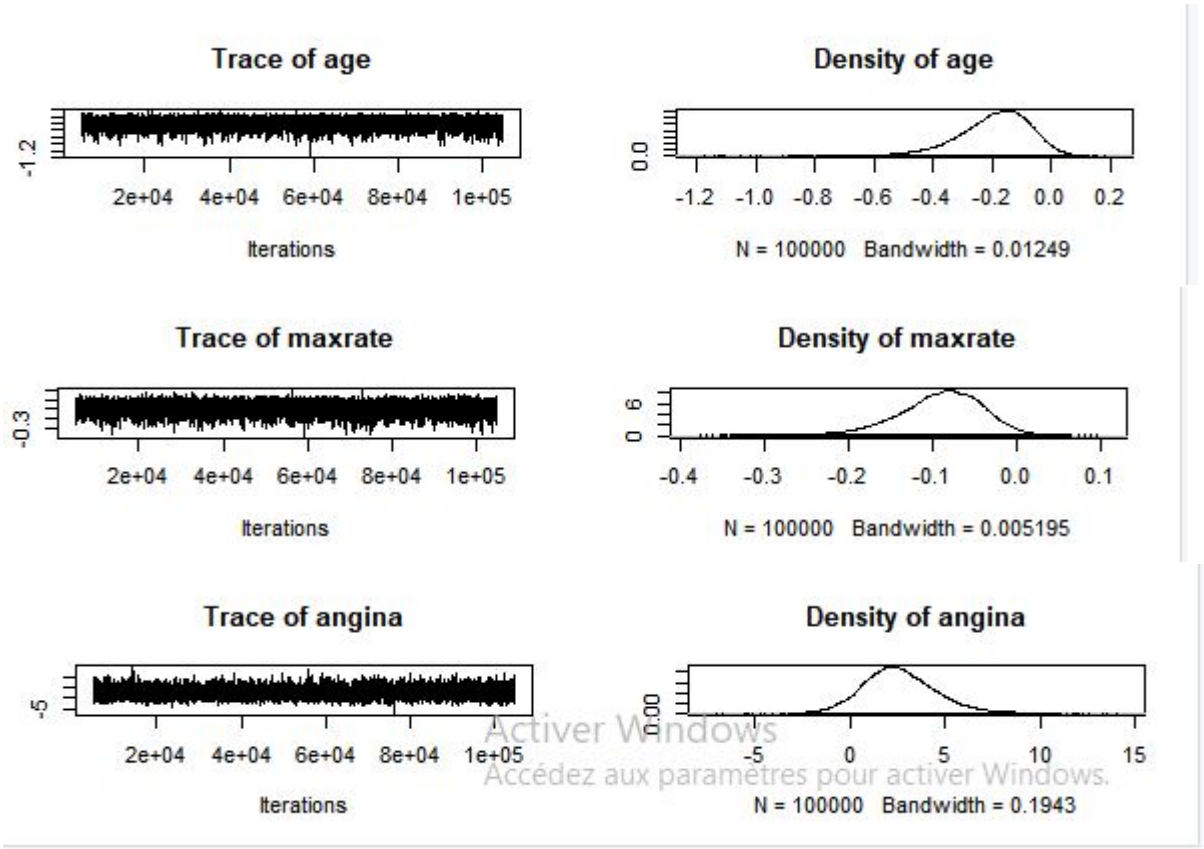


Figure 3.10: R trace and density plots of model with noninformative prior

From figure (3.10) we deduce that the Markov chain is relatively stationary this implies that the chain has reached or is close to its stationary distribution, we can therefore reasonably assume that our simulations are taken from the distribution a desired posteriori $\pi(\beta|y)$.
The quality of the graphs in figure (3.10) is satisfactory. Because the densities of the laws has posteriori are close to the Gaussian density.

Conclusion

The study of logistic regression is important to understand that the goal of analysis using the model is the same as that of any model-building thechnique used in statistic.
In this work we have studied the logistic model for the Bayesian approach enhanced by MCMC approximation techniques.
The quality of the estimate is analyzed by quantities such as R^2 in the case of logistic models, the matrix confusion and *ROC* diagrams in order to evaluate the quality of the logistic model.

An example and an application are presented to illustrate binary logistic modeling by using R software which allows to implement the Metropolis-Hasting , from this we can see that the performance of computers has made feasible efficient simulation procedures and the availability of computer programs to make easy the posterior probability calculation which was until now of discouraging complexity, and we obtained the same conclusion by the two estimation methods.

We have presented an application whose data from [22]. We have established ourse of the estimates presented made it possible to accept the same variable but the value of the coefficient of this variable is larger in the Bayesian estimate, which allowed us to have in the forecast results, less chance of failing than the result obtained with the estimate classic.
Let us conclude with this observation that regression methods are very powerful methods but which must be used with great discernment and caution. Obviously the polytomous logistic regression remain very rich models to develop and that we have not been able to do as part of this brief work.

Annex

$*$ Some others estimators:

- Minimum Mean Square Error estimator ($MMSE$)

  The $MMSE$ estimator of θ, noted $\widehat{\theta}_{MMSE}$, associated with the quadratic loss function, relatively to the prior distribution π, is the posterior mean of θ, that is to say:

$$\widehat{\theta}_{MMSE}(x) = E(\theta|x), \quad x \in \mathbb{R}$$

- The posterior median

  The Bayesian estimator associated with the prior distribution π and with the absolute loss function is the fractile of order $\frac{1}{2}$ of the posterior distribution. It is then the posterior median which is given by:

$$P(\theta|\delta) = P(\theta < \delta|x) = \frac{1}{2}$$

- Maximum Posterior estimator (MAP)

  The MAP estimator of θ is obtained by maximizing the posterior distribution:

$$\widehat{\theta}_{MAP}(x) = arg_{\theta} \max(\pi(\theta|x))$$

Remark 3.9.1 *The MAP estimator is not a Bayesian estimator because it does not verify the definition (1.3.7).*

Bibliography

[1] ANAS ALTALEB, CHRISTIAN P. ROBERT, Analyse bayesian anlyse of the logit mode: algorithm of Metropolis-Hastings ,rvue of statistic tome 49,n°9,2001,53-54

[2] CHEN, M.H and al. Properties and Implementation of Jeffreys's Prior in Binomial Regression Models. Journal of the American Statistical Association, 2008, V(108), 1659-1664.

[3] Christian P. Robert,The Bayesian Choice From Decision-Theoretic Foundations to Computational Implementation,Second Edition,Springer,2007.

[4] CHEN, MH and al. Properties and Implementation of Jeffreys's Prior in Binomial Regression Models. Journal of the American Statistical Association, 2008, V (108), 1659page-1664.

[5] FIRTH, D. Bias Reduction of Maximum Likelihood Estimates. Biometrika, 1993, V (80), 27-38

[6] FOONG, A.P, HU, Y.H and HEISEY, D.M. Logistic regression in an adaptive Web cache. IEEE Internet Computing, 1999, V(3), 27-36.

[7] GAMERMAN, D. Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference. 2 th ed. London, Chapman and Hall CRC Press, 2006, 342.

[8] GELFAND, AE and ADRIAN, FM SMITH. Sampling-based approaches to calculating mar- ginal densities. Journal of the American Statistical Association, 1990, V (85), 398-409.

[9] GENKIN, A and al. Large-scale bayesian logistic regression for text categorization. Technometrics, 2007, V(49), 291-304.

[10] GORDOVIL, M, GUARDIA, O, PERO, C and 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.

[11] GOURIEROUX, M. Asymptotic Properties of the Maximum Likelihood Estimator in Dicho- tomous Logit Models. J Econom, 1981, V (17), 83-97.

[12] GILLET, A and al. Main models used in logistic regression. Biotechnol. Agron. Soc. Environ, 2011, V (15), 425-433.

[13] HILBE, JM Practical Guide to Logistic Regression. Taylor and Francis Group, 2016, 130.

[14] HOSMER, DW and LEMESHOW, S. Applied logistic regression. 2 th ed. John Wiley and Sons, 2000, 49-56.

[15] MILA, A and 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.

[16] NEAL, R.M. Slice sampling. Annals of statistics,2003, V(31), 705-74.

[17] NELDER, J.A and WEDDERBURN, R.W.M. Generalized Linear Models. Journal of the Royal Statistical Society, 1972, V(135), 370-384.

[18] PALMA, L, BEJA, P and RODRIGUES, M. The use of sighting data to analyse Iberian lynx habitat and distribution. Journal of Applied Ecology, 1999, V(36), 812-824.

[19] RAKOTOMALALA R. Econometrics Simple and multiple linear regression. Lyon 2, Course de License, 2018, 183.

[20] ROBERT, CP and CASELLA, G. Methods of Monte-Carlo with R. 1 st ed. Paris, Springer,2011, 273p

[21] RUCH, JJ and CHABANOL, ML Markov chain. Bordeaux 1, Preparation course for aggregation, University, 2013, 27.

[22] ROUVIERE.Laurent ,Logistic regression with R , UFR Sciences Sociales,Rennes 2 uinversity

[23] THOMAS, P. Ryan, Some issues in logistic regression. Communications in Statistics-Theory and Methods, 2000, V(29), 9-10.

[24] 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.