RESEARCH ARTICLE OPEN ACCESS

# Bayesian Methods for Logistic Regression with Application to Albanian State Matura

Klodiana Bani*, Markela Muca**
*(Department of Applied Mathematics, Faculty of Natural Sciences, University of Tirana)
** (Department of Applied Mathematics, Faculty of Natural Sciences, University of Tirana)

## Abstract:

Bayesian methods are focused in some important issues. First, the incorporation of prior information that is an opinion given from experts or prior data. The prior information is usually specified in the form of a distribution that might be normal/Gaussian, Poisson, binomial etc. and represents a probability distribution for a coefficient. The prior is combined with a likelihood function that is based on the data. Then combining the results of this function with the prior distribution creates the posterior distribution of the coefficient values.

In this article we will consider the results of the simple logistic regression combined with bayesian estimation of the coefficients of the model to evaluate the probability of acceptance of graduates in universities classifing them as winning / non-winning.

The analysis is based on a sample taken from the database of the MSH 2011 (State Matura). All analysis is performed using the language R.

*Keywords* — **logistic regression (LR), Bayesian estimation, prior and posterior distribution.**

## I. INTRODUCTION

Bayesian statistical analysis has strongly benefited from the development of computer science for more than two decades. It is because Bayesian techniques can be easily applied to complex modelling problems even though they have not been implemented before. This shows that these techniques are becoming quite competitive compared to the classic statistical ones. In this article, the logistic regression model will be dealt with known techniques against Bayesian estimation techniques.

The treatment of Bayesian proper techniques is also given in various books by different authors ([3]-[7], [11]). In general, Bayesian efforts in inferential statistics differ from those of classical statistics assuming that the data is fixed (not random) but the model parameters are random variables which poses the probability problem for a hypothesis to be true. Traditional statistics instead, the model parameters are fixed while the data is considered random.

Estimation from Bayesian techniques are based on five stages:

- The prior information it is commonly received by experts of the field or literature for similar studies. This information is provided by probability distributions (normal, binomial, Poisson, etc.) and presents the probability distribution for a coefficient that interests us in the model.
- The prior distribution is combined with the maximum likelihood function, where this function represents the data, which means that the probability distribution of the estimate produced by the data.

- The combination of prior distribution with the results of the maximum likelihood function to create the posterior distribution of the coefficient needed in the model.
- The simulations are taken from the posterior distribution to create an empirical distribution of values similar to population parameters.
- Simple statistics are used for a summary of the empirical distribution after the simulations by the posterior distribution. The mean, mode or median of this empirical distribution represents the estimation of the coefficient taken from the maximum likelihood function for the correct population parameters and can also be given confidence intervals for the exact values of all unknown coefficients of the model.

The beginning of all Bayesian analysis is based on a prior probability distribution that may be strong or weak; this is based on the belief we have on prior distribution ([8]). The weak prior distribution relate to a wide area of distribution and background information, where the likelihood function will have a stronger impact on the creation of posterior distribution. While a strong prior distribution has a stronger impact and it is inherent in the creation of the posterior, limiting the range of the possible values and hence it has a lower impact of the likelihood function on the posterior distribution.

## II. THEORY

### A. *Binary Logistic Regression*

Instead of modeling the probability *p* directly with a linear model, we firstly consider the odds ratio as follows**:**

$$odds = \frac{\mathrm{P(A)}}{1-\mathrm{P(A)}} = \frac{prob\ of\ 1}{prob\ of\ 0} \quad (1)$$

In logistic regression for binary variable, we model the natural logarithm of odds ratio which is called logit (p):

$$logit(p) = \ln(odds) = \ln\left(\frac{P(A)}{1-P(A)}\right) =$$

$$= \ln\left(\frac{prob\ of\ 1}{prob\ of\ 0}\right) \quad (2)$$

As it is seen logit is a function of the probability p and it can be calculated very simply from the contingency table when we have n pairs of values ($x_i$,$y_i$). The simple logistik model has the form (3):

$$\log it(p) = \ln(odds) = \ln(\frac{p}{1-p}) = \beta_0 + \beta_1 X_1 \quad (3)$$

The regression coefficients $\beta_0, \beta_1$ are used to set a model that would classify objects in one group or in two groups. If the object has the same probability to belong to two groups, which means that p = q = 0.5 or by equation (6) it follows that:

$\beta_0 + \beta_1 x_1 = 0 \Leftrightarrow x_1 = -\frac{\beta_0}{\beta_1}$ and $A = (-\frac{\beta_0}{\beta_1}; 0.5)$ is the symmetry point (see figure 1)

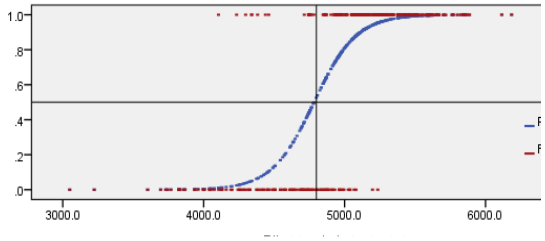

Fig.1 The diagram of points versus the probability of joining the group

So, an object will be classified in the first group or in the second group according to the following rules:

- Classified in the first group if: $\beta_0 + \beta_1 x_1 > 0$
- Classified in the second group if: $\beta_0 + \beta_1 x_1 < 0$

The rules (10) and (11) are based on the cut-off probability equal $p_c$ to 0.5. If we change the value of the $p_c$, then the general rule of classification are:

- Classified in the first group if: $\beta_0 + \beta_1 x_1 + ... + \beta_p x_p < \ln \frac{p_c}{1-p_c}$

- Classified in the second group if:

$$\beta_0+\beta_1 x_1+\ldots+\beta_p x_p<\ln\frac{p_c}{1-p_c}$$

As in our case we have yet another factor $X_2$ which shows the total score on four tests then the logistik model takes the form (4):

$$\log it(p)=\ln(odds)=\ln(\frac{p}{1-p})=\beta_0+\beta_1 X_1+\beta_2 X_2 \quad (4)$$

As alternative hypothesis we mark at least one of the parameters to be equal to zero, so that the logistic regression model predicts the probability of response better than the average of the response variable Y.

### B. Bayesian estimation approach

Bayesian inference is based on the well known theorem of Bayes that helps us calculate the probability of related event as in form (5):

$$P(A|B)=\frac{P(B|A)\times P(A)}{P(B)} \quad (5)$$

where A and B are events, P(A|B) is the conditional probability that event A occurs given that event B has already occurred and P(A) and P(B) are the marginal probabilities of A and B respectively.

As the statistical inference is considered the process of deducing about probability distribution from data or different population properties, usually using the maximum likelihood function, in fact Bayesian inference is considered the process of doing the same thing but using Bayes' theorem.

If we consider the data $D=(y_1,y_{2,\ldots,}y_n)$ and θ is the parameter that we need find, the Bayes' theorem model is:

$$P(\theta|data)=\frac{P(data|\theta)\times P(\theta)}{P(data)} \quad (6)$$

where *P(θ)* is the prior distribution and represents the beliefs we have about the true value of parameter and *P(data|θ)* is the posterior distribution of our belief of the parameter after calculations based on the data.

The Bayesian inferential analysis goes through three steps:

- It is given the prior distribution p(θ), which can be built in different ways, usually from a given value ([1]).
- The conditional distribution p(D|θ) is computed and then is made its substitution with maximum likelihood function ([2]).
- The posterior probability distribution we are interested in is p(θ|D).

Assume that parameter θ is a random variable and the prior distribution p(θ) is known, than the Bayesian estimation is presented from the distribution from (7):

$$p(x|D)=\int p(x|\theta)p(\theta|D)d\theta \quad (7)$$

There is an important equation that makes the connection between *p(x|D)* and *p(θ|D)* based on the Bayes' theorem:

$$p(\theta|D)=\frac{p(D|\theta)p(\theta)}{p(D)}=c\prod_{k=1}^{n}p(y_k|\theta)p(\theta) \quad (8)$$

and *c* is a proportional coefficient dependent from the data.

In this way, Bayesian estimation approach gives a distribution rather than making point estimations as the maximum likelihood method. If we suppose that x is another observation, $\hat{\theta}$ is the best point estimation of $\theta$ and $p(\hat{\theta})\neq 0$ than we can make the approximation:

$$p(x|D)=\int p(x|\theta)p(\theta|D)d\theta\propto p(x|\hat{\theta}) \quad (9)$$

This is the maximum likelihood solution because p(D|θ) peaks at $\hat{\theta}$ too.

## III. THE SET OF DATA AND APPLICATIONS

We will use a sample from the set of data extracted from the database of the results of high school graduates in Matura 2011, at the Faculty of Natural Sciences, University of Tirana, to test how the admission of the candidates is affected by several factors. This set consists of n = 400 graduates high school students who have chosen the Form A2 to study the programs of this faculty, there are one dependent variable and the predictive variables (Predictor variables). The dependent variable indicates whether a candidate has won or not depending on the two independent variables.

The two factors are gender and the total points on 4 tests (Points). The variables of this dataset are given in Table 1.

TABLE I
DESCRIPTION OF THE DATASET

| | Variables | Description |
|---|---|---|
| 1 | Admit | Qualitative r. v. with two values 0- is not admitted and 1- admitted. |
| 2 | Gender | Qualitative r .v .with two values 0-male and 1- female. |
| 3 | Points(Total points in four tests) | Countinous r.v with values in $R^+$ |

In Table 2, it is showed that 191 (77.33%) females are winners in of the branches and 89 (58.2%) males are winners. The total score ranges from 3049 to 6456.6, the arithmetic average and the standard deviation are 5003.02 and 488.36 points respectively. 61.75% ($n_f$ = 247) are female and 38.25% ($n_m$ = 153) are males.

TABLE 2
CONTINGENCY TABLE OF THE DATA

| | | Gender | | Total |
|---|---|---|---|---|
| | | F | M | N |
| Admit | Not wins in the II stage | 56 | 64 | 120 |
| | Wins in a branch | 191 | 89 | 280 |
| Total | | 247 | 153 | 400 |

### A. *Logistic Regression Model*

The logistic regression is part of a class of algorithms called Generalized Linear Model (GLM) because it can be modelled as:

$$g\big(E(Y)\big) = a + bx_1 + cx_2 \quad (10)$$

where *E(Y)* is the expectation of the dependent variable Y and $a + bx_1 + cx_2$ is the linear predictor and *a, b, c* are going to be predicted by Bayesian estimators.

GLM was proposed by Nelder and Wedderburn in 1972([12]) to solve problems that are not connected with linear regression, which included logistic regression as a case of this class.

There some points to take in consideration:

- Logistic regression model does not assume a linear relationship between dependent and independent variables but a linear relationship of the log(y) and the independent variables.
- It is not fundamental for the dependent variable to be normally distributed.
- The Ordinary Least Square (OLS) is not used for parameter estimation, but instead is used the maximum likelihood estimation.
- It is necessary for the errors to be independent but not normally distributed.

The linear logistic regression model was adapted into the set of the data presented in Table 1. The results below are generated with the Generalized Linear Model function with the specification of the Binomial family distribution calculated in R with the code in (11):

```
model.1<glm(Admit~Points+Gender,family=bi
nomial, data=DATA)         (11)
```

The logistic regression model is:

$$\textit{Logit prediction}= -31{,}816 +0{,}007 *(\textit{Points}) + 0{,}023 *(\textit{Gender}) \qquad (12)$$

According to the model, the natural logarithm of the odds ratio that a candidate will be admitted to a branch is positively associated with both variables (total points and gender) with p = .000 and p = 0,935 respectively. In other words, the more points a candidate has accumulated, the greater will be the chance of being a winner.

### B. *Bayesian GLM*

To create the Bayesian GLM, it is necessary to load the package ‘arm’ ([9]) which contains the `bayesglm` function. It is conducted the Bayesian Generalized linear model with binomial family distribution and the code in R is (13):

```
model.2<bayesglm(Admit ~ Points + Gender,
family=binomial, data=DATA,
prior.scale=Inf,
prior.scale.for.intercept=Inf)      (13)
```

It is obtained the same logistic regression model as in (12):

$$\textit{Bayesglm prediction}= 31{,}816 +0{,}007 *(\textit{Points}) + 0{,}023 *(\textit{Gender}) \qquad (14)$$

As it is seen from the results, the model is the exactly the same with the traditional GLM model, but with an increment of the sample size the result should converge to the same values.

The `bayesglm` function is a sort of short cut of the Bayesian approach to the inferential analysis. Instead of using the posterior distribution to make inferences, it is build an empirical distribution based on the drawings from the posterior and it is this empirical distribution which conduct to the inference. In this way, `bayesglm` function gives the simulation distribution that will be used instead of the needed empirical distribution.

After the simulation of interception and the coefficient of the model, with the command '`head`' can be given the first 10 rows as below:

```
> simulates<-coef(sim(model.3))
> head(simulates,10)
      (Intercept)     Points        Gender
 [1,]  -36.56609  0.007637015  -0.31210627
 [2,]  -35.84152  0.007435138   0.29163831
 [3,]  -31.50487  0.006523644  -0.03241929
 [4,]  -36.82089  0.007696719  -0.26739455
 [5,]  -34.60980  0.007249084   0.17782494
 [6,]  -31.26711  0.006551134  -0.13684369
 [7,]  -35.96443  0.007454721   0.14924036
 [8,]  -28.60803  0.005958682   0.35853323
 [9,]  -32.53459  0.006792468   0.18446132
[10,]  -31.01768  0.006407207   0.86913280
```

The histogram of the posterior distribution for the coefficient of variable 'Points' is shown in Fig.2.

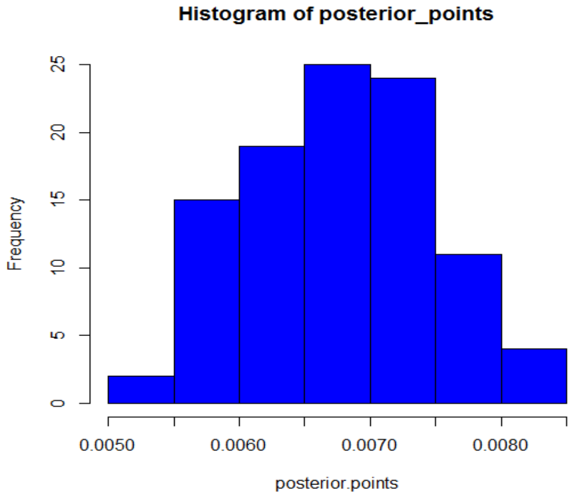

Fig. 2 The histogram of the coefficient of variable 'Points'

Actually, to display the posterior distribution of this coefficient is used '`density`' and the density plot can be considered as histogram with 100 bins. The density plot of the coefficient that is taken into consideration is shown in Fig 3.

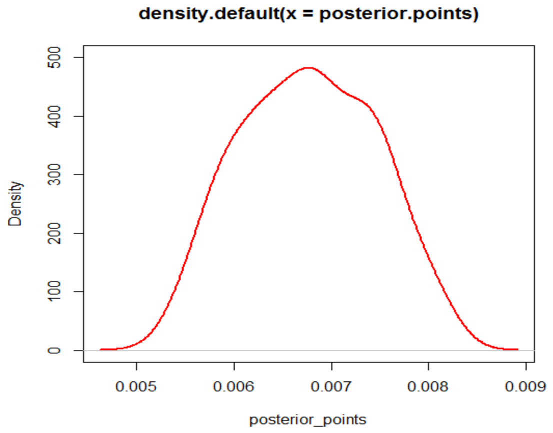

Fig. 3 The density plot of the coefficient of variable 'Points'

To make the Bayesian inference for all the coefficients of the model, we need to create the empirical distribution based on the draws from the posterior distribution. For this purpose, it is used the R package '`MCMCpack`' ([10]) and the function `logmcmc` as we have the logistic regression model. This package makes possible the creation of the empirical distribution by Markov Chain Monte Carlo simulation method, that is the mean, median or mode of the empirical MCMC simulates' distribution is the maximum likelihood estimate for one of the coefficients. The function `logmcmc` gives also the credible intervals for each coefficient.

The results after the simulations are given at table 4 for each coefficient of the regression model, including the credible intervals for the true coefficients value. So, we can say for example that the true value of the coefficient of the variable 'Points' is between -0.705 and 0.702 with 95 % probability.

TABLE 3
EMPIRICAL MEAN AND STANDARD DEVIATION OF EACH COEFFICIENT

| | **Mean** | **Standard deviation** | **2.5% quantile** | **97.5% quantile** |
|---|---|---|---|---|
| (intercept) | -32.685 | 3.620 | -40.293 | -26.089 |
| Points | 0.007 | 0.001 | 0.005 | 0.008 |
| Gender | 0.016 | 0.36 | -0.705 | 0.702 |

The credible intervals obtained for each coefficient are similar with the confidence intervals but they are really believed to contain the true value of the coefficient with a certain probability. As we are using a 95% credible interval to each coefficient this means that there is 95% probability that the coefficients are included in the credible intervals.

The logistic model is taken from the mean or other indicators of the empirical distribution of the coefficients after the simulations with MCMC.

## CONCLUSIONS

The maximum likelihood estimate, we look for the maximization point of the likelihood and the parameters are considered as fixed so we cannot inject our prior beliefs in the estimation calculations.

In contrast with this, Bayesian estimation fully calculates the posterior distribution $p(\theta|data)$ and the Bayesian inference treats the parameter as a random variable. So, when using Bayesian estimation we put in probability density functions and get out probability density functions, and not a single point as Maximum Likelihood Estimation.

What makes it useful is that it allows the usage of information or belief that we already have as prior distribution to calculate the probability density known as posterior density.

Another benefit of the Bayesian approach for any analysis is that it allows us to make credible intervals that are similar to the confidence intervals with the difference that in the Bayesian framework, is believed to contain the true value of the parameter. This is because that the credible interval is based on the posterior distribution of the parameter that is one of the coefficients of the logistic linear model.

## REFERENCES

[1] Bradley A. Carlin, Thomas A. Louis, Bayesian Methods for Data Analysis, third edition. Chapman & Hall/CRC Texts in Statistical Science, 2008.

[2] Christopher M. Bishop, Pattern Recognition and Machine Learning. Information Science and Statistics, Springer, 2011.

[3] J. Albert, *Bayesian Computation with R*. New York, Springer Science +Business Media, LLC, 2007.

[4] Donald A. Berry, *Statistics: A Bayesian perspective*. Belmont, California, Wadsworth Publishing Company, 1996.

[5] Benjamin M. Bolker, *Ecological Models and Data in R*. Princeton, New Jersey, Princeton University Press, 2008

[6] William M. Bolstad, *Introduction to Bayesian Statistics*. Hoboken, New Jersey, John Wiley & Sons, Inc, 2004.

[7] A. Gelman, J. B. Carlin, H. S. Stern & D.B Rubin, *Bayesian Data Analysis*, second edition. Boca Raton, Florida, Chapman & Hall/ CRC, 2004.

[8] A. Gelman, A. Jakulin, M. G. Pittau, & Y. Su, *A Weekly Informative Default Prior Distribution for Logistic and other Regression Models*. The Annals of Applied Statistics, 2(4), 1360- 1383, 2009.

[9] Andrew Gelman, Yu-Sung Su, Masanao Yajima, Jennifer Hill, Maria Grazia Pittau, Jouni Kerman, Tian Zheng & Vincent Dorie , *Package 'arm'*, 2018. [Online]. Available: https://CRAN.R-project.org/package=arm.

[10] Andrew D. Martin, Kevin M. Quinn, Jong Hee Park, Ghislain Vieilledent, Michael Malecki, Matthew Blackwell, Keith Poole, Craig Reed, Ben Goodrich, Ross Ihaka, The R Development Core Team, The R Foundation, Pierre L'Ecuyer, Makoto Matsumoto, Takuji Nishimura, *Package 'MCMCpack'*, 2018. [Online]. Available: https://CRAN.R-project.org/package=MCMCpack.

[11] Peter D. Hoff, A First Course in Bayesian Statistical Methods. New York, Springer Science +Business Media, LLC, 2009.

[12] J. A. Nelder, R. W. M. Wedderburn, Generalised linear Models. Royal Statistical Society, vol 135, nr. 3 (p 370-384), 1972.