African Journal of Applied Statistics
Vol. 4 (1), 2017, pages 219–229.
DOI: http://dx.doi.org/10.16929/ajas/2017.219.211

AJAS / SPAS
ISSN 2316-0861

Non-response process in Bayesian modeling and application to the INVA-ENPMO database.

CHELLAI Fatih[1*],

[1] High National School of Statistics and Applied Economics, (ENSSEA). Algeria.

Received July 10, 2017; November 1, 2017; Published Online November 15, 2017

Abstract. In this paper, we deal with the issue of missing data, in particular in the database (ENPMO) of the national survey of the shortage in agriculture workers in Algeria carried out by the Institut National de la Vulgarisation Agricole (INVA), where the main concern is the worker's assurance. We concluded with the existence a systematic process of non-response by the farmers which were surveyed, which indirectly violated the random distribution hypothesis of missing data. By using **R** and **Sphinx** softwares, we build a **logistic** model in the frame of the **Bayesian inference** to estimate the effect of three variables : *Age*, *Education level* and the *farm superficies*, which ares supposed have an impact of the non-response process.

Key words: Non-Response; Missing Data; Survey; Bayesian Analysis; Logistic modeling
AMS 2010 Mathematics Subject Classification : 62F15; 62DXX; 91G70.

Presented by Professor Gane Samb LO
University Gaston Berger, Saint-Louis (Sénégal)
Associate Professor, African University of Sciences and Technology, Abuja, Nigeria
Associated with LSTA, University UMPC, Paris VI (France)
Member of the Editors Board (Chief Editor).

[*] CHELLAI Fatih: abouchaima.1988@yahoo.com

Résumé. Dans cette étude, nous traitons de la problématique des données manquante, en relation avec la base de données (ENPMO) de l'Institut National de la Vulgarisation Agricole (INVA), récoltée après une enquête sur la pénurie de main d'œuvre agricole en Algérie. Après, dépouillement et analyse du volet de l'assurance de la main d'œuvre, nous avons constaté la présence d'un processus systématique de non-réponse des enquêtés (Agriculteurs), un résultat primaire qui viole l'hypothèse de distribution aléatoire de données manquantes. A l'aide des packages de logiciel **R**, et dans un cadre d'*inférence bayésienne*, nous avons construit un modèle **logit** afin d'estimer l'effet de trois variables : *l'Age, le niveau d'instruction de l'agriculteur et la superficie de l'exploitation agricole* sur les données manquantes.

1. Introduction and Background

Observations with missing values represent an important challenge, because conventional modeling procedures generally eliminate these observations from the analysis. Ibm.com (2001). The presence of a high rate of missing data may lead to erroneous results; which complicates the procedure for verifying research hypotheses. This major problem is recurrent in almost studies and researches, especially in the Social Sciences as in cite15. In the Political Sciences field, Gary *et al* (2001) estimate that almost 50% of respondents in opinion polls in the electoral elections have missing data. This phenomenon is also remarkable in large-scale public surveys and censuses. Our *objective*, in this study is not the computation methods of the missing data, as in Little *et al*(1987), Anderson *et al* (1983), Ibm.com (2001). Rather, we will have to check the validity of the hypothesis of the violation of the random distribution of the missing data (**MCAR**). More precisely, the question we have to give a clear answer is the following : ***is there an effect of the variables of the survey units on the missing data?***.

In the practical part of this study, we handle data from a national survey on the agricultural labor shortage named in French as Enquête Nationale sur la Pénurie de la main d'œuvre Agricole, (ENPMO)(2016) and carried out by the *INVA*. After a counting and exploratory analysis, we found that a clear systematic process of **non-response** of the respondents about the question on workers's social conver, spelled as follows : *Do you insure your workers?*. After a first step of descriptive analysis Matching Analysis (**ACP**) and Classification methods, we have identified an effect (*or of a relation*) between the answers to this question and three characteristics of the units surveyed which are: **Age** of agriculture, **instruction level** and **area of the operation**.

In order to quantify this relationship, we are going to use a statistical modeling based on the **regressive logistic model** within the **Bayesian approach**, in which the dependent variable δ_i is binary): $\delta_i = 0$, if an answer is given and $\delta_i = 1$ otherwise , and the explanatory variables are denoted by: X_{ik} to $k \in \{1, 2, 3\}$. To achieve the objectives of our study, it is divided into two sections: the next section presents a brief literature review of the topic studied, then it develops the regressive logistic model, in the second section and by using software **R**, we presented and interpreted the estimation results of the model. finally, the conclusion of this study.

2. Non-Response Process and Covariate Effect: Bayesian Logit Model Estimation

2.1. Literature review

It is common when we analyze the survey data, that the explained variable takes only a finite number of modalities. For example, one may want to model, for example,

- The electoral behaviors : $\delta_i = 1$ if the voter i voted for the candidate **X**, $\delta_i = 0$ if he voted for the candidate **Y**.
- The Scoring credits : $\delta_i = 1$ if customer i repaid a loan he contracted, $\delta_i = 0$ was unable to repay it.
- The non-answer to a question in a survey : $\delta_i = 1$ if the individual i refused to answer, $\delta_i = 0$ if he replied.

In the literature, several theoretical, practical methods and techniques have been developed. Here, we restrict ourselves to studies and methods which are conducted and carried out in the Bayesian modeling of logistic models. The first study was realized by James *et al* (1993). They developed a Bayesian estimation technique for **probit** models, where the dependent variable takes two modalities $k = 2$ so for $k > 2$, and applied their study to examples : the first concerned ata from a clinical trial and was about modeling biological factors with two outcomes : *onset* or *no* relatively to a blood disease, see Finney (1947); the second example used data from a survey on the prediction of the results of the USA presidential elections in **1976** on the basis of six (6) socioeconomic and geographic characteristics factors of the respondents. In another study, Tsai (2004)e applied the Bayesian logistic model to estimate and predict the results of the elections in **Taïwan** or **Taipei**, where two candidates were involved.

Starting from different articles and books dealing with logistic regression in a Bayesian Inference, and inspired the methodological steps by these two studies; an attempt was made to apply the Bayesian logistic model to a set of real data from a survey of agricultural labor shortages in Algeria.

2.2. Principle of Bayesian Inference

Suppose we have a sample $Y = (Y_1, Y_2, ..., Y_n)$ of size $n \geq 2$, meaning that Y_1, Y_2, ...,Y_n are independent random variables defined on the same probabilistic space $(\Omega, \mathcal{A}, \mathcal{P})$, and identically distributed probability law depending on a parameter $\theta \in \Theta \subset \mathbb{R}$. In the Bauesian setting, the parameter is supposed to be a random variable with a *prior* and an *posteriori* distribution. We try to estimate these θ and this from the observations $y = (y_1, y_1, ..., y_n\ )$ and also from information called *a prior* on the parameter θ. Bayesian inference assumes that the following quantities are known:

- The *prior density* of the parameters Θ denoted $\pi(\theta)$ which summarizes the information available on the vector of parameters θ.
- The conditional density $f(y_i/\theta)$ of the random variable Y_i on θ.

From the Bayes theorem we deduce the posterior distribution of the parameters Θ:

$$\mathbb{P}(\theta \setminus y) = \frac{\mathbb{P}(y \setminus \theta)\,\mathbb{P}(\theta)}{\mathbb{P}(y)},$$

with

$$\mathbb{P}(y \setminus \theta) = \textstyle\prod_{i=1}^{n} \mathbb{P}(y_i \setminus \theta)$$

$$\mathbb{P}(y) = \textstyle\int \mathbb{P}(y \setminus \theta)\mathbb{P}(\theta) d\theta.$$

The Bayesian estimators vector of Θ parameters are derived from the posterior distribution: $\mathbb{P}(\theta \setminus y)$ at the basis of the optimization (*Minimization*) of an appropriate loss function. The most common application functions are:

1. **The quadratic loss function**, whose estimators are the means of the posterior vector law of the parameters θ; these estimators denoted $\hat{\theta}_{MMSE}(y)$, minimize the function:

$$l\big(\hat{\theta}(y)\big) = \mathbb{E}\big[(\hat{\theta}(y) - \theta)^2\big].$$

2. **The absolute loss function**, for the same estimation procedures, the vector of the estimated θ parameters is the median of the posterior distribution that minimize function:

$$l\big(\hat{\theta}(y)\big) = |\hat{\theta}(y) - \theta|.$$

2.3. Bayesian Estimation of Logit Model

Knowing the factors likely to influence the behavior "*response* , *non-response*", these influences can be modeled using the econometrics techniques of qualitative (dummies) variables (see Gourrieroux (1984) and Christophe (2003)). Mathematically, we denote :

$$\delta_i = \begin{cases} 1 & \text{if , non-response} \\ 0 & \text{if , response} \end{cases}$$

The main interest of coding (or the quantitative representation of this qualitative variable) is that it can be reduced to discrete distributions on $\mathbb{R}^2$. Indeed, this writing allows us to define the probability of occurrence of the event ("*non-response*") as the expectation of the coded variable δ_i, since :

$$\mathbb{E}(\delta_i) = \mathbb{P}(\delta_i = 1) \times 1 + \mathbb{P}(\delta_i = 0) \times 0 = \mathbb{P}(\delta_i = 1) = p_i.$$

That is, there exists a relation between the explanatory variables X_{ik} and the dependent variable p_i. Therefore, a natural approach to presenting this relationship is *regression*, which allows us to quantify the effect of these variables on p_i, and even to predict future values of this explained variable given the X_{ik}'s. Since p_i are probabilities, it does not make sense to apply the linear regression methods. On the other hand, this relation could be modeled with a *Link Function* $\mathcal{H}$ which satisfies: $p_i = \mathcal{H}(X\beta^t)$. At this level, $\mathcal{H}$ is a distribution function, and β is a vector of unknown parameters to be estimated. If $\mathcal{H}$ is invertible, this gives: $\mathcal{H}^{-1}(p_i) = (\beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik})$. If the distribution function of a *standard normal law* has been imposed on $\mathcal{H}$, the relation is called a **Probit** model. It is called a **Logit** model, if $\mathcal{H}(x) = (\frac{e^{-x}}{1+e^{-x}})$ is a the Logistic distribution.

By another mathematical writing, and for the purpose of statistical modeling, we can write: $(\delta_i \setminus X) = X\,\beta + \epsilon$, we have assumed ϵ follows a logistic distribution with mean : $\mu_\epsilon = 0$ and variance $\sigma^2_\epsilon = \frac{p_i^2}{3}$. Therefore, the distribution of δ_i is given as follows:

$$\mathbf{z_i} = \ \mathbf{P}(\delta_i = 1 \setminus X_{ik}) = \frac{1}{1 + \exp(X\,\beta)}. \tag{1}$$

2.3.1. The Likelihood of the Model

To estimate logistic regression parameters using the maximum likelihood method, we first need to determine the distribution law of $\mathbf{P}(\delta_i \setminus X_{ik})$, We know that δ_i is a binary random variable. For an individual **i**n we model the probability using the binomial $\beta(1, z_i)$, which gives:

$$\mathbf{P}(\delta_i \setminus X_{ik}) = z_i^{\delta_i}\,(1 - z_i)^{1-\delta_i}. \tag{2}$$

The generalization to a sample of size **n** is straightforward. Denoting $\boldsymbol{X_i} = (1, X_{i1}, ..., X_{ik})$, $i = 1, ..., n$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_k)$, we get the standard form of the likelihood of the model as

$$\begin{aligned}\mathcal{L}(\beta; X_i, \delta_i) &= \prod_{i=1}^{n} \left[z_i\right]^{\delta_i} \left[1 - z_i\right]^{1-\delta_i} \\ &= \prod_{i=1}^{n} \Big(\frac{e^{\beta^t\,X_i}}{1 + e^{\beta^t\,X_i}}\Big)^{\delta_i} \Big(\frac{1}{1 + e^{\beta^t\,X_i}}\Big)^{1-\delta_i} \\ &= \prod_{i=1}^{n} \frac{e^{\beta^t\,X_i}\delta_i}{1 + e^{\beta^t\,X_i}}.\end{aligned} \tag{3}$$

The **log-likelihood** is given by:

$$\log\,\mathcal{L}(\beta; X_i, \delta_i) = \sum_{i=1}^{n} \beta^t X_i \delta_i - \sum_{i=1}^{n} log\big(1 + e^{\beta^t\,X_i}\big). \tag{4}$$

These two functions must be maximized with respect to the vector of $\boldsymbol{\beta}$ parameters.

2.3.2. Bayesian Estimation of the Model

According to **the Bayesian Approach**, we have to optimize our estimator of β if we have *a prior information* on them, since in this approach (β) is a random variable. We denote $\pi(\beta)$ the *prior* density of β. The *posterior* density is defined as follows:

$$\pi(\beta \setminus X_i, \delta_i) \propto \pi(\beta) \prod_{\mathbf{i=1}}^{\mathbf{n}}\ \mathcal{H}(\beta^t x_i)^{\delta_i}(1 - \mathcal{H}(\beta^t x_i))^{1-\delta_i}. \tag{5}$$

Then, the major problem and the most credible point in the Bayesian Approach is the choice of the *prior distribution*. However, in some situations, we require a partially automated determination of *prior* distribution, like in the case here where prior information is completely absent. This is the case in our study, because it is the first survey carried out on the subject on the agricultural workforce, and this, the quantification of a prior information is difficult even impossible. The practical solution in this case is the use of non-informative *prior* distributions. (See Robert (2006) for more mathematical details on such modeling).

Another method of specifying the *prior* distribution is the use of parametric distributions (Gamma, Beta, Normal, etc.) which are usually called *conjugate distributions*. After our different hypotheses on the nature of the prior $\pi(\theta)$, the *posterior* distribution takes the following form:

$$\mathbf{P}(\beta \setminus X\,,\delta_i) \propto \ (\sigma)^{-n-2}\ \exp[\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(\delta_i - x_i\beta)^2].$$

For the estimation computations, we used the software **R** (R Software (2009)) and other packages including the INLA one.

3. Results and Discussion

3.1. Description of database

The data we are using in this applied statistics research is part of a national survey on the shortage of agricultural labor named in French as Enquête Nationale sur la Pénurie de la main d'œuvre Agricole, (ENPMO)(2016), which has been carried out by the National Institute of Agricultural Extension (**INVA**) during the period **2016-2017**. The final report of the results is not yet published, so that we are not able to cite it in this paper. The current available, in particular the tables and graphs summarizing such results, are outcomes of the software **Sphinx** under a license assigned to the *INVA*. To avoid any conflict of interest with the **INVA** and **Sphinx**, the Institute granted us a permission to use this database in our researches, provided that the final report be published. We are not allowed to go beyond the aforementioned description of the objectives and of the variables, due to the company **Sphinx** restriction.

The objective of the National Agricultural Worker Shortage Survey is to highlight the socio-economic and environmental factors that prevent (or demotivate) young people from working in the agricultural sector. The target population at the beginning was only young people. Afterwards, it was open to members if the agricultural sector professionals. Additional questions on the availability of labor at the level of their farms, especially on the potential *potential practical causes that would prevent young people from working in agriculture.*
It is that part of the Survey which concerns this paper. Specifically, we are interested in questions on the social security of the workforce working on these farms. We observed that the non-response rate to the question *do you ensure your workforce?*, was to too high, around **35 %**. From this remark, we proceeded to an analysis in which we applied the

applications of the methods of Analysis of correspondence (**ACP**) and the Discriminant Factorial Analysis (**AFD**) and other methods of classification. We suspected that, probably, there is an effect of some variables on the response or non-response on this question. We have considered three variables: **1** the **Age** of agriculture which is divided into four classes [**20-30**], [**31-40 41-50**] and [**+ 50 years**]. **2)**. The level of farming was also divided into three classes: *Primary*, *Middle*, *secondary or higher*.

We re-collected the data on a two-stratification scheme :

(1) In the first stage, the stratification level is the type of agricultural activity (Cereale, vegetable, ...), according to which the *wilayas* (areas) have been classified.

(2) In each member of this class, pilot wilayas were selected.

A sample of size **N** = **700** has been collected distributed over **15 Wilayas** (areas). We will describe the outcomes of ur study in the next section.

3.2. Results of Bayesian Logit estimation

For the application we used the Integrated Nested Laplace Approximation (INLA) packages, which is indeed a complete package for Bayesian inference (See R Software (2009)). The generalized standardized form of the underlying commands is defined as:

```
Inla ( Assurance 1 +as.factor(Age)+as.numeric(SEXp)+as.factor(NivInstr), family = "logistic", contrasts = NULL, data=data, quantiles=c(0.025, 0.5, 0.975),E = NULL, offset=NULL, scale = NULL, weights = NULL, Ntrials = NULL, strata = NULL,link.covariates = NULL, verbose = FALSE, lincomb = NULL,control.compute = list(), control.predictor = list(), control.family = list(), control.inla = list(), control.results = list(), control.fixed = list(), control.mode = list(), control.expert = list(), control.hazard = list(), control.lincomb = list(), control.update = list(), only.hyperparam = FALSE, inla.call = inla.getOption("inla.call"), inla.arg = inla.getOption("inla.arg"), num.threads = inla.getOption("num.threads"), keep = inla.getOption("keep"), working.directory = inla.getOption("working.directory"), silent = inla.getOption("silent"), debug = inla.getOption("debug"), .parent.frame = parent.frame() )
```

The commands used in our case are the following:

```
library(INLA)

library(sp)
library(Matrix)
data = read.table("enss.csv",sep= ";", h=T); attach(data)
model = inla(Assurance 1 +as.factor(Age)+as.numeric(SEXp)+as.factor(NivInstr), data = data, family
= "logistic",control.compute = list(cpo=TRUE))
summary(model)
```

```
plot(model,col=5)
```

The outcomes of our computations according the the chosen model are summarized in the table below. The vector of estimated parameters $\hat{\beta}$ is of dimension $1\times$ 9. Of course we proceeded to the discretization of the categorical variables (**Age** and **Instruction level**, using binary encoding of all modalities $(i.e)$, takes as values **1** or **0**.

Variables	Coefficient	E.T (*)	0.025 Q	0.5 Q	0.975 Q	mode
Constante	-0.2471	0.4735	-1.1725	-0.2487	0.6864	-0.2520
[**20-30**]	0.2699	0.4231	-0.5608	0.2698	1.1005	0.2696
[**30-40**]	0.3380	0.4112	-0.4695	0.3379	1.1449	0.3379
[**40-50**]	0.3877	0.4098	-0.4171	0.3877	1.1918	0.3876
[+ **50 ans**]	0.5477	0.4095	-0.2566	0.5477	1.3511	0.5478
SupExploi	**-0.0003**	0.0004	-0.0010	-0.0003	0.0005	-0.0003
Primaire	-0.0494	0.2434	-0.5451	-0.0431	0.4117	-0.0303
Moyen	0.0737	0.2447	-0.4242	0.0799	0.5375	0.0924
Secondaire +	0.2635	0.2442	-0.2335	0.2697	0.7263	0.2822

Table 1. Posterior Estimation of parameters vector.(*): Standard deviation $\hat{\sigma}_{\hat{\beta}}$, Quantiles and Mode of $\hat{\beta}$.

On the basis of these results,

(a) the first conclusion to be drawn is that the non-response effect to the labor insurance question - despite being low - is proportional to the farmer's age. To .. this, we may have a look at the coefficients in blue : their is a positive nonresponse trend based on the farmer's age.

(b) In terms of odds-ratio, the class of age [+ **50 years**] has the maximum non-response risk with $= \exp(0.55) = 1.73$.

(c) For the area of the farm, this variable with an estimated coefficient $\hat{\beta} = -0.0003$, has no effect on whether or not to answer the questio . *This, even indirectly, negates an idea circulating in the agricultural sector in Algeria, according to which: small-scale farming does not offer social coverage to its labor force..*

(d) There is a low effect of the farmer's level of education on the non-response process, with a coefficient *hat beta* $= 0.07$ average level, a coefficient *hat beta* $= 0.26$ of secondary level and +, and an almost null effect of primary level.

For the various graphs of *posterior* distributions of the estimated parameters and the adjusted model graph, see the figures below, 4.

As a conclusion of our work, we can say :

(A) that we have confirmed the hypothesis of a relation between the non-response process on the issue of labor insurance and the explanatory variables: **Age** and **education level** of the farmer;

(B) that there is no significant effect recorded for the **area of the farm**.

We think this work has two major advantages.

(I) It was reported that decision-makers and practitioners in the field of surveys and opinion polls emphasized the importance of *non-response* and *factors likely to influence the respondents to not answer a given question.*

(II) We have identified a standard logistic model based on *Bayesian inference* that could be applied in different domains, especially, if we want to understand the effect of individual characteristics on opinion and /or response on a question or even on a series of questions.

Ackonwledgement The author expresses his thanks to the *Institut National de la Vulgarisation de l'Agriculture de la République d'Algérie* (INVA) for permission of partial use of part of the (ENPMO, 2016) database. He also feels indebted to the presenter for a constant assistance in better presenting the paper and for valuable comments on the methodology.

References

Little, R.J.A., Rubin, D.B. (1987). *Statistical Analysis with Missing Data.* New York: John Wiley-Sons.

Anderson, A.B., Basilevsky, A. Hum, D.P. (1983). *Missing data: A review of the literature.* In J.D. W.P.H. Rossi A.B. Anderson (Eds.), Handbook of survey research. New York: Academic Press.

www.ibm.com/spss.(2011). IBM SPSS Missing Values 20. A file attached in SPSS softawr.

Robert C. (2006). *Le choix bayésien Principes et pratique* , Springer-Verlag, Paris, France.

Gary K. (2003).*Bayesian Econometrics*, John Wiley - Sons Ltd, England.

R Development Core Team (2009). *A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing. Vienna, Austria. ISBN : 3-900051-07-0, http://www.R-project.org.

Le Sphinx iQ. Logiciel pour enquêtes quantitatives et qualitatives : Manuel de référence, *www.lesphinx.eu.*Visited 12/12/2016.

Gourrieroux C. (1984). *Économétrie des variables qualitatives*, Économica.

Christophe H.(2003). Économétrie des Variables Qualitatives : Modèles Dichotomiques Univariés, Polycopié de Cours, *www.univ-orleans.fr/deg/masters/ESA/.* Visited 02/01/2017.

Ghosh J.K, Delampady M.(2006). *An Introduction to Bayesian Analysis Theory and Methods*,Springer.

James H.A and Siddhartha.C .1993. Bayesian Analysis of Binay and Polytchotomous Response Data, *American Statistical Association Journal*.June 1993, Vol.88,No, 422.

Finney D.J. (1947). The Estimation from Individual Records of the Relationship between Dose and Quantal Response, *Biometrika*,No 34.

Tsai C.H. (2004). Bayesian Inference in Binomial Logistic Regression : A Case Study of the 2002 Taipei Mayoral Election,*Election Study Center*,National Chengchi University.

Gary K. and James H. (2001), Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation. *American Political Science Review*. Vol. 95, No. 1.

Juster, T., Smith, J. P. (1997). Improving the quality of economic data: Lessons from the HRS and AHEAD. *Journal of the American Statistical Association*, 92, 1268 –1278.

(2016). **Enquête Nationale sur la Pénurie de la main d'œuvre Agricole, (ENPMO)**, 2016-2017, **L'Institut National de la Vulgarisation Agricole (INVA)**, Algérie. website : www.inva.dz

4. Appendix

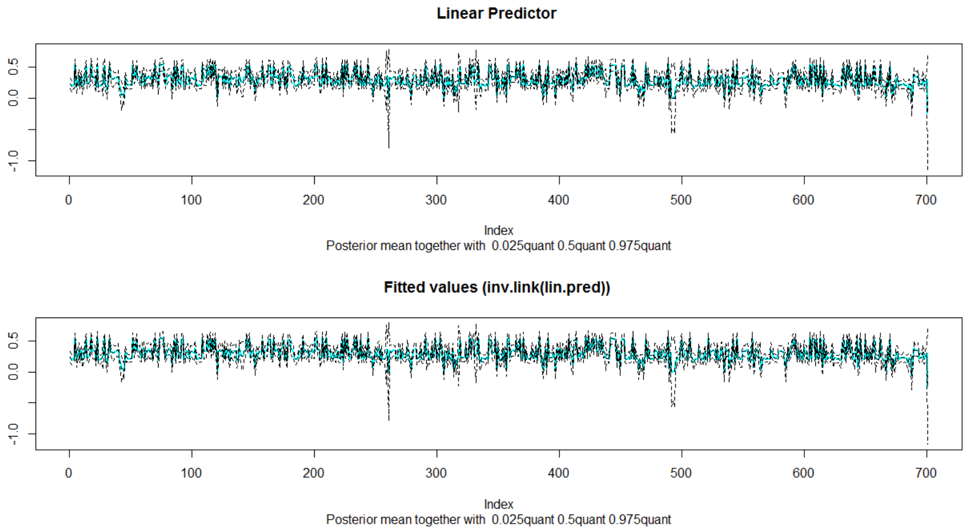

Fig. 1. Averages values of *a posteriori* probabilities $\mathbf{P}(\delta_i = 1 \setminus X_i, \hat{\beta})$ of the logistic function

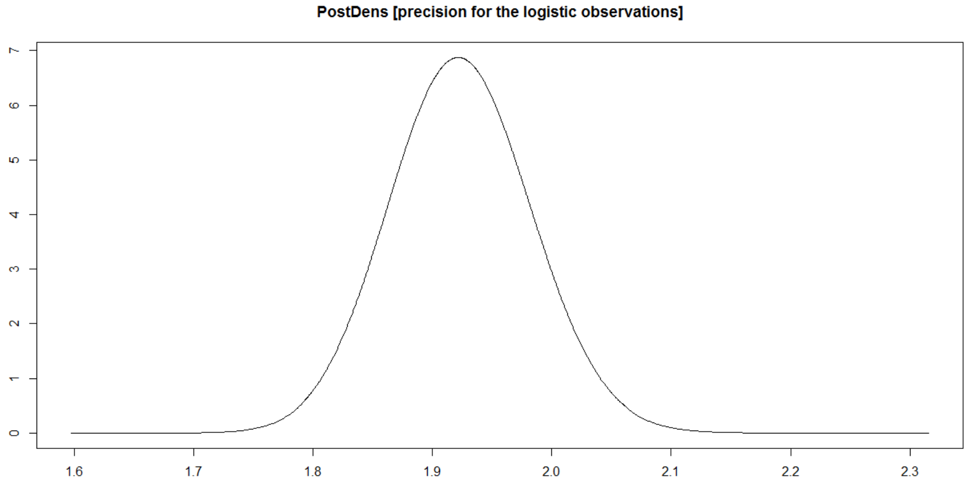

Fig. 2. Graph of *posterior distribution* of accuracy (i.e) $\frac{1}{\hat{\sigma}}$

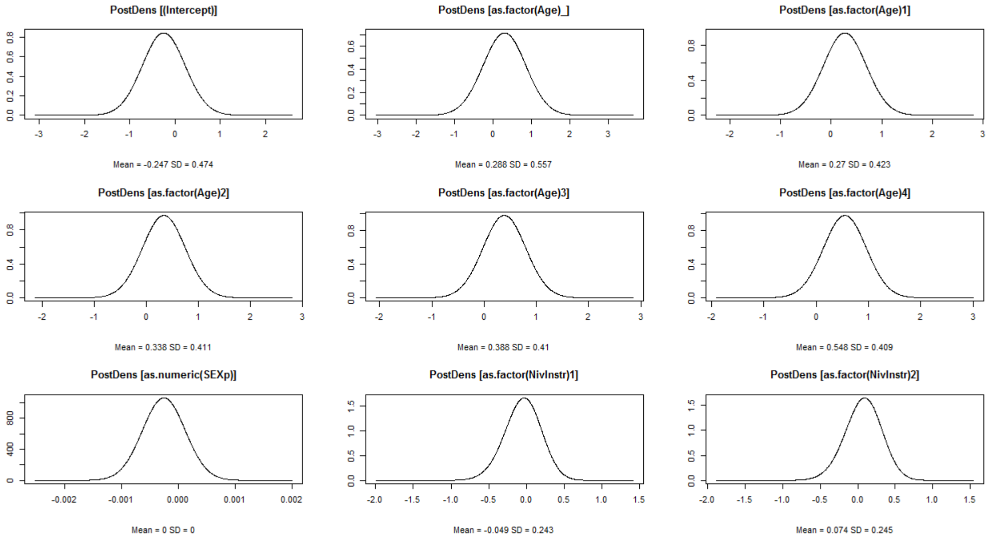

Fig. 3. The graph of the *posterior* estimated beta distributions $\hat{\beta}$

.