
SELF-SELECTING ROBUST LOGISTIC REGRESSION MODEL

GBOHOUNME IDELPHONSE LEANDRE TAWANOU

MS 300-0002/15

**A Thesis submitted to Pan African University Institute for Basic
Sciences, Technology and Innovation in partial fulfillment of the
requirement for the award of the degree of Master of Science in
Mathematics-Statistics Option**

2017

DECLARATION

This research thesis is my own work and has not been presented elsewhere for a degree award.

Signature Date.....

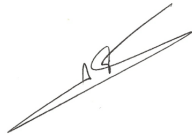
Idelphonse Léandre Tawanou GBOHOUNME

Declaration by supervisors.

This research thesis has been submitted for examination with our approval as university supervisors.

Signature Date.....

Dr. Oscar NGESA,
Taita Taveta University, KENYA



Signature Date.....

Dr. Jude EGGOH,
Angers University, FRANCE

ACKNOWLEDGMENT

The completion of this project has been possible through the help of many individuals who gave their support. I thank the almighty God for granting me the knowledge to carry out this study. I can't overlook the love and support from my family and friends who encouraged me throughout the research.

I wish to gratefully acknowledge financial support from the African Union Commission. I am also grateful to my supervisors, all the lecturers and entire staff of PAUSTI.

TABLE OF CONTENT

Declaration	i
Acknowledgment	ii
List of Tables	v
List of Figures	vi
Acronyms	vii
Abstract	viii
1 INTRODUCTION	1
1.1 Background of the Study	1
1.2 Statement of the Problem	3
1.3 Justification	4
1.4 Objectives of the Study	4
1.4.1 General Objective	4
1.4.2 Specific Objectives	5
1.5 Significance of the Study	5
1.6 Organization of the Study	5
2 LITERATURE REVIEW	6
2.1 Generalized Linear Model	6
2.2 Binary Logistic Regression	8
2.2.1 Mathematical Principles and Properties	8
2.3 Tobit Model	10
2.4 Probit Model	13
2.5 Robust Logistic Regression	14
2.5.1 Loss Function	15
2.5.2 M-Estimator	16
2.5.3 Trimming Approach	18
2.6 Estimation using Bayesian Approach	20
2.6.1 Bayesian philosophy	21
2.6.2 Bayesian Estimation	22
2.6.3 Gibbs Sampler	23
2.6.4 Principle of Gibbs Sampler	23
2.6.5 Metropolis-Hastings Algorithm	25
3 METHODOLOGY	27

3.1	Proposed Model	27
3.2	Parameters Estimation	28
4	SIMULATION RESULTS	30
4.1	Methodology	30
4.1.1	Introduction	30
4.1.2	Simulation Set up	30
4.1.3	Model diagnostics	32
4.2	Results	33
4.2.1	Introduction	33
4.2.2	Model assessment and Comparison	33
4.2.3	Discussion	37
5	CONCLUSION AND RECOMMENDATION	41
5.1	Conclusions	41
5.2	Recommendations	43
	REFERENCE	44
	APPENDIX	47

LIST OF TABLES

4.1	Description of variables X and assumed values of parameters manipulated in simulation	39
4.2	Simulated results of all models for Data with Leverage Points(0% and 3%)	40
4.3	Simulated results of all models for Data with Leverage Points (5% and 7%)	40

LIST OF FIGURES

4.1	Histogram of X_1	34
4.2	Histogram of X_1 with outliers	35
4.3	Histogram of X_2	36
4.4	Histogram of X_2 with outliers	37
4.5	Relationship between Response Variable and Endogenous Variable X with outliers	38

ABBREVIATIONS AND ACRONYMS

AIC	Akaike information criterion
BIC	Bayesian information criterion
DIC	Deviance information criterion
GRLR	Gelman robust logistics regression
LAD	Least absolute deviation
LR	Logistic regression
McMC	Markov chain Monte Carlo
MAD	Median absolute deviation
MLE	Maximum likelihood estimation
RLR	Restricted logistic regression
S	Loss function
SsRLR	Self selecting robust logistics regression
WEMEL	Weighted maximum estimated likelihood

ABSTRACT

Binary data is a common response data in many fields of research including finance, social sciences, psychology and medicine. The most common model used for the analysis of binary data is the logistic regression model. However, the problem of identification and corresponding treatment of influential outliers still remains to be well studied to check the adequacy of the fitted binary logistic models. Many researchers have developed robust statistical model to solve this problem related to the presence of atypical observations in the data. Gelman (2004) proposed a model that dealt with outliers problem by trimming the probability of success in logistic regression. The trimming values in this model are fixed and the user is required to specify this value well in advance. We explore this work and other robust logistic regression models then extend this work to allow for the trimming value to be estimated from the data. In particular, this research work presents a self selecting robust logistic regression (SsRLR) model. We proved that the SsRLR model is more robust to the presence of leverage points in the data. Parameter estimations is done using a full Bayesian approach, implemented in WinBUGS 14 software.

Chapter 1

INTRODUCTION

1.1 Background of the Study

Many dependent variables of interest in the field of social sciences are usually not continuous variables. In most cases, the outcomes are categorical with two levels, namely, yes/no, success/failure, 0/1. Such variables are called binary or dichotomous. Binary logistic regression is a helpful way of explaining the relationship between one or more independent variables and a binary response. The most attractive characteristic of a logistic regression model is that it neither assumes the linearity in the association between the independent and the outcome variable, nor necessitates normally distributed variables (Pregibon, 1981). It does not assume homoscedasticity as well and generally admits less rigorous constraint than linear regression models.

Most of the works linked to the logistic regression occurs in the experimental epidemiological study but in the past decade it has also been utilized in observational studies. Studies of residuals and the identification of outliers and influential cases are not carried out so regularly to examine the suitability of the fitted model. Data arising from observational studies sometimes can be thought of as bad from the point of view of outlying responses. The ordinary method of implementing logistic regression models with maximum likelihood, has good optimality properties in ideal settings, but is very sensitive to bad data obtained from observational studies (Pregibon, 1981).

Often in logistic regression analysis applications, the real data set contains outliers; the observations for these cases are well separated from the rest of the data.

These outlying cases may produce residuals that often have dramatic influence on the fitted maximum likelihood linear predictor (Hilbe, 2009). Therefore, it is essential to study the outlying situation sensibly and choose whether they should be kept or removed, and if kept, whether their effect should be diminished in the fitting procedure and/or the logistic regression model should be reviewed ((Menard, 2002) and (Hosmer & Lemeshow, 2000)).

There exist three ways that an observation can be thought of as unusual, namely, outlier, influential and leverage (Sarkar & Rana, 2011). In logistic regression, a set of observations whose values diverge from the expected range and produce residuals and may signify a samples particularity are called outliers. These outliers can excessively influence the results of the analysis and lead to wrong inferences (Jennings, 1986).

Detection of outliers and consistent treatment are a very important task of any modelling exercise, failure of which can lead to severe distortion of the validity of the inferences drawn from such modeling exercise (Sarkar & Rana, 2011). Therefore the use of robust logistic regression methods appears necessary and reasonable in order to obtain reliable estimates of parameters (Ritschard, 1990). These regression methods derive their strength from various tools to help detecting outliers, reduce their influences or to remove them from the data set (Maronna & Yohai, 2000).

Robust statistics, which have experienced a very important growth over the last fifteen years, are developed to eliminate adverse effects that may be experienced due to the presence of atypical observations in the data. The robustness is obtained through mechanisms that automatically reduce or discontinue the importance of atypical data in the estimation. In disregarding the implementation costs, it is generally a matter of choice between degree of robustness and efficiency (Ritschard,

1990).

In the context of regression, a range of robust methods have been developed, among them, the trimming robust approach (Gordaliza, 1991). This robust technique allows one to handle a proportion, α of contaminating data to guarantee the robustness of parameter estimation. The research work extends the literature by elucidating a part of some earlier work in robust estimation. This project is focused on the work of Gelman (2004) related to the trimming approach.

1.2 Statement of the Problem

Robustness is a subject highly developed in the fields of estimation of the position and scale of simple and multiple regression. Attention has been paid to the robust logistic regression, which is an area where outliers may also appear. Pregibon (1981) started by developing an analytical measure to assist in the detection of outliers and leverage points and quantify their effect on diverse aspects of the maximum likelihood fit. Thereafter, a good number of robust estimation procedures in the context of logistic regression have been examined. Gelman (2004) proposed a robust logistic regression model using a trimming approach. This approach used a trimming value, 0.01 which gives the chance of random error in both direction in the interval $[0, 1]$. Therefore classical logistic regression and Gelman's Robust Logistic Regression model (GRLR) (the linear predictor part only) respectively, are given by,

$$\pi = \text{logit}^{-1}(X^T \beta) \tag{1.1}$$

$$\pi = 0.01 + 0.98 \text{logit}^{-1}(X^T \beta), \tag{1.2}$$

where 0.01 and 0.98 are fixed. The model requires that the statistician specify these values beforehand. SsRLR model solves this problem in the GRLR model by relax-

ing this restriction and letting these probabilities to be self selected by the data at hand so that only a prior distribution, say Uniform $[a, b]$ with a and b belonging to $[0, 1]$ is given. A full Bayesian approach in parameter estimation given their prior distribution is used.

1.3 Justification

As more and more social scientists, scholars, practitioners and students are interested in explaining and predicting phenomena that can be characterized by a binary variable (Kateri & Agresti, 2010), it is apparent that research has to be carried out in details to properly deal with binary data with outlier cases. There is, therefore, need to develop robust model that is easy to understand.

The results of this work will contribute a lot to enhancing the method of selecting the probability values in GRLR model, obtaining a reliable fitted logistic model and will motivate more use of logistic regression with the presence of outliers in different areas such as epidemiology, social sciences, psychology, where logistic regression is commonly used.

1.4 Objectives of the Study

1.4.1 General Objective

The general objective is to develop a robust regression model for binary response data.

1.4.2 Specific Objectives

1. To use a Bayesian approach to select the trimming probability in logistic regression.
2. To develop a self-selecting robust logistic regression model.
3. To improve the robustness of Gelman's robust logistic regression model.

1.5 Significance of the Study

This study will provide more understanding on robust binary model. With the self-selecting robust logistic regression, we will offer a theoretical result, that will help practitioners in handling outliers in logistic regression. The findings from this research will enable researchers to implement the robust logistic regression model without the trouble of having to specify the trimming probability beforehand.

1.6 Organization of the Study

The rest of this work is organized as follows: chapter two presents a review of literature relating to our research objectives by describing from the generalized linear models to robust logistic regression. In chapter three, we discuss the methodology in which we develop the model for handling logistic regression with outliers and give a detailed procedure of the estimation of the robust logistic model. Simulation study is carried out in chapter four while the last chapter offers conclusions and suggestions based on our research.

Chapter 2

LITERATURE REVIEW

2.1 Generalized Linear Model

The generalized linear model originally developed by Nelder & Wedderburn (1972), presented in great detail in Enderlein (1987) and Antoniadis (1992), is used when the distribution of the error is not normal. Models catalogued in the class of generalized linear models are characterized by three components as discussed below.

The random component identifies the probability distribution of the explanatory variable. We assume that the statistical sample consists of n random variables $\{Y_i; i = 1, \dots, n\}$ independent admitting distributions from an exponential structure. This means that the laws of these variables are dominated by one measure called reference and the family of their densities relative to this measurement is given by

$$f(y_i, \theta_i, \phi) = \exp \left\{ \frac{y_i \theta_i - v(\theta_i)}{u(\phi)} + w(y_i, \phi) \right\}. \quad (2.1)$$

This formulation includes most usual distributions having one or two parameters, namely, Gaussian, inverse Gaussian, Gamma, Poisson, Binomial. In this notation, θ_i is the natural parameter of the exponential family (Antoniadis, 1992). For some distribution, the function u takes the form:

$$u(\phi) = \frac{\phi}{\omega_i}, \quad (2.2)$$

where ω_i are the known weights of observations, fixed here to 1 for simplicity and ϕ is the dispersion parameter. This is a nuisance parameter arising, for instance when the variances of the Gaussian distributions are unknown, but equal to 1 for distributions of single parameter (Poisson, Binomial). The exponential structure in

equation (2.1) can be expressed in canonical form by letting

$$Q(\theta) = \frac{\theta}{\phi} \tag{2.3}$$

$$a(\theta) = \exp \left\{ -\frac{v(\theta)}{\phi} \right\}$$

$$b(y) = \exp \{w(y, \theta)\}. \tag{2.4}$$

From this we obtain,

$$f(y_i, \theta_i) = a(\theta_i)b(y_i) \exp \{y_i Q(\theta_i)\}. \tag{2.5}$$

The planned observations of explanatory variables are organized in the matrix of design X . Let β be a vector of p parameters such that the linear predictor, deterministic component model, is a vector with n components,

$$\eta_i = \beta_0 + \beta_1 X_{i1} + \dots + \beta_p X_{ip}; i = 1, \dots, n. \tag{2.6}$$

The third component of the generalized linear models expresses a functional relationship between the component random and the linear predictor. By considering $\{\mu_i = E(Y_i); i = 1, \dots, n\}$, we set

$$\eta_i = g(\mu_i); i = 1, \dots, n, \tag{2.7}$$

where g , called the link function is assumed monotone and differentiable. This yields a model in which a function of the mean belongs to the subspace generated by the

explanatory variables:

$$g(\mu_i) = X' \beta, i = 1, \dots, n. \quad (2.8)$$

The link function that associates the mean μ_i to the natural parameter is called canonical link function. In that case,

$$g(\mu_i) = \theta_i = X' \beta \quad (2.9)$$

2.2 Binary Logistic Regression

Historically, logistic regression and binomial regression were the first methods used, particularly in marketing, for scoring, and in epidemiology to address the problem of modelling of a binary variable, binomial (number of trials or success) or Bernoulli (with $n_i = 1$), such as, possessing or not possessing a product, a good or bad customer, death or survival of a patient, absence or presence of a disease (Antoniadis, 1992).

Thus, in the generalized linear model, improvements are continually being made to the logistic regression (McFadden, Nobel Prize in Economics was awarded in 2000 for his work on this), confirming it as one of the more reliable modelling methods, with several statistical indicators which allow easy control the robustness (LR ratio, R squared of McFadden, Hosmer-Lemeshow test).

2.2.1 Mathematical Principles and Properties

When one wishes to model a binary response variable, the form of the relationship is often not linear. It is convenient to use a non-linear function, of type logistics. The principle of binary logistic regression is to consider a binary variable to predict (target variable admitting only two possible modalities) as $y = 0$ or 1 and p explana-

tory variables noted $X = (X_1, X_2, \dots, X_j)$, continuous, binary or qualitative (Hilbe, 2009).

The goal of logistic regression is to model the conditional expectation $E(Y|X = x) = \mu$, by estimating a mean value of Y for all values of X . For a value of Y being 0 or 1 (Bernoulli distribution), this mean value is the probability that $Y = 1$. Otherwise it is to explain the probability

$$\mu = Pr(Y = 1) \quad \text{or} \quad 1 - \mu = Pr(Y = 0),$$

or rather a transformation of the latter by a mutual observation of explanatory variables. The idea is indeed to involve a real function g monotonous operating from $(0,1)$ to \mathfrak{R} and therefore seek a linear model of form:

$$g(\mu_i) = X' \beta, \tag{2.10}$$

with β vector of unknown parameters associated to the vector X and of dimension $(p, 1)$ if the vector X is of dimension $(1, p)$ (Antoniadis, 1992). There are many functions whose graphs have a sigmoidal shape and are candidates for this role, among them we can have:

Probit: if one chooses the normal distribution $N(0,1)$, then the corresponding probability model is called Probit model, and it is given by:

$$\begin{aligned} \mu &= \int_{-\infty}^{X' \beta} \phi(t) dt = \Phi(X\beta) \\ 1 - \mu &= \int_{X' \beta}^{+\infty} \phi(t) dt = 1 - \Phi(X\beta) \end{aligned} \tag{2.11}$$

where Φ is the cumulative distribution function of normal and ϕ its density function.

Logit: one model more easier to use is the logit model defined as:

$$g(\mu) = \text{logit}(\mu) = \ln \left(\frac{\mu}{1 - \mu} \right) \quad (2.12)$$

with probability ,

$$\begin{aligned} \mu &= \frac{\exp(X\beta)}{1 + \exp(X\beta)} = g^{-1}(X\beta) \\ 1 - \mu &= \frac{1}{1 + \exp(X\beta)} = 1 - g^{-1}(X\beta) \end{aligned} \quad (2.13)$$

Then the quotient $\frac{\mu}{1 - \mu}$ is called Odds and the function $\ln \left(\frac{\mu}{1 - \mu} \right)$ is called logit. The computation of the odds ratio allows for consideration of this transformation as noted earlier (Fougere, 2008).

2.3 Tobit Model

We will now consider the case of limited dependent variable models. These are models where the dependent variable is continuous but is observable on a certain interval. They are models which lie midway between the linear regression models, where the endogenous variable is continuous and observable, and qualitative models.

Indeed, the basic structure of limited dependent variable models is represented by the Tobit. The Tobit model refers generally to regression models in which the dependent variable definition area is constrained in one form or another (Harari-kermadec, 2009). In economics, such models were introduced by Tobin (1958). His analysis focused on durable goods and consumption expenses and was based on a regression taking into account specifically the fact that these expenses can not be negative.

The dependent variable was therefore subject to a constraint of non negativity. The model and its generalizations are better known among economists as Tobit

model. This term was introduced by Goldberger (1964) because of the similarities with the probit model. However, these models are also called censored regression models or truncated regression models. For precision we introduce the distinction between truncated and censored samples:

A regression model is said to be a truncated regression model when all observations of the explanatory variables and the dependent variable set outside a certain interval are completely lost.

A regression model is said to be censored regression model when one has at least observations of explanatory variables on the entire sample (Harari-kermadec, 2009).

More formally, consider N pairs of variables (x_i, y_i^*) where the variable y_i^* is generated by a random process such that $E(y_i^*|x_i) = x_i\beta$, where $\beta \in \Re^k$ is a vector of unknown parameters. It is assumed that the variable y_i^* is not always observable, it is observed only if its value is greater than a certain threshold. Hence it is possible to build a variable y_i , which is equal to y_i^* when it is observable and it is equal to a constant c by convention when y_i is not observable. The Tobit model is a censored model, contrary to y_i^* , one observes x_i for the entire sample (Tobin, 1958).

$$y_i = \begin{cases} y_i^* & \text{if } y_i^* > c_i \\ c_i & \text{otherwise.} \end{cases} \quad (2.14)$$

The constant may be identical for all individuals. Two cases can arise depending on the nature of the observations:

If the vector x_i is observable for all individuals, regardless of the fact that the variable y_i^* is observable or not, then it is a censored sample. Only the variable y_i^* is observed over an interval $(c_i, +\infty)$.

If the vector x_i is observable only for the individuals for whom the variable y_i^* is observable, then it is a truncated sample. It does not have observations (x_i, y_i^*) for individuals where $y_i^* > c_i$.

The linearity assumption is put into question and shows that ordinary least squares is not the appropriate method for estimating such a relationship. In a general way, here one cannot use a continuous density to explain the conditional distribution of expense relative to income: a continuous distribution is incompatible with the fact that several observations of expenses are zero. It is in this context that Tobin (1958) proposed his limited dependent variable model.

The economic analysis of this situation is that the agent chooses consumption level that optimizes utility under a budget constraint. If the optimum consumption is positive, the optimal amount is consumed, and if it is negative, it is not consumed. Therefore the agent is limited by a constraint of positivity. The model is given by

$$y_i^* = x_i\beta + \varepsilon_i, \forall i = 1, \dots, N,$$

$$y_i = \begin{cases} y_i^* & \text{if } y_i^* > 0 \\ 0 & \text{if } y_i^* \leq 0, \end{cases} \quad (2.15)$$

where disturbances ε are normally distributed, that is, $\varepsilon \sim N(0, \sigma_\varepsilon)$.

The censored and truncated models have been used in other disciplines especially in epidemiology and engineering sciences. In epidemiology, such models were used to represent survival time of patients in terms of certain characteristics. The samples were indeed censored or truncated as soon as the patient remained alive at the last sample observation date or if the patient could not be auscultated on that date for any reason (Harari-kermadec, 2009).

Similarly in engineering, censored and truncated models are used to analyse time survival of a material or system in terms of characteristics. Such models are then qualified as survival models. Economists and sociologists have also used survival models to estimate the duration of phenomena such as unemployment, marriage, duration of residence in certain places (Goldberger, 1964).

2.4 Probit Model

Consider the observed data with n independent observations $\{(x_i, y_i) : i = 1, \dots, n\}$ with the covariate vector x_i of p dimension and y_i either 1 or 0 the binary response. The logistic regression model is defined as

$$\text{logit}(Pr(y_i = 1|x_i, \beta)) = \log \frac{Pr(y_i = 1|x_i, \beta)}{1 - Pr(y_i = 1|x_i, \beta)} = x_i' \beta. \quad (2.16)$$

The probit model where

$$Pr(y_i = 1|x_i, \beta) = 1 - Pr(y_i = 0|x_i, \beta) = \Phi(x_i' \beta), \quad (2.17)$$

is obtained by substituting the logistic distribution for the latent error terms ε_i with the standard normal distribution (Albert & Chib, 1993). The expectation maximization (EM) algorithm can be utilized to get the maximum likelihood estimates of β (Rubin, 1977). In addition, $\Phi(x)$ and $\phi(x)$ are distribution and density function respectively for standard normal distribution.

Logistic regression models and probit models have the estimates of regression coefficient which are not robust in the presence of outliers (Pregibon, 1982). Robit regression model for binary data is a robust alternative to the more common probit and logistic models. By proceeding like Lange et al. (1989) who substituted the nor-

mal distribution in the linear regression model with a t-distribution to get robust estimators of linear regression coefficients, one replaces the normal distribution in the probit regression model with t-distribution with ν degree of freedom to obtain a robust model. This model is called robit regression and denoted by robit (ν) (Liu, 2004).

Generally, the robit regression model for $\{(x_i, y_i) : i = 1, \dots, n\}$ is,

$$Pr(y_i = 1|x_i, \beta) = 1 - Pr(y_i = 0|x_i, \beta) = F_\nu(x_i\beta), i = 1, \dots, n, \quad (2.18)$$

where $F_\nu(x_i)$ is the cumulative distribution function of t random variables with centre zero and scale parameter one. The density function is given by

$$f_\nu(x) = \frac{\Gamma(\frac{\nu+1}{2})}{(\pi\nu)^{\frac{1}{2}}\Gamma(\frac{\nu}{2})\left(1 + \frac{x^2}{\nu}\right)^{\frac{\nu+1}{2}}}, (x \in (-\infty, +\infty)), \quad (2.19)$$

As $\nu \rightarrow \infty$, the robit(ν) model becomes the probit regression model.

Gelman & Hill (2007) have shown that in the presence of outliers, the model of robit, contrary to probit and logit models, can actually relieve the points of some conflicting data, for a better fit of the model. Therefore, if the degrees of freedom parameter is chosen appropriately, robit model will replicate the logistic models or probit if data follows one of these models, but will provide a robust alternative when outliers are present.

2.5 Robust Logistic Regression

Robust signifies the characteristic of remaining resistant against some irregular deviations. In statistics, models are a simple estimation of reality. The models that

underlie numerous statistical process are very optimistic and in real data big errors happen with unpredictable large frequency. An observation that lies an abnormal distance from other values in the data set is an outlier. This can later disturb statistical models causing results in an expected model to differ significantly from the exact one. Robustness means insensitivity against some divergence from the right model. Robust process was initiated in the works of Tukey (1960) and further, formal models of robustness have been expanded in 1970's. In regression models, the purpose of robust methods is to detect the outliers and extremely influential data points, leverage points, and to end by describing the goodness of fit for the data. One of the first result works related to Least Square, as a robust estimator, was carried out by Edgeworth (1887), who enhanced the proposal of Boscovich (1757) (Koenker & Bassett, 1985). This estimator is the least absolute deviation (LAD). The following improvement was Huber's M-estimator (Huber, 1973) and (Huber, 1981). In our literature on robust logistic regression, we review the M-estimators robust method by first explaining the loss function and we end by the trimming approach.

2.5.1 Loss Function

The loss function defines a cost or loss for each data with respect to an intermediate solution of the regression process. This is a function of the geometric distance between the data and this intermediate solution, i.e. the residue. Thus, for each intermediate solution is associated a total loss, which is the sum of the losses caused by each of the given data set.

Logistic regression introduces an extra non-linearity over a linear classifier f , by

using a logistic (or sigmoid) function, ω defined as,

$$\omega(f(x_i)) = \begin{cases} \geq 0.5 & \text{if } y_i = +1 \\ < 0.5 & \text{if } y_i = -1, \end{cases} \quad (2.20)$$

where

$$\sigma(f(x_i)) = \frac{1}{1 + e^{-f(x)}}.$$

with

$$f(x) = \beta^T x.$$

Here, we assume y is the label of data, x is a feature vector and β is a coefficient vector. The loss function S is defined as,

$$S(y, f(x)) = \log(1 + e^{-yf(x)}). \quad (2.21)$$

2.5.2 M-Estimator

Huber (1973) introduced the concept of M-estimator, say maximum likelihood estimator to limit the influence of erroneous data on the estimate. Estimating β by the method of maximum likelihood, it is proposed as value of β which maximizes the likelihood, namely the probability of observing the data as the realization of a sample according to a certain probability distribution. To calculate the maximum likelihood, determine the values for which the derivative of the likelihood vanishes.

This is the simplest method both computationally and theoretically. It is still widely used in the field of data analysis where contamination is mainly located in the Y -response vector. Instead of using the quadratic loss function, like in least squares, which is associated with a Gaussian probability distribution, the M-estimator of

Huber minimizes a sum of residual values calculated by using a loss function S increasing less rapidly than quadratic one.

$$\begin{aligned}\hat{\beta}_M &= \min_{\beta} Q_M(\beta) \\ Q_M(\beta) &= \sum_{i=1}^n S(r_i)\end{aligned}\tag{2.22}$$

The optimal estimate is determined from the derivative of the sum with respect to p coefficients of β , let,

$$\begin{aligned}\frac{\partial S(r_i)}{\partial \beta_j} &= \psi(r_i)X_{ij} \\ \sum_{i=1}^n \psi(r_i)X_{ij} &= 0, \forall j = 1, \dots, p\end{aligned}\tag{2.23}$$

where the function ψ is the derivative of the loss function S . M-estimation is obtained by solving the system of p non-linear equations. However, the solution is not equivariantly relative to the scale.

Indeed, if the residues are multiplied by an arbitrary value, meaning when the scale is changed, the resulting solution will be different. We must standardize residue with an estimate of the standard deviation σ . Thus the solution can be written as

$$\sum_{i=1}^n \psi(r_i|\hat{\sigma})X_i = 0,\tag{2.24}$$

where the standard deviation σ has to be estimated simultaneously. One option often used for its estimation is to use a multiple of the median absolute deviation(MAD). This implicitly assumes that the use of contamination rate due to noise is 50%. The median absolute deviation is defined by

$$MAD(X_i) = med_i \{|X_i - med_j(X_j)|\}.\tag{2.25}$$

and the estimator of standard deviation is given by

$$\hat{\sigma} = \beta.MAD. \tag{2.26}$$

To reduce the influence of contaminated data, the loss function S must be chosen according to the density of probability that defined the distribution of measurement errors. The loss function must meet certain conditions. It should be symmetrical, positive with a single minimum in zero and a growth slower than the quadratic function.

2.5.3 Trimming Approach

Logistic regression is concerned with explaining the probability of a specific response in terms of a number of regressors using a sample of relevant data. Pregibon (1981) affirmed that the estimated logistic regression correlation may be extremely influenced by outliers; this stimulates the necessity for robust logistic regression methods. Researches in this direction have been conducted by Pregibon (1981), Copas (1988), Rousseeuw & Christmann (2003), Huber (1973), Rousseeuw & Leroy (1987) and Yohai (1987).

Trimming is an extensive approach to robustifying of statistical process. It permits one to detect outliers and eliminate them from the data exploited in the estimation procedure. Trimming has been expanded highly by different authors in least squares regression, multivariate analysis and other areas (Rousseeuw (1984), Rousseeuw & van Driessen (1999), where additional mentions can be obtained). It appears attractive to apply trimming also in logistic regression to find outliers and to control their influences

When trimming, a subset of the data which is extremely probable to be free from the outliers is essential and a method is required to choose such a subset. One option is to employ maximum likelihood considerations, but this method has the tendency to run into the separation problem. This challenge is that those observations which are thought of as outliers are mostly the same observations which will give some overlap in the data.

Thus, trimming these observations eliminates the overlap and can lead to the indeterminacy of the maximum likelihood estimator (MLE) applied to the remaining data as indicated by Christmann & Rousseeuw (2001). They gave a procedure to evaluate this overlap, allowing the user to decide the closeness to indeterminacy. In complement Rousseeuw & Christmann (2003) overpowered the non-existence challenge by proposing the hidden logistic regression model with an associated estimator referred to as the maximum estimated likelihood (MEL) estimator which always exists even in absence of overlap in the data.

They also introduced a robustified system of the (MEL) estimator, denoted the weighted maximum estimated likelihood (WEMEL) estimator. But WEMEL instead of trimming, downweights leverage points, where the selection of leverage points is focused on the robust distances in the regressor space. They prove from a simulation study that WEMEL behaves extremely well as a robust method compared to its competitors. WEMEL does not consider outliers in the response direction; it is not an outlier detection technique in the sense that it gives a subset of the observations that can be considered as outliers.

On the other hand, the outlier can disturb statistical models and results in an expected model differ significantly from the exact one. Outliers in LR may occur in the Y -space called misclassification-type error (Copas, 1988), the X -space con-

sidered as leverage points or in both spaces. Outlying cases in this work are only focused on the covariate corruptions.

In this study, the robust logistic regression is based on the approach of trimming probability whose estimation procedure is related to the Bayesian inference using Gibbs sampler and Metropolis-Hastings Algorithm.

2.6 Estimation using Bayesian Approach

Consider the set of observations denoted by x with $x = (x_1, \dots, x_n)$. In other words, we have a sample of size n , where the observations x_i are considered as realization of random variables denoted X_i .

Priori information on the parameter θ means any available information on θ apart from the one brought by the observations. This contained uncertainty, otherwise the parameter θ would be known with certainty and we would not estimate it. It is natural to model this information through a probability distribution, called prior. Its density is denoted by $\pi(\theta)$.

In Bayesian statistical approach, one needs the prior distribution and the observation distribution which is the conditional distribution of X given θ . Its density is denoted by $f(x|\theta)$, for the random variable X either discrete or continuous. If X is discrete, $f(x|\theta)$ represents $Pr(X = x|\theta)$. Consider the hypothesis where, knowing θ , the variables X_i are independent. In other words, we have

$$f(x|\theta) = \prod_{i=1}^n f(x_i|\theta). \quad (2.27)$$

Posterior distribution: it is the conditional of θ knowing x . Its density function

is denoted $\pi(\theta|x)$. Using the Bayes formula, we have

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int_{\Theta} f(x|\theta)\pi(\theta) d\theta}, \quad (2.28)$$

with Θ the parametric space.

The couple distribution (θ, X) : its density is denoted

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

The marginal distribution of X : its density is denoted

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

2.6.1 Bayesian philosophy

In classical statistics, the parameters in the models are considered to be fixed, while in Bayesian analysis, the parameters are treated as random variables. The parameters being random variables, they are given distributions. Prior distribution of the parameters is one before data is collected while posterior distribution is one realized after scaling the prior distribution with new information obtained. The posterior can be interpreted as the summary (in a probabilistic sense) of the available information on θ , once x is observed. The Bayesian approach realizes somewhat the updating of prior information by observation of x , through $\pi(\theta|x)$ (Tanner & Wong (2010)).

It is sometimes possible to avoid the computation of $\int_{\Theta} f(x|\theta)\pi(\theta) d\theta$. In fact, if we let f and g be two real functions defined on the same space U . We say f and g are proportional, denoted $f \propto g$, if there exists a constant \mathbf{a} such that $f(y) = \mathbf{a}g(y)$ for all $y \in U$. It is clear that the relationship \propto is a relationship of equivalence. In

particular, if $f \propto g$ and $g \propto h$, then $f \propto h$.

In Bayesian context, we have, $\pi(\theta|x) \propto f(x|\theta)\pi(\theta)$. As a function of θ , both expressions $\pi(\theta|x)$ and $f(x|\theta)$ are effectively proportional. The constant \mathbf{a} which appeared in the earlier definition is equal to $\frac{1}{m(x)}$ here. Note that, this quantity is a constant in the sense that it does not depend on θ . The notation $\pi(\theta|x) \propto f(x|\theta)\pi(\theta)$ is often written as $\pi(\theta|x) \propto L(\theta; x)\pi(\theta)$, where $L(\theta; x)$ denotes the Likelihood. Recall that $L(\theta; x) = f(x|\theta)$ (by definition).

2.6.2 Bayesian Estimation

Bayes estimator in one-dimension

We assume that the parameter θ is real. Recall that $\pi(\theta|x)$ is interpreted as a summary of the available information once observed. To have an estimate of the parameter θ , one usually retains the average of the posterior distribution. Therefore by definition, the Bayesian estimation of the parameter θ , is the mean of the posterior distribution. This mean is denoted $E[\theta|x]$. Formally, we have

$$E[\theta|x] = \int_{\Theta} \theta \pi(\theta|x) d\theta = \frac{\int_{\Theta} \theta f(x|\theta) \pi(\theta) d\theta}{\int_{\Theta} f(x|\theta) \pi(\theta) d\theta}. \quad (2.29)$$

with $\hat{\theta}$, the estimator of θ defined by $\hat{\theta} = E[\theta|x]$.

Bayes estimator in multi-dimension

In the multi-dimensional case where $\theta = (\theta_j, j = 1, \dots, J)$, the posterior mean $E[\theta|x]$ is equal to the vector $(E[\theta_j|x], j = 1, \dots, J)$ with,

$$E[\theta_j|x] = \int_{\Theta_j} \theta_j \pi(\theta_j|x) d\theta_j. \quad (2.30)$$

$\pi(\theta_j|x)$ is obtained by integrating $\pi(\theta|x)$ on all components of θ other than θ_j .

Most often, Bayes estimators of θ_j can not be computed explicitly and we have to obtain them using the Monte Carlo simulation method, where computation of Bayes estimators does not pose great difficulty.

2.6.3 Gibbs Sampler

Introduced by Geman and Geman(1984) within the framework of the image restoration, this algorithm can be seen as an extension of the Tanner and Wong's algorithm. The principle is still based on a decomposition of the general problem (simulating following a certain distribution) into a series of basic problems (simulating following conditional distributions). Consider the density $f(x, y_1, \dots, y_p)$. We are interested in the marginal distribution:

$$f(x) = \int \dots \int f(x, y_1, \dots, y_p) dy_1 \dots dy_p. \quad (2.31)$$

In particular, one wishes to obtain the mathematical expectation and variance. Evaluating this integral can be difficult and complicated to evaluate. However, it is assumed that the conditional densities are available. The Gibbs sampler allows us to generate x following $f(x)$ directly without using its expression which is thought difficult to handle, but by using conditional densities. Thus by producing a sample (x_1, \dots, x_m) large enough we can approximate the mean, variance, and other characteristics using a law of large numbers,

$$\lim_{m \rightarrow +\infty} \frac{1}{m} \sum_{i=1}^m g(x_i) = E[g(x)]. \quad (2.32)$$

2.6.4 Principle of Gibbs Sampler

Consider the basic case of $f(x, y)$. Assume $f(x|y)$ and $f(y|x)$ available. We can then generate what we call a Gibbs sequence by starting from a value x_0 , and generating y_0 with $\pi(\cdot|x_0)$, then x_1 with $\pi(\cdot|y_0)$, and y_1 with $\pi(\cdot|x_1)$ and so on.

After M iterations of this scheme, we obtain a sequence $(x_0, y_0, x_1, y_1, \dots, x_M, y_M)$. For M large enough, x_M is a realization of X .

In the Bayesian framework, the Gibbs algorithm will allow us to obtain a realization of the parameter $\theta = (\theta_1, \dots, \theta_m)$ following the posterior distribution $\pi(\theta|x)$ as soon as one is capable of expressing the conditional distributions: $\pi(\theta_i|\theta_j; x)$, $j \neq i$. Thus, Gibbs sampling involves starting from an initial vector $\theta^{(0)} = (\theta_1^{(0)}, \dots, \theta_m^{(0)})$. At the $(p+1)^{th}$ step, with the vector $\theta^{(p)} = (\theta_1^{(p)}, \dots, \theta_m^{(p)})$, simulating

$$\begin{aligned}
\theta_1^{(p+1)} &= \pi(\theta_1|\theta_2^{(p)}, \theta_3^{(p)}, \dots, \theta_m^{(p)}; x) \\
\theta_2^{(p+1)} &= \pi(\theta_2|\theta_1^{(p+1)}, \theta_3^{(p)}, \dots, \theta_m^{(p)}; x) \\
&\dots \\
\theta_m^{(p+1)} &= \pi(\theta_m|\theta_1^{(p+1)}, \theta_2^{(p)}, \dots, \theta_{m-1}^{(p)}; x)
\end{aligned} \tag{2.33}$$

Successive iterations of this algorithm successively generate the states of a Markov chain $\{\theta^p, p > 0\}$ for values $\mathfrak{N}^{\otimes m}$. The transition probability from θ' to θ is expressed as:

$$K(\theta', \theta) = K_1(\theta', \theta) \times K_2(\theta', \theta) \tag{2.34}$$

where:

$$\begin{aligned}
K_1(\theta', \theta) &= \pi(\theta_1|\theta'_2, \dots, \theta'_m) \times \pi(\theta_2|\theta_1, \theta'_3, \dots, \theta'_m) \\
K_2(\theta', \theta) &= \pi(\theta_3|\theta_1, \theta_2, \theta'_4, \dots, \theta'_m) \times \dots \times \pi(\theta_m|\theta_1, \dots, \theta_{m-1}).
\end{aligned}
\tag{2.35}$$

This shows that the chain admits an invariant measure which is the posterior. For a sufficiently large number of iterations, the vector θ thus obtained may be considered as a realization of the posterior (Albert & Chib (1993)).

2.6.5 Metropolis-Hastings Algorithm

Originally developed in 1953 for the treatment of physical problems by Metropolis, this algorithm was thereafter used extensively in statistical physics to simulate complex systems. Currently, in the statistical literature, this algorithm is presented as a method for producing a Markov chain which has been assigned stationary law π . Its implementation has the advantage of not requiring the definition of π as a nearly constant (Tanner & Wong (2010)).

From the target density $\pi(x)$ (possibly large), one chooses a conditional density $q(x, y) = q(y|x)$ from which it is quite easy to simulate. Starting with a value x_0 (possibly vector), the algorithm passes through the following steps at each iteration. Knowing that the chain is in the state x_t at the t^{th} iteration,

- Generating $y_{t+1} \sim q(x_t)$

- Calculating the probability of acceptance

$$\alpha(x_t, y_{t+1}) = \min \left\{ \frac{\pi(y_{t+1})q(y_{t+1}, x_t)}{\pi(x_t)q(x_t, y_{t+1})}, 1 \right\} \quad (2.36)$$

- Taking

$$x_{t+1} = \begin{cases} y_{t+1}, & \text{with probability } \alpha \\ x_t, & \text{with probability } 1 - \alpha \end{cases}$$

- Repeating these steps for t going from 0 to N^3 .

Chapter 3

METHODOLOGY

3.1 Proposed Model

In this work, we improve the model of Andrew Gelman (2004) by developing a self-selecting robust logistic regression. Suppose $y = (y_1, y_2, \dots, y_n)$ are n independent observations where y_i are binary responses data defined as:

$$y_i = \begin{cases} 1 & \text{if success} \\ 0 & \text{otherwise} \end{cases}$$

Binary regression models assume that $y_i \sim \text{Ber}(\pi_i)$ with $\pi_i = \text{Pr}(y_i = 1)$ the probability of success for each observation.

From that, the robust model we are developing is as follows:

$$SsRLR : \pi_i = \alpha + (1 - 2\alpha)\text{logit}^{-1}(X^T\beta) \quad (3.1)$$

where X is a vector of p independent variables, β is a p dimensional vector of regression coefficients for the predictor variables and α the random chance.

As opposed to other studies where the value of α is set beforehand by the statistician, we allow this to be determined from the data itself. In particular since we are working in the Bayesian paradigm, we give this value α a uniform prior distribution.

3.2 Parameters Estimation

This section provides estimation procedure for our model. That is, we derive estimates of the parameters β , α and σ .

In this study, we use logistic regression which is a particular model when one deals binary response data. A full Bayesian approach in estimation is used to minimize risk estimation and to obtain the optimal estimates. To proceed to the Bayesian inference for logistic analysis, we follow the usual pattern for all Bayesian analyses by writing down the likelihood function of the data, forming a prior distribution over all unknown parameters and using Bayes theorem to find the posterior distribution over all parameters:

Likelihood function

In particular, once the probability of success (which depends on the covariates) is obtained, the likelihood function is given by

$$L(\beta, \alpha | X, y) = \prod_{i=1}^n [(\pi_i)^{y_i} (1 - \pi_i)^{1-y_i}] \quad (3.2)$$

where π_i represents the probability of success and y_i the binary responses data.

In our model we have:

$$\pi_i = \alpha + (1 - 2\alpha) \frac{e^{\beta_0 + \beta_1 X_{i1} + \dots + \beta_p X_{ip}}}{1 + e^{\beta_0 + \beta_1 X_{i1} + \dots + \beta_p X_{ip}}} \quad (3.3)$$

Hence the likelihood function of the binary responses data of n independent observations is:

$$L(\beta, \alpha, \sigma | X, y) = \prod_{i=1}^n A_i B_i \quad (3.4)$$

where:

$$A_i = \left[\alpha + (1 - 2\alpha) \frac{e^{\beta_0 + \beta_1 X_{i1} + \dots + \beta_p X_{ip}}}{1 + e^{\beta_0 + \beta_1 X_{i1} + \dots + \beta_p X_{ip}}} \right]^{y_i}$$

$$B_i = \left[1 - \left(\alpha + (1 - 2\alpha) \frac{e^{\beta_0 + \beta_1 X_{i1} + \dots + \beta_p X_{ip}}}{1 + e^{\beta_0 + \beta_1 X_{i1} + \dots + \beta_p X_{ip}}} \right) \right]^{1 - y_i}.$$

Prior distributions

For the prior distribution, we assign the normal distribution for the logistic regression parameters, says $\beta \sim N(0, \sigma)$, where σ is assigned inverse gamma prior distribution.

The parameter under study α is given the uniform prior distribution $U[a, b]$.

Posterior distribution

To derive the posterior distribution, we multiply the prior distribution over all parameters by the likelihood function. Thus we have:

$$P_{post}(\theta|X, y) = L(\beta_j, \alpha, \sigma|X, y)P_{pri}(\beta_j)P_{pri}(\alpha) \quad (3.5)$$

where:

$$P_{pri}(\beta_j) = \prod_{j=0}^p \frac{1}{\sqrt{2\pi}\sigma_j} \exp \left[-\frac{1}{2} \left(\frac{\beta_j}{\sigma_j} \right)^2 \right] \quad (3.6)$$

$$P_{pri}(\alpha) = \frac{1}{b - a}, a \leq \alpha \leq b \quad (3.7)$$

are β and α prior distributions respectively.

Inferences under the model are carried out using Bayesian approach implemented in WinBUGS to obtain the marginal posterior distributions for each parameter. The model will be well specified in the simulation study.

Chapter 4

SIMULATION RESULTS

4.1 Methodology

4.1.1 Introduction

This section presents simulation of the contaminated binary response data. We then simulate by using the data to illustrate estimation of the models, and getting summary statistics to make inference.

4.1.2 Simulation Set up

We carried out a simulation study to investigate the robustness of the three models namely: the Self-Selecting Robust Logistic Regression (SsRLR) model, Gelman's Robust Logistic Regression (GRLR) model and the ordinary Logistic Regression (LR) model .

Following the work of (Croux, 2003), a logistic regression model is generated with two independent normally distributed covariates. The additive noise ε_i is drawn from a logistic distribution defined as:

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + (\varepsilon_i \geq 0) \quad (4.1)$$

The true parameter values are $\beta = (0, 2, 2)$ with sample size $n = 200$. The study was based under a variety of situation. First, we considered data without contamination with two independent normally distributed covariates with zero mean and unit variance.

Second, to examine the robust properties of all models, we introduced outliers by contaminating the data similarly to the idea proposed by (Victoria-Feser, 2002). We generated the outliers in R software by corrupting the covariates. This consists of randomly choosing a certain t proportion (3%, 5%, 7%) from both covariates and replace them with a sample X_i drawn from $N(t, 10, 2)$. The response variable for each proportion was then generated from the new corrupted covariates. Further details are given in the appendix.

Finally the generated binary response data was contaminated under different percentages of leverage points. Thereafter, the three logistic models were applied to these data generated.

The proposed self-selecting robust logistic regression model with the ordinary one as described in the previous chapter and Gelman's robust model are:

$$LR : \pi = \text{logit}^{-1}(X^T \beta) \quad (4.2)$$

$$GRLR : \pi = 0.01 + 0.98 \text{logit}^{-1}(X^T \beta) \quad (4.3)$$

$$SsRLR : \pi = \alpha + (1 - 2\alpha) \text{logit}^{-1}(X^T \beta), \quad (4.4)$$

where X is a vector of p independent variables, β is a p dimensional vector of regression coefficients for the predictor variables, 0.01 and 0.98 are the fixed alpha probability value that Gelman introduced, to fit his robust model to binary response data in the presence of outliers.

In order to better handle those outliers, our robust model proposed to the contaminated binary data response itself to select the value of the probability alpha. After getting that significant alpha value for the robust model, we compared the

goodness of fit of the three logistic regression models.

To carry out that model estimation, Bayesian approach was used. All parameters in the models were assigned prior distributions. In this analysis, a non informative normal prior was assigned to the regression coefficients β , the α interest parameter was assigned Uniform prior depending on the percentage of contamination; the variance parameters were assigned inverse gamma distributions. The models were implemented using WinBUGS version 1.4.

For each model, we ran 10,000 Markov chain Monte Carlo (McMC) iterations, with the initial 1,000 discarded to cater for the burn-in period and thereafter keeping every tenth sample value. McMC convergence of all models parameters were accessed by checking trace plots and autocorrelation plots of the McMC output. The WinBUGS code used during the analysis is detailed in the appendix.

4.1.3 Model diagnostics

The models goodness of fit were compared using the Deviance Information Criterion (DIC) as suggested by Spiegelhalter (2002). The best fitting model is one with the smallest DIC. The DIC value is given by $DIC = \bar{D}(\theta) + pD$, where \bar{D} is the posterior mean of the deviance that measures the goodness of fit, and pD gives the effective number of parameters in the model which penalizes for complexity of the model. However, several authors have stated that a difference in DIC of 3 between two models can not be distinguished while a difference between 3 and 7 can be weakly differentiated.

For further model assessment, we used the Bayesian Information Criterion (BIC). In statistic, the Bayesian information criterion or Schwarz criterion is a criterion for

model selection among a finite set of models and the model with the lowest BIC is preferred (Schwarz. 1978). It is based, in part, on the likelihood function and it is closely related to the Akaike Information Criterion (AIC). BIC value is given by $BIC = \hat{D} + 2p \log(n)$ where $\hat{D} = -2 \log L(\theta^*|y)$ with $L(\theta^*|y)$, the likelihood of each model, p the number of parameters and n the sample size.

4.2 Results

4.2.1 Introduction

The specific purpose of these simulations is to analyze the robustness of the previous logistic regression models under different contamination proportions of the binary data. For each simulated data set, we estimated and recorded the parameters β and α . In particular we focus on investigating how much each model performs in presence of leverage points in the binary response data. In assessing that performance, we compute and compare their DIC and BIC.

4.2.2 Model assessment and Comparison

The first finding involved the classical LR model. In fact, the generated outliers values between 5 and 10 caused the LR not to run, giving "Trap Message" and no output while the SsRLR model takes care of the leverage points without any problem. We got output and summary statistics by using a Restricted Logistic Regression (RLR) model defined as:

$$y_i \sim \text{Ber}(\pi_1)$$

$$RLR: \quad \text{logit}(\pi) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \quad (4.5)$$

$$\pi_1 = \min(1, \max(0.001, \pi))$$

It can be deduced that fitting ordinary logistic regression with outliers can get "Trap Message" and no output without using Minmax in WinBUGS. Figure 4.1-5 show a visual representation of the distribution of the data set. It is clearly confirmed in the figure 4.2 and 4.4 the presence of outliers localized between 5 and 10 as earlier said .

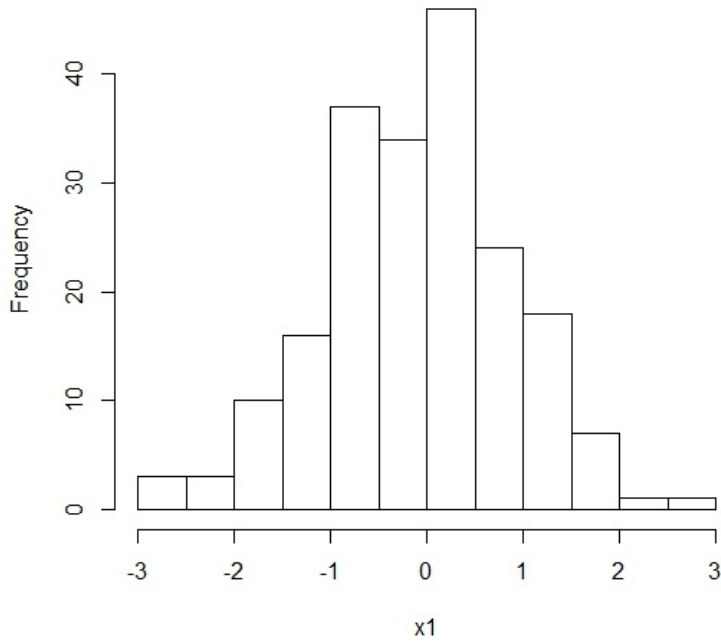


Figure 4.1: Histogram of X_1

Table 4.2-3 show the simulated results of all the fitted models for data with various percentages of leverage points. In absence of outliers (0% of lev pt), it can be observed that, there is no a significant difference between the restricted logistic and the robust models based on the DIC value. But the SsRLR model seems to give better estimated values of the parameters.

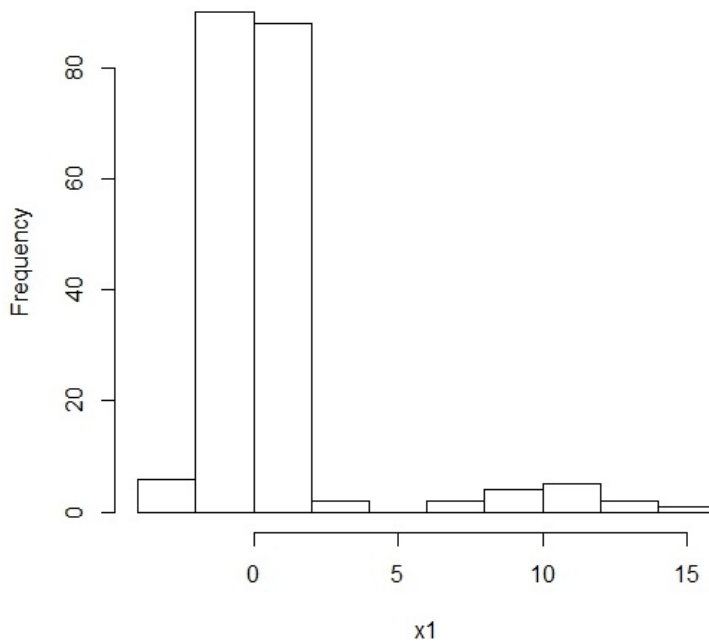


Figure 4.2: Histogram of X_1 with outliers

The Restricted LR model was immediately affected by 3% of leverage points giving the highest DIC value. Gelman's model was influenced as well showing parameter estimates which were not stronger than the expected one, while the SsRLR model let the data itself to select $2.664E-5$ alpha value that improved the parameter estimated values.

It is interesting to observe that the 5% of leverage points do not have effect on the SsRLR model. This latter confirms its robustness giving much better simulated result with the smallest DIC value.

The α values $5.014E-5$ and $2.059E-3$ respectively self selected in the presence of outliers (5% and 7% of lev pt) allowed the data to minimize the influence of those

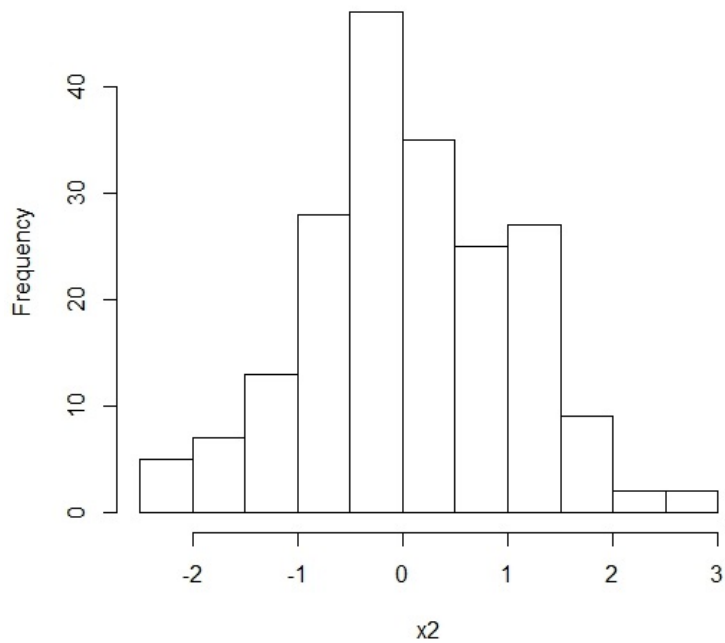


Figure 4.3: Histogram of X_2

latter in the parameter estimation.

Based on the criterion that a difference in DIC values from 3, 4 between two models provides a better fit, it can be clearly concluded that the best fitting model is the Self Selecting Robust Logistic Regression (SsRLR) model with small DIC value when there is presence of outliers in the binary response data.

Furthermore, based on the BIC, the SsRLR model with the lowest BIC is the preferred best fitting model (Schwarz. 1978).

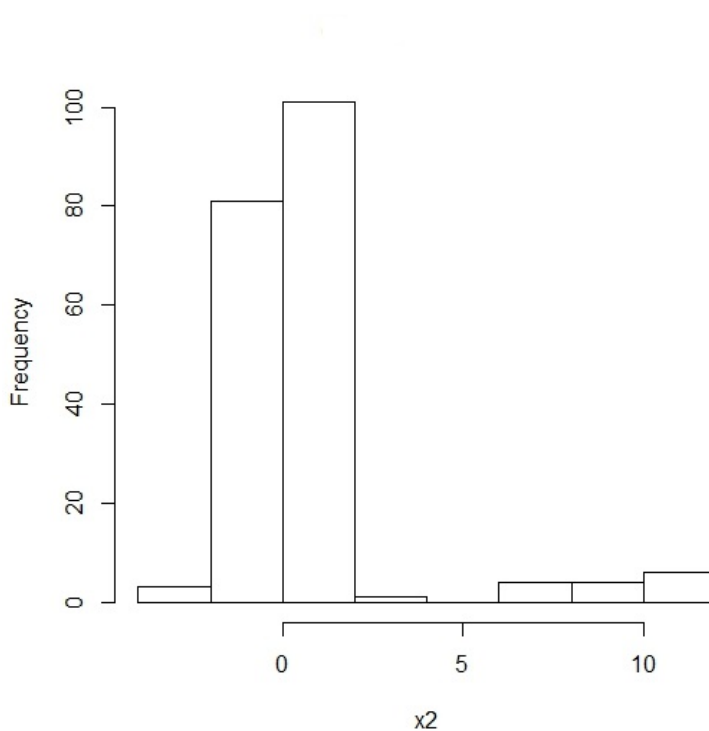


Figure 4.4: Histogram of X_2 with outliers

4.2.3 Discussion

This study uses Bayesian techniques to develop robust logistic regression model when outliers are present in binary response data. The study develops robust logistic model to help improve parameter estimation fitting. In this study, the approach used in the robust model is based on a trimming value, α chance of random error in both direction of the interval $[0,1]$.

From the existing contribution of Gelman (2004) that fixed α and $(1 - 2\alpha)$ in his model, we extended by self selecting these probability values depending on the data at hand and gave them a Uniform $[a,b]$ prior distribution.

In this study, we clearly confirmed that these probability values could also be

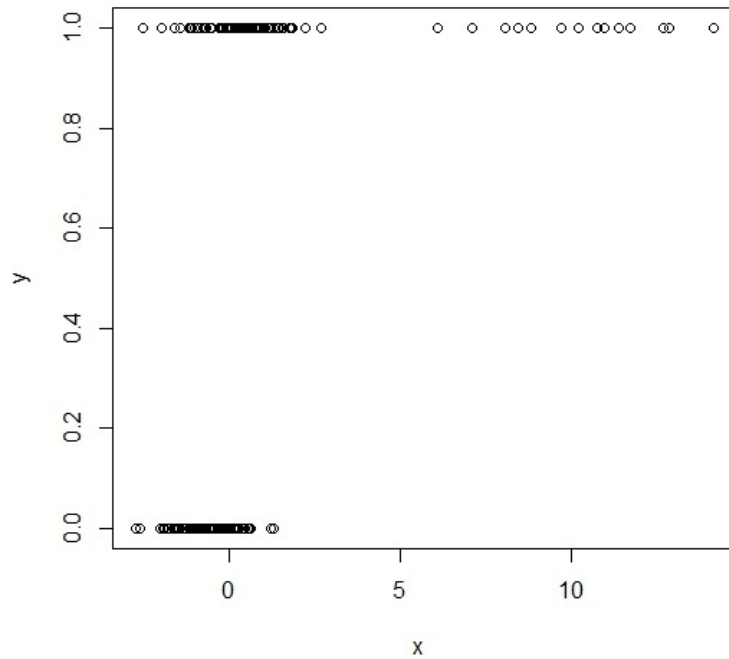


Figure 4.5: Relationship between Response Variable and Endogenous Variable X with outliers

determined by the data itself. In other words, depending on the binary data at hand, this latter could itself select α and $(1 - 2\alpha)$.

We found that the smaller the assigned values of a and b , the smaller the self selected α , and the more efficient the estimates obtained from simulation results will be, compared to the ones obtained from both the GRLR and the LR models when the data is either clean or contaminated.

Another finding is that the self selecting robust logistic regression model is a better fitting model compared to the Restricted LR model based on DIC value using Bayesian approach implemented in WinBUGS.

The SsRLR model provides a reliable fitting model based on the lowest BIC

value compared to the RLR and GRLR models.

We also found that the Restricted LR model has minimized the effect of the outliers present in the data and allowed achievement of better results. Despite this, the Self Selecting Robust Logistic Regression model presented more reliable results in comparison to the Restricted LR contrary to Gelman's robust logistic regression model.

Table 4.1: Description of variables X and assumed values of parameters manipulated in simulation

Variables and Parameters	Assumed values
n	200
x_1	$x_1 \sim N(n, 0, 1)$
x_2	$x_2 \sim N(n, 0, 1)$
β_0	0
β_1	2
β_2	2
α	$\alpha \sim U(a, b)$ with a and b belonging to $[0,1]$

Table 4.2: Simulated results of all models for Data with Leverage Points(0% and 3%)

% of lev pt	0%			3%		
	Estimate	SsRLR	GRLR	LR	SsRLR	GRLR
β_0	0.1074	0.1261	0.1072	0.02201	-0.05214	0.00607
σ_0	0.2237	0.1944	0.223	0.0216	0.2401	0.2034
β_1	2.371	2.201	2.363	2.035	1.714	1.875
σ_1	0.3139	0.3702	0.3618	0.3101	0.2016	0.3185
β_2	2.326	2.107	2.304	2.017	1.556	1.571
σ_2	0.2987	0.3731	0.3704	0.3021	0.263	0.3083
α	1.004E-5	0.01	-	2.664E-5	0.01	-
σ_α	5.835E-6	-	-	2.578E-5	-	-
Dhat	136.124	135.211	139.987	146.573	147.123	152.890
pD	2.917	2.820	2.978	2.765	2.781	3.061
BIC	154.532	153.619	153.793	164.981	165.531	166.696
DIC	142.208	141.958	143.098	153.624	153.686	157.812

Table 4.3: Simulated results of all models for Data with Leverage Points (5% and 7%)

% of lev pt	5%			7%		
	Estimate	SsRLR	GRLR	RLR	SsRLR	GRLR
β_0	0.06466	0.02474	0.06374	-0.1779	-0.1898	-0.1183
σ_0	0.2315	0.2292	0.2123	0.2202	0.2271	0.2333
β_1	2.194	1.859	2.069	2.344	2.198	2.217
σ_1	0.3014	0.2943	0.4068	0.3097	0.3634	0.4015
β_2	2.289	2.125	2.184	2.25	1.951	2.087
σ_2	0.3722	0.3104	0.3841	0.3475	0.3842	0.3907
α	5.014E-5	0.01	-	0.002059	0.01	-
σ_α	2.888E-5	-	-	0.001145	-	-
Dhat	125.280	126.038	131.056	113.339	114.058	119.009
pD	2.849	2.898	2.845	2.801	2.943	3.077
BIC	143.688	144.446	144.862	131.747	132.466	132.815
DIC	132.199	132.376	136.645	119.395	119.524	124.164

Chapter 5

CONCLUSION AND RECOMMENDATION

5.1 Conclusions

This work aims to extend the performance of logistic regression for binary data. Ordinary LR with arbitrary outliers was shown to fail. We proposed a robust SsRLR model that dealt with such contamination. It was also observed that by fixing the value of alpha, GRLR model was not that robust to the influential observations.

We proposed in this study a novel robust (LR) model to solve this issue. To proceed, we developed a self selecting robust logistic model, then investigated the robustness of this latter. We proposed a clear way of specifying the trimming values as required by the user, as opposed to fixing it.

One finding indicated across the simulation results that SsRLR model performs well in its specificity of letting the binary data itself to select the alpha value necessary to better improve the quality of the parameter estimates. Based on the smallest DIC and BIC value respectively, our SsRLR model was found to be the best fitting model under contaminated binary data sets.

We found that as long as the α value is the smallest self selected by the data at hand, the robustness of the SsRLR model is more improved. That is our contribution to Gelman's robust logistic regression model.

Inferences under the models are carried out by using Bayesian approach implemented in WinBUGS to obtain the marginal posterior distributions for each parameter. Another finding is that, when the covariates are corrupted, the use of the

Minmax through the Restricted LR model can somewhat help the classical LR to be robust.

5.2 Recommendations

Gelman (2004) introduced robust logistic regression model based on the approach of fixing a trimming probability α value that treats binary response data contaminated by outliers. In this work, we came up with an extension of that robustness. This study proposed to develop the robust model by letting the data itself to select the α value.

The SsRLR model behaved robustly with a lowest self selected α value compared to the GRLR model with a fixed α value. By carrying out this study, we showed the insufficiency contained in the Gelman's model when it comes to dealing with binary response data in presence of leverage points.

From there, we recommend that statisticians use the SsRLR model when modelling binary data in the case of covariate corruption and to further investigate its robustness in future research.

We also recommend that future researchers focus more on the robustness of the GRLR model by studying the behaviour of the trimmed probability α when the outliers occur in the Y -space called misclassification-type error or in both Y -space and X -space. The next author can widen this work by discussing the problem of improper prior of α .

References

- Albert, J. H., & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422), 669-679.
- Antoniadis, M., and Cullagh. (1992). Introduction to general linear model. *Wiki Stat*(1), 1-7.
- Christmann, A., & Rousseeuw, P. (2001). Measuring overlap in binary regression. *Computational Statistics And Data Analysis*, 37(1), 65-75.
- Copas, J. (1988). Binary regression models for contaminated data. with discussion. *Journal of the Royal Statistical Society. Series B (Methodological)*, 50(2), 225-265.
- Croux, G., Christophe Haesbroeck. (2003). Implementing the bianco and yohai estimator for logistic regression. *Computational Statistics and Data Analysis*, 44(1-2), 273-295.
- Edgeworth, F. Y. (1887). On observations relating to several quantities. *Hermathena*, 279-285.
- Enderlein, G. (1987). Generalized linear models. *Biom. J.*, 29(2), 206-222.
- Fougere, D. (2008). Introduction To Econometrics:Probit and Logit Models. *Journal of Econometrics*.
- Gelman, A. (2004). Parameterization and bayesian modeling. *Journal of the American Statistical Association*, 99(466), 537-545.
- Gelman, A., & Hill, J. (2007). Data analysis using regression and multi-level/hierarchical models. *J Educational Measurement*, 45(1), 94-97.
- Goldberger, A. S. (1964). *Econometric theory*. J. Wiley.

- Gordaliza, A. (1991). On the breakdown point of multivariate location estimators based on trimming procedures. *Statistics and Probability Letters*, 11(5), 387-394.
- Harari-kermadec, H. (2009). *Econometrics:Tobit Model*. URCA.
- Hilbe, J. M. (2009). *Logistic regression models*. CRC Press.
- Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression*. Wiley.
- Huber, P. J. (1973). Robust regression:asymptotics,conjectures and monte carlo. *Ann.Statist.*, 1(5), 799-821.
- Huber, P. J. (1981). Robust statistics. *New York: John Wiley and Sons*, 163-175.
- Jennings, D. E. (1986). Outliers and Residual Distributions in Logistic Regression . *Journal of the American Statistical Association*, 81(396), 987-990.
- Kateri, M., & Agresti, A. (2010). A generalized regression model for a binary response. *Statistics and Probability Letters*, 80(2), 89-95.
- Koenker, R., & Bassett, G. (1985). On boscovich's estimator. *Ann. Statist.*, 13(4), 1625-1628.
- Lange, K. L., Little, R. J. A., & Taylor, J. M. G. (1989). Robust statistical modeling using the t-distribution. *Journal of the American Statistical Association*, 84(408), 881.
- Liu, C. (2004). *Robit regression: a simple robust alternative to logistic and probit regression*.
- Maronna, R. A., & Yohai, V. J. (2000). Robust regression with both continuous and categorical predictors. *Journal of Statistical Planning and Inference*, 89(1-2), 197-214.
- Menard, S. W. (2002). *Applied logistic regression analysis*. Sage Publications.

- Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3), 370.
- Pregibon, D. (1981). Logistic regression diagnostics. *The Annals of Statistics*, 9(4), 705-724.
- Pregibon, D. (1982). Resistant fits for some commonly used logistic models with medical applications. *Biometrics*, 38(2), 485.
- Ritschard, G. (1990). Robust regression and problem of collinearity. *Journal of Statistics And Data Analysis*, 77-96.
- Rousseeuw, P. (1984). Least median of squares regression. *Journal of the American Statistical Association*, 79(388), 871-880.
- Rousseeuw, P., & Christmann, A. (2003). Robustness against separation and outliers in logistic regression. *Computational Statistics And Data Analysis*, 43(3), 315-332.
- Rousseeuw, P., & Leroy, A. (1987). *Robust regression and outlier detection*. J. Wiley.
- Rousseeuw, P., & van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3), 212.
- Sarkar, H., S.K.and Midi, & Rana, S. (2011). Detection of outliers and influential observations in binary logistic regression: An empirical study. *Journal of Applied Sciences*, 11(1), 26-35.
- Tanner, M. A., & Wong, W. H. (2010). From em to data augmentation: The emergence of mcmc bayesian computation in the 1980s. *Statistical Science*, 25(4), 506-516.
- Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica*, 26(1), 24.

- Tukey, J. (1960). A survey of sampling from contaminated distributions. *Stanford University Press, Contributions to Probability and Statistics*, 448-485.
- Victoria-Feser, M.-P. (2002). Robust inference with binary data. *Psychometrika*, 67(1), 21-32.
- Yohai, V. (1987). High breakdown-point and high efficiency robust estimates for regression. *The Annals of Statistics*, 15(2), 642-656.

Appendix A

For further enquiry contact gitawanou@gmail.com

Revised R code for simulation of binary response data with
different percentages of contamination.#####

```
## Simulated binary data without contamination
set.seed(0000)
x1 <- rnorm(200,0,1)
x2 <- rnorm(200,0,1)
z = 2*x1 + 2*x2
pr = 1/(1+exp(-z))
y <- rbinom(200,1,pr)
Data.R=c(x1,x2,y)
hist(x)
plot(x,y)

## Simulated binary data sets with t% of contamination
set.seed(1111)
x1 <- rnorm(200,0,1)
e=sample(1:200,t)
f<-rnorm(t,10,2)
x1[e]=f
x2 <- rnorm(200,0,1)
g=sample(1:200,t)
h <- rnorm(t,10,2)
x2[g]=h
```

```
z = 2*x1 + 2*x2
pr = 1/(1+exp(-z))
y <- rbinom(200,1,pr)
Data.R=c(x1,x2,y)
```

Appendix B

WinBUGS Codes for all three models.

Summary LR WinBUGS code simulation using data with 0% of contamination

```
# Ordinary Logistic Regression (LR) model
{
# Likelihood
for (i in 1 : 200)
y[i] ~ dbern(p[i])
logit(p[i])<-beta0+beta1*x1[i]+beta2*x2[i]
# Prior distributions
beta0 ~ dnorm(0.0001,0.0001)
beta1 ~ dnorm(0.0001,0.0001)
beta2 ~ dnorm(0.0001,0.0001)
# Data
# Initial values
list(beta0=0, beta1=0, beta2=0)
}
```

Summary RLR WinBUGS code simulation using each of all data sets with 3%,
5% and 7% of contamination

```
# Restricted Logistic Regression (RLR) model
{
# Likelihood
for (i in 1 : 200)
y[i] ~ dbern(p1[i])
p1[i]<-min(1,max(0.001,p[i]))
}
```

```

logit(p[i])<-beta0+beta1*x1[i]+beta2*x2[i]
# Prior distributions
beta0 ~ dnorm(0.0001,0.0001)
beta1 ~ dnorm(0.0001,0.0001)
beta2 ~ dnorm(0.0001,0.0001)
# Data
# Initial values
list(beta0=0, beta1=0, beta2=0)
}

```

Summary GRLR WinBUGS code simulation using each of all data sets with t% of contamination

```

# Gelman Robust Logistic Regression (GRLR) model
{
# Likelihood
for (i in 1 : 200)
y[i] ~ dbern(NewP[i])
logit(p2[i])<-beta0+beta1*x1[i]+beta2*x2[i]
NewP[i]<-(p2[i]-0.01)/(0.98)
# Prior distributions
beta0 ~ dnorm(0.0001,0.0001)
beta1 ~ dnorm(0.0001,0.0001)
beta2 ~ dnorm(0.0001,0.0001)
# Data
# Initial values
list(beta0=0, beta1=0, beta2=0)
}

```

Summary SsRLR WinBUGS code simulation for data sets with 0% of contamination

```
‡ Self Selecting Robust Logistic Regression (SsRLR) model
{
‡ Likelihood
for (i in 1 : 200)
y[i] ~ dbern(NewP[i])
logit(p2[i])<-beta0+beta1*x1[i]+beta2*x2[i]
NewP[i]<-(p2[i]-alpha)/(1-2*alpha)
‡ Prior distributions
alpha ~ dunif(0,0.00002)
beta0 ~ dnorm(0.0001,0.0001)
beta1 ~ dnorm(0.0001,0.0001)
beta2 ~ dnorm(0.0001,0.0001)
‡ Data ‡ Initial values
list(beta0=0, beta1=0, beta2=0, alpha=0.00001)
}
```

Summary SsRLR WinBUGS code simulation for data sets with 3% of contamination

```
‡ Self Selecting Robust Logistic Regression (SsRLR) model
{
‡ Likelihood
for (i in 1 : 200)
y[i] ~ dbern(NewP[i])
logit(p2[i])<-beta0+beta1*x1[i]+beta2*x2[i]
NewP[i]<-(p2[i]-alpha)/(1-2*alpha)
‡ Prior distributions
```

```

alpha ~ dunif(0,0.00007)
beta0 ~ dnorm(0.0001,0.0001)
beta1 ~ dnorm(0.0001,0.0001)
beta2 ~ dnorm(0.0001,0.0001)
# Data # Initial values
list(beta0=0, beta1=0, beta2=0, alpha=0.00001)
}

```

Summary SsRLR WinBUGS code simulation for data with 5% of contamination

```

# Self Selecting Robust Logistic Regression (SsRLR) model
{
# Likelihood
for (i in 1 : 200)
y[i] ~ dbern(NewP[i])
logit(p2[i])<-beta0+beta1*x1[i]+beta2*x2[i]
NewP[i]<-(p2[i]-alpha)/(1-2*alpha)
# Prior distributions
alpha ~ dunif(0,0.0001)
beta0 ~ dnorm(0.0001,0.0001)
beta1 ~ dnorm(0.0001,0.0001)
beta2 ~ dnorm(0.0001,0.0001)
# Data
# Initial values
list(beta0=0, beta1=0, beta2=0, alpha=0.000099)
}

```

Summary SsRLR WinBUGS code simulation for data with 7% of contamination

```

# Self Selecting Robust Logistic Regression (SsRLR) model

```

```

{
# Likelihood
for (i in 1 : 200)
y[i] ~ dbern(NewP[i])
logit(p2[i])<-beta0+beta1*x1[i]+beta2*x2[i]
NewP[i]<-(p2[i]-alpha)/(1-2*alpha)
# Prior distributions
alpha ~ dunif(0,0.004)
beta0 ~ dnorm(0.0001,0.0001)
beta1 ~ dnorm(0.0001,0.0001)
beta2 ~ dnorm(0.0001,0.0001)
# Data
# Initial values
list(beta0=0, beta1=0, beta2=0, alpha=0.00099)
}

```