

**Republic of Iraq
Ministry of Higher Education
And Scientific Research
University of Al-Qadisiyah
College of Administration
And Economics
Department of Statistics**



**Bayesian Estimation for Semiparametric Logistic Regression
with an Application**

**A Thesis submitted to the Council of the College of Administration
and Economics in Partial Fulfillment of the Requirements for the
M.S.C.**

By

Zainab Sami Turki

Supervised by

Dr. Taha Hussein Alshaybawee

2021 A.D

1443 A.H

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

(وَلَقَدْ آتَيْنَا دَاوُودَ وَسُلَيْمَانَ عِلْمًا وَقَالَا الْحَمْدُ لِلَّهِ الَّذِي فَضَّلَنَا

عَلَى كَثِيرٍ مِنْ عِبَادِهِ الْمُؤْمِنِينَ)

(صَدَقَ اللَّهُ الْعَلِيِّ الْعَظِيمِ)

سورة النمل

الآية (15)

ACKNOWLEDGEMENTS

All praise and thanks belong to Allah Who enables me to reach this stage in my education. Peace and Blessings upon the Prophet Mohammed (Peace be upon him) and his pure progeny from whom we derive the principles of being right persons before being educated.

I would like to express my deep thanks to my supervisor Dr. Taha Alshaybawee, for his valuable observations and assistance during the study period and writing the thesis.

My thanks and gratitude are due to all the staff of the Department of Statistics.

Also, I am so grateful to Asst.Lect. Ahmed Mansour Hadi for his great effort and touch.

I would also like to express my thanks to Assist.Lect Samer Al-Aalaq for his assistance.

I would also like to express my thanks and gratitude Hiba Hameed Hamzah for her great assistance during the research period.

Special thanks to my husband, for his tireless encouragement and motivation to get a higher degree in my studies.

Accordingly, thanks and love to my dear brothers who always hope the best for me.

DEDICATION

To

my parents

TO

my supervisor, Dr. Taha Alshaybawee

With

All Love, Respect and Gratitude

ABSTRACT

Semi-parametric model analysis is one of the most interesting topics in recent studies due to the precise way in which it describes the statistical data that provide effective parameters. In some studies, the response variable takes two values, either zero - for no response - or one for response. So, a logistic regression is used to model this data. Based on the Bayesian method, two new methods of estimation are proposed in this thesis.

The first one is the Bayesian estimation method which is used for estimating the unknown function and the coefficient vector in semi-parametric logistic regression (BSLR). The second one is the Bayesian lasso method which has been proposed for estimating and selecting significant variables for a single-index logistic regression (BSLLR) model. In BSLR method, normal distribution is set as prior distribution to the coefficient vector whereas Laplace distribution is considered as prior in BSLLR method. Gaussian process is set as prior for the unknown nonparametric function. The Markov Chain Monte Carlo (MCMC) algorithm is adopted for posterior inference. The different estimation methods were compared by comparing the use between the estimation methods using the mean squared error, the mean absolute error, Bias, and the standard deviation (SD). Using three simulation examples and with different sample sizes ($N=50,150,250$).

To test the efficiency of the proposed methods (BSLLR, BSLR), real data is used by adopting a set of indicators for the purpose of comparing the proposed methods with a set of pre-existing methods. To apply the estimation methods, a simple random sample of (260) was taken to study the factors affecting infection with the Coronavirus (response variable). While the explanatory variables are (gender, age, weight, pressure, diabetes, lung problems, weak immune system, vitamin D deficiency, workplace, previous surgeries, smoking, psychological state, nutrition, living condition). The study showed that the performance of Bayesian methods provides substantial improvements compared to other methods.

TABLE OF CONTENTS

No.	Subject	Page
I	ACKNOWLEDGEMENTS	i
II	DEDICATION	ii
III	ABSTRACT	iii
	Chapter One	1
1-1	Introduction	2
1-2	Aims of thesis	5
1-3	Literature Review	5
	Chapter Two	13
2-1	Introduction	14
2-2	Logistic Regression	15
2-3	Conditions for applying logistic regression	16
2-4	Justification for using logistic regression	17
2-5	Estimating the logistic regression parameters	17
2-5-1	Maximum Likelihood Estimation Method (MLE)	17
2-5-2	Bayesian Logistic Regression	23
	Chapter Three	26
3-1	Introduction	27
3-2	Single index models	29
3-3	Estimation of semi parametric single index model	29
3-3-1	Semiparametric Least Squares (SLS)	30
3-3-2	Pseudo Likelihood Estimation	31
3-4	Bayesian single index models	31
3-4-1	Bayesian Semi Parametric Logistic Regression	32
3-4-1-1	Hierarchical model and MCMC algorithm	35

No.	Subject	Page
3-4-2	Bayesian Variable Selection for Semiparametric Logistic Regression	38
3-4-2-1	Hierarchical model Posterior distribution	38
	Chapter Four	41
4-1	Simulation study	42
4-1-1	Example 1	42
4-1-2	Example 2	49
4-1-3	Example 3	53
	Chapter Five	57
5-1	Real data analysis	58
	Chapter Six	66
6-1	Conclusion	67
6-2	Recommendations	68
	Appendix A	69
	REFERENCES	70

LIST OF TABLES

No.	Table	Page
3-1	shows some commonly used semi-parametric models	28
4-1-1-1	The average SD of the parameter estimates of BSLR ,BSLLR, BLR,BPR and BBQR based on 100 replications when $n=(50,150,250)$ (Example 1)	43
4-1-1-2	bias of the parameter estimates of BSLR ,BSLLR, BLR,BPR and BBQR based on 100 replications when $n=(50,150,250)$ (Example 1)	45
4-1-1-3	The values of MSE and MAE of BSLLR,BSLR, BLR, BPR and BBQR methods for each sample (Simulated Example 1)	47
4-1-2-1	The average SD of the parameter estimates of BSLR ,BSLLR, BLR,BPR and BBQR based on 100 replications when $n=(50,150,250)$ (Example 2)	49
4-1-2-2	bias of the parameter estimates of BSLR ,BSLLR, BLR,BPR and BBQR based on 100 replications when $n=(50,150,250)$ (Example 2)	50
4-1-2-3	The values of MSE and MAE of BSLLR, BSLR, BLR, BPR and BBQR methods for each sample (Example 2)	51
4-1-3-1	The average SD of the parameter estimates of BSLR, BSLLR, BLR, BPR and BBQR based on 100 replications when $n=(50,150,250)$ (Example 3)	53
4-1-3-2	bias of the parameter estimates of BSLR ,BSLLR, BLR,BPR and BBQR based on 100 replications when $n=(50,150,250)$ (Example 3)	54
4-1-3-3	The values of MSE and MAE of BSLLR,BSLR, BLR, BPR and BBQR methods for each sample (Example 3)	55
5-1-1	The parameter estimates of BSLLR, BSLR, BLR, BPR and BBQR methods for the real data	61
5-1-2	The values of MSE and MAE of BSLLR, BSLR, BLR, BPR and BBQR methods for the real data	62

LIST OF FIGURES

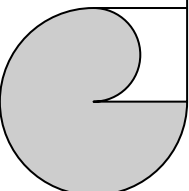
No.	Figure	Page
4-1-1-1	shows the bias values for BSLR, BSLLR, BLR, BPR and BBQR methods at three samples size (Example 1)	48
4-1-2-1	shows the bias values for BSLR, BSLLR, BLR, BPR and BBQR methods at three samples size (Example 2)	52
4-1-3-1	shows the bias values for BSLR, BSLLR, BLR, BPR and BBQR methods at three samples size (Example 3)	56
5-1-1	shows the MSE and MAE for BSLR, BSLLR, BLR, BPR and BBQR methods for real data	63
5-1-2	Traces plots for BSLR in the real data	64
5-1-3	Traces plots for BSLLR in the real data	65

ABBREVIATIONS

The term	Abbreviation
Bayesian logistic regression	BLR
Bayesian Binary Quantile Regression	BBQR
Bayesian Probit regression	BPR
Bayesian Semi parametric Logistic Regression	BSLR
Bayesian semiparametric lasso logistic regression	BSLLR
Frequentist Logistic Regression	FLR
Gaussian process	GP
independent explanatory variables	X
Least absolute shrinkage and selection operator	Lasso
Logistic Regression	LR
Markov Chain Monte Carlo	MCMC
Maximum Likelihood Estimate	MLE
Mean Absolute Error	MAE
Mean Square Error	MSE
Normal Distribution	ND
Pseudo Likelihood Estimation	PLE
Response Variable	Y
scale Mixture Uniform	SMU
Semiparametric Least Squares	SLS
Single index models	SIM
Standard Deviation	SD
Variable Selection	VS



CHAPTER ONE

- **Introduction**
 - **Aims of thesis**
 - **Literature Review**
- 

1-1 Introduction

Regression analysis aims to describe and analyze the relationship between a group of variables through a mathematical equation to link those variables. In addition to that, regression analysis methods are fundamental in analyzing the relevant data by describing the relationship between a set of independent variables and the dependent variable (Kerlinger & Pedhazur, 1973).

Although regression analysis achieves most of the aims of scientific research, whereas its methods represent the main part for any data analysis that aims to study and explain the relationship between the dependent variable and the independent variables. However, it is unable to describe and explain the relationships between the explanatory variables and the response variable when the later has binary value .

This type of dependent variable is common in the studies of a large number of humanitarian and social issues (Lea, 1997 ,Poston, 2004). This is why the need has arisen for developing new statistical methods that have the power of linear regression in reaching the best equations and dealing with them. At the same time, it recovers the problem of the inability to apply the usual linear regression analysis models in the case of dependent variables that have a binary value (Lea ,1997).

The outcome variable is discrete, taking on two possible values. Binary can have only two possible outcomes which we will denote as 1 and 0. Two relevant binary regression models, logit (logistic) and probit regression when the dependent variable is a binary response and take two values: 0 and 1

$$y = \begin{cases} 0 & \text{if } no \\ 1 & \text{if } yes \end{cases}$$

y : is a response variable distributed as Bernoulli with probability of success P .

A problem with the regression model is that the expected probabilities will not be limited between 0 and 1. Binary regression model is defined as $p_i = F(x'_i\beta)$, $i=1, \dots, N$ where β is a $k \times 1$ vector of unknown parameters, $x'_i = (x_{i1}, \dots, x_{ik})$ is a vector of covariates, and $F(\cdot)$ is a known cumulative distribution function (cdf) linking function. The linear structure of $(x'_i\beta)$ is the logit model when F is the logistic cdf, where the predicted probabilities are limited between 0 and 1. Whereas the probit model is obtained if F is the standard Gaussian cdf. The predicted probabilities are limited between 0 and 1.

Sometimes the explanatory variables are non-linear, which led researchers to find another method that deals with the nonlinear effect of these variables or nonparametric regression. It was proposed by the researcher (Jacob) in 1942. There are some problems with the nonparametric regression model, including the problem of dimensions (the curse of dimensionality). Therefore, the attractive features of single index model have motivated the researchers to extend this model for modelling a binary data. Kong & Xia (2008) suggest that the single-index model is one of the most general semiparametric models in econometrics. Single index models suppose that the response interest depends on a linear combination of covariates through an unknown link function (Hu, et al., 2013).

There are some statistical analyses when the information is not available about the sample under study. Therefore, researchers resort to making assumptions that the parameters to be assessed about random variables that

require obtaining prior information about them .It is based on previous experiences about the phenomenon by formulating them in the form of a prior probability distribution.

Prior distribution, which is combined with the probability observation distribution to obtain the posterior distribution, which contains all information about the parameters to be evaluated; this is called ‘the Bayesian’.

In this thesis we use the Bayesian lasso penalty approach for estimating and selecting variables in a single index logistic regression model. And, we will estimate the parameters of semi-parametric logistic regression using Bayesian inference because the classical methods have approximate results.

The remainder of this thesis is organized as follows:

Chapter 2, deals with the logistic regression model and its estimation methods, as well as the Bayesian logistic model recognizing the parameter models, where the researcher is concerned with the studied model (Single Index), and mentions some methods of its estimation, as well as the semi-parametric logistic model in Chapter 3, Simulation study is included in Chapter 4, The practical and applicable part of the study is presented in Chapter 5,Some conclusions and recommendations are given in Chapter 6.

1-2 . Aims of thesis

The aim of this study is to utilize the Bayesian method in estimating and variables selecting in the semi-parametric logistic model, Bayesian estimation approach is introduced to estimate the unknown link function and the coefficient vector in the semiparametric logistic regression. This thesis aims to at:

1-Constructing Bayesian hierarchical model for the semiparametric logistic regression (BSLR)

2-Constructing hierarchical Bayesian lasso semiparametric logistic regression model (BSLLR).

1-3. Literature Review

The most important studies related to each of the two topics of the semi-parametric logistic regression model and the methods for estimating the parameters of this model, represented by the Bayesian method, will be mentioned, and the following is a presentation of some of these studies:

Carroll and Wand (1991) studied the estimation of the semiparametric logistic regression model with measurement error in the prediction. This model was estimated by using (Kernel) methods as well as numerical prediction by averaging the error squares to estimate the model for obtaining the best estimate bandwidth.

(Bython, N.) 1992 studied the logistic regression model by relying on the simulation method. Four statistics were used to test the hypotheses of the model parameters to study the properties of the test strength. The differentiation was made to test the better statistics as well as to build confidence limits for estimating the parameters.

Binary response regression model is suggested by Newton, et al. (1996). That places no structural restrictions on the link function except monotonicity and known location and scale. Predictors enter linearly. They demonstrate Bayesian inference calculations in this model by modifying the Dirichlet process, we obtain a natural prior measure over this semiparametric model, and they use Polya sequence theory to formulate this measure in terms of a finite number of unobserved variables. They design a Markov chain Monte Carlo algorithm for posterior simulation and apply the methodology to data on radiotherapy treatments for cancer.

Bayesian hierarchical method proposed by (Hsu & Leonard 1997) for the semi-parameter logistic regression model by estimating the shrinkage parameters of the model and then the model was estimated. This method was applied to the mortality rate of mice exposed to nitrogen.

In 2001 Dominicil & Parmigianil studied congenital malformations through the Bayesian semi-parametric model of logistic regression, they found that the response variable fits the non-parametric part through the Dirichlet process, and the explanatory variables represent the parameter part. In this case, a semi-parametric model is formulated for them, So the parameters are estimated by merging the initial information with the possibility function for the parameters of the model. As a result, they have the posterior distribution, and this is called the Bayesian method.

In 2002, (Horowitz, et al.) reviewed several semi-parametric methods to estimate the function of conditional expectation. They made clear that these methods have broader flexibility than the parametric methods and provide greater accurate appreciation than the fully non-parametric

methods. They also clarified different methods of estimation by employing data on the salaries of the experienced baseball players in the United States.

In 2002, (Richardson et al.) used the Bayesian theory to solve the problems of measurement error by defining a prior distribution of parameters. A semi-parametric model for this distribution was presented on the basis of merging the normal distribution with unknown variables and the theory was applied Bayesian on a logistic regression model on coronary heart disease and cholesterol levels in blood byusing (MCMC).

In 2003 Titma et al. suggested to use of the polynomial logistic regression model, to study the effect of a group of factors on the development of the society of the former Soviet Union, based on two criteria to test the appropriate model. The first represents the Bayesian Information Criterion, and the second represents the statistic of the possible percentage using data for the professional class for the year (1991). They concluded that the father's education is the first influential factor, followed by the gender factor, and finally comes the heredity factor, which is less influential than the rest of the factors.

In 2003, Millmet et al. presented a study that included a comparison between the parametric and semi-parametric model about nitrogen oxide emissions and sulfur dioxide emissions, which is the air pollution problem in the United States. Through official statistical comparisons of the results, they proved simulating superiority of the semi-parametric model in data representation. They overwhelmingly rejected the parametric approach and explained the effect of the above problem on the US economy.

Dunson in(2004) studied the Bayesian theory of semi-parametric models to infer the regression function and to describe the relationship between the explanatory variable (x) and the dependent variable (y). Then, he transformed the data by using the standard natural transformation function (Z). He also used the prior distributions for the parameters by using the distributions of ‘hyper prior’, where the dependent variable data were applied to Poisson regression models and logistic regression models.

In 2005, (Lam & Xue) studied mixing the logistic regression with the semi-parametric model, since the semi-parametric regression model belongs to a flexible category with linear and non-linear models with the response variable. The capabilities of this method are consistent and approximate to the function. Its distribution is close to the normal. Simulation studies were carried out to investigate the performance of the proposed method and the model is fitted to a dataset of calcification of the hydrogel intraocular lenses, a complication of cataract treatment.

Bayesian approach used by Holmes and Hedy (2006) to study Binary regression models as well as multinomial or polychotomous regression models using two regression models: the first of the (probit regression model); the second is the (logistic regression model), and they added an auxiliary variable. Through the simulation process, Markov chain Monte Carlo method (MCMC) to compare the two models.

In 2007, (Haggag) studied the general semi-parametric linear model of logistic regression by using the greatest possibility method. The model was estimated on the credit evaluation data.

In the same year, (De Blasi & Hjort) studied the survival function analysis of the Bayesian theory in the relative risk models with logistic relative risk through series methods based on the partial probability capabilities. Simulation was used to take samples from the posterior distributions of parameter estimation, and they concluded that the Bayesian method is the better.

In 2013,(Acquah) estimated the parameters of the logistic regression model using the Bayesian method and compared it with the classical methods for the purpose of studying the relationship between per capita income and the trade of countries. The comparison was done using the simulation method and through the application of the MCMC algorithm has reached the preference of the Bayesian method over the classical methods of estimating the model parameters, this means that the trade of countries has an effect on the rise in percapita income.

In 2014 (Wang,et al.) used semi-parametric methods in logistic regression with error measurements through the conditional semi-probability method (PCL) and using the core functions (Kernel).

In 2015 (Michelot,et al.) estimated the parameters of the semi-parameter model of logistic regression by estimating the part of non-parameters in a method of Maximum Penalized Likelihood estimation.They used the Cross Validation method to smooth out the parameters by applying it to the data of pregnancy in sheep, the independent variable was taken as the environment conditions with time, non-parametric variable, and weight as a parameter variable.

In 2015, Emenyonus et al. talk about the risk diabetes. They state that significant variables were determined by the logistic regression model, which were then estimated using the Bayesian Logistic

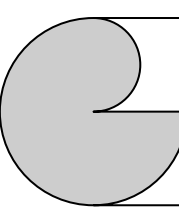
Regression (BLR) model. A flat non-informative prior, together with a non-informative non-flat prior distribution were used. These results were compared with those from the frequentist logistic regression (FLR) based on the significant factors. It was shown that the Bayesian logistic model with the non-informative flat prior distribution and frequentist logistic regression model yielded similar results, while the non-informative non-flat model showed a different result compared to the (FLR) model. Hence, non-informative but not perfectly flat prior yielded better model than the maximum likelihood estimate (MLE) and Bayesian with the flat prior.

In 2016, Salah used some modern semi-parametric methods in order to estimate and select variables in one for a single index model semi-parametric. He built an effective model by reaching the best method by depending on the average squares of error (MSE) for real data and the rate of squares of error (AMSE) by using simulation. The researcher concluded that the method (LASSO-MAVE) is the best in selecting and estimating the variable for most cases of the model in simulation experiments. He also observed that there is fluctuation in values (AMSE) with increasing sample size. Some of which decreased with increasing sample size; others increased with increasing sample size for different values of standard deviations and the number of variables for all estimation methods and semi-parametric. He suggested using the method (LASSO-MAVE) in the analysis of a single index model semi-parametric for its efficiency and for its being highly compared to acceptable ways of appreciation for the other semi-parametric.

In 2018, a single index estimation method was proposed by Alhamzawi, R, & Mohammad Ali for ordinal data. In that work, a simple and effective MCMC algorithm for calculating the posterior was developed by them through the use of ‘the normal exponential mixture representation of the skew LD’.



CHAPTER TWO

- **Introduction**
 - **Logistic Regression**
 - **Characteristics of Logistic Regression**
 - **Justification for Using Logistic Regression**
 - **Estimating the Logistic Regression Parameter**
 - **Bayesian Logistic Regression**
- 

2.1 Introduction

The ‘logistic regression model’ is used since 1845. At the beginning, it was a good model in studying, mathematically, the population growth during that period (Gürcan, 1998). Hair, et al. (2006) see that “the term logistic regression analysis rises from logit transformation, which is applied to the dependent variable. At the same time, this case causes certain differences both in estimation and interpretation”. Additionally, ‘Logistic Regression analysis’ can also be given several names, like: ‘Binary Logistic Regression Analysis’, ‘Multinomial Logistic Regression Analysis’ and ‘Ordinal Logistic Regression Analysis’.

This depends on the scale type where the dependent variable is measured and the number of categories of the dependent variable. Stephenson (2008) thinks that “Logistic regression is divided into two types: ‘univariate logistic regression’ and ‘multivariate logistic regression’.” In this model classifying the individuals in different groups is the main focus of logistic regression analysis. Kayri and Okut (2008) explained that “the individuals in special ability exam of a university for the Department of Physical Education and Sports Teaching were modeled using mixed logistic regression analysis as those achieved or not (or those who get into the department or not)”. These individuals were modeled in accordance with gender.

2-2. Logistic Regression

Logistic regression model is one of the nonlinear regression models, in which the relationship between the binary response variable (y), and the independent explanatory variables (x_1, x_2, \dots, x_k) is nonlinear. The logistic regression model is based on the basic assumption that the dependent variable (y) is binary take either of the two values (1, 0), either success with probability (P_i) or failure with probability ($1-P_i$).

A logistic regression model can be expressed by the formula:

$$p(y) = P_i^{y_i}(1 - P_i)^{1-y_i} \quad (2 - 1)$$

$$P_i = \frac{e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik}}}{1 + e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik}}}$$

$$1 - P_i = \frac{1}{1 + e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik}}}$$

Where y binary response variable (1,0), $x_i, i = 1, 2, \dots, k$ explanatory variables P_i the probability of the response occurring when $y=1$, $1 - P_i$ The probability of lack of response when $y=0$, $\beta_0, \beta_1, \dots, \beta_k$ unknown parameters of the logistic regression model.

It is known in simple linear regression that the dependent variable (y) takes values from $(-\infty, +\infty)$. But when the dependent variable is binary so, simple linear regression is not appropriate so, the value of the dependent variable takes the two values either zero or one. So, the model is not applicable from a simple linear regression perspective. One of the ways to solve this problem is to enter an appropriate mathematical transformation on the dependent variable (y). It is known that ($0 \leq P \leq 1$) and hence, the ratio $\left(\frac{P}{1-P}\right)$ is an integral positive expression Between

$(0 \leq \frac{P}{1-P} \leq \infty)$, and taking the natural logarithm of the expression $(\frac{P}{1-P})$, the field becomes enclosed by $(-\infty \leq \ln \frac{P}{1-P} \leq \infty)$, so the regression model is in the case of one independent variable as in Equation No (2_2):
(Athman,2018)

$$\ln\left(\frac{P}{1-P}\right) = b_0 + b_{1X} \quad (2 - 2)$$

If we have more than one independent variable, the model takes the form expressed in Equation No (2-3):

$$\ln\left(\frac{P}{1-P}\right) = b_0 + \sum b_j x_{ij} \quad (2 - 3)$$

Where: $i = 1, 2, \dots, n, j = 1, 2, \dots, k$

2-3. Conditions for applying logistic regression.(Saeed ,2015)

1. Logistic regression assumes that there is no linear relationship between the dependent variable and the independent variables.
2. The dependent variable must be a binary containing two classes.
3. Logistic regression does not require that the independent variables be of a continuous type or that they follow a normal distribution, nor that the relationship between the dependent variable and the independent variables be linear, and it is not assumed that the variance within each category is equal. This makes the logistic regression model more flexible than the rest prediction and classification models
4. The sample size used in the logistic regression must be larger than the sample size used in linear regression because the coefficients of the logistic regression model are approximated by using the method of the maximum likelihood method which requires a relatively large sample size.

2-4. Justification for using logistic regression (Athman,2018)

1. Heterogeneity :The error variance is not constant for the categorical variables. The unobserved errors are not normally distributed.
2. It is not possible to interpret the predicted values as probabilities, since these values cannot be limited between 0 and 1. Therefore, other statistical methods can be used, including the logistic regression model.

2-5. Estimating the logistic regression parameters:

2-5-1. Maximum Likelihood Estimation Method (MLE):

The maximum likelihood method is one of the most popular and the most suitable estimation methods for all linear and nonlinear models.

The maximum likelihood method is also known as an iterative method in which the mathematical operations are repeated several times until the best estimate of the parameters is reached.

As this method relies on finding values (β), which are estimates of the vector ($\widehat{\beta}$), that make the probability as great as possible, and it has the characteristics of a good estimator, i.e. it is characterized by unbiased, consistency and efficiency. The basic assumption of the logistic regression is that the dependent variable (Y) response variable is a binary variable that follows the Bernoulli distribution, taking the rank (1) with probability (P) and rank (0) with probability (Q= 1-P). Alshaybawee, T. (2006)

$$p_r(r_i/P_i, n_i) = C_{r_i}^{n_i} P_i^{r_i} (1 - P_i)^{n_i - r_i} \quad (2 - 5)$$

$$r_i = 0, 1, 2, \dots, n_i$$

n_i : Represents the number of attempts.

we know:

$$P_i = \frac{e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik}}}{1 + e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik}}} \quad (2 - 6)$$

$$1 - P_i = \frac{1}{1 + e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik}}}$$

Therefore, the function is:

$$p_r(r_i/p, n_i, \beta_0, \beta_1) = C_{r_i}^{n_i} \left[\frac{e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik}}}{1 + e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik}}} \right]^{r_i} \left[\frac{1}{1 + e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik}}} \right]^{n_i - r_i}$$

As for the maximum likelihood function, it will be:

$$L(\beta_0, \beta_1, n_i/r_i) = \ln \prod_{i=1}^g C_{r_i}^{n_i} + \ln \prod_{i=1}^g [P_i]^{r_i} + \ln \prod_{i=1}^g [1 - P_i]^{(n_i - r_i)} \quad (2 - 7)$$

$$L(\beta_0, \beta_1, n_i/r_i) = \ln \sum_{i=1}^g C_{r_i}^{n_i} + \sum_{i=1}^g r_i \ln [P_i] + \sum_{i=1}^g (n_i - r_i) \ln [1 - P_i]$$

And by derivation for β_0, β_1 according to the chain rule

$$\frac{\partial L}{\partial \beta_0} = \frac{\partial L}{\partial p} * \frac{\partial p}{\partial \beta_0} \quad (2 - 8)$$

$$\frac{\partial L}{\partial p} = \sum_{i=1}^g \frac{r_i}{p_i} - \sum_{i=1}^g \frac{(n_i - r_i)}{1 - p_i}$$

The derivative with respect to β_0 is:

$$\frac{\partial p_i}{\partial \beta_0} = \frac{e^{(\beta_0 + \beta_1 x_{i1})}}{(1 + e^{(\beta_0 + \beta_1 x_{i1})})^2} = p_i(1 - p_i)$$

And substituting the above two equations into equation (2 – 8)we get to

$$\frac{\partial L}{\partial \beta_0} = \sum_{i=1}^g \frac{(1-p_i)r_i - \sum_{i=1}^g p_i(n_i-r_i)}{p_i(1-p_i)} * p_i(1-p_i) \quad (2_9)$$

And from here we get to:

$$\frac{\partial L}{\partial \beta_0} = \sum_{i=1}^g (r_i - n_i p_i) \quad (2 - 10)$$

In the same way and according to the chain rule, the derivation is done with respect to β_1 :

$$\frac{\partial L}{\partial \beta_1} = \frac{\partial L}{\partial p_i} * \frac{\partial p_i}{\partial \beta_1} \quad (2 - 11)$$

And by solving equation (2 – 11), we get

$$\frac{\partial L}{\partial p} = \sum_{i=1}^g \frac{r_i}{p} - \sum_{i=1}^g \frac{(n_i - r_i)}{1 - p}$$

Since:

$$\frac{\partial p_i}{\partial \beta_1} = \frac{e^{(\beta_0 + \beta_1 x_{i1})} x_i}{(1 + e^{(\beta_0 + \beta_1 x_{i1})})^2} = p_i(1 - p_i)x_i$$

$$\frac{\partial L}{\partial \beta_1} = \sum_{i=1}^g \frac{(1-p_i)r_i - \sum_{i=1}^g p_i(n_i-r_i)}{p_i(1-p_i)} * \hat{p}_i(1 - \hat{p}_i)x_i \quad (2 - 12)$$

So the derivative with respect to β_1 is:

$$\frac{\partial L}{\partial \beta_1} = \sum_{i=1}^g X_i(r_i - n_i \hat{p}_i) \quad (2 - 13)$$

Equate equations (2 – 10) and (2 – 13) to zero, then

$$\sum_{i=1}^g (r_i - n_i p_i) = 0 \quad (2 - 14)$$

$$\sum_{i=1}^g X_i (r_i - n_i p_i) \quad (2 - 15)$$

Let $r_i = n_i p_i$, the equations (2 – 14) and (2 – 15) can be written as follows:

$$\sum_{i=1}^g n_i (p_i - \hat{p}_i) = 0 \quad (2 - 16)$$

$$\sum_{i=1}^g n_i x_i (p_i - \hat{p}_i) = 0 \quad (2 - 17)$$

The attention is focused on solving equations (2 – 16) and (2 – 17). The solution of these equations depends on the method of repetition (Iterative) proposed by Berkson in 1957.

The method of iterations requires initial values, so assuming that β_0, β_{01} represent the initial values for each of β_0, β_{01} then

$$l_0 = \beta_0 + \beta_{01} x_i$$

Where l_0 : represents an initial value to l_i and $(\hat{p}_0 - \hat{p}_i)$ represents the difference between the initial value to the ratio p and the next value is close to the first limit of the Tyler series as it is:

$$\hat{P}_i = \hat{P}_0 + (\hat{l}_i - \hat{l}_0) \hat{P}_0 \hat{Q}_0 \quad (2 - 18)$$

Substituting equation (2 – 18) into both equations (2 – 16) and (2 – 17), we conclude that:

$$\sum_{i=1}^g n_i \{p_i - (\widehat{p}_0 + (\widehat{l}_i - \widehat{l}_o)\widehat{P}_0\widehat{Q}_0)\} = 0$$

$$\sum_{i=1}^g n_i x_i \{p_i - (\widehat{p}_0 + (\widehat{l}_i - \widehat{l}_o)\widehat{P}_0\widehat{Q}_0)\} = 0$$

From it we find that:

$$\sum_{i=1}^g n_i \{(p_i - \widehat{p}_0) - (\widehat{l}_i - \widehat{l}_o) \widehat{P}_0\widehat{Q}_0\} = 0 \quad (2 - 19)$$

$$\sum_{i=1}^g n_i x_i \{(p_i - \widehat{p}_0) - (\widehat{l}_i - \widehat{l}_o) \widehat{P}_0\widehat{Q}_0\} = 0 \quad (2 - 20)$$

Since the:

$$(\widehat{l}_i - \widehat{l}_o) = (\beta - \beta_0) - (\beta_1 - \beta_0) x_i = \delta\beta_0 - \delta\beta_1 x_i \quad (2_21)$$

$$\delta\beta_0 \left(\sum_{i=0}^g n_i \widehat{P}_0\widehat{Q}_0 \right) + \delta\beta_1 \left(\sum_{i=0}^g n_i \widehat{P}_0\widehat{Q}_0 x_i \right) = \sum_{i=1}^g n_i p_i - \sum_{i=1}^g n_i p_0$$

$$\begin{aligned} & \delta\beta_0 \left(\sum_{i=0}^g n_i \widehat{P}_0\widehat{Q}_0 x_i \right) + \delta\beta_1 \left(\sum_{i=0}^g n_i \widehat{P}_0\widehat{Q}_0 x_i^2 \right) \\ & = \sum_{i=1}^g n_i p_i x_i - \sum_{i=1}^g n_i p_0 x_i \quad (2 - 22) \end{aligned}$$

which can be written in the following form:

$$\delta\beta_0 \frac{\partial^2 L}{\partial \beta_0^2} + \delta\beta_1 \frac{\partial^2 L}{\partial \beta_0 \beta_1} = - \frac{\partial L}{\partial \beta_0} \quad (2 - 23)$$

$$\delta\beta_0 \frac{\partial^2 L}{\partial \beta_0 \beta_1} + \delta\beta_1 \frac{\partial^2 L}{\partial \beta_1^2} = - \frac{\partial L}{\partial \beta_1}$$

And according to the matrices, you write:

$$\begin{bmatrix} \frac{\partial^2 L}{\partial \beta_0^2} & \frac{\partial^2 L}{\partial \beta_0 \beta_1} \\ \frac{\partial^2 L}{\partial \beta_0 \beta_1} & \frac{\partial^2 L}{\partial \beta_1^2} \end{bmatrix} \begin{bmatrix} \delta\beta_0 \\ \delta\beta_1 \end{bmatrix} = - \begin{bmatrix} \frac{\partial L}{\partial \beta_0} \\ \frac{\partial L}{\partial \beta_1} \end{bmatrix} \quad (2 - 24)$$

The matrix on the left side is the information matrix (I), and it is possible to obtain from this matrix the standard errors of each of β_0, β_1 , either each of the $\delta\beta_0, \delta\beta_1$, it is easy to obtain them from equation (2_23) or by using the matrices by taking the inverse of the information matrix in equation (2_24) multiply it by the vector on the right side, and this process is repeated until few or no differences are reached between two successive stages.

2-5-2. Bayesian Logistic Regression

Press (1989) mentions in his work that “Bayesian inference provides a useful way to combine between (prior belief) with data to arrive at some posterior belief. Bayesian inference is conducted through the use of Bayes’ theorem.”

Press (1989) also mentions that Bayes’ theorem gives a mathematical procedure for updating the prior belief to arrive at a posterior distribution; this can happen when there is a prior belief (called a prior distribution) before observing the data. The following conditional probabilities are used by the Bayes’ theorem:

$$\pi(\theta / y) = f(y / \theta) \pi(\theta) / f(y) \quad (2 - 25)$$

$$\text{Where } f(y) = \int f(y / \theta) \pi(\theta) d(\theta)$$

Equation (2 – 26) the basis of Bayesian statistics and econometrics.

$\pi(\theta / y)$: is the posterior density function .

$f(y / \theta)$: is the density function of the observed data when the parameter value is θ .

$f(y / \theta)$, which is called the likelihood function.

$\pi(\theta)$: is called the prior density and represents beliefs about the distribution of (θ) before seeing the data. These beliefs can come from the researcher’s knowledge or from other external sources.

Thus, as new information becomes available, the posterior distribution becomes the prior distribution for the next experiment. Bayesian inference for the logistic regression model requires priors on the

model parameters. Wilhelmsen *et al.* (2009) and Ziemba (2005) both use normally distributed priors for the model parameters and represented as follows: Greenberg, E. (2008)

$$\pi(\beta_i) = N(0, \sigma_i^2) \quad (2 - 26)$$

The posterior distribution is proportional to the product of the prior distribution and Likelihood

$$\pi(\beta/y) \propto l(y/\beta) \pi(\beta)$$

$$l(y/\beta) = \prod_{i=1}^n p_i^{y_i} (1 - p_i)^{1-y_i}$$

$$= \prod_{i=1}^n \left(\frac{\exp(\beta_0 + \sum_{j=1}^k \beta_j x_{ij})}{1 + \exp(\beta_0 + \sum_{j=1}^k \beta_j x_{ij})} \right)^{y_i} \left(1 - \frac{\exp(\beta_0 + \sum_{j=1}^k \beta_j x_{ij})}{1 + \exp(\beta_0 + \sum_{j=1}^k \beta_j x_{ij})} \right)^{1-y_i} \quad (2 - 27)$$

Therefore, from Equations(2 - 26) and (2 - 27) we have

$$\pi(\beta/y) \propto \prod_{i=1}^n \left(\frac{\exp(\beta_0 + \sum_{j=1}^k \beta_j x_{ij})}{1 + \exp(\beta_0 + \sum_{j=1}^k \beta_j x_{ij})} \right)^{y_i} \left(1 - \frac{\exp(\beta_0 + \sum_{j=1}^k \beta_j x_{ij})}{1 + \exp(\beta_0 + \sum_{j=1}^k \beta_j x_{ij})} \right)^{1-y_i} \prod_{j=1}^k \frac{1}{\sqrt{2\pi\sigma_j^2}}$$

$$\exp\left(-\frac{(\beta_j)^2}{2\sigma_j^2}\right) \quad (2 - 28)$$

$$\pi(\beta/y) = \frac{\prod_{i=1}^n \left(\frac{\exp(\beta_0 + \sum_{j=1}^k \beta_j x_{ij})}{1 + \exp(\beta_0 + \sum_{j=1}^k \beta_j x_{ij})} \right)^{y_i} \left(1 - \frac{\exp(\beta_0 + \sum_{j=1}^k \beta_j x_{ij})}{1 + \exp(\beta_0 + \sum_{j=1}^k \beta_j x_{ij})} \right)^{1-y_i} \prod_{j=1}^k \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left(-\frac{(\beta_j)^2}{2\sigma_j^2}\right)}{\int_{-\infty}^{\infty} \prod_{i=1}^n \left(\frac{\exp(\beta_0 + \sum_{j=1}^k \beta_j x_{ij})}{1 + \exp(\beta_0 + \sum_{j=1}^k \beta_j x_{ij})} \right)^{y_i} \left(1 - \frac{\exp(\beta_0 + \sum_{j=1}^k \beta_j x_{ij})}{1 + \exp(\beta_0 + \sum_{j=1}^k \beta_j x_{ij})} \right)^{1-y_i} \prod_{j=1}^k \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left(-\frac{(\beta_j)^2}{2\sigma_j^2}\right) d\beta} \quad (2 -$$

29)

Equation (2 – 28) reveals that the prior does not show any relation to the conjugate family. Actually, no conjugate prior for the Bayesian logistic regression model exists. Explicit calculation cannot be carried out for the integral in the denominator Equation (2 – 29). Here, there is a need to use simulation methods so as to achieve the posterior distributions of the parameters. Markov Chain Monte Carlo (MCMC) methods are also used. A Markov chain is generated, where the stationary distribution is equal to the posterior distribution of the vector β

CHAPTER THREE

- **Introduction**
- **Semi Parametric Regression Models**
- **Single Index Models**
- **.Estimation of Semi Parametric Single Index Model**
- **Bayesian Single Index Models**
- **Bayesian Semi Parametric Logistic Regression**
- **Bayesian Variable Selection Semiparametric Logistic Regression**

3-1 Introduction

The first appearance of the term semi-parametric in literature in 1980 by (Gail) and others in the field of vital statistics, and as this term is also attributed to (Oakes), in 1981 the term was used in the demographic side by (Finnas and Hoem) and in the same year the researcher (Whitehead) used in his book in the field of mathematics called (partial parametric). Semi-parameter models are used to study the relationship between the dependent variable (Y) and a set of explanatory variables (X) when the mathematical form of the relationship between the dependent variable and a set of explanatory variables is known with the existence of an explanatory variable, at least, the relationship between it and the dependent variable is unknown.

The semi-parameter models are considered a compromise between the constrained parameter models and the flexible; non-parametric models. Also, semi-parametric models contain two parts, one for parametric and the other for non-parametric, it is also an economic model which has been widely studied by economists in econometric research because it allows the inclusion of multiple explanatory variables without taking into account the problem of dimensions (curse of dimensionality), in other words, without being restricted to increasing the number of explanatory variables.

Some of the commonly used semi-parametric models can be displayed as shown in the table (3_1).

Table (3-1) shows some commonly used semi-parametric models.(Falsal,R(2020)).

Model name	Mathematical formula	Parametric part	The non-parametric part
Single index model	$E(Y/X) = m(X_i^T \beta)$	β	$m(\cdot)$
Partial additive model	$E(Y X, Z)$ $= ZX^T \beta + C$ $+ \sum_{j+1}^K m_j(Z_j)$	β	$m_j(\cdot)$
Partially linear model	$E(Y X_i, Z_i)$ $= X_i^T \beta$ $+ m(Z_i)$	β	$m(\cdot)$
Generalized partial linear model	$E(Y X_i, Z_i)$ $= g(Z_i^T \beta + m(Z_i))$	β	$g(\cdot), m(\cdot)$

In our study, we will consider single index model

3-2 . Single Index Models

Single-index model (SIM) introduce an efficient manner of handling high dimensional nonparametric estimation problems (Härdle et al., 1993; Yu and Ruppert, 2002) and avert the ‘curse of dimensionality’ (Bellman et al., 1966). Nonparametric problems assume that the response is just associated with a single linear set of the covariates. It is one of the most common and necessary semiparametric models in statistics as well as applied sciences like econometrics and psychology due to its ability to reduce dimensions (Ichimura, 1993). The semiparametric single index regression model is:

$$y = m(X_i^T \beta) + \varepsilon_i \quad (3 - 1)$$

where y is a response variable, β is a parameter vector (Parametric part), m : is an unknown link function (nonparametric part) and ε_i = errors are assumed to be iid.

3-3 .Estimation of semi parametric single index model

Several estimation methods have been proposed to estimate the unknown link function and the parameters vector the Single Index Model (SIM) .

Two essentially different approaches exist for this purpose:

- An iterative approximation of β by semiparametric least squares (SLS) or pseudo maximum likelihood estimation (PMLE),

3-3-1.Semiparametric Least Squares(SLS) (Härdle.et al)2004

SLS and its weighted version (WSLS) have been introduced by Ichimura (1993). As SLS is just a special case of WSLS with a weighting equal to the identity matrix, we concentrate here on WSLS. An objective function of least squares type can be motivated by minimizing the variation in the data that cannot be explained by the fitted regression. This “left over” variation can be written as :

$$var\{Y/(X_i^T \beta)\} = E[\{Y - E(Y/(X_i^T \beta))\}^2 / X_i^T \beta] \quad (3 - 2)$$

The previous equation leads us to a variant of the well-known LS criterion

$$min_{\beta} = E[\{Y - E(Y/(X_i^T \beta))\}^2] \quad (3 - 3)$$

Define the WSLS estimator as:

$$\hat{\beta} = min_{\beta} \sum_{i=1}^n \{Y_i - \widehat{m}_{\beta}(x_i)\}^2 w(x_i) I(x_i \in x) \quad (3 - 4)$$

Where

$I(x_i \in x)$ is a trimming factor

\widehat{m}_{β} leave-one-out estimator of m assuming the parameter β would be known.

$$\widehat{m}_{\beta}(x_i) = \frac{\sum_{j \neq i} Y_j K_h\{X_i^T \beta - X_j^T \beta\} w(x_j) I(x_j \in x_n)}{\sum_{j \neq i} K_h\{X_i^T \beta - X_j^T \beta\} w(x_j) I(x_j \in x_n)} \quad (3-5)$$

h denoting a bandwidth , K_h scaled (compact support) kernel.

3-3-2. Pseudo Likelihood Estimation: (Härdle.et al)2004

Gill (1989) and Gill & van der Vaart (1993) explain this as follows:

A sensibly defined nonparametric MLE can be seen as a MLE in any parametric sub model which happens to include or pass through the point given by the PMLE. For smooth parametric sub models, the MLE solves the likelihood equations. Consequently, also in nonparametric problems the PMLE can be interpreted as the solution of the likelihood equations for every parametric sub model passing through it.

We can now define the pseudo log-likelihood version:

$$\frac{1}{n} \sum_{i=1}^n w(x_i) \{ Y_i \log[\hat{G}_{\epsilon/x_i}\{X_i^T \beta\}]^2 + (1 - Y_i) \log[1 - \hat{G}_{\epsilon/x_i}\{X_i^T \beta\}]^2 \} \quad (3 - 6)$$

where $G_{\epsilon/x}$ is the conditional distribution of the error term. an estimate for $G_{\epsilon/x}$

is given by:

$$\hat{G}_{\epsilon/x_i}\{X_i^T \beta\} = \frac{\sum_{j \neq i} I(Y_j=1) K_h\{X_j^T \beta - X_i^T \beta\}}{\sum_{j \neq i} K_h\{X_j^T \beta - X_i^T \beta\}}$$

3-4. Bayesian single index models

Bayesian single index is introduced in Antoniadis et al. (2004). The Bayesian approaches can be classified by the different methods by which the link function m is modeled and different prior distributions assigned on the link function.

In Bayesian literature of SIM, the nonparametric link function m is either modeled by a basis representation like splines or wavelets or by

assigning a Gaussian process prior. Antoniadis et al. (2004) and Wang (2009) used spline-based basis representation of the link function for estimation. While using splines as basis for m , selecting the number of knots and the position of knots are computationally expensive. Even Reversible Jump Markov Chain Monte Carlo algorithms involving movable knots (Wang, 2009) suffer from computationally expensive variable dimensional iterations. Park et al. (2005) modeled the link function using wavelets. Using wavelets have the similar problem of selecting the number of basic functions.

Gaussian process priors were used to model the link function in Choi et al. (2011), Gramacy and Lian (2012) and Hu et al. (2013). Even though the Gaussian process priors do not have issues of selecting number of basis functions, using a Gaussian process prior on $m(\cdot)$ necessitates the inversion of $(n \times n)$ dimensional covariance matrix. Since the kernel matrix is a function of the index vector α , every iteration of the Markov chain Monte Carlo (MCMC) involves inverting a new variance-covariance matrix. This makes the algorithm computationally intensive even when the sample size is moderately large.

3-4-1. Bayesian Semi Parametric Logistic Regression

The topic of semi-parametric model analysis, which combines parametric and non-parametric models, has a clear interest in most studies of a more advanced nature in the process of accurate statistical analysis, which aims to obtain estimators of a high level of efficiency.

The semi-parametric regression model has gained widespread popularity in recent years, due to its advantage in integrating parametric regression models with non-parametric regression models at the same time.

This feature that made the new model exceeded the problem of dimensions in the case of non-parametric models completely, and also gave more space for application than that in the case of parametric models of regression, because the latter can be affected by some independent (illustrative) variables that do not have a known parameter distribution, and sometimes the function under study may not be fully represented due to the fact that some of the variables behave as parametric and the other is non parametric.

In some studies, the dependent variable of the semi-parametric binary-response model is either equal to one for the occurrence of the response or zero for the non-occurrence of the response, and this is called the Logistic Regression model, of an experimental nature, as it is one of the suitable models for binary data.

In this study, we will consider the dependent variable Y with a binary response, either with a probability of equal to one to obtain the response or zero to not obtaining the response, and this is called the logistic regression model.

The semiparametric logistic regression based on single index regression model is :

$$P_i = p(y = 1) = \frac{\exp (m(x'_i\beta))}{1 + \exp (m(x'_i\beta))} \quad (3 - 7)$$

where β parameters vector (Parametric part) and $m(\cdot)$ is an unknown link function (non-parametric part) and

$$1 - P_i = p(y = 0) = \frac{1}{1 + \exp (m(x'_i\beta))}$$

The likelihood function is the probability density function of the data which is seen as a function of the parameter treating the observed data as fixed quantities. For a given sample size n , the likelihood function is given as:

$$L(y|m, \boldsymbol{\beta}) = \prod_{i=1}^n f(y_i)$$

$$L(y|m, \boldsymbol{\beta}) = \prod_{i=1}^n P_i^{y_i} (1 - P_i)^{1-y_i}$$

Therefore, the likelihood function is of the form:

$$L(y|m, \boldsymbol{\beta}) = \prod_{i=1}^n \left(\frac{\exp(m(x'_i \boldsymbol{\beta}))}{1 + \exp(m(x'_i \boldsymbol{\beta}))} \right)^{y_i} \left(\frac{1}{1 + \exp(m(x'_i \boldsymbol{\beta}))} \right)^{1-y_i} \quad (3 - 8)$$

As in (Choi et al. (2011)) and (Gramacy and Lian (2012)), the Gaussian process prior distribution is considered as a prior for the unknown nonparametric link function $m(\cdot)$. More specially, the previous distribution of $m(\cdot)$ is GP, with zero mean and square exponential covariance function is written as follows:

$$m \sim GP(0, C(\cdot, \cdot)), \quad \text{where} \quad C(x, x') = \partial \exp \frac{-(x - x')^2}{w}$$

where ∂ and w are Hyperparameters. Writing this out in the single-index model framework using the observed covariates x_i , we have,

$$\pi(m_n | \beta, \partial) = \det[C_n]^{-1/2} \exp \left\{ - \frac{m_n' C_n^{-1} m_n}{2} \right\}$$

C_n is the covariance matrix with dimension $(n \times n)$ and elements $C(\cdot, \cdot)$ given in Equation

$$C(x_i, x_j) = \partial \exp\{-(x_i - x_j)' \beta \beta' (x_i - x_j)/w\}$$

Follow Gramacy and Lian (2012), and when we use the Gaussian process as a prior distribution to the nonparametric link function, then $\frac{\beta}{\sqrt{w}}$ is identifiable without the necessity for the constraint $\|\beta\| = 1$. Therefore, we will instead of $\frac{\beta}{\sqrt{w}}$ by β and the covariance function is reformulated as follows:

$$C(x_i, x_j) = \partial \exp\{-(x_i' \beta - x_j' \beta)^2\} \quad (3 - 9)$$

The inverse gamma distribution is set as a hyper prior for which implies that $\partial \sim \text{Inv. Gamma}(a_\partial, b_\partial)$ where a_∂ and b_∂ are the hyperparameters.

Follow Wilhelmsen *et al.* (2009) and Ziemba (2005) we will set a normal distribution as prior for the model parameters vector $\beta \sim N(0, \sigma^2)$, the inverse gamma distribution is set as a prior for $\sigma^2 \sim \text{Inv. Gamma}(a_\sigma, b_\sigma)$.

3-4-1-1 Hierarchical model and MCMC algorithm

The hierarchic model for the Bayesian single index logistic regression can be written as follows (3-10):

$$f(y|m, \beta, \partial, \sigma^2) = \prod_{i=1}^n \left(\frac{\exp(m(x_i' \beta))}{1 + \exp(m(x_i' \beta))} \right)^{y_i} \left(1 - \frac{\exp(m(x_i' \beta))}{1 + \exp(m(x_i' \beta))} \right)^{1-y_i}$$

$$m|\beta, \partial \sim GP(0, C(\cdot, \cdot)),$$

$$\beta|\sigma^2 \sim N(0, \sigma^2) \quad (3 - 10)$$

$$\partial \sim \text{Inv. Gamma}(a_\partial, b_\partial)$$

$$\sigma^2 \sim \text{Inv. Gamma}(a_\sigma, b_\sigma)$$

Based on the Bayesian hierarchical model (3 – 10), the full conditional posterior distributions are as follows:

$$\begin{aligned}
 p(m, x_i, \beta, \partial, \sigma^2 | y) & \\
 & \propto \prod_{i=1}^n \left(\frac{\exp(m(x'_i \beta))}{1 + \exp(m(x'_i \beta))} \right)^{y_i} \left(1 - \frac{\exp(m(x'_i \beta))}{1 + \exp(m(x'_i \beta))} \right)^{1-y_i} \\
 & \times \det[C_n]^{-1/2} \exp \left\{ -\frac{m_n' C_n^{-1} m_n}{2} \right\} \times \prod_{j=1}^p \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{(\beta_j)^2}{2\sigma^2} \right\} \\
 & \times \left(\frac{1}{\sigma^2} \right)^{a_{\sigma}+1} \exp \left\{ -\frac{b_{\sigma}}{\sigma^2} \right\} \times \left(\frac{1}{\partial} \right)^{a_{\partial}+1} \exp \left\{ \frac{b_{\partial}}{\partial} \right\}
 \end{aligned}$$

The conditional posterior distributions for all parameters can easily derived for the Bayesian single index logistic regression:

Sample the link function $m | x_i, \beta, \partial, \sigma^2, y$ from the following conditional posterior distribution:

$$\begin{aligned}
 \pi(m | x_i, \beta, \partial) & \propto \prod_{i=1}^n \left(\frac{\exp(m(x'_i \beta))}{1 + \exp(m(x'_i \beta))} \right)^{y_i} \left(1 - \frac{\exp(m(x'_i \beta))}{1 + \exp(m(x'_i \beta))} \right)^{1-y_i} \\
 & \times \det[C_n]^{-1/2} \exp \left\{ -\frac{m_n' C_n^{-1} m_n}{2} \right\}
 \end{aligned}$$

Sample the parameters vector $\beta | m, x_i, \partial, \sigma^2, y$ from the following conditional posterior distribution:

$$\begin{aligned} \pi(\beta|m, \sigma^2) & \propto \prod_{i=1}^n \left(\frac{\exp(m(x'_i\beta))}{1 + \exp(m(x'_i\beta))} \right)^{y_i} \left(1 - \frac{\exp(m(x'_i\beta))}{1 + \exp(m(x'_i\beta))} \right)^{1-y_i} \\ & \times \det[C_n]^{-1/2} \exp \left\{ -\frac{m_n' C_n^{-1} m_n}{2} \right\} \\ & \times \prod_{j=1}^p \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{(\beta)^2}{2\sigma^2} \right\} \end{aligned}$$

The conditional distribution of $\sigma^2|m, x_i, \partial, \beta, y$ is the Inverse Gamma distribution

$$\pi(\sigma^2|\beta) \propto \prod_{j=1}^p \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{(\beta)^2}{2\sigma^2} \right\} \times \left(\frac{1}{\sigma^2} \right)^{a_\sigma+1} \exp \left\{ -\frac{b_\sigma}{\sigma^2} \right\}$$

So that we will sample $\sigma^2|m, x_i, \partial, \beta, y$ from the following conditional posterior distribution $IG\left(\frac{p}{2} + a_\sigma, \frac{\sum \beta^2}{2} + b_\sigma\right)$

We will sample ∂ from the following posterior conditional distribution

$$\pi(\partial|m_n) \propto \det[C_n]^{-1/2} \exp \left\{ -\frac{m_n' C_n^{-1} m_n}{2} \right\} \times \left(\frac{1}{\partial} \right)^{a_\partial+1} \exp \left\{ \frac{b_\partial}{\partial} \right\}$$

An efficient Gibbs sampler algorithm is used to sample σ^2 , whereas a Metropolis-Hastings algorithm is used to sample β_τ, m_n and ∂ . We set the initial values for the hyperparameters $a_\sigma, b_\sigma, a_\partial$ and b_∂ as (0.1).

3-4-2. Bayesian Variable Selection Semiparametric Logistic Regression

Subset selection by regularization has attracted much interest recently (see for example, lasso (least absolute shrinkage and selection operator) by Tibshirani, 1996). Tibshirani, R. (1996) proposed that lasso estimates will be taken as posterior mode estimates once the regression parameters are assigned independent and corresponding standard. Park and Casella (2008) introduced the Bayesian lasso regression, using a conditional Laplace prior distribution represented as a scale mixture of normal with an exponential mixing distribution. Bayesian analysis method has become very widely applicable, as a result of its ability to benefit from all available information in the analysis.

Bayesian variable selection is a flexible method for translating prior information into a selection of variables (Fridley, 2009). Several variable selection methods are used with a Bayesian framework.

In this proposed method Gaussian process is considered as prior for the unknown link function. we will set a Laplace distribution as prior for the model parameters vector β . The gamma distribution as prior for shrinkage parameter λ .

3-4-2-1. Hierarchical model Posterior distribution

Bayesian hierarchical model for single index logistic regression model regularize by lasso is provided as follows:

$$f(\mathbf{y} | \mathbf{m}, \boldsymbol{\beta}, \partial, \sigma, \lambda) = \prod_{i=1}^n \left(\frac{\exp(m(x'_i \boldsymbol{\beta}))}{1 + \exp(m(x'_i \boldsymbol{\beta}))} \right)^{y_i} \left(1 - \frac{\exp(m(x'_i \boldsymbol{\beta}))}{1 + \exp(m(x'_i \boldsymbol{\beta}))} \right)^{1-y_i}$$

$$m|\beta, \partial \sim GP(0, C(.,.)),$$

$$\beta|\sigma, \lambda = \prod_{j=1}^p \frac{\lambda}{2\sigma} \exp(-\lambda|\beta_j|/\sigma) \quad (3 - 11)$$

$$\partial \sim Inv. Gamma(a, b)$$

$$\sigma \sim Inv. Gamma(e, f)$$

$$\lambda \sim Gamma(c, d)$$

By using MCMC algorithm the researchers have found the conditional distribution for all parameters. The conditional posterior distribution for all parameters has been derived as follows:

- link function $m|x_i, \beta, \partial, \sigma, \lambda, y$ can be sample from the following conditional distribution:

$$\pi(m|x_i, \beta, \partial)$$

$$\begin{aligned} &\propto \prod_{i=1}^n \left(\frac{\exp(m(x'_i\beta))}{1 + \exp(m(x'_i\beta))} \right)^{y_i} \left(1 - \frac{\exp(m(x'_i\beta))}{1 + \exp(m(x'_i\beta))} \right)^{1-y_i} \\ &\times \det[C]^{-1/2} \exp \left\{ -\frac{m' C^{-1} m}{2} \right\} \end{aligned}$$

- the conditional distribution of the parameter vector can be shown as:

$$\pi(\beta|x_i, m, \lambda, \sigma, y)$$

$$\begin{aligned} &\propto \prod_{i=1}^n \left(\frac{\exp(m(x'_i\beta))}{1 + \exp(m(x'_i\beta))} \right)^{y_i} \left(1 - \frac{\exp(m(x'_i\beta))}{1 + \exp(m(x'_i\beta))} \right)^{1-y_i} \\ &\times \det[C]^{-1/2} \exp \left\{ -\frac{m' C^{-1} m}{2} \right\} \times \prod_{j=1}^p \frac{\lambda}{2\sigma} \exp(-\lambda|\beta_j|/\sigma) \end{aligned}$$

- The conditional distribution function of λ can be written as:

$$\pi(\lambda | x_i, \mathbf{m}, \boldsymbol{\beta}, \partial, \sigma, \mathbf{y}) \propto \prod_{j=1}^p \frac{\lambda}{2\sigma} \exp(-\lambda |\beta_j| / \sigma) \times \lambda^{c-1} \exp(-d\lambda)$$

Therefore, the conditional posterior of λ is Gamma distribution ($p + c, d + \sum |\beta_j| / \sigma$).

- The conditional distribution of ∂ is given as:

$$\pi(\partial | x_i, \mathbf{m}, \boldsymbol{\beta}, \lambda, \sigma, \mathbf{y}) \propto \det[C]^{-1/2} \exp\left\{-\frac{\mathbf{m}' C^{-1} \mathbf{m}}{2}\right\} \times \left(\frac{1}{\partial}\right)^{a,+1} \exp\left\{-\frac{b}{\partial}\right\}$$

- The conditional distribution σ is given as:

$$\pi(\sigma | x_i, \mathbf{m}, \boldsymbol{\beta}, \lambda, \partial, \mathbf{y}) \propto \prod_{j=1}^p \frac{\lambda}{2\sigma} \exp(-\lambda |\beta_j| / \sigma) \times \left(\frac{1}{\sigma}\right)^{e+1} \exp\left\{-\frac{f}{\sigma}\right\}$$

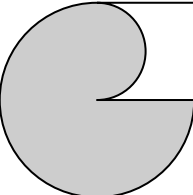
The posterior distribution of σ is Inverse Gamma ($p + e, f + \lambda \sum |\beta_j|$)

An efficient Gibbs sampler algorithm is considered to sample σ and λ , whereas a Metropolis-Hastings algorithm is used to sample $\boldsymbol{\beta}_\tau, \mathbf{m}_n$ and ∂ . The researchers set the initial values for the hyper parameters a, b, c, d, e and f as (0.1).



CHAPTER FOUR

Simulation

- Example 1
 - Example 2
 - Example 3
- 

4.1 Simulation study

In this chapter simulation examples are considered to evaluate the performance of our proposed methods Bayesian semiparametric logistic regression (BSLR) and Bayesian semiparametric lasso logistic regression (BSLLR). We have compared our proposed method to some existing methods Bayesian Logistic regression (BLR), Bayesian Probit regression (BPR) these functions are reported in MCMC pack R package and Bayesian Binary Quantile regression BBQR ($\tau = 0.5$) was reported in bayesQR R package. We have used three simulation examples as same as the examples that are used by (Hu et al. (2013), Alshaybawee et al. (2016), Zhao and Lian (2015), Alkenani and Yu (2013), Lv et al. (2014) and Kuruwita (2015)). We have constructed an R code to implement MCMC algorithm. MCMC algorithm are run 20000 iteration and remove the first 4000 as burn in.

4-1-1. Example 1

In this example, three samples size is considered (N=50, 150 and 250) and the following model used to generate our data:

$$y^* = g(t) + 0.1\varepsilon, \quad g(t) = \sin\left\{\frac{\pi(t - H)}{M - H}\right\},$$

$$y_i = \begin{cases} 1 & \text{if } y_i^* \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Where $t = \mathbf{x}_i^T \boldsymbol{\beta}$, $i = 1, 2, \dots, 10$, distributed as uniform $[0, 1]$, $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_{10})^T = \frac{1}{\sqrt{3}}(1, 1, 0, 0, 1, 0, 0, 0, 0, 0)^T$, $H = \frac{\sqrt{3}}{2} - \frac{1.645}{\sqrt{12}}$, $M = \frac{\sqrt{3}}{2} + \frac{1.645}{\sqrt{12}}$ and the error term ε_i is distributed as standard normal $N(0, 1)$. The result of this study based on 100 replications for each sample size.

Table (4-1). The average SD of the parameter estimates of BSLR, BSLLR, BLR, BPR and BBQR based on 100 replications when $n=(50,150,250)$ (Example 1)

N	Methods	$SD. \beta_1$	$SD. \beta_2$	$SD. \beta_3$	$SD. \beta_4$	$SD. \beta_5$	$SD. \beta_6$	$SD. \beta_7$	$SD. \beta_8$	$SD. \beta_9$	$SD. \beta_{10}$
50	BSLR	0.2384	0.1857	0.3904	0.1996	0.2014	0.5506	0.3317	0.2548	0.4536	0.5763
	BSLLR	0.3618	0.1421	0.3189	0.1713	0.2230	0.3240	0.1193	0.1835	0.2660	0.1252
	BLR	182.8170	74.6994	387.1729	199.7872	77.8861	61.3603	397.4594	238.5268	118.2653	251.8660
	BPR	52.9347	6.7497	134.0652	72.4283	32.8346	42.9802	134.5749	73.8978	40.3407	89.5585
	BBQR	2.7155	3.2060	4.9196	1.0394	4.5203	1.7437	2.1512	1.8854	1.0583	2.6110
150	BSLR	0.3130	0.1433	0.1847	0.1661	0.2297	0.3639	0.4360	0.3138	0.3521	0.1282
	BSLLR	0.2099	0.1372	0.1424	0.3313	0.1769	0.2625	0.2376	0.2298	0.2944	0.1837
	BLR	1.0509	0.5542	0.9899	0.2576	0.5582	0.9152	0.2375	0.9238	0.5833	1.5527
	BPR	0.4968	0.3991	0.4720	0.1275	0.1897	0.4709	0.2180	0.4099	0.1984	0.7903
	BBQR	1.8866	1.0541	2.3090	1.2864	0.9131	0.6188	1.7602	0.7019	1.6567	1.2298
250	BSLR	0.2362	0.2268	0.0807	0.1359	0.2212	0.1490	0.1688	0.2073	0.2408	0.1133
	BSLLR	0.1282	0.1298	0.1947	0.1105	0.0708	0.0671	0.1179	0.2242	0.2155	0.1102
	BLR	0.4834	0.4497	0.3030	0.3445	0.5416	0.3714	0.2010	0.5284	0.5704	0.3499
	BPR	0.2892	0.1644	0.1592	0.1621	0.2904	0.1878	0.1890	0.2529	0.2608	0.2686
	BBQR	0.5066	0.4573	0.3859	0.6919	0.3781	0.8217	0.6724	0.4177	0.4756	0.4015

Table (4-1) shows the standard division to the estimates parameters that estimate by the proposed and existing methods BSLLR ,BSLR, BLR,BPR and BBQR at three samples size 50, 150 and 250. We can see that the proposed method BSLLR and BSLR have get the smallest values over all the samples that mean this method more consistent compared to the other methods. BLR method has get high values of SD when the

Chapter four :

sample small (N=50) but these values are decreased when the sample size increase. The other two methods BBQR and BPR have high values of SD when the sample is small whereas these values decrease when N=150 and 250. Over all samples SD values of BPR method smaller than the values of BLR method.

Table (4-2) : Bias of the parameter estimates of BSLR ,BSLLR, BLR,BPR and BBQR based on 100 replications when n=(50,150,250)(Example 1)

N	Methods	<i>Bias. β_1</i>	<i>Bias. β_2</i>	<i>Bias. β_3</i>	<i>Bias. β_4</i>	<i>Bias. β_5</i>	<i>Bias. β_6</i>	<i>Bias. β_7</i>	<i>Bias. β_8</i>	<i>Bias. β_9</i>	<i>Bias. β_{10}</i>
50	BSLR	0.4565	0.20328	0.00746	0.06829	0.50394	0.19189	0.18210	0.15846	0.17762	0.02334
	BSLLR	0.3051	0.16552	0.00594	0.00447	0.42641	0.10280	0.15709	0.07988	0.16041	0.01191
	BLR	24.8464	17.4713	40.5725	20.5693	14.2326	12.01385	22.0306	17.2716	17.0540	15.3543
	BPR	9.5992	11.68162	10.8771	7.39331	8.33325	3.83632	8.3044	7.79916	9.65791	7.14968
	BBQR	6.9934	6.69474	1.84589	0.60420	3.80963	6.24124	0.58593	1.25144	1.57506	0.48388
150	BSLR	0.46727	0.25825	0.05687	0.13870	0.47328	0.01193	0.12769	0.01857	0.27568	0.06768
	BSLLR	0.63888	0.23519	0.18552	0.06368	0.24563	0.02205	0.23559	0.15214	0.12725	0.01565
	BLR	0.05600	0.54797	0.70250	0.26318	0.82760	0.89859	1.88796	1.05063	0.32237	2.13883
	BPR	0.42027	0.04728	0.36790	0.21707	0.73279	0.35535	1.02857	0.45824	0.32517	1.15661
	BBQR	0.05303	1.17797	0.75250	0.18650	1.66633	0.53418	2.52137	0.96667	0.10533	1.91601
250	BSLR	0.46729	0.38103	0.12891	0.14048	0.39222	0.07275	0.01337	0.13508	0.13930	0.12263
	BSLLR	0.28055	0.27641	0.09868	0.02908	0.20551	0.05337	0.01647	0.09038	0.10834	0.11370
	BLR	1.09552	1.10116	0.87614	0.03731	1.24578	0.61162	0.43838	0.69809	0.82416	0.97952
	BPR	0.75996	0.74141	0.42661	0.00645	0.85092	0.27681	0.21368	0.36279	0.53036	0.33877
	BBQR	0.52872	0.78159	1.79664	0.82294	0.83474	2.58975	0.51931	0.13129	0.71421	0.49498

In Table (4-2) we summarize the bias to the parameters that are estimated by all methods under study, the existing methods BLR,BPR and BBQR and proposed methods BSLR BSLLR. At the three samples size we can see that very clearly the proposed method get the smallest values of bias for all estimated parameters that mean the estimated parameters are very close to the true parameters. On the other hand, we can see that the BLR method get the largest values of bias when the sample small (N=50) but these values are decreased when the sample size increase. For the other methods, we can see that the BBQR method get bias values smaller than the BPR methods for most estimated parameters and at all samples size.

Table (4-3). The values of MSE and MAE of BSLLR,BSLR, BLR, BPR and BBQR methods for each sample (Simulated Example 1).

N	Methods	MSE	MAE
50	BSLR	6.543509	0.7653807
	BSLLR	3.418928	0.5615657
	BLR	12.60041	5.6988078
	BPR	9.457564	4.7282388
	BBQR	8.586480	6.0059584
150	BSLR	0.2805301	0.4934567
	BSLLR	0.1077895	0.3282214
	BLR	3.5822892	1.6687781
	BPR	3.2703202	1.3993889
	BBQR	9.6605426	2.8949750
250	BSLR	0.3294628	0.5637773
	BSLLR	0.1899602	0.4124804
	BLR	2.5053917	1.5057131
	BPR	1.2311219	1.0644121
	BBQR	6.2452349	2.2959856

Table (4-3) shows MSE and MAE values for all methods in this study. The proposed methods BSLLR,BSLR get the smallest values of MSE and MAE compare to the other methods. The existing method BBQR gets the largest values of MSE and MAE compare to the other two methods BLR and BPR when the samples size (150 and 250) but it is smaller than these methods when the sample size is 50. BPR method get MSE and MAE values smaller than BLR method at all cases.

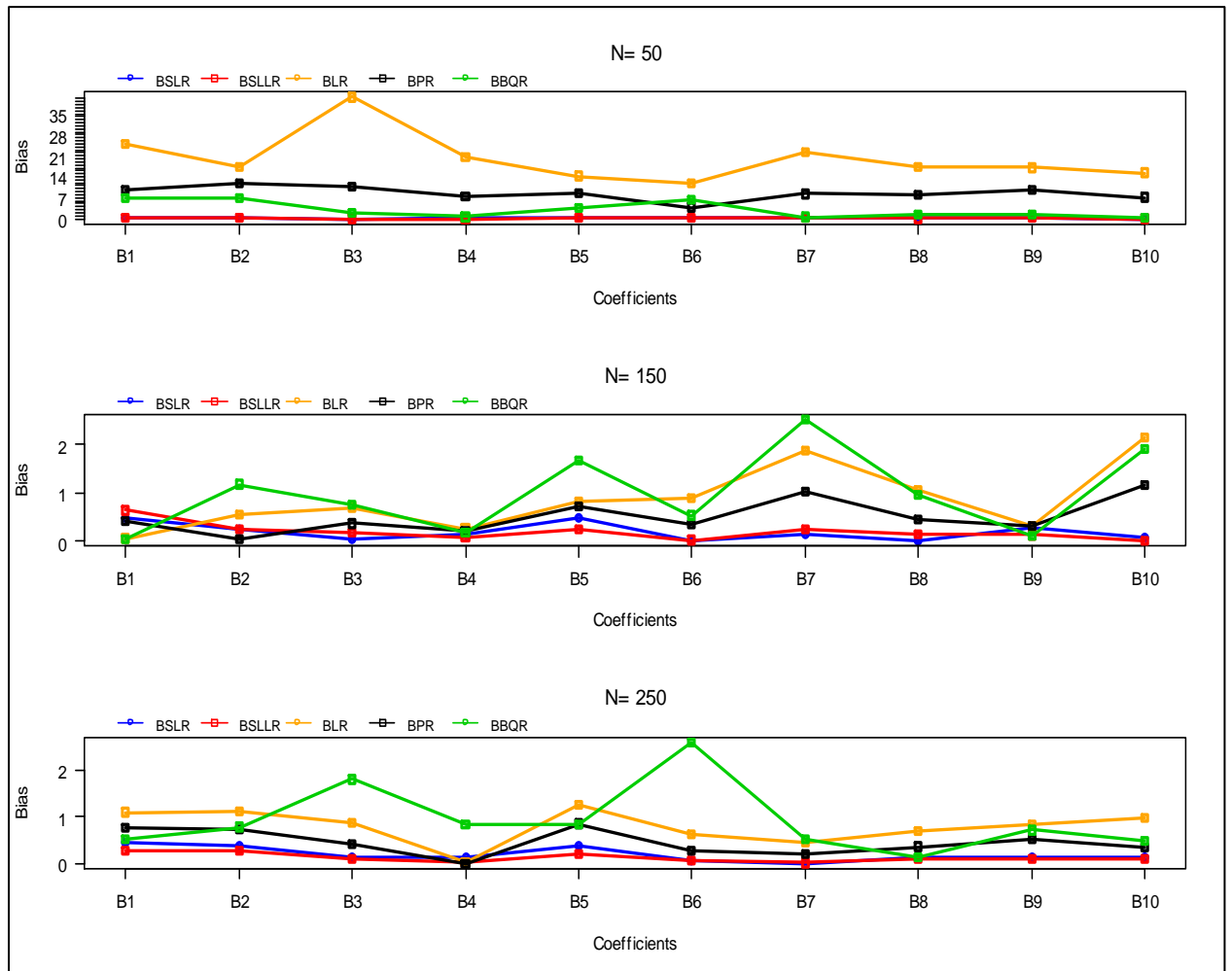


Figure (4-1). shows the bias values for BSLR, BSLLR, BLR, BPR and BBQR methods at three samples size (Example 1).

As same as Figure (4-1) shows that the proposed method BSLLR get the smallest bias for most parameter . Followed by the proposed BSLR method. In addition, the BLR method get the largest bias for all parameter estimates and when sample size($n= 50$). Also we can see that when the sample size increases, the bias for the BLQR increases.

4-1-2 Example:2

In this example, data are generated with three samples size (N=50, 100 and 250) from the following model:

$$y_i^* = g(t) + \sqrt{(\sin(t) + 1)} \varepsilon, \text{ where } g(t) = 10 \sin(0.75t), \quad y_i = \begin{cases} 1 & \text{if } y_i^* \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Where $t = \mathbf{x}_i^T \boldsymbol{\beta}$ ($i = 1,2, \dots,6$) are i.i.d. generated from a normal distribution $N[0, (1/4)^2]$, $\boldsymbol{\beta} = \frac{1}{\sqrt{5}}(0,1,0,0,1,0)^T$, and standard normal distribution use for the error term.

Table (4-4) .The average SD of the parameter estimates of BSLR ,BSLLR, BLR,BPR and BBQR based on 100 replications when n=(50,150,250)(Example 2)

N	Methods	SD. β_1	SD. β_2	SD. β_3	SD. β_4	SD. β_5	SD. β_6
50	BSLR	0.2885311	0.2266289	0.4793101	0.1845194	0.39226222	0.1755926
	BSLLR	0.1883157	0.3053995	0.2173249	0.1441359	0.08978273	0.0608119
	BLR	10.1661866	11.7392847	6.3448430	9.9168538	7.28521899	4.4024231
	BPR	5.3680380	6.5542677	3.8626995	5.3547960	3.88658103	2.3372606
	BBQR	7.2353892	9.0316138	4.3981164	8.4775600	5.26683087	3.9120461
150	BSLR	0.3511978	0.1410557	0.2068314	0.2933123	0.2249511	0.3088045
	BSLLR	0.3129664	0.1118578	0.1794991	0.2537305	0.2169618	0.2281425
	BLR	2.8701891	3.3438091	4.1820764	3.4957595	1.5805986	4.0580278
	BPR	1.6145123	1.8290660	2.5104131	2.0392099	0.8910264	2.4318003
	BBQR	3.0832892	4.0249954	4.0492844	3.8745189	1.4502137	5.2580175
250	BSLR	0.2920818	0.3232477	0.3863615	0.4507027	0.3459609	0.2682559
	BSLLR	0.2551161	0.2333616	0.1869012	0.1741956	0.3201834	0.2616635
	BLR	3.1323130	5.0057876	4.2350599	4.0338804	3.9066599	3.4286648
	BPR	1.6100048	2.7778162	2.3422853	2.3634299	2.3137920	1.9545385
	BBQR	2.9310397	6.0947630	4.4404595	4.0257083	4.7411032	3.2821200

Table (4-4) shows the standard deviation to the estimates parameters that estimate by the proposed and existing methods BSLLR ,BSLR, BLR,BPR and BBQR at three samples size 50, 150 and 250. We can see that the proposed method BSLLR and BSLR have get the smallest values over all the samples that mean this method more consistent compared to the other methods. BLR method has got high values of SD when the sample small (N=50) but these values are decreased when the sample size increase. The other two methods BBQR and BPR have high values of SD when the sample is small whereas these values decrease when N=150 and 250. Over all samples SD values of BPR method smaller than the values of BBQR method.

Table (4-5) :Bias of the parameter estimates of BSLR ,BSLLR, BLR,BPR and BBQR based on 100 replications when n=(50,150,250)(Example 2)

N	Methods	<i>Bias. β_1</i>	<i>Bias. β_2</i>	<i>Bias. β_3</i>	<i>Bias. β_4</i>	<i>Bias. β_5</i>	<i>Bias. β_6</i>
50	BSLR	0.26699515	0.7966988	0.51909302	0.3857604	0.2905819	0.30782838
	BSLLR	0.1767178	0.4973725	0.2699876	0.30846286	0.2727542	0.4199030
	BLR	7.45684148	11.0708382	1.87629177	4.5483811	4.9526617	4.58174800
	BPR	4.02339060	5.9492531	1.08940267	2.4292862	2.3746975	2.39837699
	BBQR	5.39733793	8.0652850	1.15491116	4.2964429	4.1158289	2.92241662
150	BSLR	0.1767178	0.4973725	0.2699876	0.30846286	0.2727542	0.4199030
	BSLLR	0.1054145	0.1481568	0.1127463	0.02869436	0.0944716	0.1797773
	BLR	1.1453204	13.930302	0.8187995	0.94511981	5.7732122	2.9468537
	BPR	0.7092354	7.9175284	0.5107800	0.58591506	3.2559084	1.7310921
	BBQR	1.4729005	16.431547	0.6240802	1.06342392	5.4612556	3.5349964
250	BSLR	0.1767178	0.4973725	0.2699876	0.30846286	0.2727542	0.4199030
	BSLLR	0.09021105	0.2121002	0.13553744	0.1204466	0.2179308	0.1987142
	BLR	2.47157904	9.9645115	0.52388608	1.2157478	6.0840021	0.6215322
	BPR	1.42960515	5.3674101	0.23826074	0.7146429	3.1499358	0.3124599
	BBQR	2.37095908	11.1744380	0.56601928	0.8773253	6.1934518	0.6491614

In Table (4-5) we summarize the bias to the parameters that are estimated by all methods under study, the existing methods BLR,BPR and BBQR and proposed methods BSLR BSLLR. At the three samples size, we can see that very clearly the proposed method gets the smallest values of bias for all estimated parameters that means the estimated parameters are very close to the true parameters. On the other hand, we can see that the BLR method get the largest values of bias when the sample small (N=50) but these values are decrease when the sample size increases. For the other methods, we can see that the BPR method got bias values smaller than the BBQR methods for most estimated parameters and at all samples size.

Table (4-6). The values of MSE and MAE of BSLLR,BSLR, BLR, BPR and BBQR methods for each sample (Simulated Example 2).

N	Methods	MSE	MAE
50	BSLR	0.3495810	0.3881093
	BSLLR	0.3427362	0.3760001
	BLR	1.0650195	0.8131756
	BPR	0.4811834	0.5704363
	BBQR	0.6912490	0.6671856
150	BSLR	0.4481509	0.5697954
	BSLLR	0.3453956	0.4681879
	BLR	1.1036038	0.8476163
	BPR	0.6492208	0.6699304
	BBQR	1.3548498	0.9258817
250	BSLR	0.4767392	0.5480149
	BSLLR	0.41319028	0.4678268
	BLR	0.9795460	0.8042298
	BPR	0.6384873	0.6527334
	BBQR	1.0698380	0.8405053

We reported the values of the MSE and MAE In Table (4-6) for all methods in this study.

The MSE and MAE values for the proposed methods are the smallest compared to the other three existing methods at all samples. BPR method got small MSE and MAE compare to BLR and BBQR methods at all samples. MSE and MAE for BLR method are smaller than BBQR method in the case of $N=50, N=150$.

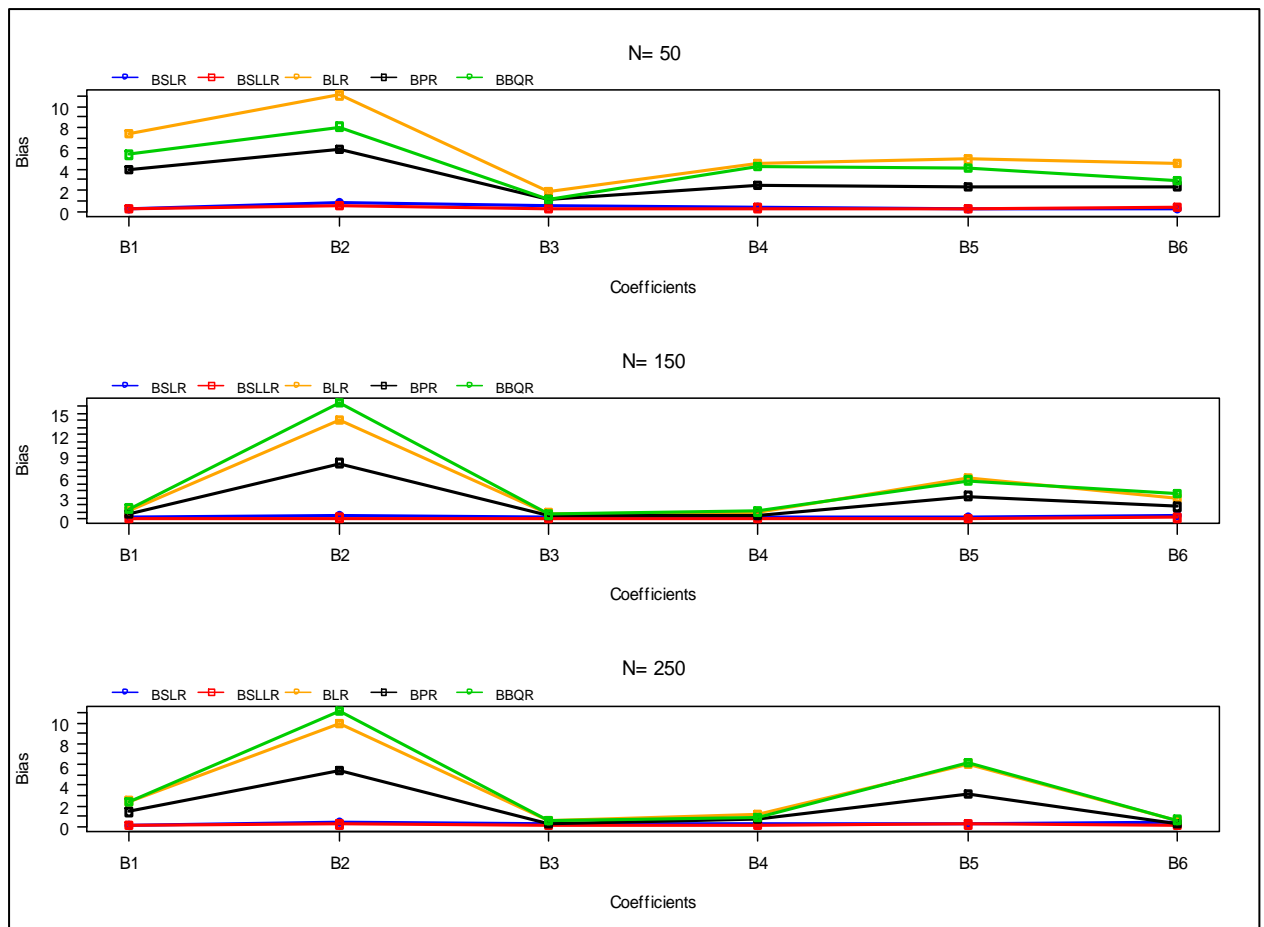


Figure (4-2). show the bias values for BSLR, BSLLR, BLR, BPR and BBQR methods at three samples size (Example 2).

4-1-3.Example 3:

Three samples size (N=50, 150 and 250) with 100 replications are generated from the following regression model:

$$y_i^* = g(x_i^T \beta) + \varepsilon, \quad y_i = \begin{cases} 1 & \text{if } y_i^* \geq 0 \\ 0 & \text{otherwise} \end{cases},$$

Where $g(t) = \exp(t)$, $x_i (i = 1,2, \dots,5)$ where the independent variables are generated from $N(0,1)$, $\beta = (\beta_1, \beta_2, \dots, \beta_5)^T = \frac{1}{\sqrt{3}}(1,1,0,0,1)^T$ and the error term generated from standard normal distribution.

Table (4-7) .The average SD of the parameter estimates of BSLR ,BSLLR, BLR,BPR and BBQR based on 100 replications when n=(50,150,250)(Example 3)

N	Methods	SD. β_1	SD. β_2	SD. β_3	SD. β_4	SD. β_5
50	BSLR	0.3779795	0.1923527	0.2751070	0.2453410	0.3561139
	BSLLR	0.2331255	0.1008542	0.2492716	0.2072045	0.3061027
	BLR	86.0604362	70.5038761	37.9116530	5.5831063	4.9809374
	BPR	17.9860155	19.7405085	14.6407443	4.1116885	8.9315752
	BBQR	4.4912284	3.3778778	1.7038149	2.3932587	3.3801078
150	BSLR	0.4724897	0.07640786	0.4719294	0.4225184	0.3882771
	BSLLR	0.4297358	0.21896945	0.3404264	0.1535764	0.2816215
	BLR	0.9698475	0.65539688	0.5872489	0.9636557	1.2316414
	BPR	0.5521709	0.30773616	0.5041294	0.4949085	0.4771225
	BBQR	0.8810987	0.56151079	1.4873602	1.8510636	0.7756866
250	BSLR	0.2788696	0.3109719	0.1565683	0.1371485	0.2577051
	BSLLR	0.2149166	0.2083046	0.1197567	0.1462992	0.2178211
	BLR	0.5019365	0.5731193	0.2172868	0.3370688	0.4172520
	BPR	0.4452285	0.3687280	0.2074063	0.2869451	0.3345502
	BBQR	0.6844260	0.6896022	0.4415480	0.5806473	0.4550241

Table (4-7) shows the standard division to the estimates parameters that estimate by the proposed and existing methods BSLLR ,BSLR, BLR,BPR and BBQR at three samples size 50, 150 and 250. We can see that the proposed method BSLLR and BSLR have get the smallest values over all the samples that mean this method more consistent compare to the other methods. BLR method has got high values of SD when the sample small (N=50) but these values are decreased when the sample size increase. The other two methods BBQR and BPR have high values of SD when the sample is small whereas these values decrease when N=150 and 250. Over all samples SD values of BPR method smaller than the values of BBQR method when N=150,250.

Table (4-8) : Bias of the parameter estimates of BSLR ,BSLLR, BLR,BPR and BBQR based on 100 replications when n=(50,150,250)(Example 3)

N	Methods	<i>Bias. β_1</i>	<i>Bias. β_2</i>	<i>Bias. β_3</i>	<i>Bias. β_4</i>	<i>Bias. β_5</i>
50	BSLR	0.3654700	0.3471845	0.04033032	0.10648746	0.7121747
	BSLLR	0.1749267	0.1041210	0.13912011	0.04160419	0.4745929
	BLR	12.3633514	9.3379315	10.19195796	10.15885297	11.8380268
	BPR	6.3335796	5.8199950	7.19230134	5.97017943	7.4432875
	BBQR	6.3481030	1.3731649	4.58732804	2.02696842	2.8459228
150	BSLR	0.2545773	0.2048565	0.22627144	0.06867267	0.3635006
	BSLLR	0.1576164	0.1424606	0.04170413	0.07613136	0.2122997
	BLR	2.4276604	1.9546028	0.18913836	0.87672881	2.2457348
	BPR	0.6398191	0.3010055	0.21612929	0.52864735	0.3821795
	BBQR	3.2447763	1.8441018	1.17326747	0.18052870	3.1885781
250	BSLR	0.3255397	0.1003537	0.34651618	0.16604541	0.2841740
	BSLLR	0.2055563	0.1317658	0.02213048	0.01012073	0.1391845
	BLR	1.8516120	1.5708966	0.87933972	0.40713401	1.3835634
	BPR	1.0971777	0.9210109	0.44643393	0.17997494	0.8569160
	BBQR	2.3399300	1.8162443	1.69752515	1.12259031	1.4193984

In Table (4-8) we summarize the bias to the parameters that estimate by all methods under study, the existing methods BLR,BPR and BBQR and proposed methods BSLR BSLLR. At the three samples size we can see that very clearly the proposed method get the smallest values of bias for all estimated parameters that mean the estimated parameters are very close to the true parameters. On the other hand, we can see that the BLR method gets the largest values of bias when the sample small (N=50) but these values are decreased when the sample size increase. For the other methods we can see that the BBQR method get bias values smaller than the BPR methods for most estimated parameters when(N=50).

Table (4-9). The values of MSE and MAE of BSLLR,BSLR, BLR, BPR and BBQR methods for each sample (Simulated Example 3).

N	Methods	MSE	MAE
50	BSLR	6.358681	2.5728602
	BSLLR	5.834332	2.2684232
	BLR	22.05150	13.4242123
	BPR	14.71437	6.8327421
	BBQR	8.102702	3.8911789
150	BSLR	0.4558037	0.6479127
	BSLLR	0.2101322	0.4326607
	BLR	13.8081717	3.4640445
	BPR	7.8910163	2.7837473
	BBQR	15.3014884	3.5209226
250	BSLR	0.4922937	0.6793419
	BSLLR	0.1933612	0.3310181
	BLR	3.6290291	1.8119455
	BPR	1.4290452	1.1472240
	BBQR	2.8464702	1.4863885

Table (4-9) shows MSE and MAE values for all methods in this study. The proposed methods BSLLR,BSLR get the smallest values of MSE and MAE compare to the other methods. The existing method BBQR gets the largest values of MSE and MAE compare to the other two methods BLR and BPR when the samples size (150 and 250) but it is smaller than these methods when the sample size is 50. BPR method gets MSE and MAE values smaller than BLR method at all cases.

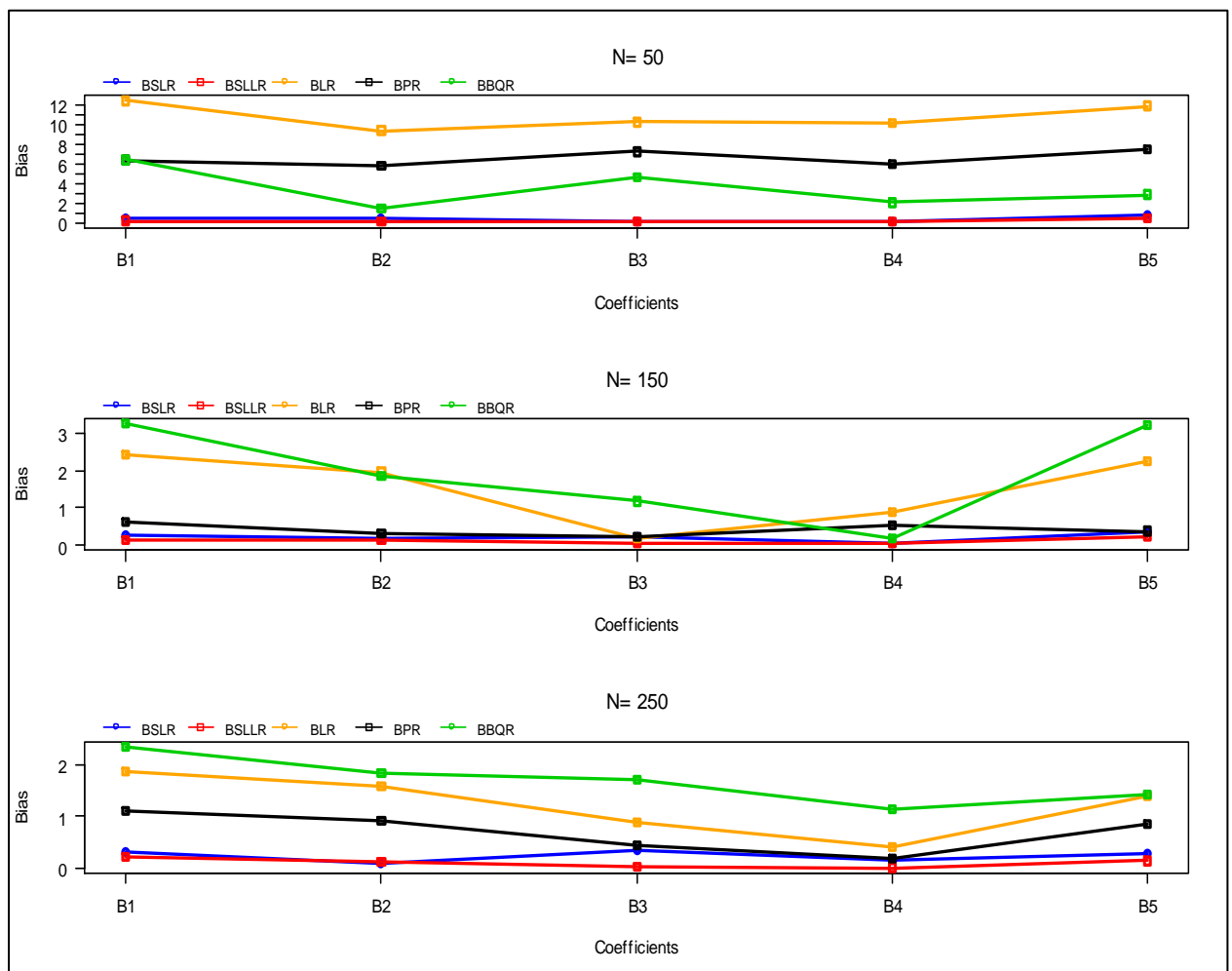


Figure (4-3). shows the bias values for BSLR, BSLLR, BLR, BPR and BBQR methods at three samples size (Example 3).

A decorative graphic at the top of the page, resembling a scroll. It consists of a horizontal line with rounded ends, and a small, shaded, teardrop-shaped element on the left side that overlaps the line.

CHAPTER FIVE

Real data analysis

A decorative graphic at the bottom left of the page, resembling a scroll. It consists of a horizontal line with rounded ends, and a large, shaded, teardrop-shaped element on the left side that overlaps the line.

5-1. Real data analysis

COVID-19 is the disease caused by Coronavirus called SARS-CoV-2. This virus was discovered by the World Health Organization (WHO) on December 31, 2019, after many complaints received from Wuhan, for people who had severe pneumonia, in PRC

The news was spread, about 150 countries closed all schools, imposed the cancellation of gatherings and events, and more than 80 countries closed all workplaces to contain and control the spread of the virus. Travel restrictions have been imposed on citizens in many countries of the world. The forced closure by governments, in addition to the automatic imposition of social distancing by consumers and producers allowed to work outside, had a significant impact on activity and trade in the world, accompanied by fluctuations in financial markets, and a sharp decline in the prices of oil and other industrial mineral Global stock markets fell on February 24, 2020, due to the significant rise in cases of coronavirus in many countries of the world, and by February 28, 2020, stock markets around the world experienced their largest decline in one week since the financial crisis that occurred in 2008.

The market collapsed Stocks globally in March 2020 with a decline in ratios in many major global indices. With the spread of the virus, all global conferences and events in the field of technology, fashion and sports will be canceled or postponed until further notice, although the monetary impact on travel and industry has not yet been estimated or known approximate value, but it is likely to be in the billions and continues to increase.

Symptoms of COVID-19 in humans can range from very mild to severe depending on the person's exposure to the virus. Some people may have only a few symptoms, while others have no symptoms at all. Some people may experience worsening symptoms that can lead to death. Therefore, the researcher tried to shed light on the reasons that lead to the major cases of infections and the minor ones as well. The data was collected by questioning people via Google Forms to measure the impact of the virus on them and the major influencing factors that lead to the infection. The data represent a sample group of 260 infected persons. The sample was taken from people in the city of Al-Diwaniyah within four months by using a form

In this study the dependent variable is binary; it either takes 'zero' in case of major infection or death, or 'one' in case of moderate or minor infection. The independent variables are 14 variables which represent the factors influencing the infection of Coronavirus

X1: Represents gender, male = 1, female = 2

X2: Represents age

X3: represents the weight

X4: represents pressure, None = 1, Found = 2, Decrease = 3, Medium = 4, Height = 5

X5: represents diabetes, None=1, Found=2, Descending=3, Ascending=4

X6: Represents lung problems, None = 1, found = 2

X7: Represents a weak immune system, None = 1, There is = 2

X8: Represents vitamin D deficiency, None = 1, Fond = 2

X9: represents the workplace, Housewife or not working = 1, Employee = 2, Wage earner = 3, Students = 4, Hospital and medical clinics = 5

Chapter five :

X10: Represents previous surgical operations, No operations = 1, Previous operations performed = 2,

X11: represents smoking, Non-smoker = 1, Smoker = 2

X12: Represents the psychological state, Not good = 1, Medium = 2, Good = 3

X13: represents nutrition, Not good = 1, Medium = 2, Good = 3

X14: living status ,Poor = 1, Medium = 2, Good or Rich = 3

The data was programmed using the R program, and the proposed methods were compared with three existing methods, as shown in the following table:

Table 5-1. The parameter estimates of BSLLR,BSLR, BLR, BPR and BBQR methods for the real data .

Methods	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}	β_{11}	β_{12}	β_{13}	β_{14}
BSLLR	0.1127	0.2714	0.1260	0.2598	0.4356	0.1271	0.0049	0.2092	0.1615	-0.2173	0.5063	0.2959	0.1197	0.2404
BSLR	0.0683	0.1984	0.3245	0.2834	0.3377	-0.0005	0.2152	0.2719	-0.1291	0.5253	0.0001	0.1801	0.0033	0.3335
BLR	0.3199	0.1265	-0.2063	0.0538	-0.1588	-0.0743	-0.2767	0.0104	0.1384	-0.2558	-0.0699	-0.2328	0.0906	0.4356
BPR	0.1850	0.0774	-0.1183	0.0265	-0.1020	-0.0402	-0.1581	0.0153	-0.0782	-0.1442	-0.0565	-0.1432	0.0670	0.2421
BBQR	0.4298	0.1644	-0.2407	0.0561	-0.2022	-0.0322	-0.3285	0.0576	-0.1950	-0.2879	-0.1229	-0.3708	0.0261	0.6868

In Table (5-1) Shows the results for the Bayesian semi parametric Lasso logistic regression (BSLLR), Bayesian semi parametric logistic regression (BSLR), Bayesian logistic regression (BLR), Bayesian probit regression (BPR), Bayesian Binary Quartile regression (BBQR).

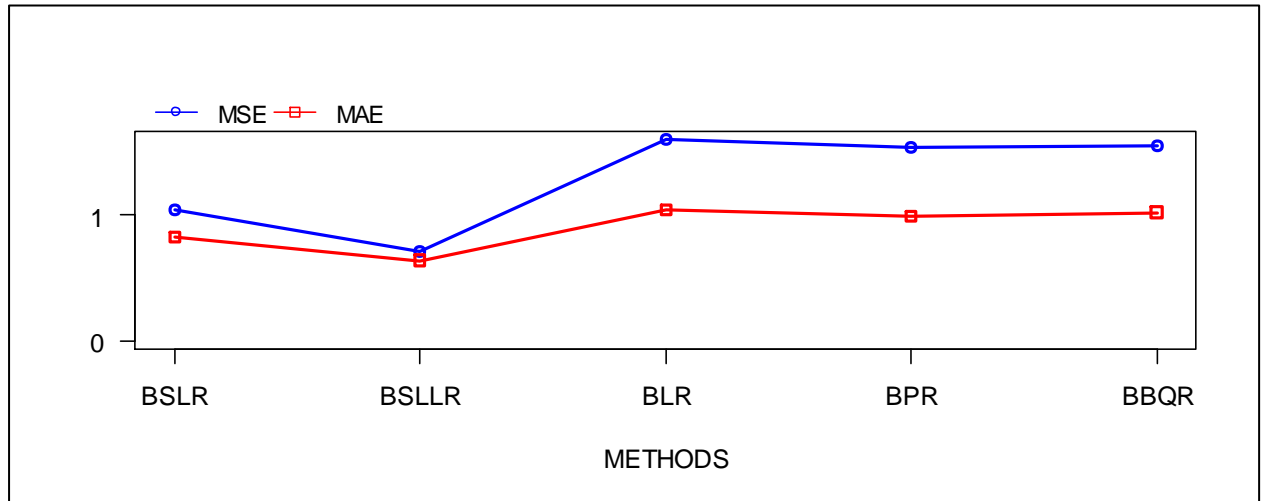
The proposed model (BSLLR) gave non-zero coefficients to (gender, age, weight, pressure, diabetes, lung problems, vitamin D, workplace, previous surgical operations, smoking, the psychological state, nutrition and living status).

BSLLR, BLR, BPR and BBQR model choose the same predictable variable, except (the weak immune system), variable was selection to be as irrelevant predictor variable on the response variable (infection status). So, we can say that the proposed model (BSLLR) works as variable selection procedure and that is cope with the nature Lasso method, but its dense vectors of parameter estimates. As well as, the results of the second proposed model (BSLR)are comparable with the other models and works well.

Table 5-2. The values of MSE and MAE of BSLLR, BSLR, BLR, BPR and BBQR methods for the real data

Methods	MSE	MAE
BSLR	1.0401	0.8251
BSLLR	0.7055	0.6360
BLR	1.5966	1.0421
BPR	1.5342	0.9866
BBQR	1.5454	1.0183

In Table (5-2) the MSE and MAE values are summarized. It can be seen that the proposed methods get the smallest values of MSE and MAE compared to other methods. The largest values of MSE and MAE are for BBQR method BLR method gets MSE and MAE values larger than BPR.



Figure(5-1). show the MSE and MAE for BSLR,BSLLR, BLR, BPR and BBQR methods for real data

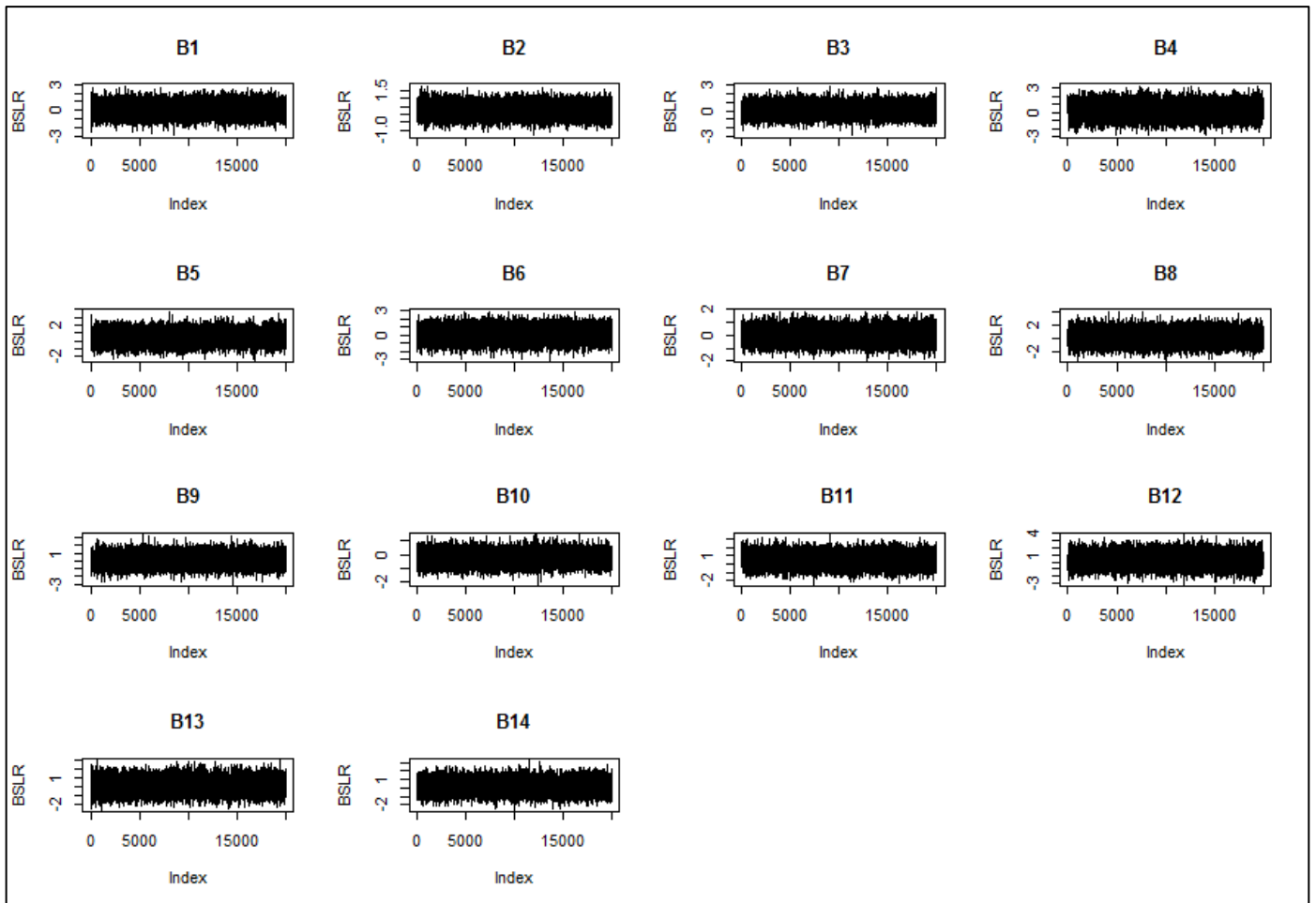


Figure (5-2). Trace plots for BSLR in the real data

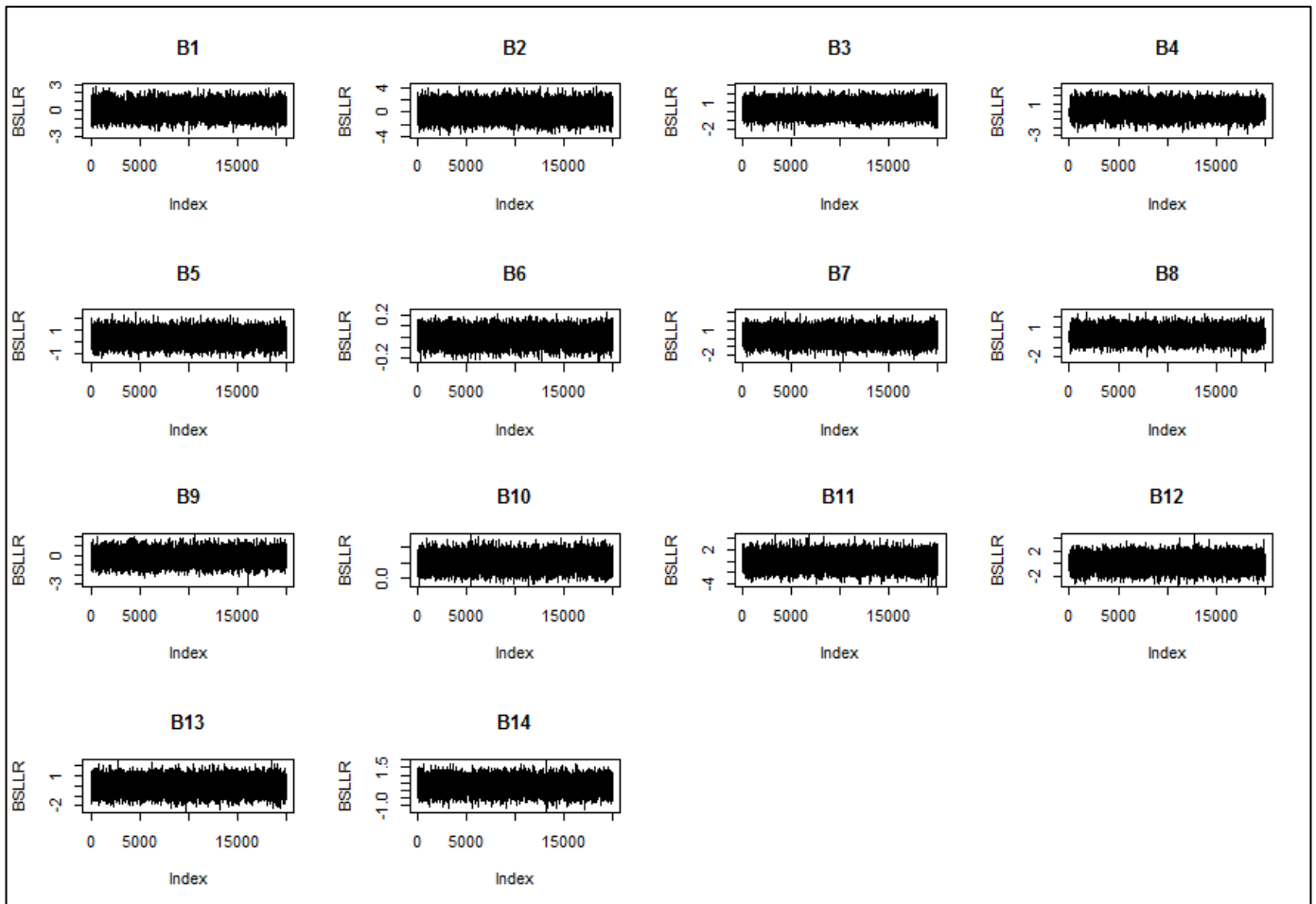


Figure (5-3). Trace plots for BSLLR in the real data.



CHAPTER SIX

- Conclusions
 - Recommendations
- 

6-1. Conclusions

In this thesis, the Bayesian estimation approach is introduced to estimate the unknown link function and the coefficient vector in the semiparametric logistic regression. The normal distribution prior is considered to the coefficient vector and Gaussian process prior is set for the unknown link function, Bayesian estimation and variable selection approach are suggested to estimate the parameters and link function and select the important variables for single index logistic regression model. Laplace distribution is set as prior to the coefficients vector and prior to the unknown link function (Gaussian process) we have developed a Bayesian hierarchical model for the single index logistic regression model and lasso semiparametric logistic regression model. This is done by using MCMC algorithm which is adopted for posterior inference.

Three simulation examples are used to compare our proposing methods, BSLR and BSLLR, with the other three methods, BLR, BPR and BBQR. We derived our conclusions from the simulation examples and the practical side that these methods have presented better results than their predecessors.

- 1- Throughout the simulation, we find that BSLLR method is better than the rest of the methods because it obtained the lowest value for SD and Bias.
- 2 - Throughout the simulation we get that the method BSLR is better than the rest of the methods, but it is not better than the method BSLLR.
- 3- Throughout the applied example, the results are similar to those in the simulation and

characterize by BSLLR method better than the rest of the methods are also obtained the lowest values of MSE followed by BSLR and then BPR.

4- Finally, we have concluded that the performance of our suggested methods are better than the other existed methods.

6-2. Recommendations

According to what have been stated in this study, the researcher comes up with these recommendations:

1- use the semi-parametric model because it is more flexible and gets the better results.

2- use of the Bayesian in estimation. We also recommend using the Bayesian semi-parametric logistic regression (BSLR) because it is of great importance in the estimation process and also gives good results.

3- use Bayesian Lasso for semi-parametric logistic regression because it works better than other methods and has less bias, standard deviation, and less mean error, and because it is of great importance in estimating and selecting variables.

References

- Acquah, H. (2013). Bayesian Logistic Regression Modelling via Markov Chain Monte Carlo Algorithm. In *Journal of Social and Development Sciences*. Vol. 4, No. 4, pp. 193-197.
- Albert, J. H., & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, No. 88(422), pp. 669-679.
- Al-Farhoud, S. (2014). The Use of Logistic Regression in studying the Factors Influencing the Performance of Stocks: An Empirical Study on the Kuwait Stock Exchange. In *Journal of Al Azhar University-Gaza (Natural Sciences)*, vol.16, pp.47-68
- Alhamzawi, R. and Ali, H. (2018): Bayesian single-index quantile regression for ordinal data, Communications. In *Statistics – Simulation and Computation* DOI: 10.1080/03610918.2018.1494283.
- Alkenani, A., & Yu, K. (2013). Penalized single-index quantile regression. *International Journal of Statistics and Probability*, 2(3), 12.
- Akkus, O. (2011). XploRe Package for the Popular Parametric and the Semiparametric Single Index Models. *Gazi University Journal of Science*, no. 24(4), pp. 753-762.
- Antoniadis, A., Grégoire, G., & McKeague, I. W. (2004). Bayesian estimation in single-index models. *Statistica Sinica*, 1147-1164.
- Athman , Ibraheem. (2018). A practical study on logistic regression using R. Sabah University.

References :

- Alshaybawee, T. (2006). Maximizing the Efficiency of the Analysis for the Logistic Curve by using Power Transformation, (Unpublished MA.Thesis), AL- Mustansiriya University.
- Berkson, J. (1957). Tables for the maximum likelihood estimate of the logistic function. *Biometrics*, 13(1), 28-34.
- Bellman, R., Kalaba, R., & Sridhar, R. (1966). Adaptive control via quasi-linearization and differential approximation. *Computing*, 1(1), 8-17.
- Bernardo, J.M. and Smith, A.F.M. (2000). *Bayesian Theory*. John Wiley & Sons, Chichester.
- Carroll, R. and Wand, M. (1991). Semiparametric Estimation in Logistic Measurement Error Models. In *J. R. Statist. Soc.* vol. 53, No.3, pp. 573-585.
- Chen, C., Zhang, G., Liu, X. C., Ci, Y., Huang, H., Ma, J., ... & Guan, H. (2016). Driver injury severity outcome analysis in rural interstate highway crashes: a two-level Bayesian logistic regression interpretation. *Accident Analysis & Prevention*, 97, 69-78.
- Chiaka, E. S., Adam, M. B., Krishnarajah, I., Shohaimi, S., & Guure, C. B. Bayesian logistic regression model on risk factors of type 2 diabetesmellitus. URL:https://www.researchgate.net/publication/272172873_Bayesian_Logistic_Regression_Model_on_Risk_Factors_of_Type_2_Diabetes_Mellitus.
- Christensen, R. (2006). *Log-linear models and logistic regression*. Springer Science & Business Media
- Choi, T., Q. Shi, J. & Wang, B. (2011). A Gaussian process regression approach to a single-index model. In *Journal of Nonparametric Statistics*. Vol, 23, no 1, pp. 21-36.

References :

- ÇOKLUK, Ö. (2010). Logistic Regression: Concept and Application. In *Eğitim Danışmanlığı ve Araştırmaları İletişim Hizmetleri Tic. Ltd. Şti.*
- Czepiel, S. A. (2002). Maximum likelihood estimation of logistic regression models: theory and implementation. *Available at czep.net/stat/mlelr.pdf*, 83.
- De Blasi, P. and Hjort, (2007). Bayesian Survival Analysis in Proportional Hazard Models with Logistic Relative Risk. In *Board of the Foundation of the Scandinavian Journal of Statistics*, Vol 34, pp. 229–257
- Delgado, M. A., & Robinson, P. M. (1992). Nonparametric and semiparametric methods for economic research. *Journal of Economic Surveys*, 6(3), 201-249.
- Dhara, K. (2018). *Shape Constrained Single Index Models for Biomedical Studies* (Doctoral dissertation, The Florida State University).
- Dhara, K., Lipsitz, S., Pati, D. & Sinha, D. (2020). A New Bayesian Single Index Model with or without Covariates Missing at Random. In *International Society for Bayesian Analysis*. Vol. 15, No 3, pp. 759–780.
- Dominicil, F. and Parmigianil, G. (2001). Bayesian Semiparametric Analysis of Developmental Toxicology Data. In *International Biometric Society, Wiley*. Vol. 57, No. 1, pp. 150-157.
- Dunson, D. (2003). Bayesian Isotonic Regression for Discrete Outcomes. In <https://www.researchgate.net/publication/2477708>
- Falsal,R (2020) .using some of the methods for estimating non parametric and semi parametric regression function with an application.
- Fisher, T. J. (2006). *Simulation Study for Single-Index Models* (Doctoral dissertation, Clemson University).

References :

- Fridley, B. L. (2009). Bayesian variable and model selection methods for genetic association studies. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, 33(1), 27-37.
- Gramacy R. and Lian, H. (2012). Gaussian Process Single-Index Models as Emulators for Computer Experiments, in *Technometrics*. Vol. 54, no.1, pp.30-41.
- Greenberg, E. (2008). *Introduction to Bayesian Econometrics*. Cambridge University Press, New York Haggag,
- Haggag, M.M. (2007). Estimation of Parametric and Semiparametric Logistic Regression Models Using Credit Scoring Data. In [http: \(PDF\) Estimation of Parametric and Semiparametric Logistic Regression Models Using Credit Scoring Data \(researchgate.net\)](http://researchgate.net)
- Hair, J. F., Black, W. C., Babin, B., Anderson, R. E., & Tatham, R. L. (2006). *Multivariate data analysis* (6th ed). Upper Saddle River, NJ: Prentice-Hall.
- Härdle, W., Müller, M., Sperlich, S. & Werwatz, A. (2004). *Nonparametric and Semiparametric Models*. New York: Springer.
- Hardle, W., Hall, P., & Ichimura, H. (1993). Optimal smoothing in single-index models. *The annals of Statistics*, 157-178.
- Holmes, C. and Heldy, L. (2006). Bayesian Auxiliary Variable Models for Binary and Multinomial Regression. In *International Society for Bayesian Analysis*. Vol. 1, Number 1, pp. 145- 168
- Horowitz, J. and Lee, S. (2002). Semiparametric methods in applied econometrics: do the models fit the data? In *Statistical Modelling*. Vol. 2, pp. 3–22

References :

- Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression*. John Wiley & Sons.
- Hsu, J. and Leonard, T. (1997). Hierarchical Bayesian semiparametric procedures for logistic regression. In *Biometrika* , vol. 84,1, pp. 85-93.
- Hu, Y., Gramacy, R. B., & Lian, H. (2013). Bayesian quantile regression for single-index models. *Statistics and Computing*, 23(4), 437-454
- Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *Journal of econometrics*, 58(1-2), 71-120.
- Kayri, M. ve Okut, H. (2008). Özel yetenek sınavındaki başarıya ilişkin risk analizinin karışımli lojistik regresyon modeli ile incelenmesi. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, 35, 227-239.
- Kerlinger, F. & Pedhazur, E. (1973). "Multiple Regression Behavioral Research", New York: Holt, Rinehart and Winston, Inc.
- Kong, E., & Xia, Y. (2008). Estimation of single-index quantile regression Model. *arXiv preprint arXiv:0803.2474*
- Kopytov, E., & Santalova, D. (2007). Application of the single index model for forecasting of the inland conveyances. In *Recent Advances In Stochastic Modeling And Data Analysis* (pp. 268-276).
- Kim, Y. and Park, J. (2019). Incorporating prior knowledge with simulation data to estimate PSF multipliers using Bayesian logistic regression. In *Reliability Engineering and System Safety*. Vol. 189, pp. 210–217
- Kuruwita, C. N. (2016). Non-iterative Estimation and Variable Selection in the Single-index Quantile Regression Model. *Communications in Statistics-Simulation and Computation*, 45(10), 3615-3628.

References :

- Lam, H and Xue, H. (2005). A semiparametric regression cure model with current status data. In *Biometrika* . Vol. 92, no. 3, pp. 573–58
- Lea, S. (1997). " Multivariate Analysis II: Manifest variables analysis. Topic 4: Logistic Regression and Discriminant Analysis ", University of EXETER, Department of Psychology. Available at: www.exeter.ac.uk/~SEGLea/multivar2/diclogi.html.
- Lee, S. (2004). "Application of likelihood ratio and logistic regression models to landslide susceptibility mapping using GIS", *Environmental Management*, Vol. 34, No. 2, 223-232.
- Lee, P.M. (2004). *Bayesian Statistics, An Introduction*. 3rd Edition. Hodder Arnold, London.
- Ly, Y., Zhang, R., Zhao, W., & Liu, J. (2014). Quantile regression and variable selection for the single-index model. *Journal of Applied Statistics*, 41(7), 1565-1577.
- Maneejuk, P., Yamaka, W., & Nachaingmai, D. (2019, January). Bayesian analysis of the logistic kink regression model using metropolis-hastings sampling. In *International Econometric Conference of Vietnam* (pp. 1073-1083). Springer, Cham
- Michelot, T., Langrock, R., Kneib, T., & King, R. (2016). Maximum penalized likelihood estimation in semiparametric mark-recapture-recovery models. *Biometrical Journal*, 58(1), 222-239.
- Mallick, H., & Yi, N. (2014). A new Bayesian lasso. *Statistics and its interface*, 7(4), 571-582.
- Millimet, D., List, J. and Stengos, T. (2003). The Environmental Kuznets Curve: Real Progress or Mis-specified Models? In *The Review of Economics and Statistics*, Vol. 85, No. 4, pp. 1038-1047.

References :

- Newton, M., Czado, C. & Chappell, R. (1996). Bayesian Inference for Semiparametric Binary Regression. In *Journal of the American Statistical Association*. Vol. 91, no.433. pp.142-153.
- Ntzoufras, I. (2009). *Bayesian Modeling Using WinBUGS*. John Wiley & Sons, Inc., Hoboken, New Jersey.
- Park, T., & Casella, G. (2008). The Bayesian Lasso. *Journal of the American Statistical Association*, no.103(482), pp.681-686.
- Pohar, M., Blas, M. & Turk, S. (2004). Comparison of Logistic Regression and Linear Discriminant Analysis: A Simulation Study. In *Metodološki zvezki*. Vol. 1, No. 1, pp.143-161
- Poston, D.L (2004). "Sociological Research: Quantitative Methods (Lecture notes, Lecture 7)", Spring.
- Powell, J. L. (1994). Estimation of semiparametric models. *Handbook of econometrics*, no. 4, pp.2443-2521.
- Richardson, S., Leblond, L. and Jaussent, I. and Green, P. (2002). Mixture models in measurement error problems, with reference to epidemiological studies. In *J. R. Statist. Soc.* vol.165, Part 3, pp. 549-566
- Press, S.J. (1989). *Bayesian Statistics: Principles, Models, and Applications*. John Wiley & Sons, Inc., Hoboken, New Jersey
- Saeed, R. (2015) Using the Logistic Regression Model in Studying the Assistant Factors to Diagnose Bladder Cancer. In *Journal of Economic and Administrative Sciences*, vol 21, no.83.
- Salah, T. (2016) . Some of the semi-parametric methods to estimate and variable selection for single index m (Unpublished Ph. D Thesis), Baghdad University
- Shin, J. L. (2015). *A Fully Bayesian Approach to Logistic Regression*. University of California: San Diego.

References :

- Stephenson, B., Cook, D., Dixon, P., Duck worth. W., Kaiser, M., Koehler, K., & Meeker, W. (2008). Binary response and logistic regression analysis. Available at:
<http://www.stat.wisc.edu/mchung/teaching/MIA/reading/GLM.logistic.Rpackage.pdf>
- Stute, W., & Zhu, L. X. (2005). Nonparametric checks for single-index models. *The Annals of Statistics*, no.33(3), pp.1048-1083.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267-288
- Titma, M. , Brandon, N.& Roomsa, K. (2003). Education as a Factor in Intergenerational Mobility in Soviet Society. *In European Sociological Review*. Vol.19, No.3 , pp.281- 297.
- Vaart, A and Zanten, J. (2008). Rates of Contraction of Posterior Distributions Based on Gaussian Process. *In Institute of Mathematical Statistics in The Annals of Statistics*. Vol. 36, No. 3, pp. 1435–1463.
- Wang, H. B. (2009). Bayesian estimation and variable selection for single index models. *Computational Statistics & Data Analysis*, 53(7), 2617-2627.
- Wang C., Wang S., Zhao, L. & Tyan Ou, S. (2014). Weighted Semiparametric Estimation in Regression Analysis with Missing Covariate Data. *In Journal of the American Statistical Association*, Vol. 92, No. 438. pp.
- Webster, G. (2011). *Bayesian logistic regression models for credit scoring* (Doctoral dissertation, Rhodes University).
- Wilhelmsen, M., Dimakos, X.K., Husebø, T., and Fiskaaen, M. (2009). *Bayesian Modelling of Credit Risk using Integrated Nested Laplace*

References :

Approximations. Available at:

<http://publications.nr.no/BayesianCreditRiskUsingINLA.pdf> Access date: 21st November 2021.

Yatchew, A. (2003). *Semiparametric regression for the applied econometrician*. Cambridge University Press

Yu, Y., & Ruppert, D. (2002). Penalized spline estimation for partially linear single-index models. *Journal of the American Statistical Association*, 97(460), 1042-1054.

Zhao, K., & Lian, H. (2015). Bayesian Tobit quantile regression with single-index models. *Journal of Statistical Computation and Simulation*, 85(6), 1247-1263.

Zhu, L. P., & Zhu, L. X. (2009). Nonconcave penalized inverse regression in single-index models with high dimensional predictors. *Journal of Multivariate Analysis*, no. 100(5), pp.862-875.

Ziemba, A. (2005). *Bayesian Updating of Generic Scoring Models*.

Available at:

http://www.crc.man.ed.ac.uk/conference/archive/2005/papers/ziemba_arkadius.pdf. Access date: 25th September 2021.

الخلاصة

يعد تحليل النموذج شبه المعلمي أحد أكثر الموضوعات إثارة للاهتمام في الدراسات الحديثة نظرًا للطريقة الدقيقة التي يصف بها البيانات الإحصائية التي توفر معلمات فعالة. في بعض الدراسات ، يأخذ متغير الاستجابة قيمتين ، إما صفر - لعدم وجود استجابة - أو واحدة للاستجابة.

لذلك ، يستخدم الانحدار اللوجستي لنمذجة هذه البيانات. بناءً على طريقة Bayesian ، تم اقتراح طريقتين جديدتين للتقدير في هذه الأطروحة.

الطريقة الأولى هي طريقة التقدير البيزية التي تستخدم لتقدير الدالة غير المعروفة ومتجه المعامل في الانحدار اللوجستي شبه المعلمي (BSLR). الطريقة الثانية هي طريقة Bayesian lasso التي تم اقتراحها لتقدير واختيار المتغيرات المهمة لنموذج الانحدار اللوجستي ذي المؤشر الفردي (BSLLR). في طريقة BSLR ، يتم تعيين التوزيع الطبيعي كتوزيع مسبق لمتجه المعامل بينما يعتبر توزيع لابلاس سابقًا في طريقة BSLLR. تم تعيين عملية Gaussian على أنها سابقة للدالة غير المعلمية غير المعروفة. تم اعتماد خوارزمية MCMC للاستدلال اللاحق.

تمت مقارنة طرق التقدير المختلفة من خلال مقارنة الاستخدام بين طرق التقدير باستخدام متوسط مربعات الخطأ، ومتوسط الخطأ المطلق، والتحيز، والانحراف المعياري (SD). باستخدام ثلاثة أمثلة محاكاة وبأحجام عينات مختلفة (العدد = 50، 150، 250).

لاختبار كفاءة الطرق المقترحة (BSLR، BSLLR) ، يتم استخدام البيانات الحقيقية من خلال اعتماد مجموعة من المؤشرات لغرض مقارنة الطرق المقترحة مع مجموعة من الأساليب الموجودة مسبقًا. لتطبيق طرق التقدير ، تم أخذ عينة عشوائية بسيطة قوامها (260) لدراسة العوامل المؤثرة في الإصابة بفيروس كورونا (متغير الاستجابة). بينما المتغيرات التفسيرية هي (الجنس ، العمر ، الوزن ، الضغط ، السكري ، مشاكل الرئة ، ضعف جهاز المناعة ، نقص فيتامين د ، مكان العمل ، العمليات الجراحية السابقة ، التدخين ، الحالة النفسية ، التغذية ، الحالة المعيشية). أوضحت الدراسة أن أداء طرق بايزي يوفر تحسينات جوهرية مقارنة بالطرق الأخرى.



جمهورية العراق

وزارة التعليم العالي والبحث العلمي

جامعة القادسية

كلية الادارة والاقتصاد

قسم الاحصاء

التقدير البيزي للانحدار اللوجستي شبه المعلمي مع التطبيق

رسالة مقدمة الى مجلس كلية الادارة والاقتصاد/جامعة القادسية وهي جزء من نيل
متطلبات الماجستير في علوم الاحصاء

من قبل الطالبة

زينب سامي تركي

إشراف

الدكتور

طه حسين الشيباوي