REPUBLIC OF TURKEY
YILDIZ TECHNICAL UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCEINCE

BAYESIAN ESTIMATION FOR MULTINOMIAL LOGISTIC REGRESSION

KHADAR MOHAMED GAHAYR

MSc. THESIS
DEPARTMENT OF STATISTICS
PROGRAM OF STATISTICS

ADVISOR
ASSOC. PROF.DR. ATIF AHMET EVREN

ISTANBUL, 2015

REPUBLIC OF TURKEY

YILDIZ TECHNICAL UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCEINCE

BAYESIAN ESTIMATION FOR MULTINOMIAL LOGISTIC REGRESSION

A thesis submitted by Khadar Mohamed GAHAYR in partial fulfillment of the requirements for the degree of **MASTER OF SCIENCE** is approved by the committee on 25.06.2015 in the department of statistics.

Thesis Adviser

Assoc. Prof.Dr. Atif Ahmet EVREN

Yıldız Technical University

Approved by the Examining Committee

Assoc. Prof.Dr. Atif Ahmet EVREN

Yıldız Technical University _____

Prof. Dr. Ali Hakan BÜYÜKLÜ

Yıldız Technical University _____

Prof. Dr. Karun NEMLIOĞLU

Istanbul University _____

ACKNOWLEDGEMENT

I am deeply grateful to all those who have contributed this research paper to be completed. First, I would like to express my deepest gratitude to may adviser, my Instructor and my collaborator Assoc. Prof. Dr.Atif Ahmed Evren. His continuous guidance, Support and Care make this research paper to be successfully completed. I would also like to thank my friends and my fellow students for their useful suggestion, warm encouragement and shared ideas from their own experience. Their comments have inspired me to do my best. Finally, I am forever grateful to my family for their love support and understanding.

June, 2015

Khadar Mohamed GAHAYR

TABLE OF CONTENTS

Page

LIST OF SYMBOLS

N	the number of trials
P	Probability of success.
$P_{\varepsilon}(\varepsilon)$	Non normal density
μ	mean
ℓ_0	Maximized value
λ_i	Effect of classification
$\theta^{(t)}$	Current state value
$\hat{\theta}$	Proposed value

LIST OF ABBREVIATIONS

BUGS	Bayesian Inference using Gibbs Sampling
CL	Confidence Interval
DARS	Data and research Solution
IID	independent and identical distribution
MC	Markov chain
MCMC	Markov chain Monte Carlo
ML	Maximum likelihood
MLE	Maximum likelihood estimation
MNL	Multinomial logit
PDF	Probability density function
SE	Standard Error

LIST OF FIGURES

Page

LIST OF TABLES

page

ABSTRACT

BAYESIAN ESTIMATION FOR MULTINOMIAL LOGISTIC REGRESSION

Khadar Mohamed GAHAYR

Department Of Statistics

Msc. Thesis

Adviser: Assoc. Prof.Dr. Atif Ahmet EVREN

This thesis discusses the applicability of Bayesian multinomial logistic regression model on the prediction of happiness of Somaliland youth by two variables; age and educational level. It presents two approximate methods on multinomial logistic regression estimation; classical method and Bayesian method or Markov Chain Monte Carlo (MCMC), to obtain the marginal posterior density for the parameters. A comparison of these two methods is carried out to determine the usefulness of Bayesian method on multinomial logistic regression estimation. R and WinBUGS (Bayesian Inference using Gibbs Sampling) programs have been used to fit the model. As both of these two methods have suggested; happiness increases with educational level and decreases with age. In addition, it is also shown that Bayesian Multinomial logistic regression is useful in direct computations and it produces very accurate approximations to the posterior density.

Key words: Bayesian inference, Binary and Multinomial data, Gibbs sampling, Markov chain Monte Carlo, Multinomial logit model

ÖZET

BAYESÇİ MULTINOMİYAL LOJISTIK REGRESYON

Khadar Mohamed GAHAYR

İstatistik Anabilim Dali

Yüksek Lisans Tezi

Tez Danışmanı: Doç. Dr. Atif Ahmet EVREN

Bu tez, Somali'de belirli bir yaştaki gençlerin mutluluğunun, yaş ve eğitim düzeyi gibi iki açıklayıcı değişken ile incelenmesinde, Bayesçi multinomiyal lojistik regresyon modelinin uygulanabilirliğini savunmaktadır. Multinominal lojistik regresyon yöntemi, parametrelerin posterior (sonsal) yoğunluğunun kestirilmesinde, klasik yöntem ve Bayesçi (veya Markov Zinciri Monte Carlo (MCMC) yöntemi) olmak üzere iki farklı yaklaşım sunar. Bu iki yöntemin karşılaştırılması, multinominal lojistik regresyon yöntemine Bayesçi yaklaşımın katkısını belirlemek amacıyla yapılmıştır. Modeli eldeki verilere uydurmak için R ve WinBugs (Bayesçi yaklaşımı kullanarak Gibbs örnekleme) programları kullanılmıştır. Her iki yöntemde mutluluk seviyesinine ğitim seviyesi ile doğru orantılı, yaşile ise ters orantılı olduğunu göstermiştir. Klasik yönteme ek olarak bu çalışmada, Bayesçi Multinominal lojistik regresyonun, doğrudan hesaplamalarda ve posterior yoğunluğun yüksek birkesinlik derecesinde belirlenmesinde yararlı olduğu gözlenmiştir.

Anahtar Kelimeler: Bayesçi çıkarım, İkili ve multinomiyal veri, Gibbs örneklemesi, Markov zinciri Monte Carlo, Multinomiyal logit modeli

CHAPTER 1

INTRODUCTION

1.1 Literature review

The application of statistical models for categorical data has substantially developed for the last decades (Agresti 2010) [1]. The most popular models for categorical response data are binary logistic regression and multinomial logistic regression. Binary logistic regression predicts the value of one dependent variable from one or more independent variables, when the dependent variable is dichotomous. On the other hand, multinomial logistic regression is a simple extension of binary logistic regression which allows that more than two categories of dependent variable can be used in the model. Multinomial logistic regression uses maximum likelihood methodology to evaluate the probability of categorical values and to predict parameter coefficients. [2].

Besides, Bayesian methods for estimating multinomial logistic regression models have long time back in history. The starting place of such efforts can be extended to the works of Zellner and Rossi both of whom studied the generalized linear models in detail. They derived approximate posterior densities for an improper uniform prior. Refer to Agresti and Hitchcock (2005) for review [3]. Considerable researches have been performed concerning this issue. One can refer to Fruhwirth (2012) [4], Held and Holmes (2006) [5] and Hartzel et al (2006) [6]. Historical perspective can also be reviewed by the work of Agresti and Hitchcock (2005) [3].

Furthermore, Albert and Chip (1993) [7] developed the exact Bayesian methods for modeling categorical response data using the idea of data augmentation, particularly how data augmentation approach provides a general framework for analysis binary regression models. Moala et al. (2009) [8] considered the Bayesian inference for the estimation of logistic regression

model by comparing two approximate methods of Laplace's method and Markov chain Monte Carlo method. The results were also compared with maximum likelihood estimation. Migon and Souza (2004) [9] discussed the Bayesian binary regression model by using Markov Chain Monte Carlo. A model building strategy based on Bayes factor is proposed and aspects of model validation are extensively discussed in it.

Tsai (2004) [10] proposed the ability to fit a binary logistic regression model with the observed data using the Bayesian inference. Brien and Dunson (2004) [11] also proposed a new type of multivariate logistic distribution that can be used to construct likelihood for multivariate logistic regression analysis of binary and categorical data. They follow a Bayesian approach to estimation and inference for developing an efficient data augmentation algorithm for posterior computation. Frühwirth and Frühwirth- Schnatter (2006) [12] showed how the auxiliary mixture sampler is implemented for binary or multinomial logit models. They also demonstrated how to extend the sampler to mixed effect models and time-varying parameters models for binary and categorical data.

Likewise, Gelman (2006) [13] suggested various noninformtive prior distribution for scale parameters in hierarchical models. He used an example to issustrate serious problems with the inverse-gammma family of noninformative prior distributions. Arcos et al. (2014) [14] extended dual frame estimation techniques to the case of estimation of proportions when the variable of interest has multinomial outcomes. They described the joint distribution of the class indicators by multinomial logistic regression model.

1.2 Objectives of the thesis

The main objective of this thesis is to demonstrate the effectiveness of Bayesian methods for estimating the multinomial logistic regression model in the prediction of Happiness of Somaliland youth with age and educational level. It also aspires to find out the effect of age and educational level to the happiness of Somaliland youth. This leads to realize why a lot of of Somaliland youth people embark every months on a perilous journey cross the Mediterranean sea on their way to Europe while others lives hopelessly in the country. If there is relationship between them, this means higher motivation towards further educational attainment and the completion of further years of schooling in Somaliland.

1.3 Hypothesis

This thesis formulates the hypothesis that Bayesian estimation for multinomial logistic regression is useful and performs well to predict the happiness of Somaliland youth with age and educational level.

CHAPTER 2

CATEGORICAL RESPONSE DATA

2.1 Types of Data

In statistical computations and analyses, it is assumed that the variables have specific levels of measurement. According to the measurement scale there are essentially two types of data, Categorical and numerical data. Numerical data can be measured on a ratio, or interval scale. A variable measured on interval scale gives information about that one value is greater or larger or better than the other as ordinal scale of categorical type of measurement do, however, interval data has an equal distance between each value. Hence the difference between two values at one point on the scale is the same as the difference between two equidistant values at another point on the scale. Temperature using Celsius or Fahrenheit is a good example of interval data. The difference between 15 and 20 degrees is the same as the difference between 5 and 10 degrees.

The ratio scale has the same properties that an interval scale has, except that a ratio scaling, there is an absolute zero point. In addition, the number which is twice as large as another reflects the fact that the attribute being measured is twice as great. Temperature measured in Kelvin is an example, there is no value possible below 0 degrees Kelvin, and it is absolute zero. Ratio scales are rare in the social sciences unless one is using a measure of some physical feature such as height or time.

A categorical variable has a measurement scale consisting of a set of categories. Categorical variables can be divided in to two; nominal and ordinal. A variable measured on a nominal scale is a variable that does not have any evaluative distinction. One value of nominal variable is not

greater than another. A good example of a nominal variable is sex (gender) with categories of male and female, hence one category is not greater or better than another, therefore they can be ordered in any way. A nominal can take discrete or countable number of values, however, the number is simply functioning as a label and the size of the number does not reflect the magnitude. With nominal variables, there is a qualitative difference between values, not a quantitative.

An ordinal variable is similar to a nominal variable. However, variable measured on an ordinal scale is a variable that has any evaluative distinction. One value of ordinal variable is greater or larger or better than the other. Economic status with categories of low, medium and high is good example of ordinal type of variable. These three categories can be ordered as low, medium and high. Ordinal variables are sometimes called discrete variables because they only take a discrete or countable number of values which indicate the size and the magnitude of value over the other.

2.2 Probability Distribution of Categorical data

A probability distribution is a statistical model that shows the possible outcomes of a particular event or course of action as well as the statistical likelihood of each event. Probability distributions can be used to create scenario analyses. A scenario analysis uses probability distributions to create several, theoretically distinct possibilities for the outcome of a particular course of action or future event. There are three important probability distributions in Categorical response data. These are binomial, multinomial, and Poisson distributions [15].

2.2.1 Binomial Distribution

Binomial distribution is a discrete probability distribution with two parameters, traditionally indicated by *n*, the number of trials, and *P* which is the probability of success. A random variable follows a binomial distribution when each trial has exactly two possible outcomes like “yes” or “no” and “success” or “failure”. A well-known example of such an experiment is the repeated tossing of a coin and counting the number of times "heads" comes up.

Binomial distribution has two probabilistic assumptions which are; independence, the pairs of the outcomes are sorted independently, and homogeneity, the outcome probabilities are the same across individuals. With X being the number of A outcomes, these assumptions lead to X having

a Binomial distribution over the possible x values 0, 1, 2, with p = P(A). We would write this by saying the distribution of X is B (n; p), The probability mass function for the possible outcomes x is:

$$P(x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}, x = 0,1,2,....,n \quad (2.1)$$

$$E(x) = \mu = np \quad (2.2)$$

$$\sigma = \sqrt{np(1-p)} \quad (2.3)$$

The skewnessis described $\frac{E(x-p)^3}{\sigma^3} = \frac{(1-2p)}{np(1-p)}$ (2.4)

The binomial distribution approximates to normality as n increases with fixed P.

The binomial distribution allows one to compute the probability of obtaining a given number of binary outcomes. For example, it can be used to compute the probability of getting 6 heads out of 10 coin flips. The flip of a coin is a binary outcome because it has only two possible outcomes: heads and tails. The multinomial distribution can be used to compute the probabilities in situations in which there are more than two possible outcomes.

2.2.2 Multinomial Distribution

Multinomial distribution is a distribution with a statistical experiment that has the following properties: the experiment consists of n repeated trials, each trials has a discrete number of possible outcomes, on any given trial, the probability that a particular outcome will occur is constant. In addition, the trials are independent; that is, the outcome on trial does not affect the outcome on the other trial. The following formula gives the probability of obtaining a specific set of outcomes when there are more than two possible outcomes for each event.

$$P(n_1, n_2....n_c) = (\frac{n!}{n_1!, n_2!...n_c!}) p_1^{n_1} p_2^{n_2} ....p_c^{n_c} \quad (2.5)$$

P is the probability,

n is the total number of events

n_1 is the number of times Outcome 1 occurs,

n_2 is the number of times Outcome 2 occurs,

n_c is the number of times Outcome c occurs,

p_1 is the probability of Outcome 1

p_2 is the probability of Outcome 2,

p_c is the probability of Outcome c.

$$E(n_i) = np_i \tag{2.6}$$

$$\sigma = \sqrt{np_i(1-p_i)} \tag{2.7}$$

2.2.3 Poisson Distribution.

Thirdly, the Poisson distribution describes the distribution of the number of events occurring in a specified length of time or space, where the intervals of time or space are independent with respect to the occurrence of an event. Sometimes it is called the distribution of rare events. this is because the occurrence of an event is "rare" in the sense that no more than one success is improbable. Let X is the number of events occurring in a specific length of time or space, the value of X will be non negative integer, its distribution should place its mass on that range. Poisson probabilities depend on a single parameter, the mean μ. The Poisson probability mass function is:

$$P(x) = \frac{e^{-\mu}\mu^x}{x!},\ x = 0,1,2,.... \tag{2.8}$$

A very important feature of the Poisson distribution is that its variance equals its mean.

$$E(x) = V(X) = \mu \tag{2.9}$$

Its skewness is described by:

$$\frac{E(x-\mu)^3}{\sigma^3} = \frac{1}{\sqrt{\mu}} \tag{2.10}$$

Distribution approaches normality as μ increases. Poisson distribution is also applies as an approximation for the binomial distribution when *n* is large and μ is small, with $\mu = np$.

2.3 Statistical Inference for Categorical Data

The analysis of categorical data commonly involves the proportion of "successes" in a given population. This may consist of estimating a single parameter, comparing two parameters, or investigating the potential relationship between two or more categorical variables. In general given a particular probability distribution to describe the data, the probability is defined for all the potential values of *P* between 0 and 1. The likelihood function is the probability of the data as a function of the parameters. The maximum likelihood estimate (MLE) is the value of the parameter that maximizes this function. Thus, for the set of parameters MLE is the value that gives the observed data, the highest probability of occurrence.

Inference for categorical data deals with the MLE and tests derived from maximum likelihood. Generally for the binomial outcome of *x* successes in *n* trials, the maximum likelihood estimate of *P* equals $P = \frac{x}{n}$ this is the sample proportion of successes for the *n* trials.

Tests of the null hypothesis of zero-valued parameters can be carried out through the Wald Test, Likelihood Ratio Test, and the Score Test. These tests are based on different aspects of the likelihood function, but are asymptotically equivalent. However, in smaller samples they yield different answers and are not equally reliable.

The binomial parameter p $SE = \sqrt{\frac{P(1-P)}{n}}$ when $H_0 : \beta = 0$ is true, the test statistic is:

$$Z = \frac{(\hat{\beta} - \beta_0)}{SE} \tag{2.11}$$

This has approximately a standard normal distribution. Equivalently, z^2 has approximately a chi-squared distribution with *df* = 1. This type of statistic, which uses the standard error evaluated at the ML estimate, is called a Wald statistic. The *Z* or chi-squared test using this test statistic is called a Wald test.

At the same time the likelihood Ratio Test uses the likelihood function through the ratio of two maximizations of it: the maximum over the possible parameter values that assume the null hypothesis to be true and the maximum over the larger set of possible parameter values, permitting the null or the alternative hypothesis to be true. Let ℓ_0 denote the maximized value of the likelihood function under the null hypothesis, and let ℓ_1 denote the maximized value more generally. The likelihood-ratio test statistics is:

$$-2\log(\frac{\ell_0}{\ell_1}) \tag{2.12}$$

The reason for taking the log transform and doubling is that it yields an approximate chi-squared sampling distribution. Under $H_0 : \beta = 0$ the likelihood ratio test statistic has a large-sample chi-squared distribution with *df*= 1. The test statistic $-2\log(\frac{\ell_0}{\ell_1})$ must be nonnegative and relatively small values of $\frac{\ell_0}{\ell_1}$ yield large values of $-2\log(\frac{\ell_0}{\ell_1})$ *and* strong evidence against *H*0.

A third test is called the score test. It finds standard errors under the assumption that the null hypothesis holds. For the binomial distribution that uses the maximum likelihood estimator in statistical inference for the parameter *P*, the ML estimator is the sample proportion; *p* hence the sampling distribution of *p* is approximately normal for large n. The sampling distribution of the sample proportion *p* has mean and standard error

$$E(p) = P, \tag{2.13}$$

$$\sigma(p) = \sqrt{\frac{P(1-p)}{n}} \tag{2.14}$$

Consider the null hypothesis $H_0 : P = P_0$ that the parameter equals some fixed value P_0 the test statistic

$$Z = \frac{p - p_0}{\sqrt{\frac{P_0(1-P_0)}{n}}} \tag{2.15}$$

A significance test merely indicates whether a particular value for a parameter is plausible. We learn more by constructing a confidence interval to determine the range of plausible values. Let *SE* denote the estimated standard error of p(A) large sample $100(1-\alpha)$ % confidence interval for *P* has the formula

$$P \pm z_{\frac{\alpha}{2}}(SE)_{\text{with}}\, SE = \sqrt{\frac{P(1-P)}{n}} \tag{2.16}$$

2.4 Contingence Table

Contingency table is a type of table in a matrix format that displays the (multivariate) frequency distribution of the variables. It is used to summaries categorical variables. It shows the frequency distribution of the values of the dependent variable, given the occurrence of the values of the independent variable. Both variables must be grouped into a finite number of categories.

Table 2. 1 Contengency table

	A_1	**A_2**	**Total**
B_1	a	b	a+b
B_2	c	d	c+d
Total	a+c	b+d	a+b+c+d

Probabilities for contingency tables can be of three types joint, marginal, or conditional. Marginal probability or simple probability, of an event of interest is the marginal total for that particular event divided by the grand total. For example, in the above table, the marginal probability is $\frac{a+c}{a+b+c+d}$. Joint probabilities occur in the body of the two-way contingency table at the intersection of two events for each categorical variable. For example, in the above table the Joint probability is $\frac{a}{a+b+c+d}$. The final type of probability is conditional probability.

It measures the probability of an event given that another event has occurred. In the above table conditional probability is $\frac{a}{a+b}$.The conditional probability is especially important; if the row variable is fixed by design and not free to vary for each observation.

Row and column variables are independent if the conditional distribution of the column variable given the row variable is the same as the marginal distribution of the column variable (and vice versa). Equivalently, if all joint probabilities equals the product of their marginal probabilities: $P_{ij} = P_{i+}P_{+j}$ for all *I* and *j*. Thus, when the two variables are independent, knowledge of the value of the row variable does not change the distribution of the column variable, and vice versa.

Response variables having two categories are called binary Variables. The two groups on a binary response can be compared when data is displayed in a 2 × 2 contingency table, in which the rows are the two groups and the columns are the response levels of binary variable. One way in which we can compare the two groups on a binary response variable is by using difference of their proportions. Let P_1denote the probability of a success, so $1 - P_1$ is the probability of a failure. The difference of proportions $p_1 - P_2$ compares the success probabilities in the two rows. This difference falls between −1 and +1. It equals zero when $P_1 = P_2$, that is, when the response is independent of the group classification.

Let the sample sizes for the two groups be n_1 and n_2. When the counts in the two rows are independent binomial samples, the estimated standard error of $P_1 - P_2$ is:

$$SE = \sqrt{\frac{P_1(1-P_1)}{n_1} + \frac{P_2(1-P_2)}{n_2}} \qquad (2.17)$$

As the sample size increases, the standard error decreases, and hence the estimate of P_1- P_2 improves.

A large-sample 100(1 − α)% (Wald) confidence interval for P_1-P_2 is:

$$(P_1 - P_2) \pm z_{\frac{\alpha}{2}}(SE) \qquad (2.18)$$

Using the difference of proportions alone, to compare two groups can be misleading, especially when the proportions are both close to zero. This is because, a difference between two

proportions of a certain fixed size usually is more important when both proportions are near 0 or 1, than when they are near the middle of the range. In such cases, the ratio of proportions is a more relevant descriptive measure. The relative risk compares the proportions of the two groups in a ratio. If the rows of a 2x2 table represent groups and the columns represent a binary response, then the relative risk of a positive response is the ratio.

$$\text{Relative Risk} = \frac{P_1}{P_2} \tag{2.19}$$

Relative risk can be any nonnegative real number, hence relative risk of 1.00 occurs when $P_1 = P_2$, that is, when the response is independent of the group.

The odds of ratio are another measure of association for 2 × 2 contingency tables. The odds ratio is the ratio of odds of a positive response by group for a probability of success *P*, the *odds* of success are defined to be

$$\theta = \frac{P_{11} / (1 - P_{11})}{P_{12} / (1 - P_{12})} = \frac{P_{11} P_{22}}{P_{12} P_{21}} \tag{2.20}$$

When $\theta = 1$, the row and column variables are independent, as the Values of θ gets further from 1 in a given direction represent stronger association. When θ = 0.25, the odds of "success" in group 1 (row 1) are 0.25 times the odds in group 2 (row 2), or equivalently, the odds of success in group 2 are 4 times the odds in group 1 [1]. The odds ratio can be used with a joint distribution of the row and column variables too. Indeed, it can be used with prospective (rows totals fixed), retrospective (column totals fixed), and cross-sectional designs. If the rows and columns are interchanged, the value of the odds ratio does not change. The sample odds ratio uses the observed sample counts, n_{ij} However, if the probability of the outcome is close to zero for both groups, then the odds ratio can be used to approximate relative risk.

Unless the sample size is extremely large, the sampling distribution of odds ratio is highly skewed. Because of this skewness, statistical inference for the odds ratio uses an alternative but equivalent measure – its natural logarithm, $\log(\theta)$. The log odds ratio is symmetric about zero,

in the sense that reversing rows or reversing columns changes its sign. Independence corresponds to log(θ) = 0. That is, an odds ratio of 1.0 is equivalent to a log odds ratio of 0.0. An odds ratio of 2.0 has a log odds ratio of 0.7. The sample log odds ratio, log θ, has a less skewed sampling distribution that is bell-shaped. Its approximating normal distribution has a mean of log θ and a standard error of [16]:

$$SE = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}} \tag{2.21}$$

Because the sampling distribution is closer to normality for log θ than θ, it is better to construct confidence intervals for log θ. A large-sample confidence interval for log θ is:

$$\log \hat{\theta} \pm Z_{\frac{\alpha}{2}}(SE) \tag{2.22}$$

2.5 Chi-Square

For an *I* x *J* contingency table, the hypothesis that the row and column variables are independent may be tested using a chi-square test (either likelihood ratio statistic or Pearson's chi-squared statistic). Although the two statistics are asymptotically equivalent and asymptotically chi-squared, Pearson's chi-squared statistic tends to be better for smaller counts.(Thompson, 2009). Let P denote the probability of success. For the null hypothesis that $P = 0.50$ the expected frequency of success equal $e = nf = \frac{n}{2}$ which also equals the expected frequency of failures. If H_0 is true, we expect about half the sample to be of each type. To judge whether the data contradict H_0 we compare (F_{ij}) to (e_{ij}). If H_0 is true F_{ij} should be close to e_{ij} in each cell. The larger the differences $(f_{ij} - e_{ij})$, the stronger the evidence against H_0. The test statistics used to make such comparisons have large-sample chi-squared distributions. The Pearson chi-squared statistic for testing H_0 is:

$$\chi^2 = \sum \frac{(F_{ij} - e_{ij})^2}{e_{ij}} \tag{2.23}$$

This statistic takes its minimum value of zero when all $(F_{ij} = e_{ij})$. For a fixed sample size, greater differences $(F_{ij} - e_{ij})$ produce larger χ^2 values and stronger evidence against H_0

To test the independency by using Chi-squared, we have to indentify the expected.

$$e_{ij} = np_{i+}p_{+j} = \frac{n_{i+}n_{+j}}{n} \tag{2.24}$$

This is the row total for the cell multiplied by the column total for the cell, divided by overall sample size. The null hypothesis of statistical independence is: $H_0 : P_{ij} = P_{i+}P_{+j}$ for all i and j.

CHAPTER 3

CATEGORICAL RESPONSE MODELS

3.1 Logistic Regression

For the last decades, the use of generalized linear models for categorical data has increased dramatically, especially for applications in the social sciences reasonably this reflects the extent of development during in recent years of sophisticated methods for analyzing categorical data. It also shows the increasing methodological sophistication of scientists and applied statisticians, most of who now realize that it is unnecessary and often inappropriate to use methods for continuous data with categorical responses [1].

Binary data are the most common form of categorical data hence the most popular model for binary data is logistic regression. Logistic regression analysis predicts the values on one dependent variable from one or more independent (predicting) variables when the dependent variable is dichotomous (meaning that it divides the respondents or cases into two exclusive groups such as having a particular illness and not having it). [17]. Logistic regression becomes increasingly important factor for reliant on sophisticated statistical modeling techniques. It can be applied to wide range of fields, particularly in Social science like health and diseases assessment. For example to evaluate the effect of new health programs, the impact of risk factors on disease, the effects of health behaviors, and a host of other public health concerns, logistic regression is an appropriate method of statistics. Many such analyses involve an outcome or dependent variable that is dichotomous “present” or “absent”“yes” or “no” etc. In these kinds of studies the logistic regression model has become the statistical model of choice. The reasons for this are the wide area of applications of the logistic regression. The other factors that make the

logistic regression important are: the ease of interpretation of the estimated coefficients of parameters as applied in log odds ratios, the ability to estimate the probability that a particular subject will develop the outcome, and the wide availability of easily used and reliable software to perform the computations.

Generally logistic regression can be divided in to three types; direct, sequential and stepwise. In the direct method all the in dependent variable are forced into the regression equation or they are part of it. In the sequential logistic regression it depends on the researcher to determine the order in which the independent variables are included into the model. In stepwise logistic regression independent variable are included or excluded from the regression equation depends on the statistical criteria [17].

It is not appropriate to use independent variables which are highly correlated into the same model. This is because they will cover each other's effects. Therefore the correlation between independent variables should be examined to afford the issue of multi-collinearity. The other important issue is the presence of outliers which also influence the validity of results result achieved by the logistic regression model.

All general linear models have three components and logistic regression is not an exception. These components are; the component which identifies the response variable Y and also assumes the probability distribution for it. The second component is systematic component which specifies the explanatory variable; lastly the third component is Link Function, which contains both random component and systematic component. It describes the relationship between the mean of the response μ and linear predictor. The link function can be identity link, log linear link, conical link or any other link.

$$g(\mu) = \alpha + \beta x \tag{3.1}$$

3.1.1 Parameter Estimation and Interpretation

Before conducting logistic regression, first it is necessary to determine whether there is a relationship between the outcome and the predictor(s). Assuming there is relationship between a set of predictors and the dependent variable, what is important is to determine whether this

relationship can be simplified by reducing the number of predictors without decreasing the predictive value of the model.

Consider a random variable Y which can take on one of two possible values. Similarly suppose there is a single explanatory variable X, which is quantitative. P(x) is the parameter of binomial distribution. Hence it is the success probability at value of x. the logit (logarithm of odds) of this probability is:

$$\text{logit}\left[P(x)\right] = \log\left(\frac{P(x)}{1-P(x)}\right) = \alpha + \beta x \tag{3.2}$$

This formula indicated that the logit increases by β for every 1 unit increase in X.

The probability of success P(X) can be derived from the above formula.

$$P(x) = \frac{\exp(\alpha+\beta x)}{1+\exp(\alpha+\beta x)} = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}} \tag{3.3}$$

The parameter β determines the rate of increase or decrease for P(X). The sign of β shows whether P(x) increases (β >0) or decreases (β <0). Furthermore when β =0, P(x) is constant which means that P(x) is identical at all x.

When logistic regression is calculated, the exponential function of the regression coefficient e^{β} is odds ratio associated with one unit increase in explanatory variable. Using odds and odds ratio is crucial part of interpretation of logistic regression.

$$Odds = \frac{P(x)}{1-P(x)} = \exp(\alpha+\beta x) = e^{\alpha}(e^{\beta})^{x} \tag{3.4}$$

The odds multiply e^{β} for every 1 unit increase in x.

3.1.2 Inference for logistic regression

The Wald statistics of logistic regression model with hypothesis of $H_0 : \beta = 0$ which states that the probability of success is independent of X is:

$$Z = \frac{\hat{\beta}}{SE} \tag{3.5}$$

Although Wald statistics is adequate and can be used to test the logistic regression parameter particularly when sample size is large, the likelihood-ratio test is more powerful and more reliable. This is because; it uses both the null maximized likelihood values as well as the alternative maximized likelihood values instead of just one of these values. This compares the maximum ℓ_0 of the log-likelihood function when $\beta = 0$ to the maximum ℓ_1 of the log-likelihood function for unrestricted β. This has an approximate chi-squared and this approximation is better than the chi-squared approximation by each Likelihood Ratio Test statistics alone. The confidence of interval of Wald statistics for parameter β in logistic regression model is:

$$\hat{\beta} \pm Z_{\frac{\alpha}{2}}(SE) \tag{3.6}$$

Logistic regression can have a multiple predictor variables. Some or all of these predictors can be categorical rather than quantitative or continuous. When a logit model is includes categorical predictors, there is a parameter for each category of each predictor. On the other hand, one of those parameters is redundant. For instant, suppose a binary response variable Y has two binary predictors, X and Z. Hence each of these predictors has the values 0 and 1to represents the two categories of each explanatory variable. The model for $P(Y=1)$ is:

$$\text{Logit}\left[P(Y=1)\right] = \alpha + \beta_1 x + \beta_2 z \tag{3.7}$$

One way to eliminate the redundancy of parameters in the logit model is to set the parameter for the last category equal to zero, and changing the definition of the remaining parameters to that of the deviation from this last parameter.

3.1.3 Multiple Logistic Regressions

Logistic regression like ordinary linear regression, take multiple explanatory variables.

$$\text{Logit}\left[P(Y=1)\right] = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k \tag{3.8}$$

Multiple logistic regressions is the direct extension of logistic regression. The parameter β_i is the effect of x_i on the log odds that $Y = 1$ controlling other variable constant. The quantity $\exp(\beta)$ represents the multiplicative effect on the odds of a 1-unit increase in x_i, by keeping constant the other covariates.

To ascertain whether some terms in the model is needed or not, we can compare the maximum likelihood values for that model and simpler model without those terms. In other wards, the likelihood- ratio test compares the maximized log-likelihood ℓ_1 for the full model to the maximized log-likelihood ℓ_0 for the simpler model in those parameters equal to 0.

On the other hand, a logit model that doesn't contain an interaction between two categorical predictors, assumes that the odds ratio between one of the explanatory variables and the binary dependent variable is the same at each level of other predictor. In this case the predictor and the dependent variable are said to be conditionally independent. Meantime we can allow interaction between predictor variables by adding cross products of terms.

Sometimes the interpretive effect in logistic regression using multiplicative effects on the odds, which corresponds to odds ratios, is difficult to understand [1]. Therefore probability based interpretation can be used as an alternative way. Suppose a setting of predictors at which $\hat{P}(Y = 1) = \hat{P}.$ controlling for the other predictors a 1unit increase in x_j corresponds approximately to a $\hat{\beta}_j \hat{P}(1 - \hat{P})$ change in $\hat{P}$. [15] gives more details.

3.1.4 Logistic regression model checking

For a logistic regression model, it is guarantee that the model fits the data well. One of the methods used for checking model fit is likelihood- ratio model comparison test. This compares the model with more complex ones. Another important method of detecting lack of fit is goodness of fit. It compares the model fit with the data. Goodness of fit can be realized by utilizing the following statistics [18]:

$$\chi^2 = \sum \frac{(observed - \exp ected)^2}{\exp ected} \tag{3.9}$$

χ^2 has approximate chi-squared null distribution with degree of freedom subtract the number of parameters in the model from the number of parameters in the complex model. Therefore large value of χ^2 provide evidence of lack of fit.

3.2 Multinomial logistic regression

A multinomial (polytomous) logistic regression model is a form of regression where the outcome variable is binary or dichotomous and the independents variables are continuous, categorical variables, or both. Multinomial logistic regression is a simple extension of binary logistic regression that allows for more than two categories of the dependent or outcome variable. Like binary logistic regression, multinomial logistic regression uses maximum likelihood estimation to evaluate the probability of categorical membership [19].

Multinomial logistic regression does necessitate careful consideration of the sample size and examination for outlying cases. Like other data analysis procedures, initial data analysis should be though and include careful univariate, bivariate and multivariate assessment.

Specifically, multi-collinearity should be evaluated with simple correlations among the independent variables. Also, multivariate diagnostics (standard multiple regression) can be used to assess for multivariate outliers and for the exclusion of outliers or influential cases.

Sample size guidelines for multinomial logistic regression indicate a minimum of 10 cases per independent variable [20].

Multinomial logistic regression is often considered an attractive analysis because; it does not assume normality, linearity, or homoscedasticity. A more powerful alternative to multinomial logistic regression is discriminant function analysis which requires these assumptions are to be met.

Indeed, multinomial logistic regression is used more frequently than discriminant function analysis because the analysis does not have such assumptions. Multinomial logistic regression does have assumptions, such as the assumption of independence among the dependent variable choices. This assumption states that the choice or membership in one category is not related to the choice or membership of another category. Furthermore, multinomial logistic regression also assumes non-perfect separation. If the groups of the outcome variable are perfectly separated by

the predictor(s), then unrealistic coefficients will be estimated and effect sizes will be greatly exaggerated [21].

3.2.1 Multicategory logit models for nominal responses

Multinomial logit models for nominal responses use separate binary logit model for each pair of response categories. Suppose, J is the number of categories of Y. Let $P_1,...,P_J$ be the success probabilities which satisfy $\sum_j P_j = 1$. The probability distribution for this number of outcomes of the J types is multinomial. In the model of this kind, the order of listing the categories is inappropriate, since the model treats the response scale as nominal. Furthermore, logit models for nominal response variable pair each category with a baseline category. Suppose the J is the last category, the baseline category logits is:

$$\log \frac{P_j}{P_J} = \alpha_j + \beta_j x \quad \text{j} = 1,\ldots, \text{J} - 1 \tag{3.10}$$

Therefore the model has J - 1 equations.

For further illustration, when J = 3, for instance, the model uses:

$$\log \frac{P_1}{P_3} = \alpha_1 + \beta_1 x \tag{3.11}$$

$$\log \frac{P_2}{P_3} = \alpha_2 + \beta_2 x \tag{3.12}$$

The equation for other arbitrary of pair of categories of 3.11 and 3.12 can be constructed:

$$\log \frac{P_1}{P_2} = \log \frac{P_1 / P_3}{P_2 / P_3} = \log \frac{P_1}{P_3} - \log \frac{P_2}{P_3} = (\alpha_1 + \beta_1 x) - (\alpha_2 + \beta_2 x) = (\alpha_1 - \alpha_2) + (\beta_1 - \beta_2)x \tag{3.13}$$

On the other hand, success probabilities of categories of response variable can be estimated by

$$P_j = \frac{e^{\alpha_j + \beta_j x}}{\sum e^{\alpha + \beta x}} \text{ , j = 1, . . . , J} \quad (3.14)$$

The denominator is the same for each probability, and the numerator varies according to different categories.

3.2.2 Multicategory logit model for ordinal response

When the response categories are ordered, the logit considers the order by using cumulative logits. Suppose Y is response variable, the cumulative probability of response variable is the probability that Y falls at or below a specific point. Therefore the cumulative probability is :

$$P(Y \leq j) = P_1 + ... + P_j \text{ , j = 1 , . . . , J} \quad (3.15)$$

In addition to that the logit of the cumulative probability is:

$$\text{logit}\left[P(Y \leq j)\right] = \log\left[\frac{P(Y \leq j)}{1 - P(Y \leq j)}\right] \text{, j = 1 , . . . , J - 1} \quad (3.16)$$

Suppose that J = 3, the logit of the cumulative probability of $Y \leq 1$ is:

$$\text{logit}\, P(Y \leq 1) = \log\left[\frac{P_1}{P_2 + P_3}\right] \text{ and logit}\, P(Y \leq 2) = \log\left[\frac{P_1 + P_2}{P_3}\right] \quad (3.17)$$

For the explanatory variable x the logit model is:

$$\text{logit}\left[P(Y \leq j)\right] = \alpha_j + \beta x \text{, j = 1, . . . , J – 1} \quad (3.18)$$

The parameter β in the model describes the effects of x on the log odds of the response in category j or below. The model assumes that the effect of x on the log odds of the response categories is identical for all J – 1 cumulative logits, in other words the coefficients are identical. Hence it assumes the J – 1 logits have different intercept increasing in j.

The explanatory variable in cumulative logit models can be quantitative, categorical or both. When the categories are reversed in order, the result of the model fit is the same but the sign of $\hat{\beta}$ reverse.

Model interpretation can use odds ratios for the cumulative probabilities. The odds ratio comparing the cumulative probabilities for two values of x_1 and x_2 of x is:

$$\frac{P(Y \le j | X = x_2) \,/\, P(Y > j | X = x_2)}{P(Y \le j | X = x_1) \,/\, P(Y > j | X = x_1)} \tag{3.19}$$

The log of this odds ratio is the difference between the cumulative logits at these two values of X, which is proportional to the distance between the x values. For example when difference between the two x values is 1, the odds of response at any category is multiply by e^{β} for each unit increase in x. this log odds ratio can be written $\beta(x_2 - x_1)$ in which the same proportionality constant (β) applies in each cumulative probability. This property is called proportional odds assumption. Proportional odds models are fit using maximum likelihood estimation.

An alternative logits for ordered categories use pairs of categories instead of using the entire response scale in forming each logit for cumulative logit models. One approach forms logits for all pairs of adjacent categories. The adjacent categories logits are the j – 1 logits formed by the log odds of the probability of response talling in category j given that the response falls within either category j [15].

$$\log\left(\frac{P_{j+1}}{P_j}\right) = \alpha_j + \beta_j x, \text{ j = 1 , . . . , J – 1} \tag{3.20}$$

Suppose J = 3, the adjacent categories logit model has form:

$$\log\left(\frac{P_2}{P_1}\right) = \alpha_1 + \beta_1 x \text{ and } \log\left(\frac{P_3}{P_2}\right) = \alpha_2 + \beta_2 x \tag{3.21}$$

Like the baseline category logits, the adjacent categories logits resolve the logit of all pairs of response categories. The adjacent categories logit models differ from cumulative logit models in that effect of explanatory variables refer to individual response categories and not cumulative group.

Continuation-ratio logits is another approach that forms logits for ordered response categories in a chronological manner. The continuation ratio logit model is:

$$\log\left[\frac{P_1}{P_2}\right], \log\left[\frac{P_1+P_2}{P_3}\right], \ldots\ldots, \log\left[\frac{P_1+\ldots+P_{J-1}}{P_J}\right] \tag{3.22}$$

These are logits of conditional probabilities that the response falls in the j^{th}category for j = 1, . . .J – 1. The likelihood is the product of the multinomial mass functions. Some practical recommendations about the model choice and numerical examples can be found in Agresti (2002) [16].

3.3 Log-Linear Analysis

Log-linear analysis is multivariate statistical technique which can be applied to the contingency table for interpretation of categorical data. It determines the significance of associations between two or more variables for categorical data and their interactions. Log linear analysis is used instead of chi-square. This is because with chi-square you can only have two variables. If there are more than two variables, Chi-square analysis of each new variable must be performed separately. This consumes a lot of time doing so many chi-square analysis. To afford this kind of repetition, Goodman came up with a way to look at association between two or more categorical variables which he called log-linear analysis.

There are two important requirements for log-linear analysis. First requirement of log-linear analysis is that the data is independent. This means that no case or respondent can appear in more than one cell of the table which is being used. The second requirement is that the expected frequencies in the data table are sufficient. Hence these expected frequencies should all be greater than 1 and in no more than 20% of the cells should they be less than 5. In addition, there is no requirement about population distributions assumptions in log-linear analysis. In two way table when the responses are statistically independent, the joint cell probabilities (P_{ij}) are determined by the row and column marginal totals.

$$P_{ij} = P_{i+}P_{+j} \tag{3.23}$$

Log-linear analysis use expected frequency instead of cell probability, therefore the log- linear analysis under the independence the model formula is: $e_{ij} = np_{i+}p_{+j}$. By taking the log of both

sides of the equation yields an additive relation which depends on a term based on sample size, a term based on the probability in row i and a term based on the probability in column j. therefore the independence has the form:

$$\log e_{ij} = \lambda + \lambda_i^{\,x} + \lambda_j^{\,y} \qquad (3.24)$$

The parameter $\lambda_i^{\,x}$ represent the effect of classification in row i. the larger the value of $\lambda_i^{\,x}$, the large each expected frequency is in row i. similarly the parameter $\lambda_j^{\,y}$ represents the effect of classification in column j. the larger the value of $\lambda_j^{\,y}$ the large each expected frequency in column j. this kind of model is called log-linear model of independence.

The fitted (observed) values that satisfy the model is $\hat{e}_{ij} = \frac{n_{i+} n_{+j}}{n}$. This is the estimated frequencies for chi-square test. Hence the null hypothesis of independence is equivalently the hypothesis of chi-square test.

When the two variables in the contingency table are statistically dependent log-linear model takes this form:

$$\log e_{ij} = \lambda + \lambda_i^{\,x} + \lambda_j^{\,y} + \lambda_{ij}^{\,xy} \qquad (3.25)$$

The parameters $\lambda_{ij}^{\,xy}$ represent interactions between X and Y, which reflects that X and Y are dependent to each other, in other words, it reflects how far we deviate from the assumption of independence, whereby the effect of one variable on the expected cell count depends on the level of the other variables. On the other hand parameters $\lambda_{ij}^{\,xy}$ determine the log odds ratio. When these parameters are zero, the log odds ratio is zero, and X and Y are independent.

With a three way contingency tables, two factor association terms describe the conditional odds ratios between variables. The model that allows two pairs of variables to have conditional association controlling the other third variable is:

$$\log e_{ij} = \lambda + \lambda_i^{\,x} + \lambda_j^{\,y} + \lambda_k^{\,z} + \lambda_{ik}^{\,xz} + \lambda_{jk}^{\,yz} \qquad (3.26)$$

The term $\lambda_{ik}^{\,xz}$ permits association between X and Z, controlling for Y. the model also permits YZ controlling for X. it specifies the conditional independence of X and Y by controlling Z. models that contains only single-factor terms denoted by X, Y and Z which delete additional

association terms are called mutual independence models. Hence they are simple to fit most data set well compared to other models which reflect association. It treats each pair of variables as independent, both conditionally and marginally. However these kinds of models are rarely appropriate most studies, because mostly there is association between variables.

Models that permit all three pairs of variables to have conditional association are homogeneous association models and take in this form.

$$\log e_{ijk} = \lambda + \lambda_i^{x} + \lambda_j^{y} + \lambda_k^{z} + \lambda_{ij}^{xy} + \lambda_{ik}^{xz} + \lambda_{jk}^{yz} \tag{3.27}$$

These models show that conditional odds ratios between any two variables are the same at each level of the third variable. Hence this is the property of homogeneous association model. In addition, the most general, log-linear models associate the three variables and provide a perfect fit. Hence they take in this form:

$$\log e_{ijk} = \lambda + \lambda_i^{x} + \lambda_j^{y} + \lambda_k^{z} + \lambda_{ij}^{xy} + \lambda_{ik}^{xz} + \lambda_{jk}^{yz} + \lambda_{ijk}^{xyz} \tag{3.28}$$

Chi-square goodness of fit test is the most important test used for inference of log-linear models. As usual large sample chi-square statistics assess goodness of fit by comparing the cell fitted values to the observed counts. The chi-square Pearson statistic for three variable association of log-linear model is:

$$\chi^2 = \sum \frac{(n_{ijk} - \hat{e}_{ijk})^2}{\hat{e}_{ijk}} \tag{3.29}$$

The large values of χ^2 have smaller p- values and indicate poorer fit. Hence the models that lack any association tem fit poorly, having P-values below 0.001. Some practical numerical examples can be found in Alan Agresti (2002) [16].

CHAPTER 4

BAYESIAN FRAMEWORK

The Reverand Thomas Bayes (1702-1761) was the first person introduced the Bayesian framework. Similarly, Laplace was first person used the Bayesian estimation in 1786. Today, Bayesian statistics is widely used by researchers in wide range of fields. This is because the significant computational advancements including MCMC, OpenBUGS and WinBUGS software which are available in Bayesian approach make it an important tool for estimation a complex models. Researchers in many fields have preferred the Bayesian approach due to its capacity to handle complexity in real world problems [22].

The Bayesian approach has many attractive features; hence, can solve complex models over frequentist approach. These includes that the Bayesian approach gives a way to include prior knowledge concerning parameters of interest. In a frequentist approach, the data are taken as random while parameters are considered to be fixed. In a Bayesian approach, parameters follow a probability distribution and they are not considered to be fixed. Furthermore, parameters may be model parameters, missing data or events that are not observed (latent). Frequentist methods also typically rely upon approximations and asymptotic results. Hence these cases are only valid for large samples [23]. However Bayesian approach doesn't depend on approximation.

The product of the likelihood function and the prior density, are usually described in terms of summary statistics of the posterior distribution. This including the most probable values of the parameters (the mean, median or mode of the distribution) and regions of the parameter space with high posterior probability (credible intervals and regions) [13].

One of the difficulties, and most controversial components of Bayesian inference, is the selection of the prior distribution. Subjective Bayesian statisticians believe that the prior distribution should describe information about the parameters that is available before the data are collected, either from previous experiments or from expert knowledge on the system of interest data. The prior distribution is chosen to concentrate high probability in regions of the parameter space believed more plausible before collecting the data. Bayes' rule can then be viewed as an explicit formula for combining these prior beliefs with new information in the observed data to obtain new beliefs described by the posterior distribution. If the information in the data is very strong then the prior and posterior distribution may be very different; if the information is weak then the prior and posterior may be very similar.

An alternative approach is to select prior distributions that are vague or non-informative, so that inference is objective in so far as the shape of the posterior is determined almost exclusively by the data. Common choices for non-informative distributions are flat densities which are improper if the parameter's value is unbounded and admissible. This is only if the resulting posterior still integrates to distributions with large variances and families of reference distributions, like Jeffrey's priors, which are invariant to certain transformations of the parameters [22]. Be that as it mind, there is no single concept of a non-informative distribution and many formulations have been considered, both in general and for specific models [13].

A great advantage of Bayesian methods is that the prior distribution can be specified to incorporate extra structure in the model parameters that is not contained in the likelihood function.

4.1 Bayes Rule

If we have data (abbreviated as "D") at hand, and we want to estimate some parameters θ of the specified model, the likelihood function may help:

$$L(\theta) = f(D|\theta) \tag{4.1}$$

Bayesian approach takes the parameters as random samples from a certain prior distribution. $\theta \approx q(\theta)$. Therefore this makes the estimation and inference completely based on the posterior distribution $P(\theta|D)$. Posterior distribution can be calculated as follows:

$$P(\theta / D) = \frac{q(\theta) f(D/\theta)}{\int q(\theta) f(D/\theta) d\theta} \tag{4.2}$$

This is called the bayes rule

4.2 Prior distribution

Introducing prior distribution provides extra information in Bayesian approach. Hence there are three major ways to specify prior distribution:

1. Elicited priors: which are prior distributions specified by some experts or elicited people with experiences and knowledge in the field.
2. Conjugate priors: which are prior distributions selected from a members of the family which is conjugate to the likelihood function so that posterior distributions belongs to the same distributional family as the prior.
3. Non informative priors: normal distribution with infinite variance or uniform distribution over the support of θ is used as its prior distribution;

Among the three kinds of priors, Non-informative priors are simple and actually provide no extra information. In addition we can avoid the danger of introducing incorrect extra information.

4.3 Bayesian inference

Within the Bayesian framework, we can perform point estimation, interval estimation and hypothesis test based on the posterior distribution. Point and interval estimation are the most important.

4.3.1 Point estimation

In the univariate parameter case, the estimator $\hat{\theta}(D)$ can be a summary feature such as mean, median or mode of the posterior distribution $P(\theta|D)$. The based observations of the three features are:

- The posterior mode equals the maximum likelihood estimator with a flat prior.
- The mean and median are identical for symmetric posterior densities.
- All three measures are the same for symmetric unimodal posteriors.

The posterior variance with respect to $\hat{\theta}(D)$, or mean squared error can be defined as follows:

$$E_{\theta(D)}(\theta - \hat{\theta}(D))^2 \tag{4.3}$$

This provides a measure of accuracy of the point estimator. Suppose $\mu(D) = E_{\theta(D)}(\theta)$ as the posterior mean.

$$E_{\theta(D)}(\theta - \hat{\theta}(D))^2 = E_{\theta(D)}\left[\theta - \mu(D) + \mu(D) - \hat{\theta}(D)\right]^2$$

$$= E_{\theta(D)}(\theta - \mu(D))^2 + (\mu(D) - \hat{\theta}(D))^2$$

$$\geq E_{\theta(D)}(\theta - \mu(D))^2 = \text{var}_{\theta(D)}(\theta - \mu(D)) \tag{4.4}$$

The equal sign "=" is achieved if and only if $\hat{\theta}(D) = \mu(D)$, therefore; in general we use the posterior mean as the estimator to get the smallest variance.

4.3.2 Interval Estimation

The counterpart of a frequentist confidence interval (CL) in Bayesian framework is called credible set or Bayesian confidence interval. The $100(1-\alpha)\%$ credible set for θ is defined as a subset S of the support of θ such that:

$$1 - \alpha \leq P(S/D) = \int_S P(\theta/D)d\theta \tag{4.5}$$

However, there is a computation issue with Bayes approach. The posterior distribution

$$P(\theta/D) = \frac{q(\theta)f(D/\theta)}{\int q(\theta)f(D/\theta)d\theta} \tag{4.6}$$

is the basis for all estimation and inference, but with integration at the denominator is difficult to explicitly computed. In addition, it is necessary to compute the posterior mean and variance which involve additional integration. Usually in high dimensional space,

$$E_{\theta(D)}(\theta) = \int \theta P(\theta / D) d\theta \tag{4.7}$$

$$Var_{\theta(D)}(\theta) = \int (\theta - E_{\theta(D)}(\theta))^2 P(\theta / D) d\theta \tag{4.8}$$

However, we can perform the estimation and inference through sampling, which is idea of Markov Chain Monte Carlo Methods, which is a powerful tool to overcome the difficult in computing.

4.4 MCMC Methods

MCMC (Markov Chain Monte Carlo) was introduced by Metropolis et.al in 1950s in the field of statistical physics, and has been used in various fields of physics such as studies in liquid, magnets, and lattice gauge theory. In 1980s, it was introduced in statistical science, and after 1990s the combination of MCMC and hierarchical Bayesian models become a standard tool for advanced data analysis. Now MCMC is also applied to the other fields such as computer graphics and becoming recognized as a universal method for sampling from multivariate distributions with unknown normalization constants [24].

MCMC Methods are the most widely used methods for statistical computing now. They are based on a class of algorithms for sampling from probability distribution by constructing a Markov chain whose stationary distribution is the desired distribution [25]. After a large number of steps, a state of the chain can be used as a sample from the target distribution. This MCMC sampling can be used to approximate the distribution or to compute an integral.

In the steps of a particular MCMC implementation used in the multinomial logistic regression framework, a posterior Monte Carlo sample is generated by simulating a Markov chain in WinBUGS. The starting value, which is a reasonably distant from the posterior, is chosen. The second step is to sample a new value from proposed normal distribution. Then the chain is run for 10000 iterations and the first 1000 runs are discarded. The issues such as convergence are investigated to determine whether the sample can reasonably be treated as a set of random realizations from the target posterior distribution.

4.5 Gibbs sampling

Gibbs sampling provides a way to convert multivariate sampling problem into univariate sampling problem. Suppose we have:

$$\theta = (\theta_1,...,\theta_k) \approx q(\theta_1,...\theta_k). \tag{4.9}$$

If all conditional distributions $\left[q_i(\theta_i \middle| \theta_j, j \neq i), i = 1,...,k\right]$ are available for sampling, we start from an arbitrary set of initial values $(\theta_1^{(0)},...,\theta_k^{(0)}$, at iteration $t(t \geq 0)$, we proceed with the following sampling steps to get $(\theta_1^{(t+1)},...,\theta_k^{(t+1)}$

$$\theta_1^{t+1} \approx q_1(\theta_1 \middle| \theta_2^{(t)},....,\theta_k^{(t)})$$

$$\theta_2^{t+1} \approx q_2(\theta_2 \middle| \theta_1^{(t+1)}\theta_3^{(t)},....,\theta_k^{(t)})$$

.

.

.

$$\theta_k^{t+1} \approx q_k(\theta_k \middle| \theta_1^{(t+1)},....,\theta_{k-1}^{(t+1)})$$

Gibbs' convergence theorem states:

$$\theta_1^{(t)},....,\theta_k^{(t)} \to (\theta_1,...,\theta_k) \approx q(\theta_1,...\theta_k) \text{ as } t \to \infty \tag{4.10}$$

3.6 Metropolis-Hastings algorithm

Metropolis-Hastings algorithm is one of the most popular MCMC methods. By the simplest form of this method we choose auxiliary proposal function $h(x,y)$ such that $h(.,y)$ is a probability density function (pdf) with respect to X for any Y and symmetric. For example, $h(x,y) = h(y,x)$ At step t, we have a current state $\theta^{(t)}$, then we do the following:

1. Sample $\hat{\theta} \approx h(.,\theta^{(t)})$;
2. Compute the ratio r=min$\left\{1, \frac{q(\theta^{*})}{q(\theta^{(t)})}\right\}$;
3. Sample $u \approx U_{(0,1)}$ let

Set $\theta^{(t+1)} = \hat{\theta}$ with the probability ($\theta^{t}, \hat{\theta}$), if $u \leq r$ otherwise set $\theta^{(t+1)} = \theta^{t}$

Therefore it can be proved under the conditions $\theta^{(t)} \to \theta \approx q(\theta)$ as $t \to \infty$

Hastings made a simple but important generalization of the metropolis algorithm in 1970. He got that the proposal function $h(x, y)$ is symmetric. So the ratio becomes.

$$r = \min\left\{1, \frac{q(\theta^{*})h(\theta^{*}, \theta^{(t)})}{q(\theta^{(t)}h(\theta^{(t)}, \theta^{*})}\right\} \tag{4.11}$$

4.7 WinBUGS

WinBugs is the product of the BUGS (bayesian infrerence Using Gibbs Sampling) project. This project is joint program of the medical research Council of Biostatistics Unit a Cambridge Univeristy and the deparmtment of Epidemiology and public health of Imperial College at St.Mary's Hospital in London. WinBUGS is software is freely distributed. In this Software Models can be implemented in two ways [23].

1. Using the BUGS language
2. Using the graphical feature, Doodle BUGS which allows the specification of models in terms of a directed graph.

It is believed that WinBUGS is a very handly tool in fitting complex models. On the other hand, it is difficult and frustrating package to master. In additon, Bayesian analysis using WinBUGS requires three major tasks to be perfomed;

1. Model specification
2. Running Model
3. Bayesian Inference

4.8 Bayesian binary logistic regression model

In Bayesian terms, the conditional probability of a sequence of events approximates the product of the probability of events, and the probability of given event under the condition of the event of interest [26]. Assume there is an event A, and the probability that event A is true is denoted by P(A). In the meantime, P(A/y) is the conditional probability of A when the statement y is true. Here y can be visualized as the observed data or characteristics of a given sample.

The logistic regression model is used to explain the probability of a binary response variable as a function of some covariates. This is because in logistic regression, a single outcome variable $y_i (i = 1,...n)$ follows a Bernoulli probability function that either produces 1 or 0. It takes on the value 1 with probability P_i and 0 with probability $1 - P_i$. Thus P_i varies over the observations as an inverse logistic function of a vector x_i which includes a constant term and $k-1$ explanatory variables. For a sample of n observations $y_i (i = 1,...n)$ the logistic regression model for binary data is specified by

$$y_i / P_i = Bernoulli(P_i) \tag{4.12}$$

$$P_i = \Pr(y_i = 1) = F(x_i \ \beta) \tag{4.13}$$

Where

$y_i = 1$ if the response of interest is observed for the i^{th} individual, and

$y_i = 0$ zero otherwise.

$\beta = (\beta_1, \beta_2 ... \beta_k)$ is the *K* vector of unknown parameters and $x_i = (x_{i1} ... x_{ik})$ is vector of known covariate values associated to the i^{th} individual. Hence f is any transformation function assuming values in (0, 1). In addition f can be any arbitrary cumulative distribution function. Specifically in logistic regression transformation function is:

$$F(x_i \ \beta) = \frac{\exp(x_i \ \beta)}{1 + \exp(x_i \ \beta)} \tag{4.14}$$

The likelihood function is given by;

$$P(\beta / y, x) = \prod_{i=1}^{n} \left[F(x_i\ \beta)\right]^{y_i} \left[1 - F(x_i\ \beta)\right]^{1-y_i} = \prod_{i=1}^{n} \left[\frac{\exp(x_i\ \beta)}{1+\exp(x_i\ \beta)}\right]^{y_i} \left[\frac{1}{1+\exp(x_i\ \beta)}\right]^{1-y_i} \quad (4.15)$$

The next step for a Bayesian analysis is to provide a joint prior distribution over the parameter space if no prior information on the model parameters exists, or if it is difficult to elicit or formalize. Then initial uncertainty about the parameters can be quantified with a non informative prior distribution. This is identical to base the estimation on the information provided by the data only. Therefore the easiest way to provide join prior belief is to propose an informative prior, but with small precision, avoiding any complaint about the specification of subjective beliefs [9].

Hence in a Bayesian framework, the inference is based on the information provided by the posterior distribution of parameters β given by:

$$P(\beta / y, x) = P(\beta) \prod_{i=1}^{n} \left[F(x_i{}^t \beta)\right]^{y_i} \left[1 - F(x_i{}^t \beta)\right]^{1-y_i} \quad (4.16)$$

$$= P(\beta) \prod_{i=1}^{n} \left[\frac{\exp(x_i{}^t \beta)}{1+\exp(x_i{}^t \beta)}\right]^{y_i} \left[\frac{1}{1+\exp(x_i{}^t \beta)}\right]^{1-y_i}$$

where $P(\beta)$ is a prior distribution.

4.9 Bayesian Multinomial logit model

Let y_i be a sequence of categorical data, as $i = 1,..., N$, where each y_i is assumed to take a value in one of $m+1$ unordered categories labeled by $0,...m$. For each category K, with $1 \le k \le m$, the probability that y_i takes the value K depends on covariate x_i. The MNL (Multinomial logit) model takes the following way [27].

$$P(y_i = k / \beta_1,..., \beta_m) = \frac{\exp(x_i \beta_k)}{1 + \sum_{j=1}^{m} \exp(x_i \beta_j)} \quad (4.24)$$

where $\beta_1,..., \beta_m$ are category specific unknown regression coefficients of dimensions d and, x_i is a $(1 \times d)$ row vector containing covariates which are not category specific.

Furthermore, we assume that under the condition that $\beta_1,...,\beta_m$ are known, the observations are mutually independent. In addition, in order to make the model more identifiable, the parameter β_0 of the baseline category $k=0$ is set equal to 0.. Therefore, the parameter β_k is in term of the change in log-odds relative to the baseline category $K=0$. We also assume that the prior distribution $P(\beta_k)$ of β_k fits a normal distribution; $N_d(b_{k0}, B_{k0})$ with known hyper-parameters b_{k0} and B_{k0}.

CHAPTER 5

RESULTS AND DISCUSSION

5.1 Frequentist Estimation for Multinomial Logistic Regression

Data and Research solution (DARS) organization in Somaliland conducted survey about Somaliland youth life style in 2011. The aim of the survey was to realize why hundreds of Somaliland youth people embarks every months on a perilous journey cross the Mediterranean sea on their way to Europe while others lives hopelessly in the country. Hence to dig out why Most of these youth people are not totally satisfy with lives Somaliland.

This was done randomly in household level in four regions in Somaliland. The target population was general population between the age of 15 and 35. Method of data collection was face to face interview using pen and paper methodology. Hence the target sample was 800 respondents, however successfully completed interviews were 794 respondents. The first section of the survey was all about the demographics like gender, age, educational level, occupational status and marital status, however the other sections touched all aspects of Somaliland youth life style. For instant media consumption habits, leisure time activities, challenges and opportunities, preference of food and beverages and above all how happy they are in their life in Somaliland.

By the way, we analyze Somaliland youth life style dataset to apply the theory part of this paper. This is because it contains a nice collection of continuous, ordered and categorical which is applicable with this research paper. The data set contains variable on 794 respondents. The outcome variable is happiness with three categorical levels happy, neutral and unhappy. The predictor variable is educational level, with three categorical levels Primary, Secondary and University and age with continuous variable

On the other hand, multinomial logistic regression is the linear regression analysis to conduct when the dependent variable is nominal with more than two levels. Thus it is an extension of logistic regression which analyzes dichotomous (binary) dependents. Like all linear regression the multinomial logistic regression is a predictive analysis, hence it used to describe data and to explain the relationship between one dependent nominal variable and one or more nominal or continuous independent variables. Multinomial logistic regression is multi-equation model similar to multiple linear regressions for nominal dependent variable with k categories; hence the multinomial logistic regression model estimates k-1 logit equations. In addition it allows selecting freely one category as baseline category to compare the other with.

Therefore, according to this dataset, Somaliland youth are different in the way they are happy in their life in Somaliland, hence some are happy and satisfy with their life, other are neither happy or unhappy and they neutral situation, while others are unhappy and totally not satisfy with their life in Somaliland. Their happy situations might be modeled using with their age and their educational level.

The table 5.1 shows the cross tabulation of educational levels with different levels of happiness. It is clear that most youth who are unhappy in Somaliland life are those with low level of education particularly those with primary level of education. Similarly most youth with university level of education are happy with their life in Somaliland.

Table 5. 1 Crosstabulation of educational level with happiness

Happiness					Total
Educational level		Happy	Neutral	Unhappy	
	Primary	194	44	42	280
	Secondary	197	29	36	262
	University	200	24	28	252
Total		591	97	106	794

The table 5.2 shows the mean age with levels of happiness. It shows that the older youth seems are happy with their life in Somaliland compared to the younger ones.

Table 5. 2 Mean and standard deviation of age with different levels of happiness

Happiness	Mean	Standard devaition
Happy	21.9	4.9
Neutral	22.5	5.3
Unhappy	23.8	5.7

The presence of a relationship between the dependent variable (Happiness) and combination of independent variables (Educational level and age) is based on statistical significance of the final model Chi-square. In the table 5.3 the distribution reveals that the probability of the model Chi-square (23.4) is 0.001 which is less than the level of significance of 0.05. The null hypothesis that there was no difference between the model without independent variables and the model with independent variable was rejected. This suggests that the existence of a relationship between the independent variables and dependent variable is supported, hence accepting the alternative hypothesis.

Table 5. 3 Model fitting information

Model	Model fitting Criteria	Likelihood ratio Tests		
	-2 log likelihood	Chi-square	df	sig
Intercept Only	406.5			
Final	383.1	23.4	6	.001

Moreover, Table 5.4 shows testing for the overall effect of educational level and Age to the happiness; hence this reveals that effects are statistally significant.

Table 5. 4 Likelihood ratio test

Model	Model fitting Criteria	Likelihood ratio Tests		
	-2 log likelihood	Chi-square	df	sig
Intercept	383.1	.000		
Age	398.6	15.5	2	.0001
Educational level	394.2	11.2	4	0.025

Once the relationship is established, the next important thing to do is to establish the strength of multinomial logistic regression relationship. To assess the strength of these relationship, the

evaluation of the usefullness for logistic model needs to be considered. In this case using Cox and Snell R square and the Negelkerke R square values can be assessed the strength of relationship. Hence they provide the indication of the amount of variation in the dependent variable. These are considered as R square. the disribution in the table 5.5 reveals that the values are 0.029 and 0.038 respectively. This suggested that between 2.9% and 3.8% of the variability is explained by the variables used in the model.

Table 5. 5 R- square

Cox and Snell	**0.029**
Nagelkerke	0.038

First before running our model, we choose the happy type of happiness to be the level of our outcome that we wish to use our baseline by specifying it in the relevel function. then we run our model by using multinon function from nnet package to estimate a multinomial logistic regression model. We chose the multinon function because it does not require the data to be reshaped. The mutinnon package doesnot include p-value calculation for the regression coefficients, so we calculate p-values using wald tests as we see belwo in Z-test.

Table 5. 6 Parameter coefficients with standard error in brackets

Coefficients	Intercept	Secondary	University	Age
Neutral	-2.2(0.49)	-0.43(0.26)	-0.72(0.28)	0.03(0.02)
Unhappy	-3.2(0.48)	-0.15(0.25)	-0.60(0.27)	0.08(0.02)

Table 5.6 shows the output of multinomial logti model with happy category of happiness as the baseline category with two explanatory variables of age (contious variable) and education level with three level of categories with primary level of education as the baseline categoreis. The software for multinomial logit model fits all equations simultaneously. For simultaneous fitting, the same parameter estames occur for a pair of categories no mater which category is the base line. Hence the choice of the baseline category is arbitray. In our example, response variable which is happiness situation for Somaliland youth between the age of 15-35 with three categories of happy, unhappy and Neutral. Hence happy situation is the baseline category. Let secondary level of ducation be x_1, university level of education be x_2 and age be x_3. The ML prediction eqautions are:

$$\log \frac{Unhappy}{Happy} = \alpha_1 + \beta_1 x_1 + \beta_1 x_2 + \beta_1 x_3 + \varepsilon \quad (5.1)$$

$$\log \frac{Unhappy}{Happy} = -3.2 - 0.15x_1 - 0.60x_2 + 0.08x_3 \quad (5.2)$$

$$\log \frac{Neutral}{Happy} = \alpha_2 + \beta_2 x_1 + \beta_2 x_2 + \beta_2 x_3 + \varepsilon \quad (5.3)$$

$$\log \frac{Neutral}{Happy} = -2.2 - 0.43x_1 - 0.72x_2 + 0.03x_3 \quad (5.4)$$

- The log odds of being in neutral situation of happiness according to in happy situation of happiness will decrease by 0.72 by moving from primary level of education to univeristiy level of education
- The log odds of being in neutral situation of happiness according to in happy situation of happiness will decrease by 0.43 by moving from primary level of education to secondary level of education.
- The log odds of being in unhappy situation of happiness according to in happy situation of happiness will decrease by 0.60 by moving from primary level of education to univeristiy level of education.
- The log odds of being in unhappy situation of happiness according to in happy situation of happiness will decrease by 0.15 by moving from primary level of education to secondary level of education.
- A one year increase in the variable age is assocaited with the increase in the log odds of being in unhappy situation according to being in happy situation in the amount of 0.08.
- A one year increase in the variable age is assocaited with the increase in the log odds of being in neutral situation according to being in happy situation in the amount of 0.03.

On the other hand, to determine the equation of the other pairs of caterories, for instant to decide the the log odds of being in unhappy accoring to neutral situation of happiness, we can munipulate the first pairs of equations:

$$\log\frac{Unhappy}{Neutral} = \log\frac{Unhappy/Happy}{Neutral/Happy} = \log\frac{Unhappy}{Happy} - \log\frac{Neutral}{Happy} \quad (5.5)$$

$$= (\alpha_1 + \beta_1 x_1 + \beta_1 x_2 + \beta_1 x_3) - (\alpha_2 + \beta_2 x_1 + \beta_2 x_2 + \beta_2 x_3) \quad (5.6)$$

$$= (\alpha_1 - \alpha_2) + (\beta_1 - \beta_2)x_1 + (\beta_1 - \beta_2)x_2 + (\beta_1 - \beta_2)x_3 \quad (5.7)$$

$$= (-3.2 + 2.2) + (-0.15 + 0.43)x_1 + (-0.60 + 0.72)x_2 + (0.08 + 0.03) \quad (5.8)$$

$$\log\frac{Unhappy}{Neutral} = -1 + 0.28x_1 + 0.12x_2 + 0.11x_3 \quad (5.9)$$

- The log odds of being in unhappy situation of happiness according to in neutral situation of happiness will increases by 0.12 by moving from primary level of education to univeristiy level of education
- The log odds of being in unhappy situation of happinessaccording to in neutral situation of happiness will increase by 0.28 by moving from primary level of education to secondary level of education
- A one year increase in the variable age is assocaited with the increase in the log odds of being in unhappy situation according to being in Neutral situation in the amount of 0.11

Table 5. 7 Z test

	Intercept	Secondary	University	Age
Neutral	-4.48	-1.67	-2.59	1.57
Unhappy	-6.62	-0.58	-2.23	3.87

Table 5.8 reveals the p-value of paramters which shows the evidence of variable effect to the model. Hence the effect of secondary category to model is not significant. In addition the effect age to the neutral situation of happiness is not significant.

Table 5. 8 P value

	Intercept	Secondary	University	Age
Neutral	0.0007	0.09	0.009	0.12
Unhappy	0.00003	0.55	0.03	0.0001

The ratio of the probability of choosing one outcome category over the probablity of the baseline category is a relative risk and commonly refered as odds ratio, particularly when the the probabiliy of choosing category and the probability of baseline are both close to zero. The relative risk is the right-hand side linear equation exponentiated. This is leading to the fact that the exponentiated regression coefficients are relative risk ratios for a unit change in the predictor variable. Hence we can exponentiate the coefficients from our model to see these risk ratios.

Table 5. 9 Extracting the coefficient from the model and exponentiation

Program type	Intercept	Secondary	University	Age
Neutral	0.11	0.65	0.49	1.03
Unhappy	0.04	0.86	0.55	1.08

- The relative risk ratio for a one year increase in the variable age is 1.03 times for being in neutral situation according to being in happy situation.
- The relative risk ratio for a one year increase in the variable age is 1.08 times for being in unhappy situation of happiness according to being in happy situation happiness.
- The relative risk ratio by moving from primary level of education to University level of education is 0.49 times for being in neutral situation of happiness according to being in happy situation.
- The relative risk ratio by moving from primary level of education to University level of education is 0.55 times for being in unhappy situation of happiness according to being in happy situation.
- The relative risk ratio by moving from primary level of education to secondary level of education is 0.65 times for being in neutral situation of happiness according to being in happy situation.
- The relative risk ratio by moving from primary level of education to secondary level of education is 0.86 times for being in unhappy situation of happiness according to being in happy situation.

The multinomial logit model has an alternative expression in term of response probabilities. This means the predicted probabilities can be used to help to understand the model.

$$P_j = \frac{e^{\alpha_j + \beta_j x_1 + \beta_j x_2 + \beta_j x_3}}{\sum_h e^{\alpha_h + \beta_h x_1 + \beta_h x_2 + \beta_h x_3}} \quad (5.10)$$

The denominator is the same for each probability of each observation, hence it sum of the numerators for various happiness categories including baseline category with various j or observations. In this way we can calculate predicted probabilities for each of our outcome level using the fitted function. Table 5.10 shows the predicted probability for the observations in our dataset and viewing the first six rows.

Table 5. 10 Predicted probability

Observation	Happy	Neutral	Unhappy
1	0.80	0.10	0.10
2	0.75	0.11	0.14
3	0.62	0.17	0.21
4	0.81	0.10	0.09
5	0.82	0.09	0.09
6	0.79	0.10	0.12

5.2 Bayesian Estimation for Multinomial Logistic Regression

In this section we present an application of Bayesian Multinomial logistic regression in the prediction of Happiness of Somaliland youth with their age and their educational level. Bayesian Multinomial logistic regression was fitted by using WinBUGS (Bayesian Inference using Gibbs Sampling). The chains are run for 10000 iterations and the first 1000 runs are discarded as a burn in. we separated the simulation in to two models Happiness and Educational level in one model and Happiness and Age in another model.

5.2.1 Happiness and Educational Level

Table 5. 11 Posterior means, posterior standard deviations and 95% credible intervals

Node	Mean	Sd	MC error	2.5%	Median	97.5%	start	sample
alpha[2]	3.772	0.1517	0.003136	3.464	3.778	4.058	1001	10000
alpha[3]	3.722	0.156	0.002987	3.407	3.726	4.017	1001	10000
beta[2,2]	-0.4233	0.2413	0.00438	-0.9111	-0.4204	0.04505	1001	10000
beta[2,3]	-0.1521	0.2318	0.003743	-0.6131	-0.1514	0.3014	1001	10000
beta[3,2]	-0.6153	0.2547	0.004415	-1.123	-0.6119	-0.1259	1001	10000
beta[3,3]	-0.4089	0.2456	0.004051	-0.9032	-0.4074	0.06539	1001	10000

The WinBUGS result in the table 5.11 indicates the output of multinomial logti model which the outcome variable is happiness with three categorical levels happy, neutral and unhappy, with happy category of happiness as the baseline category and the explanatory variables of education level with three level of categories with primary level of education as the baseline categoreis.The table shows Node which is the name of the unknown quantity equivalent to the model parameter; the approximation of the average of the posterior distribution of the unknown quantity which is the model parameter coefficient, an approximation of the standard deviation of the posterior distribution and computational accuracy of the mean. Furthermore it shows selected percentiles which are median or 50^{th} percentile of the posterior distribution, the 97.5^{th} percentile or an approximation of the upper endpoint of the 95% credible interval and the 2.5 percentile or an approximation of the lower end point of the 95% credible interval. It also shows the starting simulation after discarding the start-up and the number of simulations used to approximate the posterior distribution

The MC error is purely technical like round-off error. It quantifies the variability in the estimate that is due to Markov chain variability, in other words it is an estimate of the difference between the mean of the sampled values and the true posterior mean. Hence it should be small, less than 5% of the posterior standard deviation for a parameter. Hence it can be made as small as possible by increasing the number of simulations. Table 5.11 suggests that MC errors are low in comparison to the corresponding estimate posterior standard deviation. This is evidence that Markov chain has converged. The posterior standard deviation is used as the standard error of the

parameter estimate. The range between 2.5^{th} and 97.5^{th} percentiles represents 95% Bayesian confidence interval and is called a credible interval. Let secondary level of ducation be x_1, university level of education be x_2;

$$\log \frac{Unhappy}{Happy} = 3.722 - 0.15x_1 - 0.41x_2 \quad (5.11)$$

$$\log \frac{Neutral}{Happy} = 3.772 - 0.42x_1 - 0.62x_2 \quad (5.12)$$

- The log odds of being in neutral situation of happiness according to in happy situation of happiness will decrease by 0.62 by moving from primary level of education to university level of education
- The log odds of being in neutral situation of happiness according to in happy situation of happiness will decrease by 0.42 by moving from primary level of education to secondary level of education.
- The log odds of being in unhappy situation of happiness according to in happy situation of happiness will decrease by 0.41 by moving from primary level of education to university level of education
- The log odds of being in unhappy situation of happiness according to in happy situation of happiness will decrease by 0.15 by moving from primary level of education to secondary level of education

Figure 5.1 shows the visual inspection of the time series plot produced by History. These plots appear to stabilize in. Hence they suggest that the Markov chain have converged.

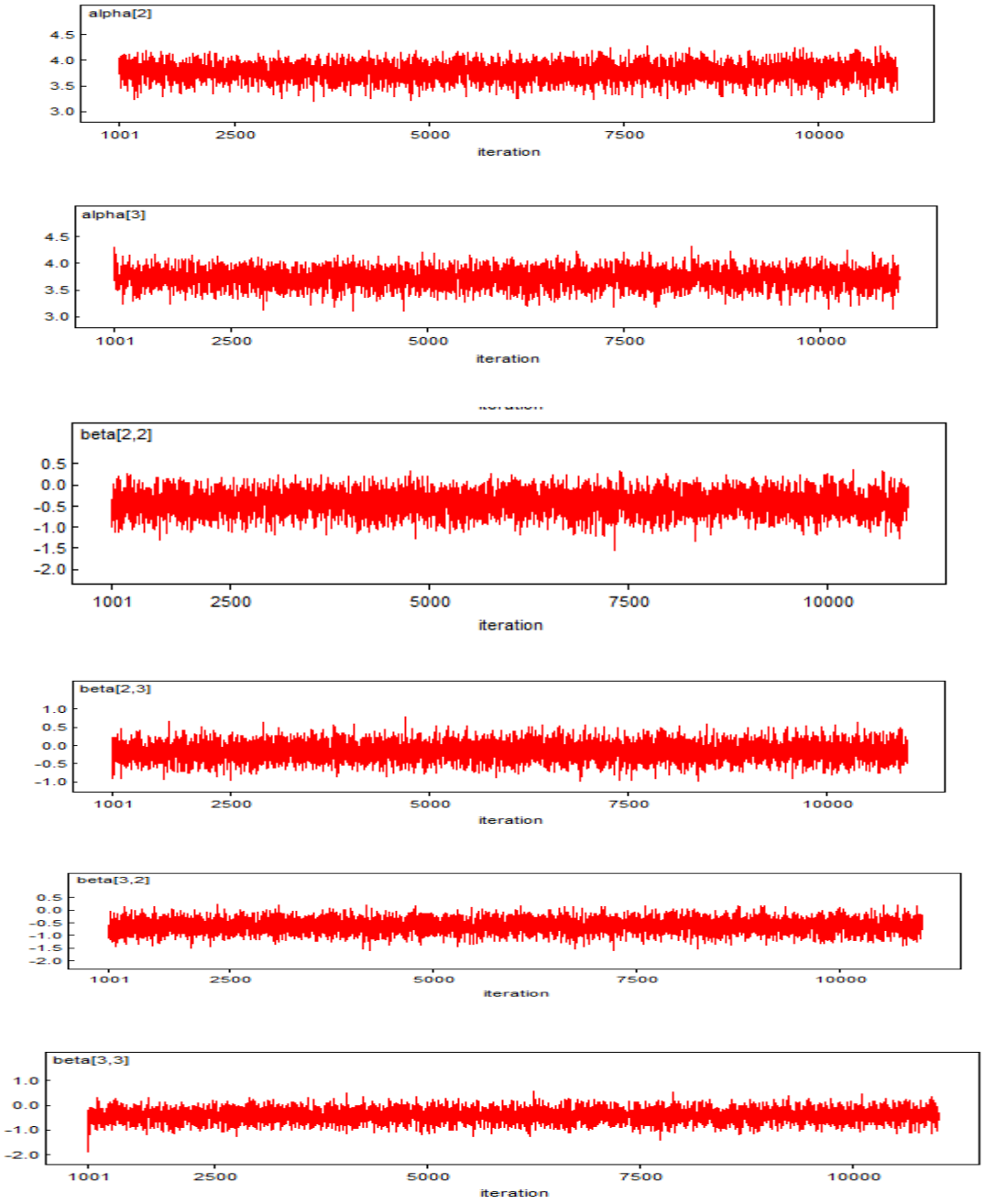

Figure 5. 1 Time series plot of MCMC output (History plot) (Continue)

Figure 5.2 shows density Plot which provides a graphical representation of the posterior density estimate for each node. Hence density plot shows that posterior distribution of each node is almost normally distributed.

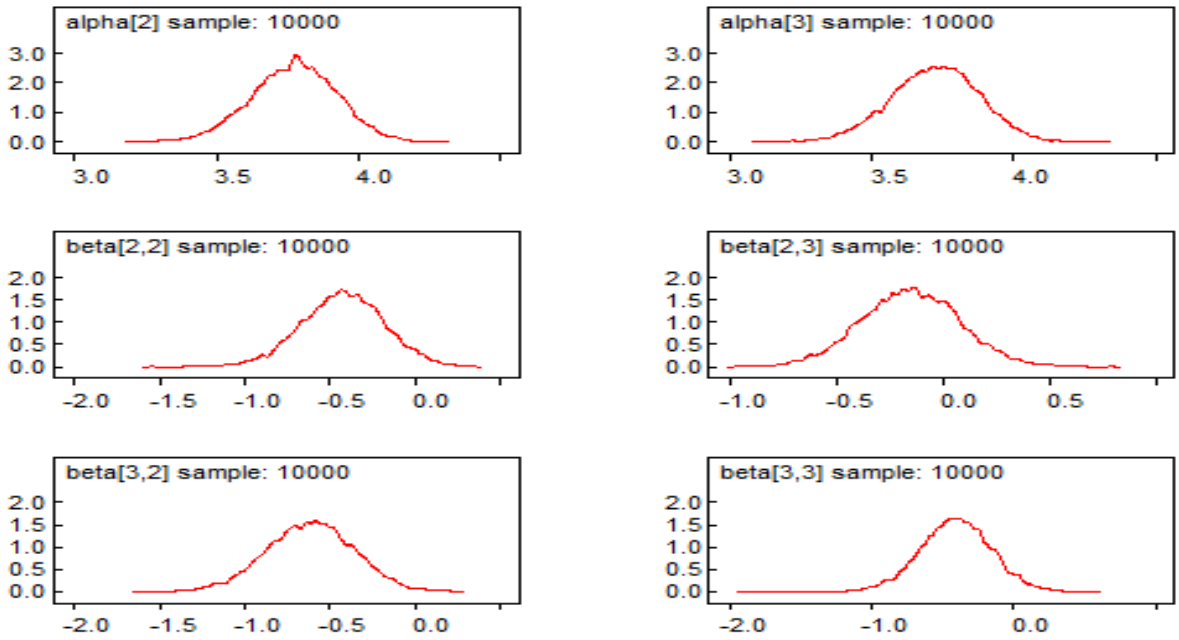

Figure 5. 2 Posterior probability density plots

Figure 5.3 shows the autocorrelation plots. These plots appear to dampen quickly; therefore this provides an evidence of the convergence of the Markov chain and also suggests that it may be appropriate to average Markov chain output.

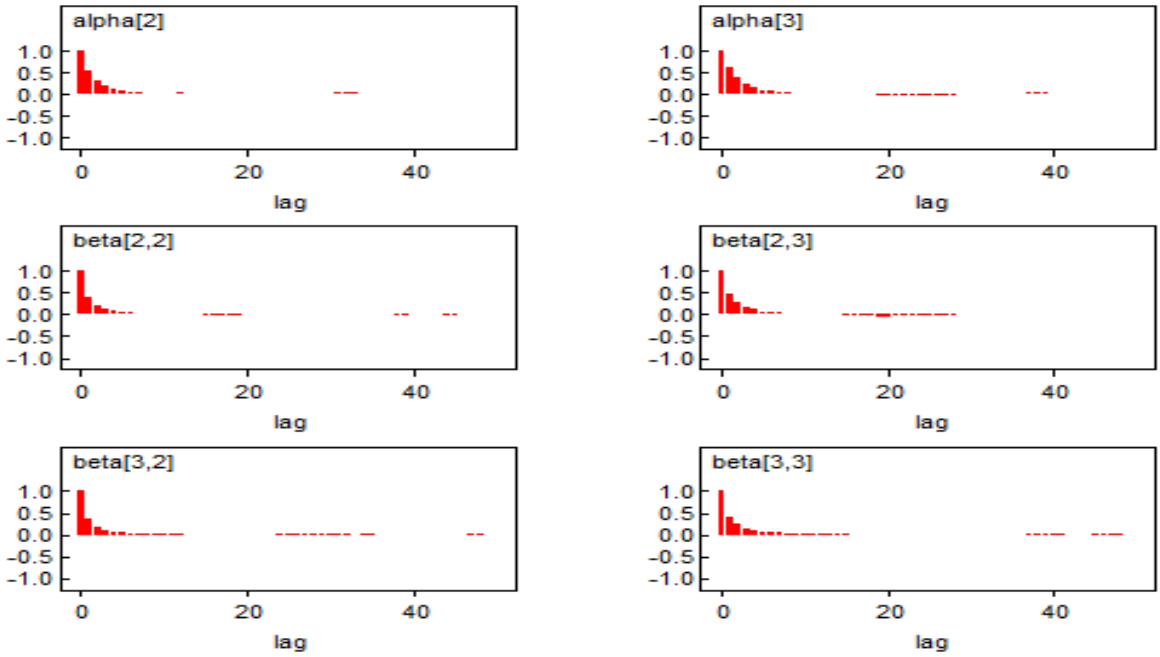

Figure 5. 3 Sample autocorrelation plots

Figure 5.4 shows quantile plot. Quantile plot shows the median and the 2.5% and 97.5% percentiles for each iteration. Hence it is a plot running mean with 95% confidence interval against iteration number. Therefore it clear that the requested quantiles have being stabilized implying that Markov chain has converged in terms of parameters.

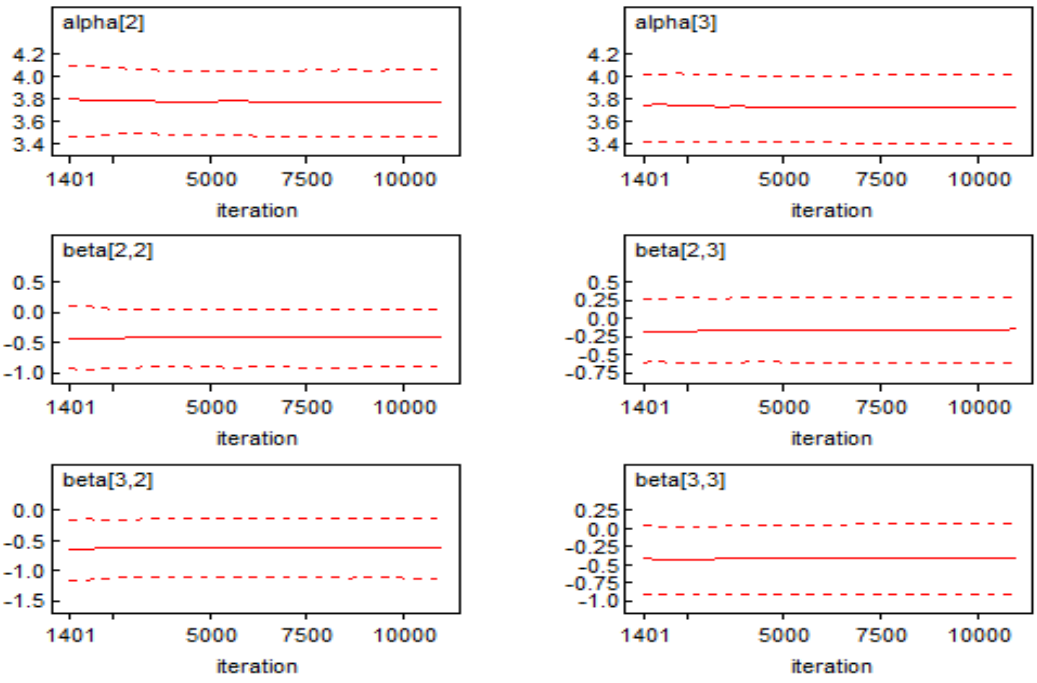

Figure 5. 4 Quantile plots

5.2.2 Happiness and Age

Table 5. 12 Posterior means, posterior standard deviations and 95% credible intervals

Node	Mean	Sd	MC error	2.5%	Median	97.5%	start	sample
Beta1[1]	-2.329	0.4967	0.005223	-3.304	-2.332	-1.364	1001	10000
Beta1[2]	0.02313	0.02158	2.314E-4	-0.0194	0.02336	0.06506	1001	10000
Beta3[1]	-3.303	0.4809	0.005259	-4.258	-3.3	-2.361	1001	10000
Beta3[2]	0.06922	0.02002	2.055E-4	0.03055	0.06904	0.1083	1001	10000

The winbugs result in Table 5.12 indicates the output of multinomial logti model which the outcome variable is happiness with three categorical levels happy, neutral and unhappy, with happy category of happiness as the baseline category and the explanatory variables of age which is continous variable. The table 13 provides that MC errors are low in comparison to the corresponding estimate posterior standard deviation. This is evidence that Markov chain has converged. Let variable age be x_3;

$$\log \frac{Unhappy}{Happy} = -3.3 + 0.07x_3 \tag{5.13}$$

$$\log \frac{Neutral}{Happy} = -2.3 + 0.023x_3 \tag{5.14}$$

- A one year increase in the variable age is assocaited with the increase in the log odds of being in unhappy situation according to being in happy situation in the amount of 0.07
- A one year increase in the variable age is assocaited with the increase in the log odds of being in neutral situation according to being in happy situation in the amount of 0.023

Figure 5.5 shows the time series plots produced by History. These plots appear to stabilize in. Hence they suggest that the Markov chain have converged.

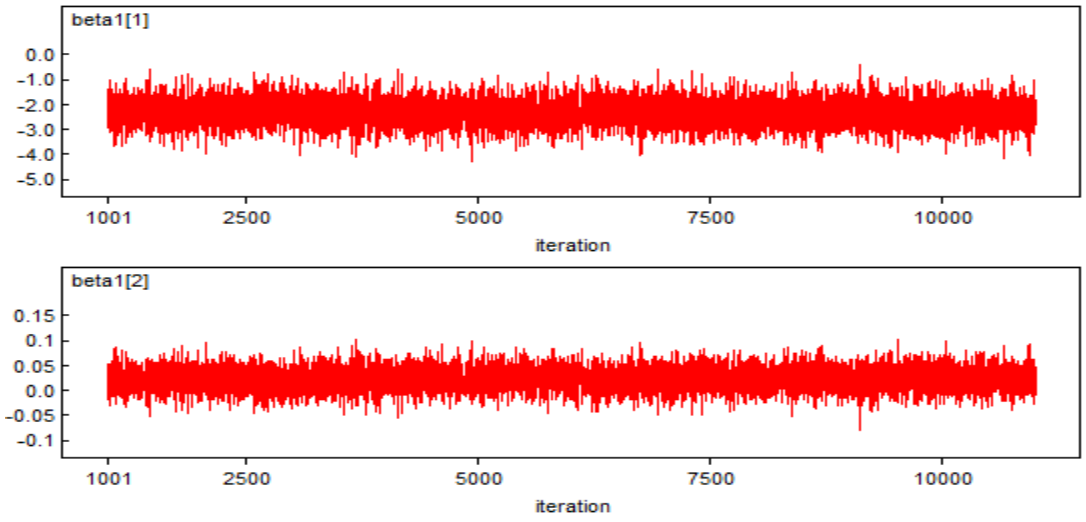

Figure 5. 5 Time series plot of MCMC output

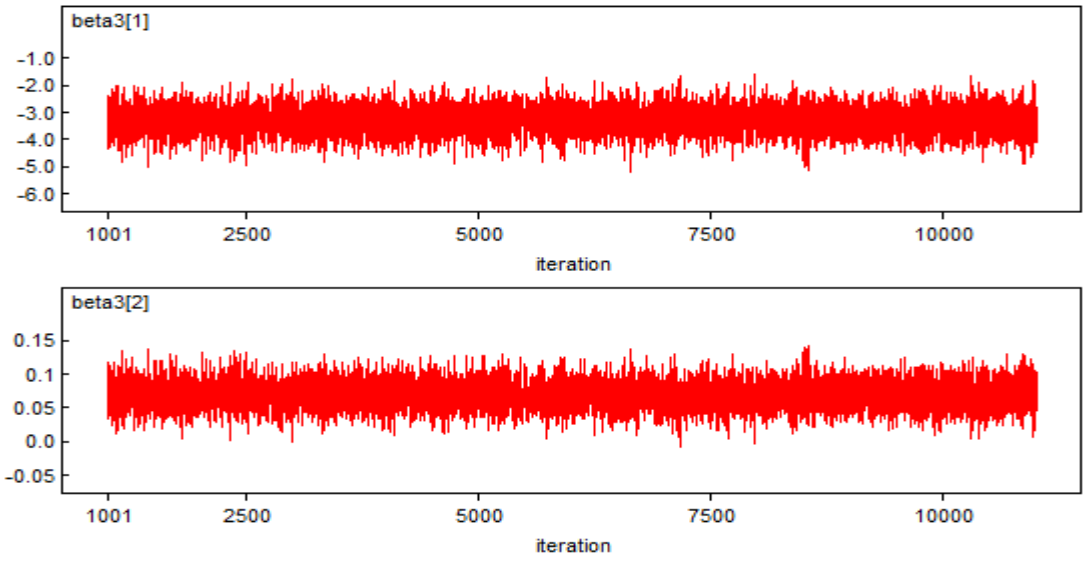

Figure 5. 6 Time series plot of MCMC output (Continue)

Figure 5.6 shows density Plot which provides a graphical representation of the posterior density estimate for each node. Hence density plot shows that posterior distribution of each node is almost normally distributed.

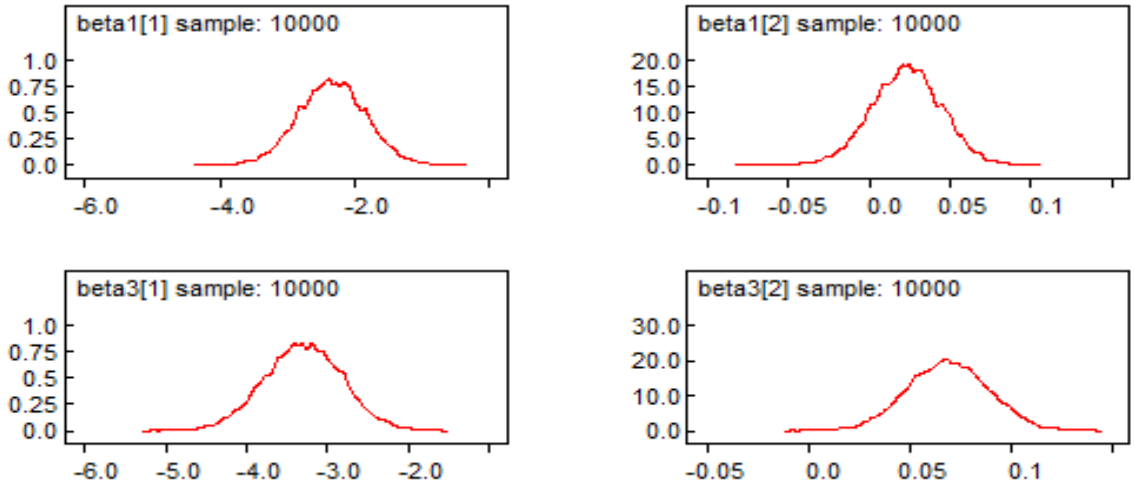

Figure 5. 7 Posterior probability density plots

Figure 5.7 shows the autocorrelation plots. These plots appear to dampen quickly; therefore this provides an evidence of the convergence of the Markov chain.

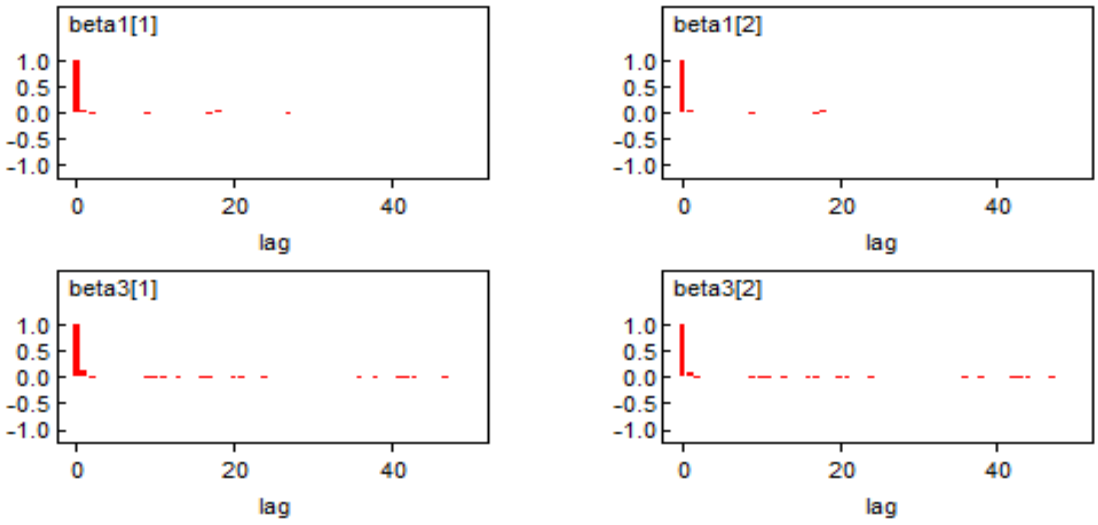

Figure 5. 8 Sample autocorrelation plots

Figure 5.8 shows quantile plot which shows the median and the 2.5% and 97.5% percentiles in each iteration. Hence it is a plot running mean with 95% confidence interval against iteration number.Therefore it provides that the requested quantiles have being stabilized implying that Markov chain has converged in terms of parameters.

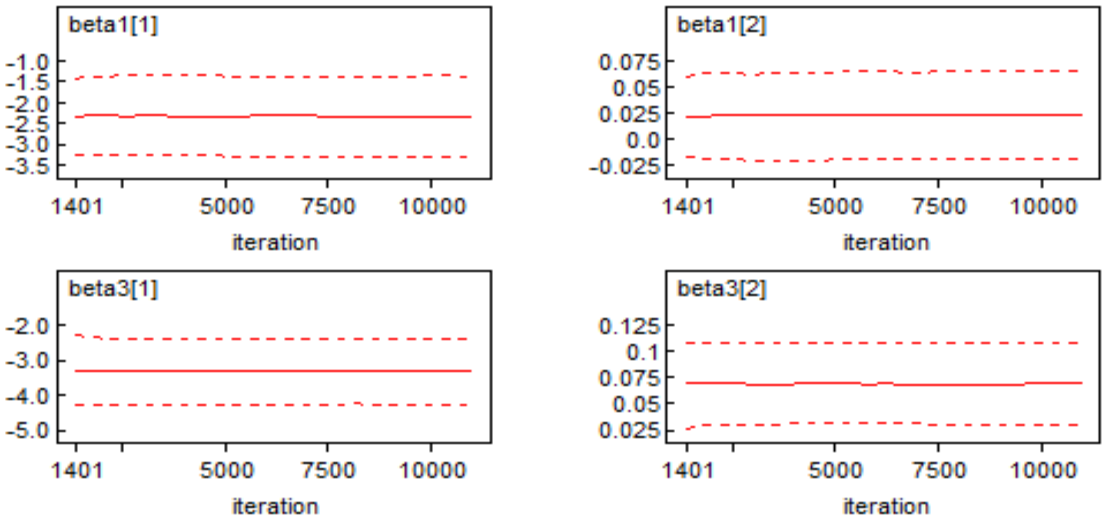

Figure 5. 9 Quantile plots

5.3 Discussion

The results supported the existence of overall relationship between the dependent variable which is happiness and the combination of two independent variables; age and educational level. This situation is shown by the model chi-square of 23.4 with p-value of 0.001 which is less than level of significance of 0.05.

At the same time, if an independent variable has an overall relationship to the dependent variable, it might or might not be statistically significant in differentiating between pairs of groups defined by the dependent variables. This is because the interpretation for an independent variable focuses on its ability to distinguish between pairs of groups and the contribution which makes to change the odds of being in one dependent variable category rather than the other. Therefore according to table 5.8 there is significant relationship between the independent variable age and unhappy category of dependent variable, however the relationship between dependent variable age and neutral category of dependent variable is not significant. As well, university category of independent variable of education level is significant in distinguished both categories of neutral and unhappy of dependent variable while secondary category is significant with neutral category, but not unhappy category of dependent variable.

Bayesian estimation for multinomial logistic regression seems to be very useful and performs very well to predict the happiness of Somaliland youth. The visual inspection of the time series plots shows that the Markov chains have converged and finally reached a stable and equilibrium distribution. Similarly, MC error is less than 5% of the posterior standard deviation for a parameter in both table 5.11 and table 5.12. Therefore, we have concluded that we have a valid sample from the posterior distribution of the parameters of the model of the mean. The node statistics give the formal parameter estimates of multinomial logistic regression.

Comparing the two results from classical estimation and Bayesian estimation for Multinomial logistic regression parameters, they produced almost the same result. Both result supported the conjecture that for those with high education and younger in age are happier than those with low education and older in age.

5.4 Conclusion

With reference to the research objectives of Bayesian estimation for multinomial logistic regression, firstly, consideration was given to the classical estimation of multinomial logistic regression. The parameters in the model were estimated using the method of maximum likelihood. Odds ratios for the response variables were calculated from the parameters of the fitted model. In order to test hypotheses in logistic regression, the study used the likelihood ratio test and Wald test. This is done to compare and contrast the results with those from Bayesian method of estimation.

In this research study, we investigated modeling Happiness of Somaliland youth with their age and their educational level. The finding revealed that the probability of the model chi-square which is less than the level of significance of 0.05, suggesting a statistically significant relationship between the independent variables of age and educational level and the dependent variable of happiness. Similarly, the odds of parameter coefficients and odds ratios which compares the odds of the response variable categories for those in sets of values of the explanatory variables, both of them suggested that happiness increases with educational level and decreases with age. To make more simple, this research study which is subject to changes if any further studies are made in the future showed that increasing the happiness of Somaliland youth would be explained by increases in educational level and younger in age.

In addition, in this research study, we showed that Bayesian Multinomial logistic regression is useful and can directly compute very accurate approximations to the posterior marginal density. This is because the MCMC simulation results almost marched to the results of estimating parameters by using maximum likelihood. Besides, our practical application showed that the Bayesian framework allows the use of additional information that can provide more accurate estimate.

REFERENCES

[1] Agresti A. (2007). Introduction to Categorical data analysis, Second Edition, Wiley, New York.

[2] Hosmer D.W and Lemeshow S. (2000). Applied logistic regression, Second Edition, Wiley.

[3] Agrenti A, and Hitchcock D.B. (2005). "Bayesian Inference for Categorical data Analysis" Journal of Statistical Methods and Applications, 14: 297-330

[4] Frühwirth R. and Frühwirth R. (2012). "Bayesian Inference in the Multinomial Logit Model", Austrian Journal of Statistics, 41(1): 27 – 43.

[5] Held L. and Holmes C.C. (2006). "Bayesian Auxiliary Variable Models for Binary and Multinomial Regression", Journal of Bayesian Analysis, 1(1):145-168.

[6] Agresti A.,Caffo B. and Hartzel J. (2001). "Multinomial Logit Random effects Models", Journal of Statistical Modeling, 1(81):102.

[7] Albert H.J and Chip S. (1993). "Bayesian Analysis of Binary and Polychotomous Response Data", Journal of the Americal Statistical Association, 88 (422): 669 – 679.

[8] Moala A. F., Santos A. M., and Tachibana M.V. (2009) "Approximate Bayesian Methods for Logistic Regression Model", Rev. Bras. Biom., Sao Paulo, 27 (2): 288-300.

[9] Migon S. A. and Souza A.D.P. (2004)."Bayesian Binary Regression Model: An Application to In-Hospital Death after Ami Prediction", Versao Online ISSN 1687-5142.

[10] Tsai C.H.(2004). "Bayesian Inference in Binomial Logistic Regression: A Case study of the 2002 Taipei Mayoral Election", Chinese Journal, 94(3):103-123.

[11] Brein S.M. and Dunson B.D. (2004), "Bayesian Multivariate Logistic Regression", Biostatistics Branch, MD A3-A3, National Institute of Environmental Health Science.

[12] Frühwirth R. and Frühwirth S. (2006)., "Auxiliary Mixture Sampling with Application to Logistic Models", IFAS Research Paper Series 2006-14, Austrian Academy of Sciences, Institute of High Engergy Physics.

[13] Gelman A. (2006)., "Prior Distributions For Variance parameters in Hierarchical Models", Journal of Bayesian Analysis, 1(3): 515-533.

[14] Arcos A., Molina D., Ranalli M.G. and Rueda M.D.M. (2014).,"Multinomial Logistic Estimation In Dual frame Surveys", Proceedings of the 14th International Conference on Computational and Mathematics Methods in Science and Engineering.

[15] Thompson L.A.(2009)., R(and S-PLUS) Manual to Accompany Agresti's Categorical data Analysis(2002), Second Edition.

[16] Agresti A. (2002). Categorical data analysis, Second Edition, Wiley, New York.

[17] Barkus E., Foster J., and Yavorsky C. (2006). Understanding and Using Advanced Statistics, first addition, Sage publications, London.

[18] Geweke J. (2013)." Bayesian Inference For Logistic Regression Models Using Sequential Posterior Simulation", University of Technology Sydney.

[19] Bayaga A. (2010)., "Multinomial Logistic Regression: Usage and Application in Risk Analysis", Journal of Applied Quantitative Methods, 5(2): 288-297.

[20] Chan Y. H. (2005), Multinomial Logistic Regression", Biostatistics 305, Singapore Med J, 46(6):259.

[21] Chip S. and Greenberg E. (1998) "Analysis of Multivariate Probit models", Biometrika, 85(2): 347-361.

[22] Muthukumarana S.(2010)., "Bayesian Methods and Applications using WinBugs", PHD Thesis, department of Statistics and Actuarial Science, Simon Fraser University.

[23] Jinhue L. (2006). "Analysis Of Longitudinal data with missing values" PHD Thesis, Department of Statistics, University of California, Los Angeles.

[24] Jackman S. (2000)., "Estimation and Inference Via Bayesian Simulation: An Introduction to Markov Chain Monto Carlo", American Journal of political Science, 44(2):375-404.

[25] Scollnik D. P.M. (1996), "An Introduction to Markov Chain Monte Carto Methods and Their Actuarial Applications", Proceedings of the Casualty Actuarial Society, LXXXIII: 114-165.

[26] Rekkas M. (2009), "Approximate Inference for the Multinomial logit model", Statistics and probability letters, Elsevier, 79(2): 237-242.

[27] Fradkin D, Genkin A., Lewis D. and. Madigan D. (2005), "Bayesian Multinomial logistic Regression For Author Identification", DIMACS, Rutgers University.

APPENDIX A

R CODE FOR MULTINOMIAL LOGISTIC REGRESSION

```
with(datta1,table(Edulevel, Happy))

with(datta1,do.call(rbind,tapply(Age, Happy,function(x)c(M=mean(x),SD=sd(x)))))

datta1$Happy2<-relevel(datta1$Happy,ref="Happy")
test<-multinom(Happy2~Edulevel+Age,data= datta1)

summary(test)

Z<-summary(test)$coefficients/summary(test)$standard.errors
z


p<-(1-pnorm(abs(z),0,1))*2
p


exp(coef(test))

head(pp<-fitted(test))
```

APPENDIX B

WINBUGS CODE: HAPPINESS AND EDUCATIONAL LEVEL MODEL

```
model
{

# PRIORS
alpha[1] <- 0;
for (k in 2 : K) { alpha[k] ~ dnorm(0, 0.00001)}
for (k in 1 : K){  beta[1, k] <- 0 }
for (i in 2 : I) {
beta[i, 1] <- 0 ;
for (k in 2 : K){  beta[i, k] ~ dnorm(0, 0.00001)}
  }

 # LIKELIHOOD
for (i in 1 : I) {

# Multinomial response
   #    X[i,1:K] ~ dmulti( p[i,1:K] , n[i]  )
   #    n[i] <- sum(X[i,])
   #    for (k in 1:K) {
   #      p[i,k]       <- phi[i,k] / sum(phi[i,])
   #      log(phi[i,k]) <- alpha[k] + beta[i,k]
```

```
  #     }

for (k in 1 : K) {
        X[i,k] ~ dpois(mu[i, k])
        log(mu[i, k]) <- alpha[k] + beta[i, k]
              }
          }
          }


        Data
        list( I   = 3,    K   = 3,
        X = structure(.Data = c(194, 44, 42, 197, 29, 36,  200,  24,  28), .Dim = c(3, 3)))


        Initials
        list(alpha = c(NA, 0, 0),
        beta = structure(.Data = c(NA, NA, NA,
                        NA, 0, 0,
                                    NA, 0, 0), .Dim = c(3, 3))
```

APPENDIX C

WINBUGS CODE: HAPPINESS AND AGE MODEL

```
model
{
for(i in 1:n)
        {
                X1[i] <- log(X[i])
                Y[i] ~ dcat(p[i,])
p[i, 1] <- ebx[i, 1]/ebxsum[i]
p[i, 2] <- ebx[i, 2]/ebxsum[i]
p[i, 3] <- ebx[i, 3]/ebxsum[i]

ebxsum[i] <-  ebx[i, 1]+ebx[i, 2]+ebx[i, 3]

ebx[i, 2]  <-exp(  b1[1] + b1[2]*(X[i]-mean(X[]))/sd(X[]))
ebx[i, 1]  <- 1 # Baseline is Category 1 for an identification
ebx[i, 3]  <-exp(  b3[1] + b3[2]*(X[i] - mean(X[]))/sd(X[]) )
}

# Compute Pr(correct classifcations | data)
for(i in 1:n)
{
```

```
ind1[i] <- equals(Y[i], 1)  # 1 if Y[i]=1, 0 otherwise.
ind2[i] <- equals(Y[i], 2)  # 1 if Y[i]=2, 0 otherwise.
ind3[i] <- equals(Y[i], 3)  # 1 if Y[i]=3, 0 otherwise.

ind.correct[i]  <- equals(p[i,Y[i]], ranked(p[i,], 3))
        # 1 if correctly classified, 0 otherwise.
        # ranked(p[i,], 3): the 3rd smallest prob. out of 3. <-> the max prob.
        ind1.correct[i] <- ind1[i]*ind.correct[i]
        ind2.correct[i] <- ind2[i]*ind.correct[i]
        ind3.correct[i] <- ind3[i]*ind.correct[i]

}
p1.correct <- sum(ind1.correct[])/sum(ind1[])
p2.correct <- sum(ind2.correct[])/sum(ind2[])
p3.correct <- sum(ind3.correct[])/sum(ind3[])
p.correct<- mean(ind.correct[])

for(i in 1:2)
        {
b1[i] ~ dnorm(0, 1.0E-6)
b3[i] ~ dnorm(0, 1.0E-6)
        }

beta1[1] <- b1[1]-b1[2]*mean(X[])/sd(X[])
beta1[2] <-     b1[2]/sd(X[])

beta3[1] <- b3[1]-b3[2]*mean(X[])/sd(X[])
beta3[2] <-     b3[2]/sd(X[])
        }
}
```

# Data 1
list(n=794)
# Data 2

Y[]	X[]
1	17
1	22
3	28
1	16
1	21
1	25
1	17
1	30
1	18
1	33
1	18
1	16
.	.
.	.
.	.
1	17
1	19

END

Init

b1[]	b3[]
0	0
0	0

END

CURRICULUM VITAE

Personal Information

Name : Khadar Mahamed GAHAYR

Date of Birth and place : 17/09/1985, Hargaisa, Somaliland

Foreign Language : English, Turkish, Arabic

Email : khadargahayr@yahoo.com

EDUCATION

Degree	**Department**	**University**	**Year of graduation**
Undergraduate:	Mathematics and Statistics	University of Hargaisa	2011
High school:	SOS Sheikh Secondary School		2007

WORK EXPERIENCE

Year	Corporation/Institute
2010 - 2011:	University OF Hargaisa
2011- 2012:	Somaliland Ministry of National Planning and Development