Journal of Social and Development Sciences
Vol. 4, No. 4, pp. 193-197, Apr 2013 (ISSN 2221-1152)

Bayesian Logistic Regression Modelling via Markov Chain Monte Carlo Algorithm

Henry De-Graft Acquah
University of Cape Coast, Cape Coast, Ghana
henrydegraftacquah@yahoo.com

Abstract: This paper introduces Bayesian analysis and demonstrates its application to parameter estimation of the logistic regression via Markov Chain Monte Carlo (MCMC) algorithm. The Bayesian logistic regression estimation is compared with the classical logistic regression. Both the classical logistic regression and the Bayesian logistic regression suggest that higher per capita income is associated with free trade of countries. The results also show a reduction of standard errors associated with the coefficients obtained from the Bayesian analysis, thus bringing greater stability to the coefficients. It is concluded that Bayesian Markov Chain Monte Carlo algorithm offers an alternative framework for estimating the logistic regression model.

Keywords: *Logistic regression, Posterior Distribution, Markov Chain Monte Carlo, Openness of Trade, Bayesian Analysis*

1. Introduction

In applied econometrics research, non linear models are essential tools for analysing empirical data. They are used when one has discrete or non linear response. One such important model is the logistic regression model which is used to explore the effect of some covariates, discrete and/ or continuous independent variables on a discrete response. The application of the logistic regression to binary response data is simple to understand, easy to compute and widely used. This classical approach fits the logistic regression by means of an iterative procedure such as the maximum likelihood, and inferences about the model are based on asymptotic theory. In some situations, due to the assumptions of iterative procedures, there may be failure in estimation convergence. Furthermore, the maximum likelihood estimation has significant bias for small samples. These limitations in the maximum likelihood estimations can be addressed by the use of Bayesian logistic regression as an alternative approach. The Bayesian estimation is flexible and does not require compliance with demanding assumptions as suggested in the maximum likelihood estimation or as in classical techniques. The flexibility of the Bayesian methodology is further enhanced by the use of the Markov Chain Monte Carlo (MCMC) based sampling methods. Progress in Markov Chain Monte Carlo (MCMC) methods has made it possible to fit various non linear regression models. Irrespective of these developments, few studies have employed the MCMC based approach to model the logistic regression. The limited application of the MCMC based approach is due to the fact that very little is understood about the concept of Bayesian analysis and its application to the logistic regression via MCMC algorithms. Although some studies have applied the Bayesian logistic regression in other fields, no empirical research has explored the application of the Bayesian logistic regression and compared it to the classical logistic regression using economic data. This study fills the gap by investigating the simple relationship between openness of trade and per capita income using classical and Bayesian logistic regression. The aim of this study is therefore to introduce Bayesian analysis and demonstrates its application to parameter estimation of the logistic regression via Markov Chain Monte Carlo (MCMC) algorithm. Fundamentally, this study presents a comparison of the Bayesian logistic regression with the classical logistic regression.

2. Literature Review

Numerous studies have applied the binary logistic regression model to study and analyse the effects of covariates on a categorical response. For example, Acquah (2013) applied the logistic regression model to investigate the relationship between openness of a country to trade and its per capita income. Conclusively, Acquah (2013) finds that higher per capita income is associated with free trade. Acquah (2011) also

investigated farmers' willingness to pay for climate change policy using logistic regression model. The logistic regression estimation finds age, years of farming experience, farm land owner, farm size and other income generating activity as significant predictors of the probability to pay for climate change policy. Han, Yang, Wang and Xu (2010) estimated publics' willingness to pay (WTP) for environment conservation and analyzed factors influencing WTP at Kanas Nature Reserve, Xinjiang, China. Logistic regression analysis was employed to compare the characteristics of those who were and were not willing to pay. Chi-square tests were administered to identify the relationships between various explanatory factors. In effect, logistic regression models have played important role in various studies. Due to the non-linearity of the logistic model, inference is made by maximum likelihood. But the maximum likelihood estimation has limitations which can be resolved by adoption of the more flexible Bayesian approach. Subsequently, some studies have estimated the Bayesian logistic regression and compared it with the classical logistic regression in other fields. For example, Mila and Michailides (2006) investigated prediction of panicle and shoot blight severity of Pistachio in California using Bayesian and classical logistic regression. They noted that the Bayesian methods gave more consistent results when used to update parameter estimates with new information and yielded predictions not statistically different from observed disease severity in more cases than the non-Bayesian analysis. Gordovil, Guardia, Pero & Fuente (2010) presented Bayesian estimation as an alternative to classical procedures in logistic regression estimation in the study of Attention Deficit Hyperactivity Disorder (ADHD) in a Mexican sample. An important data from their comparison of the classical and Bayesian estimation is the lower standard errors of the estimated coefficients in the Bayesian logistic regression. They note that this decrease is related to high coefficient's values stability. Departing from previous study, I apply the Bayesian and classical logistic regression methodology to economic data.

3. Methodology

The methodology describes Bayesian inference with emphasis on the prior distribution, likelihood function and posterior distribution for the Bayesian logistic regression. The Markov Chain Monte Carlo algorithm is also presented with emphasis on the Metropolis Hastings algorithm. Data used in the study is also described.

Bayesian Logistic Regression: In the Bayesian framework, there are three key components associated with parameter estimation: the prior distribution, the likelihood function, and the posterior distribution. These three components are formally combined by Bayes' rule as:

Posterior distribution = Prior distribution x likelihood function (1)

In simple terms, equation 1 states that the information contained in the sample (reflected in the likelihood function) is combined with information from other sources (summarized by the prior distribution) to obtain the posterior distribution. The posterior distribution contains all the available knowledge about the parameters in the model. Gill (2002) details the discussion on the concept of the Bayesian analysis.

Likelihood Function: The likelihood function used by Bayesians draws from frequentist inference. Given the probability of success (which in logistic regression varies from one subject to another, depending on their covariates), the likelihood contribution from the i^{th} subject is binomial:

$$likelihood_i = \pi(x_i)^{y_i}(1-\pi(x_i))^{(1-y_i)} \quad (2)$$

Where $\pi(x_i)$ represents the probability of the event for subject i who has covariate vector x_i, and y_i indicates the presence, $y_i = 1$, or absence $y = 0$ of the event for that subject. From the classical logistic regression, $\pi(x_i)$ is given by:

$$\pi(x) = \frac{e^{\beta_0+\beta_1 X_1+\ldots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\ldots+\beta_p X_p}} \quad (3)$$

In effect, the likelihood contribution from the i^{th} subject is

$$likelihood_i = \left(\frac{e^{\beta_0+\beta_1 X_{i1}+\ldots+\beta_p X_{ip}}}{1+e^{\beta_0+\beta_1 X_{i1}+\ldots+\beta_p X_{ip}}}\right)^{y_i}\left(1-\frac{e^{\beta_0+\beta_1 X_{i1}+\ldots+\beta_p X_{ip}}}{1+e^{\beta_0+\beta_1 X_{i1}+\ldots+\beta_p X_{ip}}}\right)^{(1-y_i)} \quad (4)$$

Given that individual subjects are assumed independent from each other, the likelihood function over a data set of n subjects is then

$$likelihood = \prod_{i=1}^{n}\left[\left(\frac{e^{\beta_0+\beta_1 X_{i1}+\ldots+\beta_p X_{ip}}}{1+e^{\beta_0+\beta_1 X_{i1}+\ldots+\beta_p X_{ip}}}\right)^{y_i}\left(1-\frac{e^{\beta_0+\beta_1 X_{i1}+\ldots+\beta_p X_{ip}}}{1+e^{\beta_0+\beta_1 X_{i1}+\ldots+\beta_p X_{ip}}}\right)^{(1-y_i)}\right] \quad (5)$$

Prior Distribution: The set of unknown parameters are made up of $\beta_0, \beta_1, \ldots, \beta_p$. Two types of prior distribution namely, informative and non-informative prior distributions can be used. Informative prior distributions are applied if something is known about the likely values of the unknown parameters. On the other hand, non-informative priors are employed if either little is known about the coefficient values or if one wishes to ensure that prior information plays very little role in the analysis. That is, the data is allowed to remain influential in the analysis. For the purpose of this study, we assume a multivariate Normal prior on β.

$$\beta_j \sim N(\mu_j, \sigma_j^2) \quad (6)$$

The most common choice for μ is zero, and σ is usually chosen to be large enough to be considered as non-informative. In this case σ is set to 1000.

Posterior Distribution via Bayes Theorem: The posterior distribution is derived by multiplying the prior distribution over all parameters by the full likelihood function, so that the posterior is given by:

$$posterior = \prod_{i=1}^{n}\left[\left(\frac{e^{\beta_0+\beta_1 X_{i1}+\ldots+\beta_p X_{ip}}}{1+e^{\beta_0+\beta_1 X_{i1}+\ldots+\beta_p X_{ip}}}\right)^{y_i}\left(1-\frac{e^{\beta_0+\beta_1 X_{i1}+\ldots+\beta_p X_{ip}}}{1+e^{\beta_0+\beta_1 X_{i1}+\ldots+\beta_p X_{ip}}}\right)^{(1-y_i)}\right]$$
$$\times\prod_{j=0}^{p}\frac{1}{\sqrt{2\pi\sigma_j}}\exp\left\{-\frac{1}{2}\left(\frac{\beta_j-\mu_j}{\sigma_j}\right)^2\right\} \quad (7)$$

The latter part of the above expression can be recognized as normal distribution for the β parameters. The above expression has no closed form expression. In this context, the Metropolis sampler is used to solve and approximate the properties of the marginal posterior distributions for each parameter. In effect, estimation of the posterior distributions of the parameters of the Bayesian logistic regression was carried out using a random walk Metropolis algorithm.

Metropolis-Hastings Algorithm: Metropolis-Hasting algorithm is an iterative algorithm that produces a Markov chain and permits empirical estimation of posterior distributions. The Metropolis-Hastings algorithm (MH) generates samples from a probability distribution using full joint density function. A basic MH algorithm consists of the following steps:

1. Establish starting values S for the parameter: $\theta^{j=0} = S$. Set $j=1$.

The starting values can be obtained via maximum likelihood estimation.

2. Draw a ´´candidate`` parameter, $\theta^c$ from a ´´proposal density,`` $\alpha(.)$.

The simulated value is considered a ´´candidate´´ because it is not automatically accepted as a draw from the distribution of interest. It must be evaluated for acceptance.

3. Compute the ratio $R = \frac{f(\theta^c)\alpha(\theta^{j-1} \mid \theta^c)}{f(\theta^{j-1})\alpha(\theta^c \mid \theta^{j-1})}$. (8)

4. Compare R with a $U(0,1)$ random draw u. If $R > u$, then set $\theta^j = \theta^c$. Otherwise, set $\theta^j = \theta^{j-1}$.

5. Set $j = j+1$ and return to step 2 until enough draws are obtained.

A detail discussion on the Metropolis Algorithm is presented in Gill (2002).

Data: International data for 1992 on the openness of trade and GDP per Capita for 20 countries was obtained from the World Bank Development Indicators. The dependent variable (openness of trade) takes the value of one for free trade and 0 otherwise whilst the independent variable of study is the GDP per Capita.

4. Results and Discussion

A Bayesian logistic regression analysis was employed to analyze the openness of a country (Y) and its per capita income in dollars (X).

Table1: Classical Logit and Posterior Moments

	Classical Logit		**Posterior**	
Variable	Mean	Std. Error	Mean	Std. Error
Intercept	-3.605	1.6800	-3.508	1.6080
GDP per capita	0.002	0.0009	0.002	0.0005

The model specification with openness of trade as the dependent variable and per capita income as the covariate was estimated for both the Bayesian and classical logistic regression. The model estimation result reveals a positive relationship between openness of trade and countries per capita income. In effect, both the classical logistic regression and the Bayesian logistic regression suggest that higher per capita income is associated with free trade of countries. The posterior moments in the Bayesian logistic estimation was obtained after a burn in period of 50,000 iterations and a follow up period of 250,000, storing every 20th iteration. Using the posterior mean as a point estimate, Table 1 compares the ordinary least squares estimates with the MCMC output. The estimated means and standard errors appear quite close with minimum difference between the classical logit estimate and MCMC output or posterior summary. Noticeably, the results show a reduction of standard errors associated with the coefficients obtained from the Bayesian analysis, thus bringing greater stability to the coefficients. Similarly, in a comparison of the classical and Bayesian estimation, Gordovil-Merino, Guardia-Olmos, Pero-Cebollero and Fuente-Solanas (2010) find lower standard errors of the estimated coefficients in the Bayesian logistic regression. They observed that this decrease is related to high coefficient's values stability. The posterior distributions of the per capita income and its corresponding quantiles given in Table 2 indicates that this parameter is mostly around 0.002 with a 2.5% probability taking a value below 0.0005 or a value above 0.0024. Graphically, all the mass of the posterior distributions of the per capita income are in the positive as illustrated in the plots of their posterior distributions in figure 1 in appendix I. These observations lead to the conclusion that higher per capita income is associated with free trade of countries.

Table 2: Posterior Distribution Summaries of parameters from MCMC Logistic regression

Variables	**Posterior Means**	**Standard Error**	**Quantiles 2.5%**	**Of Posterior 25%**	**Distributions 75%**	**97.5%**
Intercept	-3.508	1.6080	-5.787	-4.303	-2.705	-1.313
GDP per capita	0.002	0.0005	0.0006	0.0013	0.0021	0.0024

5. Conclusion

The classical estimation of the logistic regression model has some important limitations which can be resolved with possible alternative methods. The goal of this study was therefore to introduce Bayesian analysis as an alternative approach and demonstrate its application to parameter estimation of the logistic regression models for comparative analysis with the classical estimation. This study finds that the Bayesian Markov Chain Monte Carlo algorithm offers an alternative framework for estimating the logistic regression model. Both the classical logistic regression and the Bayesian logistic regression suggest that higher per capita income is associated with free trade of countries. A comparison of the classical and Bayesian approach to modelling the logistic regression reveals lower standard errors of the estimated coefficients in the Bayesian approach for the logistic regression model. Thus the Bayesian logistic regression is more stable. Importantly, the alternative methods lead to similar conclusions. Fundamentally, this study has demonstrated the application of the Bayesian MCMC algorithm to logistic regression estimation. Future research will investigate the Bayesian estimation of the multinomial logistic regression.

References

Acquah, H. D. (2013). An Introduction to Quantitative Methods. Shaker Verlag.

Acquah, H. D. (2011). Farmers Perception and Adaptation to Climate Change: A Willingness to pay analysis. *Journal of Sustainable Development in Africa*, 13(5), 150-161.

Gill, J. (2002). Bayesian Methods: A Social and Behavioral Science Approach. Boca Raton: Chapman and Hall/CRC.

Gordovil, M., Guardia, O., Pero, C. & Fuente, S. (2010). Classical and Bayesian Estimation in the Logistic Regression Model Applied To Diagnosis of Child Attention Deficit Hyperactivity Disorder. *Psychological Reports*, 106 (2), 1-15.

Han, F., Yang, Z., Wang, H. & Xu, X. (2010). Estimating Willingness to Pay for Environment Conservation: A Contingent Valuation Study of Kanas Nature Reserve, Xinjiang, China. *Environmental Monitoring Assessment*, 180(1-4), 451- 459.

Mila, A. L. & Michailides, T. J. (2006). Use of Bayesian Methods to Improve Prediction of Panicle and Shoot Blight Severity of Pistachio in Califonia. *Phytopathology*, 96, 1142-1147.

Appendix