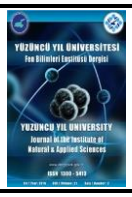
Yüzüncü Yıl Üniversitesi
Fen Bilimleri Enstitüsü Dergisi

http://dergipark.gov.tr/yyufbed



Research Article

A Bayesian Approach to Binary Logistic Regression Model with Application to OECD Data

Asuman YILMAZ*[1], Eray ÇELİK[1]

[1]Van Yüzüncü Yıl University, Faculty of Economics and Administrative Sciences, Department of Econometrics, 65080, Van, TURKEY

Asuman YILMAZ, ORCID No:000-0002-8653-6900, Eray ÇELİK, ORCID No0000-0001-7490-8124,
*Sorumlu yazar e-posta: asumanduva@yyu.edu.tr

Article Info

Recieved: 08.12.2020
Accepted: 07.07.2021
Published August 2021
DOI:10.53433/yyufbed.837533

Keywords
Binary-Logistic regression,
Maximum likelihood,
Bayesian method

Abstract: In spite of being a common method for estimating the model parameters, Maximum Likelihood (ML) method may give bias results for small sample sizes. To overcome this problem, Bayesian method is usually utilized to obtain the estimates of the model parameters as an alternative to the ML method. In this study, a real data set was analyzed by using the binary logistic regression model. Parameters of the binary logistic regression model were estimated by using ML and Bayesian methods. Modeling performance of the binary logistics regression model based on the Bayesian estimates was compared with the model based on the ML estimates. Well-known information criteria such as AIC and BIC were used in this comparison.

İki Durumlu Lojistik Regresyon Modeline Bayesci Bir Yaklaşım: OECD Örneği

Makale Bilgileri

Geliş: 08.12.2020
Kabul: 07.07.2021
Yayınlanma Ağustos 2021
DOI:10.53433/yyufbed.837533

AnahtarKelimeler
İki durumlu lojistik regresyon,
En çok olabilirlik yöntemi,
Bayesci Metot

Öz: En çok olabilirlik metodu model parametrelerini tahmin etmek için yaygın bir yöntem olmasına rağmen, küçük örneklem büyüklükleri için yanlı sonuçlar verebilir. Bu problemin üstesinden gelmek için, en çok olabilirlik yöntemine alternatif olarak model parametrelerinin tahmininde genellikle Bayes yöntemi kullanılmaktadır. Bu çalışmada, iki durumlu lojistik regresyon modeli kullanılarak gerçek bir veri seti analiz edilmiştir. İki durumlu lojistik regresyon modelinin parametreleri, en çok olabilirlik ve Bayesci yöntemler kullanılarak tahmin edilmiş, elde edilen sonuçlar Akaike bilgi kriteri (AIC) ve Bayesci bilgi kriteri (BIC) gibi kriterler kullanılarak karşılaştırılmıştır.

1. Introduction

Generalized linear models (GLM) are widely-used to define relationship between dependent and independent variables. The GLM is differed basing on the utilized function in defining the relationship between dependent and independent variables. For example, GLM becomes binary logistic regression model when dependent variable is binary and logarithmic function is utilized in defining the relationship between dependent and independent variables (Hair et al.,2006; Agresti & Hitchcock, 2005). In spite of being a common method for estimating the model parameters, Maximum Likelihood (ML) method may give bias results for small sample sizes. To overcome this problem, Bayesian method

usually considered in obtaining the estimates of the model parameters as an alternative to the ML method (Griffiths, 1973; Tektaş & Günay, 2008).

There exist many studies considering the Bayesian method in estimation procedure of the model parameters. For example, Albert & Chib, 1993 proposed a new algorithm by using the latent variables. Groenewald & Mokgatlhe, 2005 used a method suggested by Albert & Chib, 1993 to obtain a sample by using coefficients of posterior distribution through Gibbs sampling. Zelner & Rossi, 1984 considered numeric integration method and Monte Carlo integration method to obtain the posterior distribution of the model parameters in small and large sample sizes, respectively. Rashwan & El Dereny, 2012 used logistic regression model in analyzing the prostate cancer data in which Bayesian methods were used for obtaining the estimates of the model parameters. Ghosh &Mitra, 2017 investigated Bayesian logistic regression under different Cauchy prior distributions. Huggins et al., 2017 developed an efficient coreset construction algorithm for Bayesian logistic regression models. Spyroglou, et al., 2018 used Bayesian logistic regression method in analyzing the asthma persistence prediction. Dagliati, et al., 2017 proposed hierarchical Bayesian logistic regression to forecast metabolic control in type 2 DM patients. Lukman et al., 2021 used Bayesian logistic regression to analyze the hypothyroid prediction in post-radiation nasopharyngeal cancer patients. Suleiman et al., 2019 used Bayesian logistic regression approaches to predict incorrect DRG assignment.

In this study, Organization for Economic Cooperation and Development (OECD) data were analyzed via binary logistic regression model. Estimates of the model parameters were obtained using ML and Bayesian methods. Modeling performance of the model based on the Bayesian estimates was compared with the model based on the ML estimates. In this comparison, well-known information criteria Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) were used.

The rest of the paper as follows: in section 2, binary logistic regression model is briefly introduced. ML method and Bayesian method are given in section 3. In the application part, a real data set from OECD is analyzed by using the binary logistic regression model. Here, estimates of the model parameters are obtained via the ML and the Bayesian methods. Finally, section 5 is reserved for the conclusion.

2. Materials and Methods

In this section, information for the binary logistic regression, model the ML and the Bayesian methods are given briefly.

2.1. Binary logistic regression model

In binary logistic regression, dependent variable y follows a Bernoulli distribution since assumed that it takes only values 0 and 1. Here, the probability of occurrence of an event is denotes by $P\left(y_i=1\right)=\pi_i$ and probability of non-occurrence of an event is denotes by $P\left(y_i=0\right)=1-\pi_i$. If n observations are obtained for the dependent variable, i.e. $y_i\left(i=1,2,...n\right)$ binary logistic regression model can be expressed as follows:

$$\begin{aligned} & y_i \mid \pi_i: \ \ Bernoulli\left(\pi_i\right), \\ & \pi_i=P\left(y_i=1\right)=\frac{\exp(x_i^{'}\beta)}{1+\exp(x_i^{'}\beta)}, \ \ i=1,2,...,n, \end{aligned} \tag{1}$$

where $y_i=1$ if the interest response is observed for the *i-th* individual and $y_i=0$ otherwise. $\beta=\left[\beta_0\ \beta_1\ \beta_2...\beta_j\right]^{'}$ is the vector of an unknown model parameters and $x=\left[1\,x_{i1}...x_{ij}\right]^{'}$ is the vector of measurements of the *i-th* individual for the *j-th* independent variable.

2.2. ML method

The likelihood function (L) in binary logistic regression can be expressed as follows:

$$L\left(\underset{\vdots}{\beta}; \underset{\vdots}{y}, \underset{\vdots}{x}\right) = \prod_{i=1}^{n}\left[\frac{\exp\left(x_i'\beta\right)}{1+\exp\left(x_i'\beta\right)}\right]^{y_i}\left[\frac{1}{1+\exp\left(x_i'\beta\right)}\right]^{1-y_i}. \quad (2)$$

The log-likelihood equations are the first derivation of the logarithm of the likelihood function (log L) with the parameter of the interest as given below.

$$\frac{\partial \log L}{\partial \beta_j} = \sum_{i=1}^{n} x_{ij}\left(y_i - \frac{\exp\left(\sum_{k=0}^{K} x_{ik}\beta_k\right)}{1+\exp\left(\sum_{k=0}^{K} x_{ik}\beta_k\right)}\right) = 0, \quad j = 1,2,...,p. \quad (3)$$

The ML estimates of the model parameters are the simultaneous solutions of the log L equations given in Equation (3). Here, Newton-Raphson method is utilized to obtain the simultaneous solutions of these equations.

2.3. Bayesian method

ML methodology usually needs large sample size to obtain accurate estimates of the parameters. However, in some science fields such as medicine and agriculture small sample size is commonly encountered. Unlike the ML methodology, the Bayesian methodology does not need large sample size, i.e. it has not limitations regarding the size of sample. This is why the Bayesian methodology is an alternative for the ML methodology (Acquah, 2013; Tektaş & Günay 2008; Santos, 2009).
In the Bayesian methodology, there are three key parts in estimation procedure. These are (i) the prior distribution, (ii) the likelihood function, and (iii) posterior distribution. The posterior distribution is written as follows:

$$posterior\ \ distribution = \left(\ prior\ \ distribution\ \right)\left(likelihood\ \ \ function\right). \quad (4)$$

Here, prior distribution summarizes the information obtained from other sources. There are two types of prior distribution, namely, informative and non-informative prior distribution, (Acquah, 2013). In this study, we assume a normal prior on β_k.

$$\beta_k \sim N\left(0,10000\right),\ k = 1,2,...j. \quad (5)$$

The above expression is equivalent to non-informative priors of these parameters.

The likelihood function involves the information about the sample. The posterior distribution contains all the available knowledge for the model parameters. From (4), the posterior distribution is obtained by multiplying the prior distribution in (5) by the likelihood function in (3) as given below:

$$p\left(\beta; y, x\right) = \prod_{i=1}^{n}\left[\frac{\exp\left(x_i'\beta\right)}{1+\exp\left(x_i'\beta\right)}\right]^{y_i}\left[\frac{1}{1+\exp\left(x_i'\beta\right)}\right]^{1-y_i}\prod_{i=1}^{n}\frac{1}{\sigma_k\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{\beta_k-\mu_k}{\sigma_k}\right)^2\right]. \quad (6)$$

Equation (6) represents the posterior probability density function of the model parameters, and analytical solution for it cannot be obtained explicitly. The computational difficulties in obtaining the posterior distribution are disadvantage of the Bayesian method; however this problem can be solved by using the Markov Chain Monte Carlo (MCMC) simulation technique. The aim of the MCMC is to create a stationary Markov process to obtain the statistical inference for the posterior distribution. Therefore, the Markov Chain Monte Carlo (MCMC) simulation method is widely-used for getting statistical inference about the posterior distribution (Acquah, 2013).

Using the MCMC in complicated statistical problems brings some problems such as not converge to the desirable posterior distribution and determination of iteration size to obtain the stationary Markov process. The convergence to posterior distribution is necessary in the Bayesian methodology to obtain the accurate estimates of the model parameters. See (Geyer, 1992) for detailed information for the methods provides convergence in this context.

3. Results

In this section a real data set was analyzed using the binary logistic regression model. The estimates of the model parameters were obtained via the ML and the Bayesian methods. The data set includes various demographic and economic data from 34 countries which were member of the OECD. The data set was obtained from the official website of the OECD.
Following binary logistic regression model given in Equation (7) is used to analyze the data set:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \varepsilon,\ i = 1, 2, \ldots 34 \qquad (7)$$

Here, dependent variable is European Union (EU) membership (member:1, not member:0) and independent variables $(x_1, x_2, ..., x_8)$ are total number of people living, ratio of female parliamentarians, participation level of women between the ages 15 and 64 to labor force, ratio of imports, ratio of exports, life satisfaction or satisfaction with life, share allocated for health expenditure from the gross domestic product, average number of children per women between the ages 15 and 49 in each OECD country in 2013, respectively.

The estimates of the model parameters, given in Equation (7), were obtained by using the ML and the Bayesian methods and the analyses are conducted through R 3.0.3 software program. The ML estimates of the model parameters, standard errors (SEs), test statistics (z), and significant values (*p)* with corresponding to parameter estimates are given in Table 1.

Table 1. The statistics obtained with ML estimators

Variables	$\hat{\beta}_{ML}$	Exp($\hat{\beta}_{ML}$)	SE	z	*p*
Intercept	10.0493	23.139	7.57530	1.327	0.088
x_1	0.38593	1.470	0.23629	1.633	0.066
x_2	-0.06910	0.933	0.15646	-0.442	0.687
x_3	0.08177	1.085	0.08394	0.974	0.413
x_4	0.20735	1.230	0.16268	1.275	0.290
x_5	-5.69552	0.003	3.24655	-1.754	0.031
x_6	1.55750	4.746	1.05062	1.482	0.405
x_7	0.04387	1.044	0.02705	1.622	0.140
x_8	-0.80534	0.446	2.87410	-0.280	0.644

It can be seen from Table 1 that all parameter estimates for the independent variables, except x_5, are not statistically significant, since p values are greater than 0.05.

The Bayesian estimates of the model parameters along with the standard deviations (SDs), Monte Carlo simulation errors (MC Error), and confidence interval (CI) between (%2.5-%97.5) of the parameter estimates are given in Table 2.

Table 2. The statistics obtained with Bayesian estimators

Variables	Mean	Exp(Mean)	SD	MC Error	CI
Intercept	11.90223	147.72	0.72646	0.02462	(-3.11615, 29.9593)
x_1	0.05229	1.053	0.00162	0.00067	(0.01377, 0.09980)
x_2	0.41872	1.520	0.15591	0.00421	(0.11763,0.78240)
x_3	0.05399	1.055	0.16803	0.00517	(-0.25042, 0.41285)
x_4	0.15183	1.163	0.09890	0.00257	(-0.00556, 0.37750)
x_5	0.16783	1.182	0.09245	0.00220	(-0.01837, 0.33160)
x_6	-6.84683	0.001	2.60164	0.06542	(-12.94350, -2.45507)
x_7	1.42337	4.151	0.03344	0.01588	(0.09428, 2.68094)
x_8	0.42022	1.522	3.54912	0.09246	(-0.89632, 5.20511)

It can be seen from Table 2 that parameter estimates for the independent x_1, x_2, x_6 and x_7 are statistically significant at 0.05 significance level, i.e. CI between (%2.5-%97.5) of these parameter estimates not include the value 0.

According to Table 2, the ratio of the population of the OECD countries which are the member of EU to the non-member countries is 1.05. The ratio of women's participation in parliament in the OECD countries which are the member of EU to the non-member countries is 1.52. The ratio of life satisfaction of those who live in the OECD countries which are the member of EU to the non-member countries is 0.001. The ratio of health expenditures of the OECD countries which are member of EU to non-member countries is 4.52.

Burn-in procedure is applied to the first 1000 iterations in the Markov Chain for providing convergence to the posterior distribution. Autocorrelation plots, trace plots and Geweke convergence test results are also utilized to guaranteed convergence for the posterior distribution; see, Figure 1, Figure 2 and Table 3, respectively.

Autocorrelations given in Figure 1 measure the dependence between each sample value in the Markov Chain. Low correlation means that the convergence has been achieved; see (Cowles & Carlin, 1996).

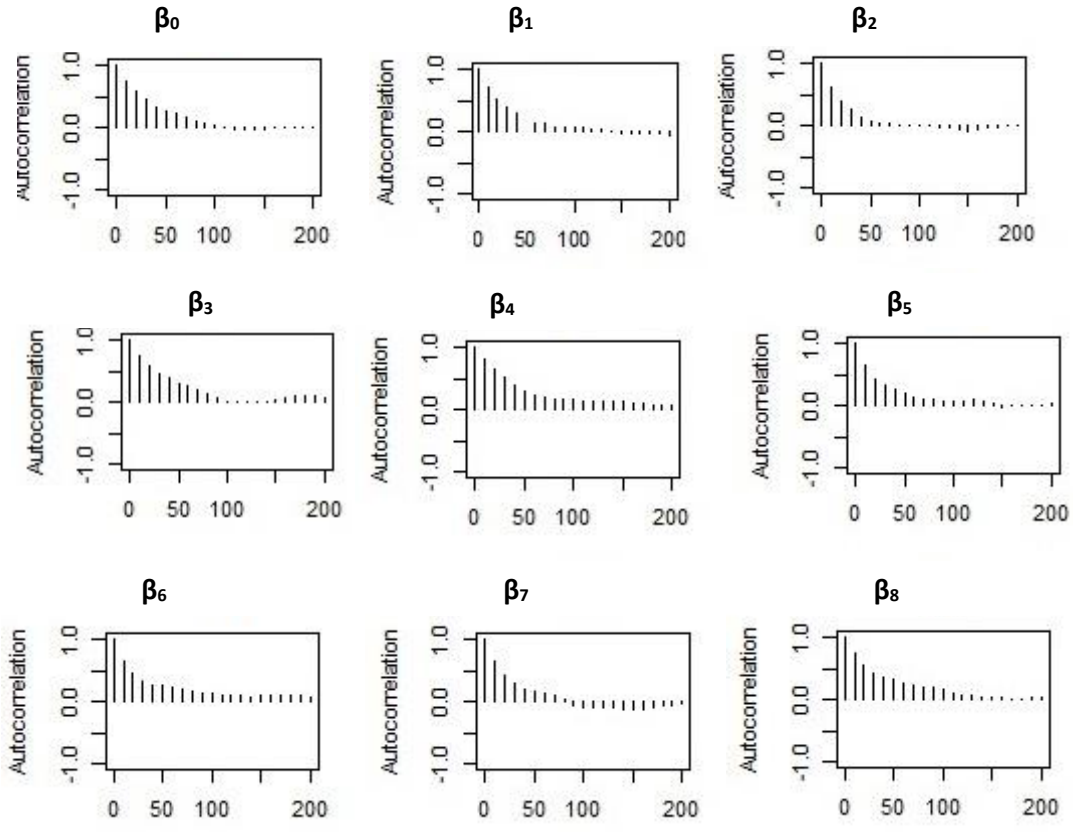


Figure 1. Autocorrelation plots within each parameter logit model

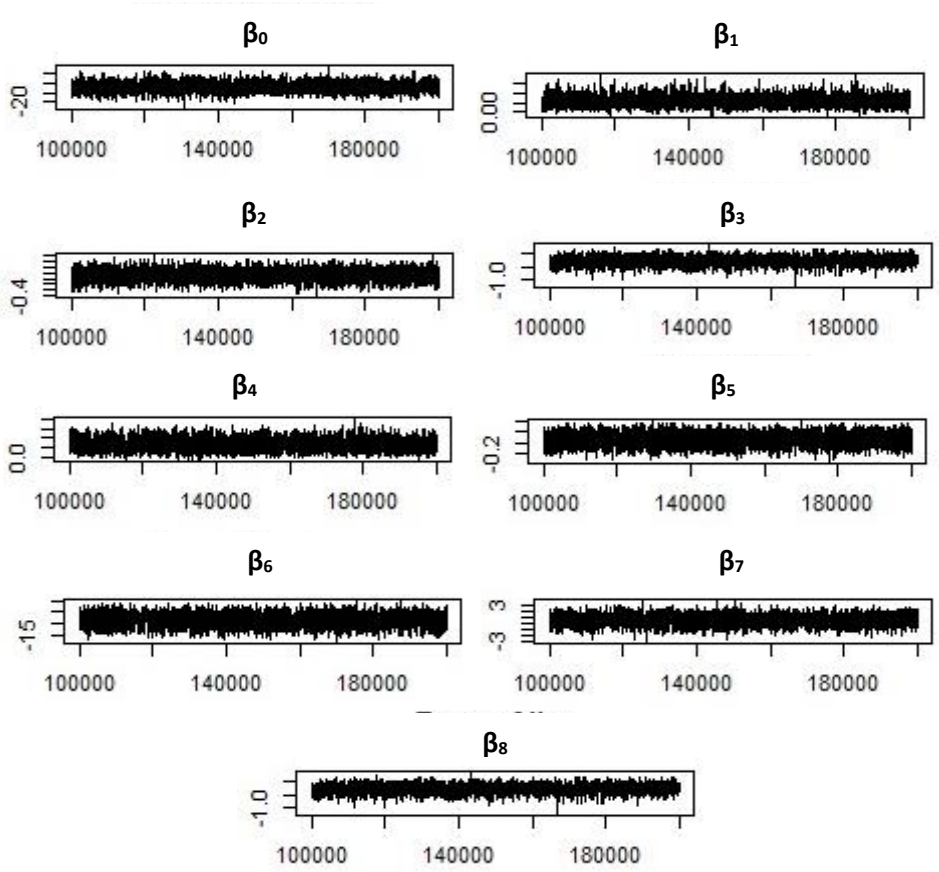


Figure 2. The traceplots within each parameter logit model

From Figure 2, it also can be said that convergence has been achieved for each parameter of the corresponding variables.

Table 3. Geweke convergence test results

Variables:	Intercept	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
z statistics:	0.693	-1.340	1.019	-1.755	0.317	0.362	0.019	-0.071	-0.990

In this study all tests were conducted under %95 confident intervals that mean critical values of test statistics is ± 1.96. It can be seen from Table 3 that Geweke test statistics for all parameters with corresponding variable is between ± 1.96 which shows the convergence has been achieved for each parameter.

4. Discussion and Conclusion

In this study, a short literature review for Bayesian logistic regression is presented. Also, the estimates of the model parameters are derived by using maximum likelihood estimation and Bayesian estimation methods. Moreover, the real data set is analyzed at for better understanding of the methods presented. This data set taken from the OECD is modeled by using the binary logistic regression model. In the estimation procedure of the model parameters, the ML and the Bayesian methods are used. In Bayesian methods, burn-in procedure is applied to the first 1000 iterations in the Markov Chain for providing convergence to the posterior distribution. Also, autocorrelation plots, trace plots and Geweke convergence test results are utilized to guaranteed convergence for the posterior distribution. It can be seen from all convergence tests that convergence was provided for each parameter. Then, the modeling performances of these two models are compared by using the well-known information criteria such as AIC and BIC given in Table (4). Also, Mc Fadden R^2 and correct classification ratio values for each model are given in Table (4). It should be noticed that smaller values of the Mc Fadden R^2, AIC and BIC are mean better fitting.

Table 4. Goodness of fit results

Criteria	Correct Classification Ratio	Mc Fadden R^2	AIC	BIC
Model based on ML estimates	89.70	0.66	38.44	51.91
Model based on Bayes estimates	94.20	0.53	32.74	46.21

According to Table (4), it can be concluded that binary logistic regression model based on the Bayesian estimates has higher correct classification ratio and smaller Mc Fadden R^2, AIC, and BIC values than the based on the ML estimates. The differences between ML and Bayesian estimates are occurred by the small sample size. Results show that for small sample size, as similar in application of this study, the Bayesian method shows better performance than the ML method based on the goodness of fit statistics given in Table (4). In this regard, it is seen that the Bayesian method is more preferable than the ML method for this data set.

References

Acquah, H. D. (2013). Bayesian logistic regression modelling via Markov chain Monte Carlo algorithm. *Journal of Social and Development Sciences*, 4, 193-197. doi: 10.22610/jsds.v4i4.751

Agresti, A., & Hitchcock, D. B. (2005). Bayesian inference for categorical data analysis. *Statistical Methodsand Applications*, *14*(3), 297-330. doi:10.1007/s10260-005-0121-y

Albert, J. H., & Chib. S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88, 669-679. doi:10.2307/2290350

Cowles, M. K., & Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association*, *91*, 883-904.

Dagliati, A., Malovini, A., Decata, P., Cogni, G., Teliti, M., Sacchi, L., & Bellazzi, R. (2016). Hierarchical Bayesian Logistic Regression to forecast metabolic control in type 2 DM patients. In *AMIA Annual Symposium Proceedings*,2016, 470-479.

Dos Santos, M. A., Moala, F. A., & Tachibana, V. M. (2009). Approximate Bayesian methods for logistic regression model. *Revista Brasileira de Biometria*, 27, 288-300.

Geyer, C. J. (1992). Practical markov chain montecarlo. *Statistical Science*, 10, 473-483.

Ghosh, J., Li, Y., & Mitra, R. (2018). On the use of Cauchy prior distributions for Bayesian logistic regression. *Bayesian Analysis*, *13*, 359-383. doi:10.1214/17-BA1051

Griffiths, D. A. (1973). Maximum likelihood estimation for the beta-binomial distribution and an application to the household distribution of the total number of cases of a disease, *Biometrics, 7*,637-648.

Groenewald, P. C., & Mokgatlhe, L. (2005). Bayesian computation for logistic regression. *Computational Statistics & Data Analysis*, 48, 857-868. doi:10.1016/j.csda.2004.04.009

Hair, F. T., William, C. B., Babin, B. T., & Anderson E. R. (2006). *Overview of Multivariate Methods*. Oxford, UK: Wiley & Sons.

Huggins, J. H., Campbell, T., & Broderick, T. (2016). Coresets for scalable bayesian logistic regression. *arXiv preprint arXiv:1605.06423*.

Lukman, P. A., Abdullah, S., & Rachman, A. (2021). Bayesian logistic regression and its application for hypothyroid prediction in post-radiation nasopharyngeal cancer patients. *In Journal of Physics: Conference Series*, 1725(1), 012010. doi:10.1088/1742-6596/1725/1/012010

Rashwan, N. I., & El Dereny, M. (2012). The comparison between result of application Bayesian and maximum likelihood approaches on logistic regression model for prostate cancer data. *Applied Mathematical Scienses*, 6, 1143-1158.

Suleiman, M., Demirhan, H., Boyd, L., Girosi, F., & Aksakalli, V. (2019). Bayesian logistic regression approaches to predict incorrect DRG assignment. *Health care management science*, *22*(2), 364-375. doi: 10.1007/s10729-018-9444-8.

Spyroglou, I. I., Spöck, G., Chatzimichail, E. A., Rigas, A., & Paraskakis, E. (2018). A Bayesian logistic regression approach in asthma persistence prediction. *Epidemiology, Biostatistics and Public Health*, *15*(1). doi:10.2427/12777.

Tektaş, D., & Günay, S. (2008). Bayesian approach to parameter estimation in binary logit models. *Hacettepe Journal of Mathematics and Statistics*, 37, 167-176.

Zellner, A., & Rossi, P.E. (1984). Bayesian analysis of dichotomous quantal response models. *Journal of Econometrics*, 25, 365-393.