**PAPER • OPEN ACCESS**

# Bayesian Computational Methods of the Logistic Regression Model

View the article online for updates and enhancements.

# Bayesian Computational Methods of the Logistic Regression Model

**Najla. A .AL-Khairullah[1], Tasnim H.K. Al-Baldawi[1].**

[1]Dept.of Math. College of Science, University of Baghdad .

nalkhairalla5@gmail.com

tasnim@csbaghdad.edu.iq

**ABSTRACT.** In this paper, we will discuss the performance of Bayesian computational approaches for estimating the parameters of a Logistic Regression model. Markov Chain Monte Carlo (MCMC) algorithms was the base estimation procedure. We present two algorithms: Random Walk Metropolis (RWM) and Hamiltonian Monte Carlo (HMC). We also applied these approaches to a real data set.

## 1. INTRODUCTION

MCMC methods are set of algorithms used in efficient counting, optimization, dynamic simulation, integral approximation, and sampling. These techniques are commonly used in problems relating to statistics, combinatorial, physics, Chemistry, probability, optimization, numerical analysis. Because of its ease of implementation, in other cases fast convergences as well as numerical stability applied mathematicians and statisticians prefer MCMC methods. But, due to its complex theoretical basis, and shortage of theoretic convergence diagnostic. MCMC methods are sometimes referred to as black boxes for sampling and posterior estimation[1].

The application of Bayesian methods in applied problems expanded during in 1990s, the basic idea of Markov chain estimation is generating approximate samples from posterior distribution of interest by generating Markov chain whose stationary distribution of which is the desired posterior. Revolutionary approach for Monte Carlo was started in the particle Physics studies by Metropolis (1953), then Hastings (1970) generalized it via more statistical setting.[2]

Gibbs sampling and Metropolis-Hastings algorithms are classical approaches to implement MCMC algorithms. A special case of Metropolis-Hastings sampling is Gibbs sampling in which all the candidate value is often accepted. Gibbs sampling can only be used if the posterior distribution has conditional distribution that is standard distribution such as Dirichlet, Gaussian, or discrete distribution. Whereas the Metropolis-Hastings (MH) algorithm is also applied to a wide range of distributions and based on the candidate values being proposed sampled of the proposal distribution. These are either rejected or accepted according to the probability rule[3].

Another common algorithm is HMC that introduces momentum variable and employs leapfrog discretization of deterministic Hamiltonian dynamics as the proposal scheme in combination with momentum resampling [4]

Difference of MCMC approaches are designed by simulation irreversible Markov chains that converge to the target distribution. One group of algorithms includes s irreversible MALA [5]and non-

reversible parallel tempering[6], another group of algorithms is included the bouncy particle [7]and Zig-Zag samplers using Poisson jump processes[8].

We are focusing in this study on estimating parameter for the logistic regression model where the dependent variable be binary data, also we discussed RWM and HMC methods to determine the parameter estimation.

## 2. LOGISTIC REGRESSION MODEL

The name of logistic regression exists from that the function $e^x/(1 + e^x)$ is named logistic function.

Logistic regression model is the special case of statistical models named generalized linear models which also include linear regression, log-linear models, Poisson regression, etc.

Logistic regression model is very commonly used in the analyzing data including binary or binomial responses as well as several explanatory variables. This model offers a strong technique analogous to ANOVA of continuous responses and multiple regression[9].

Suppose that $y_i$ $(i = 1,2, ..., n)$ independent observations on a binary response variable y, then for k-dimensional covariate X, the model is defined as:

$$\pi_i \equiv P(Y_i = 1|X = x) = \frac{e^{\beta_0 + \Sigma_{j=1}^k \beta_j x_{ij}}}{1 + e^{\beta_0 + \Sigma_{j=1}^k \beta_j x_{ij}}}$$

It can simplify to

$$\pi_i = \frac{1}{1 + e^{-(\beta_0 + \Sigma_{j=1}^k \beta_j x_{ij})}}$$

The value of $\pi_i$ be a number between 0 and 1, also (1) equivalent:

$$logit(\pi_i) = \beta_0 + \sum_{j=1}^k \beta_j x_{ij}$$

where $logit(\pi) = log\left(\frac{\pi}{1-\pi}\right)$

## 3. BAYESIAN LOGISTIC REGRESSION

Suppose we have a normal prior distribution for all the parameters, which is often used as a prior distribution for logistic regression.

$$\beta_j \sim N(\mu_j, \sigma_j^2) \qquad j = 1,2, ..., k$$

The posterior distribution is:

$$P(\beta|y) = \prod_{i=1}^n (\pi_i)^{y_i} (1 - \pi_i)^{1-y_i} \times \prod_{j=1}^k \frac{1}{\sqrt{2\pi}\sigma_j} exp\left[-\frac{1}{2}\left(\frac{\beta_j - \mu_j}{\sigma_j}\right)^2\right]$$

It seems, this expression has no closed form, and the marginal distribution of each coefficient by integration is difficult to obtain. For logistic regression, the exact numerical solution is hard to obtain. In the statistics software, the most common method used for estimating parameters is the MCMC simulation, which provides an approximate solution.

## 4. Bayesian Computational Methods

In the following we will presented the computation Bayesian algorithms in the process of obtaining Bayesian parameter estimation in Logistic regression model:

Metropolis-Hastings Algorithm

Its important algorithm can provide a general approach to generating correlated sequence drawing from the target density which may be difficult to sample using the classical independent methods.

The basic Metropolis-Hastings algorithm can be given in the following: Generate the candidate state $x$ at step t from the proposal distribution $q(.|x_t')$, for the next state in the chain the candidate state is accepted or rejected with the probabilities given by

$$x_{t+1} = \begin{cases} x_t' & ,with\ probability \qquad , \alpha(x_t, x_t') \\ x_t & ,with\ probability \qquad , 1 - \alpha(x_t, x_t') \end{cases}$$

where $\alpha(x_t, x_t') = \min\left(\frac{q(x_t|x_t')\pi(x_t')}{q(x_t'|x_t)\pi(x_t)}, 1\right)$

**Algorithm 4.1: Metropolis-Hastings**

1   Set starting point $x_0$

2   **For** t=1, ..., n

3   Generate the candidate $x'_t \sim q(.\,|x'_t)$

4   Generate $u \sim U[0,1]$

5   Calculate $r = \frac{q(x_t|x'_t)\pi(x'_t)}{q(x'_t|x_t)\pi(x_t)}$

6   **if** $u < \min(r(x_t, x'_t), 1)$

7   **then** $x_t = x'_t$

8   **else**

9   $x_t = x_{t-1}$

10  **end**

11  **end**

Random Walk Metropolis Algorithm

RWM algorithm is indeed a special case of Metropolis algorithm, using asymmetric candidate transition, that is $q(x_t, x'_t) = q(x'_t, x_t)$,

assume $q(x_t, x'_t) = q(x'_t - x_t)$ then we have $\alpha(x_t, x'_t) = \min\left(\frac{\pi(x'_t)}{\pi(x_t)}, 1\right)$.

This means the candidate $x'_t$ which has a higher value of target distribution than target distribution of the current value $x_t$ always acceptable. In contrast, the candidate with lower target distribution value will only be accepted with a probability equal to the ratio of the target distribution value to the current distribution value. However, a chain with random-walk proposal distribution will generally have several accepted candidate points, but that most moves are a short distance, that is the accepted candidate point $x'_t$ will also be close to the previous current value $x_t$. So moving around the whole state space it could take a long time. [10]

Hamiltonian Monte Carlo algorithm

HMC or (Hybrid Monte Carlo) algorithm is a MCMC technique, which uses a combination of Metropolis Monte Carlo approach and advantages of Hamiltonian dynamics [6], [11]for sampling of complex distribution.

In view of the observed data $Y = (y_1, y_2, ..., y_N)^T$ , we are interested in sampling from the posterior distribution of the model parameters $x$.

$P(x, p) \propto \exp(-V(x))$

Where $V$ function of potential energy, defined by:

$V(x) = -\sum_{i=1}^{N} \log(y_i|x) - \log P(x)$

Negative log-likelihood in the first term, $P(x)$ is assumed prior on the parameters of model and it is almost always difficult to solve the posterior distribution.

HMC expose Hamiltonian dynamics system with the auxiliary momentum variables $p$ to propose samples of $x$ in the Metropolis framework which explores parameter space more efficiently than proposals of the standard random walk.

HMC generates the proposals jointly for $(x, p)$, by using the system of differential equations as follows:

$$\frac{dx_i}{dt} = \frac{\partial H}{\partial p_i}$$

$$\frac{dp_i}{dt} = -\frac{\partial H}{\partial x_i}$$

## **Algorithm: Hamiltonian Monte Carlo**

1   Initialize $x_0$, and $p_0$

2   Set $\varepsilon$
3   **For** i=1, 2, …, n  **do**
4   │   Draw $p \sim N(0, I)$
5   │   $\left(x^{(0)}, p^{(0)}\right) = (x_{i-1}, p)$
6   │   **For** j=1, 2, …, L  **do**
7   │   │   $p^{(j-\frac{1}{2})} = p^{(j-1)} - \frac{\varepsilon}{2}\nabla V(x^{(j-1)})$
8   │   │   $x^{(j)} = x^{(j-1)} + \varepsilon p^{(j-\frac{1}{2})}$
9   │   │   $p^{(j)} = p^{(j-\frac{1}{2})} - \frac{\varepsilon}{2}\nabla V(x^{(j)})$
10  │   **end**
11  │   $(x', p') = \left(x^{(L)}, p^{(L)}\right)$
12  │   Draw $u \sim U[0,1]$
13  │   $\rho = exp\left[H(x', p') - H(x^{(0)}, p^{(0)})\right]$
14  │   **If** $u < min\{1, \rho\}$  **then**
15  │   │   $\left(x_i, p_i\right) = (x', p')$
16  │   **else**
17  │   $\left(x_i, p_i\right) = \left(x_{i-1}, p_{i-1}\right)$
18  │   **end**
19  **end**
20  return  $\{x_i, p_i\}_{i=0}^{n}$

The Hamiltonian function defined by:

$H(x, p) = V(x) + K(p)$   where   $K(p) = \frac{1}{2}p^T M^{-1} p$ is quadratic kinetic energy function which corresponds to negative log-density of the zero-mean multivariate Gaussian distribution with covariance $M$, where $M$ is called the mass matrix, which is often set to the identity matrix, but it can be used with Fisher information to precondition the sampler [12]

## 5. RESULTS AND DISCUSSION

The dataset (Heart Disease UCI Dataset) in this study consists of 14 variables with 286 response as basis for the analysis. The dataset was published by. Ronit in Kaggle website (https://www.kaggle.com/ronitf/heartdisease-uci). The variables included are:

age       :   age (years)
sex       :   the gender (value 0: female ; value 1: male)
cp        :   type of chest pain experienced (value 1: typical angina; value 2: atypical angina; value 3: non-anginal pain; value 4: asymptomatic)
tresbps   :   trestbps-resting blood pressure at admission to hospital(mm Hg)
chol      :   chol-serum cholesterol variable(mg/dl)
fbs       :   blood sugar when fasting > 120 mg/dl  (value 0 : false; value 1: true)
restecg   :   resting electrocardiographic measurement (value 0: normal; value 1: having st-t wave abnormality; value 2: showing probable or definite left ventricular hypertrophy by estes' criteria)
thalac    :   the maximum thalach heart rate variable
exang     :   exacting-exercise variable induced angina (value 0 : no; value 1: yes)
oldpeak   :   depression caused by exercise relative to the rest

| | | |
|---|---|---|
| slope | : | the slope variables of the peak training segment ST (value 1: up sloping; value 2: flat; value 3: down sloping) |
| ca | : | ca-number of main vessels with values 0 to 3; colored by fluoroscopy |
| thal | : | a blood disorder called thalassemia (3 variable : normal; variable 6 : fixed defect; variable 7 : reversible defect) |
| target | : | heart disease (value 0: no ; value 1: yes) |

Package statistical software R is used for Bayesian analysis that provide a convenient environment for the simulation of MCMC, we simulate the posterior density of the logistic regression model by using RWM and HMC algorithms .First we derived the β posterior simulation of RWM algorithm . The summary of the sample is given as in the following:

**Table 1.**Summary statistic of Bayesian logistic regression via RWM algorithm.

| **Variable** | **Mean** | **SD** | **2.5%** | **97.5%** |
|---|---|---|---|---|
| (Intercept) | 2.456069 | 2.30783 | -1.801869 | 6.952633 |
| age | 0.005533 | 0.02358 | -0.041370 | 0.053395 |
| trestbps | -0.018597 | 0.01114 | -0.040872 | 0.001648 |
| Chol | -0.001471 | 0.00385 | -0.008749 | 0.006372 |
| thalach | 0.018731 | 0.00981 | -0.000462 | 0.039198 |
| oldpeak | -0.730859 | 0.22493 | -1.203184 | -0.288489 |
| sex | -0.794047 | 0.39281 | -1.584972 | -0.015668 |
| cp | 0.857338 | 0.19375 | 0.493769 | 1.267301 |
| fbs | -0.064067 | 0.53338 | -1.104246 | 0.941653 |
| restecg | 0.538495 | 0.36561 | -0.163436 | 1.236694 |
| exang | -0.996579 | 0.42465 | -1.836700 | -0.154026 |
| ca | -0.811531 | 0.19814 | -1.201666 | -0.422844 |
| thal | -1.020371 | 0.30485 | -1.586742 | -0.412668 |
| slop | 0.473484 | 0.37066 | -0.291787 | 1.185709 |

Table1 show the posterior summaries which consist of variables, mean, standard deviation and quantiles of posterior distribution, where significant variables could be determined at the 5% significance level. Values from 2.5% to 97.5% quantiles provide 95% credibility interval for every given variable. Parameters of oldpeak, cp, exang, ca and sex are -0.730859, 0.857338, -0.996579,-0.811531 and -0.794047 respectively at the 5% significance level they are significant. An increase in (oldpeak, cp, exang, ca and sex) units by one unit with all the other variables kept fixed means which the log-odds will increase by (-0.730859, 0.857338, -0.996579,-0.811531 and -0.794047) respectively by default, since credibility interval it does not contain zero. Other variables are insignificant. Trace plots of Markov chain and the density plots of posterior distributions are shown in Figures 1 and 2:
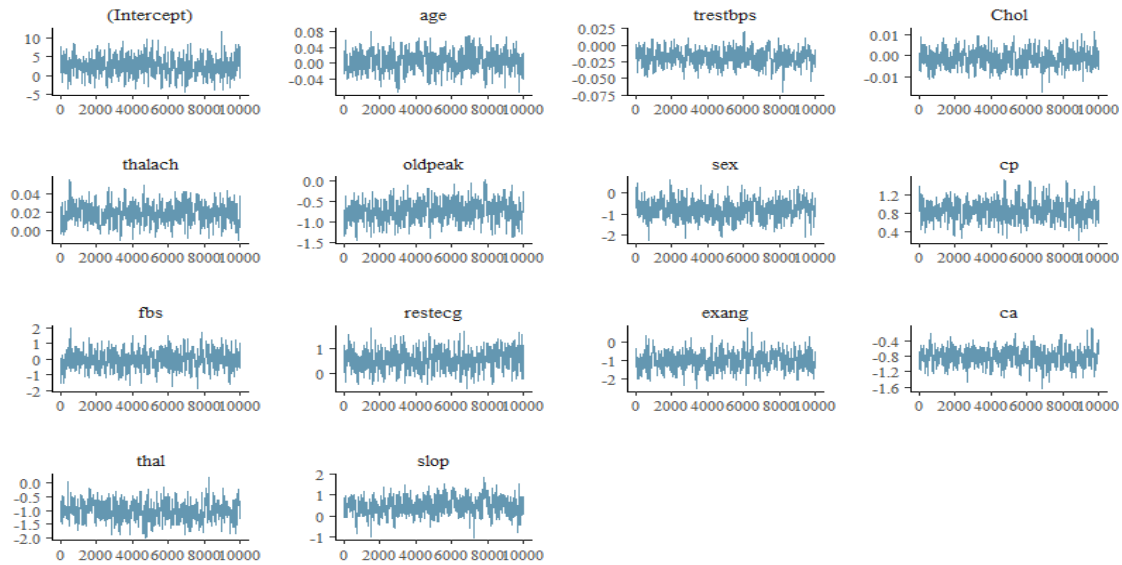
**Figure 1.** Trace plots of the corresponding posterior estimates of the Intercept, variables via RWM algorithm.
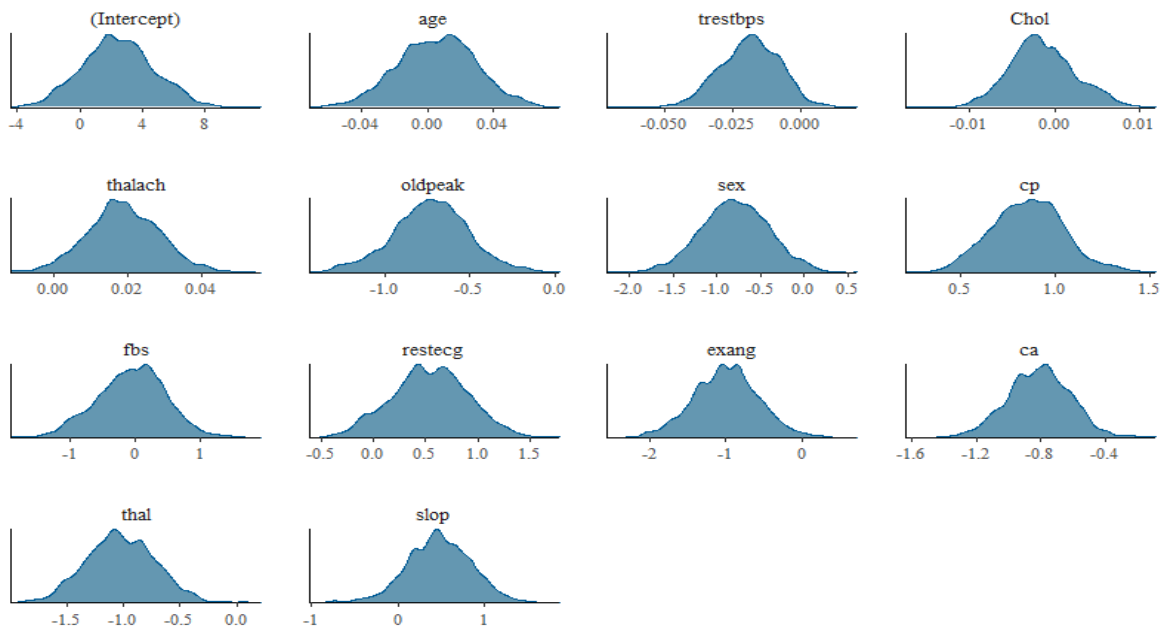


**Figure 2.** Density distribution of the corresponding posterior estimates of the Intercept, variables via RWM algorithm.

For variables, statistics of Geweke diagnostic are computed and shown in Table 2.

**Table 2.** Statistics of Geweke diagnostic for variables of the Bayesian logistic    regression model

| Variable | z |
|---|---|
| trestbps | 0.3949 |
| Chol | -0.0386 |
| thalach | 4.417 |
| oldpeak | -3.456 |
| sex | -0.8158 |
| cp | 5.669 |
| fbs | 0.1841 |
| restecg | 0.4687 |

| | |
|---|---|
| exang | -3.847 |
| ca | -5.713 |
| thal | -3.954 |
| slop | 2.682 |

Table 2 shows that variables thalach , oldpeak , cp , exang , ca, thal and slop have $|z| > 2$. Thus, these variables have not converged. All other variables have converged, according to Geweke diagnostic.

Then, we simulated the posterior density of the logistic regression model by using HMC algorithm. Five of independent variables are continuous, with wide range of the values. When tuning this model, the step size $\varepsilon$ of these types of variables is tuned separately for the specific application of the HMC. We summarize the results and plot the trace and density distribution of the simulated posteriors in the following table 3 and Figure 3:

**Table 3.** Posterior Quantiles of Bayesian logistic regression via HMC algorithm.

| Variable | 2.5% | 5% | 25% | 50% | 75% | 95% | 97.5% |
|---|---|---|---|---|---|---|---|
| (Intercept) | -1.0830 | -1.0294 | -0.8648 | -0.6150 | -0.3715 | -0.2785 | -0.2584 |
| age | -0.0181 | -0.0109 | -0.0014 | 0.0064 | 0.0126 | 0.0224 | 0.0252 |
| trestbps | -0.0266 | -0.0248 | -0.0190 | -0.0133 | -0.0091 | -0.0033 | -0.0013 |
| Chol | -0.0061 | -0.0055 | -0.0035 | -0.0019 | -0.0004 | 0.0019 | 0.0027 |
| thalach | 0.0150 | 0.0160 | 0.0208 | 0.0241 | 0.0275 | 0.0330 | 0.0342 |
| oldpeak | -0.0841 | -0.0789 | -0.0592 | -0.0418 | -0.0112 | -0.0009 | 0.0020 |
| sex | -0.8759 | -0.8461 | -0.7166 | -0.6203 | -0.3542 | -0.2142 | -0.1822 |
| cp | 0.4941 | 0.5238 | 0.6230 | 0.7024 | 0.7796 | 0.9032 | 0.9246 |
| fbs | -0.7678 | -0.7353 | -0.5804 | -0.4637 | -0.2970 | 0.0614 | 0.0761 |
| restecg | 0.0696 | 0.1113 | 0.2270 | 0.3800 | 0.4652 | 0.5950 | 0.6237 |
| exang | -1.2519 | -1.2229 | -1.1066 | -0.7512 | -0.4819 | -0.2059 | -0.1788 |
| ca | -0.9598 | -0.9374 | -0.8240 | -0.7125 | -0.5870 | -0.4707 | -0.4491 |
| thal | -1.0120 | -0.9874 | -0.8620 | -0.7866 | -0.7012 | -0.5149 | -0.4720 |
| slop | 0.4748 | 0.4997 | 0.7988 | 0.9336 | 1.0160 | 1.0923 | 1.1162 |
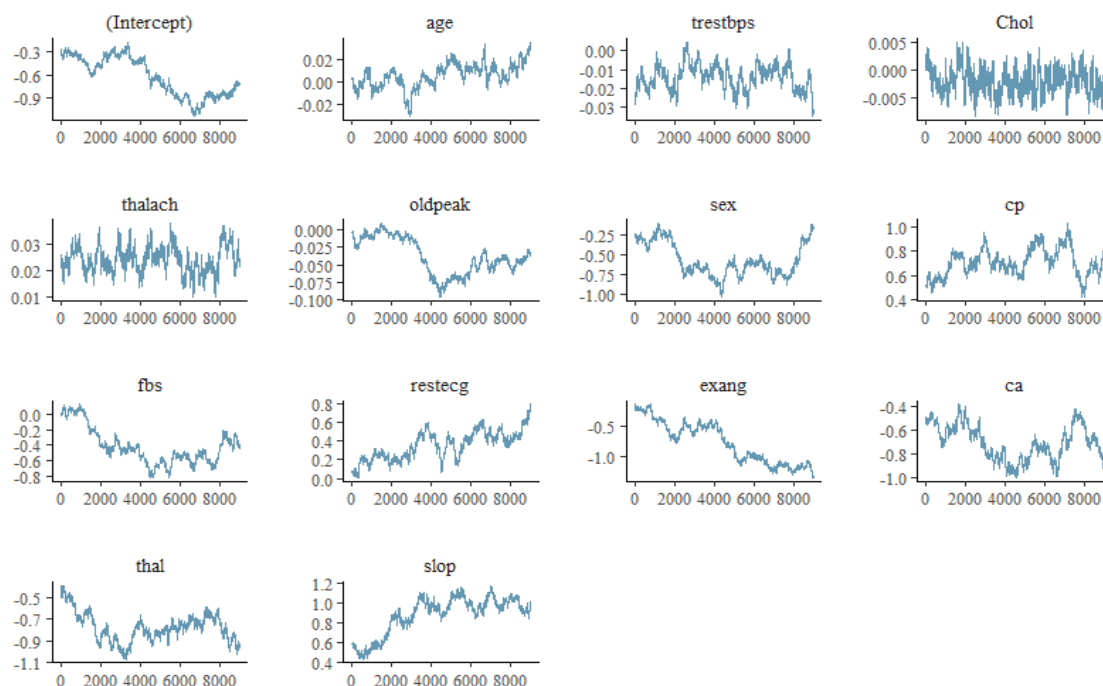


**Figure 3.** Trace plots of the corresponding posterior estimates of the Intercept, variables via HMC algorithm.
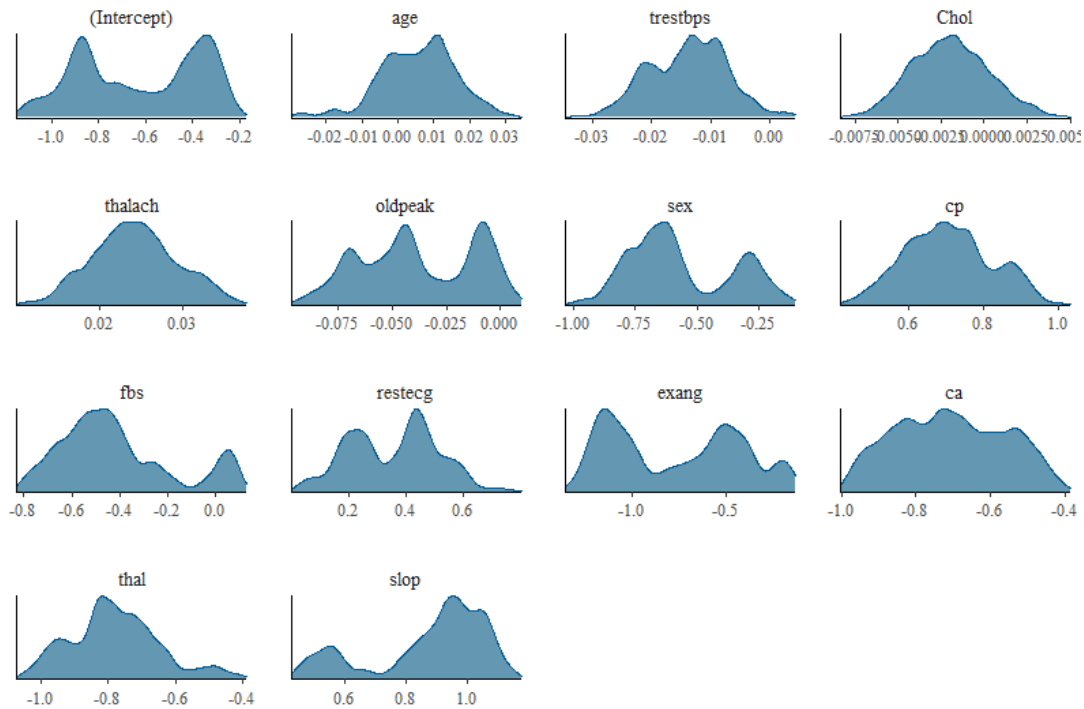
**Figure 4.** Density distribution of the corresponding posterior estimates of the Intercept, variables via HMC algorithm.

Statistics of Geweke diagnostic for variables of the Bayesian logistic regression model via HMC algorithm gives the same results where the variables thalach , oldpeak , cp , exang , ca, thal and slop have $|z| > 2$. Thus, these variables have not converged. All other variables have converged, according to Geweke diagnostic.

## 6. Conclusion

We investigated performance of two Bayesian computational methods for parameter estimation of the Logistic regression model, there is a single independent chain running for 10000 iterations in both analyses, convergence is monitored by Geweke diagnostic of samples for each chain. An acceptance rate for analyses of the RWM is 0.233 (optimal value 0.234 and range from 0.15 to 0.50 for many parameters) and acceptance rate for analyses of the HMC is 0.442 (optimal value 0.574 and range from 0.40 to 0.80 when L=1). Methods based on MCMC seems to be somewhat easier to apply in spite of convergence, given high – dimensional parametric space, can be difficult and slightly long.

**REFERENCES**
[1]    S. Brooks, A. Gelman, G. Jones, and X.-L. Meng, *Handbook of markov chain monte carlo*. CRC press, 2011.
[2]    W. K. Hastings, "Monte Carlo sampling methods using Markov chains and their applications," 1970.
[3]    N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equation of state calculations by fast computing machines," *J. Chem. Phys.*, vol. 21, no. 6, pp. 1087–1092, 1953.
[4]    R. M. Neal, "MCMC using Hamiltonian dynamics (Handbook of Markov Chain Monte Carlo) ed S Brooks et al," *A Gelman, G Jones XL Meng*, 2011.
[5]    Y.-A. Ma, E. B. Fox, T. Chen, and L. Wu, "Irreversible samplers from jump and continuous Markov processes," *Stat. Comput.*, vol. 29, no. 1, pp. 177–202, 2019.
[6]    S. Syed, A. Bouchard-Côté, G. Deligiannidis, and A. Doucet, "Non-reversible parallel tempering:

A scalable highly parallel MCMC scheme," *arXiv Prepr. arXiv1905.02939*, 2019.

[7]  A. Bouchard-Côté, S. J. Vollmer, and A. Doucet, "The bouncy particle sampler: A nonreversible rejection-free Markov chain Monte Carlo method," *J. Am. Stat. Assoc.*, vol. 113, no. 522, pp. 855–867, 2018.

[8]  J. Bierkens, P. Fearnhead, and G. Roberts, "The zig-zag process and super-efficient sampling for Bayesian analysis of big data," *Ann. Stat.*, vol. 47, no. 3, pp. 1288–1320, 2019.

[9]  A. J. Dobson and A. G. Barnett, *An introduction to generalized linear models*. CRC press, 2018.

[10]  C. Robert and G. Casella, "Monte carlo statistical methods springer-verlag," *New York*, 2004.

[11]  K. M. Hanson, "Markov Chain Monte Carlo posterior sampling with the Hamiltonian method," in *Medical Imaging 2001: Image Processing*, 2001, vol. 4322, pp. 456–467.

[12]  M. Girolami and B. Calderhead, "Riemann manifold langevin and hamiltonian monte carlo methods," *J. R. Stat. Soc. Ser. B (Statistical Methodol.*, vol. 73, no. 2, pp. 123–214, 2011.