An Application of Bayesian Methods on Binary Response Data

Sihang Jiang
Department of Applied Mathematics and Statistics
Johns Hopkins University
Baltimore, MD 21218
sjiang32@jhu.edu

December 19, 2020

Abstract

The purpose of this paper is to explore the performance of different methods while constructing Bayesian models, while the response data is binary. We consider using different link functions in the binary regression model, which is a generalization of the linear regression model. Also, we explore the difference between posterior distributions of parameters while using different prior distributions.

1 Introduction

1.1 Data set

When the response variable is binary, there are only two possible values for the response variable. In this paper, we use a data set of images of banknotes [1]. The response variable is encoded as 0 (authentic) or 1 (forgery) and there are 4 predictor variables, which are variance, skewness, kurtosis and entropy of images. There are many different techniques for a problem with a binary response variable, including k-NN, support vector machines, decision trees and random forest, and many others. In this case we consider a Bayesian regression model.

1.2 Method description

Suppose that N independent binary random variables $Y_1, ..., Y_N$ are observed, where Y_i is distributed Bernoulli with probability p_i. Define binary regression model as $p_i = H(\mathbf{x}_i^T\beta), i = 1, ..., N$, where β is a $k \times 1$ vector of unknown parameters, $\mathbf{x}_i^T = (x_{i1}, ...x_{ik})$ is an observation of the predictor variable, and H is a known cdf linking probabilities p_i with the linear structure $x_i^T\beta$. In

this paper we consider two cases. In the first case H is the cdf of Gaussian distribution and in the second case H is the cdf of t-distribution. [2]

Let $\pi(\beta)$ be the prior density of β, then the posterior is

$$\pi(\beta|\text{data}) = \frac{\pi(\beta)\prod_{i=1}^{N} H(\mathbf{x}_i^T\beta)^{y_i}(1-H(\mathbf{x}_i^T\beta))^{1-y_i}}{\int \pi(\beta)\prod_{i=1}^{N} H(\mathbf{x}_i^T\beta)^{y_i}(1-H(\mathbf{x}_i^T\beta))^{1-y_i}d\beta}$$

which is usually intractable. So we consider a simulation-based approach [2][4] to calculate the posterior distribution of β. Suppose the link function H is the standard Gaussian cdf, and the key idea is to introduce N independent latent variables $Z_1, ..., Z_N$ where Z_i is distributed $N(\mathbf{x}_i^T\beta, 1)$, and define $Y_i = 1$ if $Z_i > 0$ and $Y_i = 0$ if $Z_i \leq 0$. Gibbs sampling [3] is the key algorithm in this method, and full conditional distributions for parameters are required.

2 Modeling

2.1 the Gaussian link

Let $H = \Phi$, leading to the probit model [5]. Introduce N independent latent variables $Z_1, ..., Z_N$ where Z_i is distributed $N(\mathbf{x}_i^T\beta, 1)$, and define $Y_i = 1$ if $Z_i > 0$ and $Y_i = 0$ if $Z_i \leq 0$. It's easily shown that Y_i are independent Bernoulli random variables with $p_i = P(Y_i = 1) = \Phi(\mathbf{x}_i^T\beta)$. Then the joint posterior is

$$\pi(\beta, \mathbf{Z}|\mathbf{y}) = C\pi(\beta)\prod_{i=1}^{N}\{I(Z_i > 0)I(y_i = 1) + I(Z_i \leq 0)I(y_i = 0)\} \times \phi(Z_i; \mathbf{x}^T\beta, 1)$$

So we have the full conditional distributions. When prior of β is diffuse we have

$$\beta|\mathbf{y}, \mathbf{Z} \sim N(\hat{\beta}_Z, (\mathbf{X}^T\mathbf{X})^{-1})$$

where $\hat{\beta}_Z = (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{Z})$. When prior of β is conjugate prior $N(\beta^*, \mathbf{B}^*)$, we have

$$\beta|\mathbf{y}, \mathbf{Z} \sim N(\tilde{\beta}, \tilde{\mathbf{B}})$$

where $\tilde{\beta} = (\mathbf{B}^{*-1} + \mathbf{X}^T\mathbf{X})^{-1}(\mathbf{B}^{*-1}\beta^* + \mathbf{X}^T\mathbf{Z})$ and $\tilde{\mathbf{B}} = (\mathbf{B}^{*-1} + \mathbf{X}^T\mathbf{X})^{-1}$. Also, we have

$$Z_i|\mathbf{y}, \beta \sim N(\mathbf{x}_i^T\beta, 1)$$

truncated at the left by 0 if $y_i = 1$ and truncated at right by 0 if $y_i = 0$.

2.2 the t link

Let Z_i be independently distributed from t distributions with locations $\mathbf{x_i^T}\beta$, scale parameter 1 and degrees of freedom ν. Introduce additional random variable λ_i, we have $Z_i|\lambda_i$ is distributed $N(\mathbf{x_i^T}\beta, \lambda_i^{-1})$ and ν is distributed Gamma $(\nu/2, 2/\nu)$. We choose a uniform prior for β. Then we have the posterior distribution $\pi(\mathbf{Z}, \lambda, \beta, \nu|\mathbf{y}) = C\pi(\nu)\prod_{i=1}^{N}(\{I(Z_i > 0)I(y_i = 1) + I(Z_i \leq$

$0)I(y_i = 0)\}\sqrt{\lambda_i/2\pi} \times \exp(-\lambda_i/2(Z_i - \mathbf{x}_i^T\beta)^2)c(\nu)\lambda_i^{\nu/2-1}e^{-\nu\lambda_i/2})$ where $c(\nu) = [\Gamma(\nu/2)(\nu/2)^(\nu/2)]^{-1}$ and $\pi(\nu)$ is the prior on ν. We choose a uniform prior for β then we have the full conditional distributions:

$$\beta|\mathbf{y}, \mathbf{Z}, \lambda, \nu \sim N(\hat{\beta}_{Z,\lambda}, (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1})$$

where $\hat{\beta}_{Z,\lambda} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{W}\mathbf{Z})$ and $\mathbf{W} = \text{diag}(\lambda_i)$.

$$Z_i|\mathbf{y}, \beta, \lambda, \nu \sim N(\mathbf{x}_i^T\beta, \lambda_i^{-1})$$

truncated at the left by 0 if $y_i = 1$ and truncated at right by 0 if $y_i = 0$.

$$\lambda_i|\mathbf{y}, \mathbf{Z}, \beta, \nu \sim \text{Gamma}(\frac{\nu+1}{2}, \frac{2}{\nu + (Z_i - \mathbf{x}_i^T\beta)^2})$$

where λ_i are independent with each other. Also the pdf of $\nu|\mathbf{y}, \mathbf{Z}, \beta, \lambda$ is proportion to

$$\pi(\nu)\prod_{i=1}^{N}(c(\nu)\lambda_i^{\nu/2-1}e^{-\nu\lambda_i/2})$$

2.3 Hierarchical Analysis

We follow the introduction of normal hierarchical models by Landley and Smith [2][6]. We have $\mathbf{Z} \sim N(\mathbf{X}\beta, \mathbf{I})$, $\beta \sim N(\mathbf{A}\beta^0, \sigma^2\mathbf{I})$, and the prior of (β^0, σ^2) has the density $\pi(\beta^0, \sigma^2)$. In usual practice we assume β^0 and σ^2 are indenpendent with β^0 assigned a uniform prior and σ^2 assigned a noninformative prior. Now consider the full conditional distributions. The full conditional distribution of $\mathbf{Z}$ is same as the previous model, i.e.

$$Z_i|\mathbf{y}, \beta \sim N(\mathbf{x}_i^T\beta, 1)$$

truncated at the left by 0 if $y_i = 1$ and truncated at right by 0 if $y_i = 0$. Also, we have

$$\beta|\mathbf{Z}, \sigma^2 \sim N(\boldsymbol{\mu}, \boldsymbol{V})$$

where $\boldsymbol{\mu} = \mathbf{W}_1\hat{\theta}_1 + (\mathbf{I} - \mathbf{W}_1)\mathbf{A}\hat{\theta}_2$, $\hat{\theta}_1 = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Z}$, $\hat{\theta}_2 = [\mathbf{A}^T\mathbf{X}^T(\mathbf{I} + \mathbf{X}\mathbf{X}^T\sigma^2)^{-1}\mathbf{X}\mathbf{A}]^{-1}\mathbf{A}^T\mathbf{X}^T(\mathbf{I} + \mathbf{X}\mathbf{X}^T\sigma^2)^{-1}\mathbf{Z}$, $\mathbf{W}_1 = [\mathbf{X}^T\mathbf{X} + \mathbf{I}/\sigma^2]^{-1}\mathbf{X}^T\mathbf{X}$, $\mathbf{V} = ((\mathbf{I} - \mathbf{W}_1)\mathbf{A})[\mathbf{A}^T\mathbf{X}^T(\mathbf{I} + \mathbf{X}\mathbf{X}^T\sigma^2)^{-1}\mathbf{X}\mathbf{A}]^{-1}((\mathbf{I} - \mathbf{W}_1)\mathbf{A})^T + [\mathbf{X}^T\mathbf{X} + \mathbf{I}/\sigma^2]^{-1}$. And the pdf of $\sigma^2|\mathbf{Z}$ is proportional to

$$c(\mathbf{Z})\frac{|(\mathbf{I} + \mathbf{X}\mathbf{X}^T\sigma^2)^{-1}|^{1/2}}{|\mathbf{A}^T\mathbf{X}^T(\mathbf{I} + \mathbf{X}\mathbf{X}^T\sigma^2)^{-1}\mathbf{X}\mathbf{A}|^{1/2}} \times \exp\{-\frac{1}{2}Q(\mathbf{Z}, \mathbf{X}\mathbf{A}\hat{\theta}_2, \mathbf{I} + \mathbf{X}\mathbf{X}^T\sigma^2)\}\pi(\sigma^2)$$

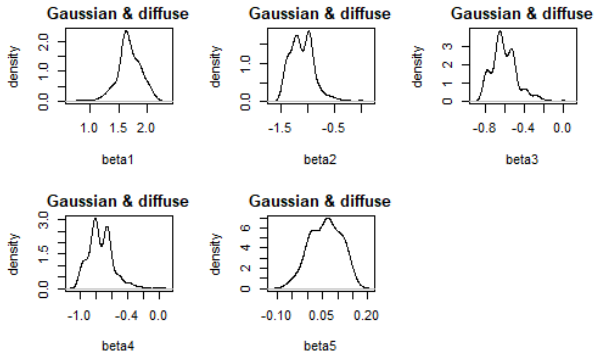
where $Q(\mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\mathbf{Z} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{Z} - \boldsymbol{\mu})$ and $c(\mathbf{Z})$ is a proportionality constant.

3 Numerical result

Now we show the numerical result of the posterior distribution of β below.

3.1 Gaussian link with diffuse prior

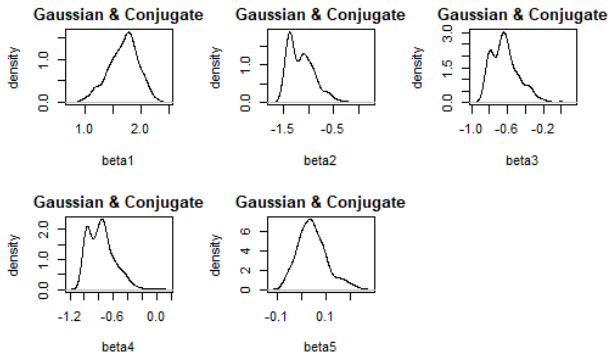
In this case we choose the Gaussian link and a diffuse prior distribution for β. The posterior is shown as below.



Posterior	Mean	Variance
β_1	1.67575271	0.04792136
β_2	-1.0915133	0.0514869
β_3	-0.60024045	0.01751719
β_4	-0.74623316	0.02457045
β_5	0.059680345	0.002604605

3.2 Gaussian link with conjugate prior

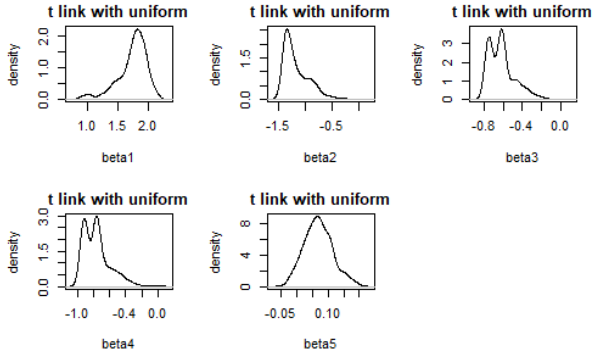
In this case we consider the Gaussian link and a conjugate prior for β described in 2.1. Specifically, $\mathbf{B}^* = \mathbf{I}$, $\beta^* = (0.5, 0.5, 0.5, 0.5, 0.5)^T$. The posterior is shown as below.



Posterior	Mean	Variance
β_1	1.68938852	0.06867735
β_2	-1.13895981	0.06835948
β_3	-0.62492134	0.02294017
β_4	-0.76939014	0.03180576
β_5	0.043121727	0.003485647

3.3 t link with uniform prior

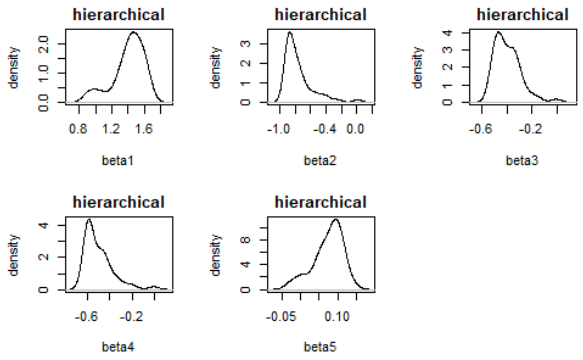
In this case we consider the t link and a uniform prior for β. Specifically, to simplify the calculation of full conditional distributions, let $\nu = 10$ in this case. In general, we can choose a discrete uniform prior for ν. The posterior is shown as below.



Posterior	Mean	Variance
β_1	1.74188693	0.06049281
β_2	-1.1653748	0.0598873
β_3	-0.6147554	0.0190480
β_4	-0.76776123	0.02822515
β_5	0.06846075	0.00202409

3.4 Hierarchical analysis

In this case we consider the hierarchical model described in 2.3. Specifically let $\mathbf{A} = \mathbf{I}$, and σ^2 has a discrete uniform prior, which is $p(\sigma^2 = 1) = p(\sigma^2 = 2) = 0.5$. Generally σ^2 can take a prior with more possible values, and the posterior of σ^2 is a multinomial distribution. The posterior is shown as below.



Posterior	Mean	Variance
β_1	1.40130711	0.03886671
β_2	-0.75666925	0.03740418
β_3	-0.39616136	0.01117426
β_4	-0.50309811	0.01671211
β_5	0.073691868	0.001568662

4 Summary

β_2, β_3, β_4, β_5 represent the coefficients of predictor variables variance, skewness, kurtosis and entropy. As a result of all methods, it's likely that the probability of a banknote to be forgery has a positive correlation with entropy, and a negative correlation with variance, skewness and kurtosis. Further, we could do hypothesis testing with respect to each coefficient of predictor variables. Also, when we are given a vector of predictor variables, we can predict the probability of a banknote to be forgery based on methods above.

References

[1] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science [http://archive.ics.uci.edu/ml].

[2] Albert, James H., and Siddhartha Chib. "Bayesian Analysis of Binary and Polychotomous Response Data." *Journal of the American Statistical Association*, vol. 88, no. 422, 1993, pp. 669–679. JSTOR, www.jstor.org/stable/2290350.

[3] Gelfand, Alan E., and Adrian F. M. Smith. "Sampling-Based Approaches to Calculating Marginal Densities." *Journal of the American Statistical Association*, vol. 85, no. 410, 1990, pp. 398–409. JSTOR, www.jstor.org/stable/2289776.

[4] Greg C. G. Wei, and Martin A. Tanner. "Posterior Computations for Censored Regression Data." *Journal of the American Statistical Association*, vol. 85, no. 411, 1990, pp. 829–839. JSTOR, www.jstor.org/stable/2290022.

[5] Tanner, Martin A., and Wing Hung Wong. "The Calculation of Posterior Distributions by Data Augmentation." *Journal of the American Statistical Association*, vol. 82, no. 398, 1987, pp. 528–540. JSTOR, www.jstor.org/stable/2289457.

[6] Lindley, D. V., and A. F. M. Smith. "Bayes Estimates for the Linear Model." *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 34, no. 1, 1972, pp. 1–41. JSTOR, www.jstor.org/stable/2985048.