STA145 GROUP PROJECT
BAYESIAN LOGISTIC REGRESSION
(On Diabetes Incidence)

March 16, 2016

Members of TEAM :
Yanlin Li 998303301
Peilin Qiu 998301989
Jun Ma 912381574
Linhao Jiang 998844857
University of California, Davis

1 Problem and Motivation

The probability of a person having diabetes is believed to be related to various explanatory variables, such as age, gender, cholesterol, weight and etc. In this project we will investigate how to perform Bayesian Logistic Regression (BLR), and apply BLR to a dataset of size 375 individuals across 10 characteristics about diabetes incidence in an African-American community. In standard logistic regression setting, in order to achieve an ideal and accurate estimate, large sample size must be acquired beforehand. In Bayesian Logistic Regression, it gives us the ability to discard such a restraint and to solely establish our focus on the information and our knowledge of the prior distribution.

2 Model Details

2.1 Describe Given Fitted Model

We are modeling the probability of a person having diabetes in an African-American community as a function of some explanatory variables, such as the location in which the person live, age, gender, height, weight, blood pressure, waist and hip measurement by a logistic regression. Mathematically, we consider a binary regression set up in which Y_1,Y_2...Y_n are independent Bernoulli random variable such that the probability that individual i has diabetes p_i will be modeled by $p_i(x_i,\beta) = \Pr(Y_i{=}1|\beta) = \mathrm{F}(x_i^T\beta)$, where x_i is a k×1 column vector of known covariates associated with Y_i, β is a k×1 vector of unknown regression coefficients we are going to work out. F here is our standard logistic distribution function, $\mathrm{F(t)} = \frac{e^t}{1+e^t}$, where t = $x_i^T\beta$. Let x_i be a (k × 1 column) vector of (k -1) explanatory variables and intercept for individual i.

2.2 Assumptions of the Model

Binary Logistic Regression requires the dependent variable to be binary. Moreover, we require the error terms to be independent. Logistic Regression also requires each observation to be independent, i.e. the model should have little or no multicollinearity. This means that the explanatory variables should be independent from each other. However, there is the option to include interaction effects of categorical variables in the the model, which we did not go into in this project. In addition, while logistic regression does not require the dependent and independent variables to be linearly related, it requires the independent variables to be linearly related to the log odds. Lastly, regular logistic regression requires quite large sample sizes, but not bayes logistic regression.

2.3 Prior and Posterior Distribution

The joint mass function of Y_1,Y_2...Y_n given $Y_i|\beta \sim \mathrm{Bin}(m_i,p_i(x_i,\beta))$, where $p_i(x_i,\beta) = \mathrm{F}(x_i^T\beta)$, and $\mathrm{F(t)} = \frac{e^t}{1+e^t}$, for i = 1,2,...,n; $m_i = 1$, ∀i, is given by:

$$\prod_{i=1}^{n} Pr(Y_i = y_i|\beta) = \prod_{i=1}^{n} \left[F\left(x_i^T\beta\right)\right]^{y_i} \left[1 - F\left(x_i^T\beta\right)\right]^{1-y_i} I_{\{0,1\}}\left(y_i\right)$$

then the posterior density of β given the data y_i is:

$$\pi(\beta|y) \propto \pi(\beta) \prod_{i=1}^{n} Pr(Y_i = y_i|\beta) \qquad where \quad \pi(\beta) \sim N(\beta_0, \Sigma_0)$$

$$\propto |\Sigma_0|^{-\frac{1}{2}} \exp\{-\frac{1}{2}(\beta - \beta_0)^T \Sigma^{-1}(\beta - \beta_0)\} \prod_{i=1}^{n} \left[F\left(x_i^T\beta\right)\right]^{y_i} \left[1 - F\left(x_i^T\beta\right)\right]^{1-y_i} I_{\{0,1\}}(y_i)$$

3 Computational Details

3.1 Sampling Method

According to Choi and Hobert (2013)'s paper, regardless of the choice of F, the posterior density is intractable, but such posterior density is required for Bayesian inference. Also unfortunately, the Gaussian distribution is not the conjugate prior of the likelihood function in logistic regression. As a result, the posterior distribution is difficult to calculate. Based on Polson, Scott, and Windel (2013)'s paper, binomial likelihoods parameterized by log-odds can be represented as mixtures of Gaussians with respect to a Pólya-Gamma distribution. Thus we use Pólya-Gamma to sample from our posterior distribution. Given prior $\pi(\beta) \sim N_p(\mathrm{b},\mathrm{B})$, there are two iterate steps involved in the procedures:

1. Draw ω_i ,...,ω_i independently with

$$(\omega_i|\beta) \sim PG(1, |x_i^T\beta|)$$

and call the observed value $\omega_i = (\omega_1, \omega_1, ..., \omega_n)^\top$

2. Draw $(\beta^{m+1} \sim N_p(\mathrm{m}(\omega),\Sigma(\omega))$

where

$$\Sigma(\omega) = (X^T\Omega(w)X + B^{-1})^{-1}$$
$$m(\omega) = \Sigma_\omega(X^T\kappa + B^{-1}b)$$
$$\Omega(\omega) = n \times n\, diagonal\, matrix\, such\, that\, D_{i,i} = \omega_i$$
$$\kappa = (y_1 - \frac{m_1}{2}, ..., y_n - \frac{m_n}{2})^T$$

3.2 Choosing Hyperparameters and Sampling Size

We start our β's with the initial value of column of 0's. Given prior $\pi(\beta) \sim N_p$(b,B), we need to choose the value for the two hyperparameters b and B. b is the prior mean of β's of dimension k×1, and B is the covariance matrix for β_i's. For simplicity, we choose this covariance matrix to be a diagonal matrix, in the form of C*I, where C is a real number, and I is the identity matrix. When C is big, i.e. our prior of β's has a large variance, we have a weak prior. Vice Versa, when C is small, the variance of our prior is small, we have a strong and informative prior of β's. We will perform our Bayesian analysis on three set of different prior. 1) We choose the prior mean, b, to be the MLE value computed from GLM function in R, and C to be 1. This is our informative prior in our analysis, we refer this choice as MLE_1 prior later in the project. 2) We choose the prior mean, b, still to be the MLE value computed from GLM function in R, but C to be 1000. This is our weak prior, and we refer this as MLE_1000. 3) We choose the prior mean, b, to be a column of 0's, and C to be 1. This is also an informative prior as the variance is very small, however, in order to distinguish this prior from MLE_1 prior, we refer this choice as the null prior.

We choose our sampling size to be 5,000 and burn-in size to be 200 to make sure that all effective sample sizes are at least above 1,000 across all variables and priors.

4 Data Analysis

4.1 Pre-Sampling Variable Selection

The dataset originally contain 10 explanatory variables. By plotting the scatter plots between all paired data (Figure 1 in Appendix), we see obvious correlation among hip, waist and weight. Hips has a strong positively linear correlation with waist; waist and hip are also positively and linearly correlate to weight; all with correlation values above 0.8. This makes sense in real life, as a person with greater hip is more likely to have a greater waist measurement, thus weight. However, we thought that blood pressure low may be correlated to blood pressure high, but in fact in our data, they do not have a strong linear correlation (correlation value of only 0.61). By the assumption of logistic model, we should not allow big multicollinearity among our explanatory variables. Thus we updated the weight in our dataset to be the average of waist, hip and weight, and discard the variables: hip and waist. Now we plot the paired scatter plot again (Figure 2), we do not see any obvious linear relationship among any two explanatory variables.

4.2 Sensitive Analysis

In general, parameter estimates and standard error of the posterior distribution with the weak prior using Bayesian should be very similar to those computed using likelihood (glm) settings. Our sampled posterior β is consistent with this. Interestingly, posterior estimates using informative prior are also very close to GLM approach, which may first lead to the false conclusion that our prior is not so sensitive. However, the informative posterior

estimate using the null prior produce somewhat different results than the previous two, and in fact it is the best prior among the three. So, our prior distribution of β is sensitive, and its sensitiveness depends on the hyperparameters we choose. In our case, it is more sensitive to changes in mean than changes in variance. Closer investigation will be performed based on the summary of the sampled posteriors of β and value computed from GLM.

β Value Across Different Sampling Methods

	MLE_1	MLE_1000	GLM	Null
Intercept, β_0	-13.7116	-15.0373	-13.5372	-0.9229
(SE)	(1.0476)	(4.529)	(4.4818)	(1.0019)
Cholesterol, β_1	0.0095	0.0094	0.0095	0.0083
(SE)	(0.0039)	(0.0039)	(0.0034)	(0.0032)
Location, β_2	-0.1537	-0.1471	-0.1711	-0.3028
(SE)	(0.2631)	(0.2873)	(0.3193)	(0.2702)
Age, β_3	0.0547	0.0567	0.0539	0.0388
(SE)	(0.0111)	(0.0123)	(0.0128)	(0.0106)
Gender, β_4	-0.3343	-0.4551	-0.1792	0.4018
(SE)	(0.3697)	(0.5329)	(0.4647)	(0.396)
Height, β_5	0.0584	0.0760	0.0550	-0.1057
(SE)	(0.0262)	(0.0658)	(0.0628)	(0.0254)
Weight, β_6	0.0347	0.0353	0.0331	0.0343
(SE)	(0.0097)	(0.0099)	(0.0097)	(0.0095)
BloodPressureHigh, β_7	0.0037	0.0033	0.0046	0.0051
(SE)	(0.0090)	(0.0091)	(0.0091)	(0.0080)
BloodPressureLow, β_8	-0.0013	0.0001	-0.0011	-0.0154
(SE)	(0.0158)	(0.0161)	(0.0162)	(0.0143)
Min ESS of β	1531	1414	NA	1773

Looking at the standard error of the posterior estimates and GLM. MLE_1000 has the overall greatest SE. The SE of GLM is very close to that of MLE_1000, weak prior. The SE of both informative prior, MLE_1000 and Null are very similar, and slightly smaller than those of the weak prior and GLM. This change can be attributed to the slight additional prior information added in the Bayesian analysis.

Effective sample size(ESS) for all β of three priors are greater than 1,000. Since we have multiple β, we look the mim ESS of the $\beta.MLE\_1000$, the weak prior, unsurprisingly has the lowest min(ESS) value, followed by informative prior MLE_1. Our Null prior gives us the greatest ESS.

Looking at the 27 trace plots (Figure 3,4,5) for β estimates of using different priors.

Note that the trace plot of beta 1 with informative prior starts at a very remote initial value, and then it travels to the target distribution indicating that we need to choose a relative burn-in sample size to exclude remote initial value, thus we choose burn-in sample size to be 200. Other 26 trace plots, however, display almost "perfect". The centers of the chains appear at some value with very small fluctuations. This indicates that the chains are mixing well. From these trace plots, we conclude that all three priors are reasonable prior.

The ACF value (Figure 6) for β's of different priors is consistent with the conclusion above. When the lag-1 and lag-2 autocorrelations of the parameter are large (or decrease very slowly), this indicates the samples are highly dependent (non-stationary). Therefore, if the estimator works reasonable, we expect there is a significant decrease between lag-1 and lag-2 autocorrelations, which happens for all priors here. Moreover, two informative prior have obvious larger change in percent compare to the flat prior, and percentage change of null prior is higher than change of MLE_1, for 8 out of 9 β's.

From the above analysis of different priors, we conclude that although all three priors are reasonable choices, the two informative MLE_1 and the null priors are for sure better than the weak priors. Moreover, informative null prior is slightly better than MLE_1.

4.3 Post-Sampling Variable Selection

Our selection rule is, we keep the parameter when the 95% confidence interval for that does not contain zero. The decision whether to include the variance across 4 methods are recorded into the table below.(We ignore decision about the intercept.)

	MLE_1	MLE_1000	GLM	Null
Cholesterol, β_1	Include	Include	Include	Include
Location, β_2	NI	NI	NI	NI
Age, β_3	Include	Include	Include	Include
Gender, β_4	NI	NI	NI	NI
Height, β_5	Include	NI	NI	Include
Weight, β_6	Include	Include	Include	Include
BloodPressureHigh, β_7	NI	NI	NI	NI
BloodPressureLow, β_8	NI	NI	NI	NI

GLM approach and the weak prior are consistent with each other, and the two informative prior are consistent with each. Since we already decide informative Null prior is the best, we will follow the decision of Null. We discard location, gender, BloodPressureHigh and BloodPressureLow and left with only four variables: cholesterol, age, weight and height, which matches the common belief in real life. We run our baysisan regression on the data set again, and finally have $\beta_{Cholesterol} = 0.0066$, $\beta_{Weight} = 0.0274$, $\beta_{Age} = 0.0431$, $\beta_{height} =$ -0.1003

4.4 Interpretation of β

Adjusting for the other variables; for every one unit increase in the level of cholesterol, the odds of having diabetes are multiplied by exp(0.0066)= 1.01; for every one year increase in the age of the subject, the odds of having diabetes are multiplied by exp(0.0431)= 1.04; for every one unit increase in the weight of the subject, the odds of having diabetes are multiplied by exp(0.0274)= 1.03; for 1 unit increase in height, the odds of having diabetes are multiplied by exp(-0.1003)= 0.9.

5 Lessons Learned

We discovered that classical Monte Carlo does not work in this case, because the posterior distribution is not a standard distribution, that is hard to sample from. And it is very time consuming to apply inverse CDF in multivariate scenario. Rejection sampling will be a good approach if we could find a reasonable proposal density. After all, we used Gibbs sampler to converge our target density distribution, and use Pólya-Gamma to mimicking the missing conditional posterior argument in the logistic case.

6 Appendices: Figures

Figure 1: MultiCollinearity1

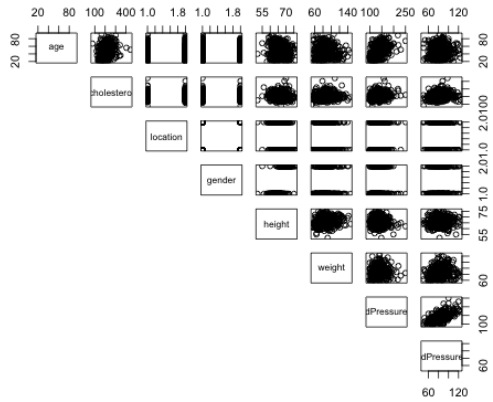

Figure 2: MultiCollinearity1

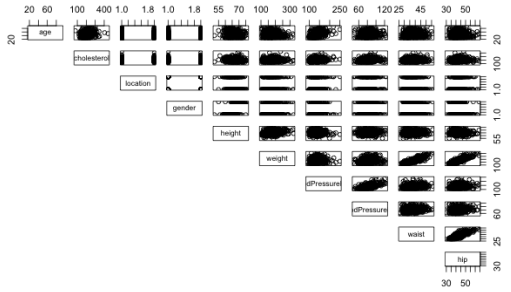

Figure 3: Informative MLE_1 Prior Trace Plot

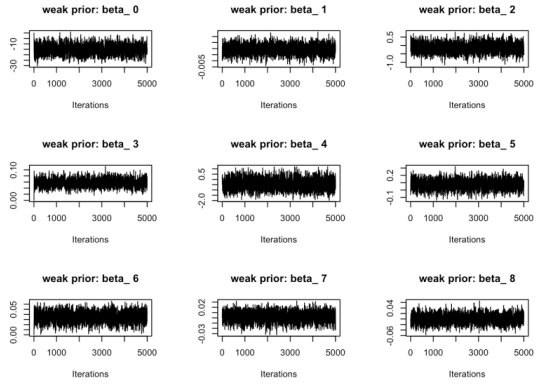

Figure 4: Weak MLE_1 Prior Trace Plot

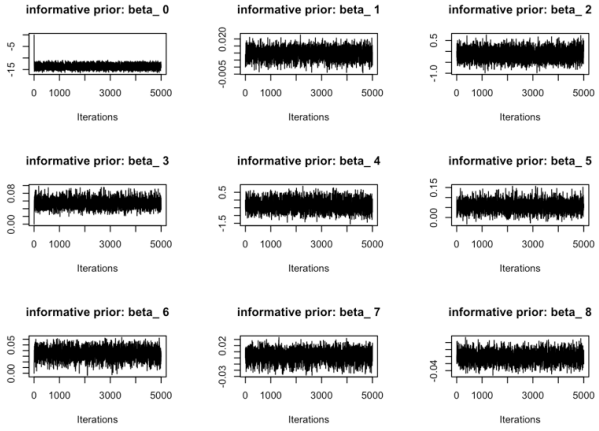

Figure 5: Null Prior Tace Plot

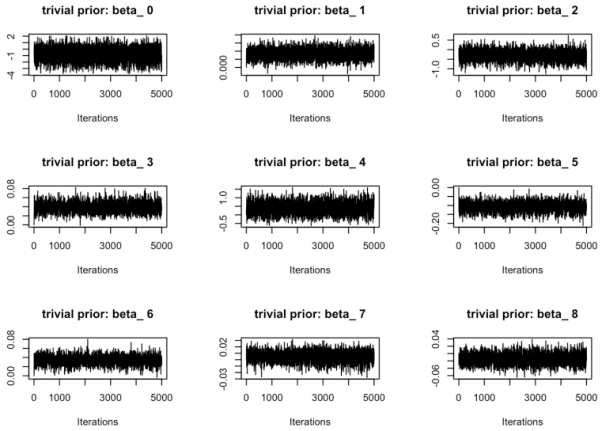

Figure 6: Percentage Change in ACF Comparision

	mle_1	mle_1000	null
intercept :beta_0	-72.9	-43.4	-91.3
cholesterol :beta_1	-58.2	-62.7	-66.5
location :beta_2	-62.5	-55.7	-65.4
age :beta_3	-66.0	-46.2	-52.6
gender :beta_4	-45.9	-45.8	-55.8
height :beta_5	-45.0	-48.9	-53.3
weight :beta_6	-54.7	-51.8	-58.1
BloodPressureHigh :beta_7	-72.2	-73.8	-70.9
BloodPressureLow :beta_8	-58.0	-55.3	-66.0