Bayesian Logistic Regression on Heart Disease: Inference, Prediction, and Comparison.

Chen Xie, Xinye Jiang, Xun Wang

2019/4/29

1. Introduction

Logistic regression is a very famous and widely applied technique in classification. It is usually used to perform predictive analysis when the response variable is binary.

Logistic regression can also be approached by bayesian modeling. In general, bayesian analysis is more flexible, and it is proved to be superior for small samples. Above all, We can incorporate prior information in bayesian modeling. For example, if we want to reduce the dimension of predictors and avoid overfitting, some shrinkage priors can be chosen to implement regularization.

In practice, predicting a binary response can be an application of the standard logistic regression as well as the bayesian approach. In this report, we delve into a data set which is about heart health condition of 303 patients. This data set was collected between May 1981 and September 1984 at the Cleveland Clinic in Cleveland, Ohio. The objective is to predict whether the patients have heart disease based on 13 independent variables, such as `age`, `sex`, `chest pain type`, etc. In this process, we will explore the different effects of specific predictors on response variable in different models, and will also compare prediction performance among models.

2. Data Exploration

2.1 Dataset

The dataset is about heart disease of patients at the Cleveland Clinic in Cleveland, Ohio. It is from UCI Machine Learning Repository, and has 303 observations and 14 variables in total. Every row is associated with a patient. The response variable `target` is whether the angiographic result is present or absent of a diameter narrowing larger than 50% (presence=1, absence=0). In other words, the patient is diagnosed as having heart disease if `target` is 1, and not if `target` is 0. To predict the heart disease, the dataset collected 3 types of independent variables. Clinical variables such as `age`, `sex`, `cp`, `trestbps`, were related to clinical effects. Predictors `chol`, `fbs`, and `restecg` were from routine tests, while variables `thelach`, `exang`, `oldpeak`, `slope`, `ca` , and `thal` were collected from noninvasive test. More detailed descriptions of all the variables can be found in Table 1.

The data collection process can be assumed as independent and without work-up bias. For each type of variables (response, clinical, routine test, noninvasive test), the data were recorded and analyzed without any knowledge of other types of variables.

2.2 Exploratory Analysis

In this section, we perform data exploration to understand possible relationships among the variables. Firstly, we check the numerical summaries of the continuous and categorical variables in Table 2 and Table 3, respectively. The dataset is very clean and has no missing value. Table 2 also shows that the continuous variables all have some extreme values, especially `chol` and `oldpeak`. The average age of the recorded patients is 54. Table 3 indicates that we have around 1/3 data from females and the rest 2/3 from males.

Table 1: Data Description

Variables	Type	Collection	Description
target	Binary	Dependent Variable	angiographic result of the presence or absence of a >50% diameter narrowing; presence = 1; absence = 0.
age	Continuous	Clinical Variable	age in years
sex	Binary	Clinical Variable	1=male; 0=female
cp	Categorical	Clinical Variable	chest pain type; 0=typical angina; 1=atypical angina; 2=non-anginal; 3=asymptomatic
trestbps	Continuous	Clinical Variable	systolic/resting blood pressure (in mm Hg on admission to the hospital)
chol	Continuous	Routine test	serum cholestoral in mg/dl
fbs	Binary	Routine test	(fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
restecg	Binary	Routine test	resting electrocardiographic results; 0=normal; 1=having ST-T wave abnormality; 2=showing probable or definite left ventricular hypertrophy by Estes' criteria
thelach	Continuous	Noninvasive test	maximum heart rate achieved
exang	Binary	Noninvasive test	exercise induced angina (1 = yes; 0 = no)
oldpeak	Continuous	Noninvasive test	ST depression induced by exercise relative to rest
slope	Ordinal	Noninvasive test	the slope of the peak exercise ST segment; 0=upsloping; 1=flat; 2=downsloping
ca	Ordinal	Noninvasive test	number of major vessels (0-3) that appeared to contain calcium
thal	Categorical	Noninvasive test	exercise thallium scintigraphic defects; 3=normal; 6=fixed defect; 7=reversable defect

Table 2: Summary of Continuous Variables

	age	trestbps	chol	thalach	oldpeak
Min.	29.00000	94.0000	126.000	71.0000	0.000000
1st Qu.	47.50000	120.0000	211.000	133.5000	0.000000
Median	55.00000	130.0000	240.000	153.0000	0.800000
Mean	54.36634	131.6238	246.264	149.6469	1.039604
3rd Qu.	61.00000	140.0000	274.500	166.0000	1.600000
Max.	77.00000	200.0000	564.000	202.0000	6.200000

Table 3: Summary of Categorical Variables

target	sex	cp	fbs	restecg	exang	slope	ca	thal
0: 138	female: 96	0: 143	0: 258	0: 147	0: 204	0: 21	0: 175	0: 2
1: 165	male: 207	1: 50	1: 45	1: 152	1: 99	1: 140	1: 65	1: 18
		2: 87		2: 4		2: 142	2: 38	2: 166
		3: 23					3: 20	3: 117
							4: 5	

Next, we take a look at the heart disease dataset through visualization. Figure 1 shows the pairwise scatterplots and histograms of continuous variables. These variables `age`, `trestbps`, `chol`, `thelach`, `oldpeak` display weak correlations between each other. In Figure 2, we try to explore the relationship between the `target` and the continuous variables by boxplots. It is noticed that the patients who are detected to have heart disease have an overall higher maximum heart rate achieved (`thalach`). It implies that `thalach` is a possibly significant preditor on `target`.

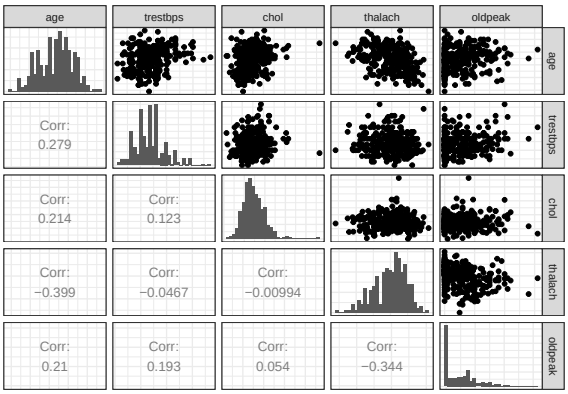

Figure 1: Pairwise scatterplots and histogram of continuous variables.

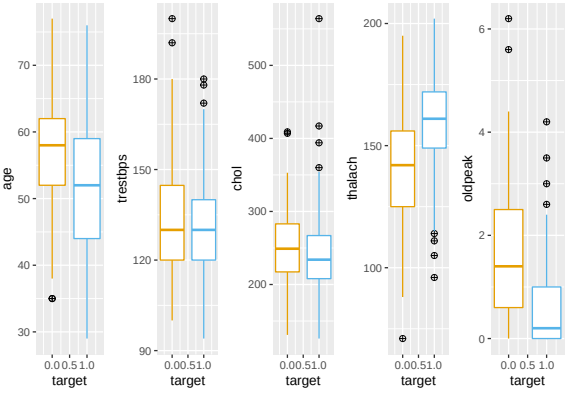

Figure 2: Boxplots of continuous variables.

Finally, let us take a quick look at the bar charts of categorical variables against response variable in Figure 3. As we could see, the response `target` is relatively balanced. In addition, we can also find that some categorical independent variables may be important to predict `target`. For instance, the distributions of `target` are quite different when `sex` is female (`sex` = 0) or male (`sex` = 1) . However, for other predictors such as `restecg`, it seems to be not very influential on `target`.

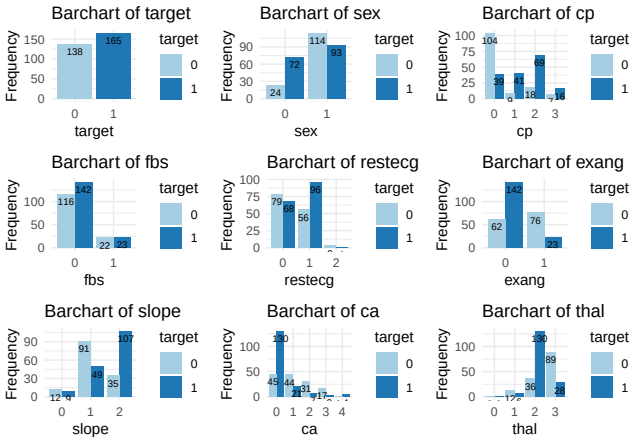

Figure 3: Bar charts of categorical variables.

3. Logistic Regressions

In this part, we will briefly introduce three logistic regression models, including a standard one and two bayesian ones. There are several advantages to use logistic regression instead of other methods here. First, it can be interpreted. It helps to explain the relationship between the predictors and the response variable. Next, the logistic regression can also handle mixed types of explanatory variables. Most importantly, we have a binary response in this problem. So logistic regression models are appropriate methods here.

3.1 Data Preparation

Before we perform logistic regressions to dataset, we need to prepare our data first. First, we turn categorical variables into dummy variables by using R function `model.matrix`. As a result, the number of independent variables grow from 13 to 23 (including intercept). In order to test the prediction performance of the models, we then randomly split the whole dataset into training (80%) and test (20%) sets. Our training dataset has 273 observations and test dataset has 30 observations.

3.2 Logistic Regression

The basic assumption of standard logistic regression model is that observations $y_1, ..., y_n$ are independent and follow binomial distribution (1, p_i), where p_i is the probability of $y_i = 1$. We could obtain point estimates of parameters β, which maximizes $\Pi_{i=1}^{n} P(Y_i = y_i)$ using maximum likelihood approach. The statistical description is as below:

$$y_1, ..., y_n, \text{ are independent, and } y_i \sim \text{ Binomial } (1, p_i)$$

$$\log(\frac{p_i}{1-p_i}) = x_i^T \beta + \beta_0$$

$$p(y_i; \beta, \beta_0) = (p_i)^{y_i}(1-p_i)^{1-y_i}, p(y_1, ..., y_n; \beta, \beta_0) = \prod_{i=1}^{n} (p_i)^{y_i}(1-p_i)^{1-y_i}$$

where x_i is the i-th row observation and β, β_0 are parameters.

Using `glm` function in R, we could easily obtain estimates of β parameters. Table 4 provides us some summary statistics of the fitted results. From p.value column, we could observe that only 6 out of 23 predictors are significant in our model. If the significance level is $\alpha = 0.1$, there are still only 9 significant predictors, which encourages us to seek sparse solutions.

Table 4: Summary of Logistic Regression

	Estimate	Std.Error	z.value	p.value	Variable	Estimate	Std.Error	z.value	p.value
(Intercept)	-14.352	1455.4	-0.0098614	0.99213	exang1	-0.87782	0.49443	-1.7754	0.075826
age	0.03118	0.027578	1.1306	0.25822	oldpeak	-0.14036	0.26153	-0.53668	0.59149
sex1	-1.799	0.62267	-2.8891	0.003863	slope1	-0.71803	0.98917	-0.72589	0.4679
cp1	1.0622	0.626	1.6969	0.089725	slope2	0.34136	1.0496	0.32523	0.74501
cp2	2.0539	0.56815	3.6151	0.00030025	ca1	-2.1807	0.55128	-3.9558	7.6278e-05
cp3	2.0025	0.72838	2.7492	0.0059741	ca2	-3.7367	0.91634	-4.0779	4.5449e-05
trestbps	-0.019333	0.01262	-1.5319	0.12554	ca3	-3.0051	1.0555	-2.8472	0.0044108
chol	-0.0062214	0.0042669	-1.4581	0.14482	ca4	1.3665	1.8836	0.72547	0.46816
fbs1	0.3258	0.64056	0.50862	0.61102	thal1	16.482	1455.4	0.011325	0.99096
restecg1	0.53266	0.42534	1.2523	0.21045	thal2	15.813	1455.4	0.010865	0.99133
restecg2	-1.2948	2.6708	-0.48482	0.6278	thal3	14.111	1455.4	0.0096957	0.99226
thalach	0.022983	0.01242	1.8506	0.064233					

3.3 Bayesian Logistic Regression with N-IG Prior

In the standard logistic regression above, we treat β as a column of unknown but fixed parameters. Actually, β, β_0 could be seen as a vector of random variables from the prospective of bayesian analysis. In this way, we could combine data (model) and the prior we build up to obtain not only a point estimate, but also the posterior samples of parameters. In addition to basic assumptions of standard logistic regression, bayesian logistic regression method with N-IG prior assumes that β, β_0 subject to a normal prior with mean μ_i and variance σ^2, where the variances follow an Inverse-Gamma distribution with hyper parameters a and b. The parameters μ_i, a, and b are assumed to be flat because we don't have much information about them. Although it probably results in improper prior, the powerful tool `Rstan` in R will help and still generate posterior samples of parameters. The statistical description is shown as below:

$$y_1, ..., y_n, \text{ are independent}$$
$$\text{Likelihood: } y_i \sim \text{ Binomial } (1, p_i)$$
$$\text{Parameters: } \log(\frac{p_i}{1-p_i}) = x_i^T \beta$$
$$\text{Prior: } \beta_i \sim N(\mu_i, \sigma_i^2)$$
$$\text{Hyper prior: } \sigma_i^2 \sim \text{Inv-Gamma}(a, b)$$
$$\mu_i, a, b \text{ are assumed to have flat distribution.}$$

Using `Rstan`, we implement Monte Carlo Markov Chain algorithm (MCMC) to our model and obtain posterior samples of interested parameters β, β_0. We naturally compute their 95% credible intervals based on the sample quantiles. The Figure 4 shows the comparision of the corresponding parameters obtained by standard logistic regression and bayesian logistic. The red lines represent the estimates of β from standard logistic regression and green lines represent the 95% credible intervals produced by posterior samples of β, while blue lines indicate the location of 0.

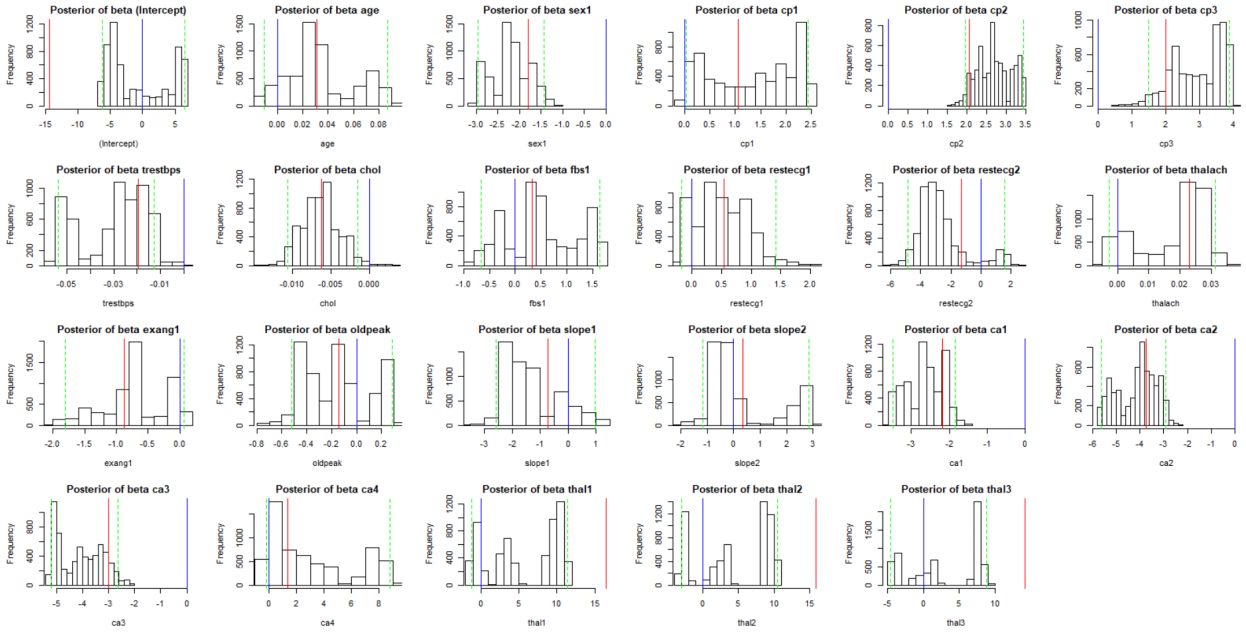

Figure 4: Estimate of parameters beta.

In fact, if the sample size is large enough compared to the number of predictors, the estimations of bayesian analysis should be very similar to those of frequentist models. Even though the dataset here is not very large

that it only has 303 observations, we can also see that most red lines, which are estimates from standard logistic regression, fall into the 95% credible intervals of posterior samples, meaning that most predictors produce similar effects in both models. In addition, comparing the blue vertical lines, which represent the '0', to the credible intervals, we can find that the significance of most parameters are not different from the results of the standard logistic regression model. In contrast, the absolute values of the coefficients of variables `intercept`, `thal1`, `thal2` and `thal3` in bayesian logistic regression are much smaller than the ones in the standard model. This may result from the fact that bayesian logistic regression could incorporate the information from priors, make full use of data and adjust the effects of these nonsignificant variables properly.

3.4 Bayesian Logistic Regression with NEG prior

Recall that in Table 4 most variables actually do not produce significant effects to the response. In order to figure out the most important predictors to the `target` response, we choose to use bayesian logistic models with shrinkage priors. There are many priors which are proved to have shrinkage effects to the parameters, for example, Cauchy prior, Laplace prior and horseshoe prior. Normal-Exponential prior, which has similar effect as LASSO regression, is a common option for logistic regression. The reason why NE prior has a shrinkage effect on β, β_0 is that the exponential distribution of variance lays a great mass of probabilities around 0. As a result, initial β, β_0 will gather around 0 with a large probability. The next problem is how we can control the shrinkage effect. In frequentist analysis, it usually uses cross-validation to choose shrinkage control parameters. But sometimes cross validation is computationally intensive. Instead, in bayesian modeling, we can build up a hyper prior usually Gamma distribution (a_0, b_0) on λ. So finally, we perform bayesian logistic regression model with NEG prior, and the detailed statistical model is shown as below.

$$y_1, ..., y_n, \text{ are independent, and } y_i \sim \text{ Binomial } (1, p_i)$$

$$\text{Parameters: } \log(\frac{p_i}{1-p_i}) = x_i^T \beta$$

$$\text{prior: } \beta_i \sim N(0, \sigma_i^2)$$

$$\text{Hyper Prior: } P(\sigma_i^2) \sim \text{Exponential } (\lambda)$$

$$\lambda \sim \text{Gamma } (a_0, b_0), \text{ where } a_0, b_0 \text{ are two fixed parameters.}$$

Another important problem for this model is how to choose hyperparameters a_0, b_0, as we do not want to too many nuisance parameters. We decide to use a method similar to model checking. First, we propose a pair of possible a_0, b_0. Second, we simulate a λ and a matrix of "Fake Data" which is generated from normal distribution $N(0, 1)$ based on our model. Then we could get a vector of generated responses following the model generating process and obtain a summary statistic (e.g., mean) of the responses. We repeat the previous steps for 1000 times and obtain the distribution of the summary statistic. Finally, comparing the distribution of `mean` summary statistic to the one of the true data, we choose reasonable values of a_0, b_0.

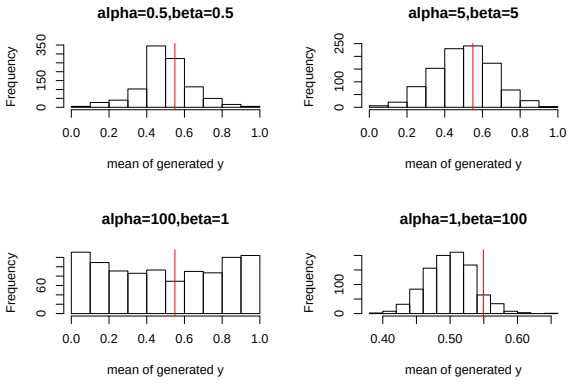

Figure 5: Selection of hyperparameters a0, b0

In this example, we try several pairs of (a_0, b_0) and their corresponding results are shown above in Figure 5. Though both $a_0 = 0.5, b_0 = 0.5$ and $a_0 = 5, b_0 = 5$ are appropriate, we finally choose $a_0 = 5, b_0 = 5$.

Implementing this model in Rstan, we obtain both the posterior distributions of β, β_0 and λ. From the posterior distribution of λ in Figure 6, bayesian method shows its power to make the posterior distribution of λ more centered around 0.8 compared to prior. And we will discuss about the differences between posterior distribution of parameters β, β_0 of bayesian logistic regression with NEG prior and the parameters from other two models in next section.

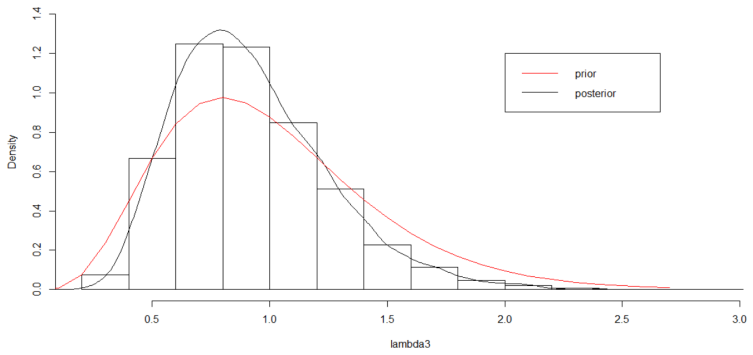

Figure 6: Histogram of the posterior distribution of lambda

4 Model Comparison

4.1 Inference

Figure 7 shows point estimates and estimated confidence intervals of parameters β, β_0 from all three different models above. Here, we take the mean of posterior samples as the point estimates for bayesian models.

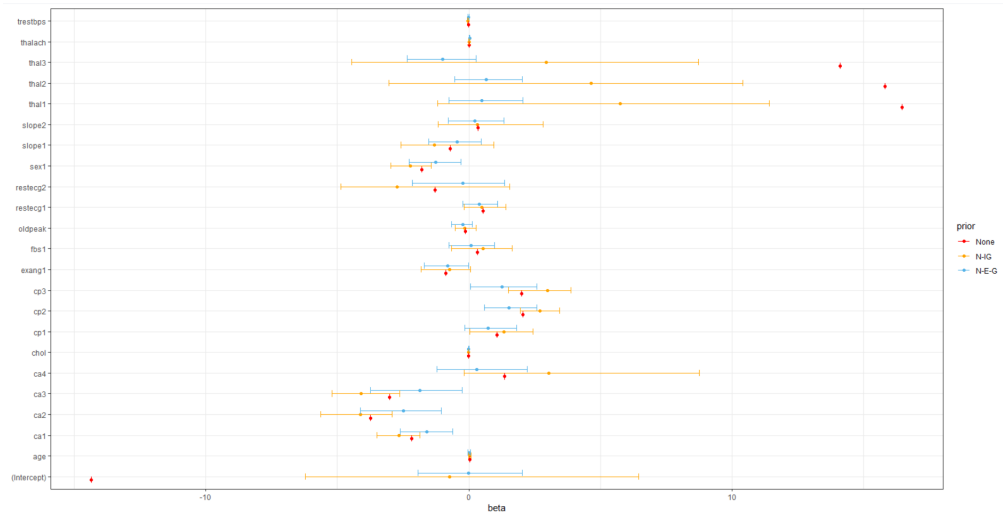

Figure 7: Confidence interval and point estimate of beta

From this plot, we could observe that NEG prior does have strong shrinkage effects on the coefficients that 0 is included in most of blue confidence intervals. We also notice that the variables kept are exactly the same significant variables in standard logistic regression, i.e., `sex1`, `cp2`, `cp3`, `ca1`, `ca2` and `ca3`. Therefore, gender,

symptom of chain pain and the number of major vessels which appeared to contain calcium are three most important factors to the `target` response.

In addition, confidence intervals with NEG prior are narrowed down compared to the ones with N-IG prior, for example, of the variables `slope1` and `intercept`, meaning that the estimates of parameters are more concentrated and precise.

In conclusion, bayesian logistic regression with NEG prior is a very efficient method which combines model fitting with variable selection.

4.2 Prediction

Based on standard logistic regression model, it is easy to make predictions using function `predict`. We make predictions and obtain the confusion matrix with 22 right predictions and 8 wrong ones as below. The `accuracy` is about 73.33%.

Table 5: Confusion matrix for logistic regression

	Predicted: No	Predicted: Yes
Actual: No	11	4
Actual: Yes	4	11

To compare the models, we are going to use two indices `log-loss` and `sensitivity` instead of `accuracy` to evaluate the prediction performance of the three models.

For prediction analysis, we should generate predictive samples from the posterior distributions of β, β_0, as the posterior distribution has been updated by incorporating information from data and prior. To do this, we first randomly draw samples of β from the posterior distribution many times. And then we generate new y which has the same size of responses as the test dataset for each sample of β. In this way, we could obtain many groups of predictions for test data. Finally, we compare each group of predicted responses with the original test data.

4.2.1 Log-Loss

To combine raw probabilities $P(y = 1|\beta, \beta_0)$ as well as the true response, we use log-loss rather than `accuracy` for predication analysis. Log-loss or logarithmic loss is a metric used in classification problems to compare the prediction performance of different models. The lower the log-loss is, the better the performance is. For logistic regression, log-loss is defined as:

$$-log(y|p) = -ylog(p) - (1-y)log(1-p)$$

where y is the target value 0 or 1, p is the probability $P(y = 1|\beta, \beta_0)$.

Using this formula, we calculate the log-loss value of the standard logistic regression and the density distributions of the log-loss values of bayesian logistic regression models and show them in Figure 8. Average log-loss values for bayesian methods are also computed and marked in the plot.

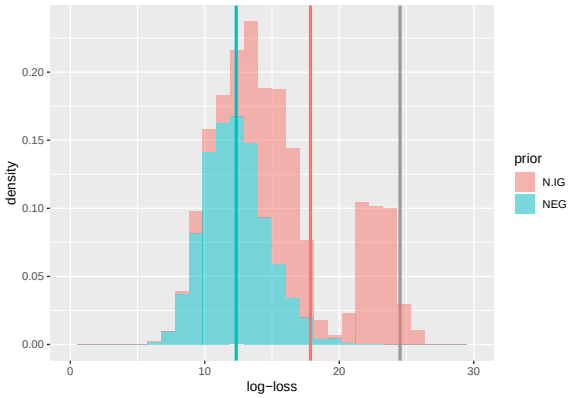

Figure 8: Histogram of predictive log-loss

It is shown that the log-loss values of bayesian models have a large improvement compared to the logistic regression, while the model with NEG prior is even better than the one with N-IG prior in log-loss. In summary, bayesian logistic regression model with shrinkage prior performs best in log-loss.

4.2.2 Sensitivity

`Sensitivity` is another index of great importance to heart disease detection. If a patient does have heart disease, it will be a big deal if doctors misdiagnose him as not having. `Sensitivity` here is used to measure the true positive rate of our prediction. That is, the proportion of the patients with heart diseases to be diagnosed correctly.

We also use a plot to illustrate the superiority of bayesian logistic regression model with NEG prior in `sensitivity`. See Figure 9. Standard logistic regression model produces a `sensitivity` of 68.25% and bayesian model with N-IG prior shows 72.05% `sensitivity`, while bayesian model with NEG prior has an average `sensitivity` of 74.56% which is higher than both of others.

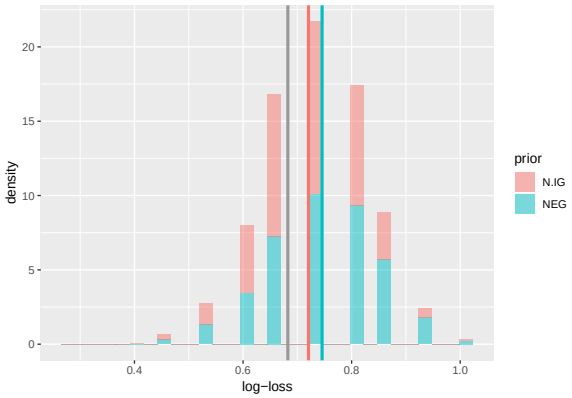

Figure 9: Histogram of predictive sensitivity

In conclusion, bayesian logistic regression model with shrinkage prior (NEG prior) outperforms other two models in prediction. The bayesian logistic regression models both have a better prediction performance than the standard logistic regression, which indicates the significance and power of the bayesian analysis.

5. Conclusion & Future Work

In conclusion, for large dataset, standard logistic regression and bayesian logistic regressions may produce very similar results. But in general, bayesian modeling is more flexible and has better performance when we do not have many observations.

We may also include sparse assumption of solutions to shrink the parameters of the non-significant variables to 0 in our model. In this way, we could reduce dimensionality and better the prediction performances, as only those factors which have important effects on the response variable would be kept. It is also easier for us to interpret the models in this case. That gender, symptom of chain pain and the number of vessels containing calcium are three most important independent variables which could affect the response greatly.

For this dataset, bayesian logistic regression with NEG prior has the best prediction performance in both log-loss and sensitivity compared to standard logistic regression and bayesian logistic regression with N-IG prior. And the two bayesian logistic regression models both have a better prediction performance than the standard logistic regression, which shows the significance and powerfulness of the bayesian analysis.

For future work, the benefit of bayesian logistic models should be verified for datasets from different clinic. Moreover, as the highest sensitivity of bayesian model with NEG is only about 74.56%, we expect that other classification methods may have better prediction performance.

6. Reference

Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J., Sandhu, S., Guppy, K., Lee, S., & Froelicher, V. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. American Journal of Cardiology, 64:304-310. http://archive.ics.uci.edu/ml/datasets/heart+disease

Wei, R. & Ghosal, S. (2017). Contraction properties of shrinkage priors in logistic regression, Preprint at http://www4.stat.ncsu.edu/~ghoshal/papers.

Genkin, A., Lewis, D. & Madigan, D. (2007). Large-scale Bayesian logistic regression for text categorization. Technometrics, 49(3): 291-304.

Kapat, P. & Wang, K. (2006). Classification Using Bayesian Logistic Regression: Diabetes in Pima Indian Women Example. Ohio State University, OH. https://www.asc.ohio-state.edu/goel.1/STAT825/PROJECTS/KapatWang_Team4Report.pdf

Li, L & Yao, W. (2017). Fully Bayesian logistic regression with hyper-LASSO priors for high-dimensional feature selection. Statistics 88, 1-25.

Park, T. & Casella, G. (2008). The Bayesian LASSO, Journal of the American Statistical Association, 103: 482, 681-686.

Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J. J., Sandhu, S., Guppy, K., Lee, S. & Froelicher, V. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. The American journal of cardiology, 64(5): 304-310.

MediaWiki (2017). Log-Loss. http://wiki.fast.ai/index.php/Log_Loss

7. Appendix

Work Division:

R codes: Xinye Jiang

Slides, Report: Chen Xie and Xun Wang