Determining the Probability of Heart Disease Using Bayesian Logistic Regression

Elliott Evans, Meng Zhang, Ziwei Zhu

April 16, 2017

1 Introduction

Cardiovascular diseases (CVDs) are a group of disorders of the heart and blood vessels. They are the number one cause of death globally, and have been a distinct concern of society and in the research community. A large number of factors can contribute to heart disease diagnosis. In this project, we examined how different factors affect cardiovascular disease, using data collected from the Cleveland Clinic Foundation (CLE) and V.A. Medical Center, Long Beach, CA (LBVA) in July 1988 [2]. We used one binary variable and five quantitative variables to fit our model. The predictors included the binary variable $HOSP1$, which is 1 when the hospital is CLE and 0 when it is LBVA (here, the reference hospital). In addition, we utilized five quantitative variables: AGE, $TRESTBPS$ (resting blood pressure), $CHOL$ (serum cholestoral), $THALACH$ (maximum heart rate achieved), and $OLDPEAK$ (ST depression induced by exercise relative to rest). The response variable is a dummy variable for whether a given patient has heart disease or not. We build a Bayesian logistic regression model that predicts the probability of having heart disease. Our primary motivation is to develop inference and analysis from our logit model to determine primary causes of heart disease, along with their impact on a given subject's probability of obtaining this disease.

2 Model/Methods

We used logistic regression to model a subjects log-odds of having heart disease. While no interaction terms are included, the dichotomous variable $HOSP1$

allows the regression parameter β_1 to represent the extra effect that being a part of the CLE hospital study has on one's odds of having heart disease. The model is presented below, and fairly week priors were used for the regression parameters, as indicated by the large variances:

$$
\begin{aligned}
[Y_i|p_i] &\sim \text{Bern}(p_i) \\
\text{logit}(p_i) &= \beta_0 + \beta_1 HOSP1_i + \beta_2 AGE_i + \beta_3 TRESTBPS_i + \beta_4 CHOL_i \\
&\quad + \beta_5 OLDPEAK_i + \beta_6 THALACH_i \\
\beta_i &\sim N(0, 1000^2) \text{ for } i = 0, 1, \ldots, 6
\end{aligned}
$$

Where:

$$
Y_i = \begin{cases} 1, & \text{if patient develops heart disease} \\ 0, & \text{otherwise} \end{cases}
$$

With regards to model fitting, we calculated WAIC [1], an extension of DIC, BIC, and AIC, created by Andrew Gelman, for the model above and with the variables $HOSP1$, $OLDPEAK$, and $THALACH$ (the variables whose credible intervals did not contain 0, as we will see in Section 3). The WAIC for the full model was 547.6 while the WAIC for the model with only significant predictors was 550.0. Since the WAIC is lower for the full model, even with a penalty for three extra parameters, it suffices to use this model for inference, and later prediction.

3 Data Analysis

Our model was run for 31,000 iterations using a 1,000 sample burn-in period. To check for convergence of our logistic regression coefficients, we observe the Geweke statistics in Table 1. We notice that these observed statistics are non-extreme, relative to a $N(0, 1)$ distribution. The most extreme of these statistics is the regression coefficient for $THALACH$, i.e. β_6, representing the effect of the patient's maximum heart rate achieved on his/her log-odds of obtaining heart disease. However, even this statistic corresponds to a two-sided p-value of 0.0659, meaning we would fail to reject the null hypothesis at the $\alpha = 0.05$ level of significance that the MCMC chain has reached the stationary posterior distribution. We also find that the chains mixed well, with the autocorrelation of the samples reaching a reasonable level by lags 3 to 5.

Before analyzing our 30,000 post burn-in samples, we perform exploratory analysis on our five quantitative variable and observe the plots in Figure 1. We notice that none of these covariates are particularly correlated with one another. In addition, we notice that the median for the variable $THALACH$ for those patients without heart disease is greater than the median for those patients with heart disease. We will see that this agrees with the results of our Bayesian analysis. In addition, the variable $OLDPEAK$, the ST depression induced by exercise relative to rest, has a greater median for those patients who had heart disease, a finding that will also come to agree with our findings in the logistic regression model.

The summary statistics for the 30,000 post burn-in samples are found in Table 1. We notice that three non-intercept coefficients have 95% credible

intervals that do *not* cover zero, namely the effects for variables $HOSP1$, $OLDPEAK$, and $THALACH$. We find that $HOSP1$, the additional effect of being a patient in the non-reference hospital (the Cleveland Clinic Foundation), has a negative posterior median of -0.8046, meaning that being in this experiment in Cleveland has the median effect of reducing the odds of heart disease by a factor of $e^{-0.8046} \approx 0.447$, relative to the patients at the LBVA hospital.

We observe the posterior densities of the effects whose credible intervals did cover zero in Figure 3. We find that the posterior medians for these effects, those of AGE, $TRESTBPS$, and $CHOL$, were all positive. Thus, we note the probabilities that the corresponding regression coefficients are positive:

$$\mathbb{P}(\beta_2|\vec{Y} > 0) \approx 0.9068 \quad \Big| \quad \mathbb{P}(\beta_3|\vec{Y} > 0) \approx 0.9512 \quad \Big| \quad \mathbb{P}(\beta_4|\vec{Y} > 0) \approx 0.9225$$

Thus, we find greater than nine-out-of-ten chances that increasing age, blood pressure, and cholestoral increase a patient's odds of obtaining heart disease.

4 Discussion

We can make several inferences from this model relating to factors that contribute to heart disease. As mentioned in Section 3, the summary statistics for the regression parameter of $HOSP1$ in Table 1 indicate that being a patient in the CLE hospital had the effect of decreasing one's chances of having heart disease. Specifically, a 95% credible interval for the factor by which it decreased one's odds of heart disease is $(e^{-1.33}, e^{-0.28})$, i.e. $(0.264, 0.757)$. Thus, being a

patient in the CLE hospital decreased one's odds of having heart disease by anywhere from a factor of one-fourth to three-fourths with 0.95 probability. We can visualize this difference in Figure 2, where we plot the credible intervals for probability of heart disease for each subject. We observe that the credible intervals for the subjects in the LBVA study are more prevalent towards the right-half of the plot, indicating probabilities of heart disease greater than one-half. This agrees with the original proportion of subjects in each study who had heart disease: 45.8% for CLE and 74.5% for LBVA.

In addition, we see that an increase in one unit of $OLDPEAK$ leads to an increase in the odds of heart disease by a factor of $(1.692, 2.672)$ with probability 0.95. Alternatively, the 95% credible interval for the factor *decrease* in the odds of heart disease for a unit increase in the significant predictor $THALACH$ is $(0.964, 0.986)$. Thus, we find that a higher maximum heart rate achieved is most likely an important indicator of *low* chances of having heart disease, while a higher ST depression induced by exercise is an important indicator of *high* chances of having heart disease. As mentioned previously, this agrees with our exploratory findings, where subjects who had heart disease had a higher median $OLDPEAK$ and those who did *not* have heart disease had a higher median $THALACH$.

While providing insightful inferential conclusions, the logistic regression model also provides a good framework for creating predictions. We partitioned the data into a training set (75% of the instances) and a testing set (25%). Posterior predictive distributions were made for each probability of heart disease for each subject in the test set. The predicted disease outcome

was set to be 1 if the posterior predictive median probability was greater than or equal to 0.5, and 0 otherwise. We then compared the predicted disease outcomes in the test set to the true values. The results are shown in Table 2. The Bayesian logistic regression model using posterior predictive median probabilities of having heart disease gave a total error rate 27.0%. We find that the model actually performed better when predicting instances where the subject actually had heart disease (the sensitivity was 20.3%) and performed worse in classifying instances of subjects without heart disease (the specificity was 35.1%). Because the error of misclassifying a subject who *does* have heart disease is most likely worse than the alternative error, a lower sensitivity in this case is indicative of a successful model.

In conclusion, our model was able to identify three significant effects of heart disease, namely the hospital in which the subject is studied, ST depression, and maximum heart rate achieved. Furthermore, we were able to quantify the probabilities that our non-significant regression parameters were positive. Finally, classification proved rather successful, with a total test error rate of 27.0%.

References

[1] Gelman A. Gabry, J. and A. Vetari. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC$*$. 2016.

[2] Pfisterer W. Janosi A., Steinbrunn W. and Detrano R. Heart Disease Databases. 1988.

5 Tables and Figures

Table 1: Summary statistics for the 30,000 post-burn-in draws from the posterior distribution of $\vec{\beta}|\vec{Y}$

	Effect	Mean	SD	2.5%	50%	97.5%	Geweke Statistic
$\beta_0\|\vec{Y}$	(intercept)	0.3759	1.3254	-2.2103	0.3791	2.9800	-1.4832
$\beta_1\|\vec{Y}$	$HOSP1$	-0.8050	0.2686	-1.3325	-0.8046	-0.2780	-1.3219
$\beta_2\|\vec{Y}$	AGE	0.0182	0.0138	-0.0088	0.0182	0.0453	0.3472
$\beta_3\|\vec{Y}$	$TRESTBPS$	0.0113	0.0068	-0.0021	0.0112	0.0248	0.6418
$\beta_4\|\vec{Y}$	$CHOL$	0.0032	0.0022	-0.0012	0.0031	0.0076	0.3365
$\beta_5\|\vec{Y}$	$OLDPEAK$	0.7514	0.1162	0.5259	0.7500	0.9829	-0.0964
$\beta_6\|\vec{Y}$	$THALACH$	-0.0252	0.0057	-0.0364	-0.0251	-0.0141	1.8388

Table 2: Error rates on the testing set using posterior predictive medians

	Predictive Error Rates
Sensitivity ($disease\_ind = 1$)	0.203
Specificity ($disease\_ind = 0$)	0.351
Total	**0.270**

Figure 1: Results of exploratory analysis on the covariates of the logistic regression model

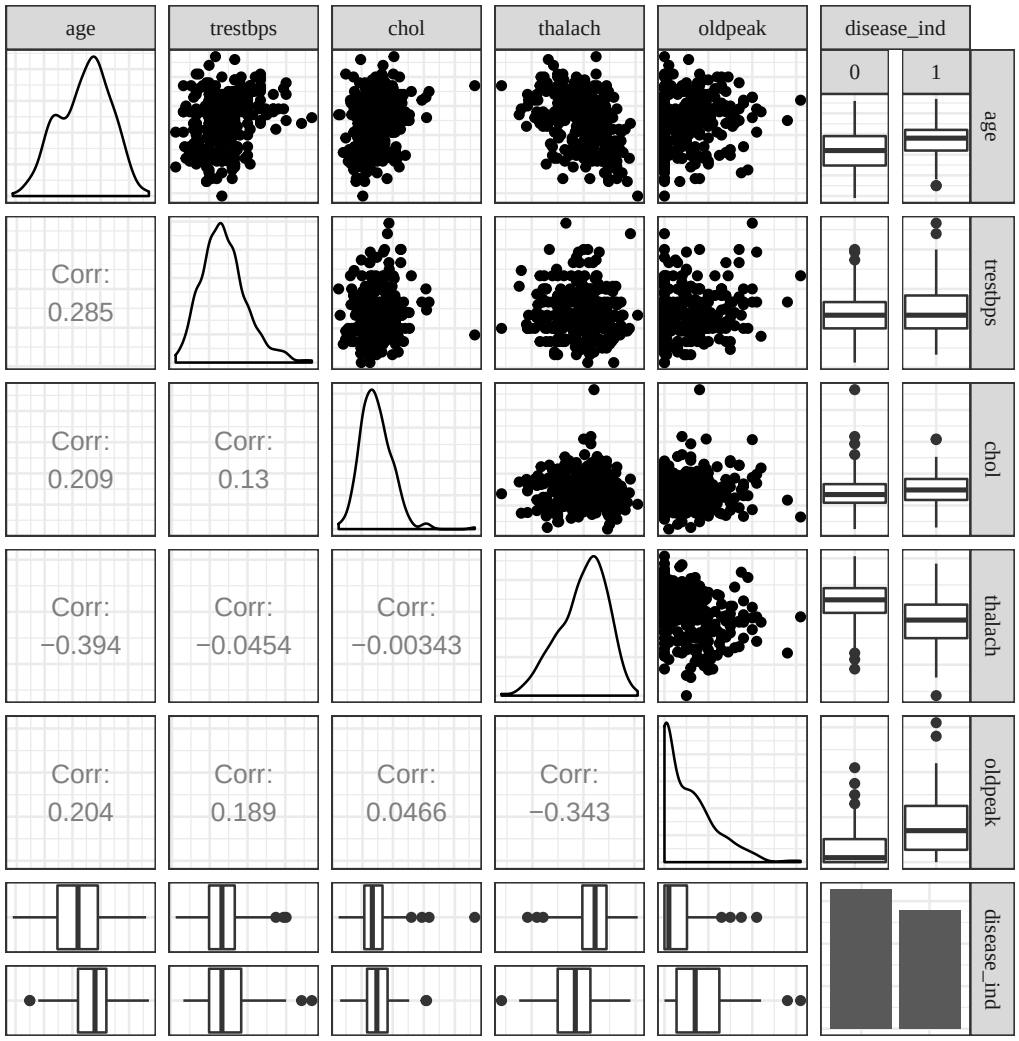

Figure 2: Ninety-five percent credible intervals for each patient's probability of heart disease

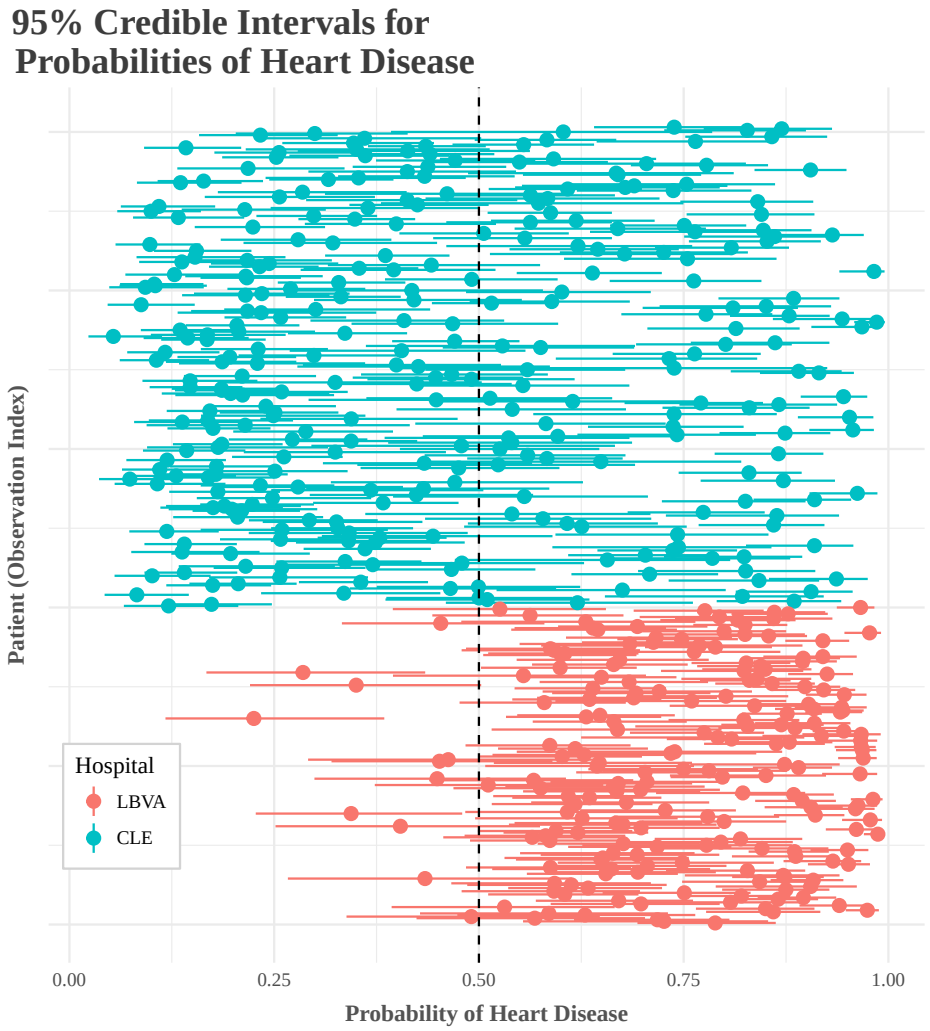

Figure 3: Posterior distributions of β_2, β_3, and β_4