A Bayesian Analysis of Heart Disease Risk

By Daniel Vitale

Introduction

According to the CDC National Center for Health Statistics (https://www.cdc.gov/nchs/), heart disease represents the number one cause of death in the United States. This project provides a rudimentary Bayesian logistic regression model of a heart disease dataset of 1025 patients with metrics on age, sex, chest pain type, resting bp, serum cholesterol, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise induced angina, oldpeak (ST depression induced by exercise relative to rest), the slope of the peak exercise ST segment, number of major vessels (0-3) colored by fluoroscopy, and thal: 0 = normal; 1 = fixed defect; 2 = reversible defect

The motivation for this project is to examine metrics that relate to heart disease across with hopes of building a model that may help predict heart disease with greater accuracy. This project shows the model that has been developed and comments on the relevance of this dataset to the overarching landscape of heart disease.

Model

The model was implemented as follows:

$\mu = \text{logistic}(\beta_0+\beta\chi_1+...+\beta\chi_n)$

$y \sim \text{Bernoulli}(\mu)$

Where:

$\text{logistic}(\chi) = 1/(1+\exp(-\chi))$

Standardization:

$z = (\chi-\bar{x}) / s_\chi$

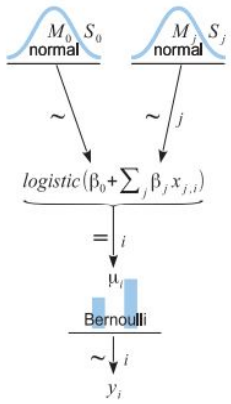


Figure 21.2: Dependency diagram for multiple logistic regression. Compare with the diagram for robust multiple linear regression in Figure 18.4 (p. 498). Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier.

This is a linear combination of metric predictors mapped to a probability value via the logistic function, and the predicted 0's and 1's are Bernoulli distributed around the probability

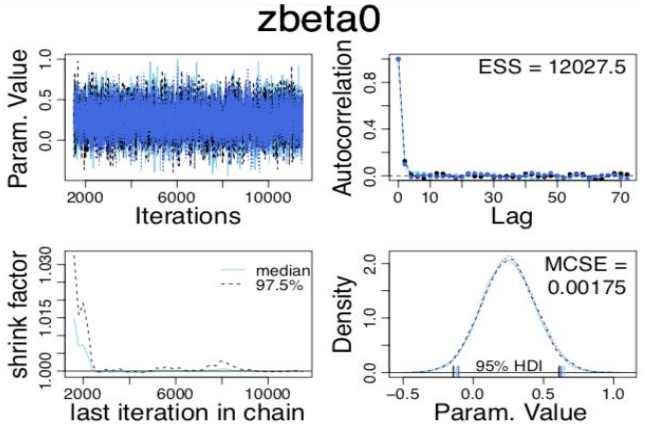
The positive/negative heart disease classes are generated from a bernoulli distribution with parameter mu. Parameter mu is generated from a logistic function with parameters beta0 (bias) and betaN in a logistic (sigmoid) activation function. The parameters beta0 and betaN are assumed to be generated from normal distributions with parameters M (mean) and S (standard deviation). Finally, a standardization is applied to put all of the outputs in a standard scale.

Due to lack of priors, β_0 and β_n were assumed to have been generated by normal distributions with mean = 0 and standard deviation = 1/2^2 (presets of DBDA logistic regression script) though this is likely not the case in real life.

Results

MCMC Diagnostics

The figure to the right provides the following diagnostics for the zbeta0 (unstandardized intercept/bias)



Param value: good overlapping of chains across later iterations

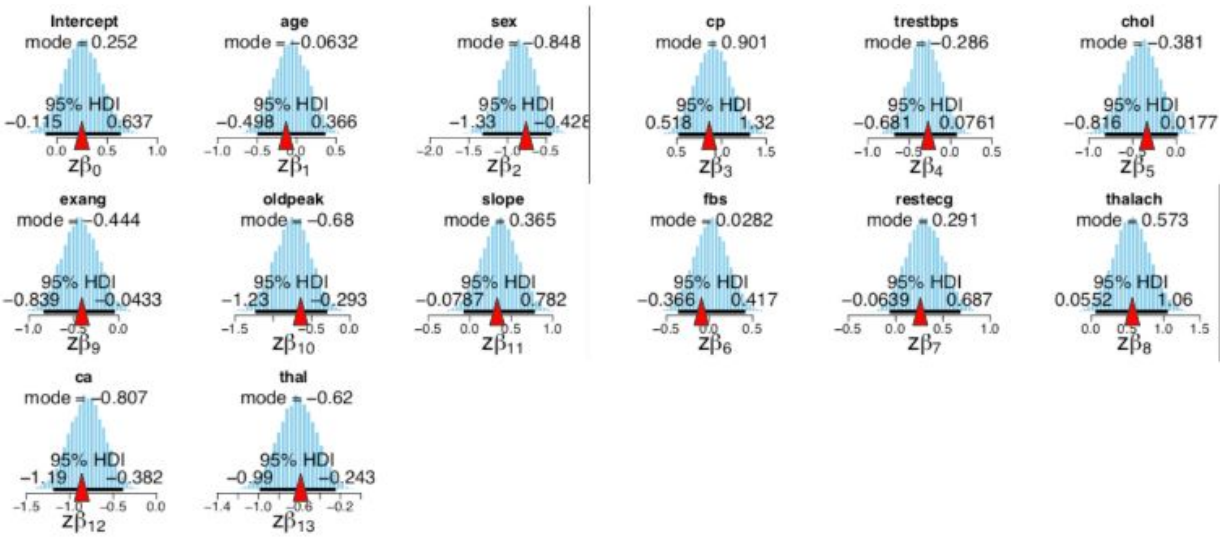
Autocorrelation (cor of chain values with the chain values k steps ahead): very little autocorrelation and ESS > 10,000

Shrink Factor: small shrink factor (< 1.1) indicates adequate convergence

Density: very similar curves for all 3 chains and a very low MCSE means that the chains found very similar posteriors (with similar HDI) and small MCSE indicates stable posterior (low estimation noise)

For brevity, I have chosen to use the bais (zbeta0) as my example because it is highly representative of the diagnostics for the other betas. All showed converging and overlapping probability densities, and very low shrink factor and autocorrelation, and many of the ESS's were above 10,000 and the ones that were not above 10,000 were very close.

Posteriors



Due to the use of generic priors (as opposed to priors of actual heart disease data), this analysis provides a look into the distributions of individual features within this dataset. Notice, the age posterior sits with a median of almost zero, showing very little effect. This is due to the fact that the ages are normally distributed around 65 years of age. This data also shows that the log-odds of heart disease in men (sex=1) versus women (sex=0) is negative, which is accounted for by the fact that 75% of women in this dataset are positive for heart disease and only 45% of men are positive. We see that the log-odds of heart disease increases with chest pain (cp) increase, which we would certainly expect. We see that resting blood pressure (trestbps) is slightly negatively related with heart disease in this data, but the resting blood pressure values are normally distributed on the high end, well above suggested blood pressure. The same effect occurs for cholesterol (chol), where the mean cholesterol value for patients with heart disease is 242.2303 and mean cholesterol value for patients without heart disease is 251.087. The heart disease patients actually have lower average cholesterol, which accounts for the effect we see in the log-odds of heart disease decreasing with increasing cholesterol. Overall, this dataset is highly biased towards unhealthy measures relating to heart disease and other factors.

Conclusions

Bias in this dataset along with a lack of priors has resulted in the production of a model that paints a highly inaccurate picture of the metrics of heart disease. This project is a testament to the necessity for more concrete priors for the parameters of a model such as this one. Going forward, the literature on heart disease biometrics metrics should be

thoroughly combed to establish reasonable prior distributions. Some interesting areas to examine would be differences in biometrics by sex, i.e. running separate models for different sexes as well as running a bayesian logistic regression model to predict sex.

Sources

Kruschke, J. K. (2014). Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition. Academic Press / Elsevier

CDC - NCHS - National Center for Health Statistics. (n.d.). Retrieved from https://www.cdc.gov/nchs/