Bayesian Logistic Regression for Heart Disease

By Daniel Vitale

Introduction

According to the CDC National Center for Health Statistics, heart disease represents the number one cause of death for both men and women in the United States.

The objective of this project is to produce a predictive model and examine the potential relevance of a handful of metrics related to heart disease in male and female patients

UCI Heart Disease Dataset

https://archive.ics.uci.edu/ml/datasets/Heart+Disease

The dataset contains metrics from 1025 individual patients including the following:

1. age
2. sex
3. chest pain type (4 values)
4. resting blood pressure
5. serum cholesterol in mg/dl
6. fasting blood sugar > 120 mg/dl
7. resting electrocardiographic results (values 0,1,2)
8. maximum heart rate achieved
9. exercise induced angina
10. oldpeak = ST depression induced by exercise relative to rest
11. the slope of the peak exercise ST segment
12. number of major vessels (0-3) colored by fluoroscopy
13. thal: 0 = normal; 1 = fixed defect; 2 = reversible defect

Bayesian Logistic Regression Model

The model was implemented as follows:

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

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

Where:

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

Standardization:

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

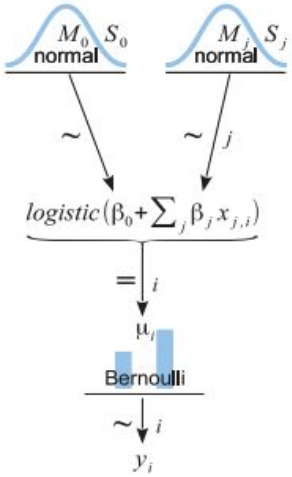
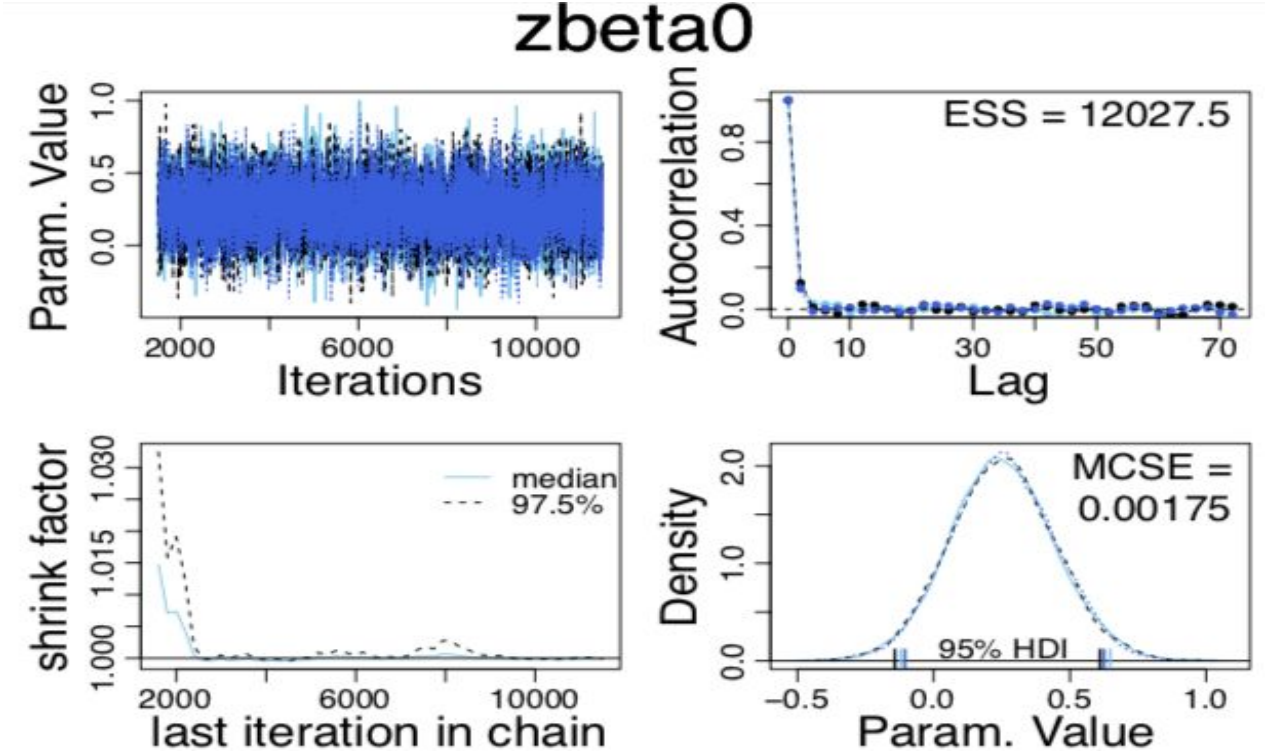


Figure 21.2: Dependency diagram for multiple logistic regression. Compare with the diagram for robust multiple linear regression in Figure 18.4 (p. 498). Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier.

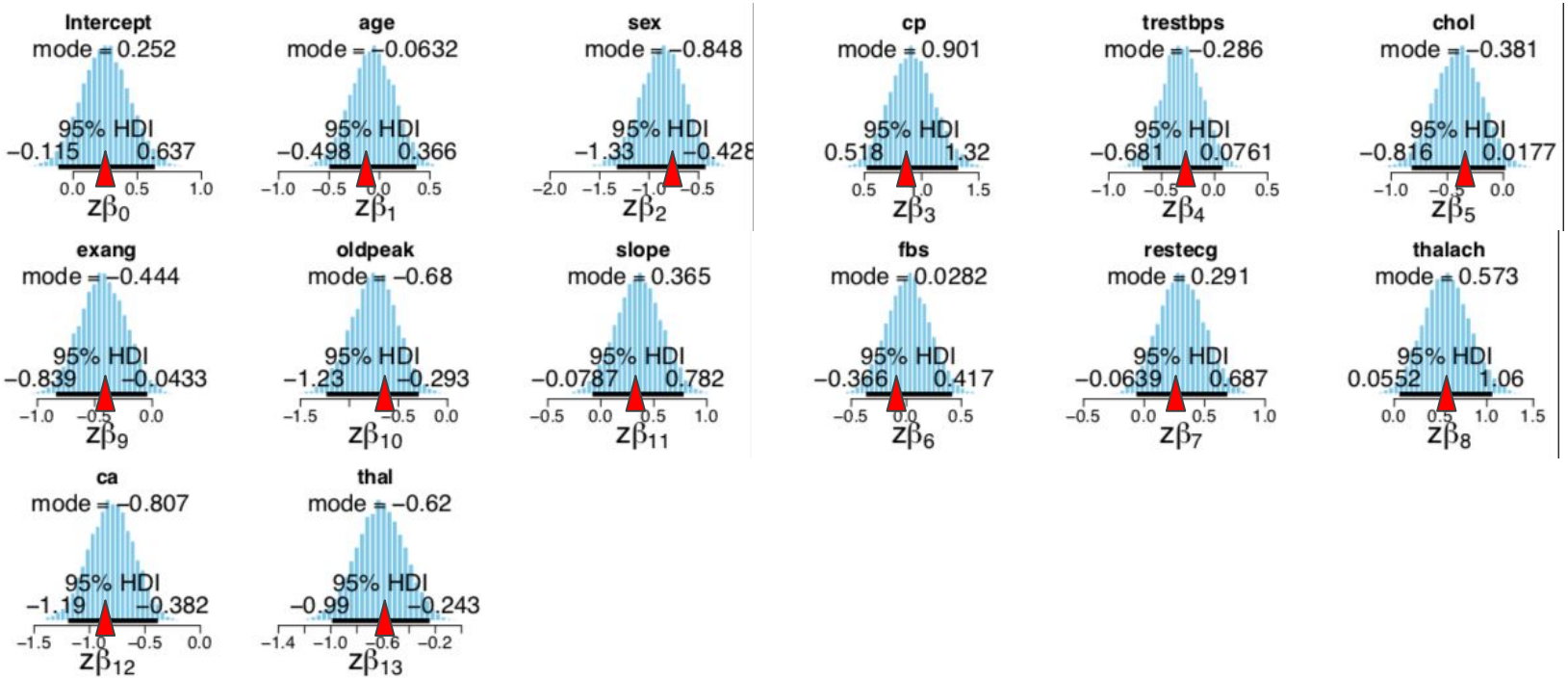
Bayesian Logistic Regression Model (continued)

- 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
- This will have effects on the posteriors that we generate!

Results: Metropolis Algorithm Diagnostics

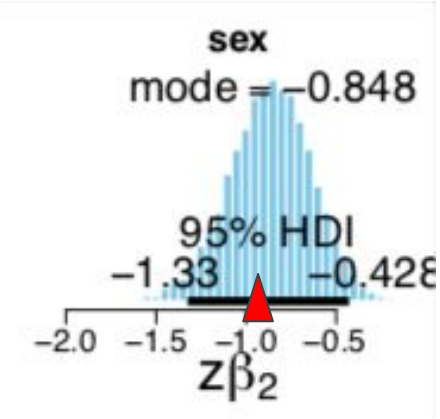


Results


Intercept
mode = 0.252
95% HDI
−0.115
0.637
zβ0
age
mode = −0.0632
95% HDI
−0.498
0.366
zβ1
sex
mode = −0.848
95% HDI
−1.33
−0.428
zβ2
cp
mode = 0.901
95% HDI
0.518
1.32
zβ3
trestbps
mode = −0.286
95% HDI
−0.681
0.0761
zβ4
chol
mode = −0.381
95% HDI
−0.816
0.0177
zβ5
exang
mode = −0.444
95% HDI
−0.839
−0.0433
zβ9
oldpeak
mode = −0.68
95% HDI
−1.23
−0.293
zβ10
slope
mode = 0.365
95% HDI
−0.0787
0.782
zβ11
fbs
mode = 0.0282
95% HDI
−0.366
0.417
zβ6
restecg
mode = 0.291
95% HDI
−0.0639
0.687
zβ7
thalach
mode = 0.573
95% HDI
0.0552
1.06
zβ8
ca
mode = −0.807
95% HDI
−1.19
−0.382
zβ12
thal
mode = −0.62
95% HDI
−0.99
−0.243
zβ13

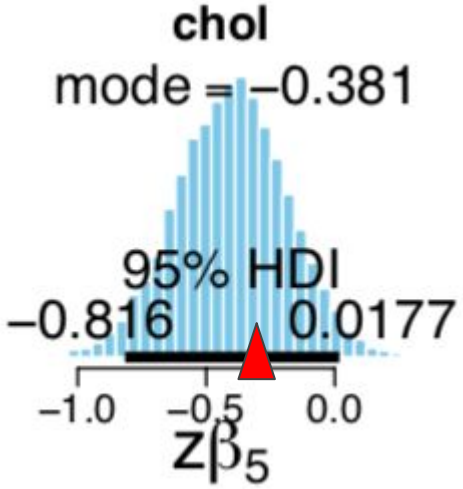
Discussion- Biological Sex and Heart Disease

- HDI for zβ falls entirely in the negative
- Log-odds of heart disease in men (sex=1) versus women (sex=0) is negative
- This is accounted for bias in the dataset:
  - 45% of men have heart disease
  - 75% of women have heart disease



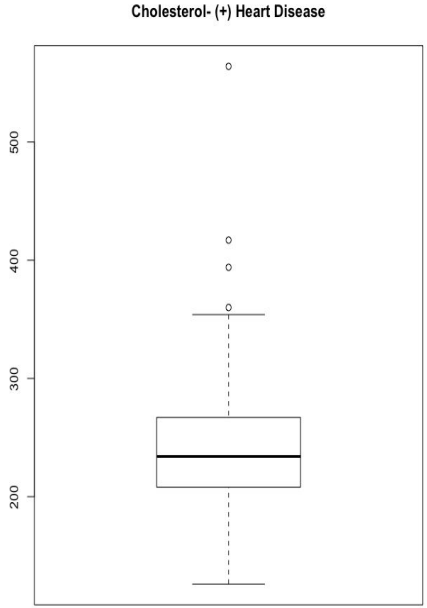
Discussion- Cholesterol and Heart Disease

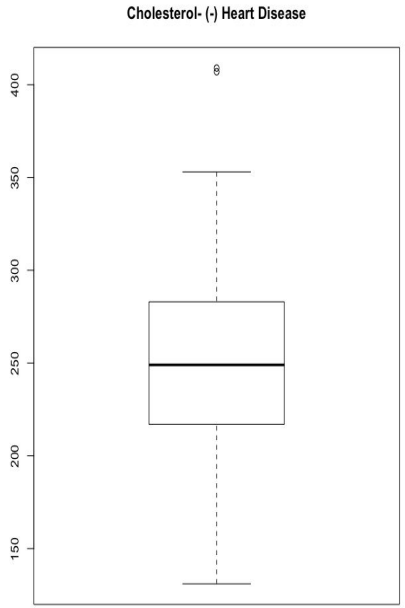
- HDI for zβ falls almost entirely in the negative
- Log-odds of heart disease and total cholesterol have negative relationship



Discussion- Cholesterol and Heart Disease (continued)

- Dataset is biased for high cholesterol
- Cholesterol means:
  - + for Heart Disease: 242.2303
  - - for Heart Disease: 251.087





Conclusions

- Bias in data matters!
- Priors matter!
  - Without including the prior distributions for specific features (cholesterol and sex) I have very likely produced a very overfit model!
  - I would expect the results to be much different for cholesterol and sex if priors were included (but I ran out of time!)

Going Forward

- Examination of sex individually
- Examination of this dataset with the inclusion of prior distributions of features

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/