Evaluating Binary Medical Data Using a Bayesian Lasso Approach

Vindyani Herath

Department of Mathematics And Statistics
Sam Houston State University

March 26, 2021

Sam Houston State University

Introduction

Cardiovascular disease (CVD) is a general term for conditions affecting the heart or blood vessels.

CVDs are the number 1 cause of death globally: more people die annually from CVDs than from any other cause.

However, it can often largely be prevented by leading a healthy lifestyle.

Causes of CVDs

- High blood pressure
- Smoking
- High cholesterol
- Diabetes
- Inactivity
- Being overweight or obese
- Family history of CVDs
- Ethnic background

Dataset Description

70,000 observations with 11 predictor variables and binary response variable

- Presence or absence of cardiovascular disease – Response variable
- Age
- Height
- Weight
- Gender
- Systolic Blood Pressure
- Diastolic Blood Pressure
- Cholesterol
- Glucose
- Smoking
- Alcohol Intake
- Physical Activity

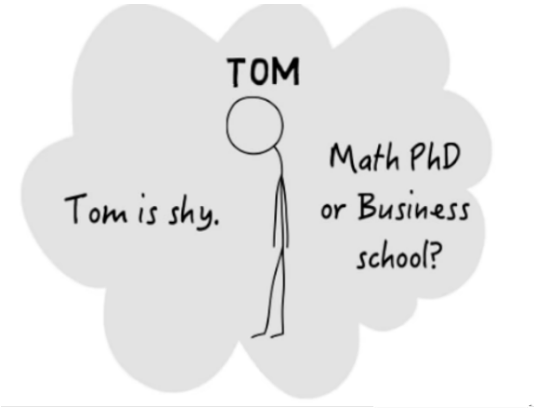
Objectives of the Study

- Predict the binary outcome (whether the patient developed cardiovascular disease)
- Dimension reduction for future model if the column size of the design matrix is too large (Variable selection)
- Bayesian thinking

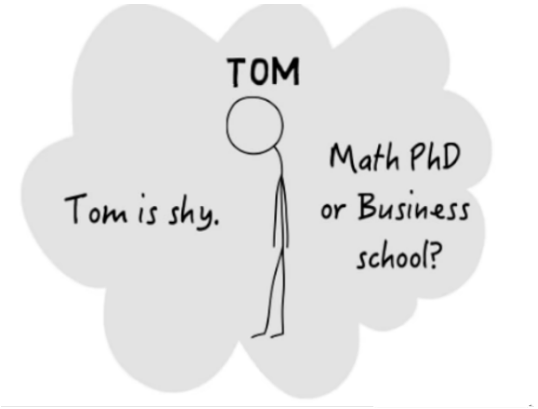
That leads to develop a hierarchical Bayesian lasso model using Gibbs sampler.

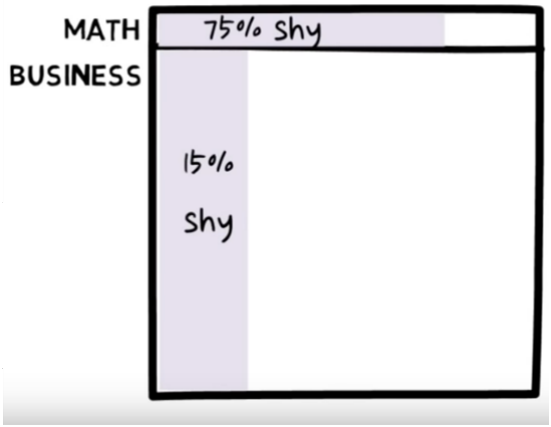
Bayesian Thinking

Bayesian Thinking

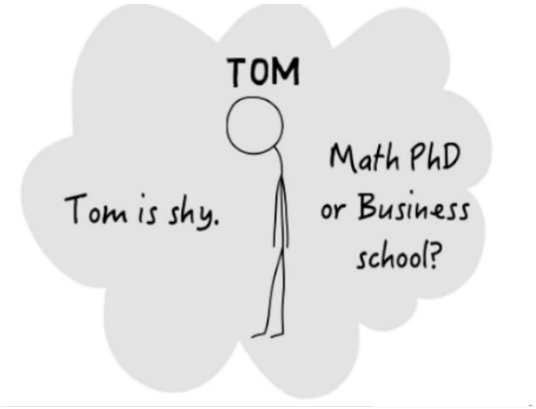


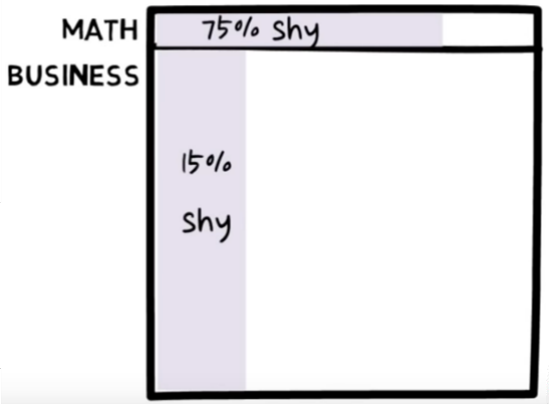
Bayesian Thinking





Bayesian Thinking





MATH : BUSINESS

Prior odds ratio 1 : 10

Likelihood ratio 75 : 15

1 : 2

Variable Selection in Regression

Identifying the best subset of predictors to include in the model

- Stepwise Selection
- Backward Selection

When number of predictors is greater than number of observations?

- Refer $\hat{\beta} = (X'X)^{-1}X'Y$
- A different technique should be used when $p >> n$

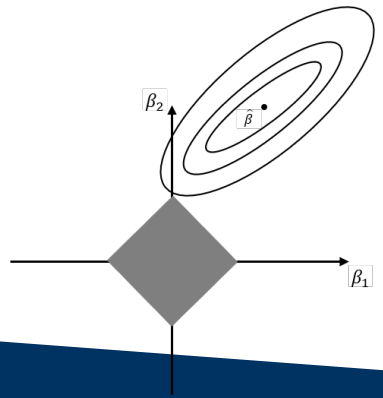
Lasso (Least Absolute Shrinkage and Selection Operator)

Proposed by Tibshirani for shrinkage and selection for regression problems

Uses L_1 regularization technique

$$(y - x\beta)^{'}(y - x\beta) + \lambda \sum_{j=1}^{d} |\beta|$$

Gives greater prediction accuracy and increase model interpretability



Bayesian Lasso

The Bayesian lasso provides valid standard errors for β and provides more stable point estimates by using the posterior center.

The lasso estimates can be estimated as posterior mode estimates when the regression parameters have independent and identical Laplace (i.e., double exponential) priors.

$$\pi(\beta) = \frac{\lambda}{2} exp(-\lambda)|\beta_j|$$

The Bayesian lasso estimates (posterior center) appear to be a compromise between the ordinary lasso and ridge regression.

Gibbs Sampling

A MCMC method for simulating a random sample from a multivariate posterior.

Generates posterior samples by sweeping through each variable to sample from its conditional distribution with the remaining variables fixed to their current values.

Methodology

Posterior: $f(\theta|y) \propto f(y|\theta)\, f(\theta)$

$$f(y|z)\, f(z|\beta)\, f(\beta|\tau)\, f(\tau|\lambda).\pi(\lambda)$$

$$y_i = \mathbb{I}_{(0,\infty)}(z_i)$$

$$z_i|\beta \sim \mathcal{N}(x'\beta, 1)$$

$$\beta|\tau_1^2, ..., \tau_d^2 \sim \mathcal{N}_d(0_d, D_\tau), \text{where } D_\tau = \text{diag}(\tau_1^2, \tau_2^2, ..., \tau_d^2)$$

$$\tau_1^2, \tau_2^2, ..., \tau_d^2 \sim \exp\left(\frac{\lambda^2}{2}\right)$$

$$\lambda \sim \pi(\lambda) \propto C$$

$$\tau_1^2, \tau_2^2, ..., \tau_d^2 > 0$$

Gibbs Sampling Structure (Marginal Distributions)

$$Z_i|\tau,\beta,\lambda,y_i = [TN(x'\beta,1,0,\infty]^{y_i} \times [TN(x'\beta,1,-\infty,0]^{(1-y_i)}$$

$$\beta|\tau,Z,\lambda,y_i \sim \mathcal{N}(A^{-1}x'Z,A^{-1}), \text{where } A = D_\tau^{-1} + X'X$$

$$\tau^2|\ Z,\beta,\lambda,y \propto \phi(\beta,0,D_\tau) \times \prod_{i=1}^{d} \frac{\lambda^2}{2} e^{\frac{-\lambda^2\tau_j^2}{2}}$$

$$\tau_j^{-2}|\ z,\beta,\lambda,y \sim \mathcal{IG}(\mu_j,\lambda^2) \text{with } \mu_j = \lambda|\beta_j|^{-1}$$

$$\lambda|\tau,\beta,z,y \propto \prod_{j=1}^{d} \frac{\lambda^2}{2} e^{\frac{-\lambda^2\tau_j^2}{2}} \times \pi(\lambda) \text{ where } \lambda|\tau,\beta,z,y \sim \text{Gamma}\left(d+1, \sum_{j=1}^{d} \tau_j^2\right)$$

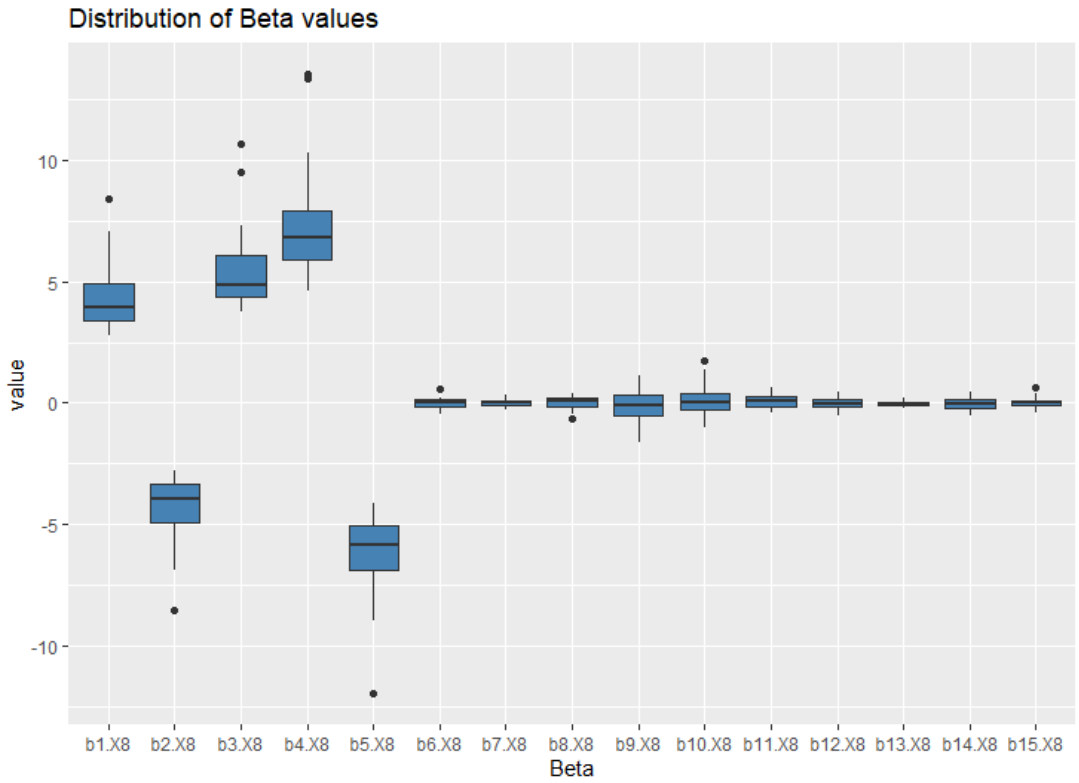
Sam Houston State University

Simulation Study

$$X_1 \sim N(n, 3, 1)$$
$$X_2 \sim N(n, X_1, 1)$$
$$X_3 \sim N(n, X_2, 2)$$
$$X_4 \sim U(n, 5, 10)$$
$$X_5 \sim U(n, X_4,\ X_4 + 3)$$
$$X_6 \sim N(n, 3.5, 1)$$
$$X_7 \sim N(n, X_6, 1)$$
$$X_8 \sim N(n, 5.2, 2)$$
$$X_9 \sim U(n, X_8, X_8 + 3)$$
$$X_{10} \sim U(n, X_9, X_9 + 1)$$
$$X_{11} \sim N(n, 5, 1)$$
$$X_{12} \sim N(n, X_{11}, 1)$$
$$X_{13} \sim N(n, X_{12}, 2)$$
$$X_{14} \sim U(n, 5, 10)$$
$$X_{15} \sim U(n, X_{14},\ X_{14} + 3)$$

Variable	True β	$\hat{\beta}$
X_1	4	3.959
X_2	-4	-4.100
X_3	5	5.184
X_4	7	6.985
X_5	-6	-5.939
X_6	0	-0.308
X_7	0	0.312
X_8	0	-0.107
X_9	0	0.773
X_{10}	0	-0.737
X_{11}	0	-0.449
X_{12}	0	0.063
X_{13}	0	0.131
X_{14}	0	0.334
X_{15}	0	-0.097

Distribution of β values



Cardiovascular Disease Data Analysis

Covariate	$\hat{\beta}$
Age	0.2778
Gender	-0.0861
Height	-0.0229
Weight	0.2135
Systolic blood pressure	0.1665
Diastolic blood pressure	0.0750
Cholesterol	0.3377
Glucose	-0.1318
Smoke	-0.0262
Alcohol intake	-0.0468
Physical activity	-0.1969

Test Data	Predicted Response Using Bayesian Lasso	Predicted response using another method
1	1	1
0	1	0
1	0	0
0	0	1
1	0	1
0	1	0
⋮	⋮	⋮
⋮	⋮	⋮
Prediction Accuracy	64.72%	49.85%

Conclusion

Bayesian lasso technique can be used for variable selection in high-dimensional data sets.

The proposed hierarchical Bayesian lasso model provides good prediction to identify the patients with cardiovascular disease.

Thank You !