Study of Variable Selection on Vertebral Column Disorder Data

Himadri Sekhar Manna [1] Ayoushman Bhattacharya [2] Ayan Paul [3]

[1]MS - 1907

[2]MS - 1904

[3]MS - 1903

May 20, 2021

Outline

Outline

Outline

Outline

Outline

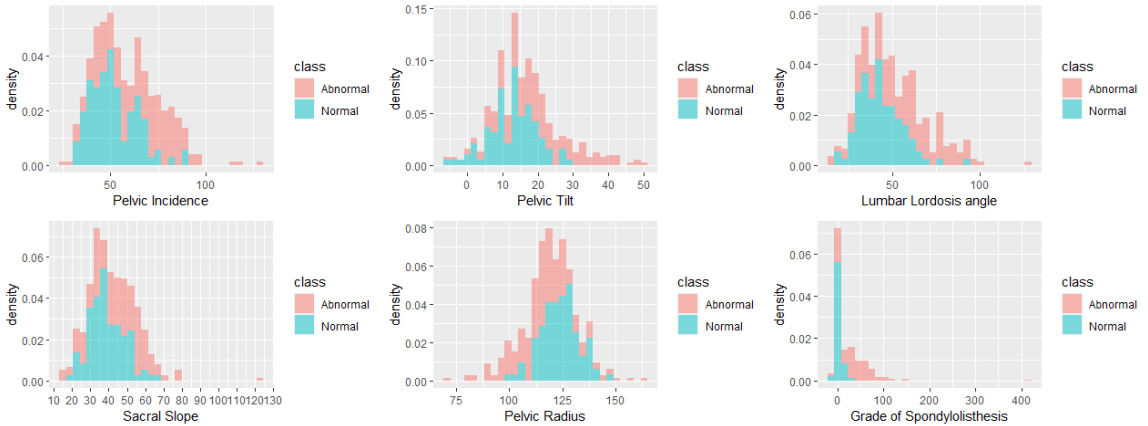
Background

- Variable selection is an important step of building models,
  1. Selecting appropriate set of variables helps in explaining the data in the simplest way — redundant explanatory variables should be removed.
  2. Selecting appropriate set of variables results in lesser computational time.
- The aim is to construct an optimal model.
- **Classical** model selection techniques help us in developing models with good predictive power but don't help in understanding the importance.
- **Bayesian** model selection techniques help us in understanding the importance of variables and as well in finding model with good predictive power.

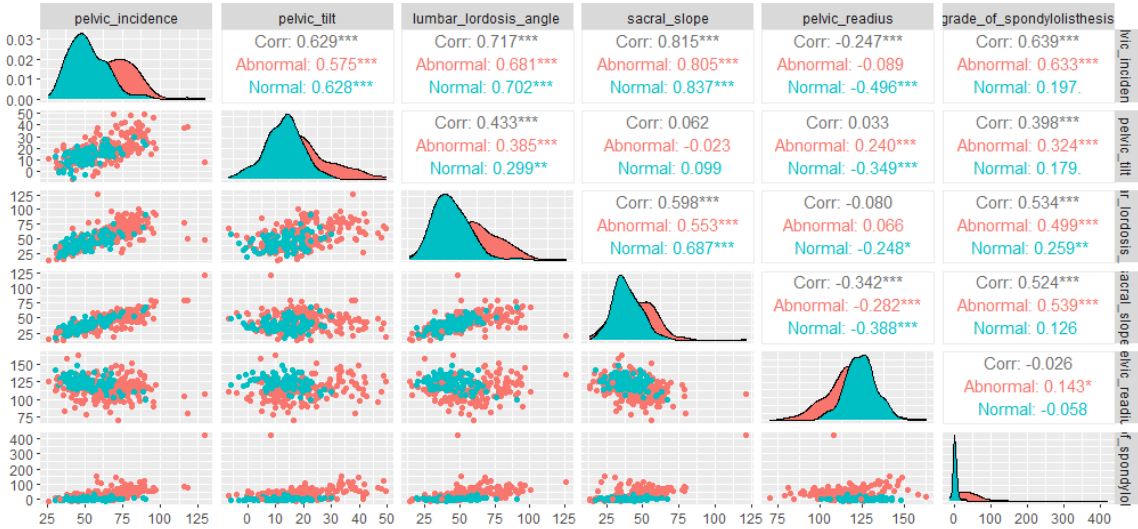
Data Description

- The Biomedical data set of our interest was built by **Dr. Henrique da Mota** during a medical residence period in Lyon, France.
- In the datset the patients are classified in the following two classes,
  1. Abnormal: **Disk Hernia** and **Spondylolisthesis** patients.
  2. Normal: Rest of the patients are coined as normal.
- For each patient we have observation on 6 biomechanical attributes derived from the shape and orientation of the pelvis and lumbar spine, those are,
  1. *pelvic incidence*
  2. *pelvic tilt*
  3. *lumbar lordosis angle*
  4. *sacral slope*
  5. *pelvic radius*
  6. *grade of spondylolisthesis*

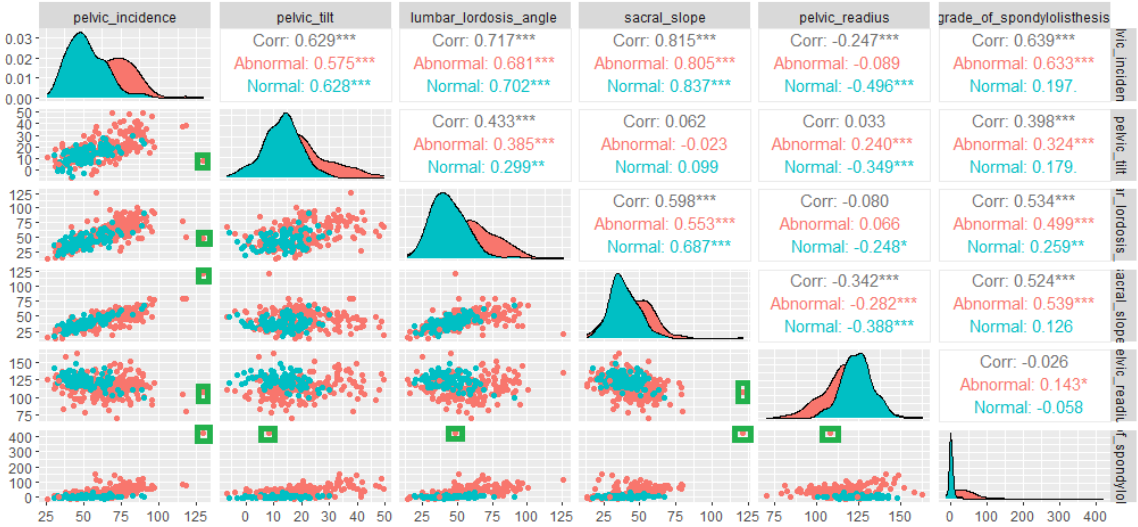
Exploratory Data Analysis



Exploratory Data Analysis



Exploratory Data Analysis



Exploratory Data Analysis

The histograms provided evidence towards existence of outlier in the dataset. The point indicated by the green box in the above plot is the outlier.

Train-Test Split

- We removed the **outlier**, i.e., the $\mathbf{116}^{th}$ observation from the dataset.
- From the remaining dataset, we considered 240 observations in the training set and the remaining observations in our test set to get approximately a 3:1 split up of the dataset into training and test respectively.

Generalized Linear Model with Probit Link

- Let Y_i for $i = 1, 2, \ldots, n$, be independent binary random variables.

$$\mathbf{E}(Y_i|\mathbf{X}) = \mathbf{P}(Y_i = 1|\mathbf{X}) = p_i$$

- Then, we define the binary regression model as,

$$p_i = g^{-1}(\beta_0 + \sum_{j=1}^{p} \beta_j x_{i,j})$$

- In Probit Regression, we take,

$$g^{-1}(\cdot) = \Phi(\cdot)$$

$$\text{i.e.} \quad \Phi^{-1}(p_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{i,j}$$

Generalized Linear Model with Probit Link-Data Augmentation

- Let us introduce n independent latent variables z_i, where each z_i follows a normal distribution. Then the augmented probit model can be written as,

$$y_i = \begin{cases} 1 & \text{, if } z_i \geqslant 0 \\ 0 & \text{, if } z_i < 0 \end{cases} \quad \text{, for all } i = 1, 2, \ldots, n.$$

with,

$$z_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{i,j} + \epsilon_i$$

$$\text{where,} \quad \epsilon_i \overset{\text{iid}}{\sim} N(0,1) \quad \text{, for all } i = 1, 2, \ldots, n.$$

- Therefore, we formulate the original problem as a missing data problem where we have a normal regression model on the latent variable z_i and the observed responses y_i are incomplete in that we only observe whether $z_i \geqslant 0$ or $z_i < 0$.

Classical Model Selection

Probit Regression-Backward Selection using BIC

- *sacral slope,pelvic radius and grade of spondylolisthesis* are the selected variables using information criteria BIC in backward selection.

Table: Coefficient Estimates of the Model selected using Backward Selection

Coefficient	Estimate
β_0	1.6009
β_4	-0.8726
β_5	-0.9249
β_6	2.6679

- Misclassification error rate for the training data is 13.33% .
- Misclassification error rate for the test data is 18.84% .

Probit Regression-Backward Selection using AIC

- *pelvic incidence, sacral slope, pelvic radius and grade of spondylolisthesis* are the selected variables using information criteria AIC in backward selection.

Table: Coefficient Estimates of the Model selected using Backward Selection

Coefficient	Estimate
β_0	1.6528
β_3	0.5596
β_4	-1.2673
β_5	-0.8208
β_6	2.6359

- Misclassification error rate for the training data is 13.75% .
- Misclassification error rate for the test data is 17.39% .

Drawbacks in classical approach

- Different model selection criterion results in different model and hence raises the question of model selection.
- We can't conclude anything on the importance of the predictor *pelvic incidence* in the model.
- The above issue can be dealt easily if we had any measure on variable importance.
- We now implement Bayesian model selection approach which can provide us relative importance of each covariates in terms of posterior inclusion probability.

Bayesian Model Selection using Indicator Variable

Model Formulation

- We define the predictor inclusion indicators,

$$\gamma_j = \begin{cases} 1 & \text{, if } j^{th} \text{ predictor is included in the model} \\ 0 & \text{, o.w.} \end{cases} \quad \text{, for all } j = 1, 2, \ldots, p.$$

- We redefine the model for latent variables,

$$z_i = \beta_0 + \sum_{j=1}^{p} \gamma_j \beta_j x_{i,j} + \epsilon_i \quad \text{, for all } i = 1, 2, \ldots, n.$$

and we assume,

$$\epsilon_i \overset{\text{iid}}{\sim} N(0, 1) \quad \text{, for all } i = 1, 2, \ldots, n.$$

Model Hierarchy

So, we have the following model hierarchy to run the Gibbs sampler and subsequent calculations

$$z_i|\mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\gamma} \overset{\text{ind}}{\sim} N\left(\beta_0 + \sum_{j=1}^{p} \gamma_j \beta_j x_{i,j}, 1\right) \quad \text{, for all } i = 1, 2, \ldots, n.$$

$$y_i|\mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\gamma} \overset{\text{ind}}{\sim} \text{Bernoulli}(p_i) \quad \text{, with } p_i = \Phi\left(\beta_0 + \sum_{j=1}^{p} \gamma_j \beta_j x_{i,j}\right) \quad \text{, for all } i = 1, 2, \ldots, n.$$

$$\boldsymbol{\beta} \sim N_{p+1}(\mathbf{0}, \Sigma) \quad \text{, where } \Sigma = \text{Diag}(1000, \ldots, 1000).$$

$$\gamma_j \overset{\text{iid}}{\sim} \text{Bernoulli}\left(\frac{1}{2}\right) \quad \text{, for all } j = 1, 2 \ldots, p.$$

Gibbs Sampler and MCMC

- We run a Gibbs sampling algorithm based on the hierarchical model we mentioned earlier to generate $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ from their respective posterior distributions.
  1. We generate 5.1×10^5 MCMC iterations.
  2. We use the first 10000 iterations as "burn-in".
  3. Based on the autocorrelation plots, we thin the chains by saving every 200^{th} observation.
- We find out posterior mean, median and 95% credible intervals for each parameter.
- We also find out posterior **marginal inclusion probability** for each covariate and the posterior probability of each unique models.

Posterior Inclusion Probability

Table: Marginal posterior inclusion probability of the covariates

Covariate	Posterior Probability
1	0.1156
2	0.0896
3	0.0076
4	0.9564
5	1
6	1

- *Sacral slope*, *pelvic radius* and *grade of spondylolisthesis* are the most important variables.

Posterior Model Probabilities

Table: Posterior Probability of Each Unique Model Obtained by MCMC Iterations

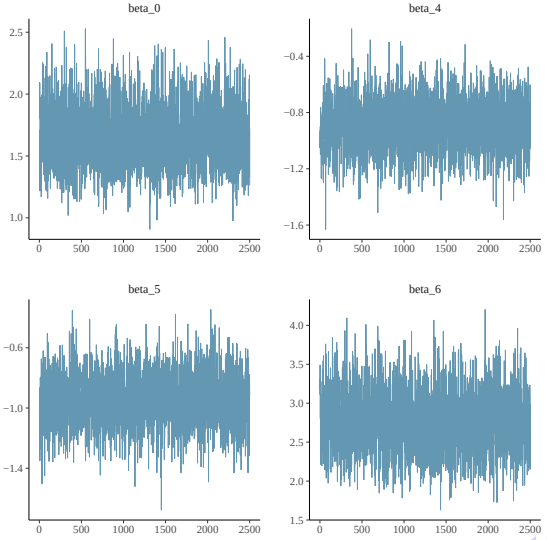
Model Covariates	Posterior Probability
0[1], 5, 6	0.0004
0, 4, 5, 6	0.8552
0, 3, 5, 6	0.0004
0, 2, 4, 5, 6	0.0220
0, 1, 4, 5, 6	0.0480
0, 1, 2, 5, 6	0.0424
0, 3, 4, 5, 6	0.0060
0, 1, 2, 4, 5, 6	0.0244
0, 2, 3, 4, 5, 6	0.0004
0, 1, 2, 3, 5, 6	0.0004
0, 1, 3, 4, 5, 6	0.0004

[1] 0 represents a model with an intercept term.

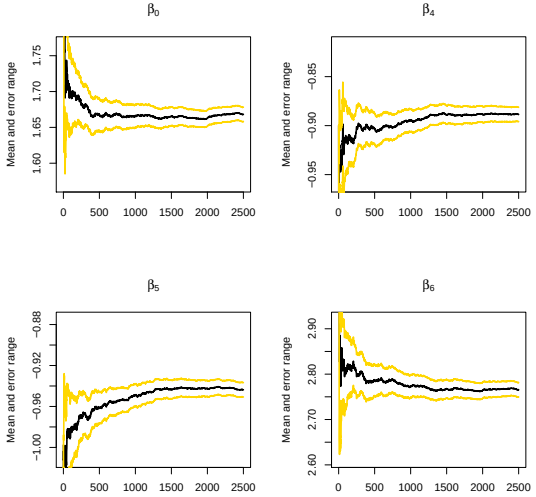
Selected Model

- We choose the model with highest posterior probability.
- Both type of posterior probabilities result in same model.
- The final model contains **three** covariates *sacral slope*, *pelvic radius* and *grade of spondylolisthesis* with an *intercept* term.

Trace Plot of Coefficients



Cumulative Mean Plot of Coefficient Estimates



Inference on posterior distribution

Table: Summary of posterior distribution

Coefficient	Posterior Mean	Posterior Median	95% HPD Credible Interval
β_0	1.6681182	1.6562169	[1.222977, 2.1860137]
β_4	-0.8885347	-0.8906949	[-1.255135, -0.5369650]
β_5	-0.9437431	-0.9438566	[-1.296979, -0.6201231]
β_6	2.7654438	2.7527849	[1.969314, 3.5006822]

Inference on posterior distribution

- The trace plots and cumulative mean plots suggest that the Gibbs sampling algorithm converges.
- Posterior mean and median are very close to each other for each coefficient.
- Any 95% HPD credible interval does not contain zero, so none of these coefficients can be dropped from the selected model.
- We will use posterior mean as the estimate of the regression coefficient for prediction.

Misclassification Error in Prediction

Table: Misclassification error

(a) Confusion matrix of the train data

Predicted Class	*Actual Class*	
	Normal	**Abnormal**
Normal	58	18
Abnormal	14	150

(b) Confusion matrix of the test data

Predicted Class	*Actual Class*	
	Normal	**Abnormal**
Normal	18	4
Abnormal	9	38

- Misclassification error rate for the train data is 13.33% .
- Misclassification error rate for the test data is 18.84% .

Bayesian Model Selection using Lasso

Model Formulation

- We define the model for latent variables,

$$z_i = \mathbf{x}_i'\boldsymbol{\beta} + \epsilon_i \quad \text{, for all } i = 1, \ldots, n.$$

  where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_6)'$ denotes the regression coefficient vector for the six predictor variables and β_0 denotes the intercept term.

- Consequently, we consider the first column of the design matrix $\mathbf{X}$ as a $n \times 1$ vector of all 1's.

- We also assume,

$$\epsilon_i \overset{\text{iid}}{\sim} N(0, 1) \quad \text{, for all } i = 1, \ldots, n.$$

Bayesian Lasso

- The Lasso is a method for simultaneous shrinkage and model selection in regression problems.
- In our case, the Lasso estimate $\hat{\boldsymbol{\beta}}$ can be described as solutions to unconstrained optimisation of the form

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^{p+1}}{\arg\min}\left(\sum_{i=1}^{n}\left(z_i-\mathbf{x}_i'\boldsymbol{\beta}\right)^2+\lambda\sum_{j=0}^{p}|\beta_j|\right)\quad;\quad\lambda\geqslant 0. \tag{1}$$

- The form of this expression suggests that the Lasso may be interpreted as a Bayesian posterior mode estimate when the parameters β_j have i.i.d. double exponential priors.

Prior Selection

- We use a prior on $\boldsymbol{\beta}$ of the form

$$\pi(\boldsymbol{\beta} \mid \lambda) = \prod_{j=1}^{p} \frac{\lambda}{2} \exp\left(-\lambda|\beta_j|\right). \tag{2}$$

- But under this prior, the posterior distribution becomes very complicated. So to develop a Gibbs sampler, we use a mixture representation of the double exponential distribution.

Double Exponential as scale mixture of Normal (Park and Caella, 2008)

- We exploit the following representation of the double exponential distribution as a scale mixture of normal:

$$\frac{a}{2}\exp(-a|z|) = \int_0^\infty \frac{1}{\sqrt{2\pi s}}\exp\left(-\frac{z^2}{2s}\right)\frac{a^2}{2}\exp\left(-\frac{a^2 s}{2}\right)\, ds, \tag{3}$$

  where $a > 0$.

- So, it is evident that, $z \mid s \sim N(0, s)$ and $s \sim$ Exponential with mean $\frac{2}{a^2}$.

Model Hierarchy

- So, we have the following model hierarchy to run the Gibbs sampler and subsequent calculations

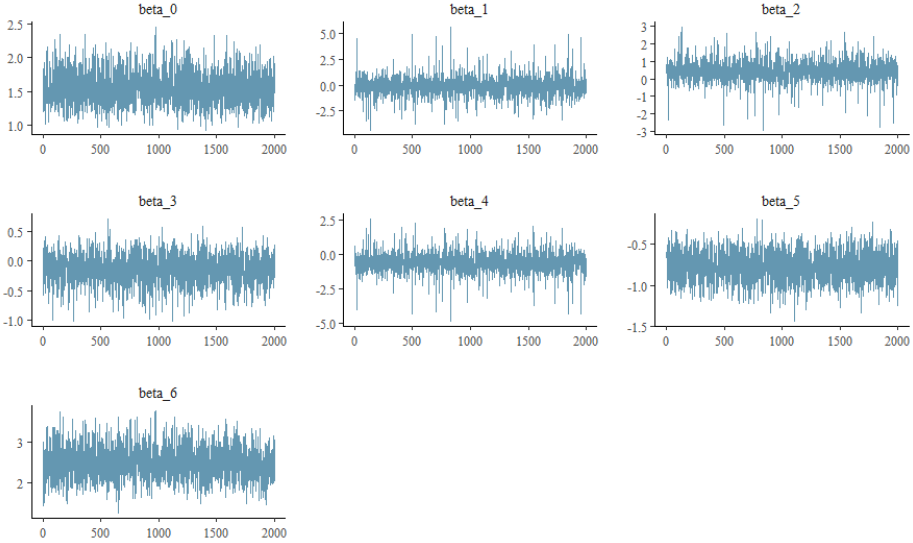
$$
\begin{aligned}
&y_i \mid \mathbf{X}, \boldsymbol{\beta} \overset{\text{ind}}{\sim} \text{Bernoulli}(p_i), \quad p_i = \Phi\left(\mathbf{x}_i'\boldsymbol{\beta}\right),\ i = 1, \ldots,\ n, \\
&z_i \mid \mathbf{X}, \boldsymbol{\beta} \overset{\text{ind}}{\sim} N\left(\mathbf{x}_i'\boldsymbol{\beta}, 1\right),\ i = 1, \ldots,\ n, \\
&\boldsymbol{\beta} \mid \tau_0^2, \ldots,\ \tau_p^2 \sim N_{p+1}\left(\mathbf{0},\ \mathbf{D}_\tau\right),\ \mathbf{D}_\tau = diag\left(\tau_0^2, \ldots,\ \tau_p^2\right), \\
&\tau_0^2, \ldots,\ \tau_p^2 \mid \lambda^2 \sim \prod_{j=0}^{p} \frac{\lambda^2}{2} \exp\left(-\frac{\lambda^2 \tau_j^2}{2}\right), \\
&\lambda^2 \sim \text{Gamma}\left(\text{shape} = a,\ \text{rate} = b\right), \qquad a > 0 \text{ and } b > 0 \text{ are known.}
\end{aligned}
$$

- Here, we consider the values of the hyperparameters as $a = b = 0.5$.

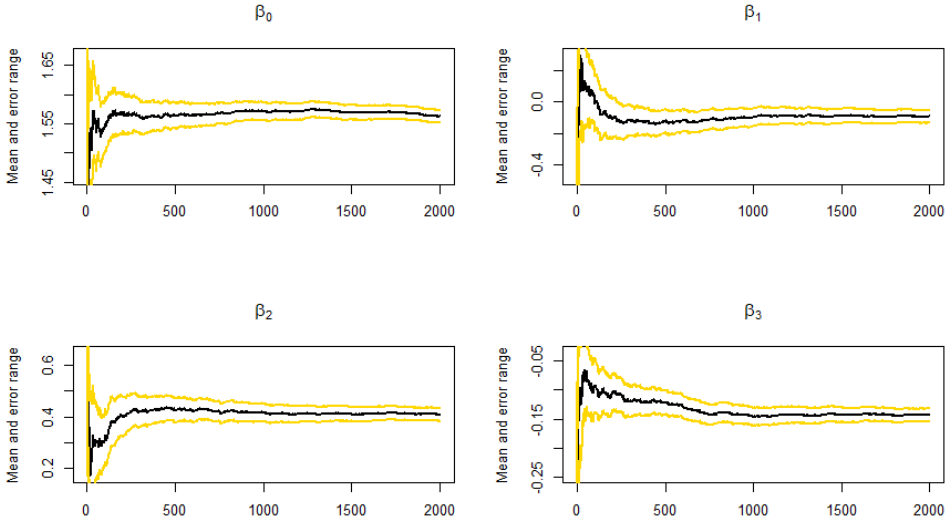
Gibbs Sampler and MCMC

- We run a Gibbs sampling algorithm based on the hierarchical model we mentioned earlier to generate $\boldsymbol{\beta}$ from its posterior distribution.
  1. We generate 2.5×10^5 MCMC iterations.
  2. We use the first 10000 iterations as "burn-in".
  3. Based on the autocorrelation plots, we thin the chains by saving every 120^{th} observation.
- We compute the posterior mean, median and 95% credible interval for each coefficient.

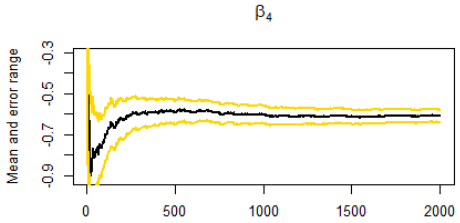
Trace Plots of Coefficients

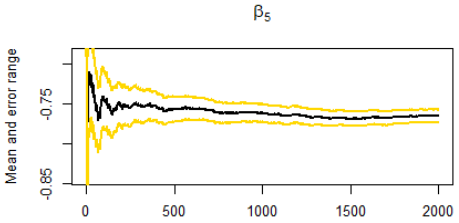


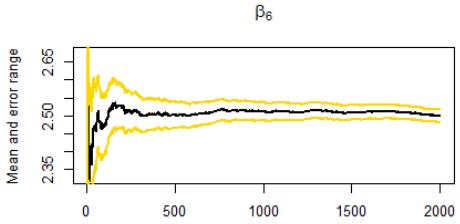
Cumulative Mean Plots of Coefficient Estimates



Cumulative Mean Plots of Coefficient Estimates







Inference on Posterior Distribution

Table: Summary of Posterior Distribution

Coefficient	Posterior Mean	Posterior Median	95% HPD Credible Interval
β_0	1.56276808	1.55055256	[1.1224095, 2.0635469]
β_1	-0.09035723	-0.08378184	[-1.6965093, 1.7962278]
β_2	0.40756949	0.39595142	[-0.6344210, 1.4592239]
β_3	-0.14285184	-0.13514894	[-0.6051095, 0.3375358]
β_4	-0.61122904	-0.60033051	[-1.9786189, 0.6408568]
β_5	-0.76579466	-0.76154425	[-1.1066248, -0.4304665]
β_6	2.49977260	2.49646424	[1.7320914, 3.2665078]

Final Selected Model

- Bayesian **Lasso** does not automatically performs model selection but it provides credible interval for each coefficient which is used to determine whether a predictor is significant or not.
- We will call a predictor insignificant if the 95% HPD credible interval of the corresponding coefficient (β_j) contains 0. Otherwise we consider the predictor significant.
- The final model contains **two** covariates *pelvic radius* and *grade of spondylolisthesis* with an *intercept* term.

Prediction Process

- We will use posterior mean as the estimate of the regression coefficients for prediction. We denote the estimate as $\hat{\boldsymbol{\beta}}$.
- We estimate p_i's as $\hat{p}_i = \Phi\left(\mathbf{x}_i'\hat{\boldsymbol{\beta}}\right)$.
- We estimate the response variables as following:

$$\hat{y}_i = \begin{cases} 1 & \text{, if } \hat{p}_i \geqslant 0.5 \\ 0 & \text{, otherwise} \end{cases}$$

Misclassification Error For Training Data

Table: Misclassification Error

(a) Confusion matrix of the Training data

Actual Class	*Predicted Class*	
	Normal	**Abnormal**
Normal	60	12
Abnormal	38	130

(b) Confusion matrix of the Test data

Actual Class	*Predicted Class*	
	Normal	**Abnormal**
Normal	21	6
Abnormal	7	35

- Misclassification error rate for the training data is 20.83% .
- Misclassification error rate for the test data is 18.84% .

Comparison

Table: Comparison of Different Approaches

Misclassification Error	Classical (Backward using BIC)	Classical (Backward using AIC)	Indicator Approach	Bayesian Lasso
Training Error	13.33%	13.75%	13.33%	20.83%
Test Error	18.84%	17.39%	18.84%	18.84%

Conclusion

- In terms of prediction accuracy, the **Bayesian** approaches are performing as good as the **classical** approaches. But in classical setting, we often can not decide between two models whereas in Bayesian approach we can compute the posterior probability of each unique model as well as the marginal inclusion probability of each covariate. And in case of Bayesian Lasso, based on the HPD credible intervals we can decide on whether a particular predictor is significant or not.
- Using the Bayesian approaches, we get two different models. Now, based on our objective, we will choose one of them
  1. If we are looking for simpler model interpretation, then we will consider the model obtained by the **Bayesian Lasso** consists of two covariates with an *intercept*.
  2. If we are interested in better prediction result, then we will consider the model obtained by the **indicator** variable approach. Although both the models are doing equally good in terms of prediction accuracy in the test data but the first model slightly performs better in prediction of the training data.

Thank You