**MATH 2269**

**Course Project**

**Group -Let’s Bayes**

**Aditi Mahajan (s3757732)**

**Archit Sharma (S3755732)**

10/18/2020



Predicting Abnormality in Lower Back

**Table of Contents**

[1. Introduction 3](#_Toc53937805)

[2. Data Source & Description 3](#_Toc53937806)

[3. The Mathematical Model 6](#_Toc53937807)

[4. Specification of Prior Distribution 6](#_Toc53937808)

[5. MCMC Analysis 7](#_Toc53937809)

[5.1. Compiling Model 8](#_Toc53937810)

[5.2. Parameter Tuning 8](#_Toc53937811)

[5.2.1. 1st Run 9](#_Toc53937812)

[5.2.2. 2nd Run 10](#_Toc53937813)

[5.2.3. 3rd Run 11](#_Toc53937814)

[5.2.4. 4th Run 13](#_Toc53937815)

[5.2.5. 5th Run 15](#_Toc53937816)

[5.2.6. 6th Run 19](#_Toc53937817)

[5.2.7. 7th Run 24](#_Toc53937818)

[5.3. Checking Predictive Power 27](#_Toc53937819)

[6. Conclusion 28](#_Toc53937820)

[References 30](#_Toc53937821)

[Appendix 30](#_Toc53937822)

**List of Tables**

[Table 1: Statistical Summary of the features in the dataset 3](#_Toc53946659)

[Table 2: 1st Run specification for MCMC model 9](#_Toc53946660)

[Table 3: 2nd Run specification for MCMC model 10](#_Toc53946661)

[Table 4: 4th Run specification for MCMC model 11](#_Toc53946662)

[Table 5: 4th Run specification for MCMC model 13](#_Toc53946663)

[Table 6: 5th Run specification for MCMC model 15](#_Toc53946664)

[Table 7: Summary Information for the prediction of abnormality in body structure 15](#_Toc53946665)

[Table 8: Summary Information for the prediction of abnormality in body structure(beta0-beta4) 23](#_Toc53946666)

[Table 9: Variation in log odds of outcome - Run 6 23](#_Toc53946667)

[Table 10: 7th Run specification for MCMC model 24](#_Toc53946668)

[Table 11: Model Evaluation 27](#_Toc53946669)

[Table 12: Comparison of Mode Run Efficiency 28](#_Toc53946670)

[Table 13: Final Specification for the MCMC Model 28](#_Toc53946671)

[Table 14: Predictions from the chosen model 28](#_Toc53946672)

[Table 15: Confusion Matrix for Model's Predictive Power 29](#_Toc53946673)

**Table of Figures**

[Figure 1: Frequency of the dependent feature in the dataset 4](#_Toc53937352)

[Figure 2: Scatter Plot of the dependent feature (Class Attribute 0: Abnormal, 1: Normal) with independent features 5](#_Toc53937353)

[Figure 3: Correlation Matrix of the predictor variables 6](#_Toc53937354)

[Figure 4: JAGS Model diagram for implementing logistic regression 7](#_Toc53937355)

[Figure 5: Screenshot of defining JAGS model in R 8](#_Toc53937356)

[Figure 6: MCMC diagnostics result for 1st run – beta4 9](#_Toc53937357)

[Figure 7: MCMC diagnostics result for 2nd run- beta2 11](#_Toc53937358)

[Figure 8: MCMC diagnostics result for 3rd run- beta10 12](#_Toc53937359)

[Figure 9: MCMC diagnostics result for 3rd Run - beta2 13](#_Toc53937360)

[Figure 10: MCMC diagnostics result for 4th run- beta7 14](#_Toc53937361)

[Figure 11: MCMC diagnostics result for 4th run- beta4 14](#_Toc53937362)

[Figure 12: Posterior Distribution plot results of the non-significant predictors (beta5 – beta10) 16](#_Toc53937363)

[Figure 13: Posterior Distribution plot results of the significant predictors (beta0 – beta4) 16](#_Toc53937364)

[Figure 14: MCMC diagnostics result for 5th run- beta0 17](#_Toc53937365)

[Figure 15: MCMC diagnostics result for 5th run- beta1 17](#_Toc53937366)

[Figure 16: MCMC diagnostics result for 5th run- beta2 18](#_Toc53937367)

[Figure 17: MCMC diagnostics result for 5th run- beta3 18](#_Toc53937368)

[Figure 18: MCMC diagnostics result for 5th run- beta4 19](#_Toc53937369)

[Figure 19: MCMC diagnostics result for 6th run- beta0 20](#_Toc53937370)

[Figure 20: MCMC diagnostics result for 6th run- beta1 20](#_Toc53937371)

[Figure 21: MCMC diagnostics result for 6th run- beta2 21](#_Toc53937372)

[Figure 22: MCMC diagnostics result for 6th run- beta3 21](#_Toc53937373)

[Figure 23: MCMC diagnostics result for 6th run- beta4 22](#_Toc53937374)

[Figure 24: Issues in MCMC diagnostics for 6th run- beta 0,3,4 22](#_Toc53937375)

[Figure 25: MCMC diagnostics result for 6th run- Guess 23](#_Toc53937376)

[Figure 26: Posterior Distribution and Diagnostics for beta0 24](#_Toc53937377)

[Figure 27: Posterior Distribution and Diagnostics for predictor- pelvic tilt 25](#_Toc53937378)

[Figure 28: Posterior Distribution and Diagnostics for predictor- sacral slope 25](#_Toc53937379)

[Figure 29: Posterior Distribution and Diagnostics for predictor- pelvic radius 26](#_Toc53937380)

[Figure 30: Posterior Distribution and Diagnostics for predictor- degree spondylolisthesis 26](#_Toc53937381)

[Figure 31: Distribution of alpha (guessing parameter) 27](#_Toc53937382)

# Introduction

With the onset of the COVID-19 pandemic, work and studies have shifted online with working on laptops from home. Being in lockdown and studies being remote students have started spending long hours in front of laptops camped out on bed, kitchen countertops or tables. Due to bad postures, people are experiencing new aches and physical strain especially lower back pains more commonly. This ignited our curiosity on whether people with certain body structure are more susceptible to have lower back pain and led to the topic for this project.

The bottom part of the back or lumbar spine has network of spinal muscles, interconnected bones, nerves and is connected to the pelvis, working in tandem. Generating lots of movements and bearing the weight of the upper body, the lower back becomes vulnerable to pain. Though one of the most common type of pain, lumbago (lower back pain) can be acute or chronic. It could be indications of degenerative or nerve and spinal problems. (Akhil Chhatre, n.d.) (John Peloza, 2017)

Lower back pain may also be caused by damages in joints, ligaments or degenerating intervertebral disc or irritation in the different components of the back. In this project, we differentiate body structures as being normal and abnormal, thereby determining susceptibility to lower back pain, based on the physical attributes of the spine and pelvis. With the information contained in the independent features or predictor variables, and the dependent variable being binary, the focus of this project would be to classify or predict the type of body structure of person using robust Bayesian binary Logistic Regression analysis methods. (Stroke, n.d.)

# Data Source & Description

The dataset for this project is sourced from Kaggle (User-sammy123, 2016) .There are 310 records, and the dataset is further spilt in the ratio of 90:10 resulting in 279 records as train and remaining 31 observations as test data. This will help the check accuracy of results from the model. There are 12 numeric predictors and one target class determining the body structure as either 1 i.e. normal or 0 i.e. abnormal.

The observations are a collection of data on the lumbar and physical spine regions. Looking at the overall statistics of all the variables:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Variables | Min | 1st Qu. | Median | Mean | 3rd Qu. | Max |
| pelvic\_incidence | 34.38 | 46.75 | 57.29 | 58.97 | 70.32 | 85.68 |
| pelvic\_tilt | -6.55 | 10.6 | 15.864 | 16.888 | 21.123 | 49.432 |
| lumbar\_lordosis\_angle | 15.59 | 37.08 | 49.28 | 51.40 | 62.35 | 125.74 |
| sacral\_slope | 13.37 | 33.42 | 42.14 | 42.08 | 51.06 | 69.51 |
| pelvic\_radius | 70.08 | 110.71 | 118.19 | 118.02 | 125.46 | 163.07 |
| degree\_spondylolisthesis | -11.058 | 1.496 | 10.443 | 22.938 | 37.795 | 148.754 |
| pelvic\_slope | 0.003 | 0.205 | 0.468 | 0.47 | 0.708 | 0.998 |
| Direct\_tilt | 7.027 | 12.950 | 22.253 | 21.498 | 29.192 | 36.744 |
| thoracic\_slope | 7.038 | 10.448 | 13.141 | 13.102 | 15.841 | 19.324 |
| cervical\_tilt | 7.031 | 9.524 | 11.954 | 11.9 | 14.354 | 16.821 |
| sacrum\_angle | -35.287 | -25.146 | -15.260 | -14.386 | -3.486 | 6.972 |
| scoliosis\_slope | 7.07 | 16.97 | 24.35 | 25.43 | 33.72 | 44.34 |
| Class\_Attribute | 0 | 0 | 0 | 0.3369 | 1 | 1 |

Table 1: Statistical Summary of the features in the dataset

1. **Pelvic incidence**: All spinal curves are based on this angle. Over 50% of the observations have pelvic incidence angle higher than 57.29 and up to 85.68
2. **Pelvic Tilt**: Represents the orientation or alignment of pelvis compared to the rest of the body, especially thighbones. The dataset has records ranging from -6.55 to 49.43 with average pelvic tilt of 16.88 degree.
3. **Lumbar Lordosis Angle**: One of the parameters for back pain, the dataset contains observations ranging from 15.59 to 125.74 with 50% observations greater than 49.28
4. **Sacral Slope**: This measure can be used to determine sagittal balance, a parameter that has been correlated to onset of spondylolisthesis. An average of 42.08 is observed.
5. **Pelvic Radius**: parameters representing spine and pelvic alignment. With 50% observation showing pelvic radius higher than 118.19, the highest record has a pelvic radius of 163.07
6. **Degree of spondylolisthesis**: Occurring due to degeneration of intervertebral disc and ligaments, spondylolisthesis is the condition when lower back vertebra slips forward or out of its position in spine. (Lancaster, n.d.). With degree ranging from -11.058 to 148.754, an average of 22.938-degree severity is observed.
7. **Pelvic Slope**: parameters representing spine and pelvic alignment, 50% of records have a pelvic slope higher than 0.468 degree.
8. **Direct Tilt**: measure determining spine and pelvic alignment, direct tilt from 7.02 to 36.744 is observed with an average of 21.498.
9. **Thoracic Slope:** parameters representing spine and pelvic alignment. With over 50% people having a thoracic slope higher than 13.141, the highest slope observed is of 19.324
10. **Cervical Tilt**: parameters representing spine and pelvic alignment. An average of 11.954 is observed.
11. **Sacrum Angle**: a measure of pine and pelvic alignment, sacrum angle ranges from -35.287 to 6.972 with an average of -14.386.
12. **Scoliosis Slope**: a measure determining presence of a sideways curve, over 50% of records have a slope higher than 24.35.
13. **Class Attribute**: Type of body structure. 0: Abnormal or 1: Normal. (Dependent Feature). The cases with abnormal body structure are twice the normal cases

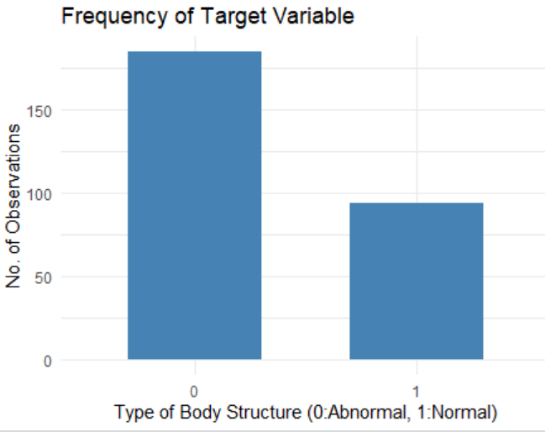


Figure 1: Frequency of the dependent feature in the dataset

**Descriptive Statistics**

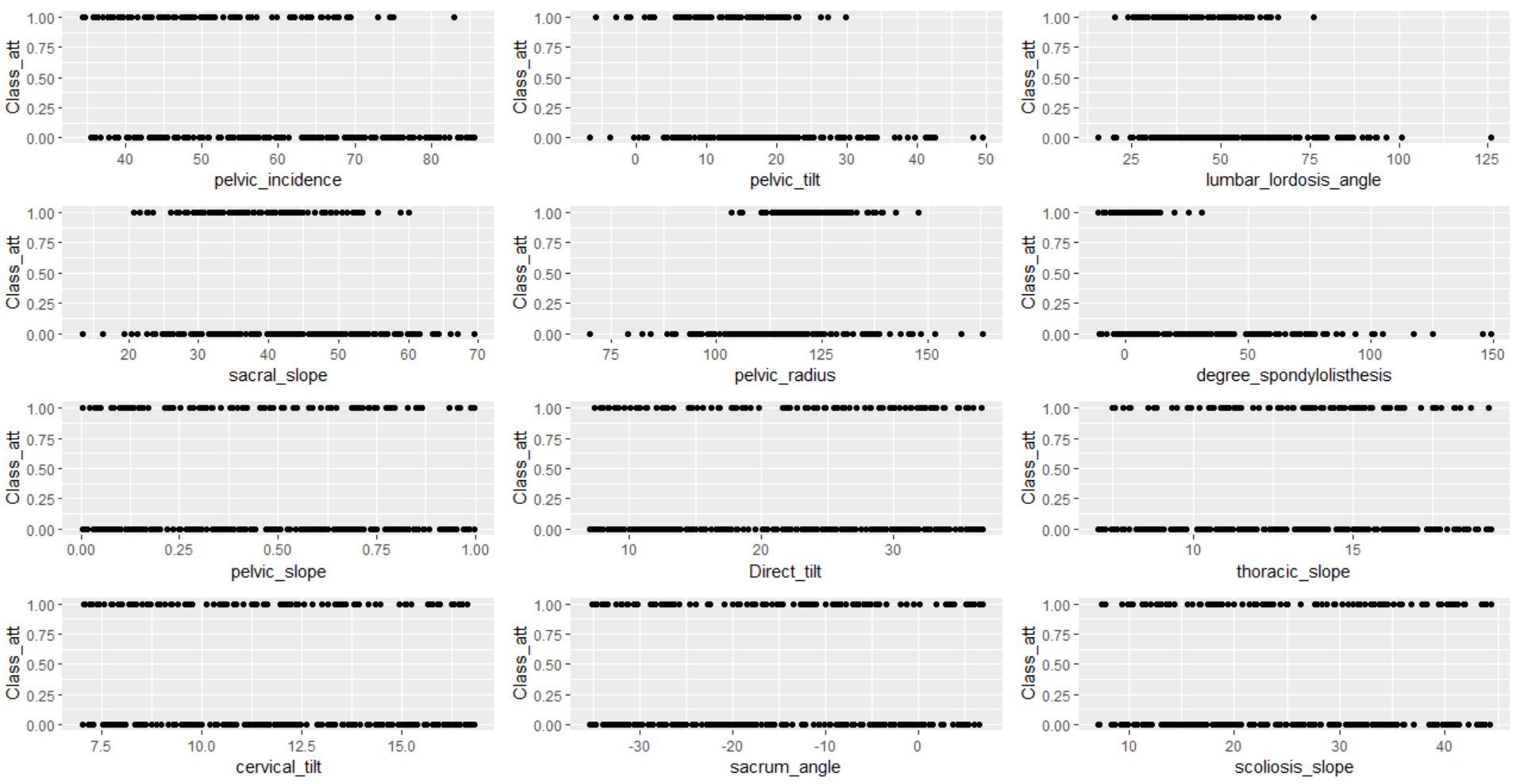


Figure 2: Scatter Plot of the dependent feature (Class Attribute 0: Abnormal, 1: Normal) with independent features

The independent features are compared against the dependent feature i.e. ‘Class Attribute’ with 1 signifying a normal body structure and 0 as abnormal body structure and more likely to experience back pain and cause disability in long run.

1. For pelvic incidence (Fig2, Top left) the normal range is between 43 degree and 62 degree and this can be seen in the scatter plot as well with most points concentrated in this range. There are also few outliers present. The abnormal body structure has a wider range of observations (Legaye, 2007)
2. A normal mean pelvic tilt angle is observed at 13 degree +/- 6 degree. The plot (Fig2, top middle) supports this as normal body structure records are concentrated towards 13 degree. While people with abnormal body structure tend to deviate from normal and have a wider range of angle. (J. C. Le Huec, 2011)
3. Previous research puts a normal lumbar lordosis angle value at 20-45 degree the figure above (Fig2, Top right) also shows people with normal body structure having lumbar lordosis angle majorly distributed around 22-50 degree whereas an abnormal body structure has range from 20 up to 100 degrees. There are also outliers present. (Lin RM, 1992)
4. Similarly, in the figure (Fig2, second row) for measures of sacral slope, pelvic radius and spondylolisthesis, a normal body structure tends to have observations in shorter range while abnormal body structure showing unusual behavior has values scattered across wider range of measure. Few outliers are also observed.
5. For the remaining measures of spine and pelvic alignment i.e. pelvic slope, direct tilt, thoracic slope, cervical tilt, sacrum angle, scoliosis slope, the figure (Fig2, last two rows), the normal body structure and abnormal body structure have similar ranges for the values. This might also mean that these measures or features may not have high significance on the type of structure of a body

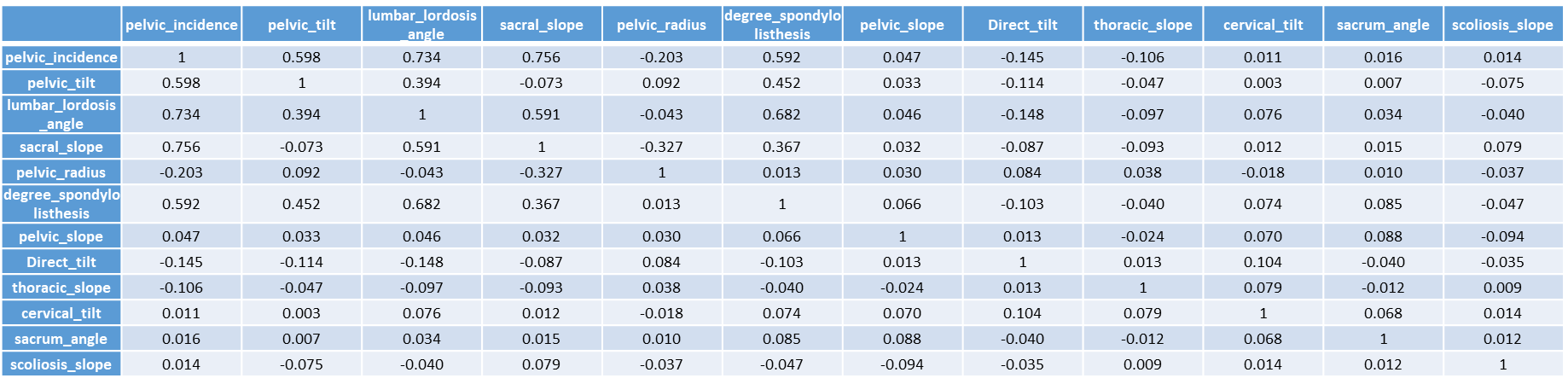
**Correlation Matrix**

Figure 3: Correlation Matrix of the predictor variables

Pelvic incidence has a good positive correlation with pelvic tilt (0.6), lumbar lordosis angle (0.74) and sacral slope (0.76). Same is the case for lumbar lordosis angle having high correlation with other predictor variables such as sacral slope (0.59) and degree of spondylolisthesis (0.68). This may cause issues in the multiple regression analysis and therefore, Pelvic incidence and LLA measures are dropped for the scope of this project.

To minimize the impact of outliers, the Bayesian logistic regression analysis done would be robust.

# The Mathematical Model

Classifying the body structure for abnormality, the model is as follows

**Y ~ *Bernoulli (Ө)***

*Where* ***Ө***  = α.(1/2) + (1- α). logistic( β0 + β1 Pelvic\_tilt + β2 Sacral\_slope + β3 Pelvic\_radius + β4 Degree\_spondylolisthesis +β5 Pelvic\_slope + β6 Direct\_tilt + β7 Thoracic\_slope + β8 Cervical\_tilt +β9 Sacrum\_angle +β10 Scoliosis\_slope )

Y: Class\_Attribute (Type of body structure)

As there are success probabilities, i.e. binary class, Bernoulli’s distribution is used. The parameter of interest ‘Ө’ is explained by logistic function i.e. defined in terms of the independent regression parameters.

To minimize the misclassification rate of the model, ‘Ө’ is defined as being generated from two sources. Either from the logistic function of the predictor or independent variables, or the other by randomness or ‘guessing’. To implement this, an additional parameter of ‘α’ is used which is the chance of being generated by guessing. This means with probability of α, observations are generated using guess else with probability (1- α), y is obtained using the model of logistic regression.

# Specification of Prior Distribution

JAGS model diagram is obtained from the distributional form of the model of logistic regression described in the previous section.

Diagram, schematic

Description automatically generated

Figure 4: JAGS Model diagram for implementing logistic regression

Ө or µ, is the parameter which is modelled. The dependent variable for each distribution (β0 to β10) returns the success probability distributed as Bernoulli. Prior distributions are fitted on the parameters of the model in the Bayesian logistic regression.

β0 to β10 can be between -∞ to +∞ and additionally, these features are continuous. Thus, the priors of these regression parameters are taken as normally distributed.

Prior for ‘α’ (i.e. guessing) has a prior distribution induced. As ‘α’ lies between 0 to 1 and can be continuous, the prior specified is beta prior. With very few values greater than 0.5, this would push most observations of ‘Y’ to be generated through the logistic regression model and only few through guesses.

Previous knowledge on the predictors are not readily available. Individual researches defining the impact of the different parameters of spine measure on normality of spine is available but cannot be trusted as they are based on individual perception. Moreover, no definite average value could be found leading to an impact on spine structure. Due to lack of credible sources of the impact of independent variables on Ө, priors taken are non-informative. With any credible information, the mean and variance can be varied in the prior specification.

# MCMC Analysis

The aforementioned methodology was carried out, and the following subsections state and discuss the results of MCMC Bayesian analysis and prediction of data.

## 5.1. Compiling Model

Since the scales for the chosen predictors were very different from each other, their standardized version was fed to the model compiler. Hence, standard deviation of each variable was introduced in the model and reflected in all the presumed expressions of distribution. Below is a snapshot of simulated model:



Figure 5: Screenshot of defining JAGS model in R

## 5.2. Parameter Tuning

The following parameters were toggled for obtaining an efficient and accurate representation of data:

* *Number of chains*
* *Burn-in period*: The length of burn-in period, i.e. moving the chain away from unrepresentative section to the most populated region of the posterior
* *Thinning*: the process of thinning reduces the autocorrelation of the chain as only every kth step is stored. The value of k is set here
* *Number of saved steps*: These refer to the number of steps wanted after the thinning process.
* *Adapt steps*: The no. of steps to be taken for adapting or tuning the samplers

Following diagnostic checks will be considered for each run

1. **Representativeness**:

The values of MCMC chain must fully cover the range of posterior distribution without any obstacle. Taking different starting points resulting chains are superimposed in visuals of trace plots. To ensure representativeness the chains must overlap each other. Other criteria to check representativeness, is the shrink factor.

If the shrink factor is less than 1.2, we can say the chains have converged adequately

1. **Accuracy:**

If MCMC method is rerun, the estimates of mean and 95% HDI should not deviate significantly. Autocorrelation plots are used to check there is no high correlation between the chains generated from the successive steps. After thinning the chain, ESS i.e. effective sample size is checked, and it must be closer to the number of iterations to ensure chains provide significant independent information.

Another accuracy measure is the Monte Carlo Standard error (MCSE). The lower the error rate and closer to 0 the better the accuracy of chain.

1. **Efficiency:**

Time taken to compute each model run

For every run, the diagnostics of β were examined.

### 5.2.1. 1st Run

|  |  |
| --- | --- |
| **Adapt Steps** | 100 |
| **Burn-in Steps** | 300 |
| **Chains** | 3 |
| **Thinning Steps** | 3 |
| **Saved Steps** | 100 |
| **Runtime** | 0:29 |

Table 2: 1st Run specification for MCMC model

For the first run, very small values were assigned to all the parameters. Moreover, the initial list of starting points for variables was not specified. This was done to check if the model was able to identify efficient starting points on its own.

The model successfully identified the starting points and took 29 seconds to run. However, the diagnostics turned out to be quite bad. For example, below is a screenshot of the diagnostics obtained for β4:

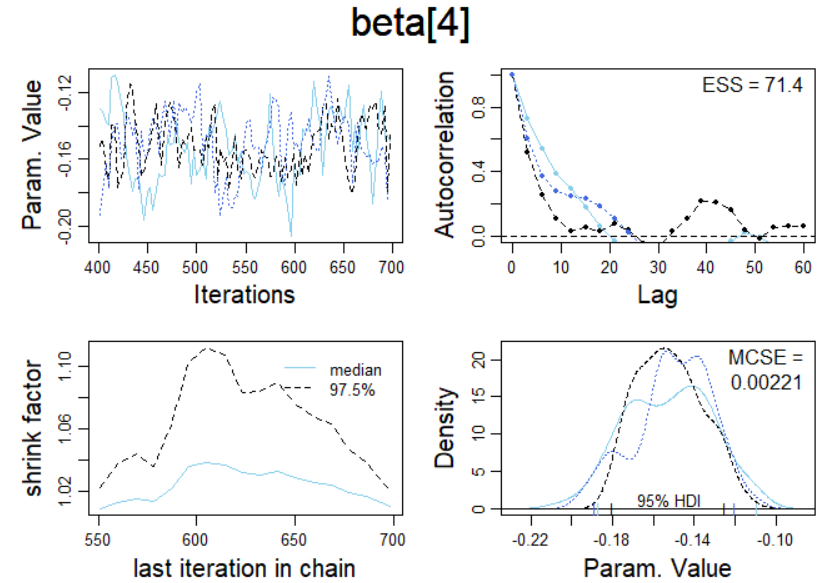


Figure 6: MCMC diagnostics result for 1st run – beta4

It can be clearly observed that:

* The chains do not converge in the trace plots
* Representativeness 🡪 Good
  + Shrink factor < 1.2, which indicates satisfactory representativeness
* Accuracy 🡪 Bad
  + Very high autocorrelation observed in observations
  + Adequately low MCSE (0.00221)
  + HDIs in density plot do not overlap
* Efficiency 🡪 Good, as the runtime was under a minute

In conclusion, the following inferences were drawn for the diagnostics:

* Acceptable 🡪 None
* Unacceptable 🡪 β0 , β1 , β2 , β3 , β4 , β5 , β6 , β7 , β8 , β9 , β10

In conclusion, the diagnostics were plagued with heavy autocorrelation, non-congruent HDIs, and visibly non-converging chains. However, the shrink factor and Monte Carlo Standard Error (MCSE) seemed acceptable. Hence the 2nd run was initialized.

### 5.2.2. 2nd Run

|  |  |
| --- | --- |
| **Adapt Steps** | 500 |
| **Burn-in Steps** | 500 |
| **Chains** | 3 |
| **Thinning Steps** | 5 |
| **Saved Steps** | 1,000 |
| **Runtime** | 1:29 |

Table 3: 2nd Run specification for MCMC model

To solve the issue of divergence in chains, adapt steps and burn-in period were increased. Thinning and saved steps were incremented to 5 and 1000 respectively, to resolve the problem of substantial autocorrelation.

The model took 1 minute and 29 seconds to run. It was observed that for the 11 βs , autocorrelation was significantly improved for β5 to β10. On the brighter side, the chains were finally converging. In β0 to β3, autocorrelations were still present, and the diagnostics were still very problematic for β4. The screenshot for β2 here highlights this issue.

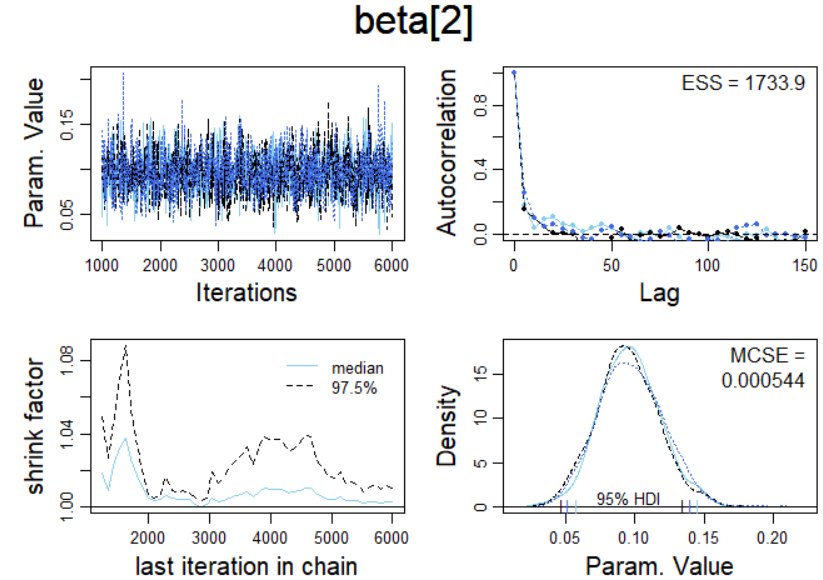


Figure 7: MCMC diagnostics result for 2nd run- beta2

It can be clearly observed that:

* The chains started converging in the trace plots
* Representativeness 🡪 Good
  + Shrink factor < 1.2, which indicates satisfactory representativeness
* Accuracy 🡪 Bad
  + Substantial autocorrelation observed in observations
  + Adequately low MCSE (0.000544)
  + HDIs in density plot do not overlap
* Efficiency 🡪 Good, as the runtime was just over a minute

In conclusion, the following inferences were drawn for the diagnostics:

* Acceptable 🡪 β0 , β1 , β3
* Unacceptable 🡪 β2 , β4 , β5 , β6 , β7 , β8 , β9 , β10

Since a majority of betas had inadequate diagnostics, 3rd run was initialized.

### 5.2.3. 3rd Run

### 

|  |  |
| --- | --- |
| **Adapt Steps** | 1,000 |
| **Burn-in Steps** | 1,000 |
| **Chains** | 3 |
| **Thinning Steps** | 7 |
| **Saved Steps** | 1,000 |
| **Runtime** | 1:45 |

Table 4: 4th Run specification for MCMC model

For a better convergence between chains, and getting rid of autocorrelation, all the parameters except no. of chains were increased. Number of chains was left unaltered to maintain efficiency of MCMC process.

The model took 1 minute 45 seconds to run on this configuration. A much better convergence with lesser autocorrelations was achieved for β1, β3 and βs between 5 to 10. The shrink factor was also reduced significantly and HDIs were found to be overlapping in density plot for half of the betas. The improvements can be seen in the screenshot below for β­10:

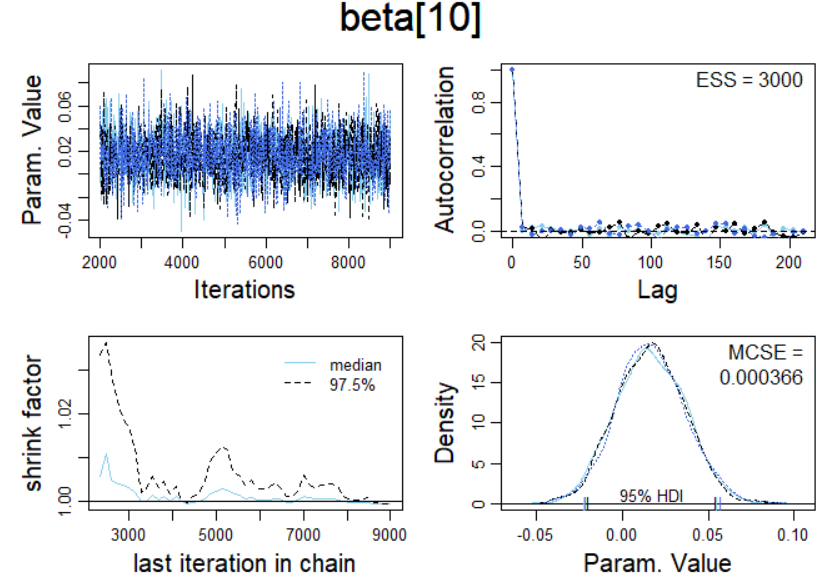


Figure 8: MCMC diagnostics result for 3rd run- beta10

It can be clearly observed that:

* The chains converged in the trace plots
* Representativeness 🡪 Good
  + Shrink factor < 1.2, which indicates satisfactory representativeness
* Accuracy 🡪 Good
  + No autocorrelation observed in observations
  + Adequately low MCSE (0.000366)
  + HDIs in density plot are close to a decent overlap
* Efficiency 🡪 Good, as the runtime was just under 2 minutes

In conclusion, the following inferences were drawn for the diagnostics:

* Acceptable 🡪 β0 , β1 , β3 , β5 , β8 , β9 , β10
* Unacceptable 🡪 β2, β4 , β6 , β7

There was a significant autocorrelation present in β2, β4 and β6, and a majority of betas had inadequate overall diagnostics. An example of autocorrelation among one of the betas is shown below:

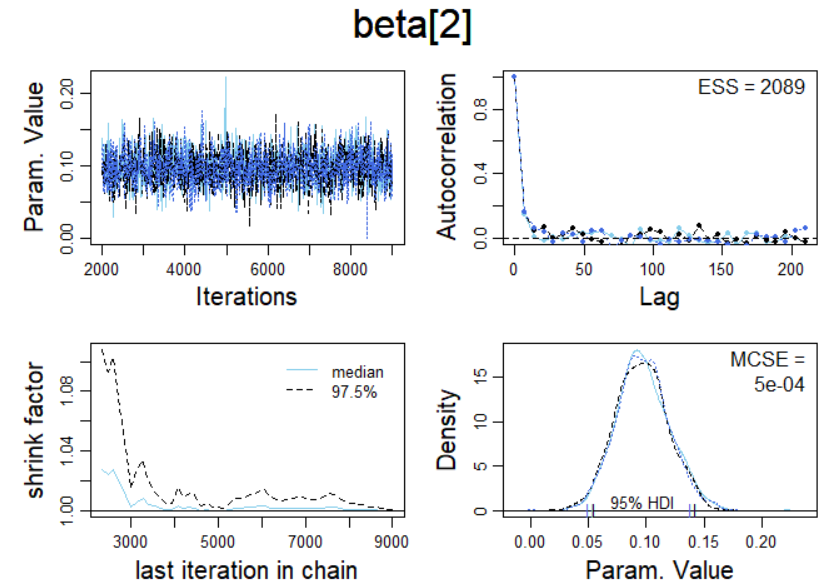


Figure 9: MCMC diagnostics result for 3rd Run - beta2

Hence, 4th run was initialized.

### 5.2.4. 4th Run

|  |  |
| --- | --- |
| **Adapt Steps** | 1,000 |
| **Burn-in Steps** | 1,000 |
| **Chains** | 3 |
| **Thinning Steps** | 10 |
| **Saved Steps** | 1,500 |
| **Runtime** | 2:53 |

Table 5: 4th Run specification for MCMC model

To tackle this autocorrelation, thinning was increased to 10, and number of saved steps was also increased. The runtime for this configuration was 2 minutes 53 seconds. As seen in the diagnostics for β7 below, a perfect effective sample size, and Monte Carlo standard error close to 0 indicated that the model was performing better with each run.

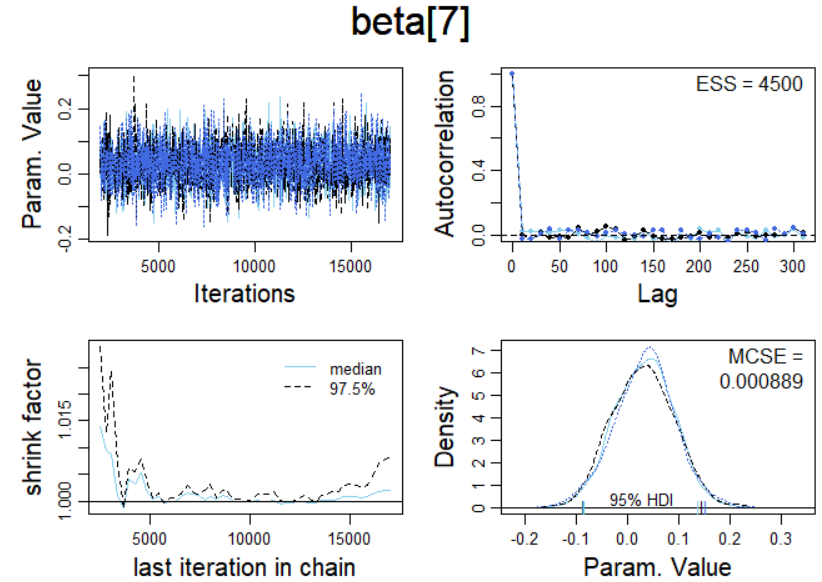


Figure 10: MCMC diagnostics result for 4th run- beta7

The following improvements were observed for all the βs:

* A shrink factor of less than 1.2 was achieved
* The chains were observed to be converging very well
* The HDIs in density plots were either extremely close or overlapping perfectly
* The MCSE was observed to be very close to 0
* ESS was observed to be decently high, but could be better

However, the problem of autocorrelation still existed in β2 and β4. This can be confirmed from the diagnostics for β4 below:

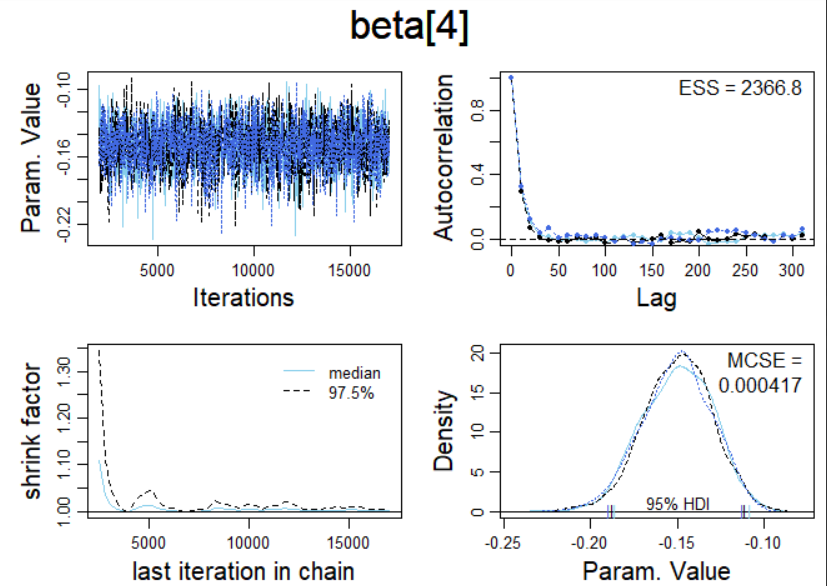


Figure 11: MCMC diagnostics result for 4th run- beta4

Hence, another configuration was initialized with increased thinning and number of chains.

### 5.2.5. 5th Run

|  |  |
| --- | --- |
| **Adapt Steps** | 1,000 |
| **Burn-in Steps** | 1,000 |
| **Chains** | 4 |
| **Thinning Steps** | 20 |
| **Saved Steps** | 1,500 |
| **Runtime** | 6:58 |

Table 6: 5th Run specification for MCMC model

Finally, autocorrelations present in β2 and β4 were successfully removed in the 5th run which took 6 minutes and 58 seconds to train. For this run, thinning was increased to 20 and 4 chains were initialized for compiling the model.

The shrink factor was less than 1.2, MCSE close to 0, chains converged and HDIs overlapped for all the 11 betas. The summary for the model is given below:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Parameter** | **Mean** | **Median** | **Mode** | **ESS** | **HDImass** | **HDIlow** | **HDIhigh** |
| beta0 | -14.6718 | -14.5215 | -14.5237 | 6000 | 0.95 | -2.20E+01 | -7.72435 |
| beta[1] | -0.07443 | -0.07348 | -0.06889 | 6000 | 0.95 | -1.36E-01 | -0.0152 |
| beta[2] | 0.096489 | 0.095998 | 0.094085 | 6000 | 0.95 | 5.26E-02 | 0.141248 |
| beta[3] | 0.107093 | 0.106304 | 0.106627 | 6000 | 0.95 | 5.82E-02 | 0.153604 |
| beta[4] | -0.15028 | -0.14938 | -0.14583 | 4851.1 | 0.95 | -1.88E-01 | -0.11029 |
| beta[5] | 0.254731 | 0.256604 | 0.226969 | 6000 | 0.95 | -1.16E+00 | 1.58329 |
| beta[6] | -0.00826 | -0.00858 | -0.00908 | 6000 | 0.95 | -5.26E-02 | 0.044221 |
| beta[7] | 0.033434 | 0.033332 | 0.020865 | 6000 | 0.95 | -8.19E-02 | 0.154277 |
| beta[8] | -0.05444 | -0.05432 | -0.05769 | 6000 | 0.95 | -1.96E-01 | 0.0803 |
| beta[9] | -0.00419 | -0.00424 | -0.00331 | 6000 | 0.95 | -3.82E-02 | 0.026043 |
| beta[10] | 0.016998 | 0.016531 | 0.014788 | 6000 | 0.95 | -2.09E-02 | 0.056438 |
| guess | 0.017557 | 0.012931 | 0.003855 | 6277.7 | 0.95 | 5.08E-07 | 0.050213 |

Table 7: Summary Information for the prediction of abnormality in body structure

It can be seen that a near perfect estimated sample size was achieved for each parameter. However, it can also be seen that in the posteriors for β5 to β10, 0 value was present inside their respective 95% HDIs. To confirm analyze the significance of these parameters, the posterior distributions were obtained as follows:



Figure 12: Posterior Distribution plot results of the non-significant predictors (beta5 – beta10)

It is evident that 0 value is present somewhere around the middle of the HDIs for β5 to β10. This is an indicator of insignificance, meaning that any change in their magnitude will not affect the outcome. On the other hand, betas between 0 to 4 were all found to be significant.

Chart, histogram

Description automatically generated

Figure 13: Posterior Distribution plot results of the significant predictors (beta0 – beta4)

The diagnostics for the significant priors are discussed below:

* β0 : The chains converged and a perfect ESS was achieved with some very minor autocorrelations, low MCSE and shrink factor less than 1.2.

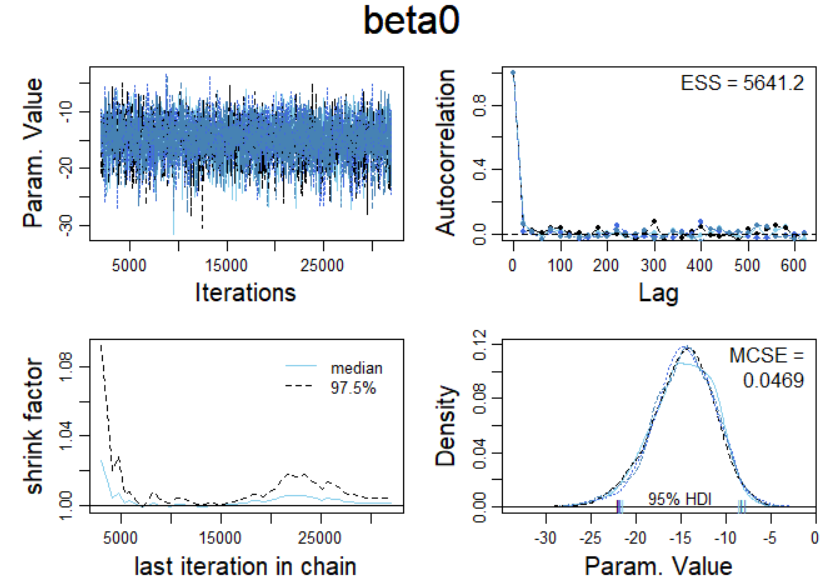


Figure 14: MCMC diagnostics result for 5th run- beta0

* β1 : The chains converged and a perfect ESS was achieved with no significant autocorrelations, low MCSE and shrink factor less than 1.2.

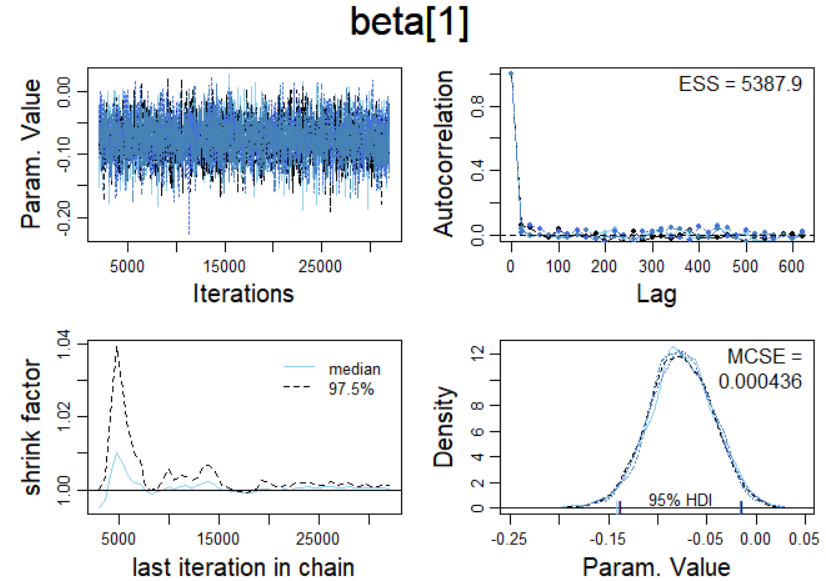


Figure 15: MCMC diagnostics result for 5th run- beta1

* β2 : The chains converged and a perfect ESS was achieved with no significant autocorrelations, low MCSE and shrink factor less than 1.2.



Figure 16: MCMC diagnostics result for 5th run- beta2

* β3 : The chains converged and a perfect ESS was achieved with some very minor autocorrelations, low MCSE and shrink factor less than 1.2.

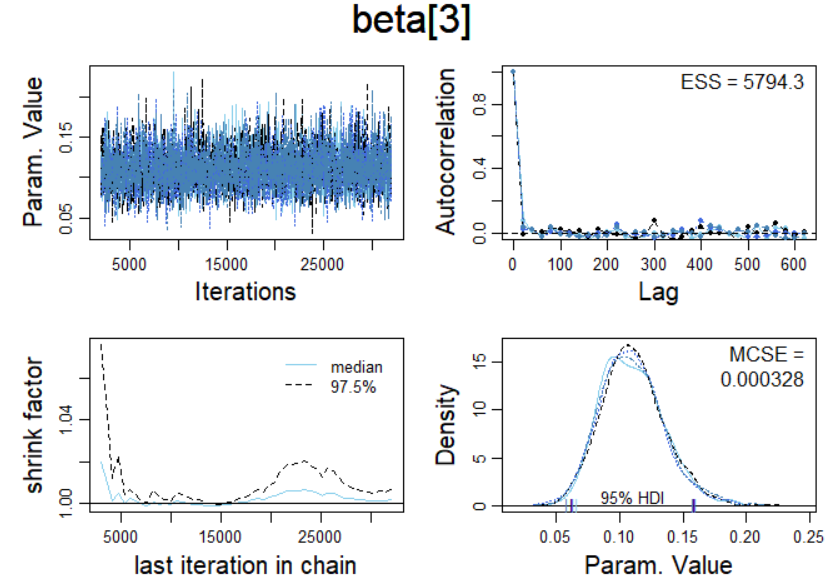


Figure 17: MCMC diagnostics result for 5th run- beta3

* β4 : The chains converged and a perfect ESS was achieved with some very minor autocorrelations, low MCSE and shrink factor less than 1.2.

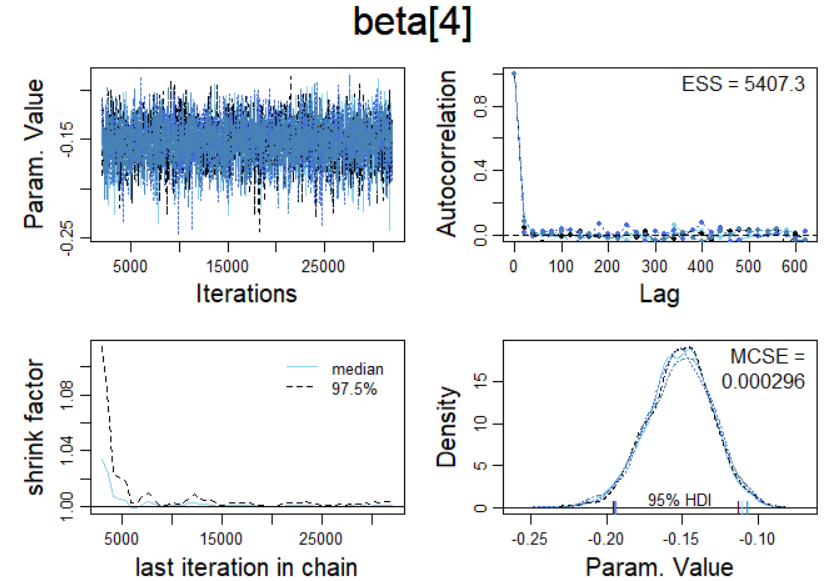


Figure 18: MCMC diagnostics result for 5th run- beta4

Since 6 out of 10 parameters were found to be insignificant, the model was re-initialized using only the significant parameters, i.e. β0 to β4 to analyze the sensitivity of the model.

### 5.2.6. 6th Run

For the 6th run, the following 5 betas were considered for the model:

1. β0 – Intercept
2. β1 – Pelvic Tilt
3. β2 – Sacral Slope
4. β3 – Pelvic Radius
5. β4 – Degree of Spondylolisthesis

MCMC parameters from the previous run (5th run) were considered for simulation as the diagnostics for all the concerned posteriors were found to be satisfactory. The diagnostics of betas are discussed below:

* β0 : The chains converged and a perfect ESS was achieved with some very minor autocorrelations, low MCSE and shrink factor less than 1.2.

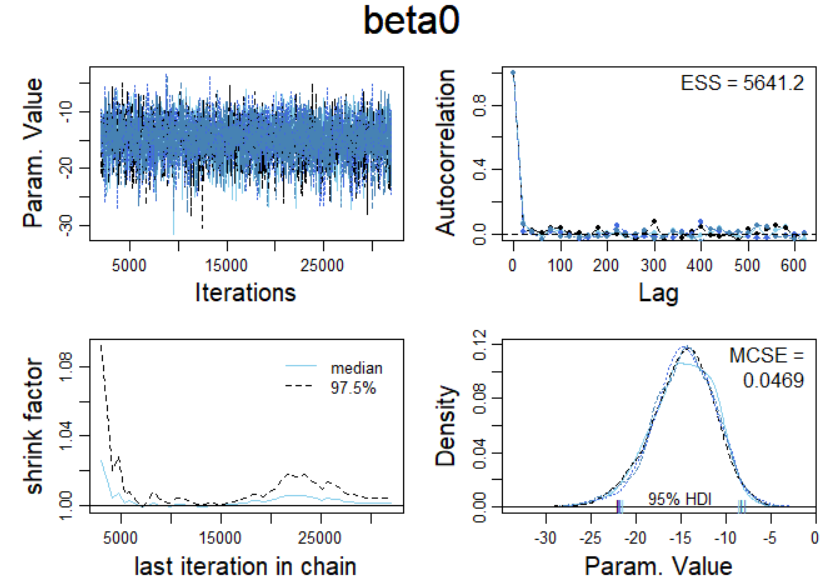


Figure 19: MCMC diagnostics result for 6th run- beta0

* β1 : The chains converged and a perfect ESS was achieved with no significant autocorrelations, low MCSE and shrink factor less than 1.2.

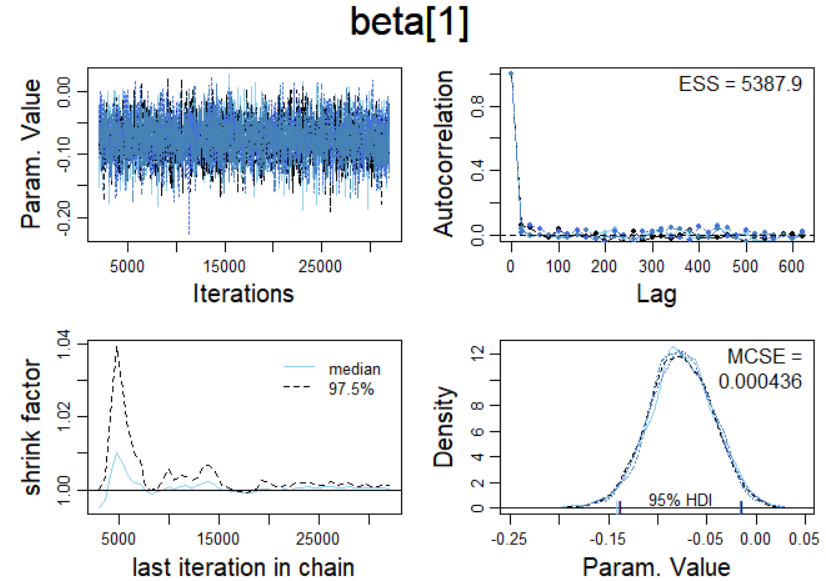


Figure 20: MCMC diagnostics result for 6th run- beta1

* β2 : The chains converged and a perfect ESS was achieved with no significant autocorrelations, low MCSE and shrink factor less than 1.2.



Figure 21: MCMC diagnostics result for 6th run- beta2

* β3 : The chains converged and a perfect ESS was achieved with some very minor autocorrelations, low MCSE and shrink factor less than 1.2.

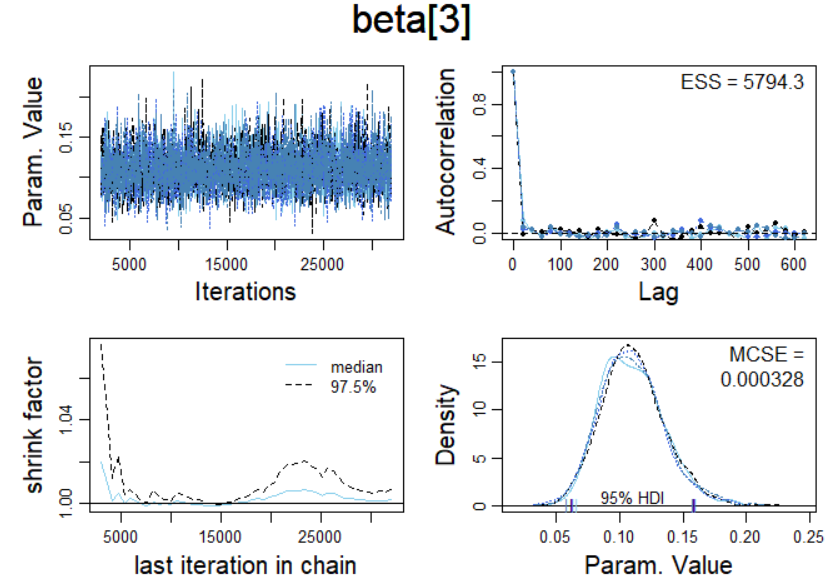


Figure 22: MCMC diagnostics result for 6th run- beta3

* β4 : The chains converged and a perfect ESS was achieved with some very minor autocorrelations, low MCSE and shrink factor less than 1.2.

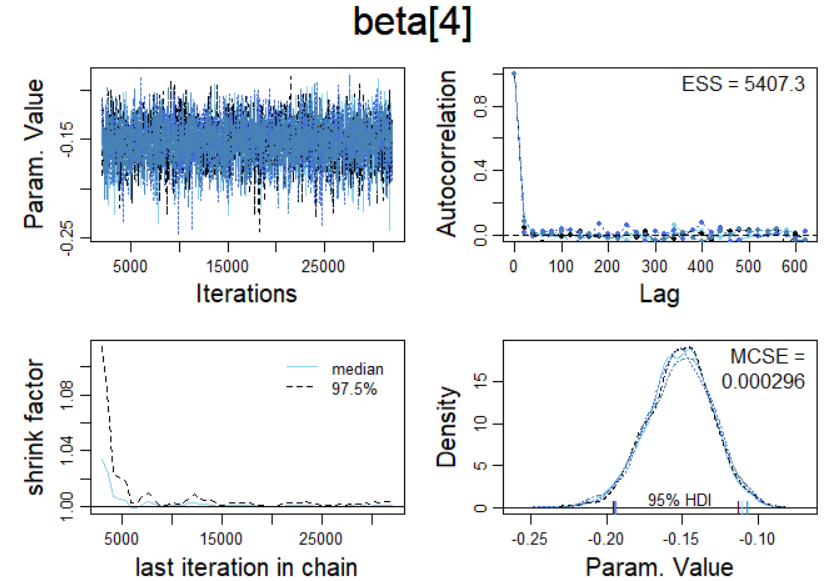


Figure 23: MCMC diagnostics result for 6th run- beta4

Lesser number of predictors meant improved computational efficiency, and the model was simulated in 2 minutes 46 seconds. The diagnostics were adequate in almost every aspect, except for slight presence of minor autocorrelations and marginally overlapping HDIs. The screenshots of diagnostics for β0 , β3 and β4 highlight the same below:

Graphical user interface, diagram, application

Description automatically generated

Figure 24: Issues in MCMC diagnostics for 6th run- beta 0,3,4

The model summary is given below:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Parameter** | **Mean** | **Median** | **Mode** | **ESS** | **HDImass** | **HDIlow** |
| beta0 | -1.47E+01 | -1.45E+01 | -1.42E+01 | 5384 | 0.95 | -2.21E+01 |
| beta[1] | -7.62E-02 | -7.66E-02 | -7.93E-02 | 5321.6 | 0.95 | -1.38E-01 |
| beta[2] | 9.42E-02 | 9.36E-02 | 9.33E-02 | 6000 | 0.95 | 4.52E-02 |
| beta[3] | 1.10E-01 | 1.09E-01 | 1.07E-01 | 5391.6 | 0.95 | 6.26E-02 |
| beta[4] | -1.52E-01 | -1.51E-01 | -1.47E-01 | 5356.3 | 0.95 | -1.95E-01 |
| guess | 1.85E-02 | 1.34E-02 | 4.15E-03 | 6000 | 0.95 | 5.47E-07 |

Table 8: Summary Information for the prediction of abnormality in body structure(beta0-beta4)

For Intercept (β0), the mode value (-14.2) shown in summary table indicates the base value for log odds for outcome regarding the condition of an individual’s lower back when all other variables are equal to 0. Similarly, the following inferences were drawn regarding the variation in log odds of outcome influenced by the predictors:

|  |  |
| --- | --- |
| **Parameter** | **Change in log odds of outcome, for 1-unit change in Parameter** |
| beta[1] | -0.08 |
| beta[2] | 0.09 |
| beta[3] | 0.11 |
| beta[4] | -0.15 |

Table 9: Variation in log odds of outcome - Run 6

Furthermore, the mode for guess parameter was computed to be 0.00415, which indicates an approximately 0.4% (~0%) chance of randomness being generated by guessing.

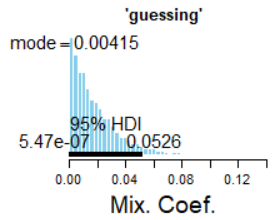


Figure 25: MCMC diagnostics result for 6th run- Guess

This implies that the outliers do not hamper the model’s performance much. The model equation could be simplified to:

*Ө* = 0%.(1/2) + (1- 0%). logistic( β0 + β1 Pelvic\_tilt + β2 Sacral\_slope + β3 Pelvic\_radius + β4 Degree\_spondylolisthesis )

Or,

***Ө* = logistic( β0 + β1 Pelvic\_tilt + β2 Sacral\_slope + β3 Pelvic\_radius + β4 Degree\_spondylolisthesis )**

The MCMC parameters were toggled further to check if perfect diagnostics could be achieved for all the predictors involved.

### 5.2.7. 7th Run

|  |  |
| --- | --- |
| **Adapt Steps** | 1,000 |
| **Burn-in Steps** | 1,000 |
| **Chains** | 4 |
| **Thinning Steps** | 25 |
| **Saved Steps** | 2,000 |
| **Runtime** | 3:12 |

Table 10: 7th Run specification for MCMC model

In order to get rid of the slight autocorrelations observed in the previous run, thinning and the number of saved steps were further increased. The model took 3 minutes and 12 seconds to train and perfect diagnostics were obtained for all the 5 betas. The diagnostics and posterior distributions for all the betas are discussed below:

#### Intercept β0

Graphical user interface, diagram

Description automatically generated

Figure 26: Posterior Distribution and Diagnostics for beta0

For intercept, it can be concluded that the chains converge and a perfect ESS is achieved with no significant autocorrelations, low MCSE and shrink factor less than 1.2. The mode value (-14.5) shown in posterior distribution indicates the base value for log odds for outcome regarding the condition of an individual’s lower back when all other variables are equal to 0.

#### Pelvic Tilt β1

Graphical user interface, diagram, schematic

Description automatically generated

Figure 27: Posterior Distribution and Diagnostics for predictor- pelvic tilt

With regards to diagnostics for pelvic tilt, it can be concluded that the chains converge and a perfect ESS is achieved with no significant autocorrelations, low MCSE and shrink factor less than 1.2. Furthermore, for a unit increase in pelvic tilt, the log odds of outcome would decrease by 0.072.

#### Sacral Slope β2

Graphical user interface, diagram

Description automatically generated

Figure 28: Posterior Distribution and Diagnostics for predictor- sacral slope

With regards to diagnostics for sacral slope, it can be concluded that the chains converge and a perfect ESS is achieved with no significant autocorrelations, low MCSE and shrink factor less than 1.2. Furthermore, for a unit increase in sacral slope, the log odds of outcome would increase by 0.0862.

#### Pelvic Radius β3

Diagram, schematic

Description automatically generated

Figure 29: Posterior Distribution and Diagnostics for predictor- pelvic radius

With regards to diagnostics for pelvic radius, it can be concluded that the chains converge and a perfect ESS is achieved with no significant autocorrelations, low MCSE and shrink factor less than 1.2. Furthermore, for a unit increase in pelvic radius, the log odds of outcome would increase by 0.106.

#### Degree of Spondylolisthesis β4

Graphical user interface

Description automatically generated

Figure 30: Posterior Distribution and Diagnostics for predictor- degree spondylolisthesis

With regards to diagnostics for degree of spondylolisthesis, it can be concluded that the chains converge and a perfect ESS is achieved with no significant autocorrelations, low MCSE and shrink factor less than 1.2. Furthermore, for a unit increase in degree of spondylolisthesis, the log odds of outcome would decrease by 0.149.

#### Guessing Parameter

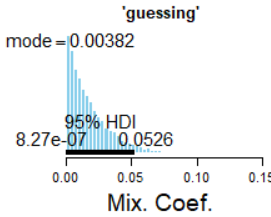


Figure 31: Distribution of alpha (guessing parameter)

The mode for guess parameter was computed to be 0.00382, which indicates an approximately 0.3% (~0%) chance of randomness being generated by guessing. This implies that the outliers do not hamper the model’s performance much. The model equation has been simplified to:

*Ө* = 0%.(1/2) + (1- 0%). logistic( β0 + β1 Pelvic\_tilt + β2 Sacral\_slope + β3 Pelvic\_radius + β4 Degree\_spondylolisthesis )

Or, ***Ө* = logistic( β0 + β1 Pelvic\_tilt + β2 Sacral\_slope + β3 Pelvic\_radius + β4 Degree\_spondylolisthesis )**

## 5.3. Checking Predictive Power

The predictive power of the models obtained in run 5, 6 and 7 were obtained to understand the impact of parameter alterations and feature selection. Hence, the accuracy, precision, recall and F-score was calculated for the 3 configurations using the 31 target observations in test data. The following results were obtained:

|  |  |  |  |
| --- | --- | --- | --- |
| **Metric** | **5th Run** | **6th Run** | **7th Run** |
| **Accuracy** | 90.30% | 90.30% | 90.30% |
| **Precision** | 1 | 1 | 1 |
| **Recall** | 0.88 | 0.88 | 0.88 |
| **F-score** | 0.936 | 0.936 | 0.936 |
| **Run Time** | 6:58 | 2:46 | 3:12 |

Table 11: Model Evaluation

Firstly, it was observed that the predictive power of the 3 models was surprisingly similar. The model was able to predict the condition of lower back of an individual with 90% accuracy. It was able to correctly identify 88% of the total abnormal lower back cases. A perfect precision score indicates that the identification of abnormal lower back was correct 100% of the time.

# Conclusion

To conclude, a robust logistic regression model for classifying the condition of an individual’s lower back was successfully built. Feature selection was performed to reduce the dimensions of model by using principles of correlation and statistical insignificance.

As per the results obtained, the model simulated in 6th run was considered ideal for prediction. The selection was based on principle of parsimony and efficiency. Of the 3 models considered for gauging predictive power, the configuration from 6th run took lesser computational time than 7th run and utilized lesser number of predictors than 5th configuration. This can be confirmed from the table below:

|  |  |  |  |
| --- | --- | --- | --- |
| **Metric** | **5th Run** | **6th Run** | **7th Run** |
| **No. of Predictors** | 11 | 5 | 5 |
| **Run Time** | 6:58 | 2:46 | 3:12 |

Table 12: Comparison of Mode Run Efficiency

The reduction of dimensions from 5th run to 6th run, yielded similar results in terms of predictive power. Hence, it was also inferred that the MCMC module automatically disregards the insignificant parameters for prediction. The MCMC configuration of chosen model are tabulated below:

|  |  |
| --- | --- |
| **Adapt Steps** | 1,000 |
| **Burn-in Steps** | 1,000 |
| **Chains** | 4 |
| **Thinning Steps** | 20 |
| **Saved Steps** | 1,500 |

Table 13: Final Specification for the MCMC Model

The final predictive model is given below:

**Y =** **-14.2 - 0.079\*Pelvic\_Tilt + 0.093\*Sacral\_Slope+ 0.107\*Pelvic\_Radius - 0.147\*Degree\_Spondylolisthesis**

For a unit increase in pelvic tilt and degree of spondylolisthesis, the log odds of outcome would decrease by 0.079 and 0.147, respectively. On the contrary, for a unit increase in sacral slope and pelvic radius, the log odds of outcome would increase by 0.093 and 0.107, respectively. Additionally, the first few predictions from the chosen model are given below:

|  |  |  |  |
| --- | --- | --- | --- |
| **Prediction** | **Mode** | **HDIlow** | **HDIhigh** |
| pred[1] | 0.00005 | 0.00000 | 0.00099 |
| pred[2] | 0.00001 | 0.00000 | 0.00015 |
| pred[3] | 0.00000 | 0.00000 | 0.00003 |
| pred[4] | 0.00000 | 0.00000 | 0.00007 |
| pred[5] | 0.00012 | 0.00000 | 0.00214 |
| pred[6] | 0.00 | 0.00 | 0.00 |
| pred[7] | 0.00 | 0.00 | 0.02 |
| pred[8] | 0.65 | 0.36 | 0.90 |

Table 14: Predictions from the chosen model

It can be observed for 1st prediction that the probability of normal lower back was 0.00005 (~0). This probability extends from 0 to 0.00099. This means that the probability of lower back being abnormal is very high. Similar inferences can be drawn for predictions 1 to 7. However, in case of 8th prediction, the probability of normal lower back was 0.65, which is well above the 0.5 threshold. This means that the individual had a higher probability of having a normal lower back.

The model was able to predict the condition of lower back of an individual with **90%** accuracy.

|  |  |  |  |
| --- | --- | --- | --- |
| **Confusion Matrix** | | **Actual** | |
| **Abnormal** | **Normal** |
| **Predicted** | **Abnormal** | 22 | 0 |
| **Normal** | 3 | 6 |

Table 15: Confusion Matrix for Model's Predictive Power

# References

Akhil Chhatre, M., n.d. *Lower Back Pain: What Could It Be?.* [Online]   
Available at: https://www.hopkinsmedicine.org/health/conditions-and-diseases/back-pain/lower-back-pain-what-could-it-be  
[Accessed 06 October 2020].

J. C. Le Huec, S. A. L. P. a. P. N., 2011. Pelvic parameters: origin and significance. *European Spine Journal,* 20(Suppl 5), p. 564–571.

John Peloza, M., 2017. *Lower Back Pain Symptoms, Diagnosis, and Treatment.* [Online]   
Available at: https://www.spine-health.com/conditions/lower-back-pain/lower-back-pain-symptoms-diagnosis-and-treatment  
[Accessed 07 October 2020].

Lancaster, O. A. o., n.d. *Spondylolisthesis - Slipped Vertebrae.* [Online]   
Available at: Spondylolisthesis - Slipped Vertebrae   
[Accessed 0.8 October 2020].

Legaye, J., 2007. The femoro-sacral posterior angle: an anatomical sagittal pelvic parameter usable with dome-shaped sacrum. *European Spine Journal,* 16(2), pp. 219-225.

Lin RM, J. I. Y. C., 1992. Lumbar lordosis: normal adults. *ournal of the Formosan Medical Association = Taiwan yi zhi,* 91(3), pp. 329-333.

Stroke, N. I. o. N. D. a., n.d. *Low Back Pain Fact Sheet.* [Online]   
Available at: https://www.ninds.nih.gov/Disorders/Patient-Caregiver-Education/Fact-Sheets/Low-Back-Pain-Fact-Sheet#top  
[Accessed 07 October 2020].

User-sammy123, 2016. *Lower Back Pain Symptoms Dataset.* [Online]   
Available at: https://www.kaggle.com/sammy123/lower-back-pain-symptoms-dataset  
[Accessed 20 Sept 2020].

# Appendix

# MATH2269 Final Project Codes in R

graphics.off() # Closing all of previous R's graphics windows.

rm(list=ls()) # CLearing R Memory

library(ggplot2)

library(ggpubr)

library(ks)

library(rjags)

library(runjags)

setwd("E:\\ABS - Applied Bayesian Stats\\z JAGS")

source("DBDA2E-utilities.R")

#setwd("~/Documents/MATH2269\_Bayesian/2020/presentations/Module 7/Application2/")

#===============PRELIMINARY FUNCTIONS FOR POSTERIOR INFERENCES====================

smryMCMC\_HD = function( codaSamples , compVal = NULL, saveName=NULL) {

summaryInfo = NULL

mcmcMat = as.matrix(codaSamples,chains=TRUE)

paramName = colnames(mcmcMat)

for ( pName in paramName ) {

if (pName %in% colnames(compVal)){

if (!is.na(compVal[pName])) {

summaryInfo = rbind( summaryInfo , summarizePost( paramSampleVec = mcmcMat[,pName] ,

compVal = as.numeric(compVal[pName]) ))

}

else {

summaryInfo = rbind( summaryInfo , summarizePost( paramSampleVec = mcmcMat[,pName] ) )

}

} else {

summaryInfo = rbind( summaryInfo , summarizePost( paramSampleVec = mcmcMat[,pName] ) )

}

}

rownames(summaryInfo) = paramName

# summaryInfo = rbind( summaryInfo ,

# "tau" = summarizePost( mcmcMat[,"tau"] ) )

if ( !is.null(saveName) ) {

write.csv( summaryInfo , file=paste(saveName,"SummaryInfo.csv",sep="") )

}

return( summaryInfo )

}

#===============================================================================

plotMCMC = function( codaSamples , data , xName="x" , yName="y" ,

showCurve=FALSE , pairsPlot=FALSE ,

saveName=NULL , saveType="jpg" ) {

# showCurve is TRUE or FALSE and indicates whether the posterior should

# be displayed as a histogram (by default) or by an approximate curve.

# pairsPlot is TRUE or FALSE and indicates whether scatterplots of pairs

# of parameters should be displayed.

#-----------------------------------------------------------------------------

y = data[,yName]

x = as.matrix(data[,xName])

mcmcMat = as.matrix(codaSamples,chains=TRUE)

chainLength = NROW( mcmcMat )

guess = mcmcMat[,"guess"]

zbeta0 = mcmcMat[,"zbeta0"]

zbeta = mcmcMat[,grep("^zbeta$|^zbeta\\[",colnames(mcmcMat))]

if ( ncol(x)==1 ) { zbeta = matrix( zbeta , ncol=1 ) }

beta0 = mcmcMat[,"beta0"]

beta = mcmcMat[,grep("^beta$|^beta\\[",colnames(mcmcMat))]

if ( ncol(x)==1 ) { beta = matrix( beta , ncol=1 ) }

#-----------------------------------------------------------------------------

if ( pairsPlot ) {

# Plot the parameters pairwise, to see correlations:

openGraph()

nPtToPlot = 1000

plotIdx = floor(seq(1,chainLength,by=chainLength/nPtToPlot))

panel.cor = function(x, y, digits=2, prefix="", cex.cor, ...) {

usr = par("usr"); on.exit(par(usr))

par(usr = c(0, 1, 0, 1))

r = (cor(x, y))

txt = format(c(r, 0.123456789), digits=digits)[1]

txt = paste(prefix, txt, sep="")

if(missing(cex.cor)) cex.cor <- 0.8/strwidth(txt)

text(0.5, 0.5, txt, cex=1.5 ) # was cex=cex.cor\*r

}

pairs( cbind( beta0 , beta , guess )[plotIdx,] ,

labels=c( "beta[0]" ,

paste0("beta[",1:ncol(beta),"]\n",xName) , "guessing" ) ,

lower.panel=panel.cor , col="skyblue" )

if ( !is.null(saveName) ) {

saveGraph( file=paste(saveName,"PostPairs",sep=""), type=saveType)

}

}

#-----------------------------------------------------------------------------

# Data with posterior predictive:

# If only 1 predictor:

if ( ncol(x)==1 ) {

openGraph(width=7,height=6)

par( mar=c(3.5,3.5,2,1) , mgp=c(2.0,0.7,0) )

plot( x[,1] , y , xlab=xName[1] , ylab=yName ,

cex=2.0 , cex.lab=1.5 , col="black" , main="Data with Post. Pred." )

abline(h=0.5,lty="dotted")

cVec = floor(seq(1,chainLength,length=30))

xWid=max(x)-min(x)

xComb = seq(min(x)-0.1\*xWid,max(x)+0.1\*xWid,length=201)

for ( cIdx in cVec ) {

lines( xComb ,

guess[cIdx]\*0.5

+ (1.0-guess[cIdx])\*1/(1+exp(-(beta0[cIdx]+beta[cIdx,1]\*xComb ))) ,

lwd=1.5 ,

col="skyblue" )

xInt = -beta0[cIdx]/beta[cIdx,1]

arrows( xInt,0.5, xInt,-0.04, length=0.1 , col="skyblue" , lty="dashed" )

}

if ( !is.null(saveName) ) {

saveGraph( file=paste(saveName,"DataThresh",sep=""), type=saveType)

}

}

# If only 2 predictors:

if ( ncol(x)==2 ) {

openGraph(width=7,height=7)

par( mar=c(3.5,3.5,2,1) , mgp=c(2.0,0.7,0) )

plot( x[,1] , x[,2] , pch=as.character(y) , xlab=xName[1] , ylab=xName[2] ,

col="black" , main="Data with Post. Pred.")

cVec = floor(seq(1,chainLength,length=30))

for ( cIdx in cVec ) {

abline( -beta0[cIdx]/beta[cIdx,2] , -beta[cIdx,1]/beta[cIdx,2] , col="skyblue" )

}

if ( !is.null(saveName) ) {

saveGraph( file=paste(saveName,"DataThresh",sep=""), type=saveType)

}

}

#-----------------------------------------------------------------------------

# Marginal histograms:

decideOpenGraph = function( panelCount , saveName , finished=FALSE ,

nRow=1 , nCol=3 ) {

# If finishing a set:

if ( finished==TRUE ) {

if ( !is.null(saveName) ) {

saveGraph( file=paste0(saveName,ceiling((panelCount-1)/(nRow\*nCol))),

type=saveType)

}

panelCount = 1 # re-set panelCount

return(panelCount)

} else {

# If this is first panel of a graph:

if ( ( panelCount %% (nRow\*nCol) ) == 1 ) {

# If previous graph was open, save previous one:

if ( panelCount>1 & !is.null(saveName) ) {

saveGraph( file=paste0(saveName,(panelCount%/%(nRow\*nCol))),

type=saveType)

}

# Open new graph

openGraph(width=nCol\*7.0/3,height=nRow\*2.0)

layout( matrix( 1:(nRow\*nCol) , nrow=nRow, byrow=TRUE ) )

par( mar=c(4,4,2.5,0.5) , mgp=c(2.5,0.7,0) )

}

# Increment and return panel count:

panelCount = panelCount+1

return(panelCount)

}

}

# Original scale:

panelCount = 1

panelCount = decideOpenGraph( panelCount , saveName=paste0(saveName,"PostMarg") )

histInfo = plotPost( beta0 , cex.lab = 1.75 , showCurve=showCurve ,

xlab=bquote(beta[0]) , main="Intercept" )

for ( bIdx in 1:ncol(beta) ) {

panelCount = decideOpenGraph( panelCount , saveName=paste0(saveName,"PostMarg") )

histInfo = plotPost( beta[,bIdx] , cex.lab = 1.75 , showCurve=showCurve ,

xlab=bquote(beta[.(bIdx)]) , main=xName[bIdx] )

}

panelCount = decideOpenGraph( panelCount , saveName=paste0(saveName,"PostMarg") )

histInfo = plotPost( guess , cex.lab = 1.75 , showCurve=showCurve ,

xlab=bquote("Mix. Coef.") , main="'guessing'" )

panelCount = decideOpenGraph( panelCount , finished=TRUE , saveName=paste0(saveName,"PostMarg") )

# Standardized scale:

panelCount = 1

panelCount = decideOpenGraph( panelCount , saveName=paste0(saveName,"PostMargZ") )

histInfo = plotPost( zbeta0 , cex.lab = 1.75 , showCurve=showCurve ,

xlab=bquote(z\*beta[0]) , main="Intercept" )

for ( bIdx in 1:ncol(beta) ) {

panelCount = decideOpenGraph( panelCount , saveName=paste0(saveName,"PostMargZ") )

histInfo = plotPost( zbeta[,bIdx] , cex.lab = 1.75 , showCurve=showCurve ,

xlab=bquote(z\*beta[.(bIdx)]) , main=xName[bIdx] )

}

panelCount = decideOpenGraph( panelCount , finished=TRUE , saveName=paste0(saveName,"PostMargZ") )

#-----------------------------------------------------------------------------

}

#===============PRELIMINARY FUNCTIONS FOR POSTERIOR INFERENCES====================

#Loading the data file

myData <- read.csv("train\_spine.csv")

all\_var = c("pelvic\_incidence",

"pelvic\_tilt",

"lumbar\_lordosis\_angle",

"sacral\_slope",

"pelvic\_radius",

"degree\_spondylolisthesis",

"pelvic\_slope",

"Direct\_tilt",

"thoracic\_slope",

"cervical\_tilt",

"sacrum\_angle",

"scoliosis\_slope", "Class\_att")

colnames(myData) <- all\_var

head(myData)

#Converting the dependent variable to numeric

# Assigning 1 as Normal body structure and 0 as abnormal body structure

myData$Class\_att <- as.numeric(as.factor(myData$Class\_att)) - 1 # To get 0/1 instead of 1/2; Abnormal = 0; Normal = 1

head(myData)

# Summary Statistics

summary(myData)

#=== Descriptive look ===

# Distribution of Class Attrbute/ Target

ggplot(myData, aes(x=factor(Class\_att)))+

geom\_bar( width=0.7, fill="steelblue")+

theme\_minimal()+ labs(title="Frequency of Target Variable",

x="Type of Body Structure (0:Abnormal, 1:Normal)", y = "No. of Observations")

# Scatter plots

myplots <- list() # new empty list

for (i in 1:12) {

p1 <- eval(substitute(

ggplot(data=myData,aes(x=myData[ ,i], y = Class\_att))+

geom\_point() +

xlab(colnames(myData)[ i])

,list(i = i)))

print(i)

print(p1)

myplots[[i]] <- p1 # add each plot into plot list

}

figure = ggarrange(myplots[[1]],myplots[[2]],myplots[[3]],myplots[[4]],myplots[[5]],myplots[[6]],myplots[[7]],myplots[[8]],myplots[[9]],myplots[[10]],myplots[[11]],myplots[[12]], ncol=3, nrow = 4)

figure

# THE DATA.

y = myData[,"Class\_att"]

x = as.matrix(myData[,c(1:12)])

cat("\nCORRELATION MATRIX OF PREDICTORS:\n ")

show(round(cor(x),3) )

cat("\n")

x = as.matrix(myData[,c(2,4:6)])

# Loading the test data i.e. for evaluating the model

PredData = read.csv("test\_spine.csv")

colnames(PredData) <- all\_var

head(PredData)

PredData = PredData[,c(2,4:6,13)]

PredData$Class\_att <- as.numeric(as.factor(PredData$Class\_att)) - 1

summary(PredData)

xPred = as.matrix(PredData[,c(1:5)])

Nx = ncol(x)

PredData

# Specifying the data in a list, for later shipment to JAGS:

dataList <- list(

x = x ,

y = y ,

xPred = xPred ,

Ntotal = length(y),

Nx = Nx,

Npred = nrow(xPred)

)

modelString = "

# Standardize the data:

data {

for ( j in 1:Nx ) {

xm[j] <- mean(x[,j])

xsd[j] <- sd(x[,j])

for ( i in 1:Ntotal ) {

zx[i,j] <- ( x[i,j] - xm[j] ) / xsd[j]

}

}

}

# Specify the model for standardized data:

model {

for ( i in 1:Ntotal ) {

# In JAGS, ilogit is logistic:

y[i] ~ dbern( mu[i] )

mu[i] <- ( guess\*(1/2)

+ (1.0-guess)\*ilogit(zbeta0+sum(zbeta[1:Nx]\*zx[i,1:Nx])) )

}

# Priors vague on standardized scale:

zbeta0 ~ dnorm( 0 , 1/2^2 )

for ( j in 1:Nx ) {

zbeta[j] ~ dnorm( 0 , 1/2^2 )

}

guess ~ dbeta(1,9)

# Transform to original scale:

beta[1:Nx] <- zbeta[1:Nx] / xsd[1:Nx]

beta0 <- zbeta0 - sum( zbeta[1:Nx] \* xm[1:Nx] / xsd[1:Nx] )

# Compute predictions at every step of the MCMC

for ( k in 1:Npred){

pred[k] <- ilogit(beta0 + sum(beta[1:Nx] \* xPred[k,1:Nx]))

}

}

" # close quote for modelString

# Write out modelString to a text file

writeLines( modelString , con="TEMPmodel.txt" )

parameters = c( "zbeta0" , "beta0")

for ( i in 1:Nx){

parameters = c(parameters, paste0("zbeta[",i,"]"))

}

for ( i in 1:Nx){

parameters = c(parameters, paste0("beta[",i,"]"))

}

for ( i in 1:nrow(xPred)){

parameters = c(parameters, paste0("pred[",i,"]"))

}

# Specification of MCMC model runs parameters

parameters = c(parameters, "guess")

adaptSteps = 1000 # Number of steps to "tune" the samplers

burnInSteps = 1000

nChains = 4

thinSteps = 25

numSavedSteps = 2000

nIter = ceiling( ( numSavedSteps \* thinSteps ) / nChains )

runJagsOut <- run.jags( method="parallel" ,

model="TEMPmodel.txt" ,

monitor=parameters ,

data=dataList ,

# inits=initsList ,

n.chains=nChains ,

adapt=adaptSteps ,

burnin=burnInSteps ,

sample=numSavedSteps ,

thin=thinSteps , summarise=FALSE , plots=FALSE )

codaSamples = as.mcmc.list( runJagsOut )

#save.image(file="run-.RData")

#load(file="run-5.RData") # Load the results with 124,000 iterations

diagMCMC( codaSamples , parName="beta0" )

for ( i in 1:Nx){

diagMCMC( codaSamples , parName=paste0("beta[",i,"]") )

}

compVal <- data.frame("beta0" = 15, "beta[1]" = 0, "beta[2]" = 0, "beta[3]" = 0, "beta[4]" = 0, "beta[5]" = 0,

"beta[6]" = 0, check.names=FALSE)

summaryInfo <- smryMCMC\_HD( codaSamples = codaSamples , compVal = NULL )

print(summaryInfo)

plotMCMC(codaSamples = codaSamples , data = myData, xName=c("pelvic\_tilt",

"sacral\_slope",

"pelvic\_radius",

"degree\_spondylolisthesis"), yName="Class\_att")

# Predictions for full records in training set

preds <- data.frame(PredProb = summaryInfo[12:42,3])

threshold <- 0.5# summaryInfo[427,3]

preds[which(preds$PredProb<threshold),2] <- 0

preds[which(preds$PredProb>threshold),2] <- 1

table(preds)

PredData[,5]

actual = PredData[,5]

# ============ Predictive check ============

confusionMatrix <- function(resp, pred){

classRes <- data.frame(response = resp , predicted = pred)

conf = xtabs(~ predicted + response, data = classRes)

accuracy = sum(diag(conf))/sum(conf)

accuracy

precision = conf[1,1]/(conf[1,1]+conf[1,2])

precision

recall = conf[1,1]/(conf[1,1]+conf[2,1])

recall

Fscore = 2\*((precision\*recall)/(precision+recall))

Fscore

return(list(accuracy = accuracy, precision = precision, recall = recall, Fscore = Fscore, conf = conf))

}

confusionMatrix(resp = PredData[,5], pred = preds[,2])