Insurance Cost Prediction using Bayesian Analysis

Project Report

Rinku Bajaj (S3672522) Siddhant Gehlot (S3620089) Namita Chibba (S3631442) Ravi Pandey (S3638787)

21 October 2018

Introduction

The aim of the project is to predict the health care costs that will be incurred by the clients of an Insurance Company using Bayesian Analysis. The data set for this project is sourced from `Machine Learning with R by Brett Lantz` and downloaded from Kaggle. The data set consist of seven variables and six of them are the independent variables and one dependent variable.

The Independent variables are:

* Age: Age of the insurance contractor

* Sex: Insurance Contractor gender (Female/Male)

* BMI: Body Mass Index in kg/m^2 (Ideally between 24.9)

* Children: Number of children covered by health insurance

* Smoker: Whether the client smokes or not (Yes/No)

* Region: Resident area in the US (NorthEast, NorthWest, SouthEast, SouthWest)

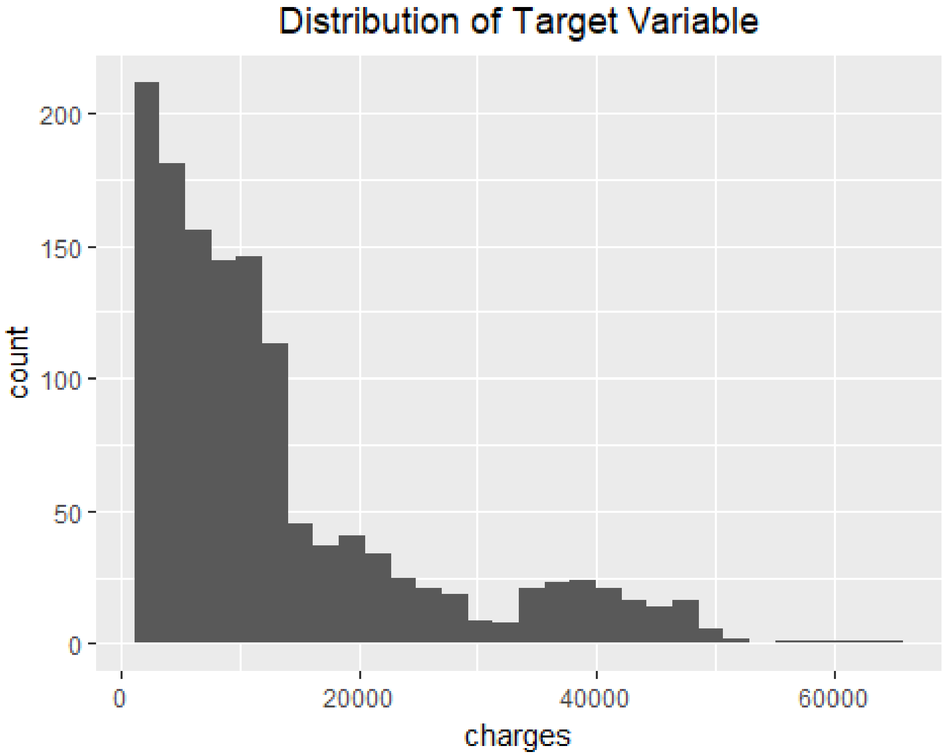
The Dependent variable is:

* Charges: Individual Medical costs billed by the client

Since the data set has three categorial variables namely `sex`, `smoker` and `region`, they will be converted into dummy variables to make them appropriate for the analysis.

Visual Analysis

Analysis of Target Variable

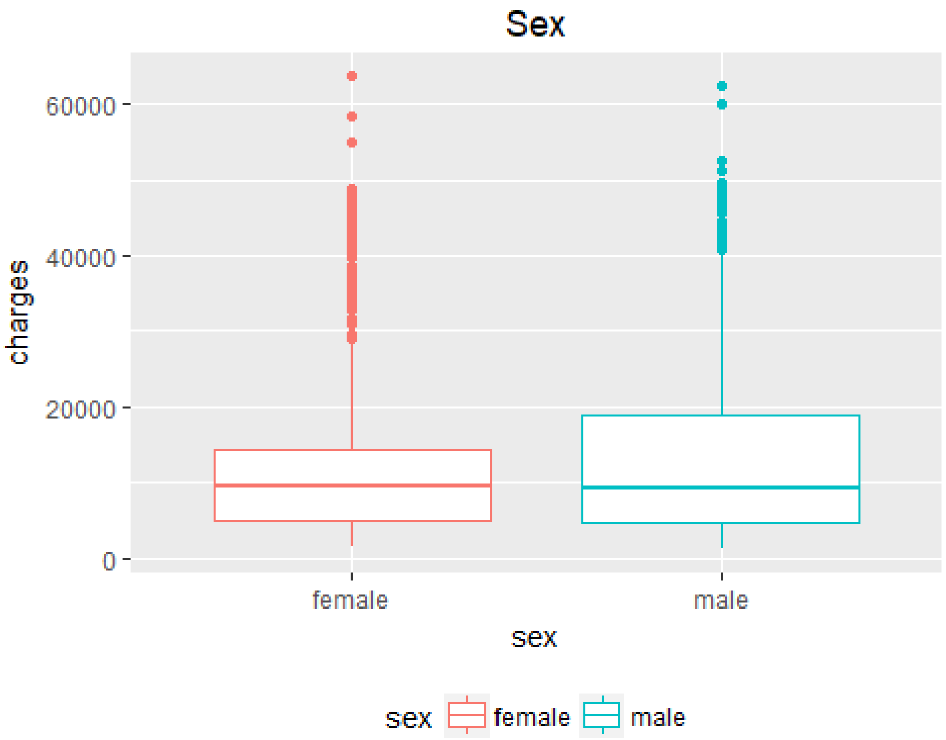


Distribution of Target Variable

Looking at the distribution of the target variable we find that it is highly skewed to the right, with value of the charges varying between 1000 to above 60000.

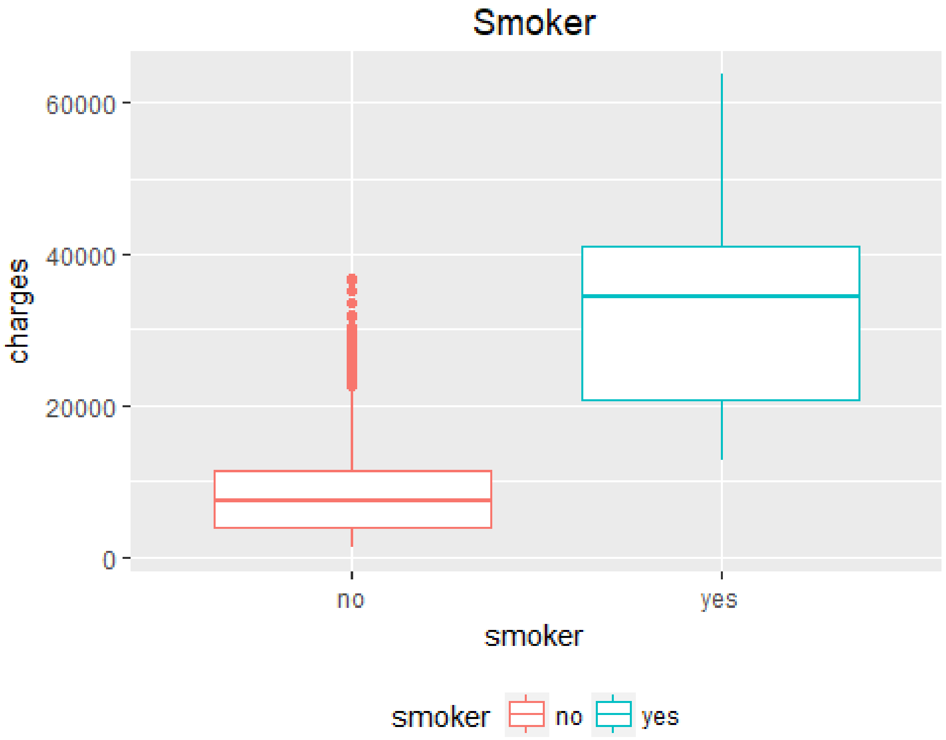
Analysis of Categorical Variables

Gender



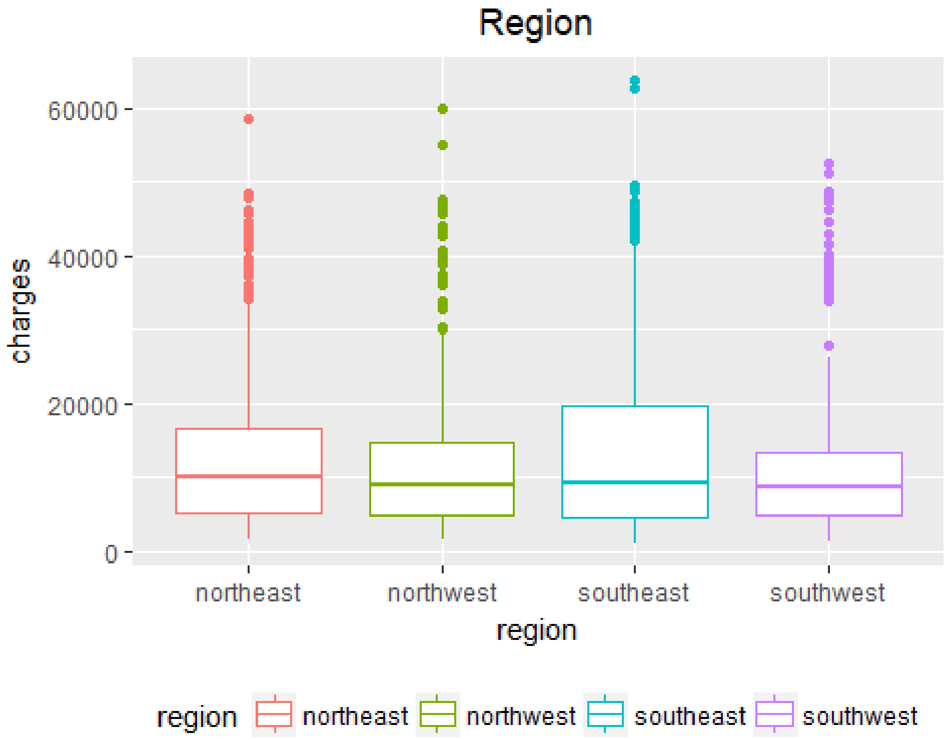
The Box plot shows that the insurance charges for the male clients are generally higher than the female clients. We also observe that gender is not much of an informing attribute for the insurance charge prediction because the distribution for both the sexes is almost similar with just the apparent difference that IQR for males is bit larger than that of the females but the median is same for both the genders. We can say from the difference in the upper quartile of males that male clients have costlier insurance. Moreover, both genders have heavy trail of outliers.

Somker/Non-smoker



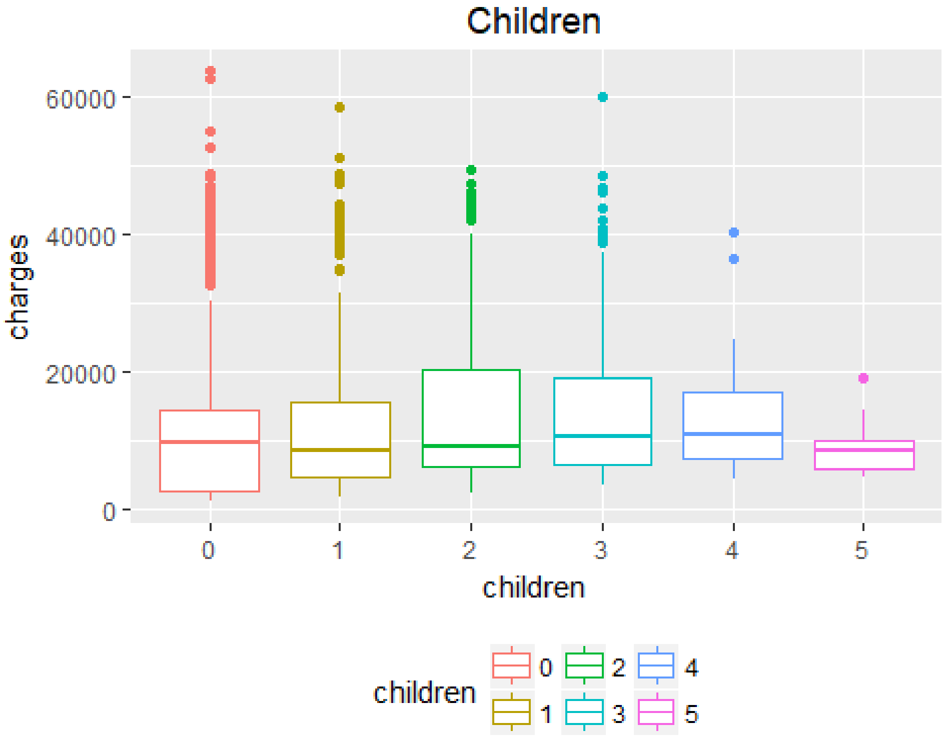
The factor that the client is a smoker or not largely affects the insurance charges which can be confirmed in the box plot above. The clients who are not smokers tend to be charged less than the clients who are smoker. The difference in the prices is also quite large as the minimum amount charged for a client who smokes is almost $10,000 more than the non smoking clients.

Region



The insurance charges for the clients also vary with different regions. Even though the minimum amount charged for clients in every region is similar, but it is clear from the plot that the clients in southeast region tend to pay more for the insurance followed by the northeastern popullation. This may be because the clients living in these regions are wealthier than the clients living in other regions. Higher crime rate in these regions may also be a reason for the higher insurance prices. Furthermore, like gender, all the regions have many outliers.

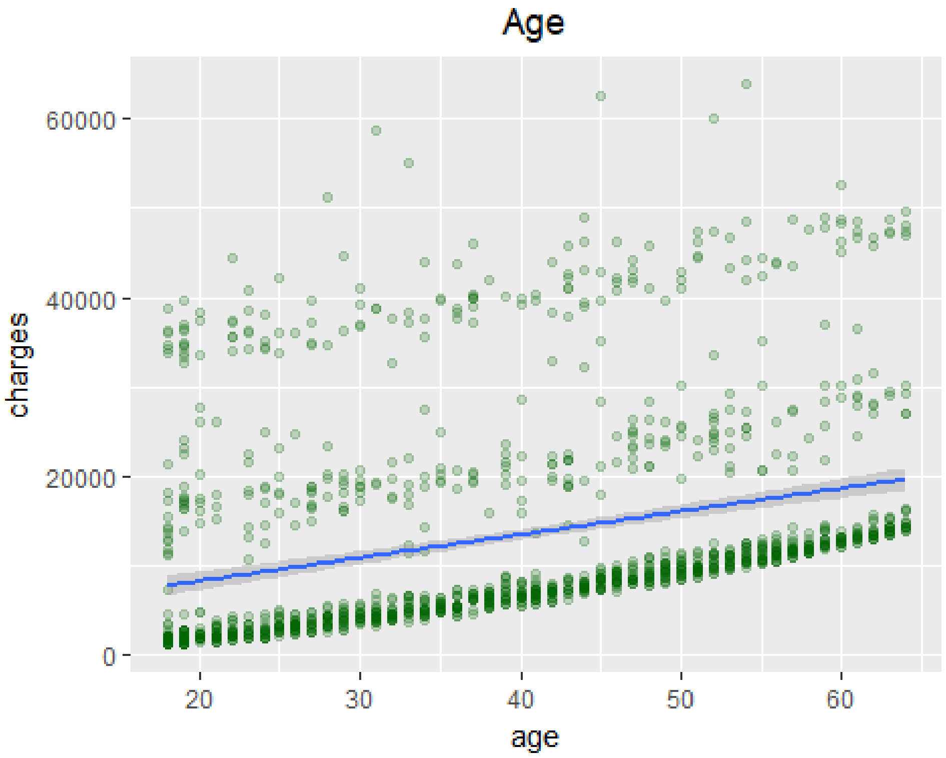
Children



This plot shows the variation in the insurance charges with number of children in the family. The maximum amount paid by the client for insurace increases with the increase in the number of children which is clear from the box plot. But we can also notice a drop in the maximum insurance charge when the number of children in the family are three or more. This may be because of the reason that not a lot of families in the client list of the insurance company have more than three children.

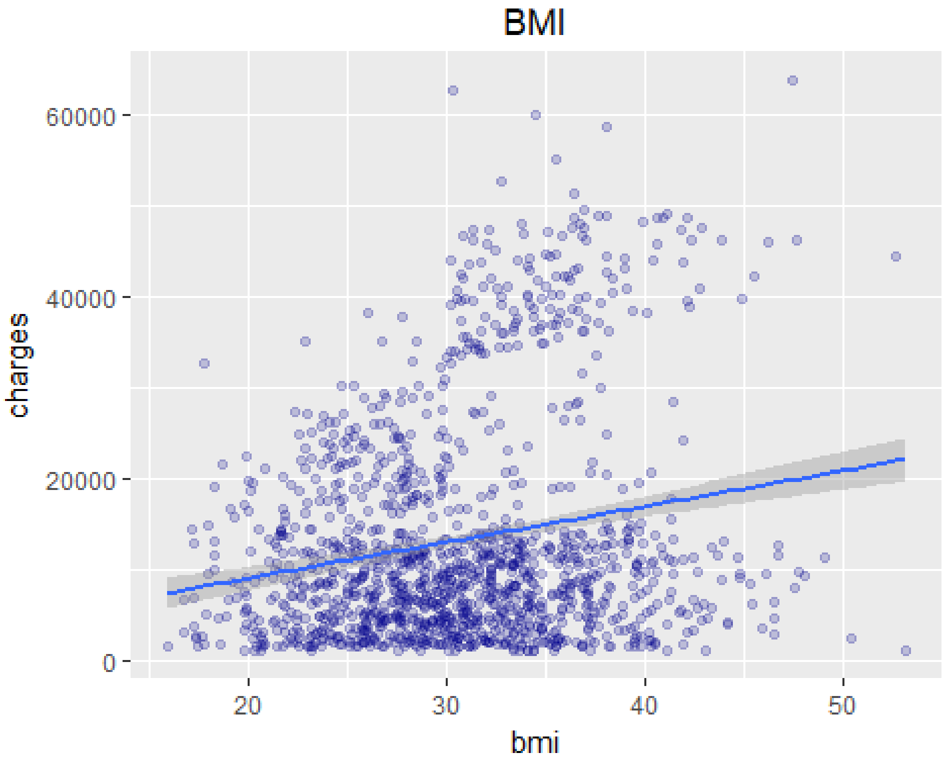
Analysis of Numeric Variables

Age



The scatter plot for the age versus insurance charges clearly shows that with increase in the age of clients, the medical cost also increases and also every age has wide range of charges associated with it. Though there seems to be the linear trend between age and charges but there seems to be some kind of segregation within every age group of people. We can state from this that the liner regression line does not fit the data and raises a question about the reason for this segregation.

Body-Mass Index



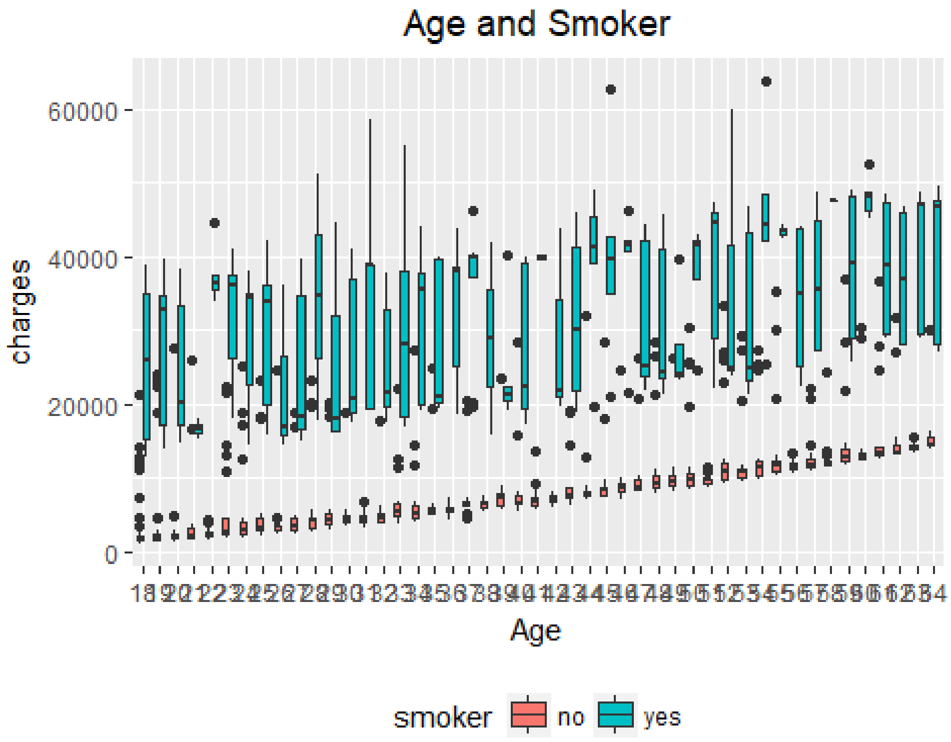
The BMI ratio also plays a role in deciding the charges for insurance of a client. We observe from the plot that there may be two separate groups. One tending towards higher insurance charges and other limited towards lower charges. This may be because healthy and fit clients tend to pay less for the insurance as the chances of them getting sick are very low as compared to that of unhealthy and obese clients.

But the reason for two different groups in BMI is still not very clear from the plot.

Multivariate Analysis

To further investigate the data, multivariate analysis for the age and smoker variables is done with the target variable to check the relationship between them.

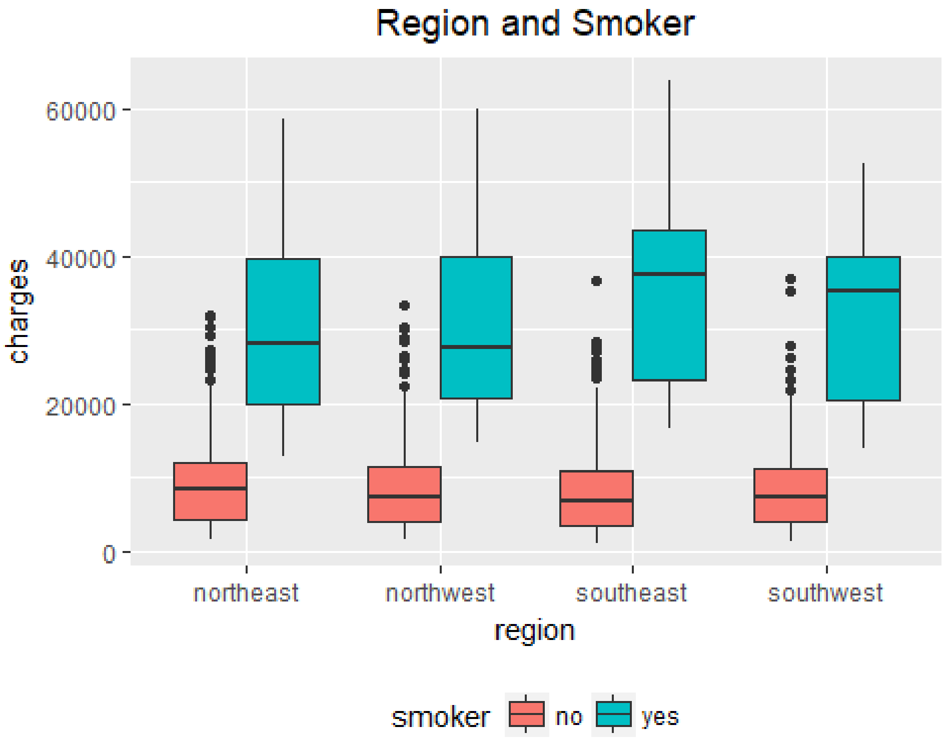
Age & Smoker vs Charges



Analysis of Age and Smoker vs Medical Costs

From the above boxplot it becomes much clearer that the insurance charges are dependent more on the fact if the client is a smoker or not because in different age groups, the smokers are charged more by the insurance company than a non-smoker. From here on we will analyze the effect of smoking habits paired with other factors on predicting the amount of insurance charges.

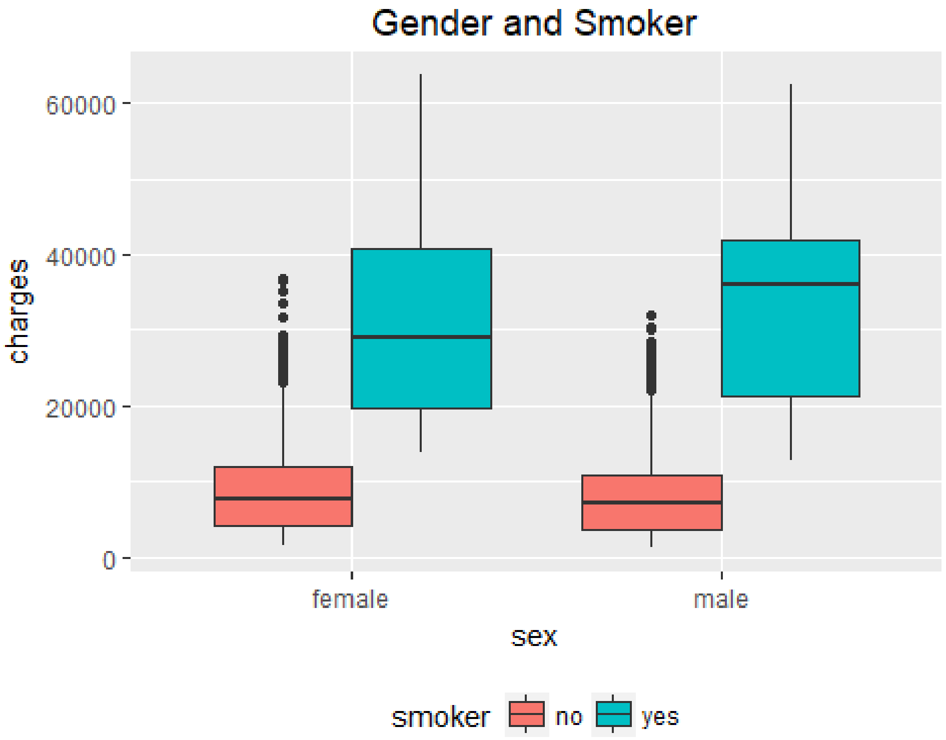
Region & Smoker vs Charges



Analysis of Region and Smoker vs Medical Costs

The boxplot above shows the clients from different regions seperated by their smoking habbits. It is clear from the plot that the smokers tend to pay more for insurance than the non-smokers. This plot now explains the outliers in the previous plot for regions where the smokers from different regions were considered as outliers.

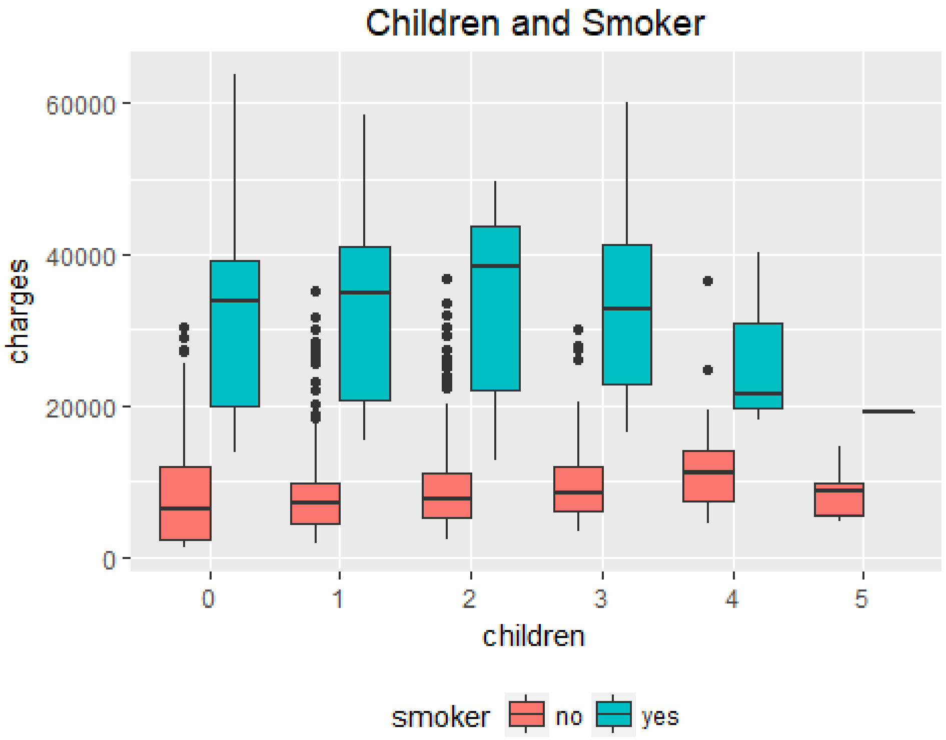
Gender & Smoker vs Charges



Analysis of Gender and Smoker vs Medical Costs

The gender and smoker vs charges boxplot explains the trail of outliers seen in the previous gender vs charges plot. It is clear here in this plot as well that the smokers are generally charged more than non-smokers no matter the gender of the client.

Children & Smoker vs Charges



Analysis of Children and Smoker vs Medical Costs

We can notice the similar outcomes for smokers as smokers with any number of children tend to pay higher insurance charges in comparison to the corresponding non-smokers.

Therefore, we can say that “smoker” is the most important attribute deciding the insurance charges of the individual. Every other predictor like age, region, gender is connected to the charges through “smoker” attribute.

Data Preprocessing

The dataset contained numerical as well as categorical attributes. The categorical attributes were converted into dummy variables as follows:

```
insurance <- insurance %>% mutate(bmi_cat = cut(bmi,
                                                breaks = c(15, 30, 55),
                                                labels = c("bmi_less_30",
"bmi_greater_30")))

insurance <- cbind(insurance, dummy(insurance$sex, sep = "_"),
dummy(insurance$region, sep = "_"),
                   dummy(insurance$smoker, sep = "_"),
dummy(insurance$bmi_cat, sep = "_"))

insurance <- insurance[,c(1,17,4,9,16,11:13,7)]

insurance_pred = insurance[1328:1338,]

insurance = insurance[1:20,]
```

The categorical variables were converted into the following dummy variables:

* Insurance_female: 0 (Male) and 1 (Female)

* Insurance_smoker: 0 (Non-Smoker) and 1 (Smoker)

* Insurance_northeast: 0 (Not NorthEast) and 1 (NorthEast)

* Insurance_northwest: 0 (Not Northwest) and 1 (Northwest)

* Insurance_southeast: 0 (Not Southeast) and 1 (SouthEast)

Also, in the case when the above three regions are all `0`, the fourth region i.e. `SouthWest` will be represented i.e.

* Insurance_northeast: 0 (Not NorthEast)

* Insurance_northwest: 0 (Not Northwest)

* Insurance_southeast: 0 (Not Southeast)

Linear Regression Model

This section deals with the Bayesian parameter estimation for the linear regression model.

The Bayesian regression equation can be defined as follows:

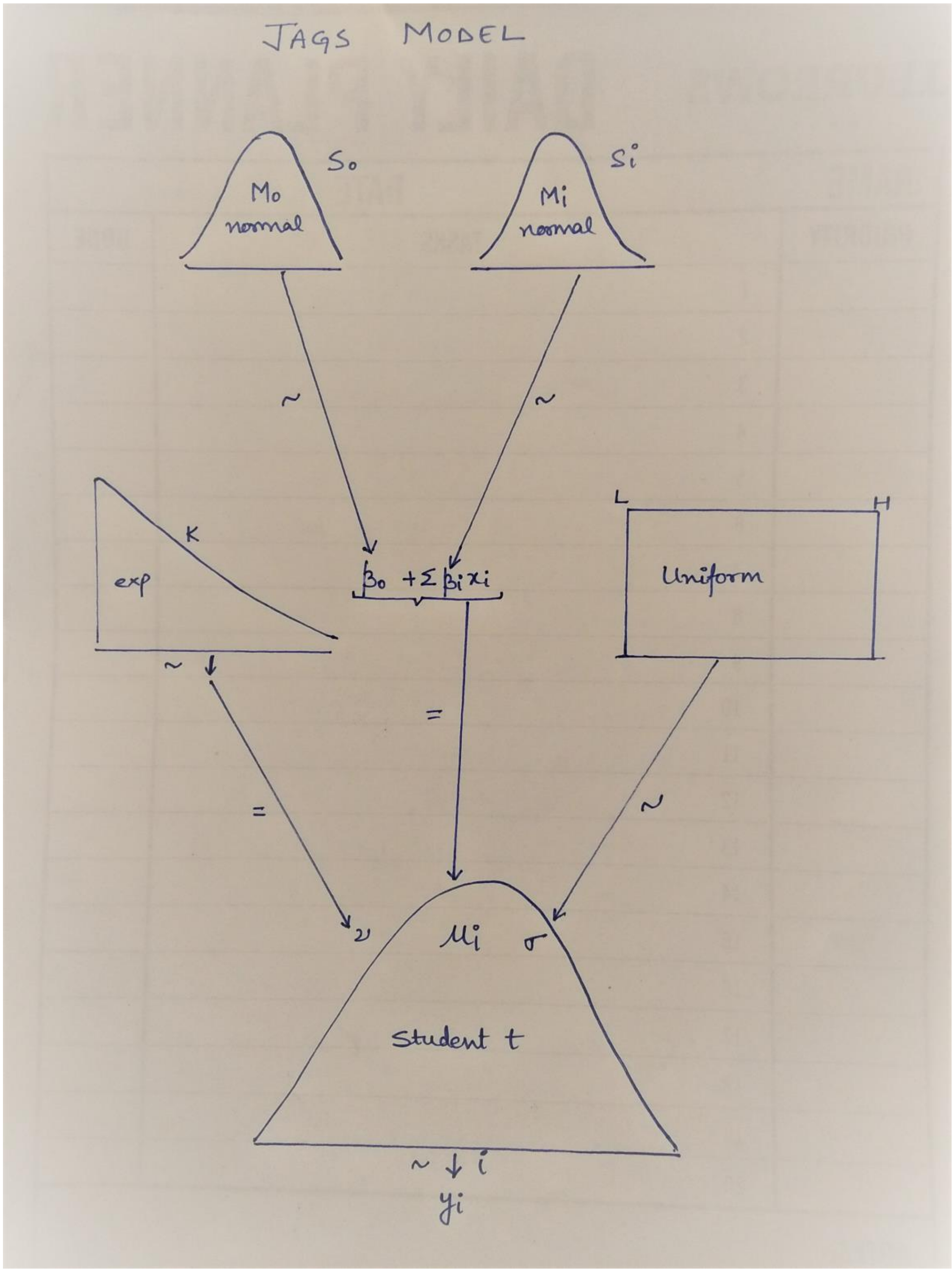
Y = beta[0] + beta[1]X[1] + beta[2]X[2] + beta[3]X[3] + beta[4]X[4] + beta[5]X[5] + beta[6]X[6] + beta[7]X[7] + beta[8]X[8]

Here the target variable Y is our predictor i.e. Insurance Charges , beta[1:NX] are the Bayesian Regression Co- Efficients, X[1:NX] are the precitor variables i.e. Age, BMI, Smoking Habits and so on.

Since it is evident from our visual analysis that the greater a person's BMI the higher is his health care cost, hence we decided to convert BMI from continuous variable to a categorical variables with two categories namely `bmi_less_30` and `bmi_greater_30` to improve the results. We chose to split the dataset at BMI of 30 and above after inspecting the scatter plot of BMI vs Medical Charges, which showed a clear existence of health costs spirally up as soon as the person's BMI crosses 30.

Model Diagram

The Model Diagram for Jags is as follows:



JAGS MODEL

Model Specification

The JAGS model has a t distribution for Bayesian Regression Model estimation, normal distribution for the bayesian co-efficient estimates, uniform distribution for bayesian sigma estimate and exponential distribution for nu parameter in the t distribution. It can be seen as follows:

```
  # Specify the model for standardized data:
# model
# {
#   for ( i in 1:Ntotal )
#     {
#       zy[i] ~ dt( zbeta0 + sum( zbeta[1:Nx] * zx[i,1:Nx] ) , 1/zsigma^2 , 
nu )
#     }
# # Priors vague on standardized scale:
#   zbeta0 ~ dnorm( muZ0 , 1/Var0 )
#   for ( j in 1:Nx )
#     {
#       zbeta[j] ~ dnorm( muZ[j] , 1/VarZ[j] )
#     }
#   zsigma ~ dgamma(0.01, 0.01)
#   nu ~ dexp(1/30.0)
#
# }
```

Data Block

The data block passed to JAGS model is as follows:

```
# Standardize the data:
# data
# {
#   ym <- mean(y)
#   ysd <- sd(y)
#   for ( i in 1:Ntotal )
#   {
#     zy[i] <- ( y[i] - ym ) / ysd
#   }
#   for ( j in 1:Nx )
#   {
#     xm[j]  <- mean(x[,j])
#     xsd[j] <-   sd(x[,j])
#     for ( i in 1:Ntotal )
#     {
#       zx[i,j] <- ifelse(j<3, ( x[i,j] - xm[j] ) / xsd[j], x[i,j] )
#     }
#   }
# }
```

Prior Estimates

The mean and sigma estimates for priors have been set as follows:

Mean:

```
+ mu0 =   0
+ mu1 = 100
+ mu2 = 150
+ mu3 = 200
+ mu4 = 300
+ mu5 = 200
+ mu6 = 200
+ mu7 = 200
```

Variance:

```
+ var0 = 0.1
+ var1 = 100
+ var2 = 500
+ var3 = 100
+ var4 = 500
+ var5 = 120
+ var6 = 150
+ var7 = 170
```

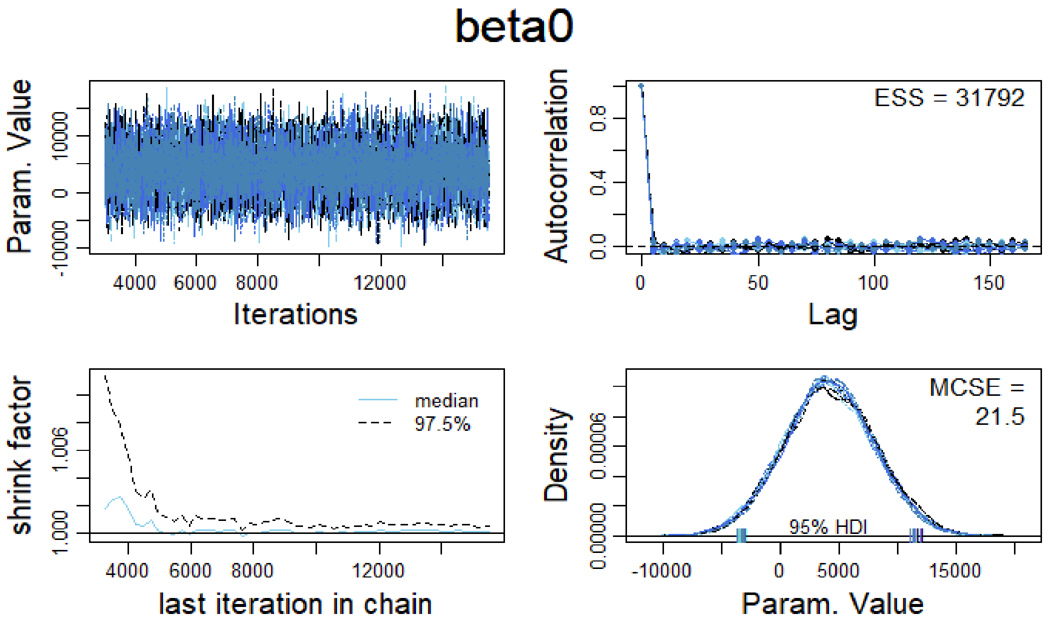
Note: For constructing the JAGS Model, we standardized the sample data and mean estimates of our prior distributions. These standardized data points are then given to the data block. However, for predictions we had to back transform the prior estimates, likelihood parameters and Bayesian Regression estimates.

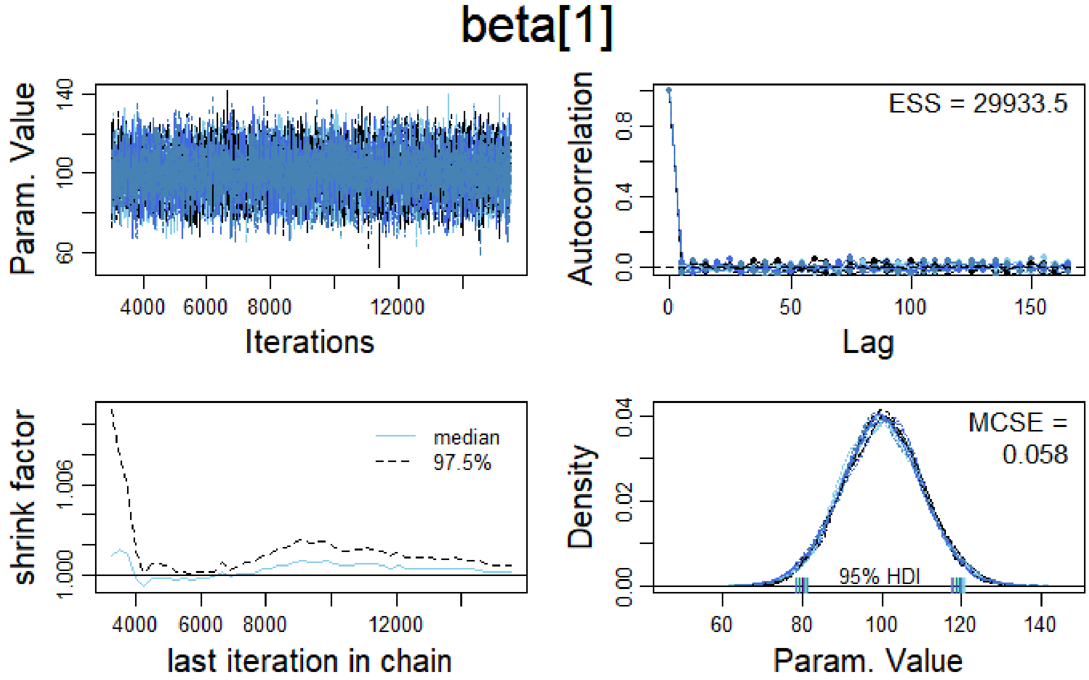
Model Parameters

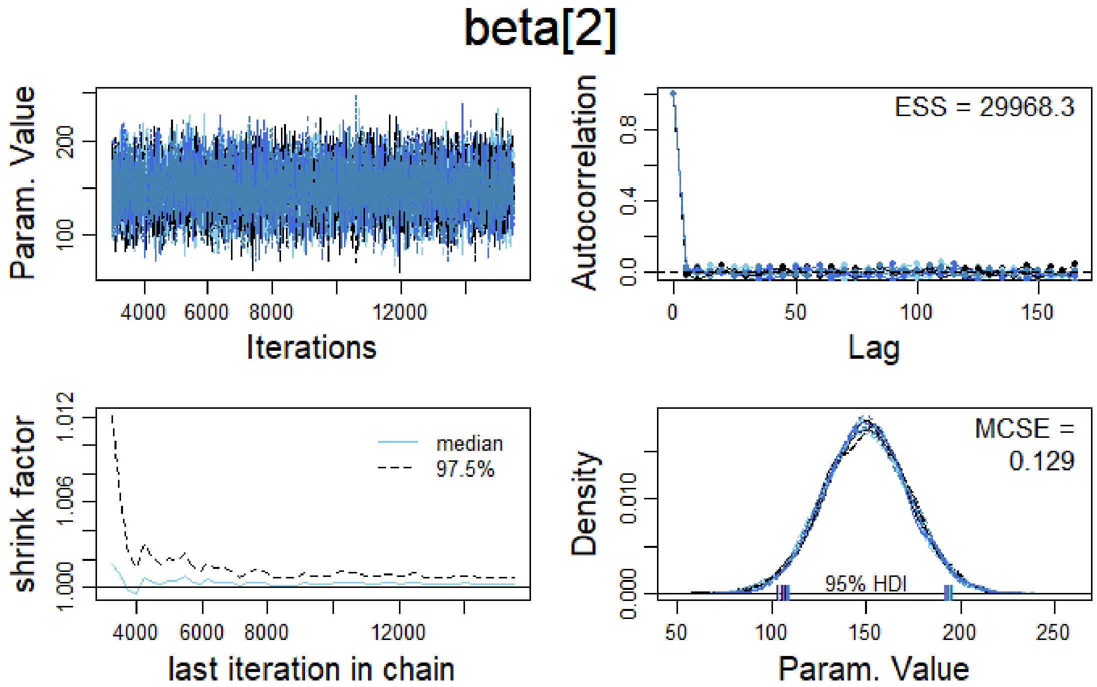
For the JAGs Model optimization, the following parameter estimates for used on the dataset:

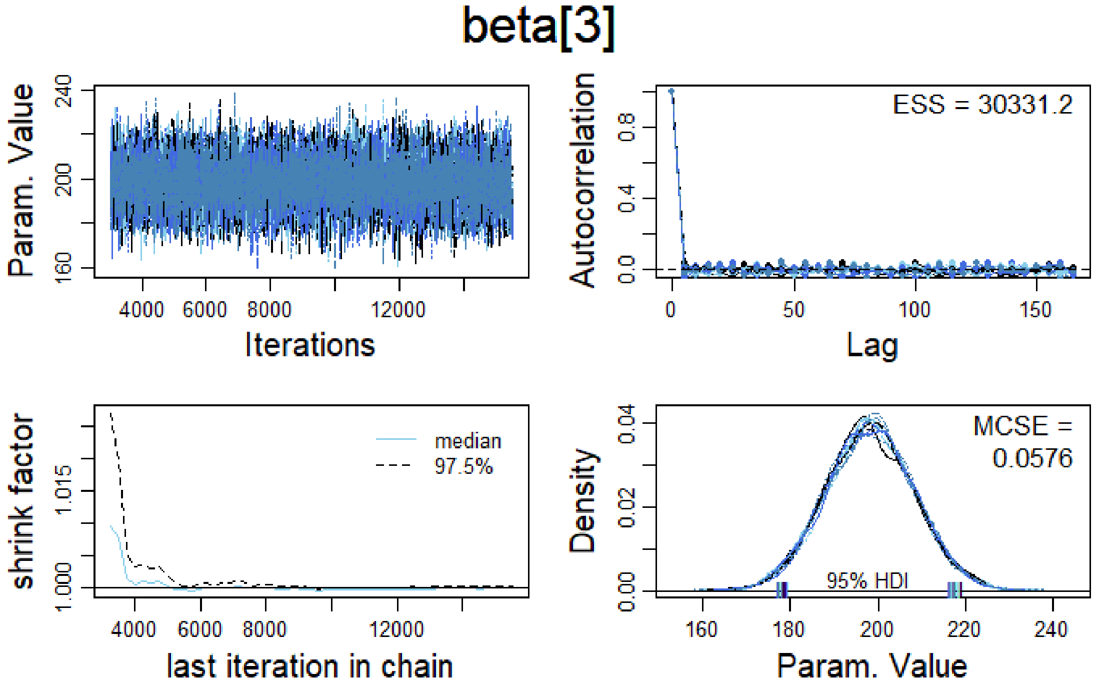
```
+ nChains = 12
+ numSavedSteps = 30000
+ thinSteps = 5
+ adaptSteps = 1000 # Number of steps to "tune" the samplers
+ burnInSteps = 2000
```

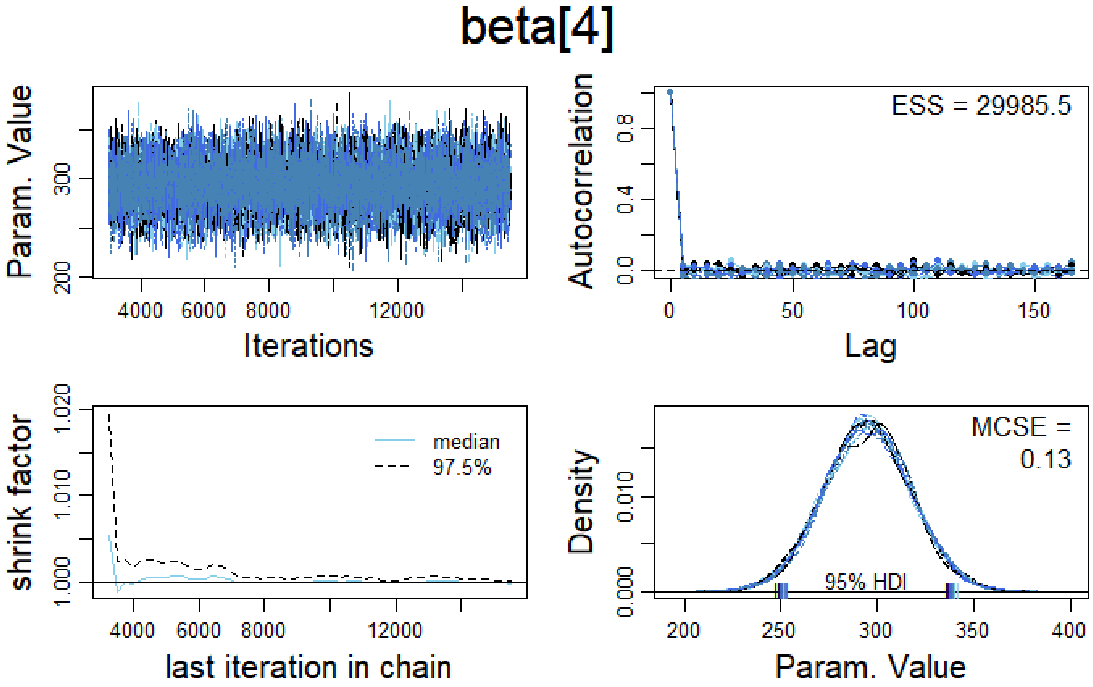
MCMC Diagnostics

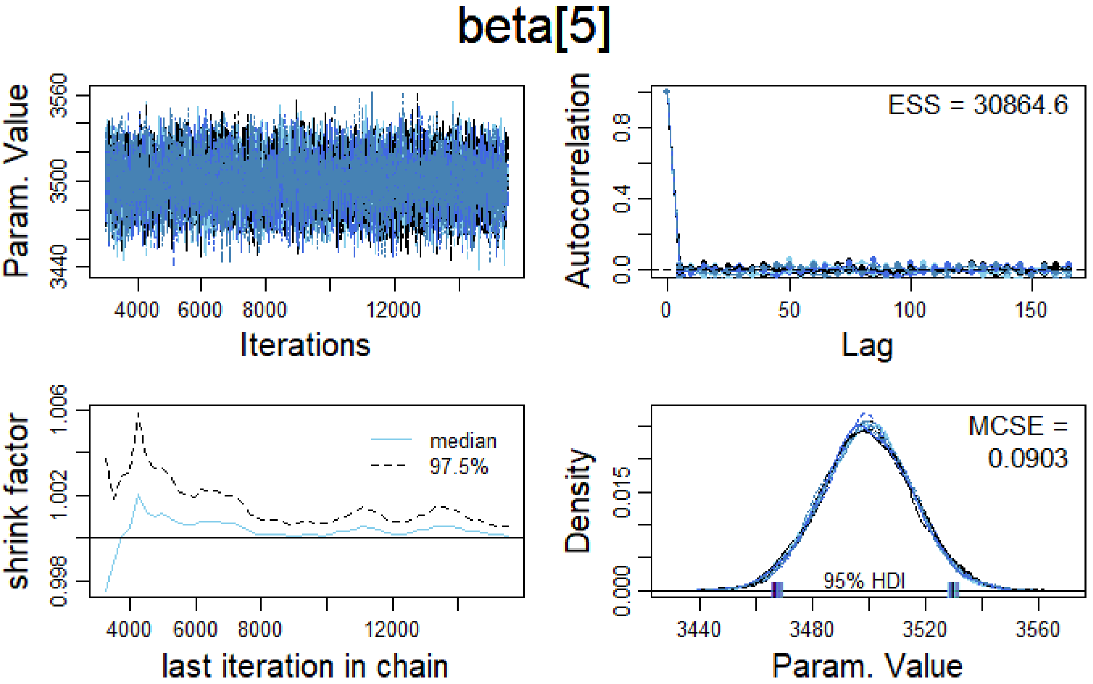


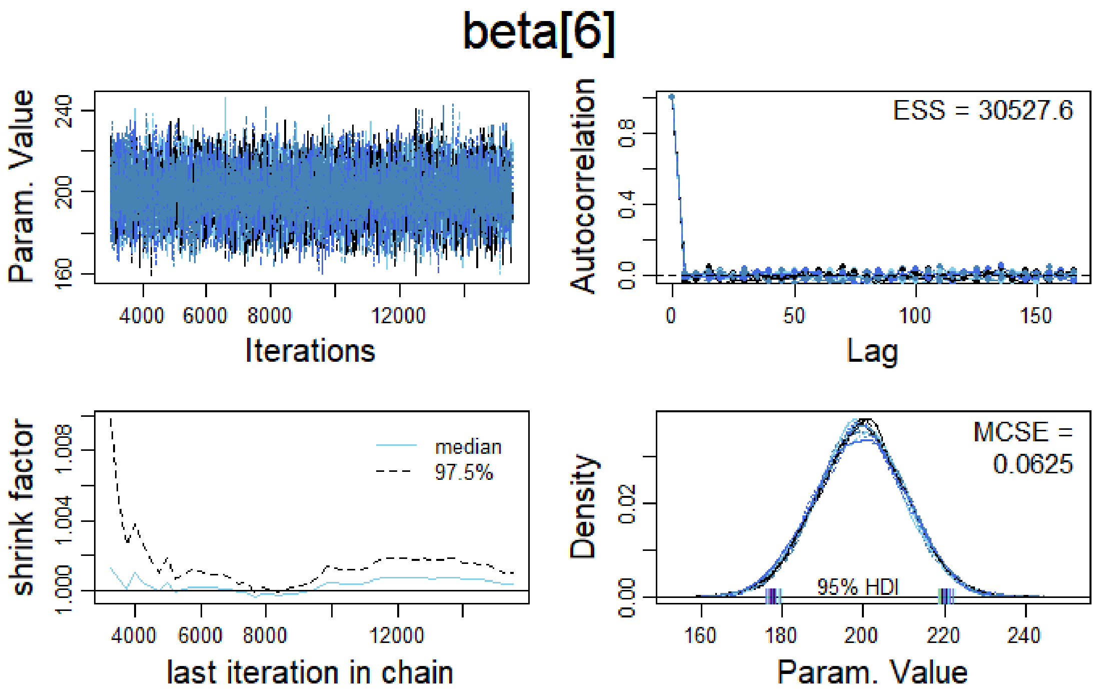


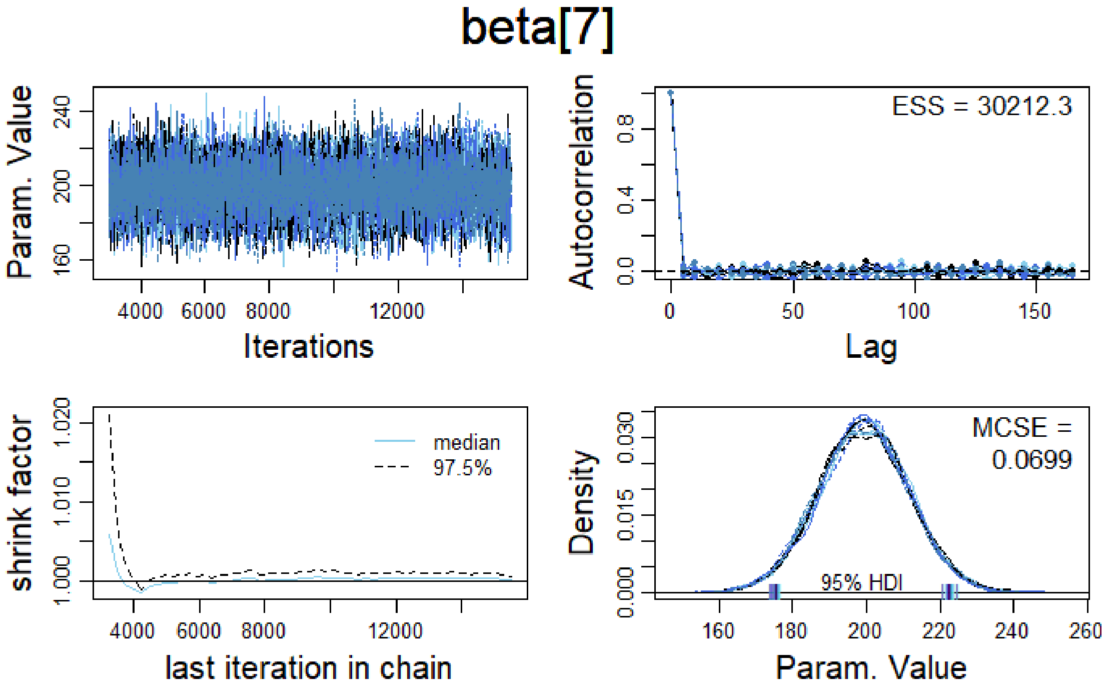

beta[2]
Param. Value
100
200
4000
6000
8000
12000
Iterations
Autocorrelation
0.0
0.4
0.8
ESS = 29968.3
0
50
100
150
Lag
shrink factor
1.000
1.006
1.012
median
97.5%
4000
6000
8000
12000
last iteration in chain
Density
0.000
0.010
MCSE =
0.129
95% HDI
50
100
150
200
250
Param. Value

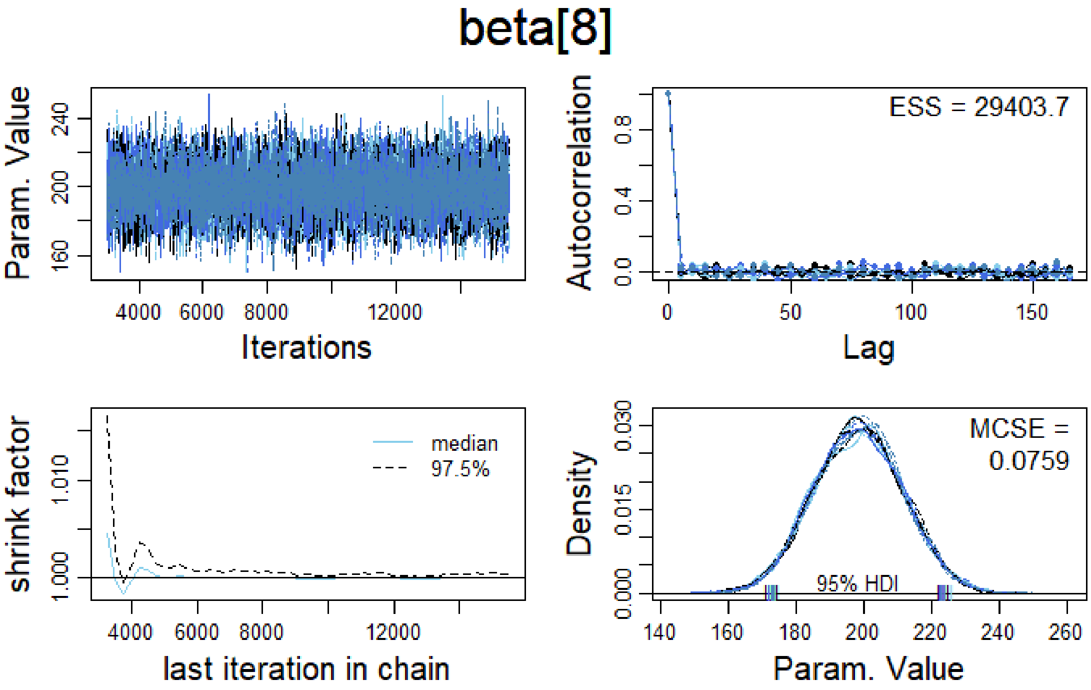

beta[3]
Param. Value
160
200
240
4000
6000
8000
12000
Iterations
Autocorrelation
0.0
0.4
0.8
ESS = 30331.2
0
50
100
150
Lag
shrink factor
1.000
1.015
median
97.5%
4000
6000
8000
12000
last iteration in chain
Density
0.00
0.02
0.04
MCSE =
0.0576
95% HDI
160
180
200
220
240
Param. Value


beta[4]
Param. Value
200
300
4000 6000 8000 12000
Iterations
Autocorrelation
0.0
0.4
0.8
ESS = 29985.5
0 50 100 150
Lag
shrink factor
1.000
1.010
1.020
median
97.5%
4000 6000 8000 12000
last iteration in chain
Density
0.000
0.010
MCSE =
0.13
95% HDI
200 250 300 350 400
Param. Value


beta[5]
Param. Value
3440
3500
3560
4000 6000 8000 12000
Iterations
Autocorrelation
0.0
0.4
0.8
ESS = 30864.6
0 50 100 150
Lag
shrink factor
0.998
1.002
1.006
median
97.5%
4000 6000 8000 12000
last iteration in chain
Density
0.000
0.015
MCSE =
0.0903
95% HDI
3440 3480 3520 3560
Param. Value


beta[6]
Param. Value
240
200
160
4000
6000
8000
12000
Iterations
Autocorrelation
ESS = 30527.6
0.0
0.4
0.8
0
50
100
150
Lag
shrink factor
1.000
1.004
1.008
median
97.5%
4000
6000
8000
12000
last iteration in chain
Density
MCSE =
0.0625
0.00
0.02
95% HDI
160
180
200
220
240
Param. Value

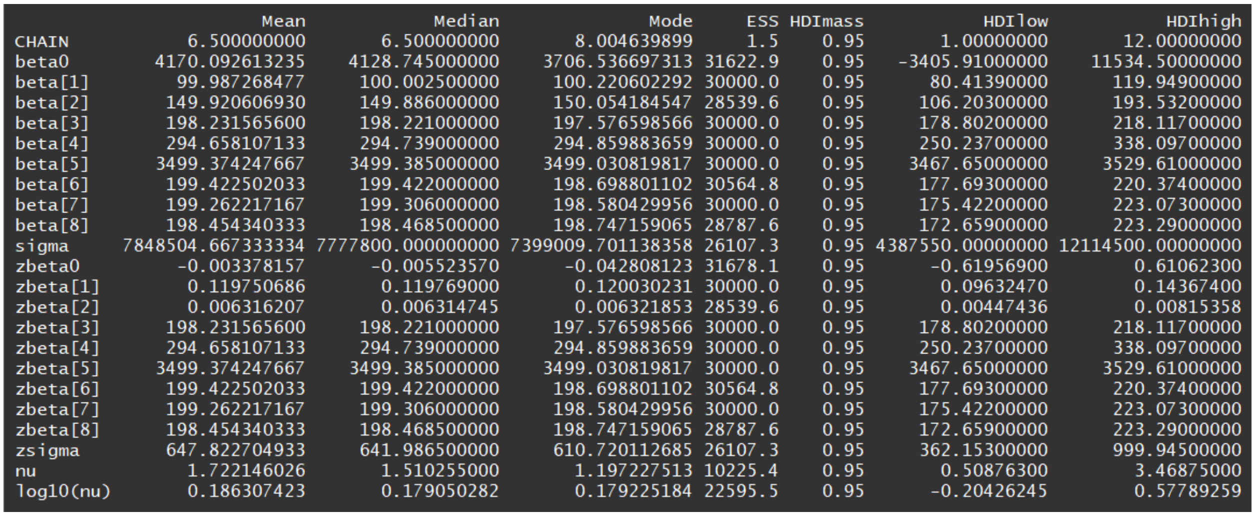

beta[7]
Param. Value
240
200
160
4000
6000
8000
12000
Iterations
Autocorrelation
ESS = 30212.3
0.0
0.4
0.8
0
50
100
150
Lag
shrink factor
1.000
1.010
1.020
median
97.5%
4000
6000
8000
12000
last iteration in chain
Density
MCSE =
0.0699
0.000
0.015
0.030
95% HDI
160
180
200
220
240
260
Param. Value



All the diagnostic plots for the analysis look satisfactory. The trace plot shows a good convergence of chains after the burn-in period and appear to be well mixing. Also, there are no orphan chains left after the convergence. The shrink factor for all the betas and zbetas looks good as the overall value of shrink factor is less than 1.1 and the lines are almost similar with a step decline as iterations increase. It is also clear from the iteration plot that there is no autocorrelation between the chains. Moreover, the trajectory of the chains in the density plot is consistant for all the betas and the distribution looks normal.

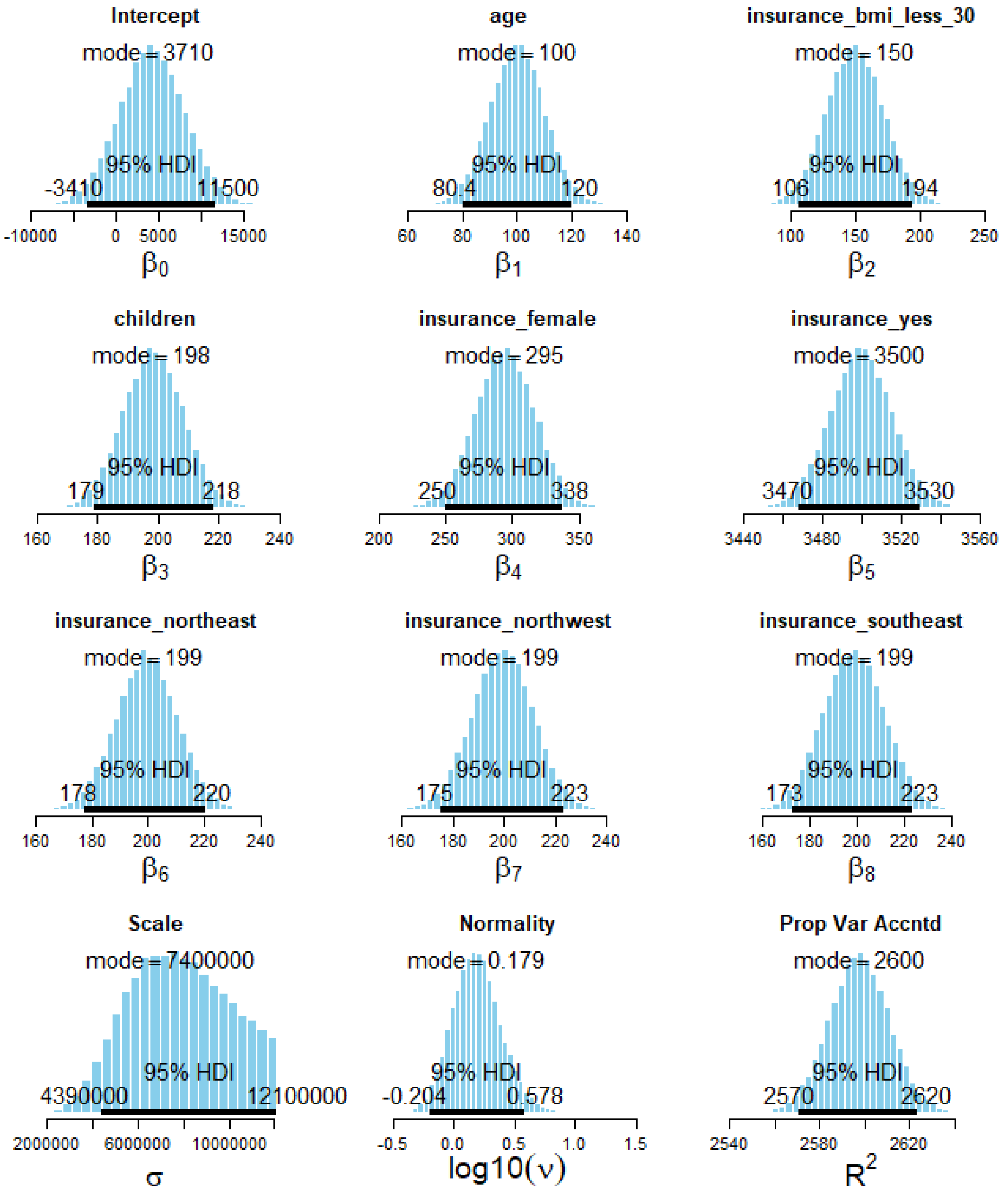
Summary Statistics

The summary statistics for the Bayesian Estimates calculated can be seen below :



```
                       Mean            Median              Mode     ESS HDImass           HDIlow           HDIhigh
CHAIN           6.500000000       6.500000000       8.004639899     1.5    0.95       1.00000000       12.00000000
beta0        4170.092613235    4128.745000000    3706.536697313 31622.9    0.95  -3405.91000000    11534.50000000
beta[1]        99.987268477     100.002500000     100.220602292 30000.0    0.95      80.41390000      119.94900000
beta[2]       149.920606930     149.886000000     150.054184547 28539.6    0.95     106.20300000      193.53200000
beta[3]       198.231565600     198.221000000     197.576598566 30000.0    0.95     178.80200000      218.11700000
beta[4]       294.658107133     294.739000000     294.859883659 30000.0    0.95     250.23700000      338.09700000
beta[5]      3499.374247667    3499.385000000    3499.030819817 30000.0    0.95    3467.65000000     3529.61000000
beta[6]       199.422502033     199.422000000     198.698801102 30564.8    0.95     177.69300000      220.37400000
beta[7]       199.262217167     199.306000000     198.580429956 30000.0    0.95     175.42200000      223.07300000
beta[8]       198.454340333     198.468500000     198.747159065 28787.6    0.95     172.65900000      223.29000000
sigma     7848504.667333334 7777800.000000000 7399009.701138358 26107.3    0.95 4387550.00000000 12114500.00000000
zbeta0         -0.003378157      -0.005523570      -0.042808123 31678.1    0.95      -0.61956900        0.61062300
zbeta[1]        0.119750686       0.119769000       0.120030231 30000.0    0.95       0.09632470        0.14367400
zbeta[2]        0.006316207       0.006314745       0.006321853 28539.6    0.95       0.00447436        0.00815358
zbeta[3]      198.231565600     198.221000000     197.576598566 30000.0    0.95     178.80200000      218.11700000
zbeta[4]      294.658107133     294.739000000     294.859883659 30000.0    0.95     250.23700000      338.09700000
zbeta[5]     3499.374247667    3499.385000000    3499.030819817 30000.0    0.95    3467.65000000     3529.61000000
zbeta[6]      199.422502033     199.422000000     198.698801102 30564.8    0.95     177.69300000      220.37400000
zbeta[7]      199.262217167     199.306000000     198.580429956 30000.0    0.95     175.42200000      223.07300000
zbeta[8]      198.454340333     198.468500000     198.747159065 28787.6    0.95     172.65900000      223.29000000
zsigma        647.822704933     641.986500000     610.720112685 26107.3    0.95     362.15300000      999.94500000
nu              1.722146026       1.510255000       1.197227513 10225.4    0.95       0.50876300        3.46875000
log10(nu)       0.186307423       0.179050282       0.179225184 22595.5    0.95      -0.20426245        0.57789259
```

Posterior Distributions

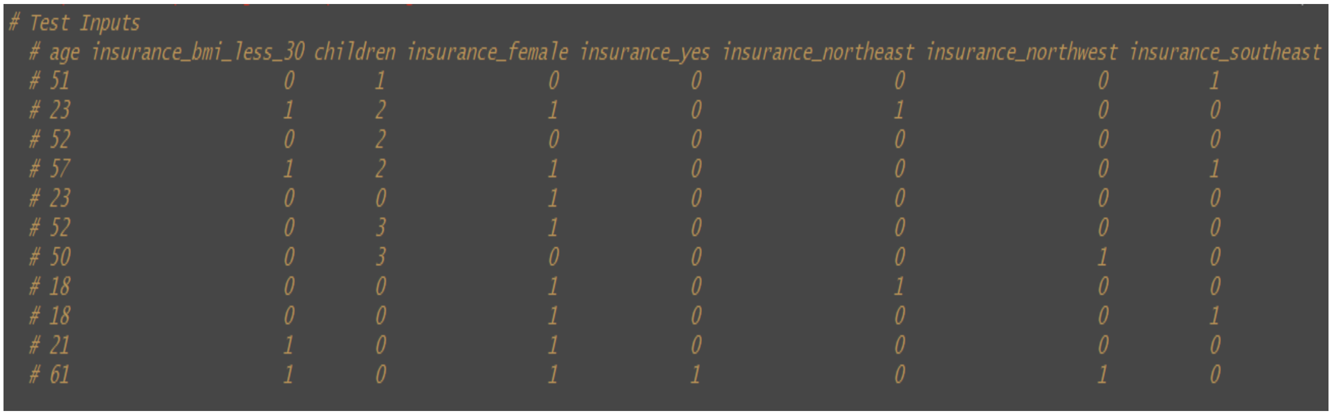


The histogram posterior distribution for the bayesian estimates between HDI intervals of 95% confidence levels can be seen above. All the estimates seem to be positive and since the diagnostics for these estimates were good, we can move onto the next stage i.e prediction using these co-efficient estimates and predictor variables.

Results and Discussion

The data was split into training and testing sets. The last 10 observations were used for checking the accuracy of the Bayesian Regression Model.

Further, the chunk below shows the Bayesian Linear Regression Equation used for building the model:



```
# Test Inputs
  # age insurance_bmi_less_30 children insurance_female insurance_yes insurance_northeast insurance_northwest insurance_southeast
  # 51                     0        1                0             0                   0                   0                   1
  # 23                     1        2                1             0                   1                   0                   0
  # 52                     0        2                0             0                   0                   0                   0
  # 57                     1        2                1             0                   0                   0                   1
  # 23                     0        0                1             0                   0                   0                   0
  # 52                     0        3                1             0                   0                   0                   0
  # 50                     0        3                0             0                   0                   1                   0
  # 18                     0        0                1             0                   1                   0                   0
  # 18                     0        0                1             0                   0                   0                   1
  # 21                     1        0                1             0                   0                   0                   0
  # 61                     1        0                1             1                   0                   1                   0
```

The final Predictions were as follows:

```
#   charges       Pred
#   9377.905  4560.880
# 22395.744 -1477.888
# 10325.206  4781.390
# 12629.166  6017.142
# 10795.937 -1611.512
# 11411.685  4781.481
# 10600.548  4340.547
#   2205.981 -2713.695
#   1629.833 -2713.758
#   2007.945 -1918.891
# 29141.360 10388.107
```

The Final predictor estimates are represented by the Pred column in the data table above. These prices are specified in USD. These values seem to be an approximate representation when compared with the original values.

The MCMC diagnostics for the beta parameters are good, indicating that the chains are converging well.

Lastly, from the current model we see that being a smoker has the largest impact on the medical costs of an individual closely followed by being overweight and older. So we plan to test build a model which penalises the client more for having these attributes by increasing the weights (beta co-efficients) associated with these factors

Testing Additional Models

The data set used for modelling consisted of both smokers and non-smokers. Since smoking habbit in a client largely affects the insurance charges, we can further divide our data set on the basis of the smoking factor creating a model for clients who are non-smokers and a seperate model for the smoker clients. We can then use these models according to the smoking factor of the client to get better predictions.

Conclusion

From this project, we can see that the Bayesian Regression model can be built by incorporating expert information gained through experience/intuition into the sample data available. Additionally, it is able to factor correlation between attributes into its model. However, the specification of sensible prior information parameters (mu and sigma) seem to the real challenge since the posterior distribution seem to be affected by it to a great extent. Another important factor is the accuracy and efficacy of Bayesian Estimates for small sample data, which then extends well when run on large population parameters.

References

- The data set for the project is sourced from Kaggle