Bayesian Linear Regression & Model Comparison

For this Bayesian final project, we choose the medical cost personal dataset from Kaggle. This dataset contains 7 variables:

- age: age of primary beneficiary
- sex: insurance contractor gender, female, male
- bmi: Body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg / m ^ 2) using the ratio of height to weight, ideally 18.5 to 24.9
- children: Number of children covered by health insurance / Number of dependents
- smoker: Smoking
- region: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
- charges: Individual medical costs billed by health insurance

We used traditional Multiple Linear Regression, Bayesian Linear Regression and also some different Bayesian models to compare the prediction results.

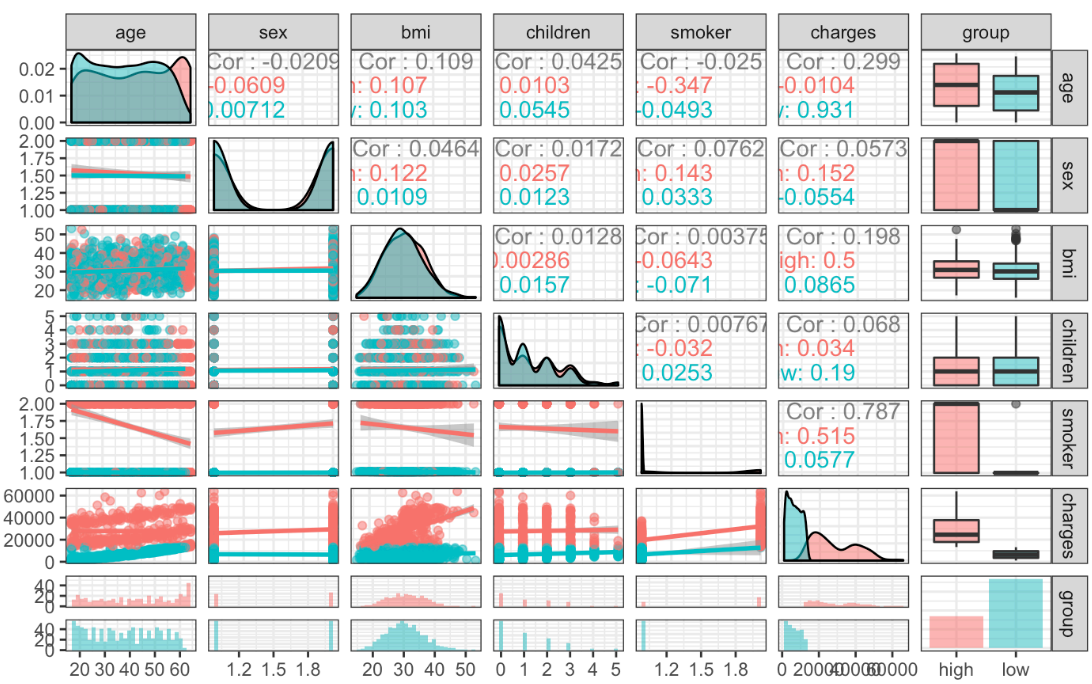
We include all necessary analysis steps in one R script file. To run the code, we primarily used the packages like ggplot2, GGally, corrplot, BAS. When run the code, please make sure place the 'insurance.csv' file and the R script file in the same directory.
The R file contains four main functions:

- dataPrepare(df): df – training data frame/ testing data frame;
  Prepare and convert the both training and testing data frame to the same format.
- EDA(traindata): traindata – training data frame;
  Exploratory data analysis for training data.
- multiLinearRegression(trainningPrepared, testingPrepared): trainningPrepared/ testingPrepared – training/ testing data frame after the regulation by dataPrepare function;
  Generate the multiple linear regression, and get the summary, plots and related analysis.
- bayesianAveraging(trainningPrepared, testingPrepared): trainningPrepared/ testingPrepared – training/ testing data frame after the regulation by dataPrepare function;
  Generate Bayesian Linear Regression and other different Bayesian models, get the summary, plots, and comparison between models.

Firstly, we observed the original dataset after loading the csv file. Our dataset has 1338 observations and 7 variables. We look at the structure and statistical summary at the beginning. Also, we checked the missing value for our dataset.

```
> Insurance <- read.csv('insurance.csv')
>
> head(Insurance)
  age    sex    bmi children smoker    region   charges
1  19 female 27.900        0    yes southwest 16884.924
2  18   male 33.770        1     no southeast  1725.552
3  28   male 33.000        3     no southeast  4449.462
4  33   male 22.705        0     no northwest 21984.471
5  32   male 28.880        0     no northwest  3866.855
6  31 female 25.740        0     no southeast  3756.622
>
> # Summary and Structure of the data -- Descriptive Statistics
> str(Insurance)
'data.frame':	1338 obs. of  7 variables:
 $ age     : int  19 18 28 33 32 31 46 37 37 60 ...
 $ sex     : Factor w/ 2 levels "female","male": 1 2 2 2 2 1 1 1 2 1 ...
 $ bmi     : num  27.9 33.8 33 22.7 28.9 ...
 $ children: int  0 1 3 0 0 0 1 3 2 0 ...
 $ smoker  : Factor w/ 2 levels "no","yes": 2 1 1 1 1 1 1 1 1 1 ...
 $ region  : Factor w/ 4 levels "northeast","northwest",..: 4 3 3 2 2 3 3 2 1 2 ...
 $ charges : num  16885 1726 4449 21984 3867 ...
> summary(Insurance)
      age            sex           bmi           children       smoker          region       charges     
 Min.   :18.00   female:662   Min.   :15.96   Min.   :0.000   no :1064   northeast:324   Min.   : 1122  
 1st Qu.:27.00   male  :676   1st Qu.:26.30   1st Qu.:0.000   yes: 274   northwest:325   1st Qu.: 4740  
 Median :39.00                Median :30.40   Median :1.000              southeast:364   Median : 9382  
 Mean   :39.21                Mean   :30.66   Mean   :1.095              southwest:325   Mean   :13270  
 3rd Qu.:51.00                3rd Qu.:34.69   3rd Qu.:2.000                              3rd Qu.:16640  
 Max.   :64.00                Max.   :53.13   Max.   :5.000                              Max.   :63770  
> sapply(Insurance, function(x) sum(is.na(x)))
     age      sex      bmi children   smoker   region  charges 
       0        0        0        0        0        0        0 
>
```

Then, we visualize our variable by groups (we assign charges above average as high, verse vice). At the same time, we convert factor features sex and smoker to numeric.



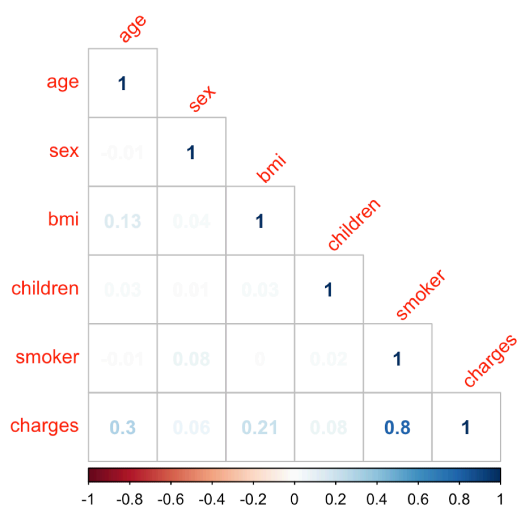
For splitting the training set and testing set, We use set.seed(101) in order to make sure the experiment is reproducible. We randomly select 75% as training set and 25% as the testing set.

```r
# Split training and testing set
set.seed(101) # Set Seed so that same sample can be reproduced in future also
# Now Selecting 75% of data as sample from total 'n' rows of the data
sample <- sample.int(n = nrow(ins), size = floor(.75*nrow(ins)), replace = F)
trainingData <- ins[sample, ]
testingData  <- ins[-sample, ]
```

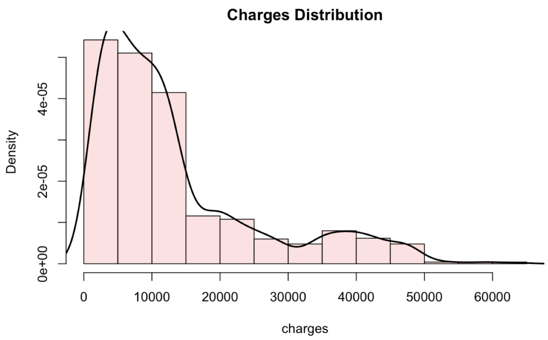
Below is the structure of training set and testing set.

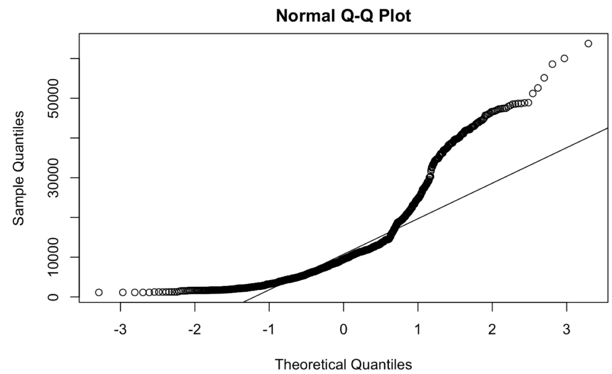
```
> # Prepare training and testing set
> trainingPrepared <- dataPrepare(trainingData)
'data.frame':	1003 obs. of  6 variables:
 $ age     : int  44 53 42 41 56 18 53 59 36 30 ...
 $ sex     : num  1 1 2 2 1 1 2 1 2 1 ...
 $ bmi     : num  24 22.9 31.3 28.8 28.8 ...
 $ children: int  2 1 0 1 0 0 0 1 1 3 ...
 $ smoker  : num  1 2 1 1 1 1 1 1 2 2 ...
 $ charges : num  8211 23245 6359 6282 11658 ...
> testingPrepared <- dataPrepare(testingData)
'data.frame':	335 obs. of  6 variables:
 $ age     : int  18 33 31 37 27 59 55 19 24 36 ...
 $ sex     : num  2 2 1 1 2 1 1 2 1 2 ...
 $ bmi     : num  33.8 22.7 25.7 27.7 42.1 ...
 $ children: int  1 0 0 3 0 3 2 0 0 0 ...
 $ smoker  : num  1 1 1 1 2 1 1 1 1 2 ...
 $ charges : num  1726 21984 3757 7282 39612 ...
```

Here, we check the correlation of numeric predictor variable with charges by drawing scatter plots.

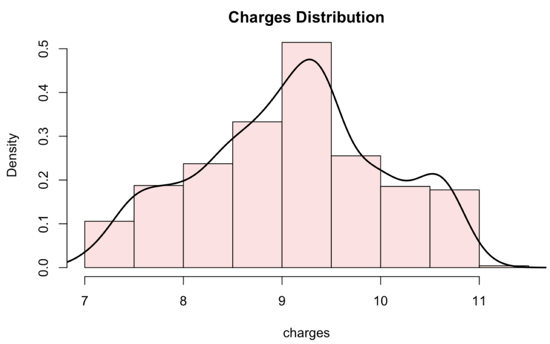


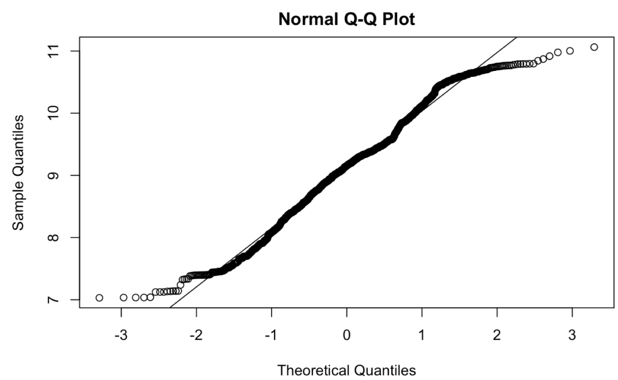
And then we checked the distribution of the charges (response variable). We find that the distribution is not ideal normal, which is skewed right. Thus we tried to apply log transformation.



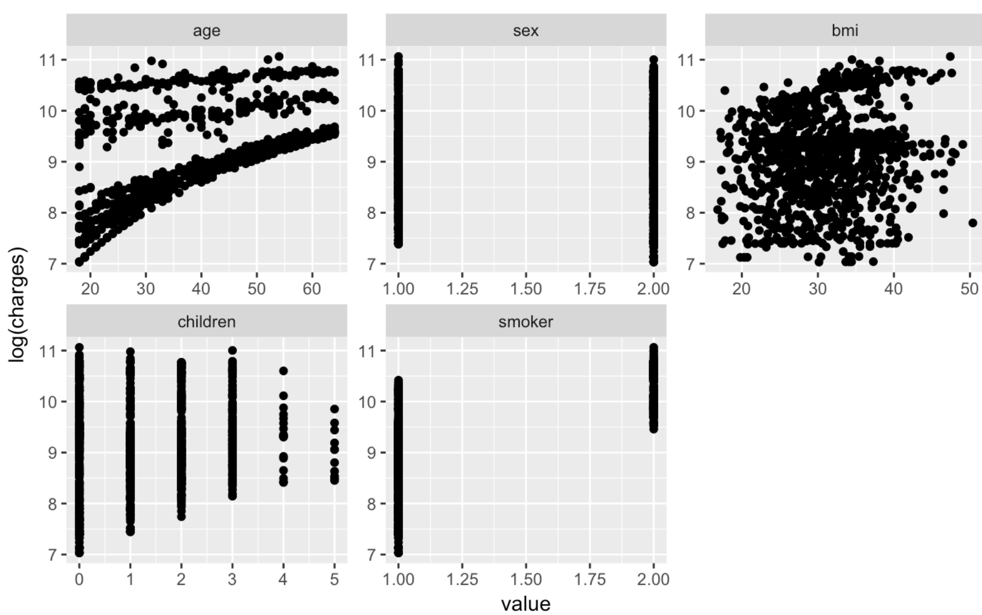


After the log transformation, the distribution is so much better.





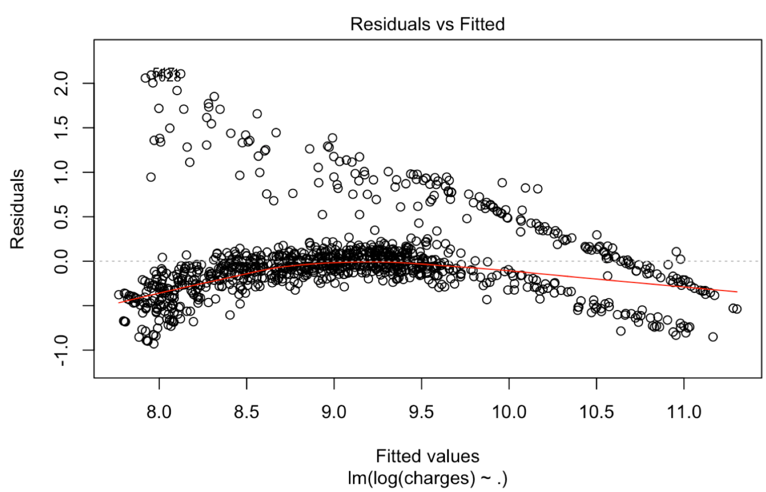
Following, we draw scatter plots for predictor variables with log transformed charges.

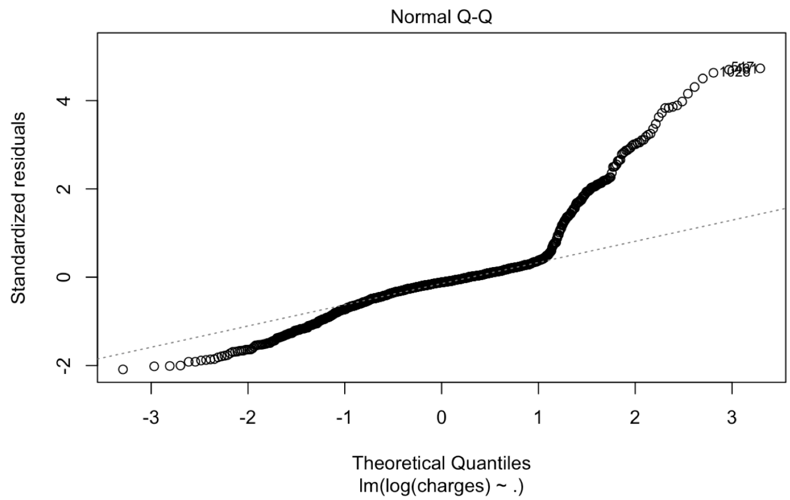


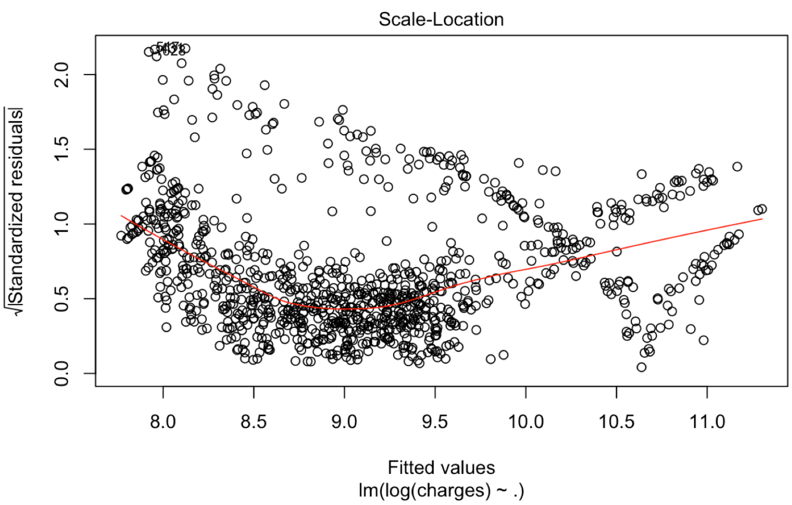
After the exploratory data analysis, we generated the multiple linear regression for our data, and below is the summary output.

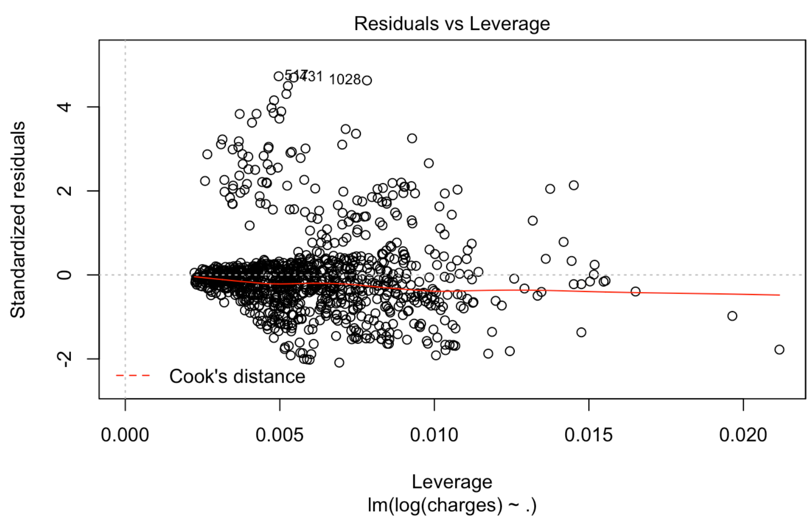
```
> multiLinearRegression(trainingPrepared,testingPrepared)

Call:
lm(formula = log(charges) ~ ., data = trainingPrepared)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.92922 -0.20950 -0.05163  0.07880  2.10684 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  5.518388   0.096638  57.104  < 2e-16 ***
age          0.034368   0.001014  33.895  < 2e-16 ***
sex         -0.079808   0.028306  -2.820   0.0049 ** 
bmi          0.011852   0.002320   5.109 3.87e-07 ***
children     0.108689   0.011849   9.173  < 2e-16 ***
smoker       1.549919   0.034829  44.501  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4465 on 997 degrees of freedom
Multiple R-squared:  0.7689,	Adjusted R-squared:  0.7678 
F-statistic: 663.6 on 5 and 997 DF,  p-value: < 2.2e-16
```

And, there is the residual related plots.




Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(log(charges) ~ .)
1028
517


Scale-Location
√|Standardized residuals|
Fitted values
lm(log(charges) ~ .)


Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage
lm(log(charges) ~ .)
1028

Finally, we calculate the RMSE for comparison.

```
Hit <Return> to see next plot:
[1] "Root Mean Squared Error 8791.53749831095"
```

Then we use the Bayesian Linear Regression and summary the results.

```
> bma_charges

Call:
bas.lm(formula = log(charges) ~ ., data = trainingPrepared, prior = "BIC",
    modelprior = uniform(), method = "MCMC")


 Marginal Posterior Inclusion Probabilities:
Intercept        age        sex        bmi   children     smoker
   1.0000     0.9922     0.6750     0.9672     0.9859     0.9969
```

```
> bayesianAveraging(trainingPrepared,testingPrepared)
          P(B != 0 | Y)    model 1      model 2       model 3       model 4       model 5
Intercept     1.0000000     1.0000    1.0000000  1.000000e+00  1.000000e+00  1.000000e+00
age           0.9859375     1.0000    1.0000000  0.000000e+00  0.000000e+00  0.000000e+00
sex           0.5875000     1.0000    0.0000000  0.000000e+00  0.000000e+00  0.000000e+00
bmi           0.9921875     1.0000    1.0000000  0.000000e+00  1.000000e+00  1.000000e+00
children      0.9953125     1.0000    1.0000000  1.000000e+00  0.000000e+00  1.000000e+00
smoker        0.9921875     1.0000    1.0000000  0.000000e+00  0.000000e+00  1.000000e+00
BF                   NA     1.0000    0.5900675 2.113381e-307 1.564587e-308 8.593733e-166
PostProbs            NA     0.5875    0.3984000  4.700000e-03  3.100000e-03  3.100000e-03
R2                   NA     0.7689    0.7671000  2.900000e-02  2.400000e-02  5.001000e-01
dim                  NA     6.0000    5.0000000  2.000000e+00  2.000000e+00  4.000000e+00
logmarg              NA -2674.7860 -2675.3135353 -3.380931e+03 -3.383535e+03 -3.054864e+03
```

The Marginal posterior probabilities for BMA, BPM, MPM, HPM.

```
 Marginal Posterior Summaries of Coefficients:

 Using  BMA

 Based on the top  8 models
           post mean  post SD   post p(B != 0)
Intercept   9.100904  0.014347  1.000000
age         0.033901  0.004172  0.985938
sex        -0.046887  0.044881  0.587500
bmi         0.011732  0.002696  0.992188
children    0.108223  0.014189  0.995313
smoker      1.534679  0.140619  0.992188
```

```
 Marginal Posterior Summaries of Coefficients:

 Using  BPM

 Based on the top  8 models
           post mean  post SD   post p(B != 0)
Intercept   9.100904   0.014347   1.000000
age         0.033901   0.004172   0.985938
sex        -0.046887   0.044881   0.587500
bmi         0.011732   0.002696   0.992188
children    0.108223   0.014189   0.995313
smoker      1.534679   0.140619   0.992188

 Marginal Posterior Summaries of Coefficients:

 Using  MPM

 Based on the top  1 models
           post mean  post SD   post p(B != 0)
Intercept   9.100904   0.014099   1.000000
age         0.034368   0.001014   1.000000
sex        -0.079808   0.028306   1.000000
bmi         0.011852   0.002320   1.000000
children    0.108689   0.011849   1.000000
smoker      1.549919   0.034829   1.000000

 Using  HPM

 Based on the top  1 models
           post mean  post SD   post p(B != 0)
Intercept   9.100904   0.014099   1.000000
age         0.034368   0.001014   0.985938
sex        -0.079808   0.028306   0.587500
bmi         0.011852   0.002320   0.992188
children    0.108689   0.011849   0.995313
smoker      1.549919   0.034829   0.992188
```

We look at the 95% confidence intervals for the coefficients.

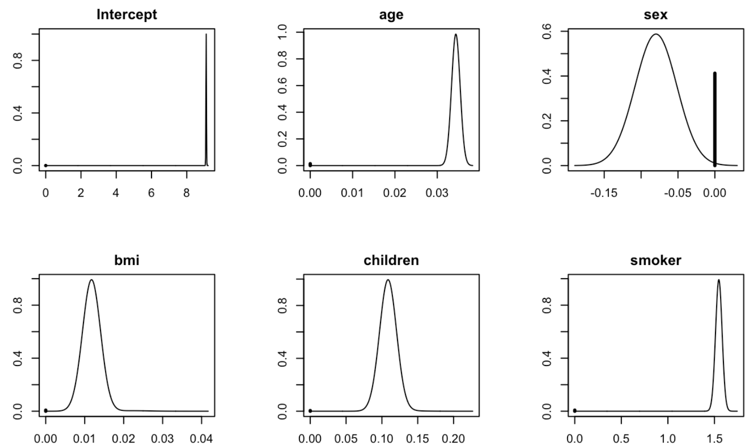
```
> confint(coef(bma_charges,estimator = estimatorName),level = 0.95)
                  2.5%       97.5%        beta
Intercept  9.073236317  9.12857144  9.10090388
age        0.032378040  0.03635749  0.03436777
sex       -0.135354008 -0.02426276 -0.07980839
bmi        0.007300012  0.01640372  0.01185187
children   0.085437845  0.13194114  0.10868949
smoker     1.481572860  1.61826502  1.54991894
attr(,"Probability")
[1] 0.95
attr(,"class")
[1] "confint.bas"
```

Finally, compare the RMSE again.

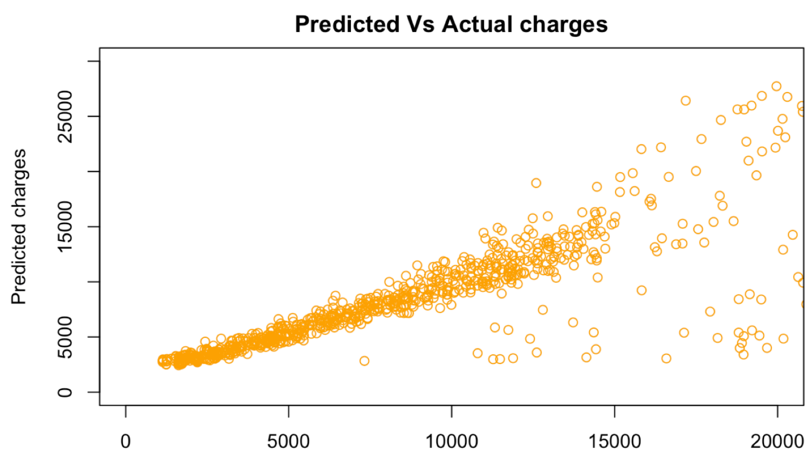
```
[1] "Root Mean Square Error BMA 8563.28017809004"
[1] "Root Mean Square Error BPM 8791.53749831119"
[1] "Root Mean Square Error MPM 8791.53749831118"
[1] "Root Mean Square Error HPM 8791.53749831119"
```

By comparison of the RMSE, we can find that Bayesian Linear Model has the lowest RMSE, which means the highest accuracy.

To further consider the Bayesian method, we plot the posteriors distribution. From those plots we learned that the feature ‘sex’ is the least significant due to the highest bar on zero; feature ‘age’ and ‘smoker’ are the most significant due to the shortest bar on zero and narrow width of the curve.



This is the prediction scatter plot.



About all above analysis, the Bayesian Linear Regression is better than the other four, and feature ‘smoker’ and ‘age’ are most significant in ‘charges’ prediction. Since there might be some correlation between features, and the high insurance charges have the different character distribution. Thus our model still has some limitation.