Coronary heart disease prediction using logistic regression and fully Bayesian

Abolfazl Joukar

Introduction

World Health Organization has estimated 12 million deaths occur worldwide, every year due to Heart diseases. Half the deaths in the United States and other developed countries are due to cardiovascular diseases. The early prognosis of cardiovascular diseases can aid in making decisions on lifestyle changes in high-risk patients and in turn reduce the complications. This research intends to pinpoint the most relevant/risk factors of heart disease as well as predict the overall risk using logistic regression.

Source

The dataset is publicly available on the Kaggle website, and it is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. The classification goal is to predict whether the patient has 10-year risk of future coronary heart disease (CHD). The dataset provides the patients' information. It includes over 4238 records and 16 attributes.

Data structure and details

Each attribute is a potential risk factor. There are demographic, behavioral and medical risk factors.

variables	Type			information
Male	Demographic	Integer	Nominal	Male or female
Age	Demographic	Integer	Continuous	Age of patient
Education	Demographic	Integer	Nominal	School or university degree
CurrentSmoker	Behavioral	Integer	Nominal	Smoker or not
CigsPerDay	Behavioral	Integer	Continuous	Number of cigarettes per day
BPMeds	Medical	Integer	Nominal	The patient was on blood pressure medication or not
PrevalentStroke	Medical	Integer	Nominal	The patient had previously a stroke
PrevalentHyp	Medical	Integer	Nominal	The patient was hypertensive or not
Diabetes	Medical	Integer	Nominal	The patient had diabetes or not
TotChol	Medical	Integer	Continuous	Total cholesterol
SysBP	Medical	Numeric	Continuous	Systolic blood pressure
DiaBP	Medical	Numeric	Continuous	Diastolic blood pressure
BMI	Medical	Numeric	Continuous	Body mass index
HeartRate	Medical	Integer	Continuous	Heart rate
Glucose	Medical	Integer	Continuous	Glucose level
TenYearCHD	Target	Integer	nominal	10 years risk of coronary heart disease CHD

```
> summary(data)
      male             age          education     currentSmoker      cigsPerDay         BPMeds       
 Min.   :0.0000   Min.   :32.00   Min.   :1.000   Min.   :0.0000   Min.   : 0.000   Min.   :0.00000  
 1st Qu.:0.0000   1st Qu.:42.00   1st Qu.:1.000   1st Qu.:0.0000   1st Qu.: 0.000   1st Qu.:0.00000  
 Median :0.0000   Median :49.00   Median :2.000   Median :0.0000   Median : 0.000   Median :0.00000  
 Mean   :0.4292   Mean   :49.58   Mean   :1.979   Mean   :0.4941   Mean   : 9.003   Mean   :0.02963  
 3rd Qu.:1.0000   3rd Qu.:56.00   3rd Qu.:3.000   3rd Qu.:1.0000   3rd Qu.:20.000   3rd Qu.:0.00000  
 Max.   :1.0000   Max.   :70.00   Max.   :4.000   Max.   :1.0000   Max.   :70.000   Max.   :1.00000  
                                  NA's   :105                      NA's   :29       NA's   :53       
 prevalentStroke     prevalentHyp       diabetes          totChol          sysBP           diaBP       
 Min.   :0.000000   Min.   :0.0000   Min.   :0.00000   Min.   :107.0   Min.   : 83.5   Min.   : 48.00  
 1st Qu.:0.000000   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:206.0   1st Qu.:117.0   1st Qu.: 75.00  
 Median :0.000000   Median :0.0000   Median :0.00000   Median :234.0   Median :128.0   Median : 82.00  
 Mean   :0.005899   Mean   :0.3105   Mean   :0.02572   Mean   :236.7   Mean   :132.4   Mean   : 82.89  
 3rd Qu.:0.000000   3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:263.0   3rd Qu.:144.0   3rd Qu.: 89.88  
 Max.   :1.000000   Max.   :1.0000   Max.   :1.00000   Max.   :696.0   Max.   :295.0   Max.   :142.50  
                                                       NA's   :50                                      
      BMI          heartRate         glucose         TenYearCHD   
 Min.   :15.54   Min.   : 44.00   Min.   : 40.00   Min.   :0.000  
 1st Qu.:23.07   1st Qu.: 68.00   1st Qu.: 71.00   1st Qu.:0.000  
 Median :25.40   Median : 75.00   Median : 78.00   Median :0.000  
 Mean   :25.80   Mean   : 75.88   Mean   : 81.97   Mean   :0.152  
 3rd Qu.:28.04   3rd Qu.: 83.00   3rd Qu.: 87.00   3rd Qu.:0.000  
 Max.   :56.80   Max.   :143.00   Max.   :394.00   Max.   :1.000  
 NA's   :19      NA's   :1        NA's   :388                     
```

Libraries:

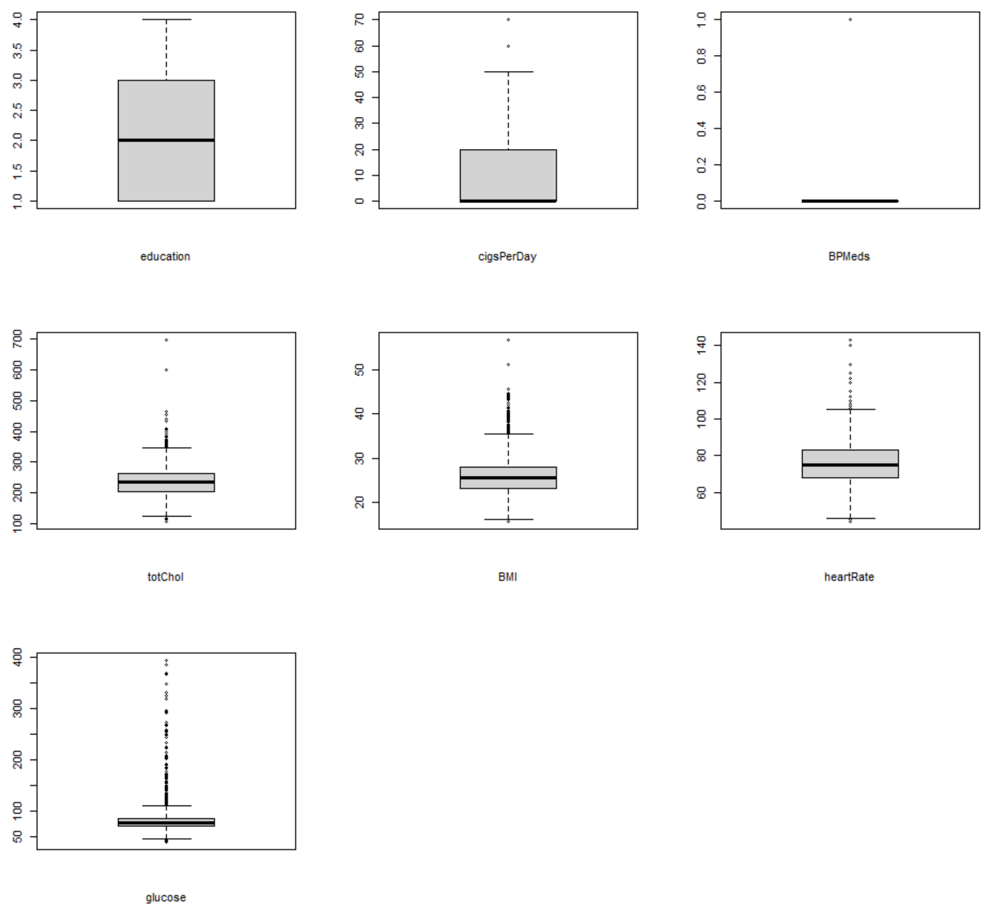
You can see the libraries which are used in this project.

```
> library(coda)
> library(R2jags)
> library(LaplacesDemon)
> library(TeachingDemos)
> library(corrplot)
> library(caTools)
> library(pROC)
```

Check NAs:

```
> colSums(is.na(data))
           male             age       education   currentSmoker      cigsPerDay          BPMeds 
              0               0             105               0              29              53 
prevalentStroke    prevalentHyp        diabetes         totChol           sysBP           diaBP 
              0               0               0              50               0               0 
            BMI       heartRate         glucose      TenYearCHD 
             19               1             388               0 
> 
```

As you can see, some variables have nan values. For fixing them, we should take a look at the boxplot of these variables.



there is no outlier data on education feature and because this variable is categorical, I replaced the nan values with the most used value.

```
> table(data$education)

   1    2    3    4 
1720 1253  687  473 
```

There are outlier data in cigsPerDay, totalChol, BMI, heartrate and glucose, So I replaced the nan values with their median.

BPModes is categorical variable which the most values of it, is 0. So, we can remove it.

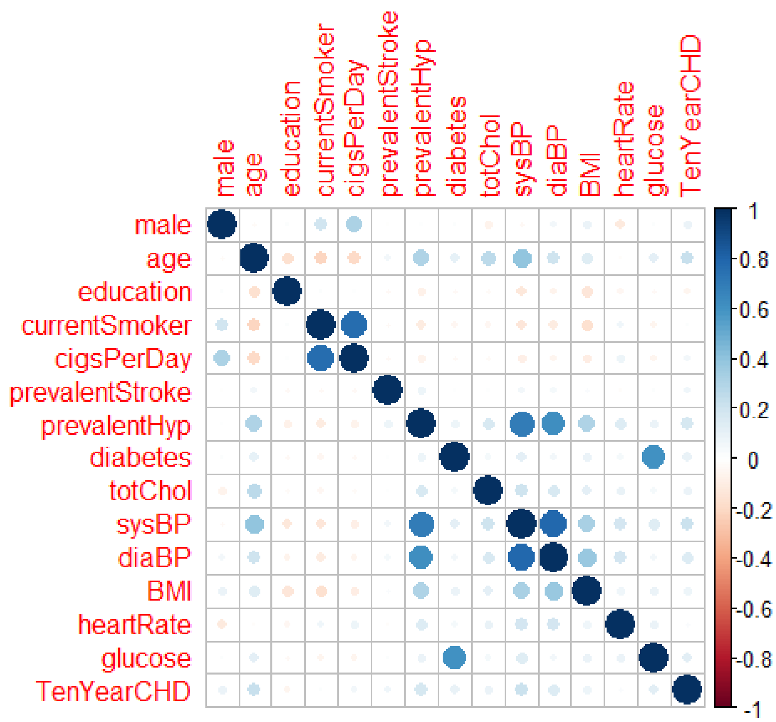
```
> table(data$BPMeds)

   0    1 
4061  124 
```

Again, we can have a look to the data after fixing the nan values.

```
> colSums(is.na(data))
           male             age       education   currentSmoker      cigsPerDay prevalentStroke    prevalentHyp 
              0               0               0               0               0               0               0 
       diabetes         totChol           sysBP           diaBP             BMI       heartRate         glucose 
              0               0               0               0               0               0               0 
     TenYearCHD 
              0 
```

Correlation

```
> cor(data)[,'TenYearCHD']
        male          age    education currentSmoker    cigsPerDay prevalentStroke  prevalentHyp
  0.08842757   0.22525610  -0.05281226    0.01945627    0.05885914      0.06180995    0.17760273
    diabetes      totChol        sysBP         diaBP           BMI       heartRate       glucose
  0.09731651   0.08156572   0.21642904    0.14529910    0.07421662      0.02285676    0.12127740
  TenYearCHD
  1.00000000
```



The age, diaBP, sysBP, prelaventHyp and glucose are the most correlated variables.

```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)    
(Intercept)      -8.161976   0.663886 -12.294  < 2e-16 ***
male              0.497756   0.100250   4.965 6.86e-07 ***
age               0.062179   0.006220   9.997  < 2e-16 ***
education        -0.010172   0.045777  -0.222 0.824157    
currentSmoker     0.013959   0.142789   0.098 0.922123    
cigsPerDay        0.021373   0.005640   3.790 0.000151 ***
prevalentStroke   1.013011   0.439301   2.306 0.021113 *  
prevalentHyp      0.243064   0.127835   1.901 0.057251 .  
diabetes          0.194796   0.293777   0.663 0.507283    
totChol           0.001852   0.001025   1.807 0.070809 .  
sysBP             0.014510   0.003525   4.116 3.86e-05 ***
diaBP            -0.002940   0.005976  -0.492 0.622757    
BMI               0.003453   0.011799   0.293 0.769812    
heartRate        -0.001653   0.003883  -0.426 0.670381    
glucose           0.006680   0.002136   3.127 0.001766 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value of male, age,cigsPerDay, prevalentStroke, sysBP and glucose are acceptable.

In the below table, you can find the p-values and correlation.

variables	p-value	correlation
male	0.00000069	0.088428
age	0.00000001	0.225256
education	0.82415721	-0.05281
currentSmoker	0.92212274	0.019456
cigsPerDay	0.00015082	0.058859
prevalentStroke	0.02111287	0.06181
prevalentHyp	0.05725103	0.177603
diabetes	0.5072829	0.097317
totChol	0.07080926	0.081566
sysBP	0.00003859	0.216429
diaBP	0.62275671	0.145299
BMI	0.76981241	0.074217
heartRate	0.67038115	0.022857
glucose	0.00176637	0.121277

I run the logistic model in three different independent variables and different split ratio and check the model's AIC and the accuracy of it.

Models:

Bernoulli

In this table you can see the result of model with different independent variables and split ratio.

	Variables	AIC	ACC	Split ratio
Model1	All	2265.7	85.12982	0.7
Model2	Correlation > 0.1	2307	84.7364	0.7
Model3	P values < 0.05	2252.6	84.97246	0.7
Model4	All	2594.7	85.37736	0.8
Model5	Correlation > 0.1	2648.4	85.25943	0.8
Model6	P values < 0.05	2584.2	85.25943	0.8
Model7	All	2920.1	85.5792	0.9
Model8	Correlation > 0.1	2982.4	85.34279	0.9
Model9	P values < 0.05	2912.9	85.5792	0.9

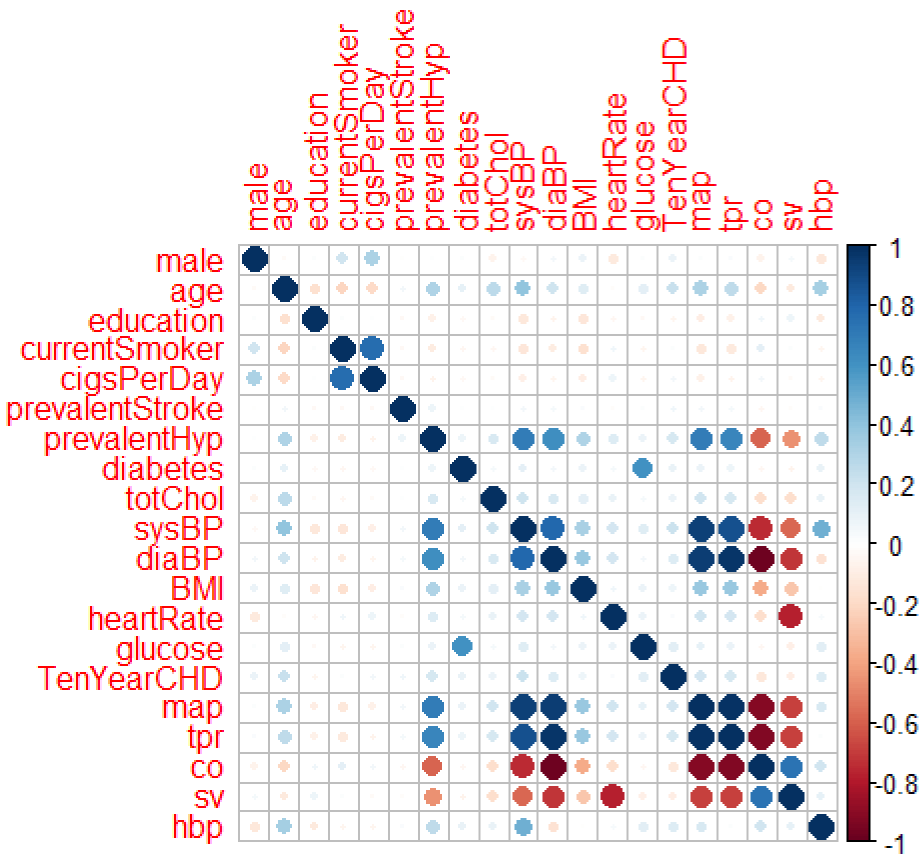
Feature engineering:

I tried to create some new variables using some equations from medical articles.

```
> #Feature engineering
> data$map = as.integer(((2*data$diaBP) + data$sysBP)/3) #mean blood pressure
> data$tpr = (data$map * data$diaBP) / 5 #Total peripheral resistance
> data$co = data$map/ data$tpr #cardiac output
> data$sv = data$co / data$heartRate
> data$hbp = data$sysBP/data$diaBP #blood pressure
```

Below you can see the p-value and correlation of new variables.

variables	p-value	correlation
map	0.08992722	0.189582
tpr	0.63100465	0.174653
co	0.23827812	-0.12575
sv	0.2946932	-0.09311
hbp	0.46402328	0.144479



As you can see, some of the new variables has the correlation more than 0.15.

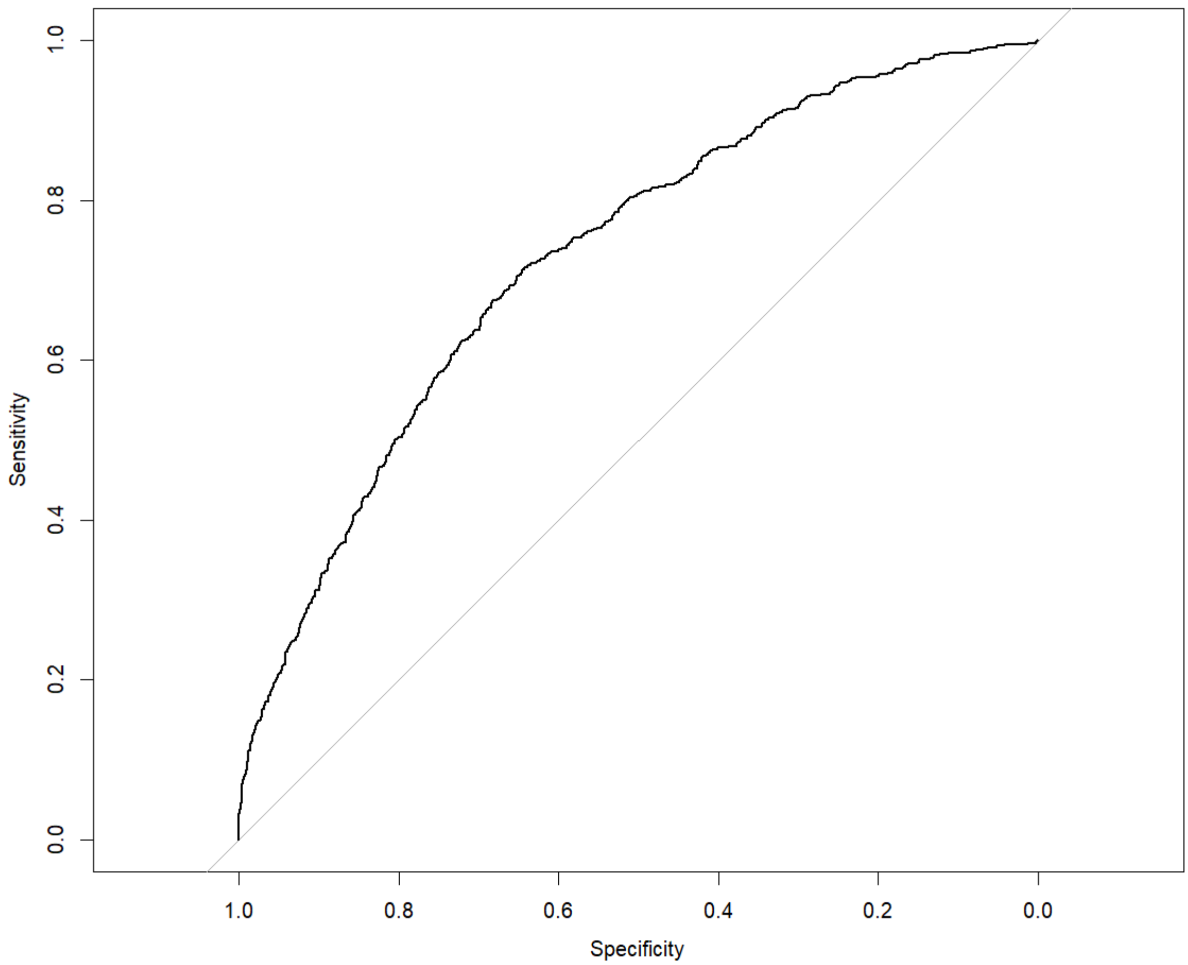
I tried again the models in different options with new variables.

	Variables	AIC	ACC	Split
Model1	All	2265.1	85.2085	0.7
Model2	Correlation > 0.1	2304.1	84.81511	0.7
Model3	P values < 0.1	2256.7	85.12982	0.7
Model4	All	2591.8	85.49528	0.8
Model5	Correlation > 0.1	2642.9	85.25943	0.8
Model6	P values < 0.1	2587.1	85.37736	0.8
Model7	All	2916.5	85.5792	0.9
Model8	Correlation > 0.1	2974.5	84.86998	0.9
Model9	P values < 0.1	2912.7	85.34279	0.9

Finally, the chosen model is:

The best model for our data is binomial because the target value is just 0 and 1.

Male, age, cigsPerDay, prevalentStroke, sysBP and glucose are independent variables and the split ratio 0.9 which the accuracy of this model is 85.5792 and the AIC is 2912.7.



Area under the curve: 0.7286.

Posterior

Y_i is Bernoulli distributed with $p_i = P(Y_i = 1), i = 1,2, \dots, n$. Then the logistic regression model for this data is:

$$\log(p_i) = \delta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5}, \qquad i = 1,2, \dots, n$$

$$p_i = \frac{e^{\delta_i}}{1 + e^{\delta_i}} = \frac{e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5}}}{1 + e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5}}}$$

The likelihood function according to the model is

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left(\frac{e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\beta_3 x_{i3}+\beta_4 x_{i4}+\beta_5 x_{i5}}}{1+e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\beta_3 x_{i3}+\beta_4 x_{i4}+\beta_5 x_{i5}}} \right)^{y_i} \left(1 - \frac{e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\beta_3 x_{i3}+\beta_4 x_{i4}+\beta_5 x_{i5}}}{1+e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\beta_3 x_{i3}+\beta_4 x_{i4}+\beta_5 x_{i5}}} \right)^{1-y_i} \quad (1)$$

The β_j parameter, j = 1,2, …, k, can be in the range $(-\infty, \infty)$ and there is no information regarding previous studies regarding $\boldsymbol{\beta}$. Therefore, the prior distribution for β_j is assumed to be normally distributed with mean μ_j and variance σ_j^2.

All β_js are assumed to be independent, so the joint prior distribution for all the regressioncoefficients can be written as:

$$\pi(\boldsymbol{\beta}) = \pi(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5) = \prod_{j=0}^{5} \frac{1}{\sigma_j\sqrt{2\pi}} exp\left[-\frac{(\beta_j-\mu_j)^2}{2\sigma_j^2}\right] \quad (2)$$

Then the posterior distribution can be denoted by $\pi(\boldsymbol{\beta}|data)$. From equation 1 and equation 2, then

$$\pi(\boldsymbol{\beta}|data) \propto L(\boldsymbol{\beta})\pi(\boldsymbol{\beta})$$

$$\propto \prod_{i=1}^{n} \left(\frac{e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\beta_3 x_{i3}+\beta_4 x_{i4}+\beta_5 x_{i5}}}{1+e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\beta_3 x_{i3}+\beta_4 x_{i4}+\beta_5 x_{i5}}} \right)^{y_i} \left(1 - \frac{e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\beta_3 x_{i3}+\beta_4 x_{i4}+\beta_5 x_{i5}}}{1+e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\beta_3 x_{i3}+\beta_4 x_{i4}+\beta_5 x_{i5}}} \right)^{1-y_i} \times \prod_{j=0}^{5} \frac{1}{\sigma_j\sqrt{2\pi}} exp\left[-\frac{\left(\beta_j-\mu_j\right)^2}{2\sigma_j^2}\right].$$

The posterior distribution is a non-closed form since it does not form a particular distribution. Thus, computational techniques are needed to obtain the Bayes estimator (in this case β). MCMC simulation with Gibbs sampling will be used to obtain the Bayes estimator.

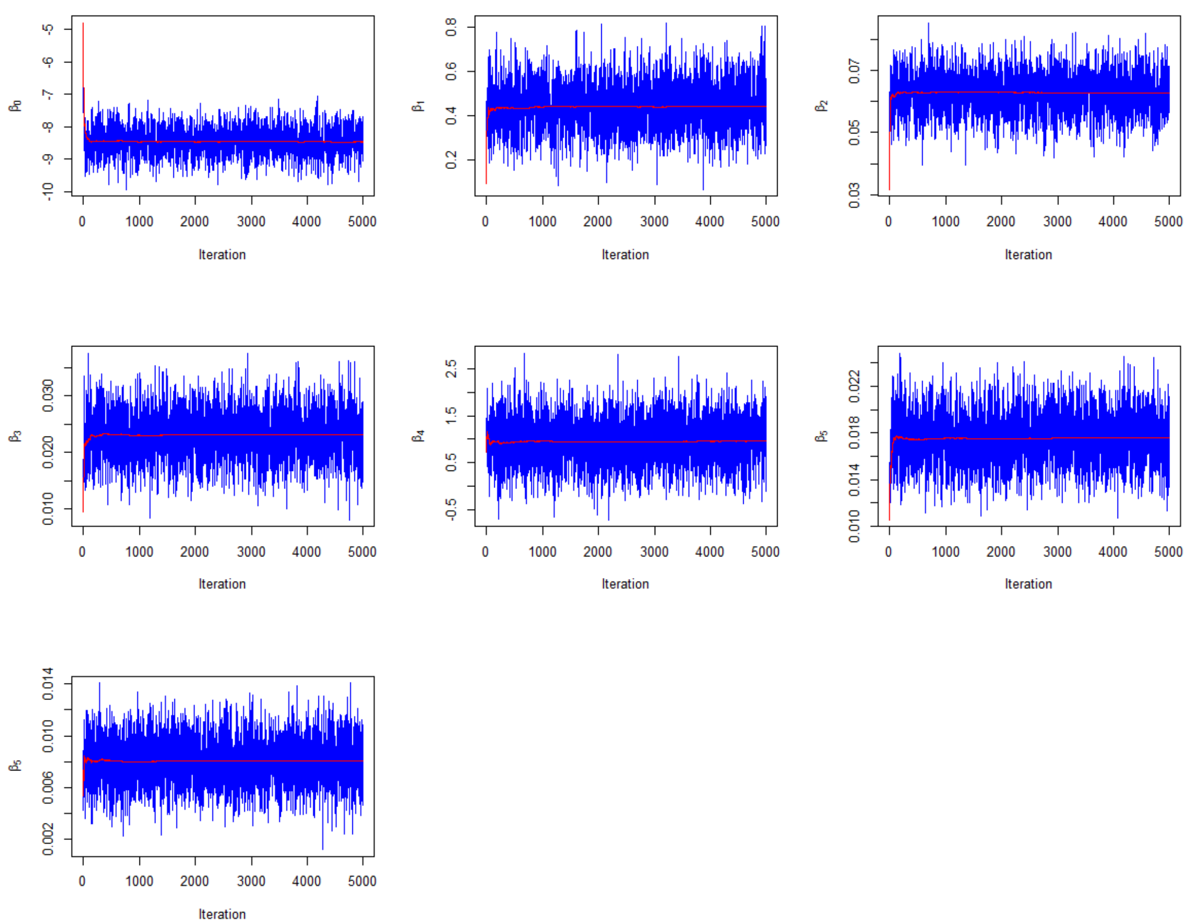
JAGS

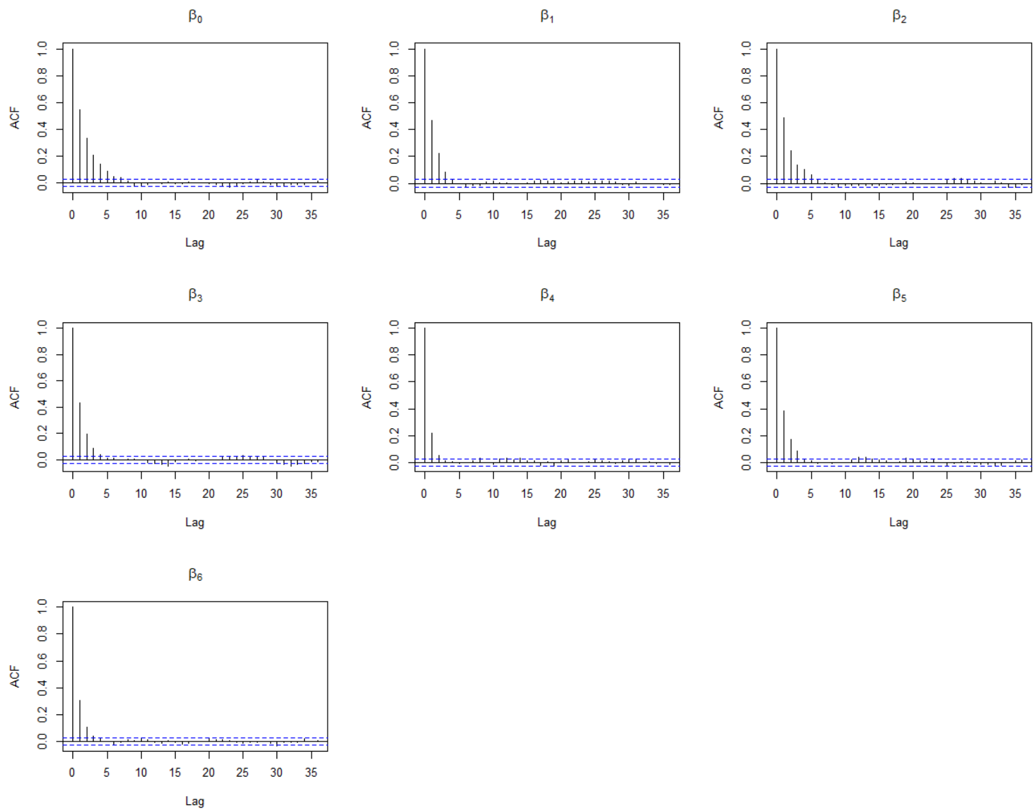
After that I run JAGS function. For Jags function first I need to create a function for model file, the model file Y_i is Bernoulli and each Beta is normal distribution as follows:

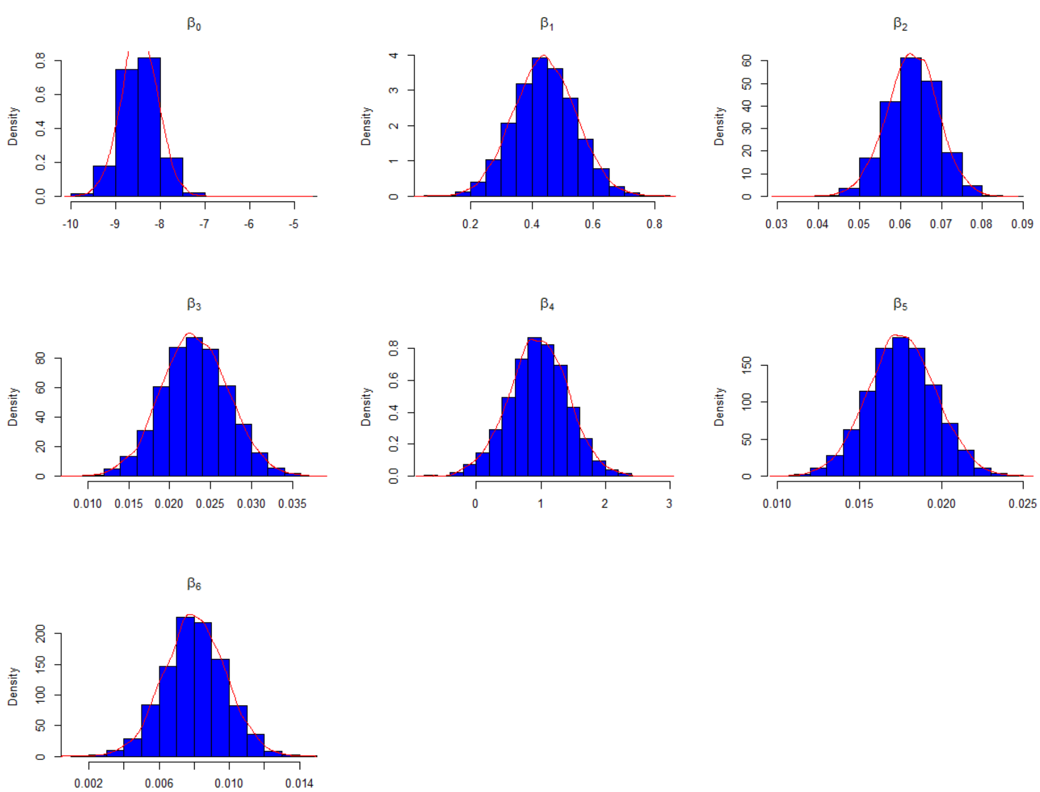
```
> bern_model <- function() {
+   for (i in 1:N) {
+     y[i] ~ dbern(p[i])
+     logit(p[i]) = beta0+(beta1*x1[i])+(beta2*x2[i])+(beta3*x3[i])+(beta4*x4[i])+
+       (beta5*x5[i])+(beta6*x6[i])}
+   beta0 ~ dnorm(0,1.0E-3)
+   beta1 ~ dnorm(0,1.0E-3)
+   beta2 ~ dnorm(0,1.0E-3)
+   beta3 ~ dnorm(0,1.0E-3)
+   beta4 ~ dnorm(0,1.0E-3)
+   beta5 ~ dnorm(0,1.0E-3)
+   beta6 ~ dnorm(0,1.0E-3)
+ }
```

For Jags function, the number of chains is 6, number of iterations is 6000 and the length of burn in 1000.

Below you can see the trace, ACF and histogram plots of the Jags result.




β_0
β_1
β_2
β_3
β_4
β_5
β_6
ACF
Lag


β_0
β_1
β_2
β_3
β_4
β_5
β_6
Density

As you see on the graphs the Convergence of the model is good enough and the histogram plot shows the normal distribution for our model.

Evaluation

For making sure that the model is good enough and, I used some other convergence algorithms like MC error, acceptance rate, Gewek's test, Raftery and Lewis's diagnostic and Heidelberger and Welch's convergence diagnostic.

```
> round(t(Result),4)
          Beta0   Beta1  Beta2  Beta3   Beta4   Beta5  Beta6
MC error 0.0112  0.0023 0.0002 0.0001  0.0081  0.0000 0.0000
geweke   0.6098 -1.5843 0.5363 0.2153 -1.7314 -1.2502 0.5142
raftery  1.5400  1.2700 1.4300 1.2900  1.1200  1.2300 1.1600
heidel   0.9927  0.1043 0.5411 0.9824  0.3659  0.2186 0.9335
```

As we can see the MCMC error is too low, so we could find that the model is convergent. You can find it in the graphs too.

Geweke proposed a convergence diagnostic for Markov chains. This diagnostic is based on a test for equality of the means of the first and last part of a Markov chain. For Geweke's test the good result for convergence should be between -.196 to 1.96. as you can see, for our model. The Geweke's is between these numbers.

Raftery and Lewis (1992) introduced an MCMC diagnostic that estimates the number of iterations needed for a given level of precision in posterior samples, as well as estimating burn-in, when quantiles are the posterior summaries of interest. If this test be less than 5, the model is good.

Heidelberger and Welch proposed a two-part MCMC convergence diagnostic that calculates a test statistic to accept or reject the null hypothesis that the Markov chain is from a stationary distribution. If the value of this test is more than 0.05, we can say that the model is convergent or not.

The DIC is 2918.1.

Confidence Interval

```
> round(t(inter_result),4)
                 Beta0  Beta1  Beta2  Beta3  Beta4  Beta5  Beta6
lower.classic  -9.2765 0.2434 0.0506 0.0151 0.0585 0.0134 0.0046
upper.classic  -7.6673 0.6441 0.0751 0.0309 1.8497 0.0217 0.0114
length.classic  1.6092 0.4006 0.0245 0.0158 1.7911 0.0083 0.0068
lower.HPD      -9.2648 0.2455 0.0503 0.0147 0.0224 0.0137 0.0048
upper.HPD      -7.6831 0.6360 0.0749 0.0308 1.8349 0.0218 0.0116
length.HPD      1.5817 0.3905 0.0245 0.0161 1.8125 0.0081 0.0068
lower.EQ       -9.2742 0.2475 0.0505 0.0150 0.0401 0.0136 0.0046
upper.EQ       -7.6862 0.6392 0.0750 0.0312 1.8538 0.0217 0.0114
length.EQ       1.5880 0.3916 0.0246 0.0162 1.8137 0.0081 0.0068
```

A confidence interval is the mean of your estimate plus and minus the variation in that estimate. This is the range of values you expect your estimate to fall between if you redo your test, within a certain level of confidence.

```
> round(t(inter_result),4)[c(3,6,9),]
                Beta0  Beta1  Beta2  Beta3  Beta4  Beta5  Beta6
length.classic 1.6092 0.4006 0.0245 0.0158 1.7911 0.0083 0.0068
length.HPD     1.5817 0.3905 0.0245 0.0161 1.8125 0.0081 0.0068
length.EQ      1.5880 0.3916 0.0246 0.0162 1.8137 0.0081 0.0068
```

The lengths of confidence interval of HPD of all the variables are the shortest one. After that the equal tail is the shortest.

Mean, standard deviation, median and quantile

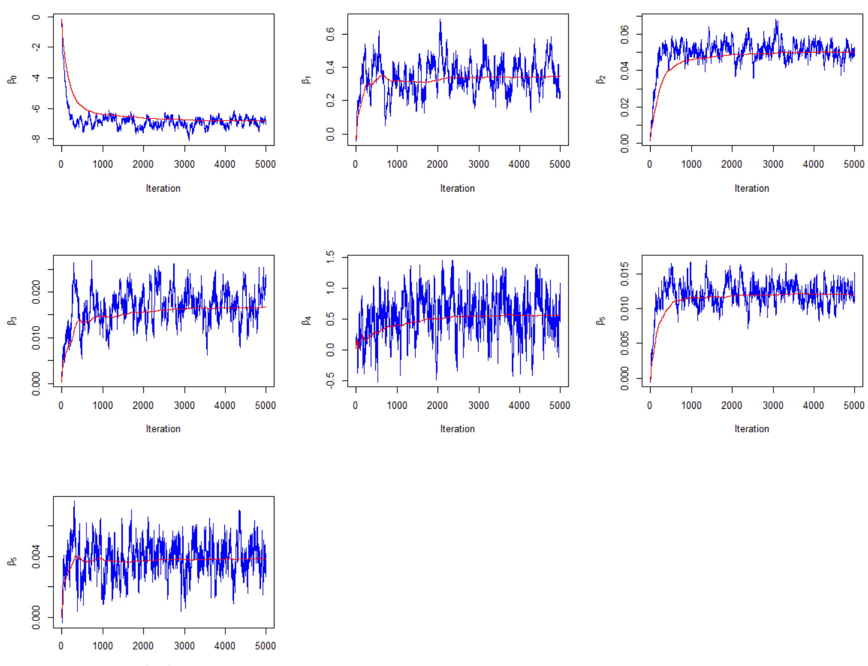
```
> round(t(other_result),4)
                 Beta0  Beta1  Beta2  Beta3  Beta4  Beta5  Beta6
beta.hat.classic -8.4611 0.4436 0.0628 0.0230 0.9650 0.0080 0.0080
beta.hat.bayes   -8.4727 0.4416 0.0628 0.0231 0.9621 0.0176 0.0080
sd                0.4089 0.1015 0.0062 0.0041 0.4588 0.0021 0.0017
2.5%             -9.2742 0.2475 0.0505 0.0150 0.0401 0.0136 0.0046
median           -8.4681 0.4412 0.0630 0.0230 0.9657 0.0176 0.0080
97.5%            -7.6862 0.6392 0.0750 0.0312 1.8538 0.0217 0.0114
```

Poisson

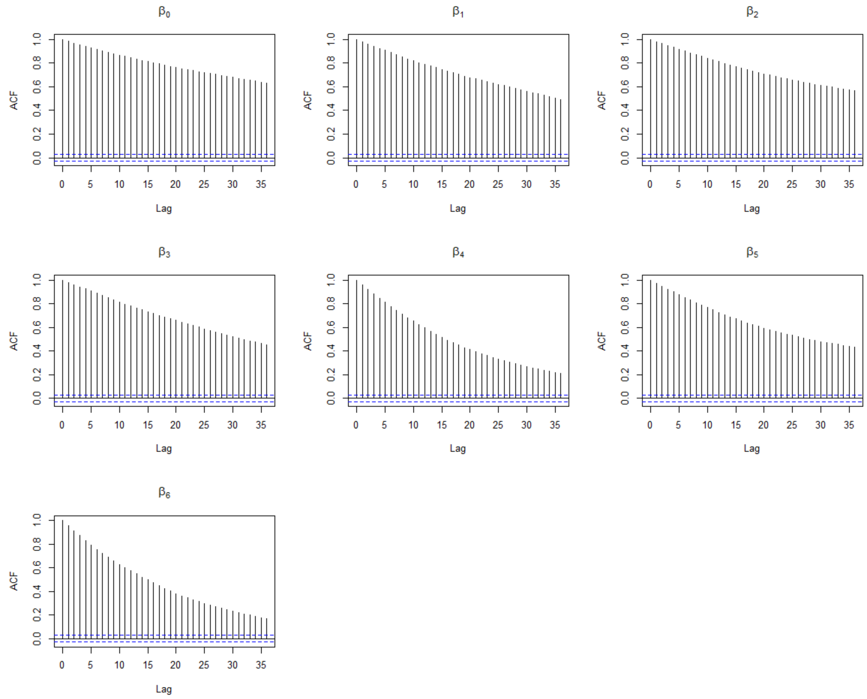
One more time, I run the logistic regression model with Poisson distribution and want to compare it with the Bernoulli model.

I chose Poisson model because the target value is 0 and 1 and Poisson distribution could be all Natural number and starts from 0, so we can use it too for our model.

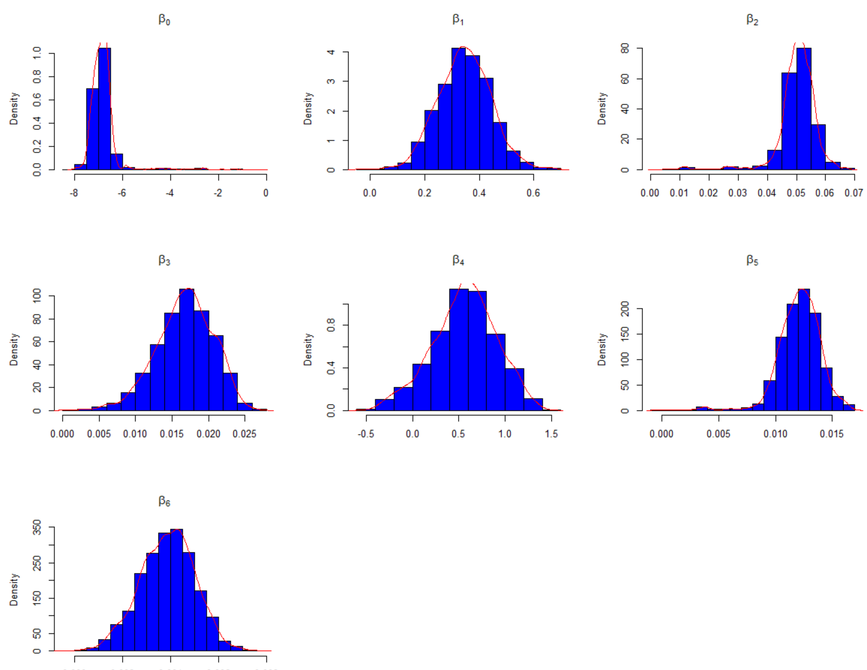
The AIC for Poisson model is 3081.2.



The iteration plot shows us that there isn't any divergence, but the plots of Bernoulli model is better than this one.



The ACF plots show that autocorrelation is large at short lags.



The ACF plots are not good.

```
> round(t(eval_result),4)
            Beta0   Beta1   Beta2   Beta3   Beta4   Beta5   Beta6
MC error   0.1050  0.0139  0.0010  0.0005  0.0338  0.0002  0.0001
geweke     1.2686 -0.4664 -1.2780 -1.1991 -3.5225 -1.0114 -0.3374
raftery   18.6000 11.3000 54.9000 36.0000  9.2800 34.2000  5.9700
heidel     0.5329  0.2256  0.8584  0.4385  0.0783  0.7129  0.3204
```

The Raftery evaluation method reject the MCMC of Poisson.

```
> round(t(inter_result),4)
                  Beta0  Beta1  Beta2  Beta3   Beta4  Beta5  Beta6
lower.classic   -7.5018 0.1730 0.0398 0.0103 -0.0806 0.0090 0.0018
upper.classic   -6.2556 0.5233 0.0611 0.0235  1.1891 0.0155 0.0060
length.classic   1.2461 0.3503 0.0214 0.0132  1.2697 0.0065 0.0042
lower.HPD       -7.5978 0.1536 0.0401 0.0092 -0.1337 0.0089 0.0015
upper.HPD       -6.2313 0.5288 0.0628 0.0237  1.2206 0.0156 0.0059
length.HPD       1.3665 0.3752 0.0227 0.0146  1.3543 0.0067 0.0044
lower.EQ        -7.4891 0.1586 0.0321 0.0086 -0.1823 0.0085 0.0015
upper.EQ        -5.5908 0.5367 0.0605 0.0234  1.1900 0.0154 0.0059
length.EQ        1.8983 0.3782 0.0284 0.0148  1.3724 0.0070 0.0044
```

```
> round(t(other_result),4)
                   Beta0  Beta1  Beta2  Beta3   Beta4  Beta5  Beta6
beta.hat.classic -6.8759 0.3481 0.0504 0.0170  0.6172 0.0041 0.0041
beta.hat.bayes   -6.8178 0.3452 0.0502 0.0167  0.5602 0.0121 0.0039
sd                0.6520 0.0974 0.0069 0.0039  0.3441 0.0019 0.0011
2.5%             -7.4891 0.1586 0.0321 0.0086 -0.1823 0.0085 0.0015
median           -6.8812 0.3457 0.0508 0.0169  0.5752 0.0122 0.0039
97.5%            -5.5908 0.5367 0.0605 0.0234  1.1900 0.0154 0.0059
```

The DIC for Poisson model is 13615 which is not good enough in compare with the Bernoulli model.

The models comparison:

Models	AIC	DIC
Bernoulli	2912.7	2918.1
Poisson	3081.2	13615

In this comparison we can completely find that the Bernoulli model in the same situation of work, is better than Poisson one.

References:

https://pubmed.ncbi.nlm.nih.gov/27803621/

https://www.ahajournals.org/doi/full/10.1161/HYPERTENSIONAHA.120.14929

https://thoracickey.com/blood-pressure-regulation-2/

https://www.kaggle.com/datasets/dileep070/heart-disease-prediction-using-logistic-regression

https://www.heart.org/en/health-topics/consumer-healthcare/what-is-cardiovascular-disease/coronary-artery-disease