Math 459: Bayesian Statistics – Homework 4

Use the `heart` data in the `ncvreg` package. This dataset contains 462 observations on 10 variables. The response variable is `chd`, which indicates whether or not coronary heart disease was present at the time of observation.

(a) Using a package of your choosing, fit Bayesian probit and logistic regression models. Explain which package and priors are used.

(b) Using a method (Laplace and/or MCMC) of your choosing, estimate the marginal likelihood for each model and hence estimate the Bayes factor. Explain how the marginal likelihood is estimated according to the method you have chosen. Give an interpretation of the computed value of the Bayes factor.

(c) Remove one data point from the original dataset. Fit the probit and logistic models again. Provide a HPD interval for the predicted value of the response using the predictor values for the one observation left out from the original dataset. This means you need to simulate from the posterior predictive distribution. Give a 95% HPD interval, as well as an estimate of the posterior predictive mean. Compare this estimate to the actual observed value of that response variable.

Solution.

(a) The `R` package `MCMCpack` is used to fit Bayesian probit and logistic regression models, with multivariate normal priors. Details are included in the following `R` code.

```
#load data package
library(ncvreg)
data(heart)
summary(heart)
```

```
##       sbp            tobacco              ldl            adiposity
##  Min.   :101.0   Min.   : 0.0000   Min.   : 0.980   Min.   : 6.74
##  1st Qu.:124.0   1st Qu.: 0.0525   1st Qu.: 3.283   1st Qu.:19.77
```

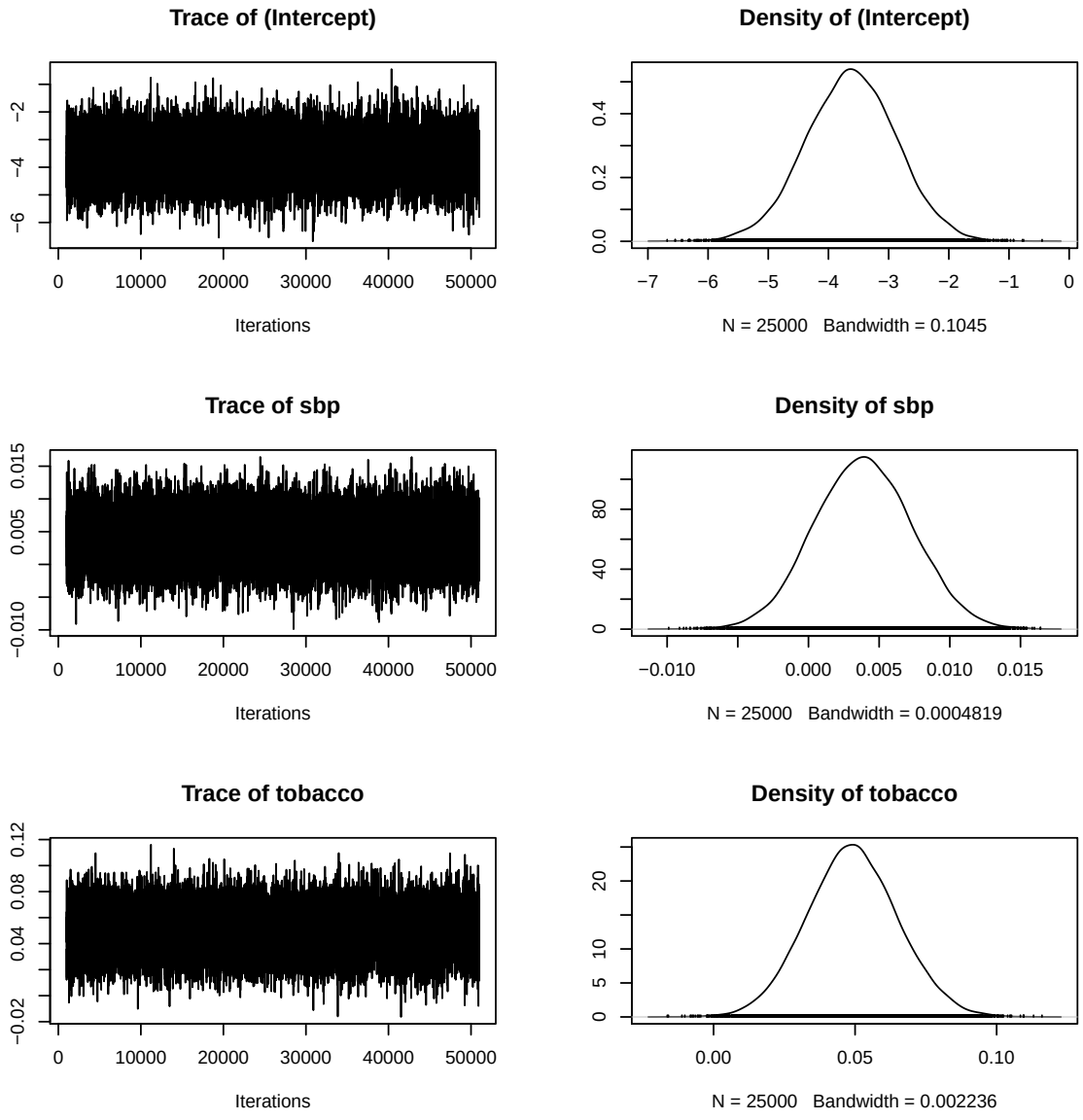
```
##  Median :134.0   Median : 2.0000   Median : 4.340   Median :26.11
##  Mean   :138.3   Mean   : 3.6356   Mean   : 4.740   Mean   :25.41
##  3rd Qu.:148.0   3rd Qu.: 5.5000   3rd Qu.: 5.790   3rd Qu.:31.23
##  Max.   :218.0   Max.   :31.2000   Max.   :15.330   Max.   :42.49
##     famhist           typea          obesity         alcohol
##  Min.   :0.0000   Min.   :13.0   Min.   :14.70   Min.   :  0.00
##  1st Qu.:0.0000   1st Qu.:47.0   1st Qu.:22.98   1st Qu.:  0.51
##  Median :0.0000   Median :53.0   Median :25.80   Median :  7.51
##  Mean   :0.4156   Mean   :53.1   Mean   :26.04   Mean   : 17.04
##  3rd Qu.:1.0000   3rd Qu.:60.0   3rd Qu.:28.50   3rd Qu.: 23.89
##  Max.   :1.0000   Max.   :78.0   Max.   :46.58   Max.   :147.19
##       age              chd
##  Min.   :15.00   Min.   :0.0000
##  1st Qu.:31.00   1st Qu.:0.0000
##  Median :45.00   Median :0.0000
##  Mean   :42.82   Mean   :0.3463
##  3rd Qu.:55.00   3rd Qu.:1.0000
##  Max.   :64.00   Max.   :1.0000
```

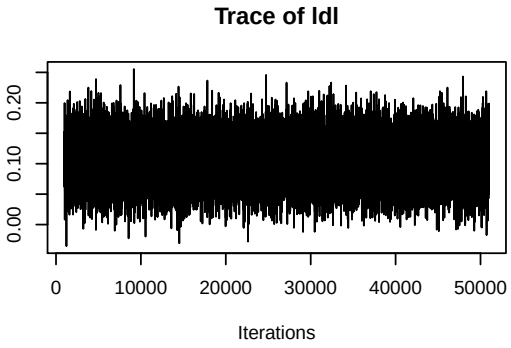
```
#use MCMCpack to do the regression
library(MCMCpack)
```

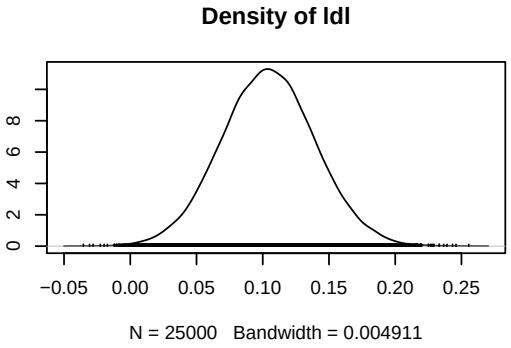
```
## Loading required package:  coda
## Loading required package:  lattice
## Loading required package:  MASS
## ##
## ## Markov Chain Monte Carlo Package (MCMCpack)
## ## Copyright (C) 2003-2016 Andrew D. Martin, Kevin M. Quinn, and Jong Hee Park
## ##
## ## Support provided by the U.S. National Science Foundation
## ## (Grants SES-0350646 and SES-0350613)
## ##
```

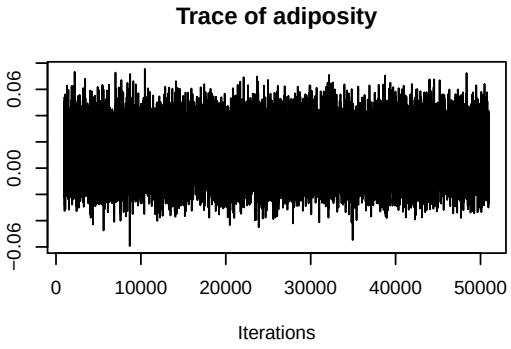
```r
#Bayesian Probit Regression Model with multivariate normal prior
fit1 <-MCMCprobit(chd~.,data=heart,thin=2, burnin=1000,mcmc =50000, b0=0, B0=.001)
summary(fit1)

##
## Iterations = 1001:50999
## Thinning interval = 2
## Number of chains = 1
## Sample size per chain = 25000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##                    Mean       SD  Naive SE Time-series SE
## (Intercept) -3.621e+00 0.747084 4.725e-03      6.956e-03
## sbp          3.892e-03 0.003445 2.179e-05      2.741e-05
## tobacco      4.913e-02 0.016011 1.013e-04      1.262e-04
## ldl          1.043e-01 0.035287 2.232e-04      2.833e-04
## adiposity    1.272e-02 0.017253 1.091e-04      1.524e-04
## famhist      5.441e-01 0.136905 8.659e-04      1.127e-03
## typea        2.395e-02 0.007185 4.544e-05      6.423e-05
## obesity     -4.123e-02 0.025773 1.630e-04      2.295e-04
## alcohol      8.071e-07 0.002723 1.722e-05      2.183e-05
## age          2.667e-02 0.007039 4.452e-05      6.514e-05
##
## 2. Quantiles for each variable:
##
##                   2.5%       25%        50%       75%     97.5%
## (Intercept) -5.105905 -4.122542 -3.615e+00 -3.118454 -2.156192
## sbp         -0.002799  0.001542  3.879e-03  0.006218  0.010672
## tobacco      0.018052  0.038333  4.901e-02  0.059754  0.081191
## ldl          0.034891  0.080571  1.041e-01  0.127618  0.174278
```

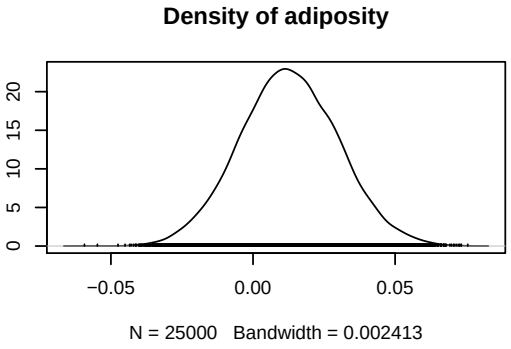
```
## adiposity   -0.020907  0.001005  1.248e-02  0.024487  0.047104
## famhist      0.275640  0.451602  5.434e-01  0.636691  0.811829
## typea        0.009912  0.019178  2.396e-02  0.028747  0.038076
## obesity     -0.091530 -0.058762 -4.107e-02 -0.023844  0.008966
## alcohol     -0.005392 -0.001825  1.519e-05  0.001832  0.005336
## age          0.013054  0.021914  2.663e-02  0.031390  0.040574

plot(fit1)
```

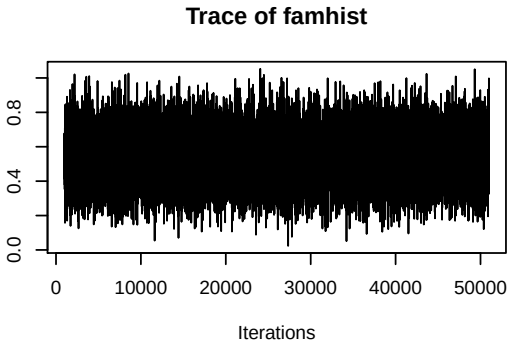


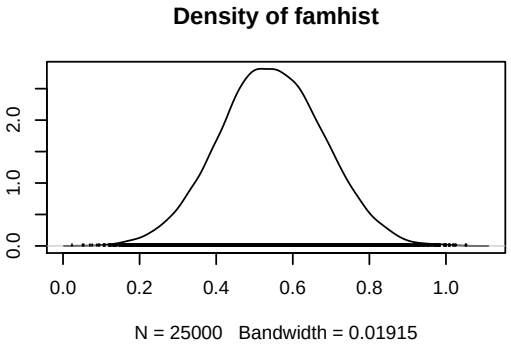

Trace of ldl
0.00
0.10
0.20
0
10000
20000
30000
40000
50000
Iterations

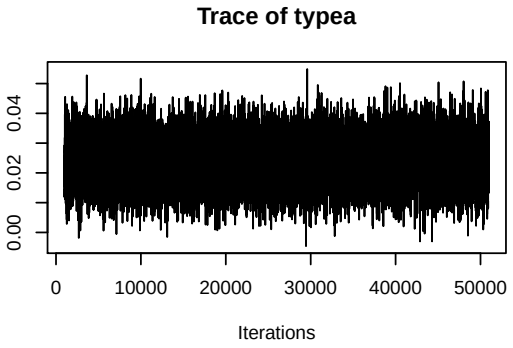

Density of ldl
0
2
4
6
8
−0.05
0.00
0.05
0.10
0.15
0.20
0.25
N = 25000 Bandwidth = 0.004911

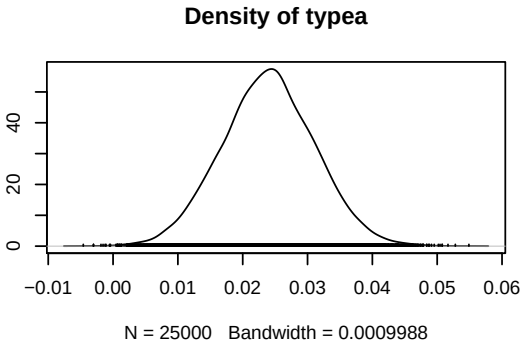

Trace of adiposity
−0.06
0.00
0.06
0
10000
20000
30000
40000
50000
Iterations

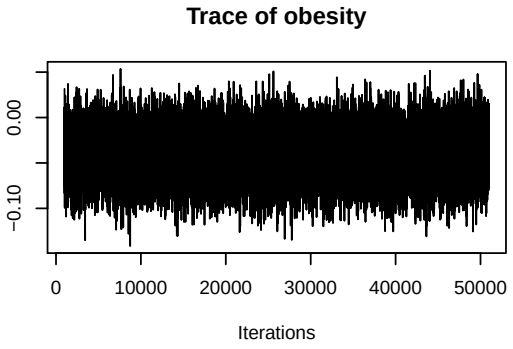

Density of adiposity
0
5
10
15
20
−0.05
0.00
0.05
N = 25000 Bandwidth = 0.002413

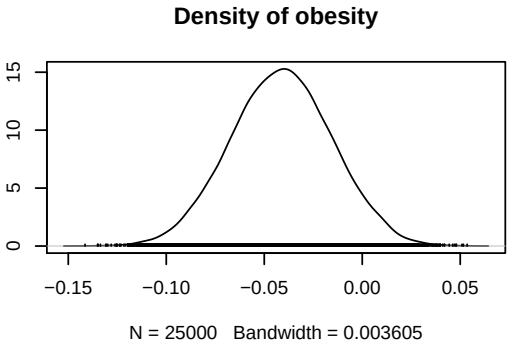

Trace of famhist
0.0
0.4
0.8
0
10000
20000
30000
40000
50000
Iterations

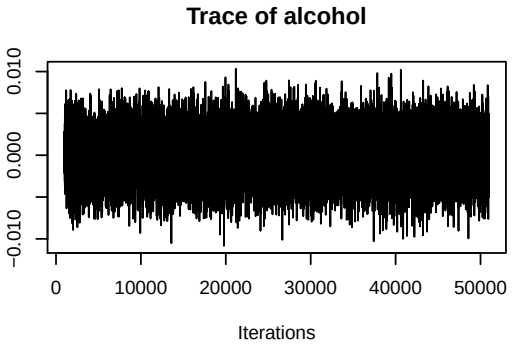

Density of famhist
0.0
1.0
2.0
0.0
0.2
0.4
0.6
0.8
1.0
N = 25000 Bandwidth = 0.01915

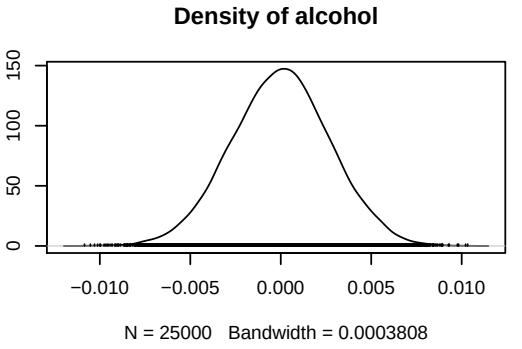

Trace of typea
0.00
0.02
0.04
0
10000
20000
30000
40000
50000
Iterations


Density of typea
0
20
40
−0.01
0.00
0.01
0.02
0.03
0.04
0.05
0.06
N = 25000 Bandwidth = 0.0009988

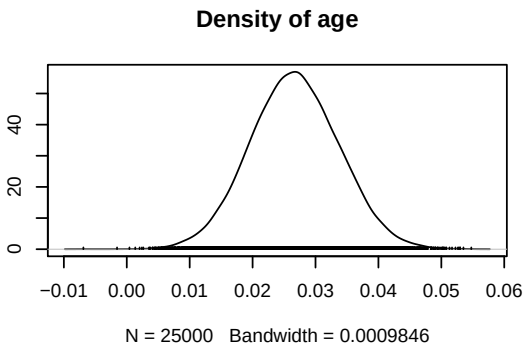

Trace of obesity
−0.10
0.00
0
10000
20000
30000
40000
50000
Iterations


Density of obesity
0
5
10
15
−0.15
−0.10
−0.05
0.00
0.05
N = 25000 Bandwidth = 0.003605


Trace of alcohol
−0.010
0.000
0.010
0
10000
20000
30000
40000
50000
Iterations


Density of alcohol
0
50
100
150
−0.010
−0.005
0.000
0.005
0.010
N = 25000 Bandwidth = 0.0003808

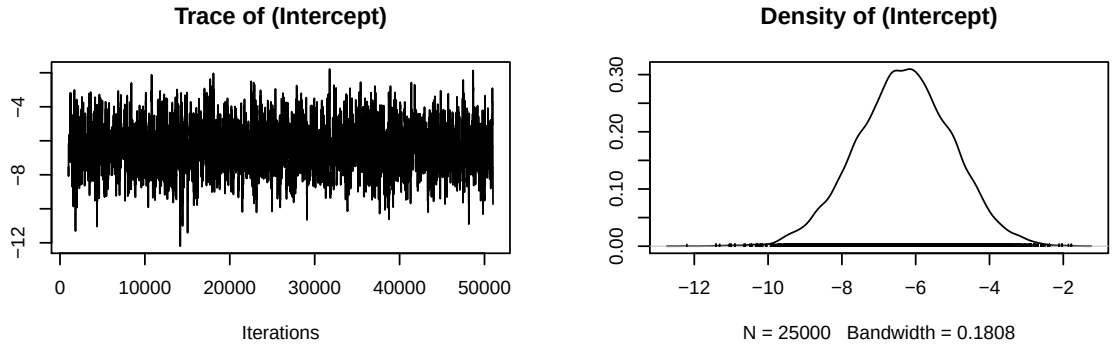

Trace of age
0.00
0.02
0.04
0
10000
20000
30000
40000
50000
Iterations

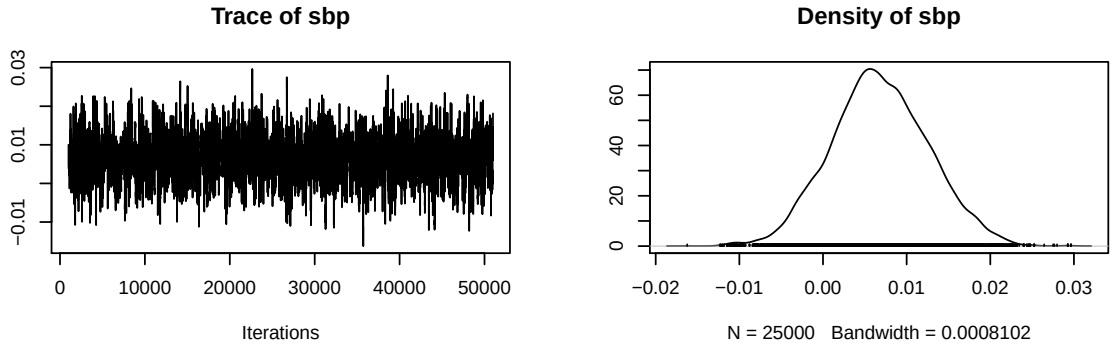

Density of age
0
20
40
−0.01
0.00
0.01
0.02
0.03
0.04
0.05
0.06
N = 25000 Bandwidth = 0.0009846

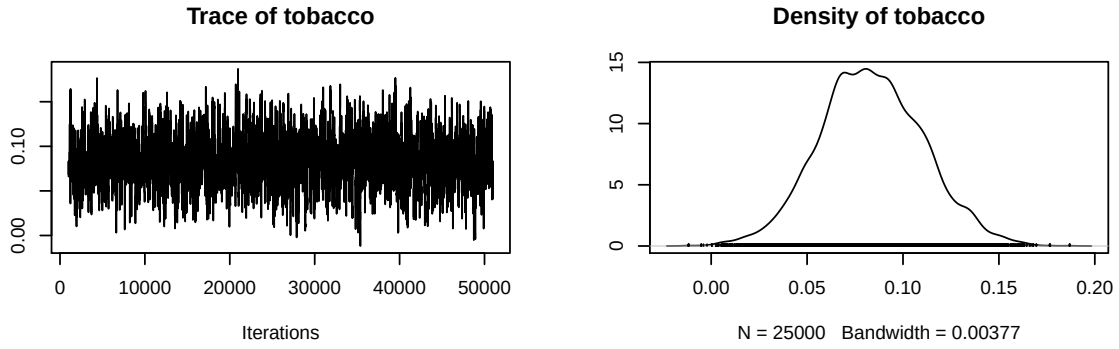
```
# Bayesian Logistic Regression Model with multivariate normal prior
fit2 <-MCMClogit(chd~.,data=heart,thin=2, burnin=1000,mcmc =50000, b0=0, B0=.001)
summary(fit2)

##
## Iterations = 1001:50999
## Thinning interval = 2
## Number of chains = 1
## Sample size per chain = 25000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##                    Mean       SD  Naive SE Time-series SE
## (Intercept) -6.309e+00 1.292887 8.177e-03      0.0349008
## sbp          6.804e-03 0.005793 3.664e-05      0.0001605
## tobacco      8.357e-02 0.026953 1.705e-04      0.0007653
## ldl          1.770e-01 0.059070 3.736e-04      0.0015856
## adiposity    2.014e-02 0.029060 1.838e-04      0.0007813
## famhist      9.474e-01 0.230186 1.456e-03      0.0063331
## typea        4.078e-02 0.012422 7.857e-05      0.0003431
## obesity     -6.606e-02 0.043531 2.753e-04      0.0011372
## alcohol     -9.108e-05 0.004472 2.828e-05      0.0001237
## age          4.615e-02 0.011917 7.537e-05      0.0003238
##
## 2. Quantiles for each variable:
##
##                  2.5%        25%       50%       75%     97.5%
## (Intercept) -8.872583 -7.1602667 -6.292613 -5.427414 -3.817365
## sbp         -0.004236  0.0029154  0.006660  0.010747  0.018402
## tobacco      0.031735  0.0654747  0.082735  0.102046  0.137569
## ldl          0.059741  0.1371210  0.176064  0.216899  0.292115
```

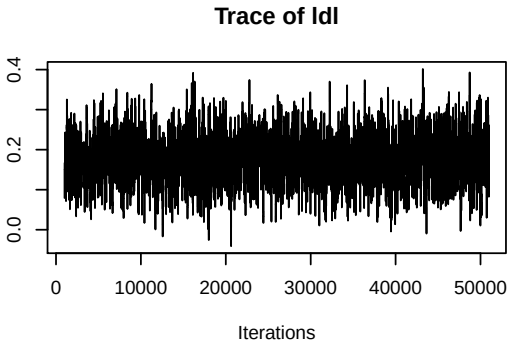
```
## adiposity   -0.035226 -0.0001532  0.019757  0.039466  0.077776
## famhist      0.493160  0.7947355  0.946496  1.097733  1.411638
## typea        0.016837  0.0326539  0.040362  0.049156  0.065421
## obesity     -0.151264 -0.0958790 -0.065044 -0.035736  0.017472
## alcohol     -0.009166 -0.0030584 -0.000122  0.002898  0.008713
## age          0.022656  0.0383769  0.045899  0.054031  0.069862
```

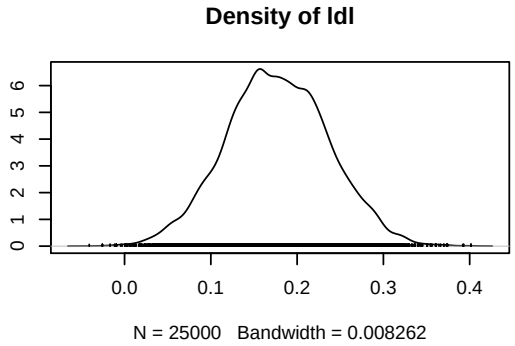
```
plot(fit2)
```

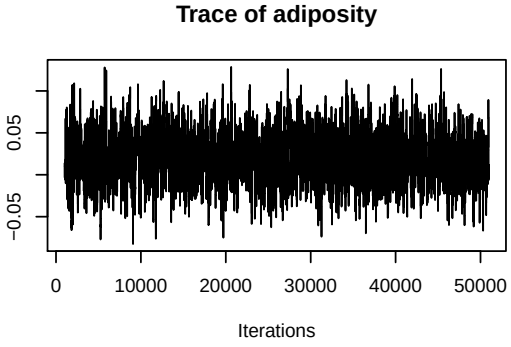


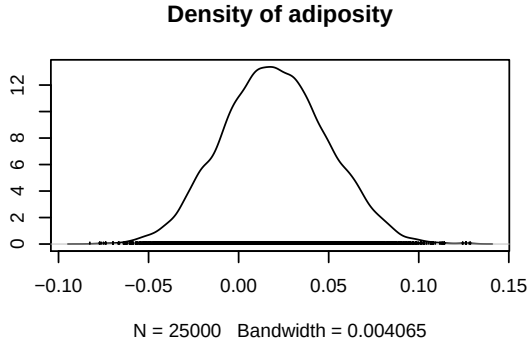


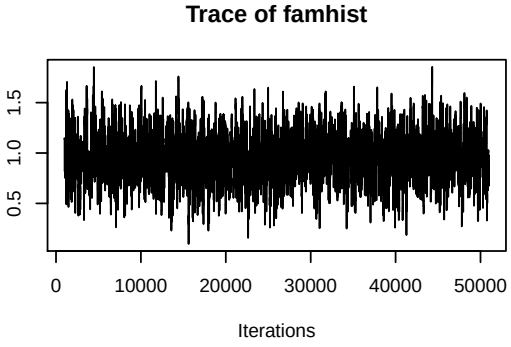


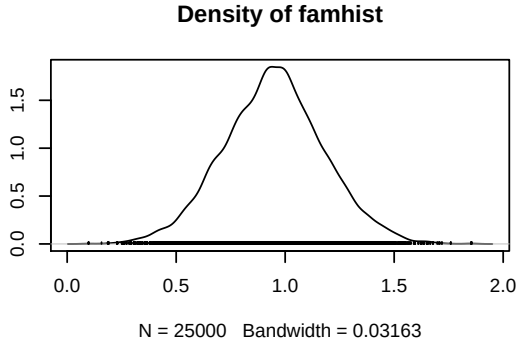

Trace of ldl
0.0
0.2
0.4
0
10000
20000
30000
40000
50000
Iterations

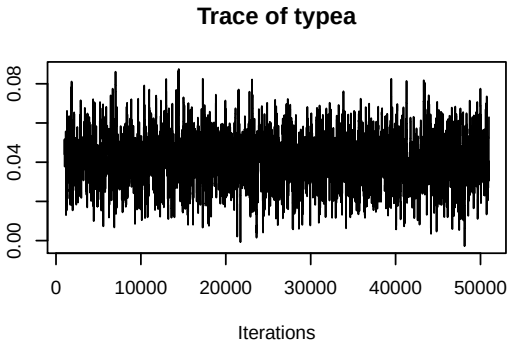

Density of ldl
0 1 2 3 4 5 6
0.0
0.1
0.2
0.3
0.4
N = 25000 Bandwidth = 0.008262

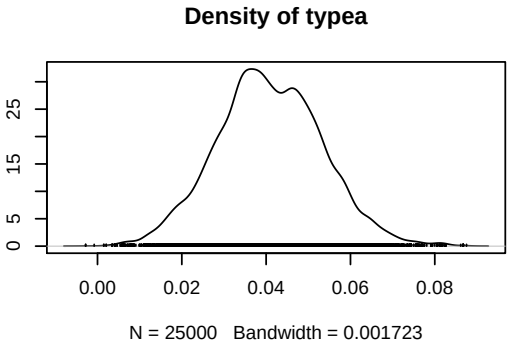

Trace of adiposity
−0.05
0.05
0
10000
20000
30000
40000
50000
Iterations

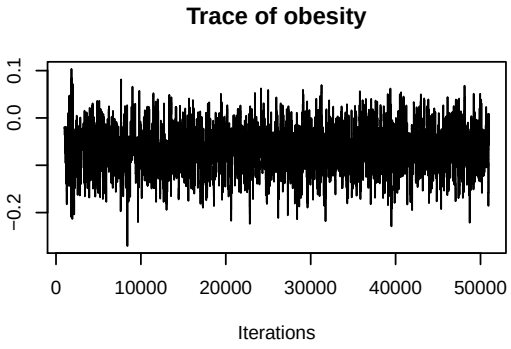

Density of adiposity
0 2 4 6 8 12
−0.10
−0.05
0.00
0.05
0.10
0.15
N = 25000 Bandwidth = 0.004065

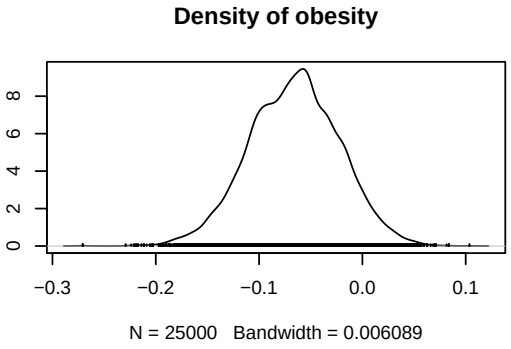

Trace of famhist
0.5
1.0
1.5
0
10000
20000
30000
40000
50000
Iterations

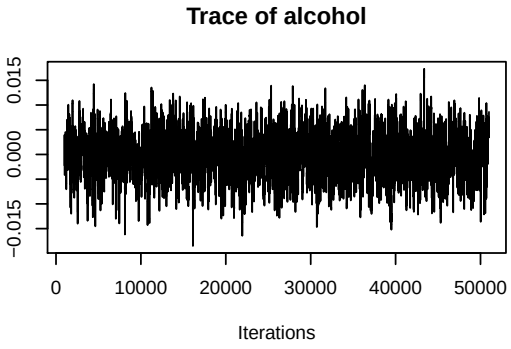

Density of famhist
0.0 0.5 1.0 1.5
0.0
0.5
1.0
1.5
2.0
N = 25000 Bandwidth = 0.03163

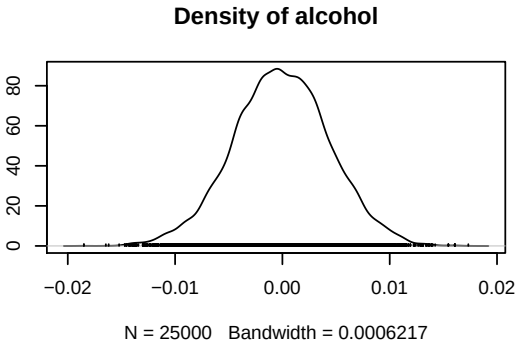

Trace of typea
0.00
0.04
0.08
0
10000
20000
30000
40000
50000
Iterations


Density of typea
0
5
15
25
0.00
0.02
0.04
0.06
0.08
N = 25000 Bandwidth = 0.001723

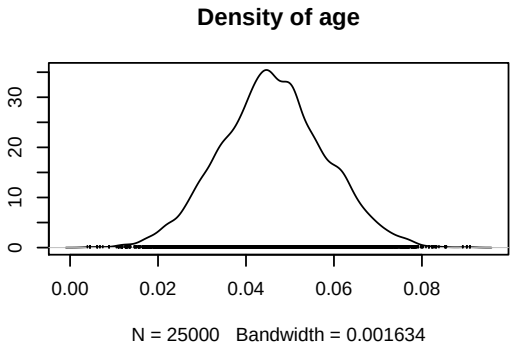

Trace of obesity
−0.2
0.0
0.1
0
10000
20000
30000
40000
50000
Iterations


Density of obesity
0
2
4
6
8
−0.3
−0.2
−0.1
0.0
0.1
N = 25000 Bandwidth = 0.006089


Trace of alcohol
−0.015
0.000
0.015
0
10000
20000
30000
40000
50000
Iterations


Density of alcohol
0
20
40
60
80
−0.02
−0.01
0.00
0.01
0.02
N = 25000 Bandwidth = 0.0006217


Trace of age
0.02
0.06
0
10000
20000
30000
40000
50000
Iterations


Density of age
0
10
20
30
0.00
0.02
0.04
0.06
0.08
N = 25000 Bandwidth = 0.001634

□

(b) To estimate the marginal likelihood, we can use Laplace's method or the method of Chib (1995), with corresponding function notations, `marginal.likelihood = c("Laplace","Chib95")`, developed in `R` package `MCMCpack`. Here, we choose Laplace's method and provide a brief explanation of Laplace's method as follows.

Let $\hat{\theta}$ denote the posterior mode and $H(\theta)$ denote the Hessian (second derivative matrix) of the log posterior density. Then the prior predictive density can be approximated as

$$m(y) \approx (2\pi)^{d/2} g(\hat{\theta}) f(y|\hat{\theta})| - H(\hat{\theta})|^{1/2}$$

where d is the number of parameters. On the log scale, we have

$$\log m(y) \approx \frac{d}{2}\log(2\pi) + \log(g(\hat{\theta})f(y|\hat{\theta})) + \frac{1}{2}\log|-H(\hat{\theta})|.$$

Once an `R` function is written to compute the logarithm of the product $f(y|\theta)g(\theta)$, then the function `laplace` can be applied and the component of the output `int` gives an estimate of $\log m(y)$. By applying this method for each model, one can use the computed values of $m(y)$ to compute a Bayes factor. Since the `MCMCpack` can provide the marginal likelihood directly by using the `BayesFactor` function, we will use it to obtain the marginal likelihood and the Bayes factor at the same time. Results are included in the following `R` code.

```
# Model comparison using Bayes factors
# sepcify the Laplace's method of approximation
fit3 <-MCMCprobit(chd~.,data=heart,thin=2, burnin=1000,mcmc =50000,
b0=0, B0=.001, marginal.likelihood="Laplace")
fit4 <-MCMClogit(chd~.,data=heart,thin=2, burnin=1000,mcmc =50000,
b0=0, B0=.001, marginal.likelihood="Laplace")
BayesFactor(fit3,fit4)

## The matrix of Bayes Factors is:
##      fit3    fit4
## fit3    1 0.00579
## fit4  173 1.00000
##
## The matrix of the natural log Bayes Factors is:
```

```
##      fit3  fit4
## fit3 0.00 -5.15
## fit4 5.15  0.00
##
##  fit3 :
##    call =
## MCMCprobit(formula = chd ~ ., data = heart, burnin = 1000, mcmc = 50000,
##     thin = 2, b0 = 0, B0 = 0.001, marginal.likelihood = "Laplace")
##
##    log marginal likelihood =  -312.7153
##
##
##  fit4 :
##    call =
## MCMClogit(formula = chd ~ ., data = heart, burnin = 1000, mcmc = 50000,
##     thin = 2, b0 = 0, B0 = 0.001, marginal.likelihood = "Laplace")
##
##    log marginal likelihood =  -307.5636
```

The marginal likelihood can easily obtained by the result of `log marginal likelihood` in the output. From the result in the matrix of Bayes Factors, we know the probit model `fit4` is much better than the logit model `fit3`, with strength of evidence, *decisive* according to Jeffreys scale and *very strong* according to Kass & Raftery scale. □

(c) Details of `R` code and output are attached as follows.

```
#Remove one data point from the original dataset and fit the probit and logistic models
heart_new = heart[1,]
x = cbind(1,heart_new[1,1:9])
y = heart_new[10]
heart = heart[-1,]
```

```r
# refit the models
fit5 <-MCMCprobit(chd~.,data=heart,thin=2, burnin=1000,mcmc =50000, b0=0, B0=.001)
fit6 <-MCMClogit(chd~.,data=heart,thin=2, burnin=1000,mcmc =50000, b0=0, B0=.001)

# predict the value at testpoint for the two models
pred <- fit6 %*% t(as.matrix(x))
p1 <- 1/(1+exp(-pred))
HPDinterval(mcmc(p1))

##       lower     upper
## 1 0.5142217 0.8613767
## attr(,"Probability")
## [1] 0.95

pred <- fit5 %*% t(as.matrix(x))
p2 <- pnorm(pred)
HPDinterval(mcmc(p2))

##       lower     upper
## 1 0.5127122 0.8605163
## attr(,"Probability")
## [1] 0.95
```

From the result, we find the predictions of both regression models are very close to the actual data 1 we removed.

□