MKTING HW 2

Gihani Dissanayake

February 18, 2018

Assignment 2

Marketing Analytics 2

Gihani Dissanayake

Linear and Hierarchical Linear Models: Bayesian Estimation

```
library(readr)
sow.data = read_csv("~/CreditCard_SOW_data.csv")
```

```
## Parsed with column specification:
## cols(
##   ConsumerID = col_integer(),
##   History = col_integer(),
##   Income = col_double(),
##   WalletShare = col_double(),
##   Promotion = col_double(),
##   Balance = col_integer()
## )
```

```
sow.data$ConsumerID = as.factor(sow.data$ConsumerID)
sow.data$logIncome = log(sow.data$Income)
sow.data$logSowRatio = log(sow.data$WalletShare/(1-sow.data$WalletShare))
head(sow.data)
```

```
## # A tibble: 6 x 8
##   ConsumerID History Income WalletShare Promotion Balance logIncome
##       <fctr>   <int>  <dbl>       <dbl>     <dbl>   <int>     <dbl>
## 1          1      55  82000       0.643       0.5     836  11.31447
## 2          1      55  82000       0.628       0.2     467  11.31447
## 3          1      55  82000       0.567       1.0    1208  11.31447
## 4          1      55  82000       0.638       0.8     792  11.31447
## 5          1      55  82000       0.554       0.7    1215  11.31447
## 6          1      55  82000       0.573       1.1    1248  11.31447
## # ... with 1 more variables: logSowRatio <dbl>
```

```
library(MCMCpack)
```

```
## Loading required package: coda
```

```
## Loading required package: MASS
```

```
## ##
## ## Markov Chain Monte Carlo Package (MCMCpack)
```

```
## ## Copyright (C) 2003-2018 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)
## ##
```

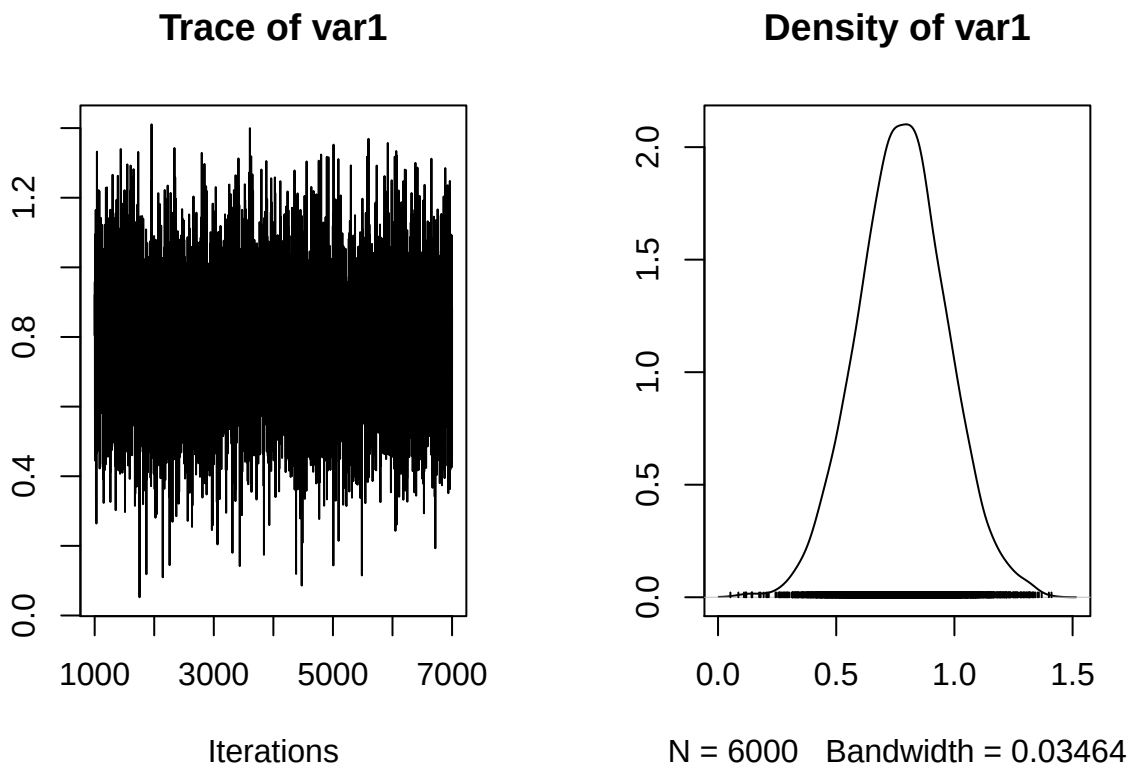
```
library(coda)
library(MASS)

reg1 = MCMCregress(logSowRatio ~ History+Balance+Promotion+History:Promotion+
                     logIncome:Promotion, sow.data, mcmc=6000)
summary(reg1)
```

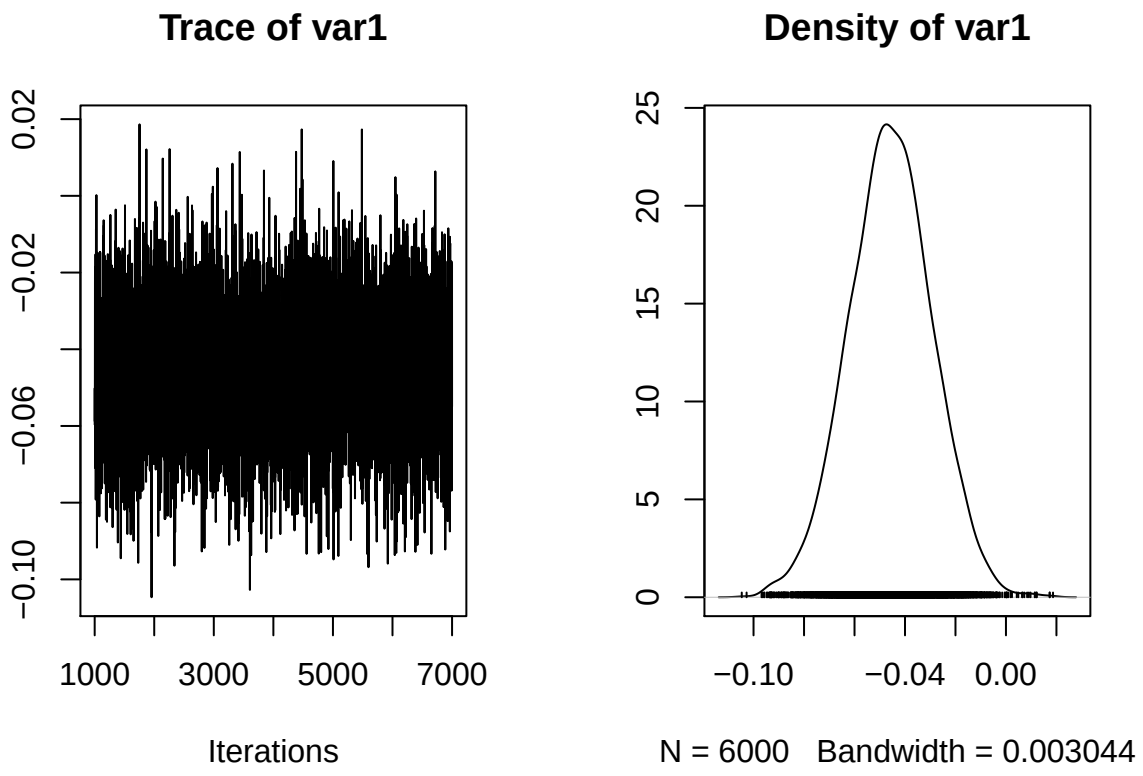
```
##
## Iterations = 1001:7000
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 6000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##                            Mean        SD  Naive SE Time-series SE
## (Intercept)           0.0886515 1.604e-02 2.071e-04      2.071e-04
## History               0.0103993 4.155e-04 5.365e-06      5.365e-06
## Balance              -0.0004959 2.888e-06 3.728e-08      3.728e-08
## Promotion             0.7796326 1.891e-01 2.441e-03      2.441e-03
## History:Promotion    -0.0026062 5.706e-04 7.366e-06      7.366e-06
## Promotion:logIncome -0.0457032 1.656e-02 2.138e-04      2.138e-04
## sigma2                0.0432230 1.022e-03 1.319e-05      1.405e-05
##
## 2. Quantiles for each variable:
##
##                             2.5%        25%        50%       75%      97.5%
## (Intercept)           0.0574844  0.0775755  0.0886492  0.099764  0.1194332
## History               0.0096083  0.0101103  0.0103974  0.010679  0.0111943
## Balance              -0.0005015 -0.0004978 -0.0004959 -0.000494 -0.0004901
## Promotion             0.4146629  0.6539999  0.7803401  0.903460  1.1558051
## History:Promotion    -0.0037377 -0.0029956 -0.0025925 -0.002226 -0.0014945
## Promotion:logIncome -0.0784528 -0.0566332 -0.0458005 -0.034715 -0.0137243
## sigma2                0.0412641  0.0425356  0.0432089  0.043890  0.0452657
```

Based on the the Bayesian posterior intervals, all the regression coefficients (history, balance, promotion, history:promotion, and promotion:logIncome) are significant at the 5% level because none of the 2.5% to 97.5% ranges inculde 0.

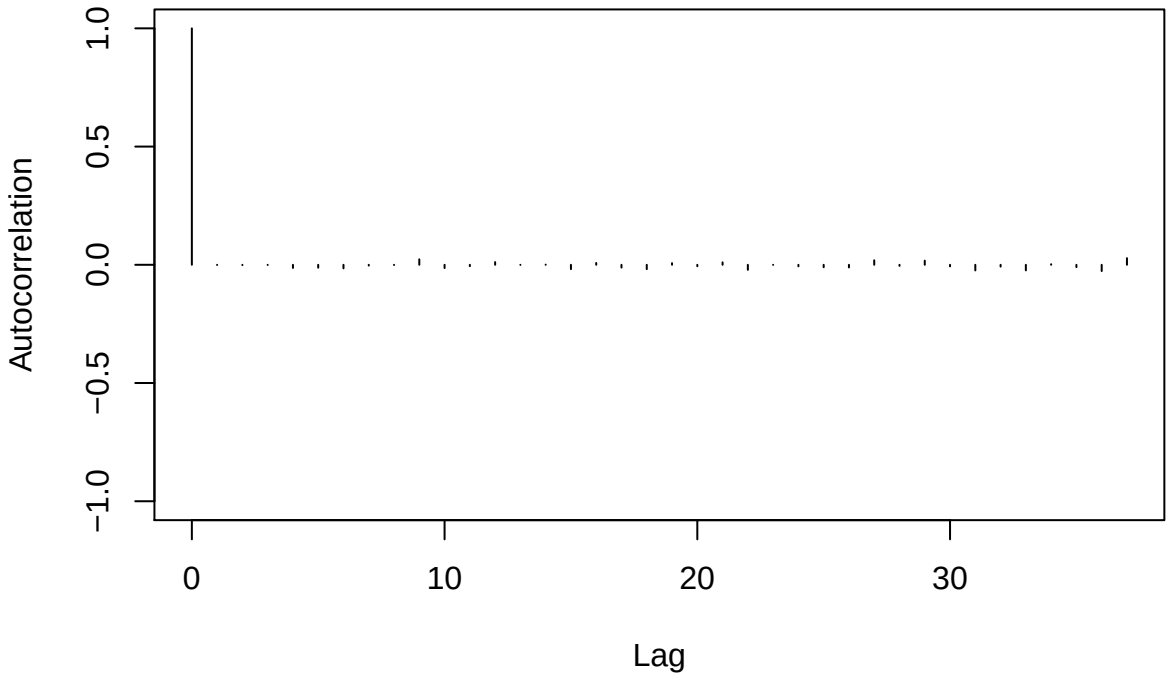
```
plot(reg1[,"Promotion"], type="l")
```



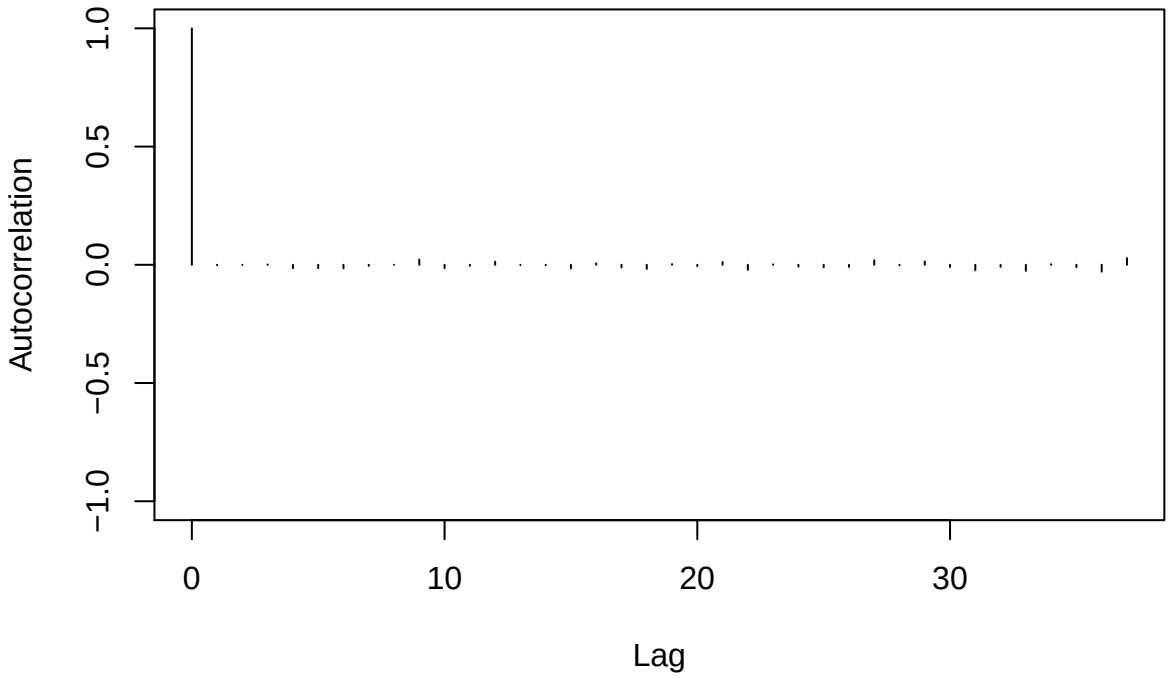

```
plot(reg1[,"Promotion:logIncome"], type="l")
```




```
autocorr.plot(reg1[,"Promotion"])
```




```
autocorr.plot(reg1[,"Promotion:logIncome"])
```



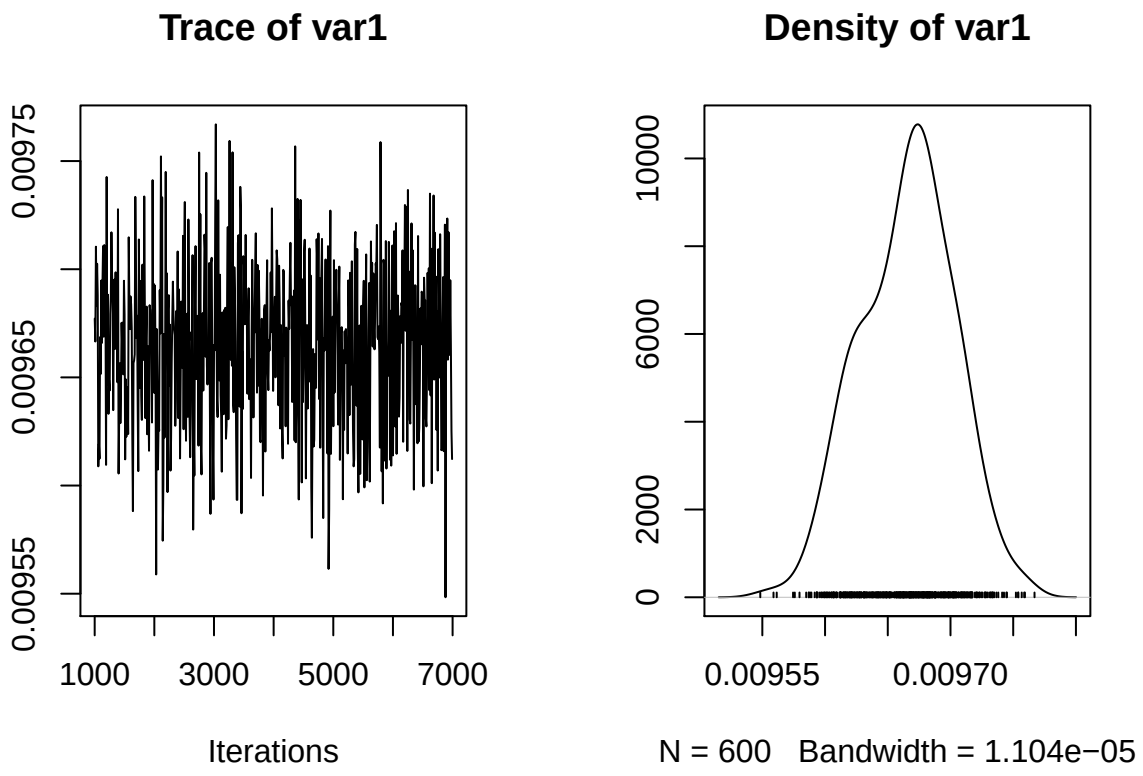

```
reg2 = MCMChregress(fixed=logSowRatio ~ History+ Balance + History:Promotion +
                      logIncome:Promotion, random=~Promotion, group="ConsumerID",
                    data=sow.data, r=2, R=diag(2), mcmc=6000)
```

```
##
## Running the Gibbs sampler. It may be long, keep cool :)
##
## **********:10.0%
## **********:20.0%
## **********:30.0%
## **********:40.0%
## **********:50.0%
## **********:60.0%
## **********:70.0%
## **********:80.0%
## **********:90.0%
## **********:100.0%
```

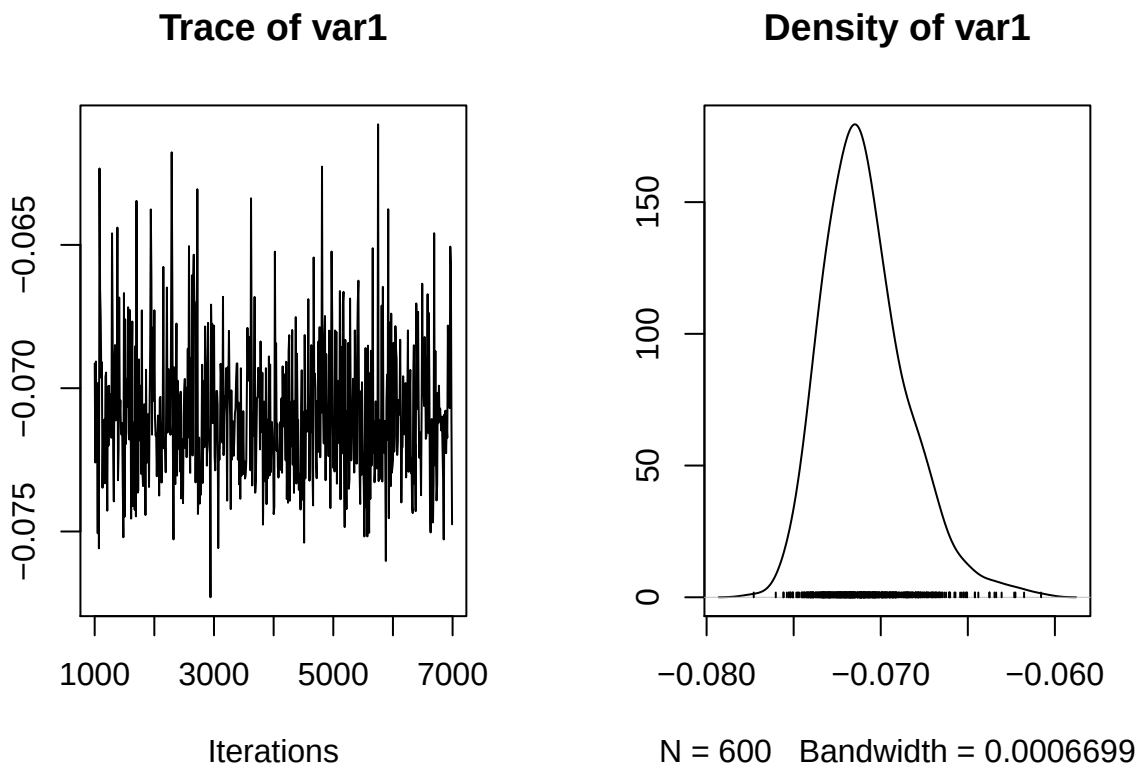
```
#summary(reg2$mcmc[,1:6])
```

All of the fixed effects are significant becase the 2.5% to 97.5% range for each effect does not include 0.

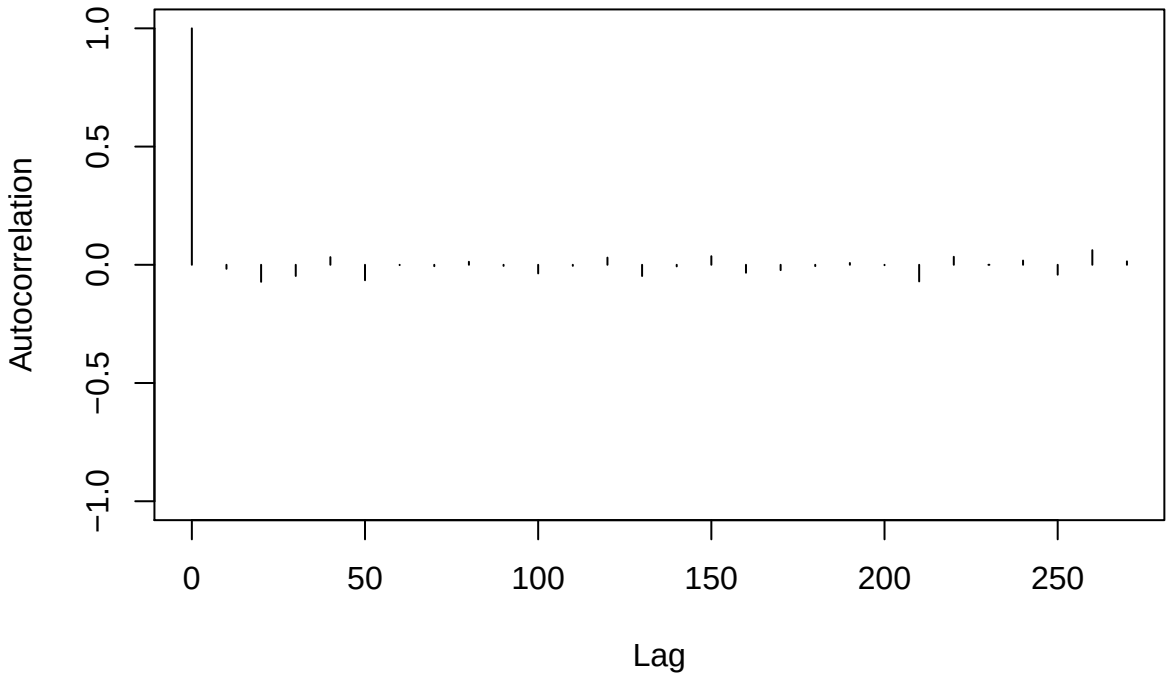
```
plot(reg2$mcmc[,"beta.History"])
```



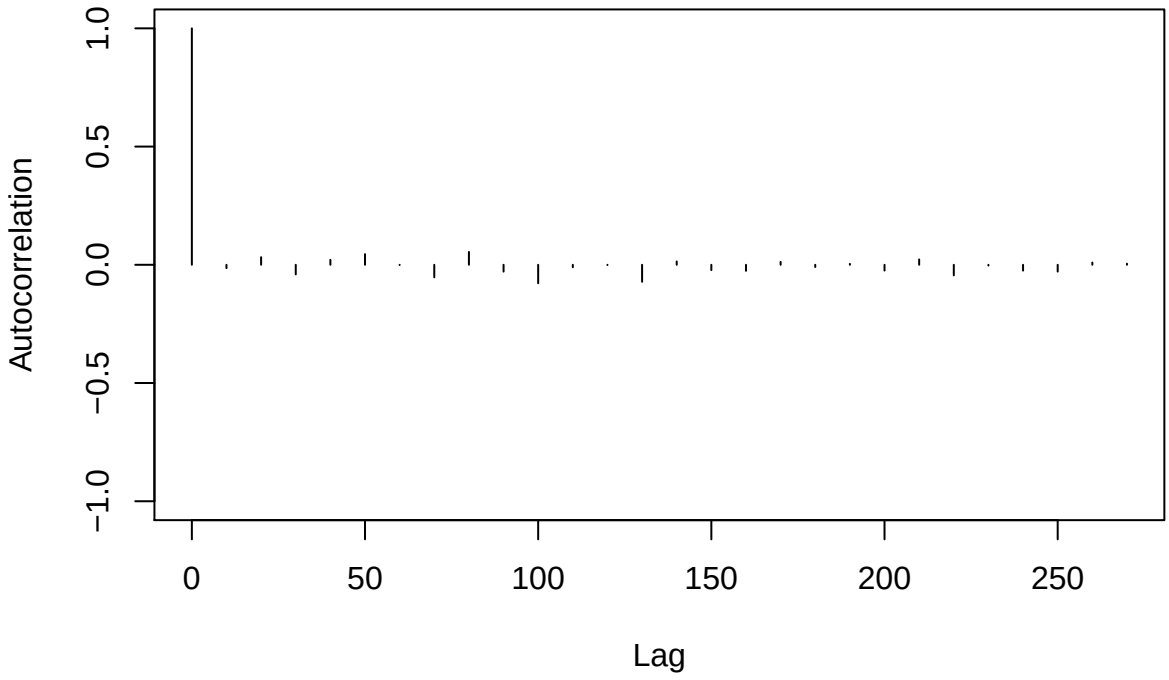

```
plot(reg2$mcmc[,"beta.Promotion:logIncome"])
```




```
autocorr.plot(reg2$mcmc[,"beta.History"])
```




```
autocorr.plot(reg2$mcmc[,"beta.Promotion:logIncome"])
```



The posterior densities can be viewed in above in the plots. The 95% interval from Question 1, is from -0.07845284 to -0.01372430 while the 95% interval from Question 2 is -0.03543907 to -0.02766906. From this, we can conclude that both are statistically significant because the ranges are all negative and do not cross the intercept.

Additionally, the greater range is in Question 1, so we can say that the model

Logistic Regression Models for Bank Customer Attrition

```
library(lme4)
```

```
## Loading required package: Matrix
```

```
library(readr)

bank <- read_csv("~/Bank_Retention_Data.csv")
```

```
## Parsed with column specification:
## cols(
##   Age = col_integer(),
##   Income = col_double(),
##   HomeVal = col_double(),
##   TractID = col_integer(),
##   Tenure = col_double(),
##   DirectDeposit = col_integer(),
##   Loan = col_integer(),
```

```
##   NumAccounts = col_integer(),
##   Dist = col_double(),
##   MktShare = col_double(),
##   Churn = col_integer()
## )
```

```
bank$TractID <- as.factor(bank$TractID)

reg3 = glm(Churn ~ Age+Income+HomeVal+Tenure+DirectDeposit+Loan+Dist+MktShare,
           data=bank, family=binomial(link="logit"))
summary(reg3)
```

```
##
## Call:
## glm(formula = Churn ~ Age + Income + HomeVal + Tenure + DirectDeposit +
##     Loan + Dist + MktShare, family = binomial(link = "logit"),
##     data = bank)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2054  -0.6823  -0.5328  -0.3401   2.6266
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -0.606224   0.296596  -2.044 0.040960 *
## Age           -0.016103   0.004150  -3.881 0.000104 ***
## Income         0.107067   0.015985   6.698 2.11e-11 ***
## HomeVal       -0.026059   0.005477  -4.758 1.95e-06 ***
## Tenure        -0.029709   0.006549  -4.536 5.73e-06 ***
## DirectDeposit -0.465836   0.110617  -4.211 2.54e-05 ***
## Loan           0.099376   0.124380   0.799 0.424310
## Dist           0.267618   0.061958   4.319 1.57e-05 ***
## MktShare      -0.082440   0.325551  -0.253 0.800089
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2355.9  on 2504  degrees of freedom
## Residual deviance: 2189.4  on 2496  degrees of freedom
## AIC: 2207.4
##
## Number of Fisher Scoring iterations: 5
```

The statistically significant coefficients here are the intercept, age, income, home value, tenure, direct deposit, loan, distance, and market share as they have p values less than 0.05. Beta6 and beta8, which are loan and market share respectively are not statistically significant because their p valies are greater than 0.05. Increasing each of the significan coeffcients value by one unit increases the probability of churn by eˆcoefficient. e.g. increasing age one unit decreases the outcome bye*-.606

```
AIC(reg3)
```

```
## [1] 2207.358
```

```
BIC(reg3)
```

```
## [1] 2259.793
```

```
reg4 = glmer(Churn ~ Age+Income+HomeVal+Tenure+DirectDeposit+Loan+Dist+MktShare +
               (1|TractID), data=bank, family=binomial(link="logit"))
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control
## $checkConv, : Model failed to converge with max|grad| = 0.00217238 (tol =
## 0.001, component 1)
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, : Model is nearly unide
##  - Rescale variables?
```

```
summary(reg4)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: Churn ~ Age + Income + HomeVal + Tenure + DirectDeposit + Loan +
##     Dist + MktShare + (1 | TractID)
##    Data: bank
##
##      AIC      BIC   logLik deviance df.resid
##   2208.7   2266.9  -1094.3   2188.7     2495
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.0912 -0.5118 -0.3895 -0.2447  5.3475
##
## Random effects:
##  Groups  Name        Variance Std.Dev.
##  TractID (Intercept) 0.01994  0.1412
## Number of obs: 2505, groups:  TractID, 26
##
## Fixed effects:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -0.564391   0.305963  -1.845   0.0651 .
## Age           -0.016479   0.004178  -3.944 8.00e-05 ***
## Income         0.107015   0.016078   6.656 2.81e-11 ***
## HomeVal       -0.026706   0.005691  -4.693 2.69e-06 ***
## Tenure        -0.029231   0.006564  -4.453 8.46e-06 ***
## DirectDeposit -0.461463   0.111004  -4.157 3.22e-05 ***
## Loan           0.099944   0.124635   0.802   0.4226
## Dist           0.266979   0.063386   4.212 2.53e-05 ***
## MktShare       0.007963   0.373360   0.021   0.9830
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) Age    Income HomeVl Tenure DrctDp Loan   Dist
## Age         -0.647
## Income      -0.221  0.055
## HomeVal     -0.206 -0.060 -0.534
## Tenure       0.014 -0.285 -0.075  0.077
## DirectDepst -0.175  0.012 -0.050  0.081 -0.115
## Loan        -0.073  0.073 -0.007 -0.059 -0.105 -0.083
## Dist        -0.324  0.000 -0.012 -0.150 -0.013 -0.008 -0.012
## MktShare    -0.359 -0.006 -0.031  0.060 -0.140  0.005 -0.008  0.260
```

```
## convergence code: 0
## Model failed to converge with max|grad| = 0.00217238 (tol = 0.001, component 1)
## Model is nearly unidentifiable: very large eigenvalue
##  - Rescale variables?
```

The age, income, homeval, tensure, direct deposit, and dist coefficients were statistically significant. Conversely, the loan and market share coefficients are not significant. This is the same conclusion reached by the model in the previous question.

```
AIC(reg3)
```

```
## [1] 2207.358
```

```
AIC(reg4)
```

```
## [1] 2208.686
```

```
BIC(reg3)
```

```
## [1] 2259.793
```

```
BIC(reg4)
```

```
## [1] 2266.947
```

The lower AIC and BIC are indicative of a superior model, and in this case, the AIC as well as BIC are lower in the first model. Therefore we should prefer that model as it is a better fit for the data.