marketing_homework1

Abhinaya

February 10, 2019

Random Effects and Hierarchical Linear Models

Business Problem

Random Effects and Hierarchical Linear Models

In this exercise, we will use hierarchical linear models and regressions with random effects for an analytics problem from a credit card company. The credit card company would like to figure out whether offering more promotions (for example, gasoline rebates and coupons for using the credit card) to their existing customers can increase the share-of-wallet of the credit card (that is, the share of a consumer's monthly spending using the credit card in her total spending). The company would also like to figure out what customer characteristics make them more responsive to promotions.

Question 1

1). Please read the data into R and create a data frame named "sow.data". Please convert consumer ID's to factors and create the following 2 variables in the data frame: logIncome = log(Income) and logSowRatio = log(WalletShare/(1-WalletShare)).

```
setwd("~/3_Spring Classes/Marketing Analytics/Assignment 1")
sow.data <- read.csv("CreditCard_SOW_data.csv")

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))

str(sow.data)
```

```
## 'data.frame':    3600 obs. of  8 variables:
##  $ ConsumerID : Factor w/ 300 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 1
1 ...
##  $ History    : int  55 55 55 55 55 55 55 55 55 55 ...
##  $ Income     : num  82000 82000 82000 82000 82000 82000 82000 82000 82000
82000 ...
##  $ WalletShare: num  0.643 0.628 0.567 0.638 0.554 0.573 0.666 0.649 0.527
0.459 ...
##  $ Promotion  : num  0.5 0.2 1 0.8 0.7 1.1 0.9 0.6 0.1 0 ...
##  $ Balance    : int  836 467 1208 792 1215 1248 197 567 1190 1709 ...
```

```
##  $ logIncome  : num  11.3 11.3 11.3 11.3 11.3 ...
##  $ logSowRatio: num  0.588 0.524 0.27 0.567 0.217 ...
```

Question 2

2). Use the function lm( ) to run the regression

Copy and paste the results here.

```
sow.lm <- lm(logSowRatio ~ History + Balance + Promotion + History:Promotion
+ logIncome:Promotion, sow.data)

summary(sow.lm)

##
## Call:
## lm(formula = logSowRatio ~ History + Balance + Promotion + History:Promoti
on +
##     logIncome:Promotion, data = sow.data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.59976 -0.14401  0.00153  0.13634  0.75883
##
## Coefficients:
##                       Estimate Std. Error  t value Pr(>|t|)
## (Intercept)          8.908e-02  1.603e-02    5.558 2.92e-08 ***
## History              1.039e-02  4.153e-04   25.027  < 2e-16 ***
## Balance             -4.959e-04  2.882e-06 -172.064  < 2e-16 ***
## Promotion            7.777e-01  1.888e-01    4.120 3.87e-05 ***
## History:Promotion   -2.598e-03  5.722e-04   -4.541 5.79e-06 ***
## Promotion:logIncome -4.558e-02  1.651e-02   -2.760  0.00581 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2078 on 3594 degrees of freedom
## Multiple R-squared:  0.8984, Adjusted R-squared:  0.8982
## F-statistic:  6353 on 5 and 3594 DF,  p-value: < 2.2e-16
```

Question 3

Estimate the following hierarchical linear model using the function lmer( ) in the R package "lme4"

$$logSowRatio_{ij} = \beta_{0i} + \beta_1 \times Balance_{ij} + \beta_{2i} \times Promotion_{ij} + \varepsilon_{ij}$$

$$\beta_{0i} = \mu_0 + \mu_1 \times History_i + \zeta_i$$

$$\beta_{2i} = \gamma_0 + \gamma_1 \times History_i + \gamma_2 \times logIncome_i + \xi_i$$

Following what we did in our class, please rewrite this hierarchical linear model as a one-level linear regression model with random effects.

$$logSowRatio_{ij} = \beta_{0i} + \beta_1 \times Balance_{ij} + \beta_{2i} \times Promotion_{ij} + \varepsilon_{ij}$$

$$\beta_{0i} = \mu_0 + \mu_1 \times History_i + \zeta_i$$

$$\beta_{2i} = \gamma_0 + \gamma_1 \times History_i + \gamma_2 \times logIncome_i + \xi_i$$

This implies,

$$logSowRatio_{ij} = (\mu_0 + \mu_1 \times History_i + \zeta_i) + \beta_1 \times Balance_{ij} + (\gamma_0 + \gamma_1 \times History_i + \gamma_2 \times logIncome_i + \xi_i) \times Promotion_{ij} + \varepsilon_{ij}$$

This implies,

$$logSowRatio_{ij} = (\mu_0 + \zeta_1) + \mu_1 \times History_i + (\gamma_0 + \xi_1) \times Promotion_{ij} + \gamma_1 \times History_i \times Promotion_{ij} + \gamma_2 \times logIncome_i \times Promotion_{ij} + \beta_1 \times Balance_{ij} + \varepsilon_{ij}$$

Which variables (and interactions) in the regression have fixed effects? Which ones have random effects?

Fixed effects: 1+ History + Promotion + History:Promotion + logIncome:Promotion + Balance

Random effects: 1 + Promotion

Specify the variables in lmer() and run the regression. Please copy and paste the summary() of the regression.

```
library(lme4)

## Warning: package 'lme4' was built under R version 3.5.2

## Loading required package: Matrix

sow.lmer = lmer(logSowRatio ~ History + Balance + Promotion + History:Promoti
on + logIncome:Promotion + (1+ Promotion|ConsumerID), data=sow.data, REML=F)
```

```
## Warning: Some predictor variables are on very different scales: consider
## rescaling
```

```r
sow.lmer = lmer(logSowRatio ~ History + Balance + Promotion + History:Promoti
on + logIncome:Promotion + (1+ Promotion|ConsumerID), data=sow.data, REML=T,
control=lmerControl(optimizer="Nelder_Mead"))
```

```
## Warning: Some predictor variables are on very different scales: consider
## rescaling
```

```r
summary(sow.lmer)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula:
## logSowRatio ~ History + Balance + Promotion + History:Promotion +
##     logIncome:Promotion + (1 + Promotion | ConsumerID)
##    Data: sow.data
## Control: lmerControl(optimizer = "Nelder_Mead")
##
## REML criterion at convergence: -6476.6
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.1051 -0.6423  0.0052  0.6335  3.4460
##
## Random effects:
##  Groups     Name        Variance  Std.Dev. Corr
##  ConsumerID (Intercept) 0.0361995 0.19026
##             Promotion   0.0005776 0.02403  0.05
##  Residual               0.0066091 0.08130
## Number of obs: 3600, groups:  ConsumerID, 300
##
## Fixed effects:
##                       Estimate Std. Error  t value
## (Intercept)          0.0959627  0.0266431    3.602
## History              0.0103948  0.0007159   14.520
## Balance             -0.0005003  0.0000018 -277.993
## Promotion            0.6128625  0.1473436    4.159
## History:Promotion   -0.0025708  0.0002414  -10.649
## Promotion:logIncome -0.0310973  0.0129495   -2.401
##
## Correlation of Fixed Effects:
##             (Intr) Histry Balanc Promtn Hstr:P
## History     -0.900
## Balance     -0.107 -0.001
## Promotion   -0.011  0.009  0.013
## Hstry:Prmtn  0.143 -0.159 -0.002 -0.153
## Prmtn:lgInc  0.001  0.000 -0.012 -0.998  0.099
## fit warnings:
## Some predictor variables are on very different scales: consider rescaling
```

Interpret the estimated fixed effects in the regression.

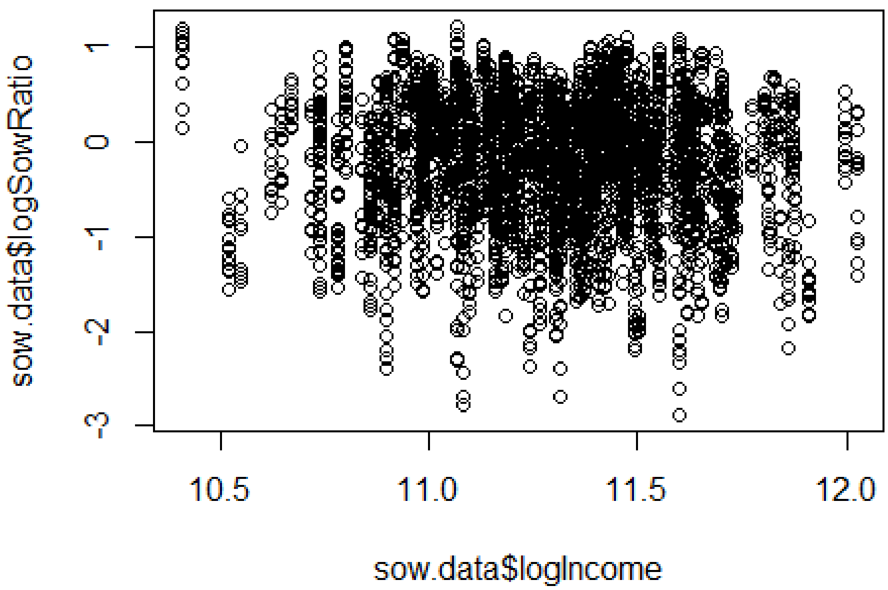
```
fixef(sow.lmer)

##        (Intercept)             History             Balance
##       0.0959627161        0.0103947609       -0.0005002994
##          Promotion   History:Promotion Promotion:logIncome
##       0.6128624920       -0.0025708298       -0.0310972585

head(coef(sow.lmer)$ConsumerID,5)

##   (Intercept)    History       Balance Promotion History:Promotion
## 1   0.1793437 0.01039476 -0.0005002994 0.6211239       -0.00257083
## 2   0.5266302 0.01039476 -0.0005002994 0.6392806       -0.00257083
## 3   0.2854206 0.01039476 -0.0005002994 0.6288758       -0.00257083
## 4   0.1389467 0.01039476 -0.0005002994 0.6020174       -0.00257083
## 5  -0.0685256 0.01039476 -0.0005002994 0.6197353       -0.00257083
##   Promotion:logIncome
## 1         -0.03109726
## 2         -0.03109726
## 3         -0.03109726
## 4         -0.03109726
## 5         -0.03109726
```

Promotion has the highest influence on Share of Wallet. 1 unit increase in index of promotions per month, can lead to customer spending 1.84 (exp(0.61)) more provided all the other factors remain the same

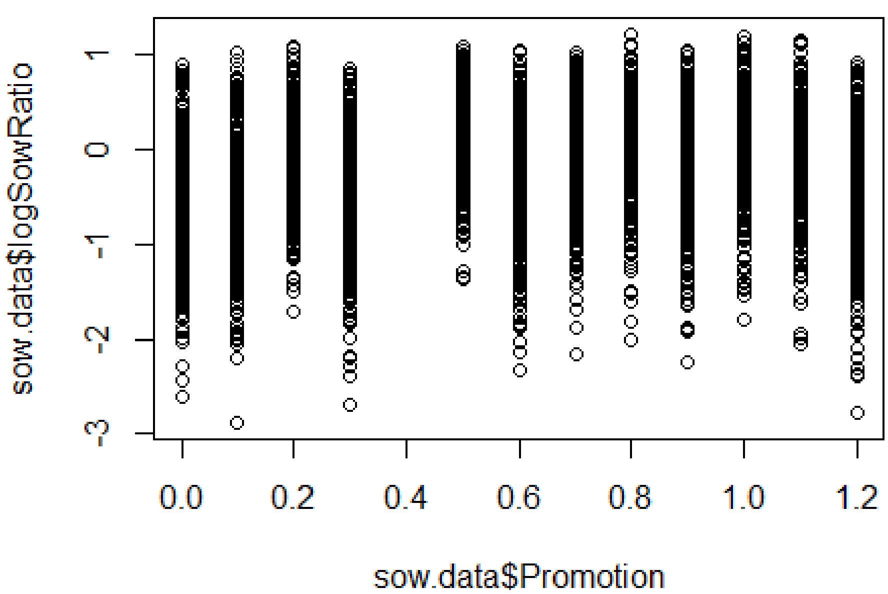
It is interesting to see that the interaction of History with Promotions has a negative coefficient, but History by itself has a positive coefficient and Promotion by itself also has a positive coefficient. It probably means that someone who is loyal to the store and shops there often is likely to spend more, but he/she also doesn't get influenced much by the promotions in the store.

Another interesting observation is the negative coefficient for interaction between Promotion and logIncome. Plotting the variables

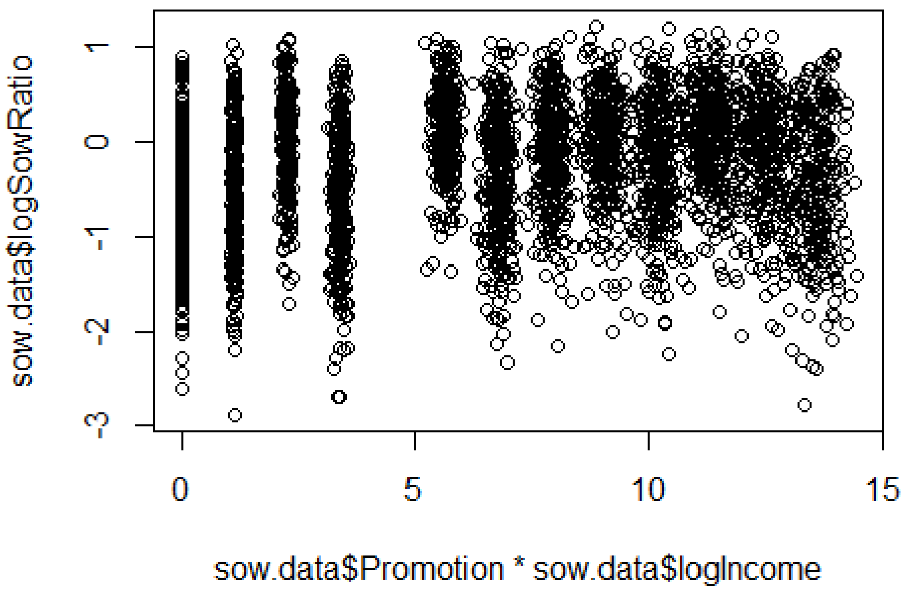
```
plot(sow.data$logIncome,sow.data$logSowRatio)
```




```
plot(sow.data$Promotion,sow.data$logSowRatio)
```




```
plot(sow.data$Promotion*sow.data$logIncome,sow.data$logSowRatio)
```



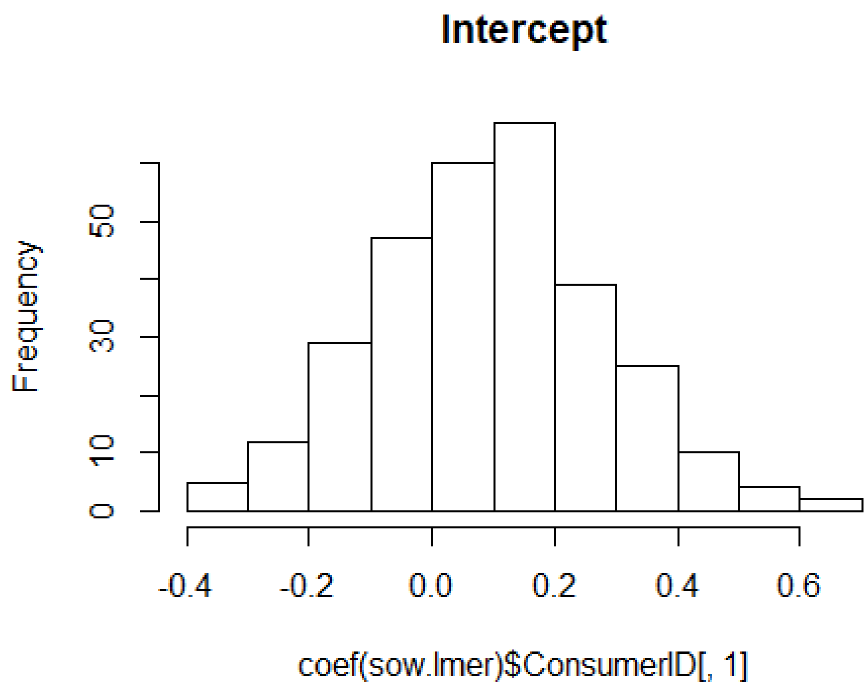
Comment: From the graph we can observe that share of waller is denser as logIncome increases.

Please plot the histograms for the random effects in the linear mixed effect model.

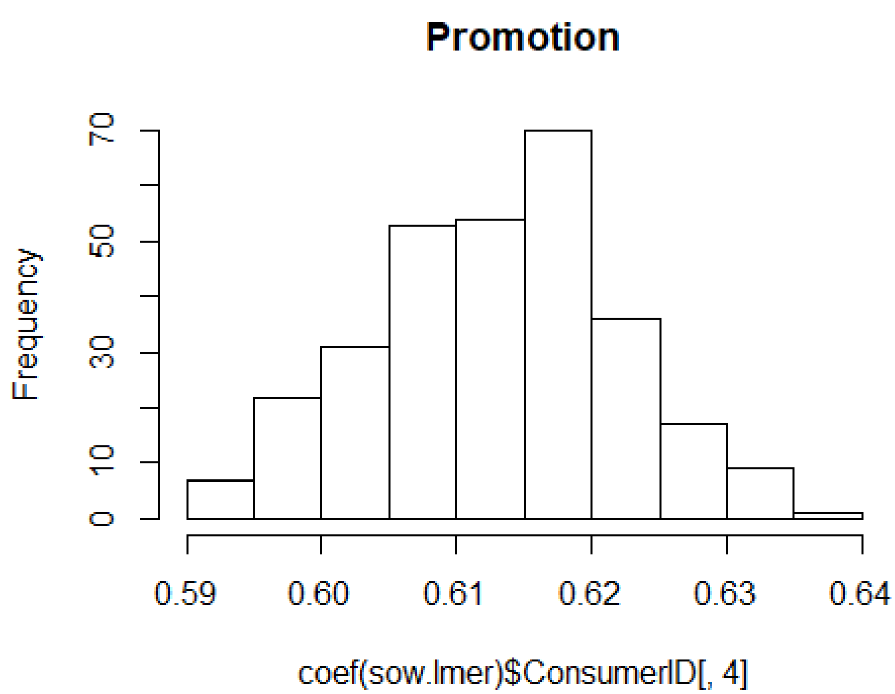
```
#checking coefficients of random effects
head(ranef(sow.lmer)$ConsumerID, 5)
```

```
##   (Intercept)    Promotion
## 1  0.08338102  0.008261421
## 2  0.43066744  0.026418146
## 3  0.18945789  0.016013326
## 4  0.04298397 -0.010845090
## 5 -0.16448832  0.006872832
```

```
#checking the assumption if random effects are normally distributed
hist(coef(sow.lmer)$ConsumerID[,1], main="Intercept")
```

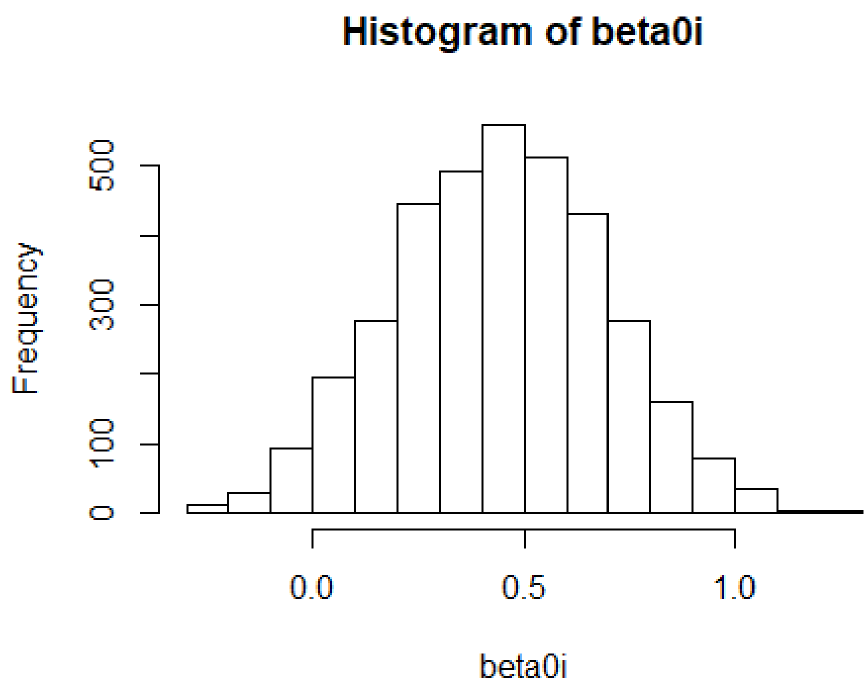



```
hist(coef(sow.lmer)$ConsumerID[,4], main="Promotion")
```

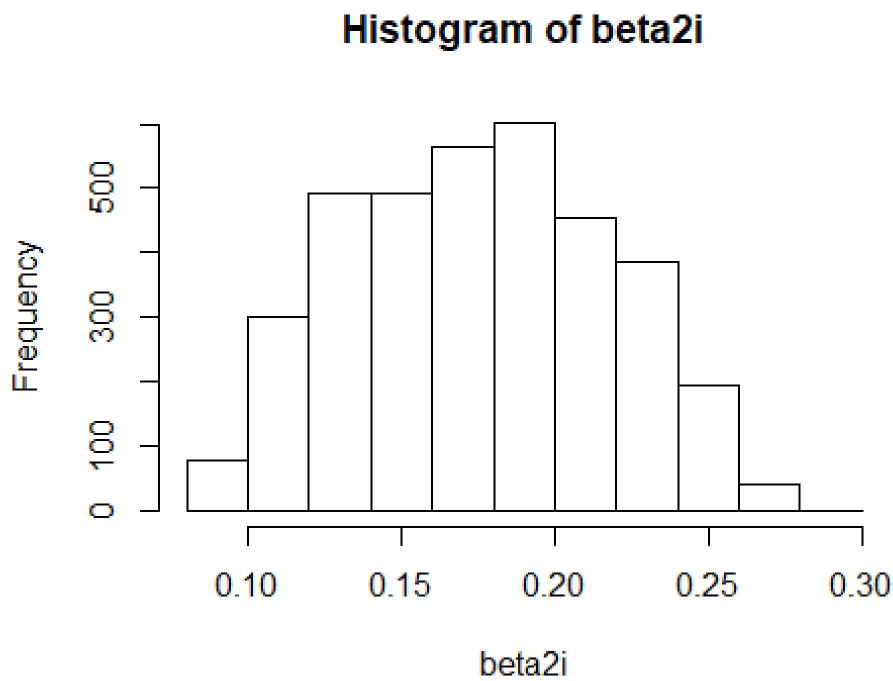


Using the estimated random effect, calculate beta0i and beta2i and plot their histograms.

```
# beta0i = intercept of fixed effects + beta for history * history + intercep
t for random effect
beta0i = fixef(sow.lmer)[1] + fixef(sow.lmer)[2] * sow.data$History + ranef(s
ow.lmer)$ConsumerID[,1]
hist(beta0i)
```




```
# beta2i = beta for promotion * promotion + beta for interaction of history a
nd promotion * history * promotion + beta for interaction of logincome and pr
omotion * logincome * promotion + promotion value for random effects
beta2i = fixef(sow.lmer)[4]+ fixef(sow.lmer)[5] * sow.data$History + fixef(so
w.lmer)[6] * sow.data$logIncome + ranef(sow.lmer)$ConsumerID[,2]
hist(beta2i)
```



Compare model fit using AIC() and BIC() with the model in (2).

```
AIC(sow.lmer)

## [1] -6456.623

AIC(sow.lm)

## [1] -1087.389

BIC(sow.lmer)

## [1] -6394.736

BIC(sow.lm)

## [1] -1044.069
```

AIC and BIC of linear mixed models are much lower than linear model. This shows that linear mixed models outperform linear models for this scenario

Linear and Hierarchical Linear Models: Bayesian Estimation

Question 4

Use the function MCMCregress() in the R package “MCMCpack” to estimate the linear regression

```
library(MCMCpack)

## Warning: package 'MCMCpack' was built under R version 3.5.2

## Loading required package: coda

## Warning: package 'coda' was built under R version 3.5.2

## Loading required package: MASS

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

## ## Copyright (C) 2003-2019 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)
## ##

sow.ba1 = MCMCregress(logSowRatio ~ History + Balance + Promotion + History:P
romotion + logIncome:Promotion,mcmc=6000, data=sow.data)
```

Use the summary() function to find the results of the estimation. Copy and pastes the results here.

```
summary(sow.ba1)

##
## 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
```

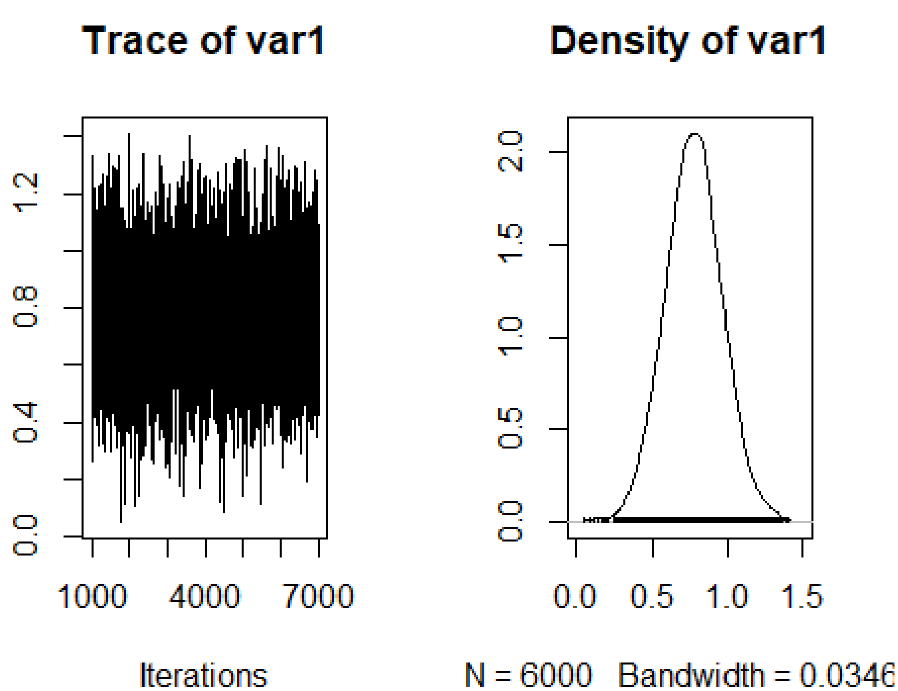
```
## sigma2                 0.0432230 1.022e-03 1.319e-05      1.405e-05
##
## 2. Quantiles for each variable:
##
##                              2.5%        25%        50%       75%      97.5%
## (Intercept)            0.0574844  0.0775755  0.0886492  0.099764  0.1194332
## History                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
```

From the Bayesian posterior intervals (use 2.5% and 97.5% quantiles of the simulated posterior distributions), are regression coefficients significant at the 5% level?

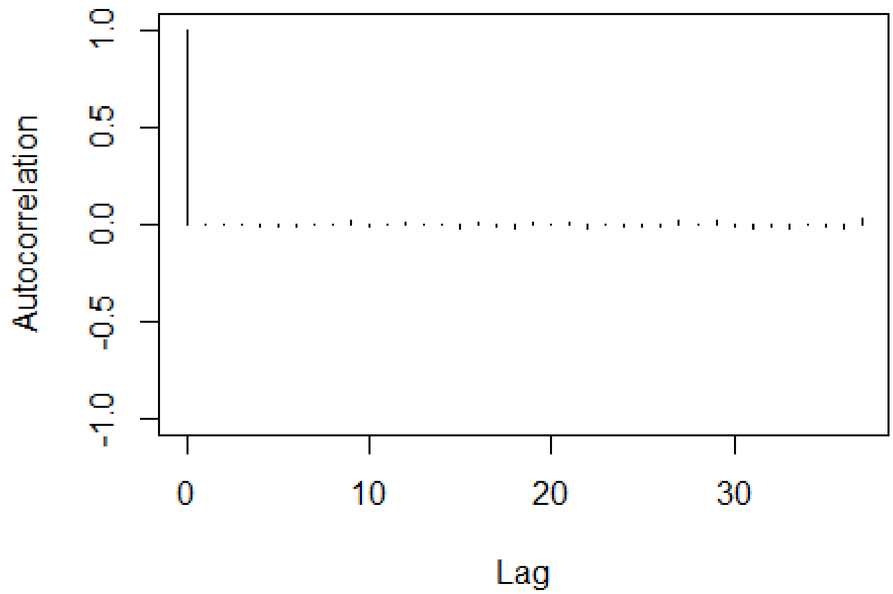
Based on the 2.5% and 97.5% quantiles, we can see that History, Balance, Promotion, interaction of History with Promotion and interaction of Promotion with logIncome are all significant

Use the plot() function to plot the posterior sampling chains and posterior densities (estimated by kernel methods automatically by R) for beta3 and beta5 copy and paste the results here. Use the autocorr.plot() function to plot the autocorrelation of the posterior sampling chains for beta3 and beta5; copy and paste the results here.

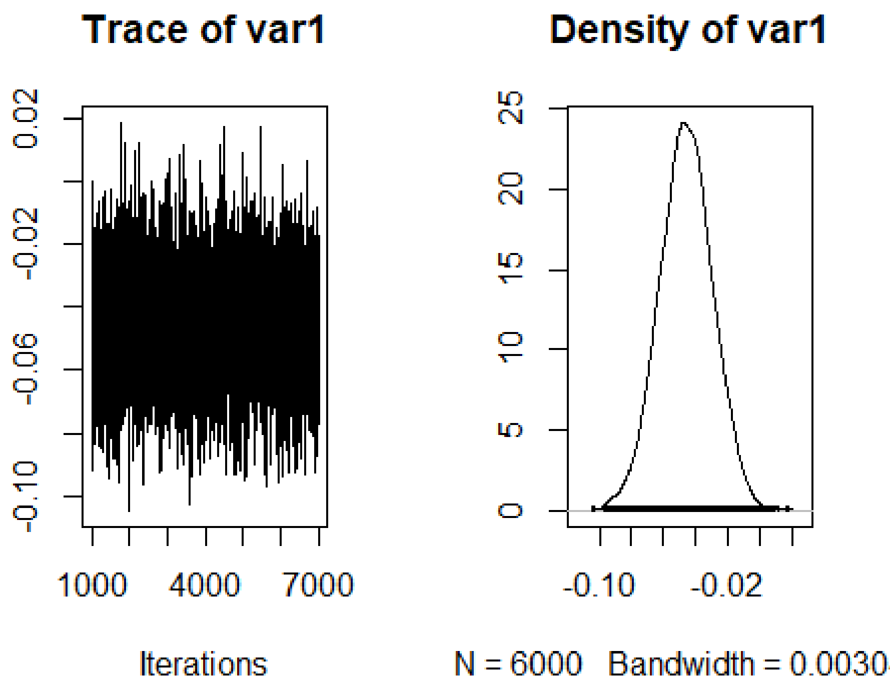
```
#beta3 = promotion
plot(sow.ba1[,4])
```



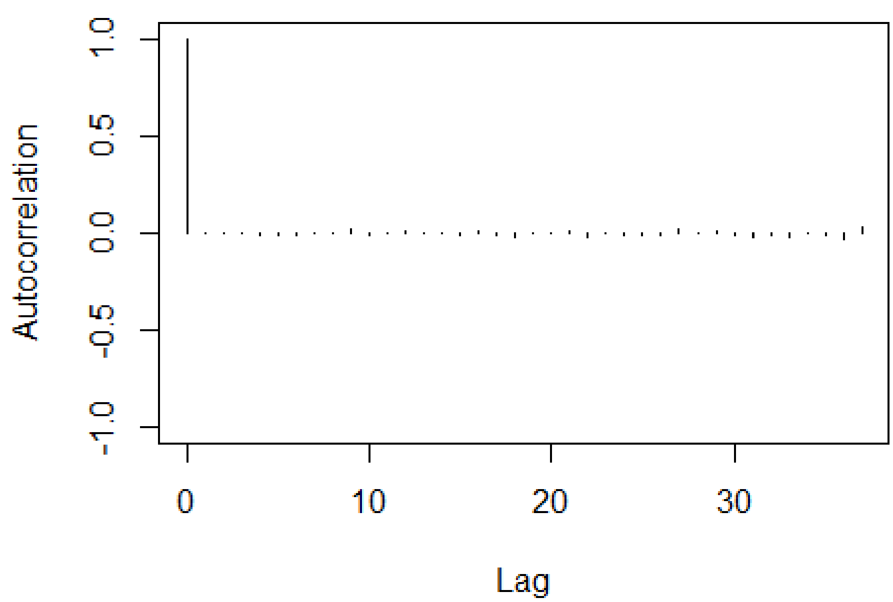
```r
autocorr.plot(sow.ba1[,4])
```




```r
#beta5 = logIncome * promotion
plot(sow.ba1[,6])
```




```r
autocorr.plot(sow.ba1[,6])
```



```r
#traceplot(m1[,"price"], main="Price Coefficient")

#densplot(m1[,"price"], main="Price Coefficient")

apply(sow.ba1, 2, quantile, probs=c(0.025, 0.5, 0.975))
```

```
##       (Intercept)     History       Balance Promotion History:Promotion
## 2.5%   0.05748437 0.009608261 -0.0005015492 0.4146629      -0.003737697
## 50%    0.08864919 0.010397351 -0.0004959350 0.7803401      -0.002592536
## 97.5%  0.11943319 0.011194338 -0.0004901459 1.1558051      -0.001494478
##       Promotion:logIncome     sigma2
## 2.5%          -0.07845284 0.04126405
## 50%           -0.04580048 0.04320888
## 97.5%         -0.01372430 0.04526569
```

Question 5

For the hierarchical linear model use the function MCMChregress( ) in the R package "MCMCpack" for its Bayesian estimation

```r
#library(MCMC)
sow.ba2 = MCMChregress(fixed=logSowRatio ~ History + Balance + Promotion + Hi
story:Promotion + logIncome:Promotion, random=~Promotion, group="ConsumerID",
data=sow.data, r=2,R=diag(2))
```

```
##
## 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%
```

Please copy and paste the Bayesian estimation results of the fixed effects (same fixed effects as in (3)) in the model using summary("yourBayesianModelName"$mcmc[,1:6]). From the Bayesian posterior intervals, are the fixed effects significant at the 5% level?

```r
summary(sow.ba2$mcmc[,1:6])
```

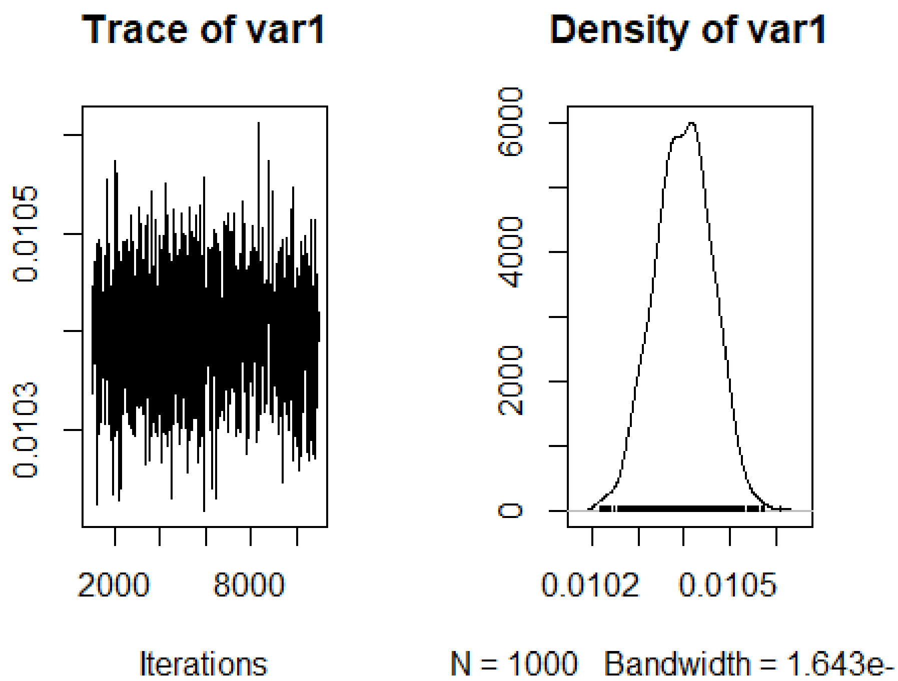
```
##
## Iterations = 1001:10991
## Thinning interval = 10
## Number of chains = 1
## Sample size per chain = 1000
##
## 1. Empirical mean and standard deviation for each variable,
```

```
##    plus standard error of the mean:
##
##                                 Mean        SD  Naive SE Time-series SE
## beta.(Intercept)           0.0965326 2.361e-03 7.467e-05      7.813e-05
## beta.History               0.0103994 6.172e-05 1.952e-06      2.059e-06
## beta.Balance              -0.0005008 1.483e-07 4.690e-09      4.690e-09
## beta.Promotion             0.6167817 2.373e-02 7.504e-04      8.401e-04
## beta.History:Promotion    -0.0025734 3.709e-05 1.173e-06      1.173e-06
## beta.Promotion:logIncome -0.0314372 2.083e-03 6.588e-05      7.353e-05
##
## 2. Quantiles for each variable:
##
##                                 2.5%        25%        50%        75%
## beta.(Intercept)           0.0918748  0.0949810  0.0964767  0.0980893
## beta.History               0.0102840  0.0103574  0.0104013  0.0104417
## beta.Balance              -0.0005011 -0.0005009 -0.0005008 -0.0005007
## beta.Promotion             0.5714261  0.6011425  0.6165593  0.6333237
## beta.History:Promotion    -0.0026485 -0.0025992 -0.0025729 -0.0025489
## beta.Promotion:logIncome -0.0354763 -0.0328847 -0.0314427 -0.0300743
##                                97.5%
## beta.(Intercept)           0.1015093
## beta.History               0.0105167
## beta.Balance              -0.0005005
## beta.Promotion             0.6628724
## beta.History:Promotion    -0.0025018
## beta.Promotion:logIncome -0.0274620
```

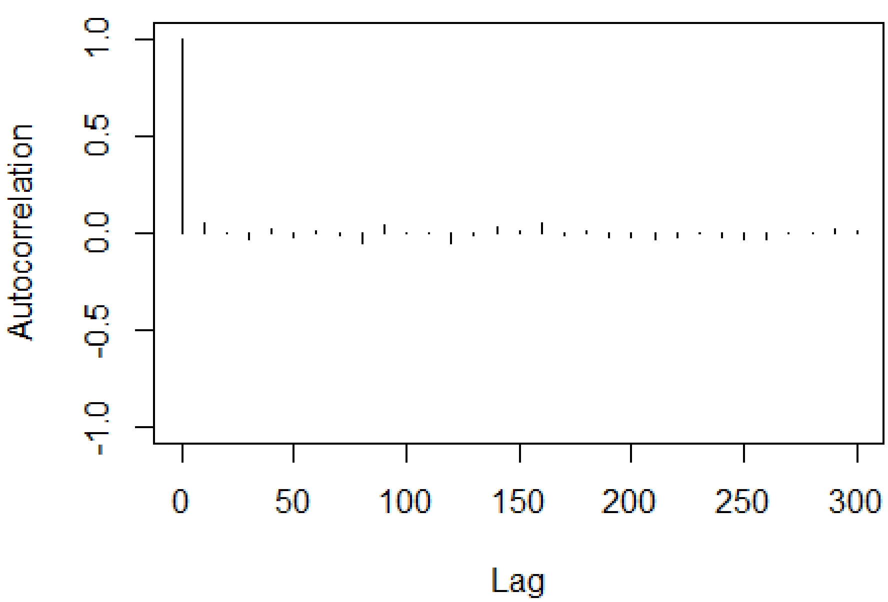
All are significant

Use the plot() function to plot the posterior sampling chains and posterior densities for mu1 and gamma2; copy and paste the results here. Use the autocorr.plot() function to plot the autocorrelation of the posterior sampling chains for mu1 and gamma2; copy and paste the results here.

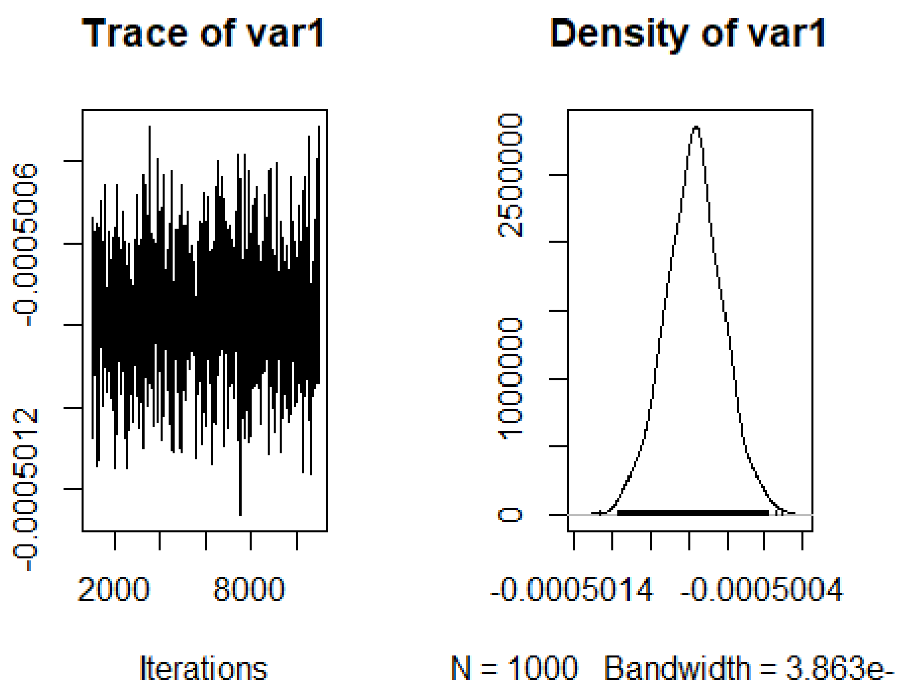
```r
#mu1 = coefficient of fixed effect history
plot(sow.ba2$mcmc[,2])
```



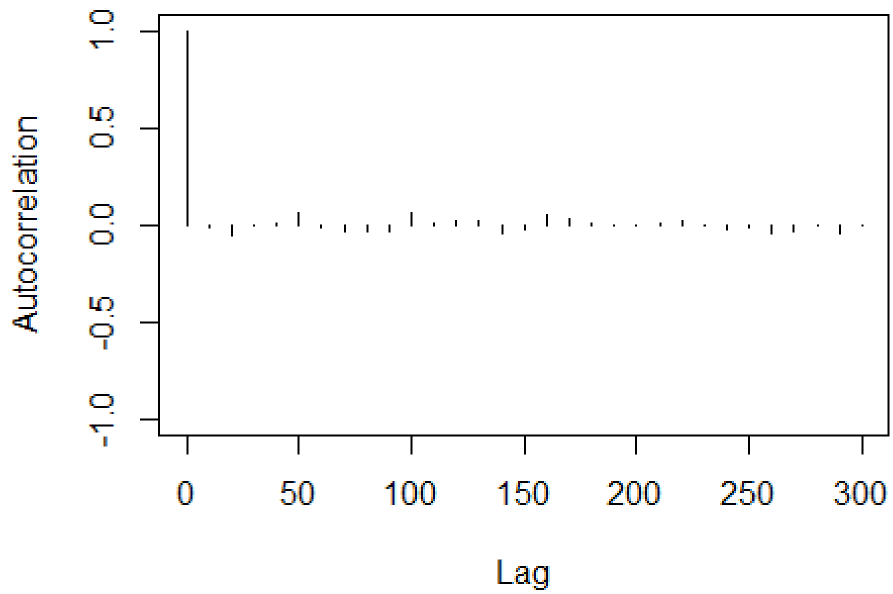

```r
autocorr.plot(sow.ba2$mcmc[,2])
```



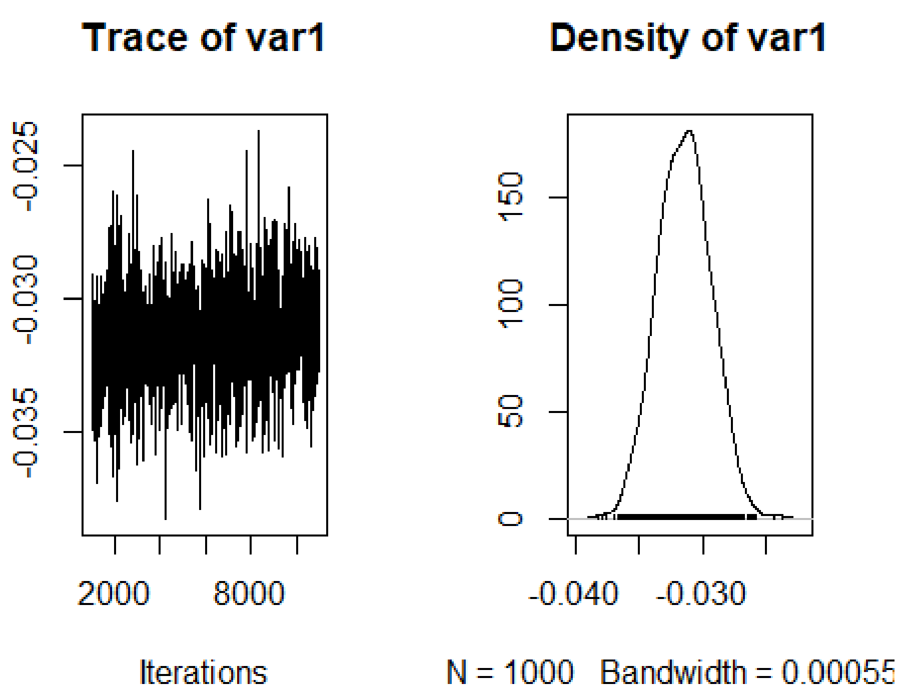
```r
#beta1 = coefficient of balance
plot(sow.ba2$mcmc[,3])
```



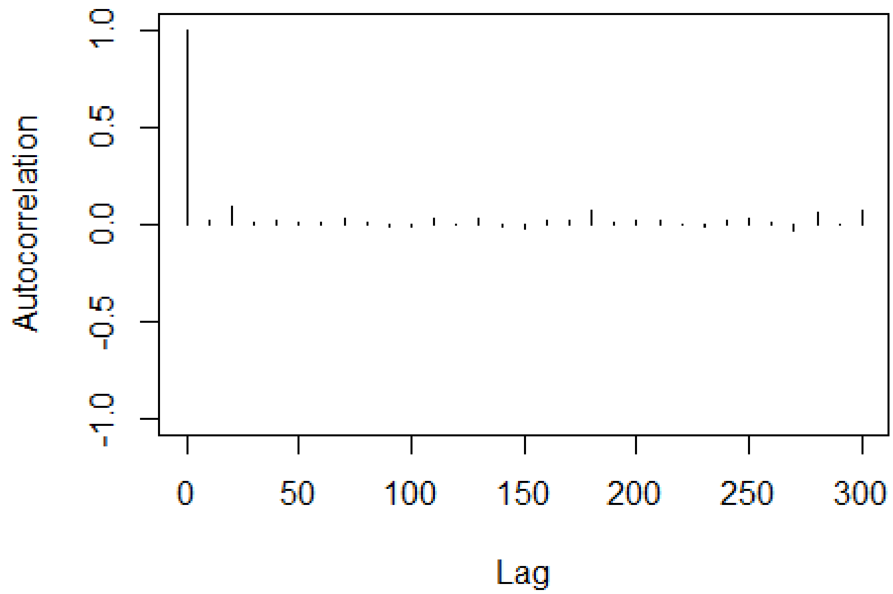

```r
autocorr.plot(sow.ba2$mcmc[,3])
```




```
#gamma2 = coefficient of interaction of logIncome and Promotion
plot(sow.ba2$mcmc[,6])
```

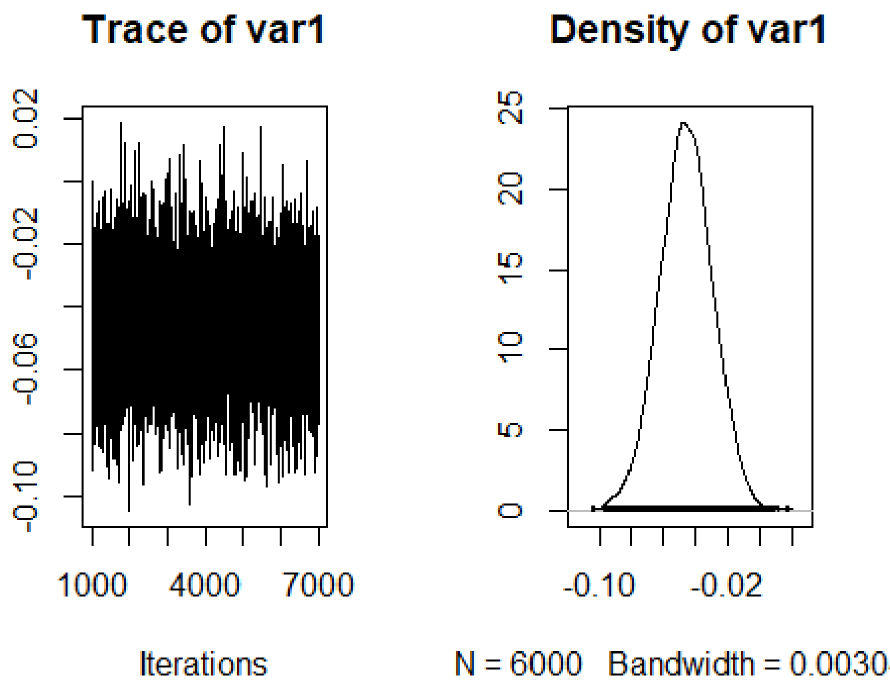


```r
autocorr.plot(sow.ba2$mcmc[,6])
```

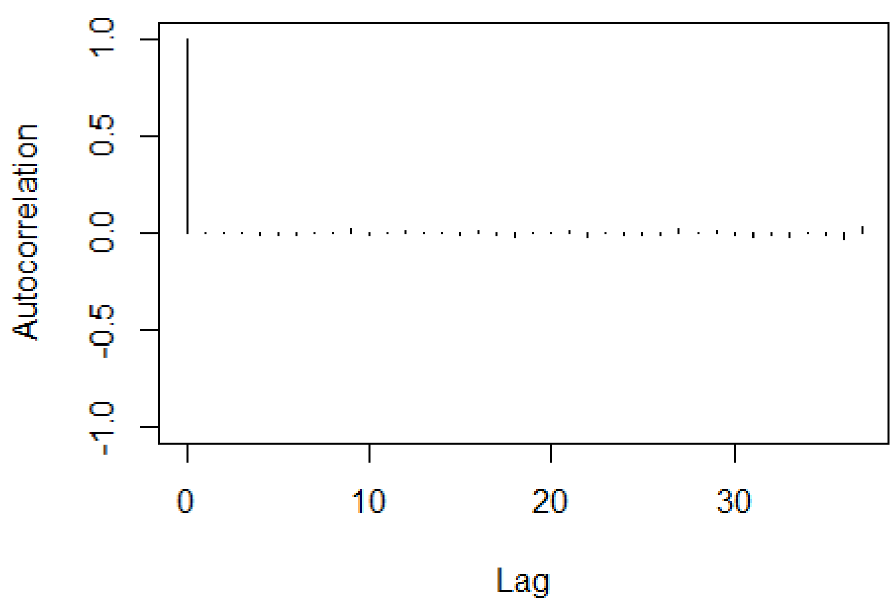


Compare the posterior densities and 95% intervals of gamma2 and beta5 in Question (4). Do the intervals include zero?

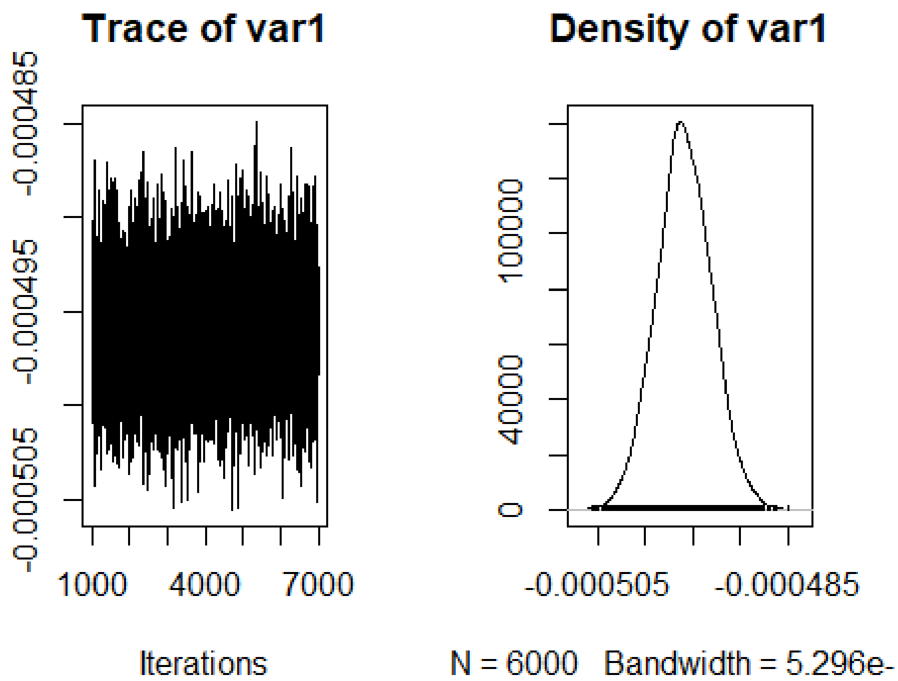
```r
#Comparing with beta5 and beta1 of linear model
#beta5 = coefficient of interaction of History and Promotion
plot(sow.ba1[,6])
```



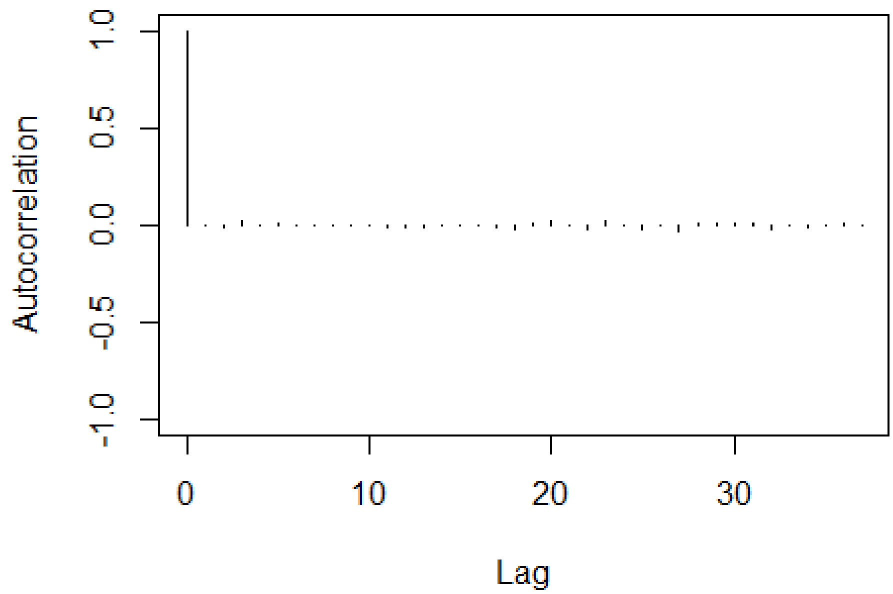

```
autocorr.plot(sow.ba1[,6])
```



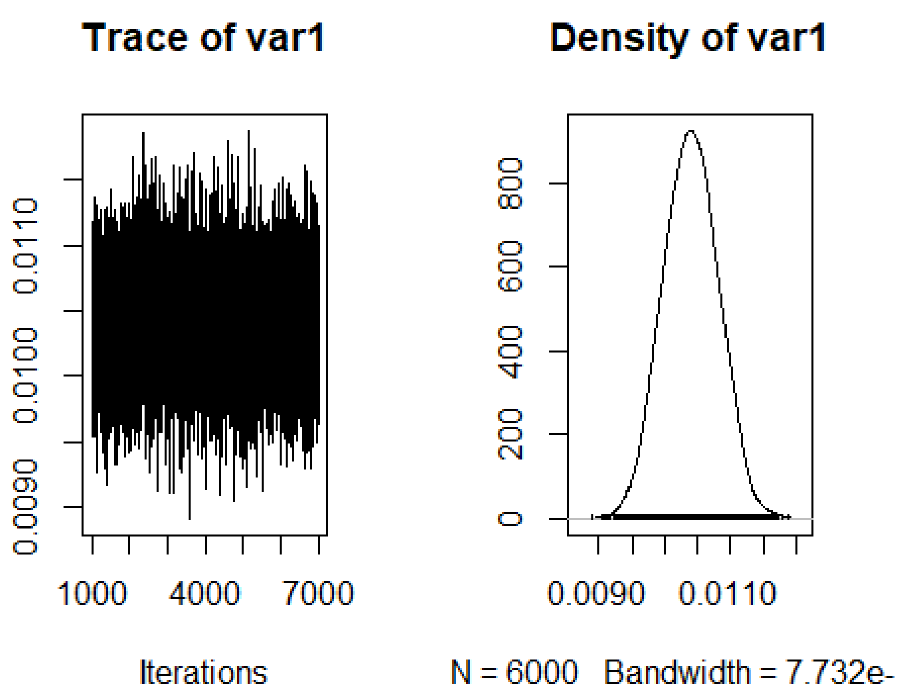
```r
#beta2 = coefficient of balance
plot(sow.ba1[,3])
```



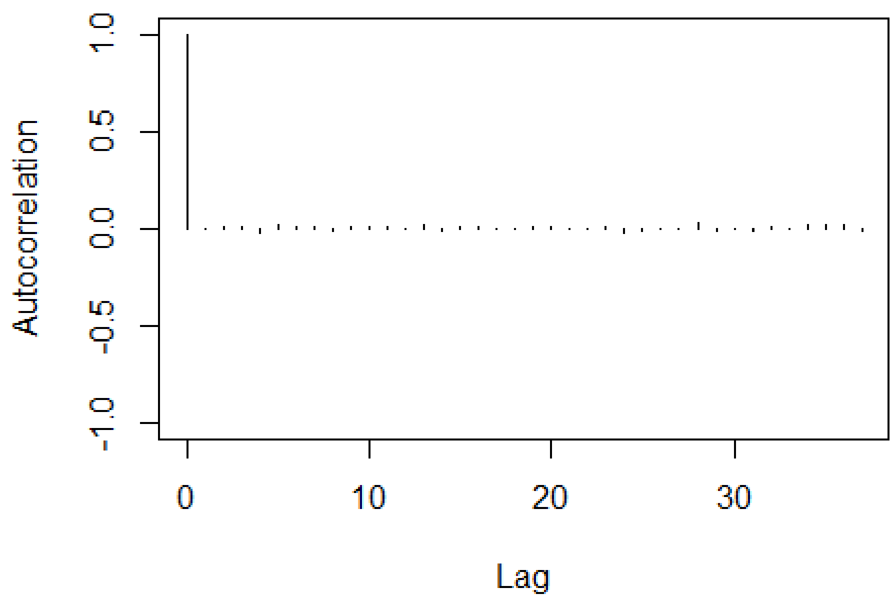

```r
autocorr.plot(sow.ba1[,3])
```




```
#beta1 = coefficient of history
plot(sow.ba1[,2])
```



```
autocorr.plot(sow.ba1[,2])
```



Comment on the differences of their estimation results.

We observe that interaction of History with Promotion is significant in the hierarchical model, but that is not the case in the Linear model. The other variables are significant in both models.

By accounting for Promotion as random effects, we are able to better estimate the model and identify the pattern in repeat purchases data.