**OM 386 Assignment 2**

**Due: March 3rd, 11:59pm**

**By: Callie Gilmore (cgg756)**

**Binary Data Regression Models for Bank Customer Attrition**

This exercise is similar to the bank customer acquisition problem that we discussed in our class. Imagine that you are hired as a consultant. For the analysis, the management has given you access to 2505 customers, among whom 449 (about 18%) have closed their accounts within one year. As a consultant, you would like to know what demographic and behavioral variables contribute to higher attrition/churn rates among these customers.

The data file is "Bank\_Retention\_Data.csv" on Canvas. It has the following variables:

|  |  |
| --- | --- |
| Age | The customer’s age |
| Income | The customer’s income |
| HomeVal | The customer’s home value |
| TractID | A label/ID of the census tract of the customer’s residence |
| Tenure | How long this person has been a customer of the bank |
| DirectDeposit | Indicator dummy=1 if the customer uses direct deposit and 0 otherwise |
| LoanInd | Loan indicator dummy = 1 if the customer has ever taken loans from her bank and 0 if not |
| Dist | Distance from customer’s home to the nearest bank branch |
| MktShare | Bank’s market share in the customer’s market |
| Churn | Indicator dummy = 1 if the customer has closed her/his accounts (s/he has churned) with the bank and 0 if not |

1). Read the data into R. Convert TractID into a factor variable. (15 points)

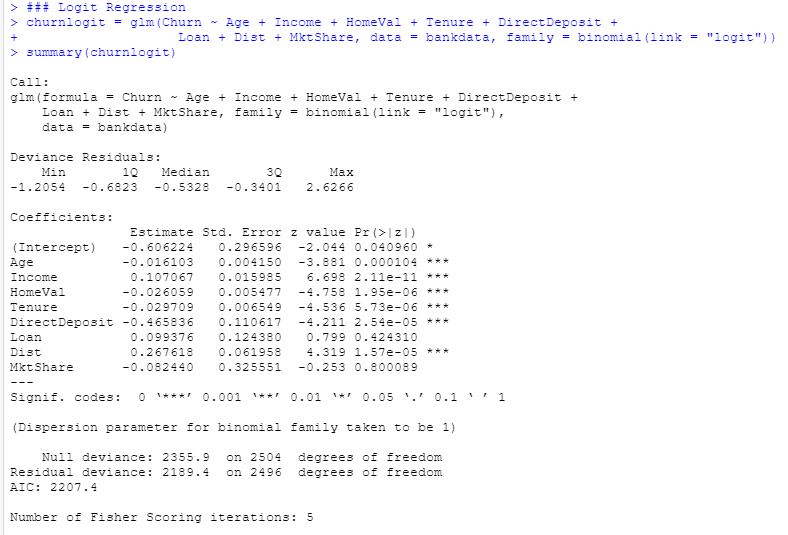


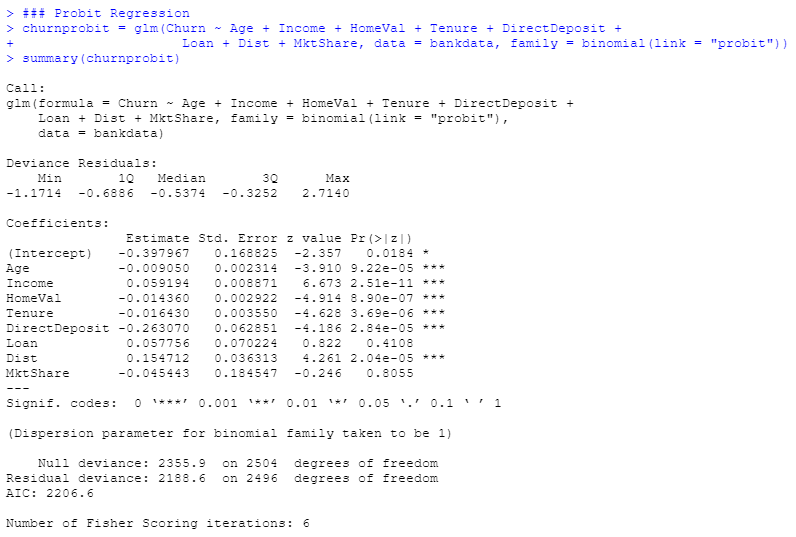
Estimate the following binary data regression model using the R function glm( ).

*Churni* ~ *β*0 + *β*1×*Agei* + *β*2×*Incomei+*β3×*HomeVali* + β4×*Tenurei*

*+β*5×*DirectDepositi* + *β*6×*LoanIndi +*β7×*Disti* + β8×*MktSharei*

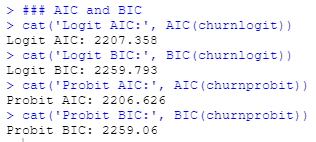
Use both of the logit (for logistic regression) and probit (for probit regression) link functions of the binomial family and paste results here.





How do you interpret *β*1, *β*2*, β*3, *β*4, *β*5, *β*6, *β*7, *β*8? Are they statistically significant in the logistic and probit models? Please also calculate the AIC and BIC of the logistic and probit models using the R functions AIC( ) and BIC( ). Which model (logistic or probit) fits the data better based on AIC and BIC?

From the results above, the following variables are statistically significant as their p-values are lower than 0.05: Age, Income, HomeVal, Tenure, DirectDeposit and Dist.



Based on results above, the Probit Model is better because its AIC and BIC are lower; however, it is important to note that they are not much lower so the Probit Model barely outperformed Logit Model.

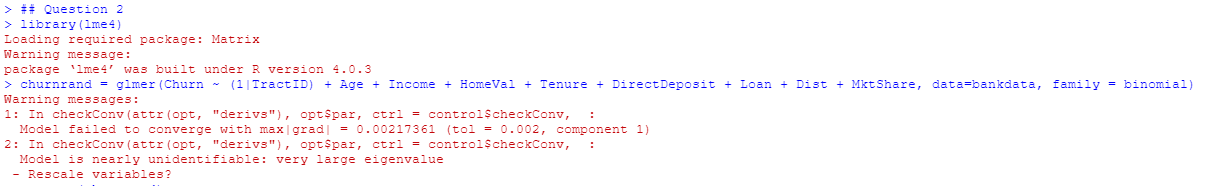
2). Next we will use a random effect grouped by TractID in the logistic regression.

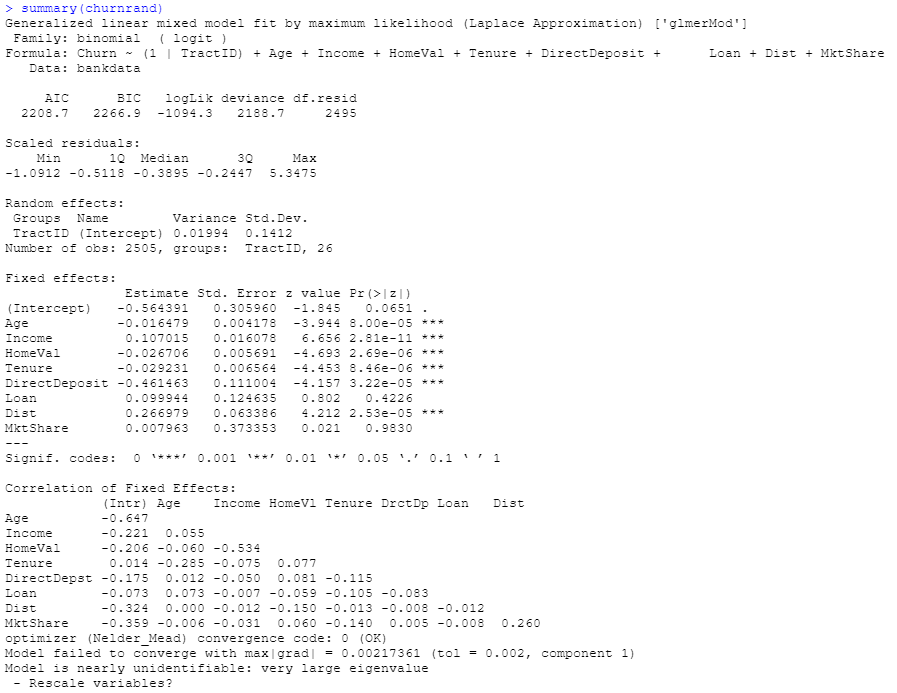
Use the function glmer( ) in the "lme4" package in R to fit

*Churni* ~ *β*0p + *β*1×*Agei* + *β*2×*Incomei +β*3×*HomeVali* + *β*4×*Tenurei*

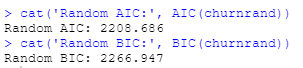
*+β*5×*DirectDepositi* + *β*6×*LoanIndi +β*7×*Disti* + *β*8×*MktSharei*

where *β*0p is the random effect for the p-th census tract (TractID). Paste results here.



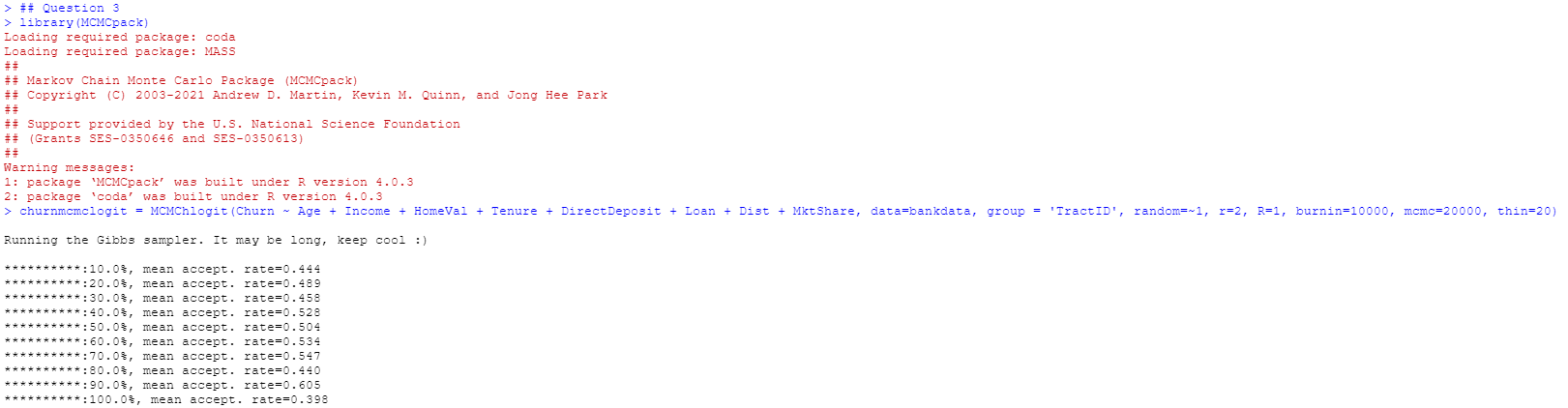


Check the fixed effect estimates of *β*1, *β*2*, β*3, *β*4, *β*5, *β*6, *β*7, *β*8 again. Are they still statistically significant? Please also calculate the AIC and BIC of this model using the R functions AIC( ) and BIC( ). Based on the AIC and BIC, compare the model fit of this model to the models in (1). (15 points)

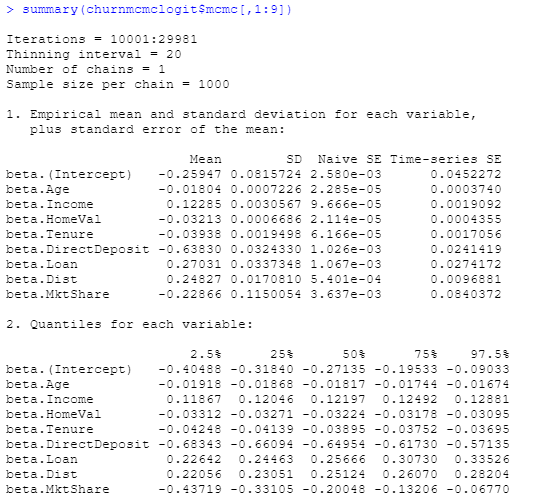


From the results above, the following variables are still statistically significant: Age, Income, HomeVal, Tenure, DirectDeposit, Dist.

3). For the model in (1), use the MCMCpack function MCMChlogit() to estimate the same parameters with Bayesian estimation. Because the model only has a random intercept, specify random=~1 and r=2, R=1 in the MCMChlogit() function. Please also set burnin=10000, mcmc=20000 and thin=20.

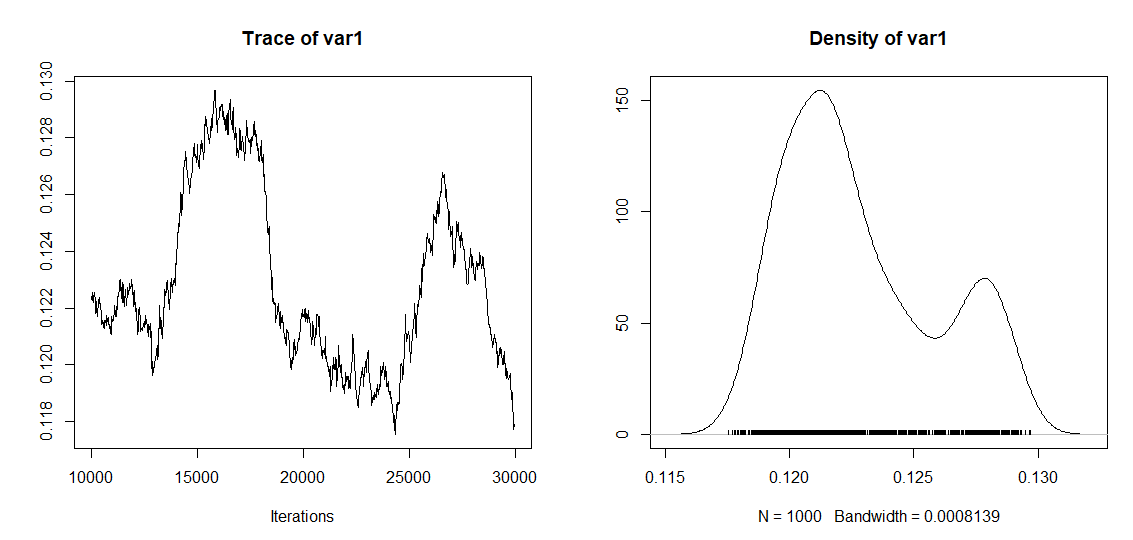


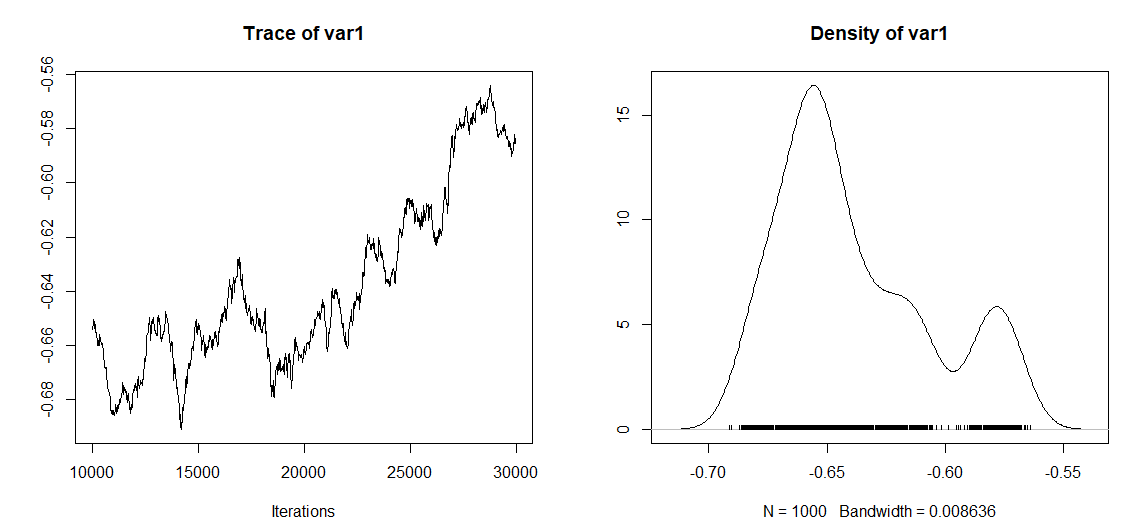
Please copy and paste the Bayesian estimation results of the fixed effects (same fixed effects as in (1)) in the model using summary("*yourBayesianModelName"*$mcmc[,1:9]). From the Bayesian posterior intervals, are the fixed effects significant at the 5% level?



From the results above, the following variables are statistically significant at the 5% level: Age, Income, HomeVal, Tenure, DirectDeposit, Loan, Dist and MarketShare. So, Betas1-8 are significant at 5%.

Use the plot() function to plot the posterior sampling chains and posterior densities for *β*2 and *β*5; copy and paste the results here. (15 points)





**Probit Regression: Bayesian Estimation**

In this exercise, we will practice coding the Gibbs sampler for a probit regression model using the dataset "CreditCard\_LatePayment\_Data.csv". The dataset has the following variables.

|  |  |
| --- | --- |
| ConsumerID | ID's of the sampled consumers |
| Latepay | Whether the consumer makes a late payment in the month |
| Usage | Monthly credit usage activities |
| Balance | The customer's outstanding balance in the month |

1). We would like fit the following probit regression model

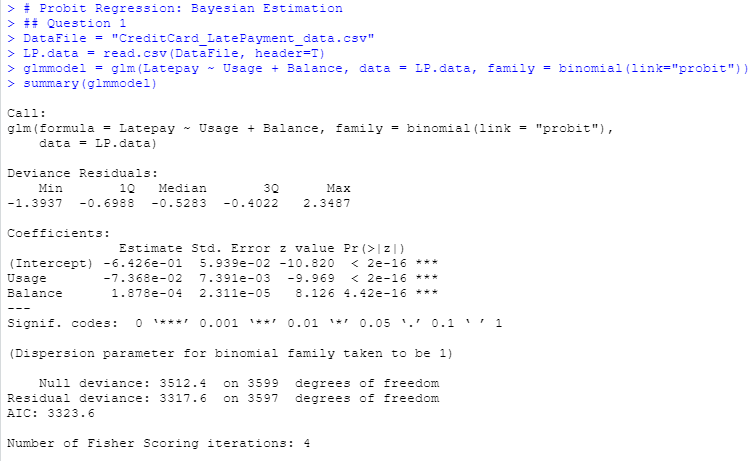
*Yij\* = β0 + β1×Usageij + β2×Balanceij + εij*

*Latepayij =*0 if *Yij\** ≤ 0

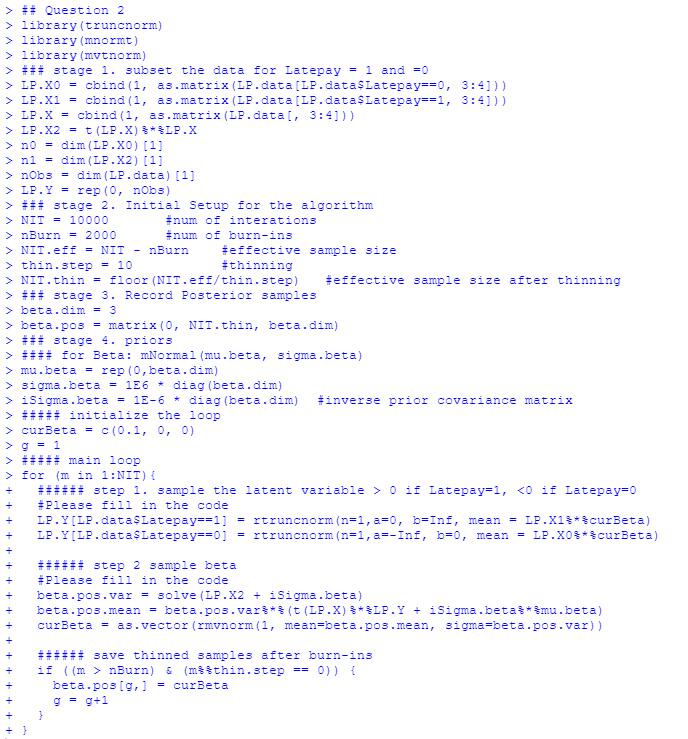
*Latepayij =*1 if *Yij\** > 0

*εij ~N*(0, 1)

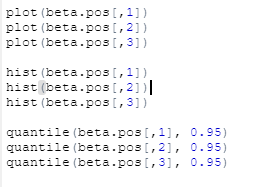
Please use the R function glm( ) to fit this model by MLE. Copy and paste the summary of the results here.

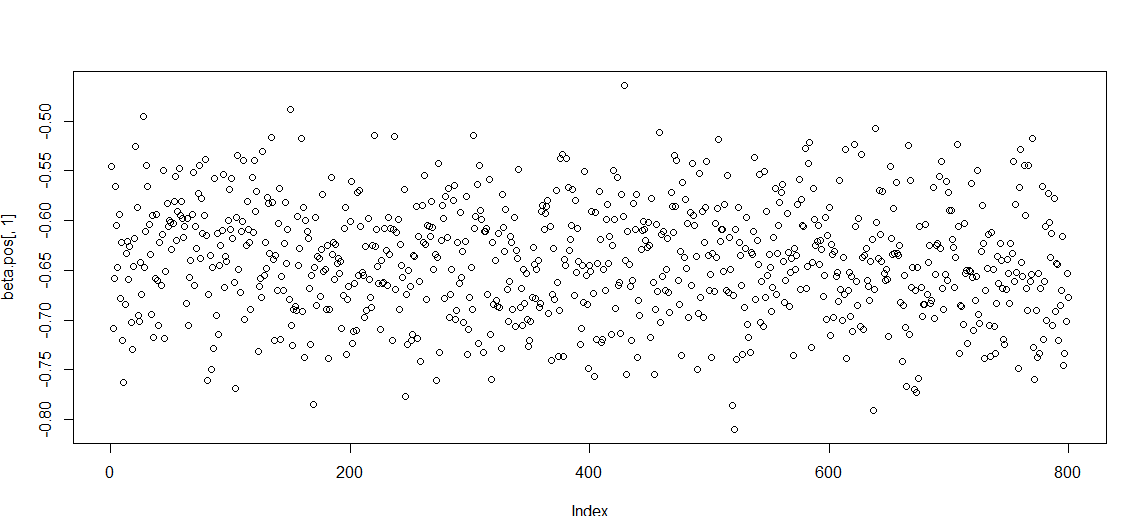


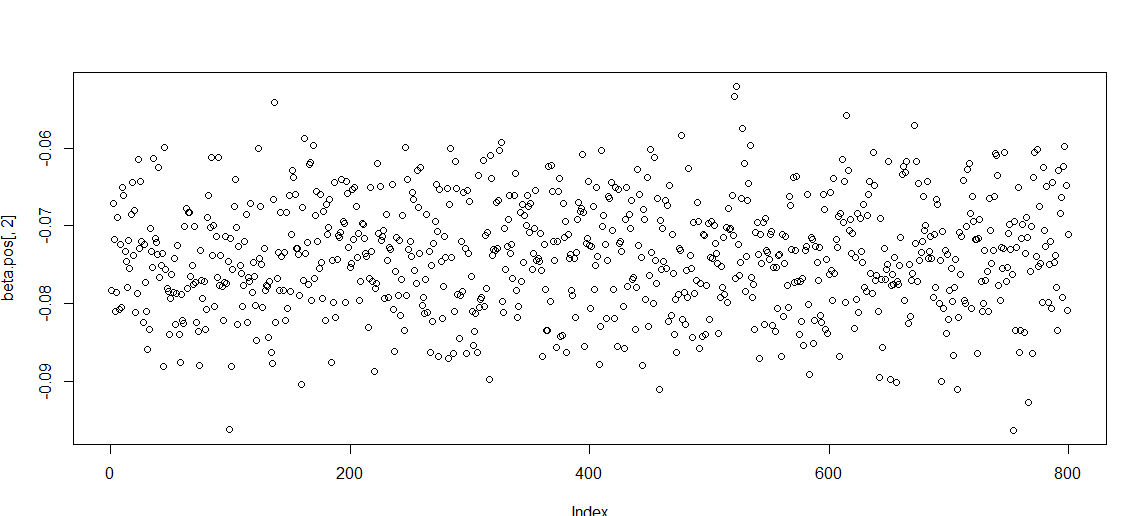
2).Next, we will fit the model above using a Gibbs sampler for Bayesian inference, which involves sampling the latent *Yij\**. Parts of the R code are in "Assignment-2\_Probit-code\_blanks.r". Please read the code carefully and fill in the code in the blanks in the file. You may use the rtruncnorm( ) function in the library(truncnorm) to sample from truncated normal distributions. For the linear regression part given the sampled latent *Yij\** in the main loop, please refer to the code BayesianLM.r on Canvas

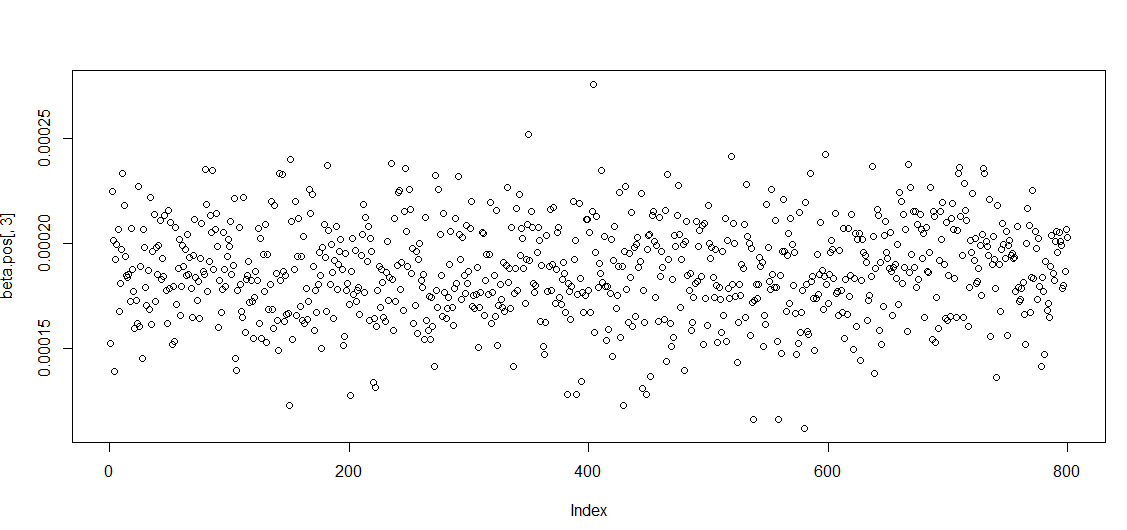


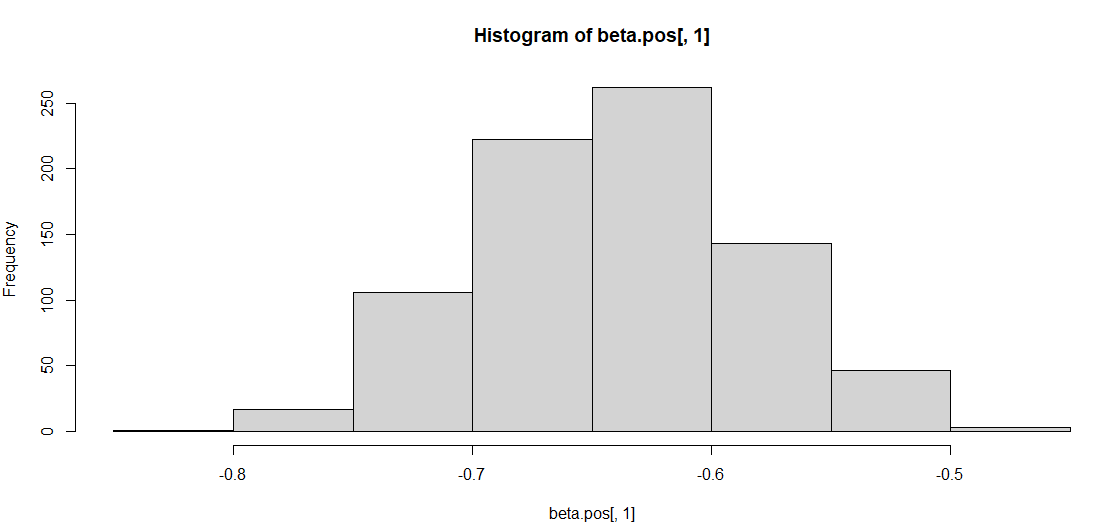
Please run the completed code. Use the plot() function to plot the posterior sampling chains and hist() to plot posterior histograms for *β0, β1, β2 .* Copy and paste the results here. Please also calculate the 95% posterior intervals for *β0, β1, β2 .* Copy and paste the results here.

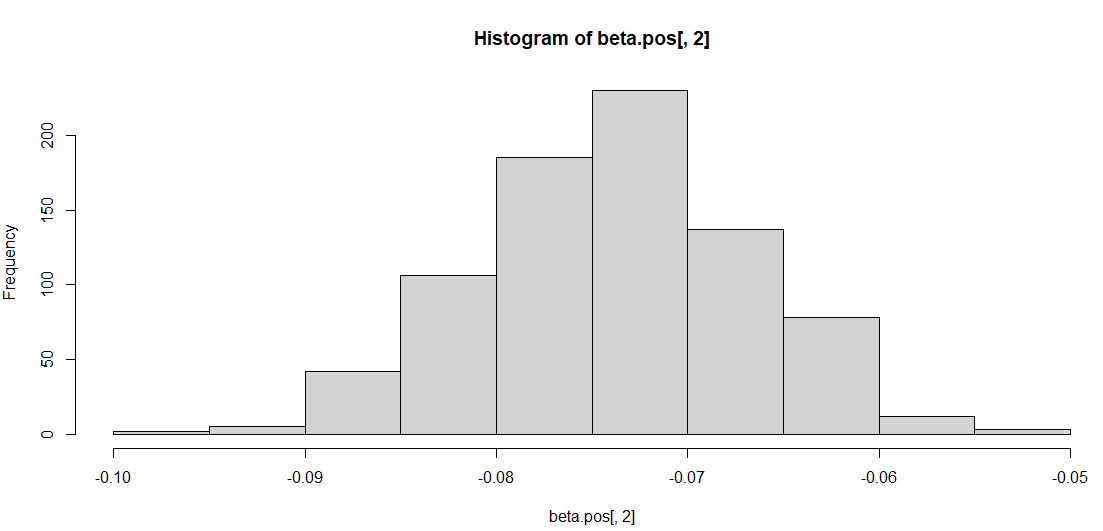
****

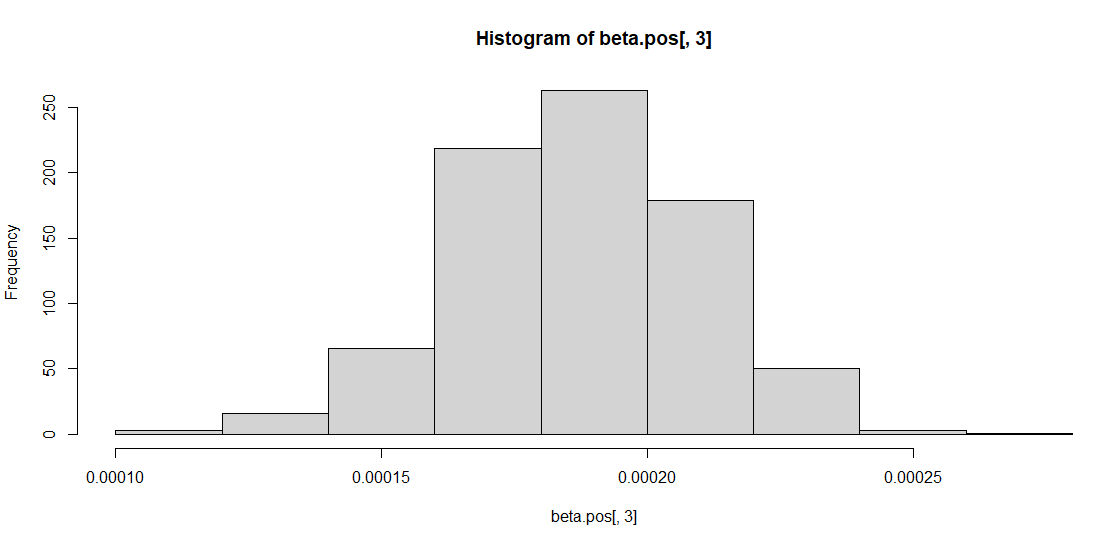
****

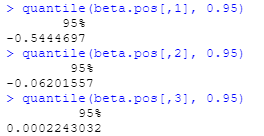
****

****

****

****

****

****