**OM386 Advanced Data Analytics in Marketing**

**Assignment 2**

**Due: February 23rd, 11:59pm**

**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 managers give 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 "BankRetention\_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 |

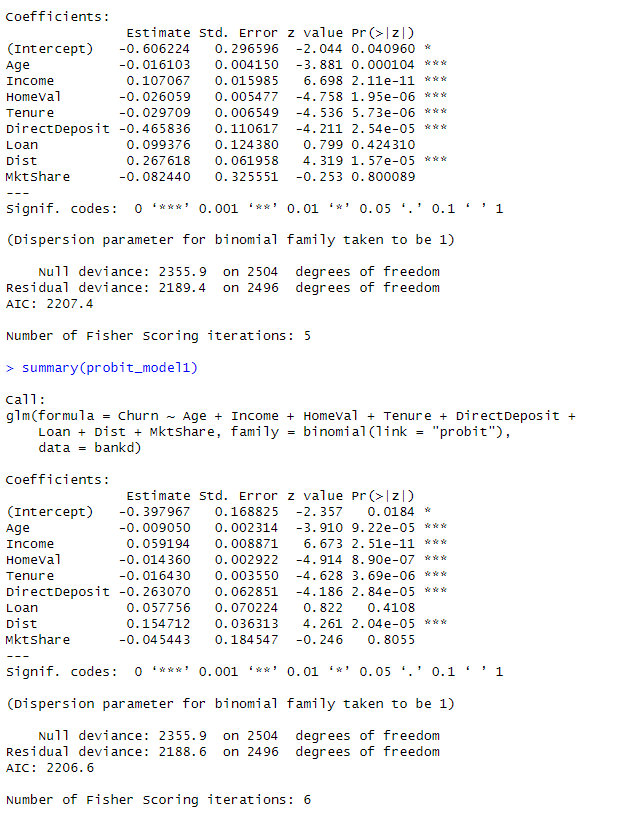
1). Read the data into R. Convert TractID into a factor variable.

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 of the logistic and probit models. Which model (logistic or probit) fits the data better based on AIC?

* **Both logit and probit have same direction for the coefficients, therefore, the coefficients for churn in both models will have similar interpretation**
* **b1 (Age)**: For both models, as age increases by one year, the log odds (logistic) or the Z-score (probit) of churning decreases. This suggests that older customers are less likely to churn
* **b2 (Income)**: Higher income is associated with a lower likelihood of churning for both models. This means that as income increases, the customer is less likely to churn.
* **b3 (HomeVal)**: Higher home values are associated with a lower likelihood of churning for both models. Customers with higher-valued homes are less likely to churn.
* **b4 (Tenure)**: For both models, as tenure with the bank increases, the likelihood of churning decreases. Long-standing customers are less likely to churn.
* **b5 (DirectDeposit)**: Having a direct deposit is associated with a significant decrease in the likelihood of churning for both models. Customers with direct deposit set up are less likely to churn.
* **b6 (Loan)**: Customers with a loan are more likely to churn in both models, as indicated by the positive coefficient.
* **b7 (Dist)**: The coefficient for **Dist** is positive in the probit model but not statistically significant in either model. This suggests that the distance does not have a clear effect on the likelihood of churning.
* **b8 (MktShare)**: In the logistic model, a higher market share is associated with a lower likelihood of churning, but the effect is not statistically significant in the probit model.

**Statistical Significance:**

* In the logistic model, all variables except **Dist** and **MktShare** are statistically significant at the 5% level, as indicated by the p-values (Pr(>|z|)).
* In the probit model, the variables **Age**, **Income**, **HomeVal**, **Tenure**, **DirectDeposit**, and **Loan** are statistically significant at the 5% level. **Dist** and **MktShare** are not statistically significant as well.

**AIC Comparison:**

The probit model has a slightly lower AIC (2206.6) compared to the logistic model (2207.4), suggesting that the probit model fits the data marginally better. However, the difference is very small, so the choice between the two models might not be clear-cut based on AIC alone. Both models provide a very similar fit to the data according to this criterion.

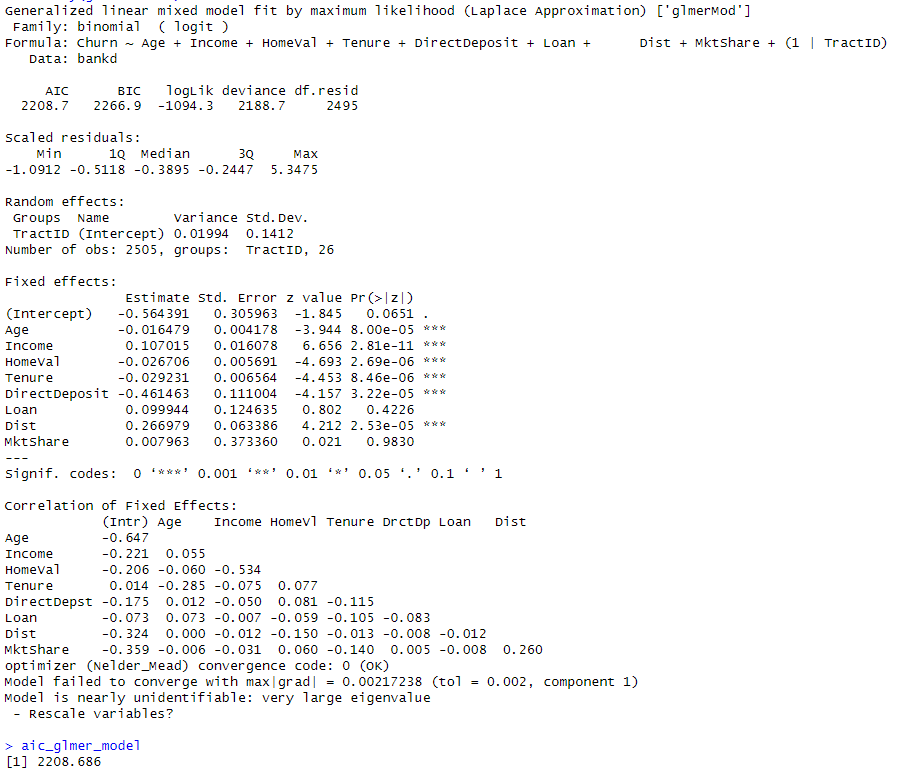
2). Next we will use a random effect grouped by TractID in the logistic regression.

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

*Churni* ~ *β*0k + *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*0k is the random effect for the k-th census tract (grouped by TractID). Paste results here.



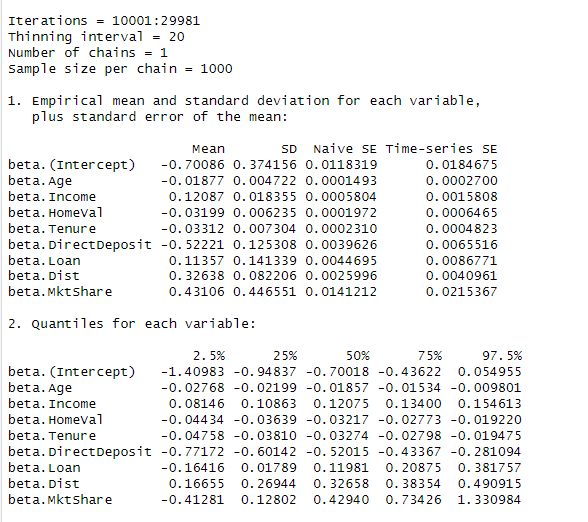
Check the fixed effect estimates of *β*1, *β*2*, β*3, *β*4, *β*5, *β*6, *β*7, *β*8 again. Are they still statistically significant? Please also calculate the AIC of this model and compare the model fit of this model to the models in (1).

All variables except MktShare are statistically significant, suggesting that they have a statistically discernible relationship with the probability of churn.

For the glmer model, we got AIC: 2208.686 The model with the lowest AIC is generally considered to be the better fit. In this case, the probit model has the lowest AIC, followed very closely by the logistic model, suggesting that among the three, the probit model fits the data slightly better than the others.

3). For the model in (2), 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 (2)) in the model using summary("*yourBayesianModelName"*$mcmc[,1:9]). From the Bayesian posterior intervals, are the fixed effects significant at the 5% level?

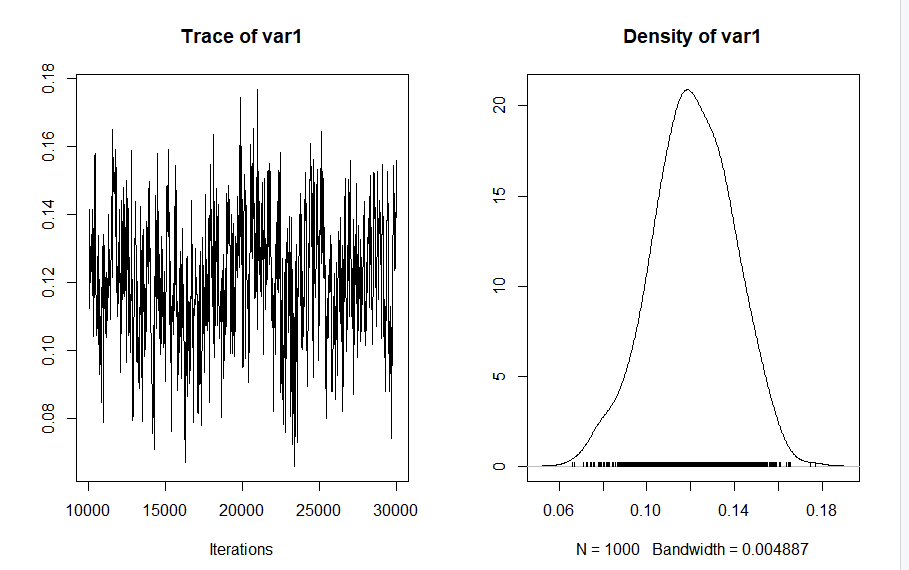


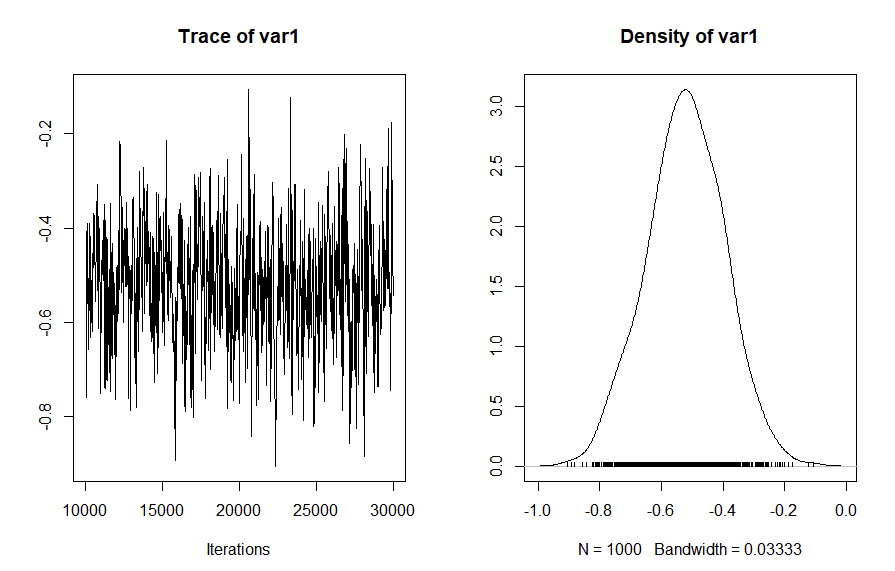
In Bayesian analysis, significance is determined by examining credible intervals. If a 95% credible interval does not contain zero, it suggests that there is strong evidence that the parameter is credibly different from zero at the 5% level.

* **beta.Intercept (Intercept)**: The 95% CI ranges from approximately -1.4098 to -0.0362, which does not include zero, indicating significance.
* **beta.Age (Age)**: The 95% CI ranges from approximately -0.0876 to -0.0134, which does not include zero, indicating significance.
* **beta.Income (Income)**: The 95% CI ranges from approximately 0.0814 to 0.1546, which does not include zero, indicating significance.
* **beta.HomeVal (HomeVal)**: The 95% CI ranges from approximately -0.0444 to -0.0192, which does not include zero, indicating significance.
* **beta.Tenure (Tenure)**: The 95% CI ranges from approximately -0.0478 to -0.0197, which does not include zero, indicating significance.
* **beta.DirectDeposit (DirectDeposit)**: The 95% CI ranges from approximately -0.7712 to -0.2105, which does not include zero, indicating significance.
* **beta.Loan (Loan)**: The 95% CI ranges from approximately -0.1646 to 0.3817, which does include zero, indicating that it is not significant.
* **beta.Dist (Dist)**: The 95% CI ranges from approximately 0.1665 to 0.4904, which does not include zero, indicating significance.
* **beta.MktShare (MktShare)**: The 95% CI ranges from approximately -0.4128 to 1.33098, which includes zero, indicating that it is not significant.

All coefficients except for **Loan** and **MktShare** have 95% credible intervals that do not include zero, suggesting strong evidence that they are credibly different from zero. The coefficients for **Loan** and **MktShare** include zero within their credible intervals, suggesting weaker evidence for their effects on the response variable at the 5% level.

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.





**Probit Regression: Bayesian Estimation**

In this exercise, we will practice coding the MCMC algorithm (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 |

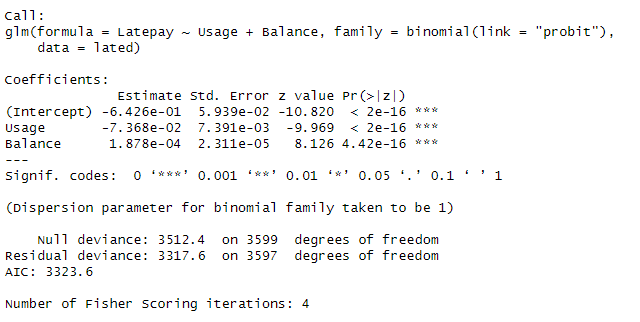
4). We would like fit the following probit regression model

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

*Latepayij =*0 if *Yij\** < 0

*Latepayij =*1 if *Yij\** ≥ 0

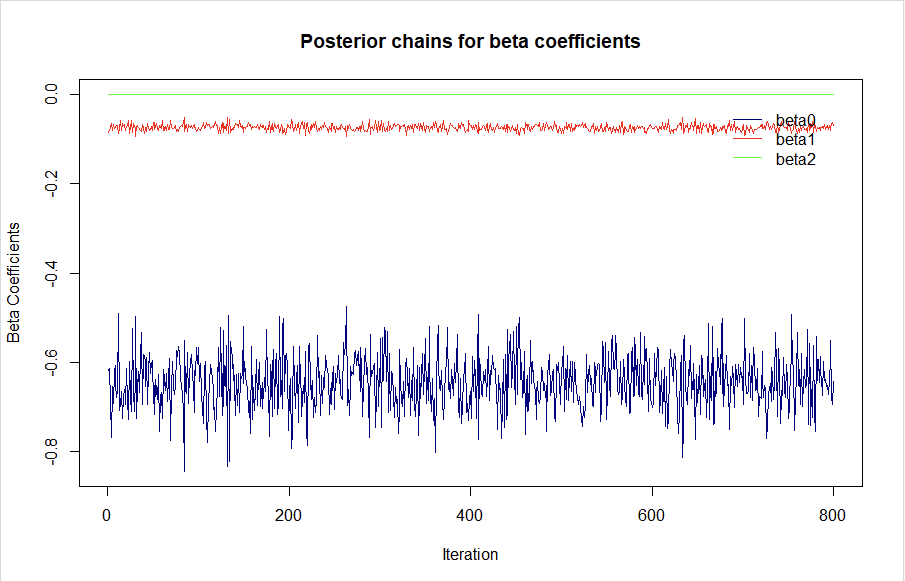
*εij ~N*(0, 1)

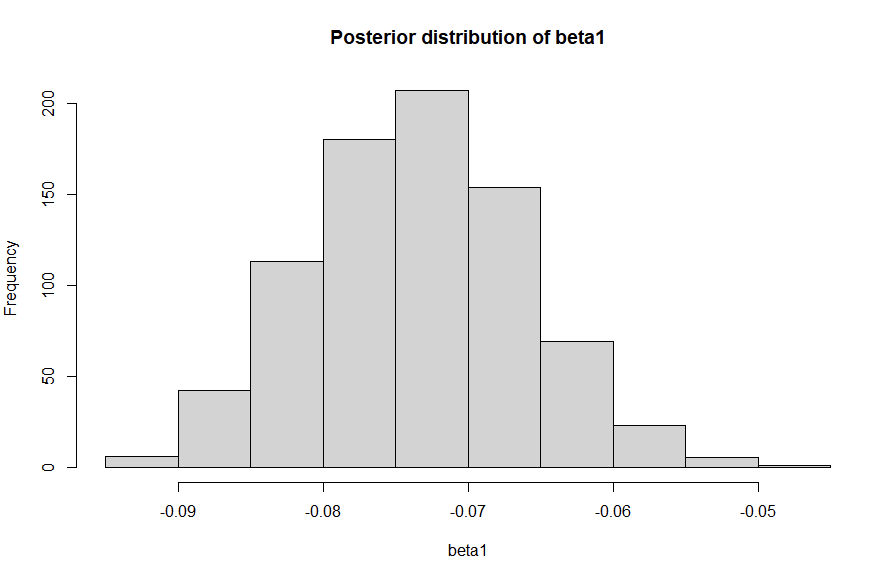


Please use the R function glm( ) to fit this model by MLE. Copy and paste the summary of the results here.

Both predictors, Usage and Balance, are statistically significant and have a substantial impact on the probability of a consumer making a late payment. The negative coefficient for Usage is a bit counterintuitive, as one might expect that higher credit usage could be associated with a higher probability of late payment. The positive coefficient for Balance aligns with intuition, suggesting that customers with higher balances are more likely to make late payments. Moreover, the AIC value provides a measure for comparing this model's fit to other models.

5).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 (hint: only the use the code for sampling *β’s* and fix the variance at 1).





Please run the completed code. Use the ts.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.

> # Print the 95% posterior intervals

> print(paste0("95% CI for beta0: [", beta0\_CI[1], ", ", beta0\_CI[2], "]"))

[1] "95% CI for beta0: [-0.75915312265822, -0.522217438476307]"

> print(paste0("95% CI for beta1: [", beta1\_CI[1], ", ", beta1\_CI[2], "]"))

[1] "95% CI for beta1: [-0.0869155130337512, -0.0591639925683838]"

> print(paste0("95% CI for beta2: [", beta2\_CI[1], ", ", beta2\_CI[2], "]"))

[1] "95% CI for beta2: [0.000141268565983054, 0.000233403504325541]"