Project Proposal

Dong Luo, u5319900

The project will use the “bank marketing” data set provided on the course website.

The main issue to be addressed with the data set is the difference of average yearly balance among married, single, or divorced client.

As we are investigating the relation between the balance and marital status, these two variables need to be used during the analysis. Age and housing will also be used as they are expected to affect the balance as well.

We will fit a linear regression model using balance as the response variable, and age, marital status, and housing as explanatory variables. We can use Bayesian model comparison method to find the model of highest posterior probability. Using g-prior for β and a weakly informative inverse-gamma prior for, we can derive the full conditional distribution of β and the marginal posterior distribution of. The posterior means and confidence intervals of β and can be obtained by generating independent Monte Carlo samples. Taking into model uncertainty, we can approximate the model averaged estimate of β and answer the main question.

As 17 variables are provided, it’s hard to decide which variables we shall use. We may get misleading result as the true model may be much more complicated than our linear regression model. We also have to notice that we may have to deal with huge computations as the data set for the project contains statistics from 45211 customers, which is much larger than any data set we’ve talked about in the class and usually means long waiting time. Hence we may have to choose a random subset of the data to use and use some other random proportion of the data to access the fitted model. We also have to pay much attention when interpreting the estimated β as the model becomes more complicated when we have factor “marital” which has more than two levels and the interaction between the two factors “marital” and “housing”.