Module 10: Logistic Regression

Rebecca C. Steorts

```r
library(dplyr)
```

Agenda

We will explore a variable selecion model for Bayesian logistic regression using the data in **azdiabetes.dat**. This closely follows the exercise 10.5 of the Hoff book.

Diabetes data

Application to diabetes (Exercise 9.2, part a)

Suppose we have data on health-related variables of a population of 532 women.

Our goal is to model the conditional distribution of glucose level (glu) as a linear combination of the other variables, excluding the variable diabetes.[1]

[1] See Exercise 7.6 for the data description.

Diabetes Data

```r
library(knitr)
```

```
## Warning: package 'knitr' was built under R version 3.5.2
```

```r
rm(list=ls())
azd_data = read.table("azdiabetes.dat", header = TRUE)
head(azd_data)
```

```
##   npreg glu bp skin  bmi   ped age diabetes
## 1     5  86 68   28 30.2 0.364  24       No
## 2     7 195 70   33 25.1 0.163  55      Yes
## 3     5  77 82   41 35.8 0.156  35       No
## 4     0 165 76   43 47.9 0.259  26       No
## 5     0 107 60   25 26.4 0.133  23       No
## 6     5  97 76   27 35.6 0.378  52      Yes
```

Diabetes Data

The dataset contains information on diabetes status of 532 individuals along with 7 covariates. We will consider building a logistic regression model for predicting diabetes as a function of the following variables.

$x_1 =$ number of pregnancies

$x_2 =$ blood pressure

$x_3 =$ body mass index

$x_4 =$ diabetes perdigree

$x_5 =$ age

Let us appropriately subset the data. Note that the piping operator %>% from the "dplyr"" package in R makes your code easy to read, but it is not necessary.

Diabetes Data

```r
X <- as.matrix(azd_data[,c(-2,-4,-8)])
y = azd_data$glu
head(X)
```

```
##      npreg bp  bmi   ped age
## [1,]     5 68 30.2 0.364  24
## [2,]     7 70 25.1 0.163  55
## [3,]     5 82 35.8 0.156  35
## [4,]     0 76 47.9 0.259  26
## [5,]     0 60 26.4 0.133  23
## [6,]     5 76 35.6 0.378  52
```

Task 1: Standardization

Center and scale each of the x-variables by substracting the sample average and dividing by the sample stanard deviation. Why is it important to standardize the x-variables?

Task 1: Solution

```
# standardize data to have mean 0 and variance 1
ys = scale(y)
Xs = scale(X)
n = dim(Xs)[1]
p = dim(Xs)[2]
```

Task 2: Logistic regression

The logistic regression model we consider is of the form $\Pr(Y_i = 1 \mid x_i, \beta, \gamma) = e^{\theta_i}/(1 + e^{\theta_i})$ where $\beta = (\beta_0, \ldots, \beta_5)$, $\gamma = (\gamma_1, \ldots, \gamma_5)$ and

$$\theta_i = \beta_0 + \sum_{j=1}^{5} \beta_j \gamma_j x_{i,j}$$

Here $\gamma_j = 1$ if the jth variable is a predictor of diabetes and 0 otherwise. For example, $\gamma = (1, 1, 0, 0, 0)$ corresponds to the model $\theta_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2}$. Obtain posterior distribution of β and γ, assuming the following independent priors.

$$\gamma_j \sim \text{Ber}(0.5), \quad \beta_0 \sim \text{Normal}(0, 16), \quad \beta_j \sim \text{Normal}(0, 4)$$

for each $j > 0$.

Task 3

Implement a Metropolis-Hastings algorithm for approximating the posterior distributions of β and γ. Adjust the proposal distribution to achieve a reasonable accepance rate, and run the algorithm long enough so that the effective sample size is at least 1000 for each parameter.

You can also use RStan or Rjags to do this step. However you still need to monitor the acceptance rate and effective sample size and report your findings. Here is a sample Rjags code.

Task 3: Solution

```
# Logistic regression:

  logistic_model <- "model{

   # Likelihood

   for(i in 1:n){
    Y[i] ~ dbern(q[i])
    logit(q[i]) <- beta0 + beta[1]*gamma[1]*X[i,1] + beta[2]*gamma[2]*X
                   beta[3]*gamma[3]*X[i,3] + beta[4]*gamma[4]*X[i,4] +
   }

   #Priors
  beta0 ~ dnorm(0,1/16)
   for(j in 1:6){
    beta[j] ~ dnorm(0,1/4)
    gamma[j] ~ dbern(0.5)
   }

  }"
```