STA 601: Lab 11

Akbota Anuarbek

Probit regression

$Pr(y_i = 1|x_i) = \Phi(x_i'\beta)$ is equivalent to
$y_i^* = x_i'\beta + \epsilon_i$, $y_i = I(y_i^* > 0)$, $\epsilon_i \sim \mathcal{N}(0, 1)$

proof:
$Pr(y_i = 1|x_i) = Pr(y_i^* > 0|x_i) = Pr(x_i'\beta + \epsilon_i > 0|x_i) = Pr(\epsilon > -x_i'\beta|x_i)$
Since $\epsilon_i \sim \mathcal{N}(0, 1)$, by symmetry of standard normal distribution we obtain $Pr(\epsilon_i > -x_i'\beta|x_i) = Pr(\epsilon_i < x_i'\beta|x_i) = \Phi(x_i'\beta)$
Thus, $Pr(y_i = 1|x_i) = \Phi(x_i'\beta)$

Logistic regression

$Pr(y_i = 1|x_i) = \text{logit}^{-1}(x_i'\beta)$ is equivalent to
$y_i^* = x_i'\beta + \epsilon_i$, $y_i = I(y_i^* > 0)$, $\epsilon_i \sim \text{logistic}(0, 1)$

proof:
$Pr(y_i = 1|x_i) = Pr(y_i^* > 0|x_i) = Pr(x_i'\beta + \epsilon_i > 0|x_i) = Pr(\epsilon_i > -x_i'\beta|x_i)$
Since $\epsilon_i \sim \text{logistic}(0, 1)$, by symmetry of the logistic distribution we obtain $Pr(\epsilon_i > -x_i'\beta|x_i) = Pr(\epsilon_i < x_i'\beta|x_i) = \text{logit}^{-1}(x_i'\beta)$
Thus, $Pr(y_i = 1|x_i) = \text{logit}^{-1}(x_i'\beta)$

$y_i^*|x_i, \beta \sim \mathcal{N}(x_i'\beta, 1)$

$y_i = I(y_i^* > 0)$

$\beta \sim \mathcal{N}(b_0, B_0)$

Full conditionals:

$\underline{\pi(\beta|y, y^*, X)} \propto \pi(\beta)L(y^*|X, \beta)L(y|y^*, \beta)$

$\propto \exp\big(-\frac{1}{2}(\beta-b_0)'B_0^{-1}(\beta-b_0)\big) \exp\big(-\frac{1}{2}(y^*-X\beta)'(y^*-X\beta)\big) \prod_{i=1}^n \big(I(y_i^* > 0)y_i + I(y_i^* < 0)(1-y_i)\big)$

$\propto \exp\big(-\frac{1}{2}(\beta-b_0)'B_0^{-1}(\beta-b_0)\big) \exp\big(-\frac{1}{2}(y^*-X\beta)'(y^*-X\beta)\big)$

$\propto \exp\big[-\frac{1}{2}(\beta'B_0^{-1}\beta - 2\beta'B_0^{-1}b_0 + \beta'X'X\beta - 2\beta'X'y^*)\big]$

$\propto \exp\big[-\frac{1}{2}(\beta'(B_0^{-1} + X'X)\beta - 2\beta'(B_0^{-1}b_0 + X'y^*))\big]$

$\propto \underline{\mathcal{N}(b, B)}$, where

$\underline{B = (B_0^{-1} + X'X)^{-1}}$

$\underline{b = B(B_0^{-1}b_0 + X'y^*)}$

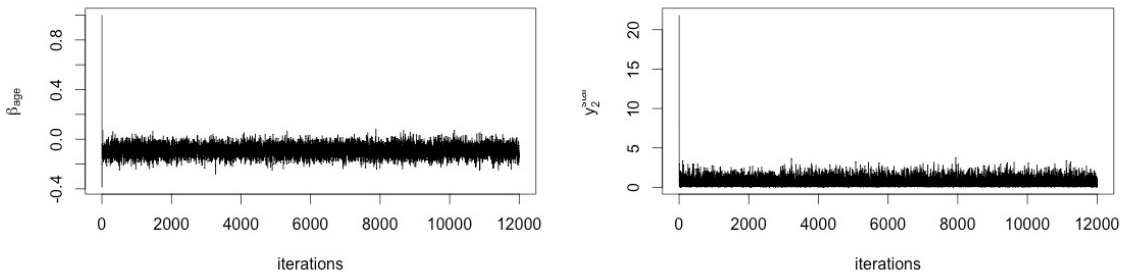
$\pi(y_i^*|y_i, x_i, \beta) \propto L(y_i^*|\beta)L(y_i|y_i^*, \beta) \propto N(y_i^*; x_i'\beta, 1)[I(y_i^* > 0)y_i + I(y_i^* < 0)(1-y_i)]$

$\underline{\pi(y_i^*|y_i = 1, x_i, \beta)} \propto N(y_i^*; x_i'\beta, 1)I(y_i^* > 0) \propto \underline{N_+(y_i^*; x_i'\beta, 1)}$

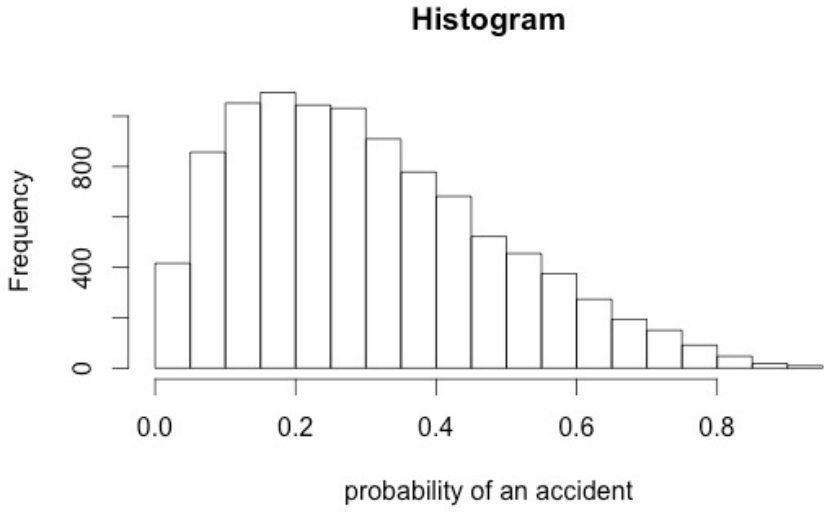
$\underline{\pi(y_i^*|y_i = 0, x_i, \beta)} \propto N(y_i^*; x_i'\beta, 1)I(y_i^* < 0) \propto \underline{N_-(y_i^*; x_i'\beta, 1)}$

Gibbs sampling is more difficult in the logistic case, because there is no conjugacy between normal and logistic distributions, and we cannot obtain full conditionals easily as in case of probit regression.

Run Gibbs sampling, R code is given at the end. Look at traceplots to ensure that sampler has converged, you can see couple of them below. Burn-in first 2000 observations.



Using $x'_{TA} = (1, 26, 0, 1)$ and posterior sample of β obtained from Gibbs, calculate posterior predictive distribution for TA using the formula $\Phi(x'_{TA}\beta^{(t)})$. Histogram is given below:



Probability of being involved in an accident in the last 6 month for TA:
$Pr(y_{TA} = 1|x_{TA}, X, y) = \Phi(x'_{TA}\hat{\beta}) = 0.275$, where $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_{age}, \hat{\beta}_{course}, \hat{\beta}_{likeStats})' =$
$= (\frac{\sum_{t=1}^{T}\beta_1^{(t)}}{T}, \frac{\sum_{t=1}^{T}\beta_{age}^{(t)}}{T}, \frac{\sum_{t=1}^{T}\beta_{course}^{(t)}}{T}, \frac{\sum_{t=1}^{T}\beta_{likeStats}^{(t)}}{T})' = (1.036, -0.090, -0.483, 0.694)'$ is a posterior point estimate for β.

$x_1 = (1, 17, 1, 0)'$, $x_2 = (1, 18, 1, 1)'$

Then $Pr(y_1 = 1|x_1, X, y) = \Phi(x_1'\hat{\beta}) = 0.166$
$Pr(y_2 = 1|x_2, X, y) = \Phi(x_2'\hat{\beta}) = 0.358$

According to the probit output, 18-year old lover of statistics has higher probability of being involved in an accident. So I would prefer 17-year old statistics hater to park my car.

```
library(truncnorm)
library(mvtnorm)
library(optimbase)

setwd('/Users/akbota/Documents/STA601/Labs/lab11')
d<- read.csv("data.csv")

n=35
y=d$y
X=matrix(data=rep(1), nrow=n, ncol=4)
X[,2]=d$age
X[,3]=d$course
X[,4]=d$likeStats

N=12000
b0=rep(0,4)
B0=diag(4)

beta=matrix(nrow=4, ncol=N)
beta[,1]=rep(1,4)
y.star=matrix(nrow=n,ncol=N)
```

```
for(t in 2:N){

  for(i in 1:n){
    if(y[i]==1){
      y.star[i,t]=rtruncnorm(1, a=0, mean=X[i,]%*%beta[,t-1])
    }else if(y[i]==0){
      y.star[i,t]=rtruncnorm(1, b=0, mean=X[i,]%*%beta[,t-1])
    }
  }

  B=solve(solve(B0)+transpose(X)%*%X)
  b=B%*%(solve(B0)%*%b0+transpose(X)%*%y.star[,t])
  beta[,t]=rmvnorm(1, mean=b, sigma=B)

  if(t%%1000==0){
    print(t)
  }
}

plot(beta[2,], type='l', ylab=expression(beta[age]), xlab="iterations")
plot(y.star[2,], type='l', ylab=expression(y[2]^star), xlab="iterations")

Beta=c(mean(beta[1,2001:N]), mean(beta[2,2001:N]), mean(beta[3,2001:N]), mean(beta[4,2001:N]))

x.TA=c(1,26,0,1)
yis1.TA=pnorm(transpose(beta[,2001:N])%*%x.TA)

hist(yis1.TA, xlab="probability of an accident", main="Histogram")
```

```
pty.yis1.TA=pnorm(x.TA%*%Beta)


x1=c(1,17,1,0)
x2=c(1,18,1,1)
ptyx1=pnorm(x1%*%Beta)
ptyx2=pnorm(x2%*%Beta)
```