S2208 MATH8050 Data Analysis - Section 001: Homework 9 Due on 11/09/22

Adithya Ravi, C09059838

2022-11-09

Solutions

Question1

```
library(knitr)
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
```

```
y = azd_data$glu
# remove glu and diabetes
X = as.matrix(azd_data[,c(-2,-8)])
head(X)
```

```
##      npreg bp skin  bmi   ped age
## [1,]     5 68   28 30.2 0.364  24
## [2,]     7 70   33 25.1 0.163  55
## [3,]     5 82   41 35.8 0.156  35
## [4,]     0 76   43 47.9 0.259  26
## [5,]     0 60   25 26.4 0.133  23
## [6,]     5 76   27 35.6 0.378  52
```

```
ys = scale(y)
Xs = scale(X)
n = dim(Xs)[1]
p = dim(Xs)[2]
```

```
g = n
nu0 = 2
s20 = 1
```

```
Hg = (g/(g+1)) * Xs %*% solve(t(Xs) %*% Xs) %*% t(Xs)
SSRg = t(ys) %*% ( diag(1,nrow=n) - Hg ) %*% ys
```

```
S = 1000
set.seed(1234)

s2 = 1/rgamma(S, (nu0+n)/2, (nu0*s20 + SSRg)/2)
head(s2)
```

```
## [1] 0.9143171 0.8318653 0.8296027 0.8872172 0.8222744 0.8786227
```

```
Vb = g*solve(t(Xs) %*% Xs)/(g+1)
Eb = Vb %*% t(Xs) %*% ys
E = matrix(rnorm(S*p, 0, sqrt(s2)),S,p)

beta_s = t( t(E %*% chol(Vb)) + c(Eb))

sd_X = apply(X,2,sd)
Beta_a = sweep(beta_s,2,sd_X,FUN = "/")
```

```
Beta_CIa = apply(Beta_a, 2, quantile, c(0.025, 0.975))
kable(data.frame(Beta_CIa), digits=4)
```

	npreg	bp	skin	bmi	ped	age
2.5%	-0.0527	0.0001	-0.0041	0.0060	0.1313	0.0149
97.5%	0.0102	0.0138	0.0160	0.0343	0.5701	0.0342

```
lpy.X <- function(y, X, g=length(y), nu0=1,
s20=try(summary(lm(y~ -1+X))$sigma^2,
silent=TRUE)){
n = dim(X)[1]
p = dim(X)[2]
if (p==0) { Hg = 0; s20 = mean(y^2)}
if (p>0){ Hg = (g/(g+1)) * X %*% solve(t(X) %*% X) %*% t(X) }
SSRg = t(y) %*% ( diag(1, nrow=n) - Hg ) %*% y -
.5*( n*log(pi) + p*log(1+g) + (nu0+n)*log(nu0*s20 + SSRg) -
nu0*log(nu0*s20)) +
lgamma( (nu0+n)/2 ) - lgamma(nu0/2)
}
```

```
g = n
nu0 = 1
z = rep(1, p)

lpy.c = lpy.X(ys, Xs[,z==1,drop=FALSE])
S = 1000
Z = matrix(NA, S, p)
B = matrix(0, S, p)
```

```
  for(s in 1:S){
    for (j in sample(1:p)){
      zp = z
      zp[j] = 1 - zp[j]
      lpy.p = lpy.X(ys,Xs[, zp==1, drop=FALSE])
      r = (lpy.p - lpy.c) * (-1)^(zp[j]==0)
      zp[j] = rbinom(1, 1, 1/(1+exp(-r)))
      if(z[j] == zp[j]) {lpy.c = lpy.p}
    }
  Z[s,] = z

  pm = sum(z==1)
  if(pm==0){
    Hg = 0
    s20 = mean(y^2)
  }
  if(pm>0){
    Hg = (g/(g+1)) * Xs[,z==1,drop=F] %*%
    solve(t(Xs[,z==1,drop=F]) %*%
    Xs[,z==1,drop=F]) %*% t(Xs[,z==1,drop=F])
    s20=summary(lm(ys ~ -1+Xs[,z==1,drop=F]))$sigma^2
  }
  SSRg = t(ys) %*% ( diag(1,nrow=n) - Hg ) %*% ys
  s2 = 1/rgamma(1, (nu0+n)/2, (nu0*s20 + SSRg)/2)

  Vb = g * solve(t(Xs[,z==1,drop=F]) %*% Xs[,z==1,drop=F])/(g+1)
  Eb = Vb %*% t(Xs[,z==1,drop=F]) %*% ys
  E = rnorm(p, 0, sqrt(s2))
  beta_z = E %*% chol(Vb) + c(Eb)
  B = t( t(beta_z))
}
```

```
pprob_Z = apply(Z,2,mean)
pprob_Z = data.frame(matrix(pprob_Z,nr=1,nc=p))
names(pprob_Z) = names(azd_data[c(-2,-8)])
row.names(pprob_Z) = 'posterior including \n probability'
kable(pprob_Z)
```

	npreg	bp	skin	bmi	ped	age
posterior including probability	1	1	1	1	1	1

```
Beta_b = sweep(B,2,sd_X,FUN = "/")
# 95% credible interval
Beta_CIb = apply(Beta_b, 2, quantile, c(0.025, 0.975))
kable(data.frame(Beta_CIb), col.names = names(azd_data[c(-2,-8)]),
digits=4)
```

	npreg	bp	skin	bmi	ped	age
2.5%	0.006	6e-04	0.0027	0.0219	0.4026	0.023
97.5%	0.006	6e-04	0.0027	0.0219	0.4026	0.023

$$p(\beta|y,X,\sigma^2) \propto p(y|\beta,X,\sigma^2) \times p(\beta) \propto \exp\{\beta^T(X^Ty/\sigma^2)) - \frac{1}{2}\beta^T(\Sigma_0^{-1} + X^TX/\sigma^2)\beta\}$$

$$O_j = \frac{Pr(z_j = 1|y,X,z_{-j})}{Pr(z_j = 0|y,X,z_{-j})} = \frac{Pr(z_j = 1)}{Pr(z_j = 0)} \times \frac{Pr(y|X,z_{-j},z_j = 1)}{Pr(y|X,z_{-j},z_j = 0)}$$

Full condition on z_j follows Binomial distribution $p(z_j = 1)$ $Bin(1, o_j/o_j + 1)$

Question2

```
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
```

```
y = azd_data$glu
# remove glu and diabetes
X = as.matrix(azd_data[,c(-2,-8,-4)])
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
```

```
ys = scale(y)
Xs = scale(X)
n = dim(Xs)[1]
p = dim(Xs)[2]
```

```
# a function to compute the marginal probability
lpy.X <- function(y, X, g=length(y), nu0=1,
s20=try(summary(lm(y~ -1+X))$sigma^2,
silent=TRUE)){
n = dim(X)[1]
```

```
p = dim(X)[2]
if (p==0) { Hg = 0; s20 = mean(y^2)}
if (p>0){ Hg = (g/(g+1)) * X %*% solve(t(X) %*% X) %*% t(X) }
SSRg = t(y) %*% ( diag(1, nrow=n) - Hg ) %*% y

return(-.5*( n*log(pi) + p*log(1+g) + (nu0+n)*log(nu0*s20 + SSRg) -
nu0*log(nu0*s20)) +
lgamma( (nu0+n)/2 ) - lgamma(nu0/2))
}

#lpy.X(ys,Xs)
```

```
g = n
nu0 = 1 # unit information prior
z = rep(1, p)
# picking a starting value for the marginal probability
lpy.c = lpy.X(ys, Xs[,z==1,drop=FALSE])
S = 1000
Z = matrix(NA, S, p)
B = matrix(NA, S, p)
S2 = matrix(0,S,1)
lpy.c
```

```
##           [,1]
## [1,] -730.0736
```

```
sig0 = matrix(0,6,6)
sig0[row(sig0)==col(sig0)] =4
#sig0[1,1] = 16
for(s in 1:S){
# if(s %% 100 ==0) {print(s)}
# sample z

#print("loop enter")
for (j in sample(1:p)){
zp = z
##print(zp)
zp[j] = 1 - zp[j]
##print(zp)
lpy.p = lpy.X(ys,Xs[, zp==1, drop=FALSE])
##print(lpy.p)
##print(lpy.c)
if(zp[j]==0){r = lpy.c - lpy.p}
if(zp[j]==1){r = lpy.p - lpy.c}
##print(paste0("Current working dir: "))
##print(r)
zp[j] = rbinom(1, 1, 1/(1+exp(-r)))
#print(zp[j])
if(z[j] != zp[j]) {
  lpy.c = lpy.p
  z = zp
  }
}
```

```
##print("loop exit")
Z[s,] = z

# sample s2
pm = sum(z==1) # number of nonzero variables in the model
if (pm==0){
Hg = 0
s20 = mean(y^2)
}
if (pm>0){
Hg = (g/(g+1)) * Xs[,z==1,drop=F] %*% solve(t(Xs[,z==1,drop=FALSE]) %*%
Xs[,z==1,drop=F]) %*% t(Xs[,z==1,drop=F])
# estimated residual variance from OLS
s20=summary(lm(ys ~ -1+Xs[,z==1,drop=F]))$sigma^2 }
SSRg = t(ys) %*% ( diag(1,nrow=n) - Hg ) %*% ys

s2 = 1/rgamma(1, (nu0+n)/2, (nu0*s20 + SSRg)/2)
S2[s,] = s2

# sample beta
Vb = g * solve(t(Xs[,z==1,drop=F]) %*% Xs[,z==1,drop=F])/(g+1)
Eb = Vb %*% t(Xs[,z==1,drop=F]) %*% ys

sig0 = matrix(0,pm,pm)
sig0[row(sig0)==col(sig0)] = 4
Vb = solve( solve(sig0) + (t(Xs[,z==1,drop=F]) %*% Xs[,z==1,drop=F])/s2)
Eb = Vb %*% t(Xs[,z==1,drop=F]) %*% ys
#print(Vb)
#print(E)
#print(chol(Vb))
#print(Eb)
c = 1;
for (j in (1:p)){
  if(z[j]==1){
    B[s,j] = rnorm(1,Eb[c],sqrt(Vb[c,c]))
    c = c+1
  }
}
c = 1
}
```

```
pprob_Z = apply(Z,2,mean)
pprob_Z = data.frame(matrix(pprob_Z,nr=1,nc=p))
names(pprob_Z) = names(azd_data[c(-2,-8,-4)])
row.names(pprob_Z) = 'posterior including \n probability'
kable(pprob_Z)
```

	npreg	bp	bmi	ped	age
posterior including probability	0.105	0.179	1	0.688	1

```
Beta_CIb = apply(B, 2, quantile, c(0.025, 0.975), na.rm = TRUE)
kable(data.frame(Beta_CIb), col.names = names(azd_data[c(-2,-8,-4)]),
digits=4)
```

	npreg	bp	bmi	ped	age
2.5%	-0.1520	-0.0232	0.0987	0.0250	0.1311
97.5%	0.0153	0.1624	0.2724	0.1781	0.3268