Introduction to Bayesian Data Analysis

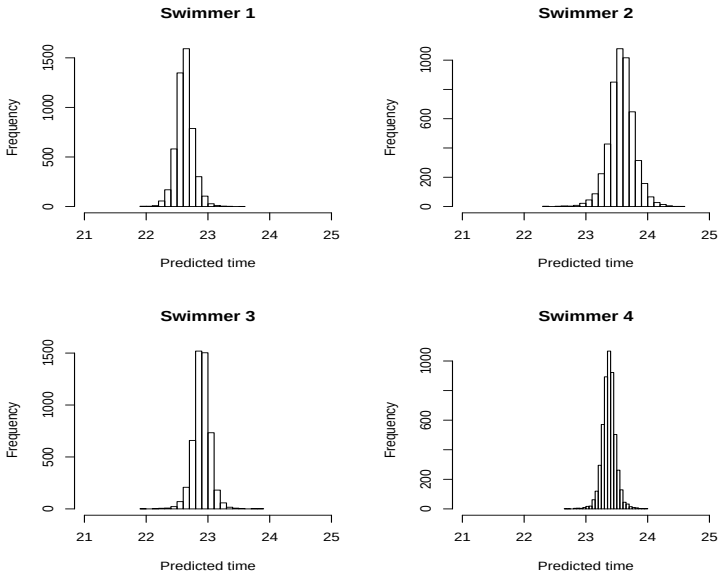
Tutorial 9 - Solutions

(1) (a) (i) Let x denote the week. The line $\beta_0 + \beta_1 x$ should produce values in the range (22,24) with high probability. With a little algebra, we can show that we need a prior distribution on (β_0, β_1) such that $21.6 < \beta_0 < 24.4$ with high probability and $-0.4 < \beta_1 < 0.4$ with high probability. This could be done by taking $\Sigma_{0,1,1} = 0.7^2$ and $\Sigma_{0,2,2} = 0.2^2$, for example. We also set $\boldsymbol{\beta}_0 = (23, 0)$. The R code is:

```
#LS each swimmer separately
X<-cbind(rep(1,6),c(1,2,3,4,5,6))
y1<-c(23.1,23.2,22.9,22.9,22.8,22.7)
y2<-c(23.2,23.1,23.4,23.5,23.5,23.4)
y3<-c(22.7,22.6,22.8,22.8,22.9,22.8)
y4<-c(23.7,23.6,23.7,23.5,23.5,23.4)
Y<-rbind(y1,y2,y3,y4)

####
#use Cholesky decomposition
rmvnorm<-function(n,mu,Sigma)
{ # samples from the multivariate normal distribution
  E<-matrix(rnorm(n*length(mu)),n,length(mu))
  t(  t(E%*%chol(Sigma)) +c(mu))
}

### store mcmc samples in these objects
S<-5000
beta.post<-matrix(nrow=S,ncol=p)
BETA<-list(beta.post,beta.post,beta.post,beta.post)
sigma2.post<-matrix(nrow=4,ncol=S)
y.pred<-matrix(nrow=4,ncol=S)
X.pred<-c(1,7)
### starting value
```

```
set.seed(1)
sigma2<- var( residuals(lm(y~0+X)) )

#semiconjugate prior distribution
### MCMC algorithm
for (i in 1:4){
y<-Y[i,]

n<-length(y)
fit.ls<-lm(y~-1+ X) #~-1 because X contains a column of intercept values
#prior based on information that competitive times for this age group are in
#the range 22 to 24 seconds
beta.0<-c(23,0) ; Sigma.0<-diag(c(0.7,0.2)^2,p)
nu.0<-1 ; sigma2.0<- summary(fit.ls)$sigma^2

### some convenient quantites
p<-length(beta.0)
iSigma.0<-solve(Sigma.0)
XtX<-t(X)%*%X
sigma2<- var( residuals(lm(y~0+X)) )

for( scan in 1:S) {
#update beta
V.beta<- solve(  iSigma.0 + XtX/sigma2 )
E.beta<- V.beta%*%( iSigma.0%*%beta.0 + t(X)%*%y/sigma2 )
beta<-t(rmvnorm(1, E.beta,V.beta) )

#update sigma2
nu.n<- nu.0+n
ss.n<-nu.0*sigma2.0 + sum(  (y-X%*%beta)^2 )
sigma2<-1/rgamma(1,nu.n/2, ss.n/2)

#save results of this scan
beta.post[scan,]<-beta
sigma2.post[i,scan]<-sigma2

y.pred[i,scan]<-X.pred%*%beta+rnorm(1,0,sqrt(sigma2))
                          }
BETA[[i]]<-beta.post

}
```

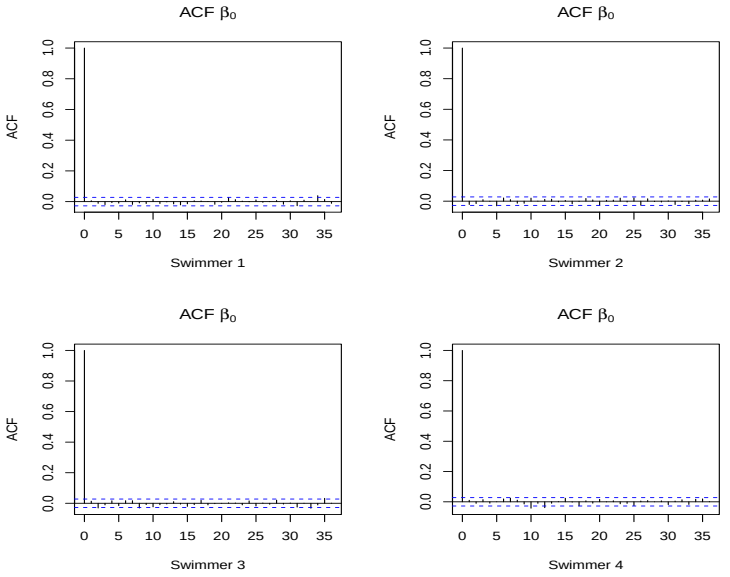
(ii) Histograms of the posterior predictive distribution of swim times two weeks from the last recorded time for each swimmer are below. From visual inspection of the histograms, the posterior predictive variance is the least for Swimmer 4, and it appears that either Swimmer 1 or Swimmer 3 are predicted to record the faster times.

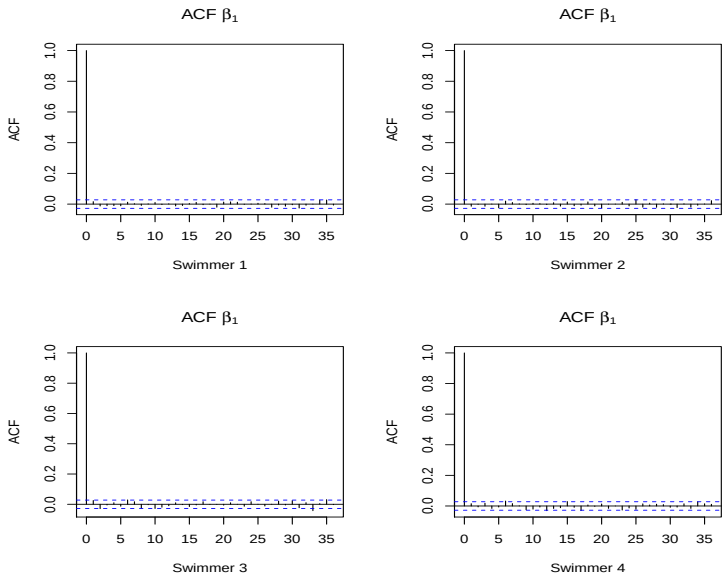


Autocorrelation plots and effective size calculations show now MCMC diagnostics to be concerned about.

```
> library(coda)
Loading required package: lattice
> for (i in 1:4){
+ effectiveSize(BETA[[i]][,1])
+ }
> ef_beta0<-NULL
> for (i in 1:4){
+ ef_beta0<-c(ef_beta0,effectiveSize(BETA[[i]][,1]))
+ }
> ef_beta0
    var1     var1     var1     var1
5000.000 5221.433 5172.701 5466.758
```

```
> ef_beta1<-NULL
> for (i in 1:4){
+ ef_beta1<-c(ef_beta1,effectiveSize(BETA[[i]][,2]))
+ }
> ef_beta1
    var1     var1     var1     var1
5000.000 5000.000 5042.192 5000.000
```





(b) Let's compute $Pr(Y_j^* = \min(\{Y_1^*, ..., Y_4^*\} \,|\mathbf{Y}))$ for each swimmer j.

```
min.pred<-NULL
for (i in 1:S){
   min.pred<-c(min.pred,which.min(y.pred[,i]))
}

pred.pr<-NULL
for (i in 1:4){
pred.pr<-c(pred.pr,mean(min.pred==i))
> pred.pr
[1] 0.9214 0.0008 0.0774 0.0004
}
```

$Pr(Y_j^* = \min(\{Y_1^*, ..., Y_4^*\} \,|\mathbf{Y}))$ is highest for $j = 1$. Therefore, we recommend to send swimmer 1 to the meet.

(2) (a) The R Code is:

```
diab<-read.table("azdiabetes.dat",header=T)
y<-diab$glu
n<-length(y)
X<-cbind(diab[,1],diab[,3:7])

#standardize
y<-(y-mean(y))/sd(y)

X<-t( (t(X)-apply(X,2,mean))/apply(X,2,sd))
X<-cbind(rep(1,length(y)),X)
tmp<-lm.gprior(y,as.matrix(X),g=dim(X)[1],nu0=2,s20=1)
#####  #g prior
beta.post<-tmp$beta
s2.post<-tmp$s2
apply(beta.post,2,function(x) quantile(x,c(0.025,0.975)))
beta.post<-tmp$beta
s2.post<-tmp$s2
```

Posterior confidence intervals are:

β_{npreg}	β_{bp}	β_{skin}	β_{bmi}	β_{ped}	β_{age}	σ^2
(-0.18, 0.026)	(-0.01,0.17)	(-0.04,0.17)	(0.03,0.25)	(0.04, 0.20)	(0.16,0.37)	(0.75,0.95)

(b) After performing the model averaging, we have the following estimates for $Pr(\beta_j \neq 0|\mathbf{Y})$ and posterior confidence intervals

```
> apply(BETA,2,function(x) mean(x!=0))
```

β_{npreg}	β_{bp}	β_{skin}	β_{bmi}	β_{ped}	β_{age}
0.0983	0.1790	0.0916	0.9833	0.7178	1.0000
(-0.11,0.00)	(0.00,0.13)	(0.00,1.12)	(0.09, 0.29)	(0.00,0.19)	(0.17, 0.35)

For β_j where $Pr(\beta_j \neq 0|\mathbf{Y}) > 0.5$, the corresponding posterior confidence interval in part (a) does not contain zero, and hence, the model averaging results agree with the results in part (a),