STA 360/602L: Module 5.2

Hierarchical normal models with constant variance: two groups (illustration)

Dr. Olanrewaju Michael Akande

No pre-recorded video for this module. To be done during discussion session.

Full conditionals recap

$$\mu | Y, \delta, \sigma^2 \sim \mathcal{N}(\mu_n, \gamma_n^2), \quad \text{where}$$

$$\gamma_n^2 = \frac{1}{\dfrac{1}{\gamma_0^2} + \dfrac{n_m + n_f}{\sigma^2}}$$

$$\mu_n = \gamma_n^2 \left[\frac{\mu_0}{\gamma_0^2} + \frac{\sum_{i=1}^{n_m} (y_{i,male} - \delta) + \sum_{i=1}^{n_f} (y_{i,female} + \delta)}{\sigma^2} \right].$$

Full conditionals

$$\delta | Y, \mu, \sigma^2 \sim \mathcal{N}(\delta_n, \tau_n^2), \quad \text{where}$$

$$\tau_n^2 = \frac{1}{\dfrac{1}{\tau_0^2} + \dfrac{n_m + n_f}{\sigma^2}}$$

$$\delta_n = \tau_n^2 \left[\frac{\delta_0}{\tau_0^2} + \frac{\sum_{i=1}^{n_m} (y_{i,male} - \mu) + (-1) \sum_{i=1}^{n_f} (y_{i,female} - \mu)}{\sigma^2} \right].$$

Full conditionals

$$\sigma^2 | Y, \mu, \delta \sim \mathcal{IG}(\frac{\nu_n}{2}, \frac{\nu_n \sigma_n^2}{2}), \quad \text{where}$$

$$\nu_n = \nu_0 + n_m + n_f$$

$$\sigma_n^2 = \frac{1}{\nu_n} \left[\nu_0 \sigma_0^2 + \sum_{i=1}^{n_m} (y_{i,male} - [\mu + \delta])^2 + \sum_{i=1}^{n_f} (y_{i,female} - [\mu - \delta])^2 \right].$$

Application to data

- The data we will use in the R package `rethinking`.

```
#install.packages(c("coda","devtools","loo","dagitty"))
#library(devtools)
#devtools::install_github("rmcelreath/rethinking",ref="Experimental")
#library(rethinking)
data(Howell1)

Howell1[1:15,]
```

```
##     height   weight  age male
## 1  151.765 47.82561 63.0    1
## 2  139.700 36.48581 63.0    0
## 3  136.525 31.86484 65.0    0
## 4  156.845 53.04191 41.0    1
## 5  145.415 41.27687 51.0    0
## 6  163.830 62.99259 35.0    1
## 7  149.225 38.24348 32.0    0
## 8  168.910 55.47997 27.0    1
## 9  147.955 34.86988 19.0    0
## 10 165.100 54.48774 54.0    1
## 11 154.305 49.89512 47.0    0
## 12 151.130 41.22017 66.0    1
## 13 144.780 36.03221 73.0    0
## 14 149.900 47.70000 20.0    0
## 15 150.495 33.84930 65.3    0
```

Application to data

- For now, focus on data for individuals under age 15.

```
htm <- Howell1$height/100
bmi <- Howell1$weight/(htm^2)
y_male <- bmi[Howell1$age<15 & Howell1$male==1]
y_female <- bmi[Howell1$age<15 & Howell1$male==0]
n_m <- length(y_male)
n_f <- length(y_female)

n_f
```

```
## [1] 84
```

```
n_m
```

```
## [1] 77
```

```
summary(y_male)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   12.07   13.87   14.63   14.84   15.53   18.22
```

```
summary(y_female)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   9.815  13.559  14.305  14.585  15.712  18.741
```

Application to data

- We will set the hyper-parameters as:
  - $\mu_0 = 15, \gamma_0 = 5,$
  - $\delta_0 = 0, \tau_0 = 3,$
  - $\nu_0 = 1, \sigma_0 = 5.$
- Do these values seem reasonable?

Application to data

```r
#priors
mu0 <- 15; gamma02 <- 5^2
delta0 <- 0; tau02 <- 3^2
nu0 <- 1; sigma02 <- 5^2

#starting values
mu <- (mean(y_male) + mean(y_female))/2
delta <- (mean(y_male) - mean(y_female))/2
#no need for starting values for sigma_squared, we can sample it first

MU <- DELTA <- SIGMA2 <- NULL
```

Application to data

```r
#set seed
set.seed(1234)

#set number of iterations and burn-in
n_iter <- 10000; burn_in <- 0.2*n_iter

##Gibbs sampler
for (s in 1:(n_iter+burn_in)) {
#update sigma2
sigma2 <- 1/rgamma(1,(nu0 + n_m + n_f)/2,
                   (nu0*sigma02 + sum((y_male-mu-delta)^2) + sum((y_female-mu+delta)^2))/2

#update mu
gamma2n <- 1/(1/gamma02 + (n_m + n_f)/sigma2)
mun <- gamma2n*(mu0/gamma02 + sum(y_male-delta)/sigma2 + sum(y_female+delta)/sigma2)
mu <- rnorm(1,mun,sqrt(gamma2n))

#update delta
tau2n <- 1/(1/tau02 + (n_m+n_f)/sigma2)
deltan <- tau2n*(delta0/tau02 + sum(y_male-mu)/sigma2 - sum(y_female-mu)/sigma2)
delta <- rnorm(1,deltan,sqrt(tau2n))

#save parameter values
MU <- c(MU,mu); DELTA <- c(DELTA,delta); SIGMA2 <- c(SIGMA2,sigma2)
}
```

Posterior summaries

```r
#library(coda)
MU.mcmc <- mcmc(MU,start=1)
summary(MU.mcmc)
```

```
## 
## Iterations = 1:12000
## Thinning interval = 1 
## Number of chains = 1 
## Sample size per chain = 12000 
## 
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
## 
##           Mean             SD       Naive SE Time-series SE 
##      14.712517       0.118765       0.001084       0.001089 
## 
## 2. Quantiles for each variable:
## 
##  2.5%   25%   50%   75% 97.5% 
## 14.48 14.63 14.71 14.79 14.95
```

```r
(mean(y_male) + mean(y_female))/2 #compare to data
```

```
## [1] 14.7127
```

Posterior summaries

```r
DELTA.mcmc <- mcmc(DELTA,start=1)
summary(DELTA.mcmc)
```

```
##
## Iterations = 1:12000
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 12000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##           Mean             SD       Naive SE Time-series SE
##       0.127657       0.119522       0.001091       0.001091
##
## 2. Quantiles for each variable:
##
##     2.5%      25%      50%      75%    97.5%
## -0.10691  0.04791  0.12743  0.20796  0.36407
```

```r
summary((2*DELTA)) #rescale as difference in group means
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -0.63464  0.09582  0.25487  0.25531  0.41592  1.23660
```

```r
mean(y_male) - mean(y_female) #compare to data
```

```
## [1] 0.2553392
```

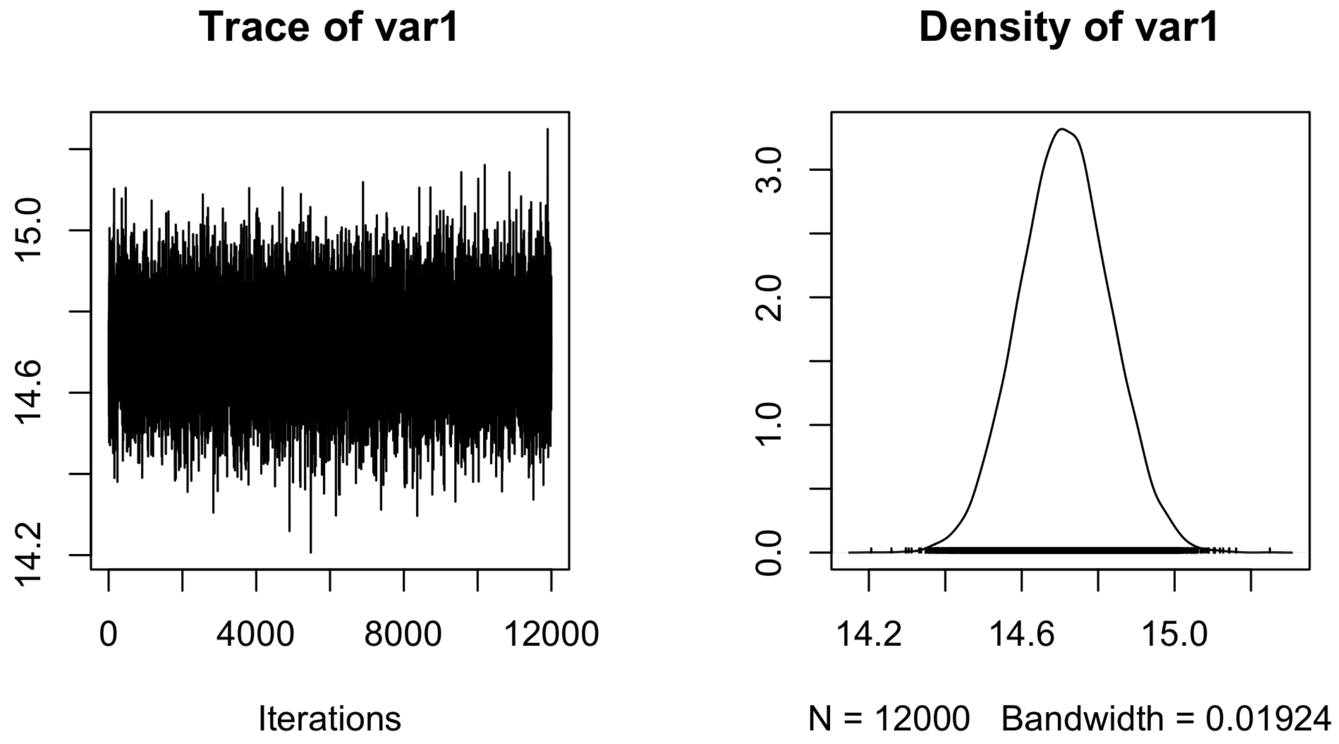
Posterior summaries

```
SIGMA2.mcmc <- mcmc(SIGMA2,start=1)
summary(SIGMA2.mcmc)
```

```
##
## Iterations = 1:12000
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 12000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##           Mean             SD       Naive SE Time-series SE
##       2.287927       0.257689       0.002352       0.002352
##
## 2. Quantiles for each variable:
##
##  2.5%   25%   50%   75% 97.5%
## 1.833 2.107 2.272 2.455 2.841
```

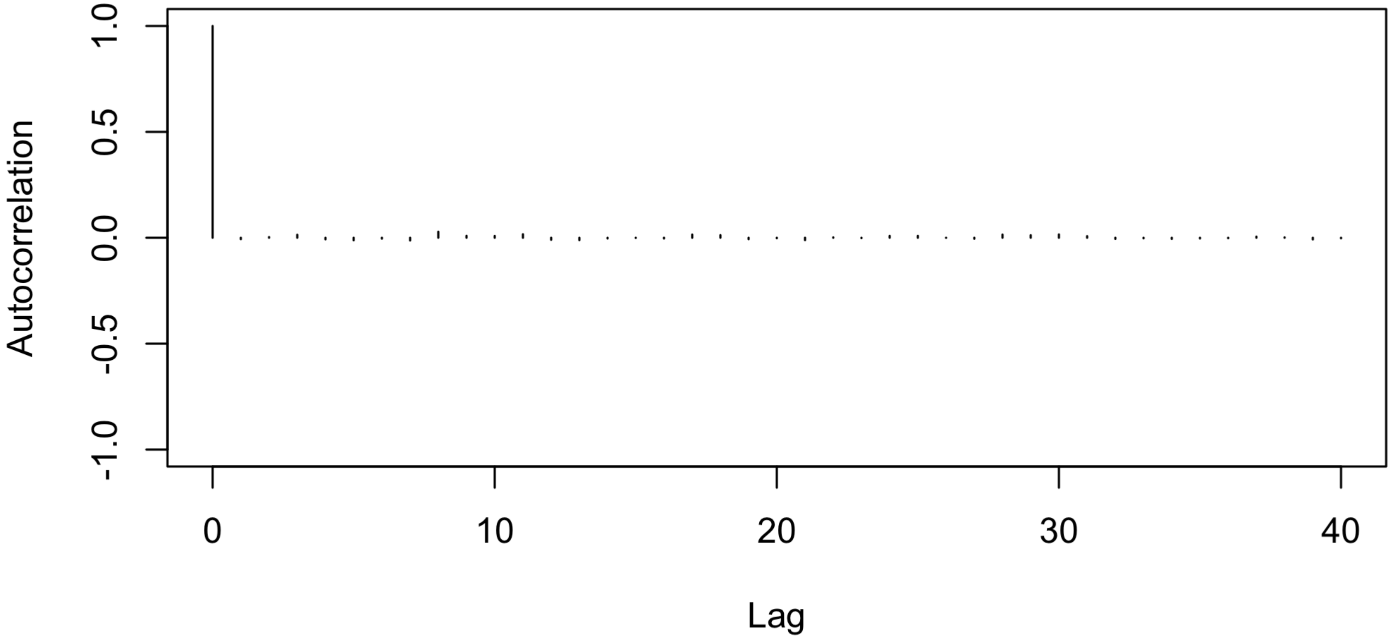
Diagnostics

```
plot(MU.mcmc)
```



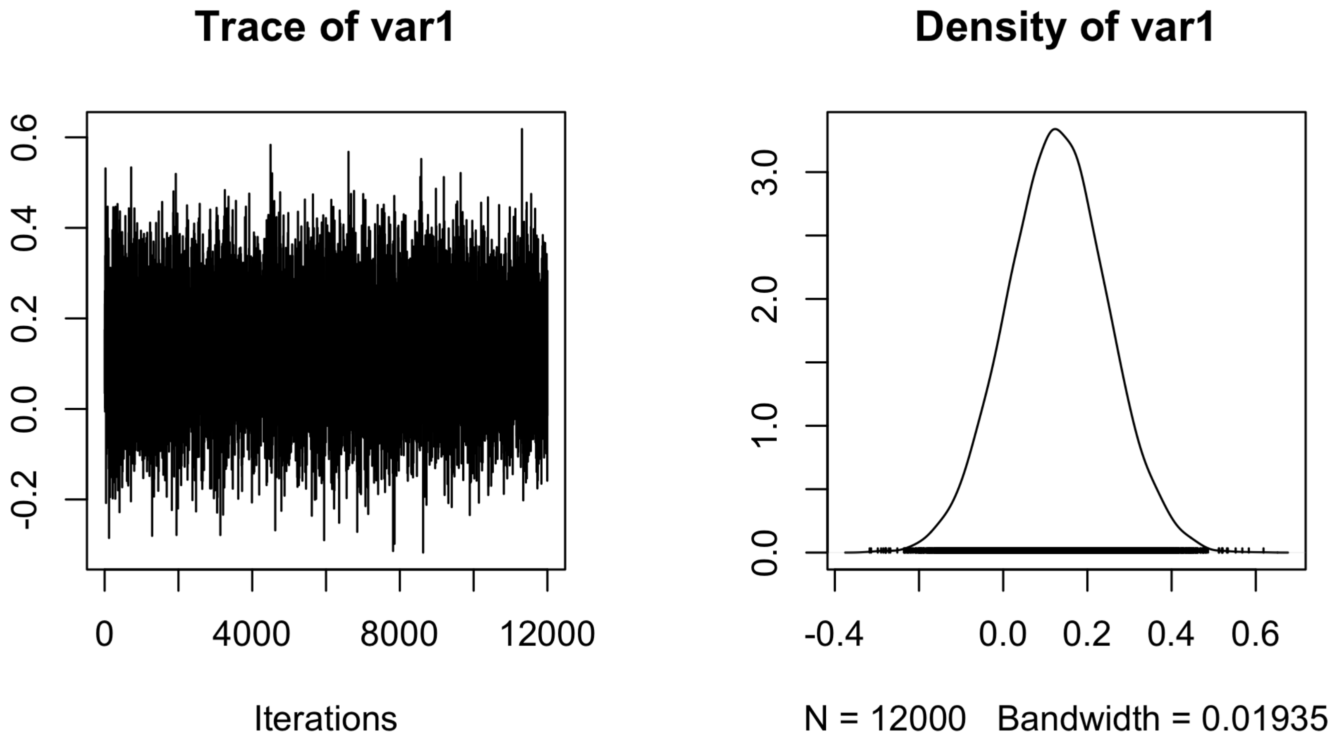
Diagnostics

```
autocorr.plot(MU.mcmc)
```



Diagnostics

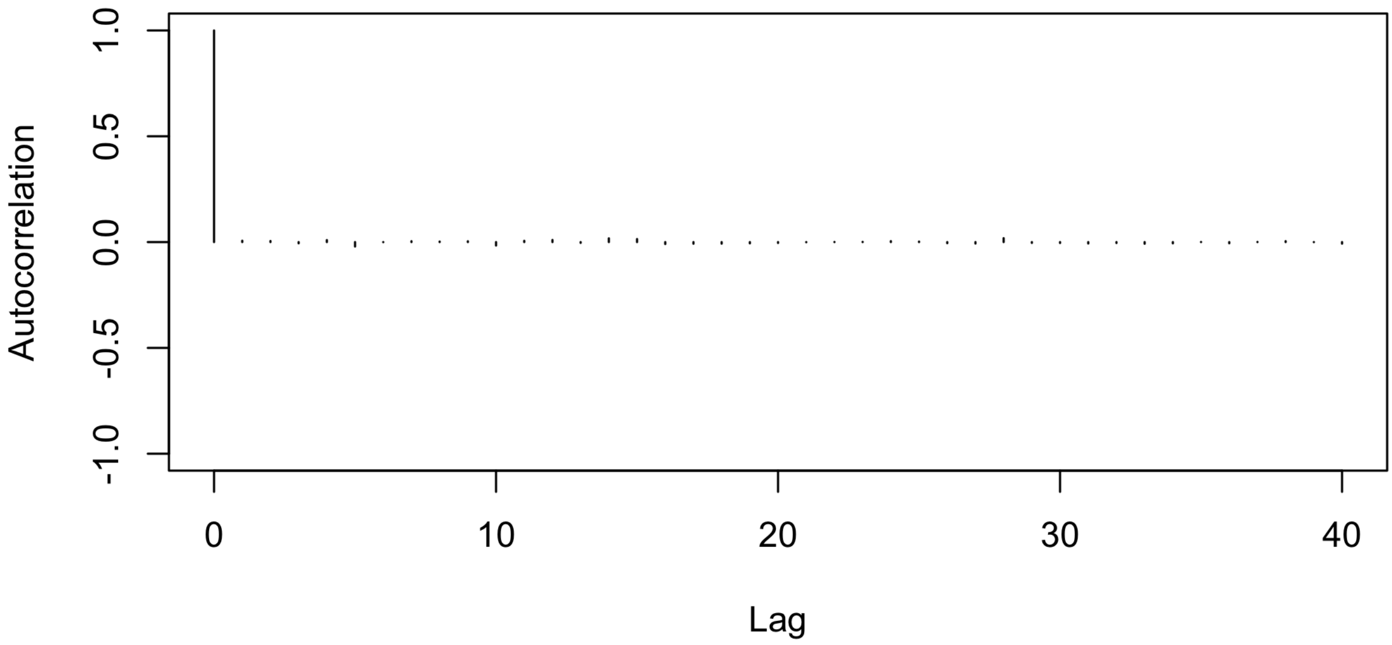
```r
plot(DELTA.mcmc)
```



STA 602L

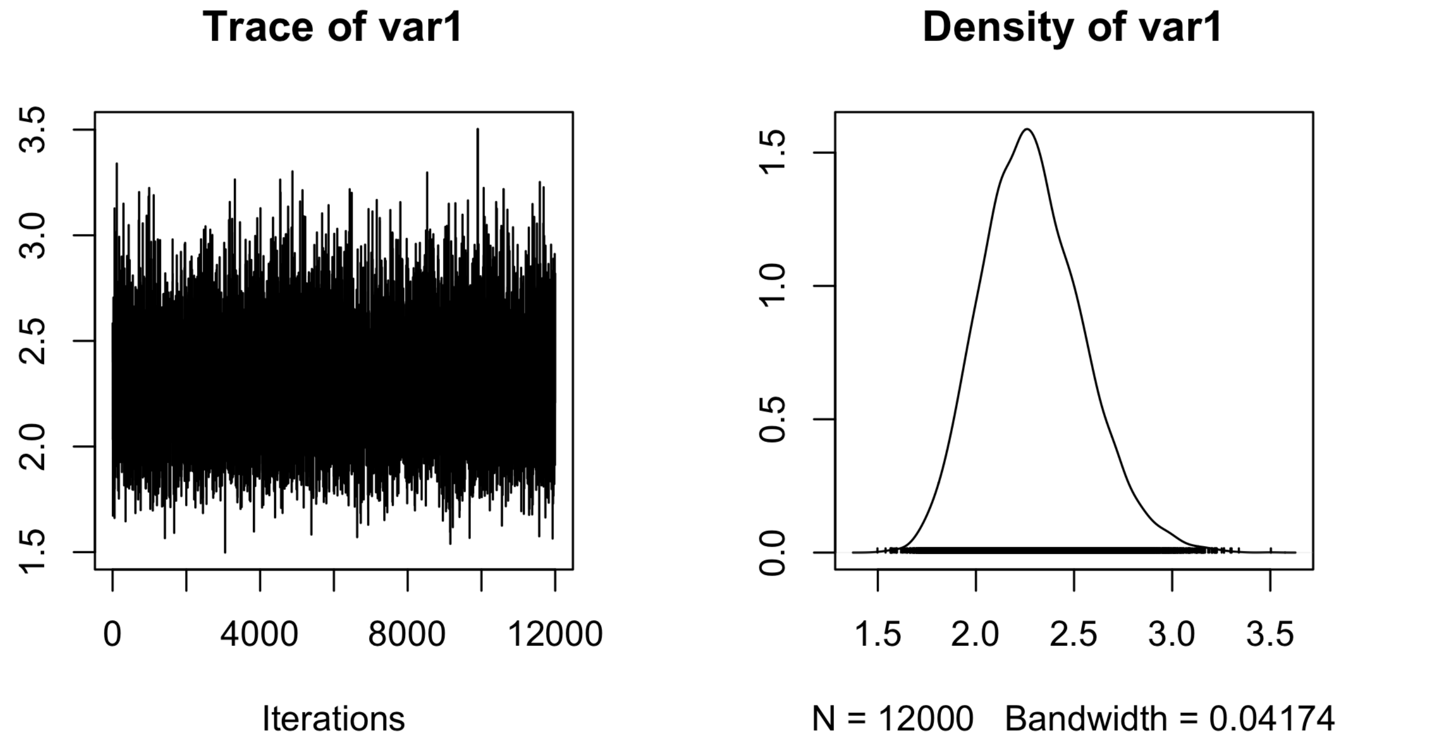
Diagnostics

```
autocorr.plot(DELTA.mcmc)
```



Diagnostics

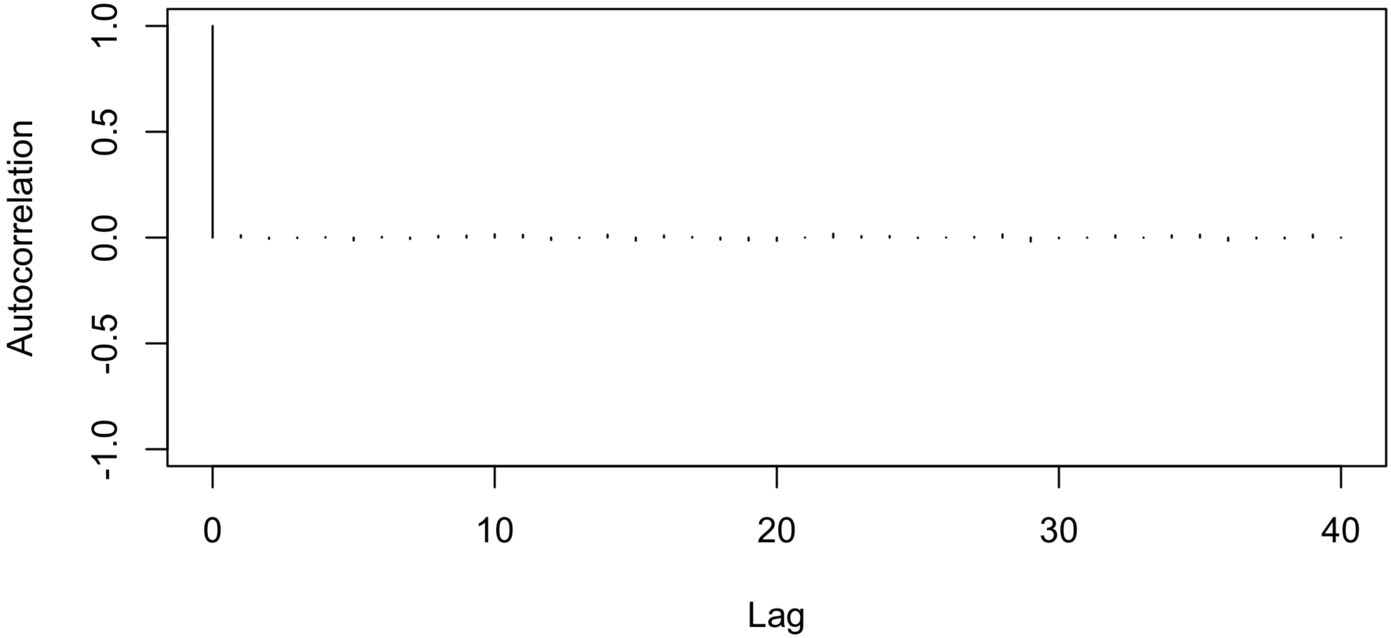
```
plot(SIGMA2.mcmc)
```



STA 602L

Diagnostics

```
autocorr.plot(SIGMA2.mcmc)
```



Application to data

- Posterior probability that boys have larger average BMI than girls is 0.86!
- Posterior medians and 95% credible intervals for the group means are actually quite similar to the unpooled gender specific intervals from classical inference (do a t-test to confirm).

```r
#mean for boys
quantile((MU+DELTA),probs=c(0.025,0.5,0.975))
```

```
##     2.5%      50%    97.5%
## 14.50255 14.84146 15.17925
```

```r
#mean for girls
quantile((MU-DELTA),probs=c(0.025,0.5,0.975))
```

```
##     2.5%      50%    97.5%
## 14.26848 14.58276 14.90761
```

```r
#posterior probability girls have larger BMI than boys
mean(DELTA > 0)
```

```
## [1] 0.8571667
```

Application to data

- Let's look at a different sub-population. For older individuals > 75, we only have 8 male and 4 female.

```
y_male <- bmi[Howell1$age > 75 & Howell1$male==1]
y_female <- bmi[Howell1$age > 75 & Howell1$male==0]
n_m <- length(y_male)
n_f <- length(y_female)
n_m
```

```
## [1] 8
```

```
n_f
```

```
## [1] 4
```

Application to data

- A 95% confidence interval for the difference between genders in BMI (estimated as 0.24) is (-4.20,4.68).

```
mean(y_male) - mean(y_female)
```

```
## [1] 0.2408966
```

```
t.test(y_male,y_female)
```

```
## 
##      Welch Two Sample t-test
## 
## data:  y_male and y_female
## t = 0.13801, df = 5.1869, p-value = 0.8954
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4.197948  4.679741
## sample estimates:
## mean of x mean of y 
##  18.06751  17.82662
```

Application to data

- Let's apply the Bayesian model with these priors:
  - $\mu_0 = 18, \gamma_0 = 5,$
  - $\delta_0 = 0, \tau_0 = 3,$
  - $\nu_0 = 1, \sigma_0 = 5.$
- The R code for running the sampler is suppressed here. Basically, just re-run the same Gibbs sampler from before on this new data.
- Using the results from the model, the posterior mean is 0.25 with 95% CI (-3.45, 3.88).

```
mean((DELTA*2))
```

```
## [1] 0.2493733
```

```
quantile((DELTA*2),probs=c(0.025,0.5,0.975))
```

```
##       2.5%         50%      97.5%
## -3.4466931  0.2758598  3.8762543
```

Application to data

- The width of this interval is smaller than that of the 95% confidence interval from before.

- In a way, precision has been improved by borrowing of information across the groups. Of course the prior is important here given the sample sizes.

What's next?

Move on to the readings for the next module!