Homework 6

Thomas Fleming- Lab time: 1:25

October 27, 2016

7.4
a)

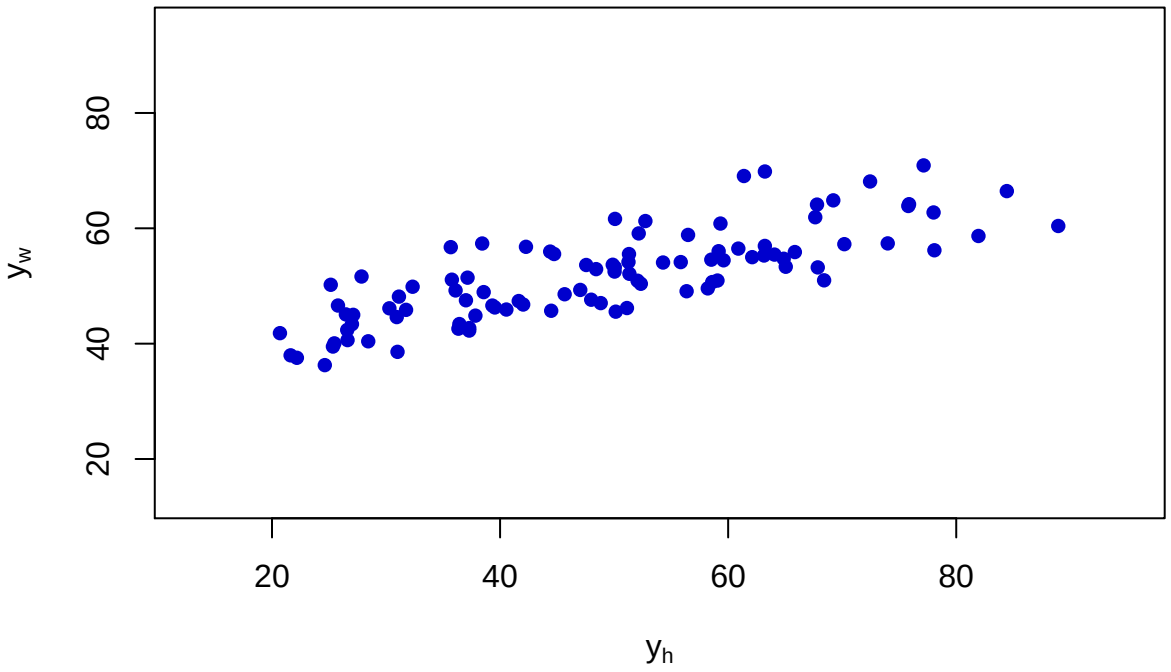
```
library(mvtnorm)
```

```
## Warning: package 'mvtnorm' was built under R version 3.1.2
```

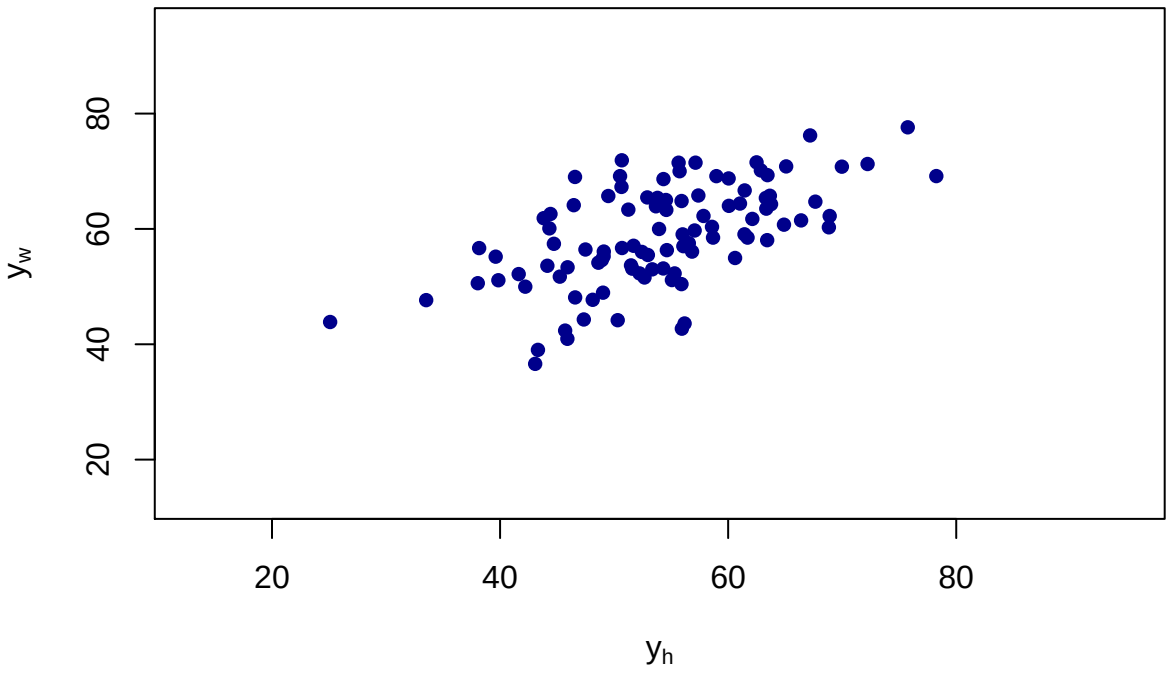
```
set.seed(1)

mu.0 <- c(55, 52)
nu.0 <- 4
lambda.0 <- matrix(c(100, 50, 50, 100), nrow = 2, ncol = 2)
sigma.0 <- matrix(c(256, 166.4, 166.4, 256), nrow = 2, ncol = 2)
```

b)

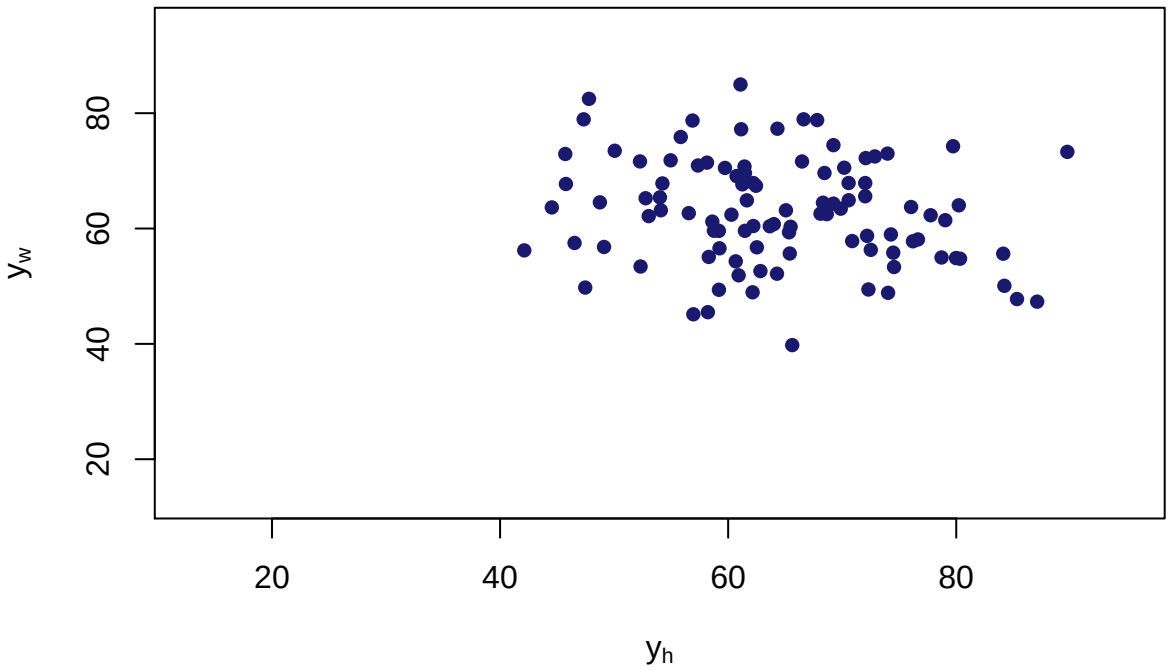
```
n <- 100
#1
y.1 <- matrix(NA, nrow = n, ncol = 2)
theta.pred.1 <- rmvnorm(1, mu.0, lambda.0)
sigma.pred.1 <- solve(rWishart(1, nu.0, solve(sigma.0))[, , 1])
y.1 <- rmvnorm(n, theta.pred.1, sigma.pred.1)
plot(y.1[,1], y.1[,2], type = "p", pch = 16,
xlab = expression(y[h]), ylab = expression(y[w]),
xlim = c(13, 95), ylim = c(13, 95), col = "mediumblue")
```



```r
#2
y.2 <- matrix(NA, nrow = n, ncol = 2)
theta.pred.2 <- rmvnorm(1, mu.0, lambda.0)
sigma.pred.2 <- solve(rWishart(1, nu.0, solve(sigma.0))[, , 1])
y.2 <- rmvnorm(n, theta.pred.2, sigma.pred.2)
plot(y.2[,1], y.2[,2], type = "p", pch = 16,
xlab = expression(y[h]), ylab = expression(y[w]),
xlim = c(13, 95), ylim = c(13, 95), col = "darkblue")
```




```r
#3
y.3 <- matrix(NA, nrow = n, ncol = 2)
theta.pred.3 <- rmvnorm(1, mu.0, lambda.0)
sigma.pred.3 <- solve(rWishart(1, nu.0, solve(sigma.0))[, , 1])
y.3 <- rmvnorm(n, theta.pred.3, sigma.pred.3)
plot(y.3[,1], y.3[,2], type = "p", pch = 16,
xlab = expression(y[h]), ylab = expression(y[w]),
xlim = c(13, 95), ylim = c(13, 95), col = "midnightblue")
```

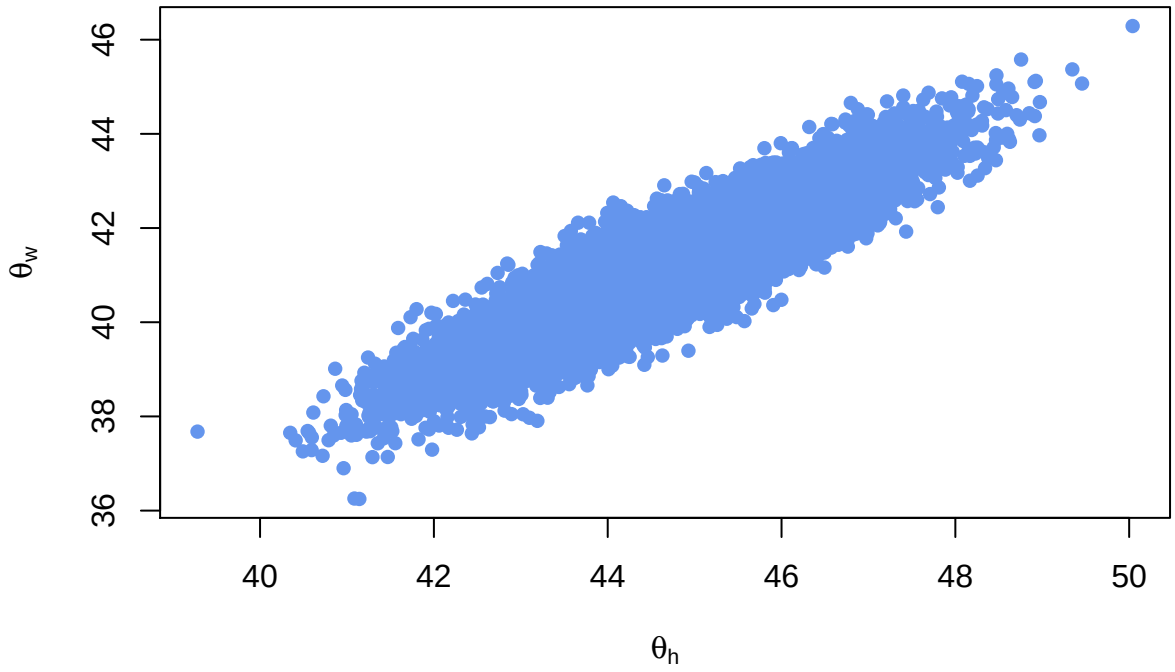


The prior distribution I chose is theta ~ mvnorm(mu.0, lambda.0) where mu.0 = 55, 52 and lambda.0 = 256, 166.4, 166.4, 256 and sigma ~ inv.wishart(nu.0, (sigma.0)^-1) where nu.0 = 4 and sigma.0 = 256, 166.4, 166.4, 256. While the theta's and covariance matrix I picked do not fully reflect all the nuances of what I would envision the true plot to look like, it is adequate (especially the first data set). I expect husband's age to be correlated with the wife's age, with a small average difference between the two. However, the distribution should also account for deviations from the norm, such as a 60 year old married to a 20 year old, which is not always reflected, depending on the sample. Nevertheless, there seems to be high variability in scatter plots from sample to sample, at least when we consider samples of n = 100, and samples I have experimented with- those having a larger n- have reflected these nuances.

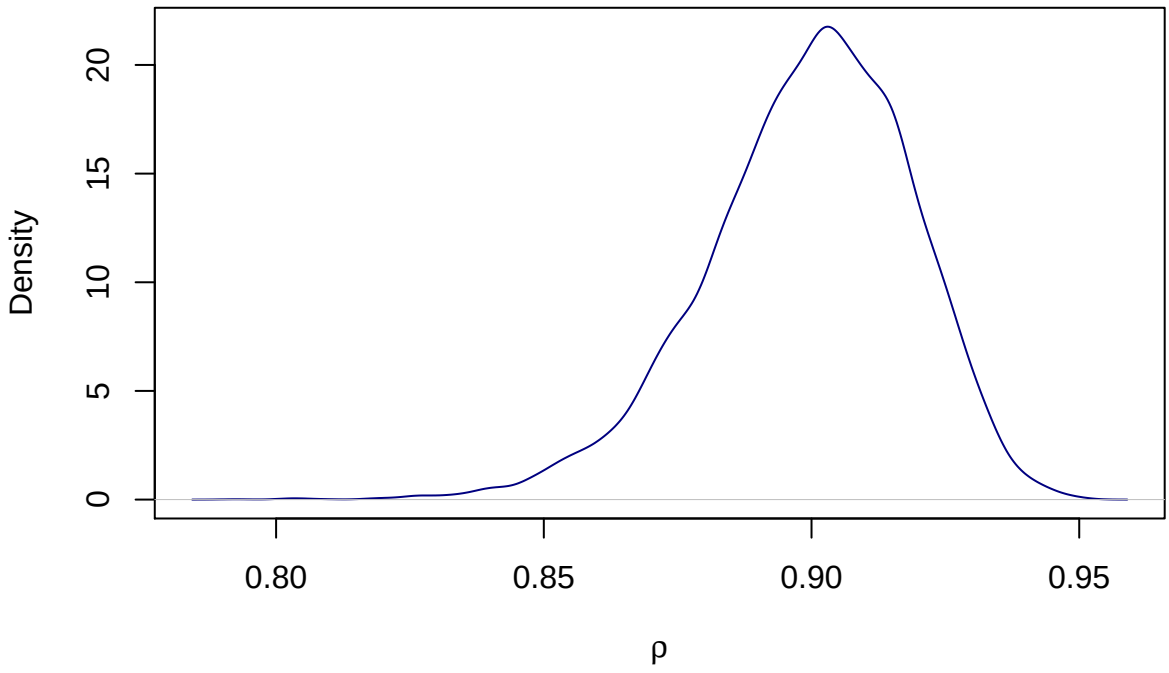
c)

```
setwd("~/Downloads")
data <- read.table("agehw.dat", header = TRUE)
n.y <- dim(data)[1]
y.bar <- apply(data, 2, mean)
sigma.y <- cov(data)
THETA <- SIGMA <- NULL

I <- 10000

for(i in 1:I) {
  #update theta
  lambda.n <- solve(solve(lambda.0) + n.y*solve(sigma.y))
  mu.n <-
    lambda.n%*%(solve(lambda.0)%*%mu.0 + n.y*solve(sigma.y)%*%y.bar)
  theta.n <- rmvnorm(1, mu.n, lambda.n)

  #updata sigma
  S.n <- sigma.0 + (t(data) - c(theta.n))%*%t(t(data) - c(theta.n))
  sigma.n <- solve(rWishart(1, nu.0 + n.y, solve(S.n))[, , 1])

  #store results
```

```
  THETA <- rbind(THETA, theta.n)
  SIGMA <- rbind(SIGMA, c(sigma.n))
}

plot(THETA[, 1], THETA[, 2], type = "p", pch = 16,
xlab = expression(theta[h]), ylab = expression(theta[w]),
col = "cornflowerblue")
```




```
plot(density(SIGMA[, 2]/sqrt(SIGMA[, 1]*SIGMA[, 4])),
main = "", col = "navyblue", xlab = expression(rho))
```




```
quantile(THETA[,1], c(.025, .975))
```

```
##     2.5%    97.5%
## 42.01336 47.31710
```

```
quantile(THETA[,2], c(0.025, .975))
```

```
##     2.5%    97.5%
## 38.64921 43.62607
```

```
quantile(SIGMA[, 2]/sqrt(SIGMA[, 1]*SIGMA[, 4]), c(.025, .975))
```

```
##      2.5%     97.5%
## 0.8564385 0.9319188
```

d)

```
mu.0.d <- c(0, 0)
lambda.0.d <- diag(x = 10^5, nrow = 2, ncol = 2)
nu.0.d <- 3
sigma.0.d <- diag(x = 1000, nrow = 2, ncol = 2)

THETA.d <- SIGMA.d <- NULL

I <- 10000

for(i in 1:I) {
  #update theta
```

```
  lambda.n.d <- solve(solve(lambda.0.d) + n.y*solve(sigma.y))
  mu.n.d <-
    lambda.n.d%*%(solve(lambda.0.d)%*%mu.0.d + n.y*solve(sigma.y)%*%y.bar)
  theta.n.d <- rmvnorm(1, mu.n.d, lambda.n.d)

  #updata sigma
  S.n.d <-
    sigma.0.d + (t(data) - c(theta.n.d))%*%t(t(data) - c(theta.n.d))
  sigma.n.d <- solve(rWishart(1, nu.0.d + n.y, solve(S.n.d))[, , 1])

  #store results
  THETA.d <- rbind(THETA.d, theta.n.d)
  SIGMA.d <- rbind(SIGMA.d, c(sigma.n.d))
}

quantile(THETA.d[,1], c(.025, .975))
```

```
##     2.5%    97.5%
## 41.75004 47.05859
```

```
quantile(THETA.d[,2], c(0.025, .975))
```

```
##     2.5%    97.5%
## 38.39662 43.38770
```

```
quantile(SIGMA.d[, 2]/sqrt(SIGMA.d[, 1]*SIGMA.d[, 4]), c(.025, .975))
```

```
##      2.5%     97.5%
## 0.7916734 0.8998533
```

e) In comparing the two posteriors, it seems that the results are very similar. This may be due to the fact that the values for nu.0 and S.0 in both priors make it so the distribution is loosely centered around the prior covariance matrix. The biggest difference is that the 95% confidence interval in c) includes a higher range of correlation values, and is also slightly more concentrated, making us more certain of where the true correlation lies. Because of the narrower confidence interval and, subjectively, more realistic range of correlational values, the prior in c) should be preferred, albeit slightly.

If we were to choose a smaller sample of n = 25, this would probably favor the prior from c), since the variances of its lambda.0 and sigma.0 are much lower at 256. The prior from d) has a lambda with variance 10^5 and sigma with variance 1000. Since we would be taking less samples, there would be more opportunity for erratic results from d) because of its large variance, as well as a wider confidence interval- making the prior from c) more helpful in this case.