Age and Marriage Bayesian Analysis

Rory Quinlan

```
library(MASS)
library(ggplot2)
library(kableExtra)
```

```
agehw <- read.table("C:\\Users\\roryq\\Downloads\\agehw.dat", header = T)
nrow(agehw)
```

```
## [1] 100
```

Descriptive Statistics

```
# Table Descriptive stat
`Mean Age of Wives`<-c(mean(agehw$agew))
`Mean Age of Husbands`<-c(mean(agehw$ageh))
table<-rbind(`Mean Age of Wives`,`Mean Age of Husbands`)
t(table) %>%
  kbl() %>%
  kable_styling()
```

Mean Age of Wives	Mean Age of Husbands
40.89	44.42

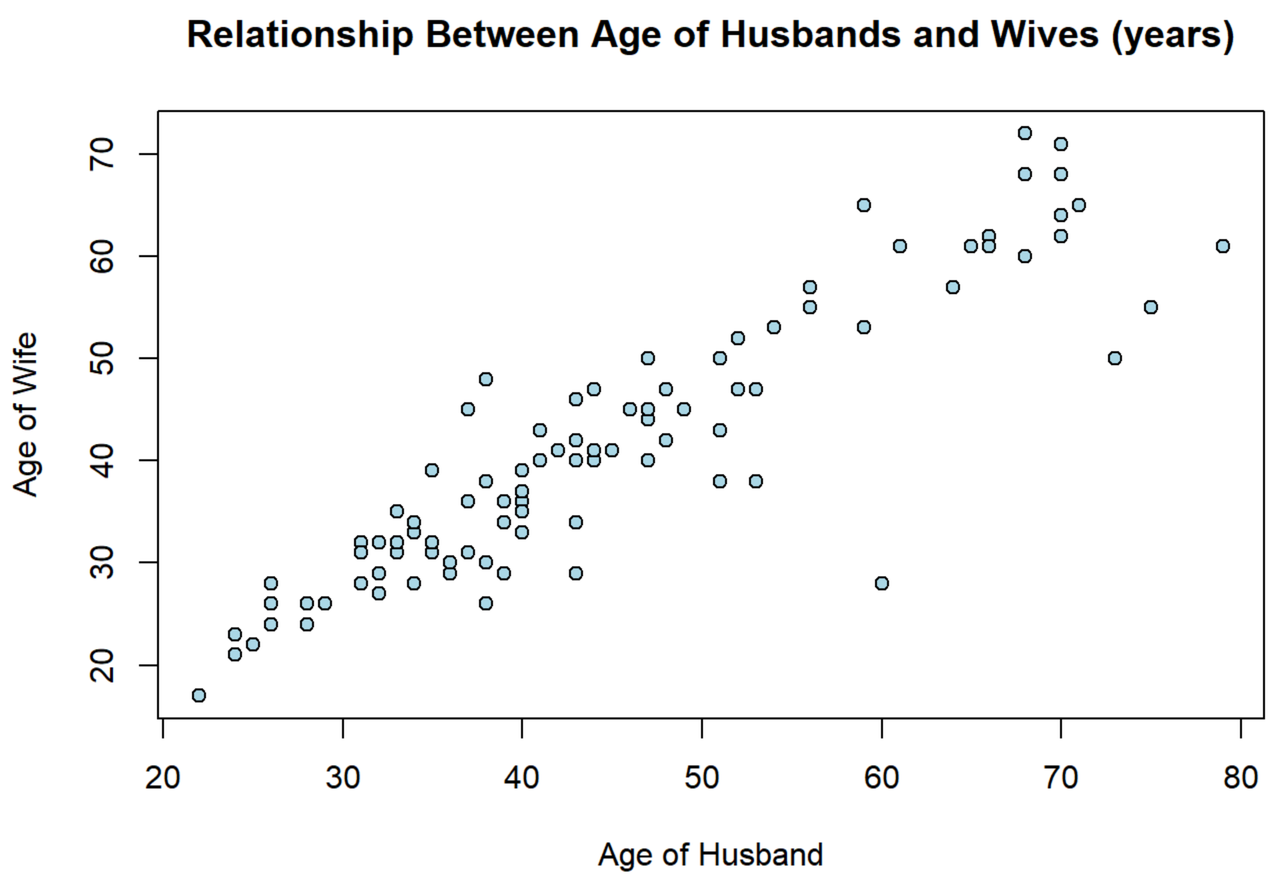
```
`Standard Deviation for Age of Wives`<-c(sd(agehw$agew))
`Standard Deviation for Age of Husbands`<-c(sd(agehw$ageh))
table<-rbind(`Standard Deviation for Age of Wives`,`Standard Deviation for Age of Husbands`)
t(table) %>%
  kbl() %>%
  kable_styling()
```

Standard Deviation for Age of Wives	Standard Deviation for Age of Husbands
12.80064	13.63239

```
# Table covariance
cov(agehw) %>%
  kbl() %>%
  kable_styling()
```

	ageh	agew
ageh	185.8420	157.6729
agew	157.6729	163.8565

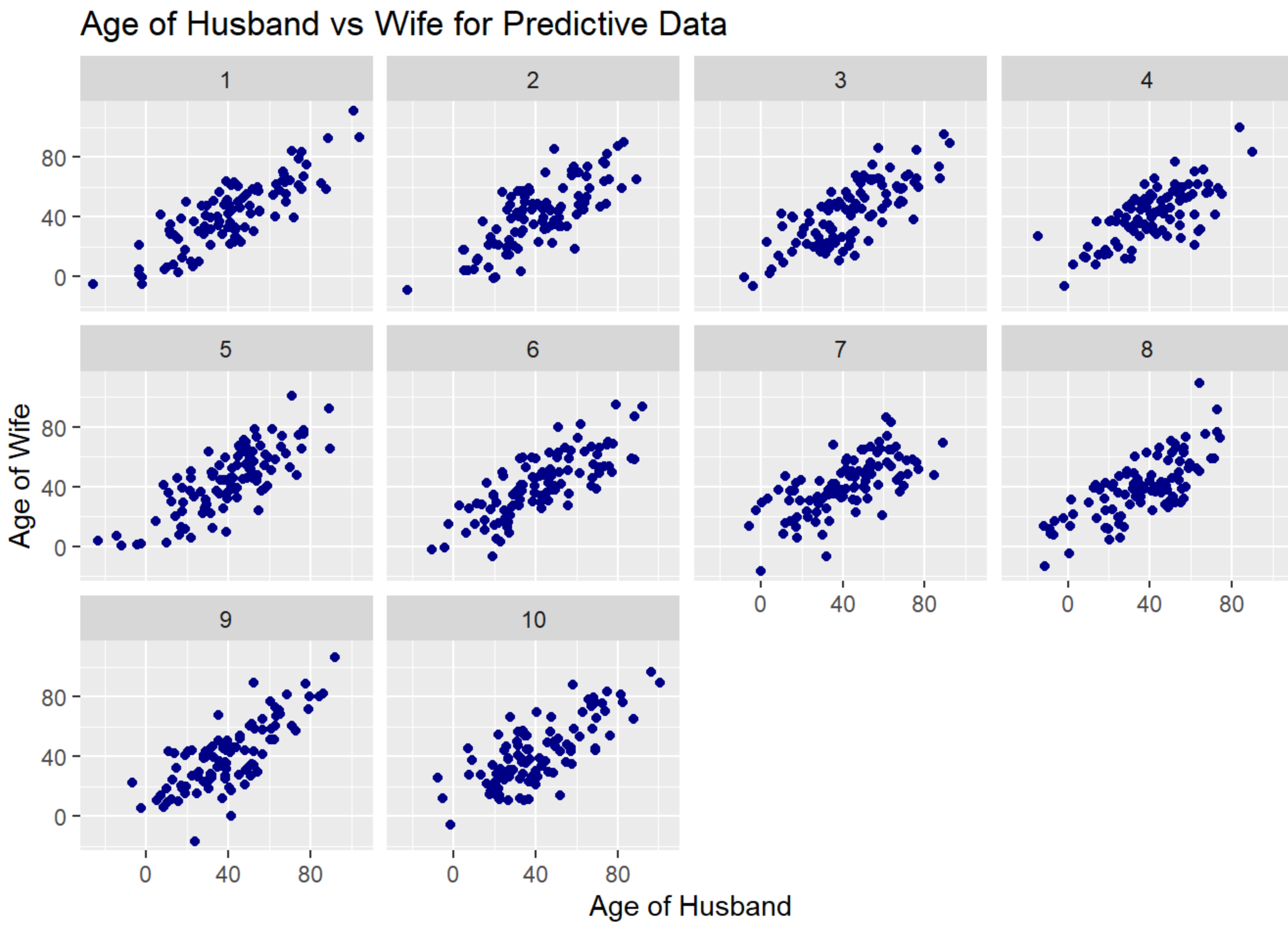
```
# Visualization of relationship between husbands and wives

plot(x=agehw$ageh, y= agehw$agew, pch=21, xlab="Age of Husband", ylab="Age of Wife", main="Relat
ionship Between Age of Husbands and Wives (years)", col="black",bg="lightblue")
```



Bayesian Analysis

```
# Generate predictive data set (size 100) from average ages (theta) and cov matrix (sigma)



n = 100
s = 10
# Mean of ages
mu0 <- c(42,42)
L0 <- matrix(c(441, 330.75, 330.75, 441), nrow = 2, ncol = 2)
# Mean ages follow multivariate norm
theta <- mvrnorm(s, mu0, L0)

# sample sigmas following inverse wishart distribution
sigmas <- list()
for(i in 1:s){
 sigma <- solve(rWishart(1, 4, solve(L0))[,,1])
 sigmas[[i]] <- sigma
}

# Sample age from multivariate normal distribution
data <- data.frame(h_age= c(), w_age = c(), dataset= c())
for(i in 1:s){
 y <- mvrnorm(100, mu0, L0)
 new <- data.frame(h_age = y[,1], w_age = y[,2], dataset =i)
 data <- rbind(data, new)

}
```

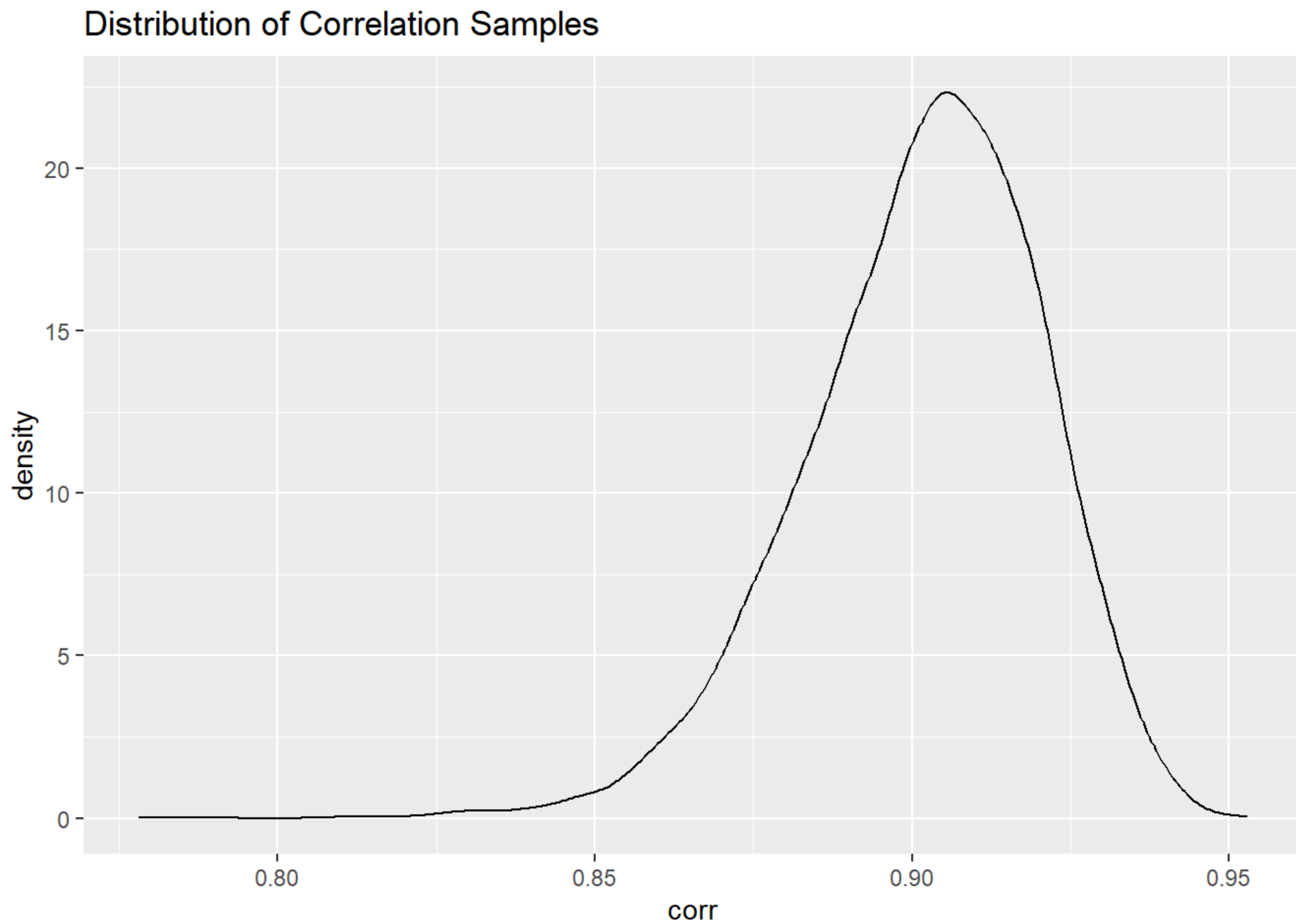
```
# Plot the 10 predictive datasets on scatter plot to see if it matches the trend of the original
and confirm our selection of prior (inverse whishart for sigma and multivariate normal for thet
a)
ggplot(data = data, aes(x = h_age, y = w_age)) + geom_point(color='darkblue') + facet_wrap(~data
set)+ labs(x=" Age of Husband", y="Age of Wife", title="Age of Husband vs Wife for Predictive Da
ta")
```



There appears to be a moderately strong positive correlation between age of spouses for each of the predictive data sets. This is the expected relationship and reflects the one we observed from the real data. The similarity confirms our selections of priors for the mean and covariance of ages.

```r
# MCMC Approximations

#prior
mu0 <- c(42,42)
nu0 <- 4
L0 <- S0 <- matrix(c(150, 112.5, 112.5, 150), nrow = 2, ncol = 2)
ybar <- apply(agehw, 2, mean)
Sigma <- cov(agehw)
n <- dim(agehw)[1]
THETA<-SIGMA <- NULL
set.seed(10000)
for(s in 1:10000){

#Update theta
 Ln <- solve(solve(L0) + n * solve(Sigma))
 mun <- Ln %*% (solve(L0) %*% mu0 + n * solve(Sigma) %*% ybar)
 theta <- mvrnorm(1, mun, Ln)

#Update Simga
 Sn <- S0 + (t(agehw) - c(theta)) %*% t( t(agehw) - c(theta))
 Sigma <- solve( rWishart(1, nu0 + n, solve(Sn))[,,1])

# save results

 THETA <- rbind(THETA, theta) ; SIGMA <- rbind(SIGMA, c(Sigma))
}
```

```r
# Make MCMC correlations into dataframe and plot
cov <- SIGMA[,2]
var_h <- SIGMA[,1]
var_w <- SIGMA[,4]
corr <- cov/sqrt(var_h*var_w)
corr <- data.frame(corr)

# Plot Sampled Correlation
ggplot(data = corr, aes(x = corr)) + geom_density()+labs(title="Distribution of Correlation Samp
les")
```




```
# Print Confidence intervals for descriptive statistics

`Average Age for Husband CI`<- c(quantile(THETA[,1],.05),quantile(THETA[,1],.95))
`Average Age of Wives CI`<- c(quantile(THETA[,2],.05),quantile(THETA[,2],.95))
`Correlation Coefficient CI`<-c(quantile(corr[,1], c(.05, .95)))

bt<- as.data.frame(rbind(`Average Age of Wives CI`,`Average Age for Husband CI`,`Correlation Coe
fficient CI`))

t(bt) %>%
  kbl() %>%
  kable_styling()
```

	Average Age of Wives CI	Average Age for Husband CI	Correlation Coefficient CI
5%	38.80692	42.2129	0.8684954
95%	42.98145	46.6178	0.9289617

Frequentist Analysis

```
result <- t.test(agehw$agew)
# Extract the confidence interval
`Avg Age of Wives CI` <- result$conf.int

result1 <- t.test(agehw$ageh)
# Extract the confidence interval
`Avg Age of Husbands CI` <- result1$conf.int

# Correlations
`fcor` <- cor.test(~ ageh + agew, data = agehw)
`correlation Coefficient CI`<-`fcor`$conf.int


ft<-as.data.frame(rbind(`Avg Age of Wives CI`, `Avg Age of Husbands CI`,`correlation Coefficient
CI`))

t(ft) %>%
  kbl() %>%
  kable_styling()
```

	Avg Age of Wives CI	Avg Age of Husbands CI	correlation Coefficient CI
V1	38.35007	41.71504	0.8597107
V2	43.42993	47.12496	0.9341782

Results

```
# Improvement with bayesian analysis

Improve<-as.data.frame(cbind(t(bt),t(ft)))

Improve$Improvement_Correlation<-abs(Improve$`Correlation Coefficient CI`-Improve$`correlation C
oefficient CI`)

Improve$Wife_Improvement<- abs(Improve$`Average Age of Wives CI`-Improve$`Avg Age of Wives CI`)

Improve$Husband_Improvement<- abs(Improve$`Average Age for Husband CI`-Improve$`Avg Age of Husba
nds CI`)

Improve[,7:9] %>%
  kbl() %>%
  kable_styling()
```

	Improvement_Correlation	Wife_Improvement	Husband_Improvement
5%	0.0087847	0.4568479	0.4978602

	Improvement_Correlation	Wife_Improvement	Husband_Improvement
95%	0.0052165	0.4484796	0.5071584

The limited sample size and no prior information about the variables in question population leads to a relatively wide confidence interval using standard frequentist approaches. Using the bayesian approach, the confidence interval for average age of husbands and wives was reduced by over a year. And the confidence interval for the correlation coefficient was reduced by 2%, compared to the frequentist.

After visualizing the relationship we formed priors for the 2 variables we are interested in predicting. Correlation is 2 dimensional so we formulated it following an inverse wishart distribution, and the mean age for spouses follows a multivariate normal distribution.

Once we established our priors, we can generate more data by taking random samples of these distributions. We confirm our priors and data validity by comparing simulated data scatter plots to the actual data scatter plots. The moderately strong positive correlation holds in each predictive data set to the original.

After confirming distributions, we use MCMC approximation to estimate the mean correlation, age of husband, and age of wife. With the prior information about the distribution of the data we are able to create new confidence intervals for average ages, and correlations.