1. Exercise 7.1 (pg 238) Jeffreys' prior: For the multivariate normal

model, Jeffreys' rule for generating a prior distribution on (θ, Σ) gives $p_J(\theta, \Sigma) \propto |\Sigma|^{-(p+2)/2}$.

a) Explain why the function p_J cannot actually be a probability density for (θ, Σ).

Since the density is uniform with respect to $\boldsymbol{\theta}$, the integral over the support of this function is infinite and cannot be 1.

b) Let $p_J(\theta, \Sigma|y_1, ..., y_n)$ be the probability density that is proportional

to $p_J(\theta, \Sigma) \times p(y_1, ..., y_n|\theta, \Sigma)$.Obtain the form of $p_J(\theta, \Sigma|y_1, ..., y_n), p_J(\theta|\Sigma, y_1, ..., y_n)$ and $p_J(\Sigma|y_1, ..., y_n)$.

$$\begin{aligned} p_J(\boldsymbol{\theta}, \Sigma \mid \boldsymbol{y}_{1:n}) &\propto p(\boldsymbol{\theta}, \Sigma) \times p(\boldsymbol{y}_{1:n} \mid \boldsymbol{\theta}, \Sigma) \\ &\propto \left(|\Sigma|^{-\frac{p+2}{2}}\right) \times \left(|\Sigma|^{-\frac{n}{2}} \exp\left[-\frac{1}{2}\text{tr}(\mathbf{S}_\theta \Sigma^{-1})\right]\right) \\ &\propto |\Sigma|^{-\frac{n+p+2}{2}} \exp\left[-\frac{1}{2}\text{tr}(\mathbf{S}_\theta \Sigma^{-1})\right] \end{aligned}$$

To obtain the full conditionals of a parameter, we treat the other parameters as constant, so

$$\begin{aligned} p_J(\boldsymbol{\theta} \mid \Sigma, \boldsymbol{y}_{1:n}) &\propto \exp\left[-\frac{1}{2}\text{tr}(\mathbf{S}_\theta \Sigma^{-1})\right] \\ &= \exp\left[-\frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{y}_i - \theta)'\Sigma^{-1}(\boldsymbol{y}_i - \theta)\right] \\ &= \exp\left[-\frac{n}{2}(\bar{\boldsymbol{y}} - \theta)'\Sigma^{-1}(\bar{\boldsymbol{y}} - \theta)\right] \\ \boldsymbol{\theta} \mid \Sigma, \boldsymbol{y}_{1:n} &\sim \text{Normal}(\bar{\boldsymbol{y}}, \Sigma/n) \\ p_J(\Sigma \mid \boldsymbol{\theta}, \boldsymbol{y}_{1:n}) &\propto |\Sigma|^{-\frac{n+p+2}{2}} \exp\left[-\frac{1}{2}\text{tr}(\mathbf{S}_\theta \Sigma^{-1})\right] \\ \Sigma \mid \boldsymbol{\theta}, \boldsymbol{y}_{1:n} &\sim \text{Inverse-Wishart}\left(n+1, \mathbf{S}_\theta^{-1}\right) \end{aligned}$$

2. Exercise 7.2 (pg 238) Unit information prior

Letting $\Psi = \Sigma^{-1}$, show that a unit information prior for (θ, Ψ) is given by $\theta|\Psi \sim$ multivariate normal$(\bar{y}, \Psi^{-1})$ and $\Psi \sim$ Wishart$(p+1, S^{-1})$, where $S = \sum(y_i - \bar{y})(y_i - \bar{y})^T/n$. This can be done by mimicking the procedure outlined in Exercise 5.6 as follows:

a) Reparameterize the multivariate normal model in terms of the precision matrix

$\Psi = \Sigma^{-1}$. Write out the resulting log likelihood, and find a probability density $p_U(\theta, \Psi) = p_U(\theta|\Psi)p_U(\Psi)$ such that $\log p(\theta, \Psi) = l(\theta, \Psi|\mathbf{Y})/n + c$, where c does not depend on θ or Ψ.

Hint: Write $(y_i-\theta)$ as $(y_i-\bar{y}+\bar{y}-\theta)$, and note that $\sum a_i^T \mathbf{B} a_i$ can be written as $\mathrm{tr}(AB)$, where $\mathbf{A}=\sum a_i a_i^T$.

$$\log p(\theta,\Psi)=\frac{1}{n}l(\theta,\Psi|\mathbf{Y})+c=\ln p_U(\theta|\Psi)+\ln p_U(\Psi)$$

$$l(\theta,\Psi|\mathbf{Y})=\ln[\prod_{i=1}^n p(y_i|\theta,\Psi)]=\frac{-np}{2}\ln(2\pi)-\frac{n}{2}\ln(|\Psi^{-1}|)-\frac{1}{2}\sum_{i=1}^n (y_i-\theta)^T\Psi(y_i-\theta)$$

$$=\frac{-np}{2}\ln(2\pi)-\frac{n}{2}\ln(|\Psi^{-1}|)-\frac{1}{2}\mathrm{tr}\left[\sum_{i=1}^n (y_i-\bar{y}+\bar{y}-\theta)^T\Psi(y_i-\bar{y}+\bar{y}-\theta)\right]$$

$$=\frac{-np}{2}\ln(2\pi)-\frac{n}{2}\ln(|\Psi^{-1}|)-\frac{1}{2}\mathrm{tr}\{\Psi[\underbrace{\sum_{i=1}^n (y_i-\bar{y})(y_i-\bar{y})^T}_{nS}+(\bar{y}-\theta)\underbrace{\sum_{i=1}^n (y_i-\bar{y})^T}_{0}+\underbrace{\sum_{i=1}^n (y_i-\bar{y})}_{0}(\bar{y}-\theta)^T+n(\bar{y}-\theta)(\bar{y}-\theta)^T]\}$$

$$=\underbrace{-\frac{1}{2}\mathrm{tr}\left[\Psi nS\right]}_{n\ln p_U(\Psi)}-\underbrace{\frac{n}{2}\ln(|\Psi^{-1}|)-\frac{n}{2}\mathrm{tr}\left[\Psi(\bar{y}-\theta)(\bar{y}-\theta)^T\right]}_{n\ln p_U(\theta|\Psi)}+\frac{-np}{2}\ln(2\pi)$$

$$\ln p_U(\Psi)=-\frac{1}{2n}\mathrm{tr}\left[\Psi\sum_{i=1}^n (y_i-\bar{y})(y_i-\bar{y})^T\right]=-\frac{1}{2}\mathrm{tr}\left[S\Psi\right]$$

$$p_U(\Psi)\propto|\Psi|^{\frac{p+1-p-1}{2}}\exp\left\{-\frac{1}{2}\mathrm{tr}\left[S\Psi\right]\right\}$$

$$\Psi\sim\mathrm{Wishart}(p+1,S^{-1})$$

$$\ln p_U(\theta|\Psi)=\frac{1}{2}\ln(|\Psi^{-1}|)-\frac{1}{2}\mathrm{tr}\left[\Psi(\bar{y}-\theta)(\bar{y}-\theta)^T\right]$$

$$p_U(\theta|\Psi)=|\Psi^{-1}|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\mathrm{tr}\left[\Psi(\bar{y}-\theta)(\bar{y}-\theta)^T\right]\right\}=|\Psi^{-1}|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(\bar{y}-\theta)\Psi(\bar{y}-\theta)^T\right\}$$

$$\theta|\Psi\sim\text{multivariate normal}(\bar{y},\Psi^{-1})$$

b) Let $p_U(\Sigma)$ be the inverse-Wishart density induced by $p_U(\Psi)$.

Obtain a density $p_U(\theta,\Sigma|y_1,...,y_n)\propto\ p_U(\theta|\Sigma)p_U(\Sigma)p(y_1,...,y_n|\theta,\Sigma)$. Can this be interpreted as a posterior distribution for θ and Σ?

$$p_U(\Sigma)\propto|\Sigma|^{-\frac{p+1+1+1}{2}}\exp\left\{-\frac{1}{2}\mathrm{tr}\left[S\Sigma^{-1}\right]\right\}$$

$$p_U(\theta|\Sigma)\propto|\Sigma|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(\bar{y}-\theta)\Sigma^{-1}(\bar{y}-\theta)^T\right\}$$

$$p(y_{1:n}|\theta,\Sigma)\propto|\Sigma|^{-\frac{n}{2}}\exp\left[-\frac{1}{2}\sum_{i=1}^n (y_i-\theta)^T\Sigma^{-1}(y_i-\theta)\right]$$

$$p_U(\theta,\Sigma|y_{1:n})\propto\ p_U(\theta|\Sigma)p_U(\Sigma)p(y_{1:n}|\theta,\Sigma)$$

$$\propto |\Sigma|^{-\frac{1}{2}-\frac{p+3}{2}-\frac{n}{2}} \exp\left[-\frac{1}{2}(\bar{y}-\theta)\Sigma^{-1}(\bar{y}-\theta)^T - \frac{1}{2}\text{tr}\left[S\Sigma^{-1}\right] - \frac{1}{2}\sum_{i=1}^{n}(y_i-\theta)^T\Sigma^{-1}(y_i-\theta)\right]$$

$$\propto |\Sigma|^{-\frac{1}{2}-\frac{n+p+3}{2}} \exp[-\frac{1}{2}(\bar{y}-\theta)\Sigma^{-1}(\bar{y}-\theta)^T - \frac{1}{2}\text{tr}\left[S\Sigma^{-1}\right] - \frac{1}{2}\sum_{i=1}^{n}(y_i-\bar{y})\Sigma^{-1}(y_i-\bar{y})^T$$

$$+\frac{1}{2}\sum_{i=1}^{n}(y_i-\bar{y})\Sigma^{-1}(y_i-\bar{y})^T - \frac{1}{2}\sum_{i=1}^{n}(y_i-\theta)^T\Sigma^{-1}(y_i-\theta)]$$

$$\propto |\Sigma|^{-\frac{1}{2}-\frac{n+p+3}{2}} \exp\left\{-\frac{1}{2}\left[(\bar{y}-\theta)\frac{\Sigma^{-1}}{n}(\bar{y}-\theta)^T + \sum_{i=1}^{n}(\bar{y}-\theta)^T\Sigma^{-1}(\bar{y}-\theta)\right] - \frac{1}{2}\text{tr}\left[S\Sigma^{-1}\right] - \frac{n}{2}\text{tr}\left[S\Sigma^{-1}\right]\right\}$$

$$\propto |\Sigma|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left[(\bar{y}-\theta)(n+1)\Sigma^{-1}(\bar{y}-\theta)^T\right]\right\} |\Sigma|^{-\frac{n+p+3}{2}} \exp\left\{-\frac{1}{2}\text{tr}\left[(n+1)S\Sigma^{-1}\right]\right\}$$

$$\propto |\Sigma|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left[(\bar{y}-\theta)(n+1)\Sigma^{-1}(\bar{y}-\theta)^T\right]\right\} |\Sigma|^{-\frac{n+p+3}{2}} \exp\left\{-\frac{1}{2}\text{tr}\left[(n+1)S\Sigma^{-1}\right]\right\}$$

$$\theta, \Sigma|y_{1:n} \sim \text{MVN}(\theta|\bar{y}, \frac{\Sigma}{n+1}) \cdot \text{Inverse-Wishart}(\Sigma|n+2, \frac{1}{(n+1)S})$$

density induced by $p_U(\Psi)$. Obtain a density $p_U(\theta, \Sigma|y_1, ..., y_n) \propto p_U(\theta|\Sigma)p_U(\Sigma)p(y_1, ..., y_n|\theta, \Sigma)$. Can this be interpreted as a posterior distribution for θ and Σ

3. Exercise 7.4 (pg 239) Marriage data

The file agehw.dat contains data on the ages of 100 married couples sampled from the U.S. population.

```
#Store the agehw.dat files in the same folder as this Rmd file
Y.marr <- read.table("agehw.dat",sep=" ",header=T)
```

a) Before you look at the data, use your own knowledge to formulate a semiconjugate prior distribution for

$\theta = (\theta_h, \theta_w)^T$ and Σ,where θ_h,θ_w are mean husband and wife ages, and Σ is the covariance matrix.

Assume the mean value is 40. The 95% range of ages is [20,60]. Variance is $10^2 = 100$. Correlation is 0.9, $\sigma_{hw} = 0.9 \times 100 = 90$. Set

$$\mathbf{S}_0^{-1} = \Lambda_0 = \begin{bmatrix} 100 & 90 \\ 90 & 100 \end{bmatrix} \qquad \nu_0 = p + 2 = 4$$

```
Y = Y.marr
p = ncol(Y.marr)
n = nrow(Y.marr)
ybar = colMeans(Y.marr)
mu0 = rep(40, p)
lambda0 = s0 = rbind(c(100,90), c(90,100))
nu0 = p + 2
```

b) Generate a prior predictive dataset of size n = 100,

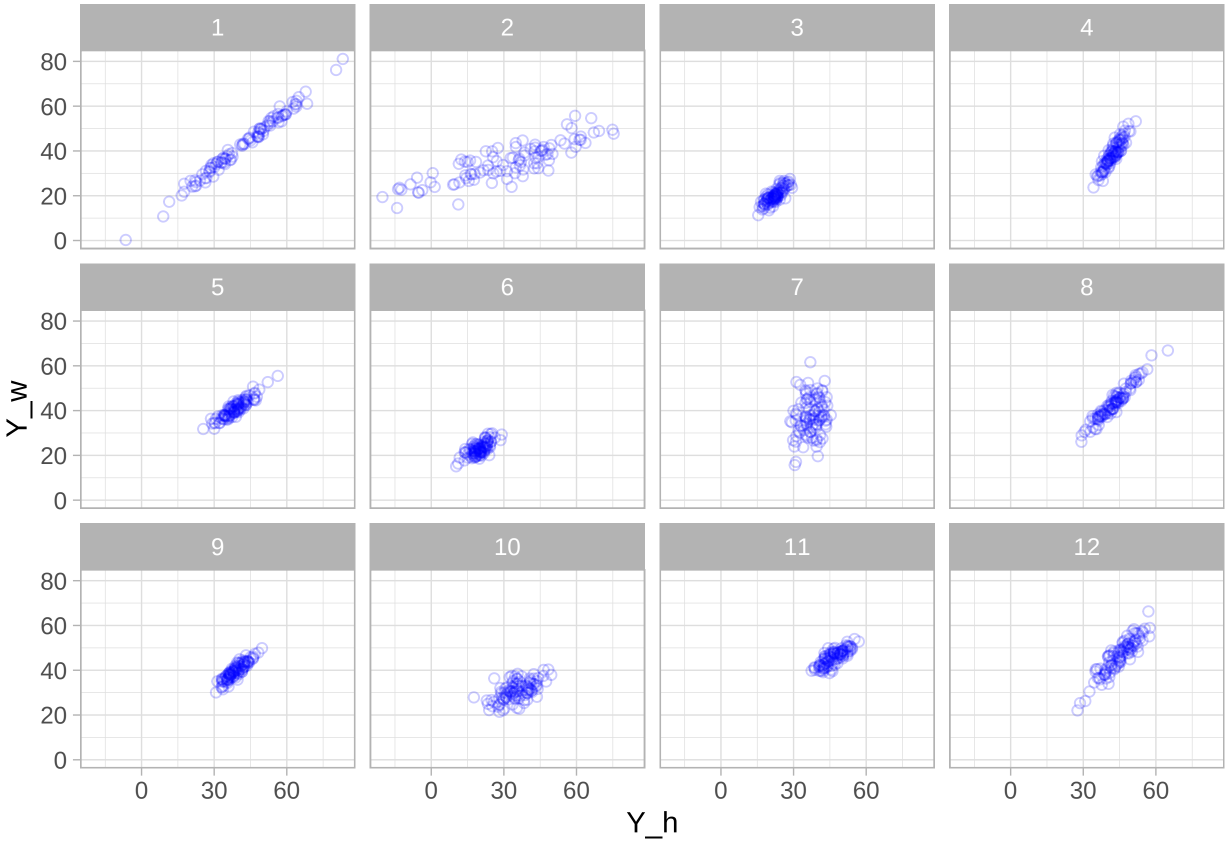
by sampling (θ, Σ) from your prior distribution and then simulating $Y_1, ..., Y_n \sim$ i.i.d. multivariate normal(θ, Σ). Generate several such datasets, make bivariate scatterplots for each dataset, and make sure they roughly represent your prior beliefs about what such a dataset would actually look like. If your prior predictive datasets do not conform to your beliefs, go back to part a) and formulate a new prior. Report the prior that you eventually decide upon, and provide scatterplots for at least three prior predictive datasets.

The wording of the question is interesting - I assume I'm supposed to sample a fixed $\boldsymbol{\theta}, \Sigma$ and from there sample 100 points all with the same parameters. If I were to do this myself, I feel like I would sample a new data point for each sample of $\boldsymbol{\theta}, \Sigma$...

In fact, because of that wording, I originally set $\nu_0 = p + 2 = 4$ to loosely center my prior. But given that this variance will often produce uncorrelated prior predictive datasets, I'm increasing ν_0 a bit...

After increasing ν_0, I'm fairly comfortable with what these posterior predictive datasets look like.

```r
# N = 100;S = 12
Y_preds = lapply(1:12, function(s) {
  # Sample THETA according to prior
  theta = mvrnorm(n = 1, mu0, lambda0)
  sigma = solve(rWishart(1, nu0, solve(s0))[, , 1])
  Y_s = mvrnorm(n = 100, theta, sigma)
  data.frame(Y_h = Y_s[, 1], Y_w = Y_s[, 2], dataset = s)
})
Y_comb = do.call(rbind, Y_preds)
ggplot(Y_comb, aes(x = Y_h, y = Y_w)) +geom_point(shape =1,alpha = 2/10,colour="blue") +facet_wrap(~ da
```



c) Using your prior distribution and the 100 values in the dataset,

obtain an MCMC approximation to $p(\theta, \Sigma | y_1, ..., y_{100})$. Plot the joint posterior distribution of θ_h and θ_w, and also the marginal posterior density of the correlation between Y_h and Y_w, the ages of a husband and wife. Obtain 95% posterior confidence intervals for θ_h, θ_w and the correlation coefficient.

```r
S = 10000
mcmc = function(Y, mu0, lambda0, s0, nu0) {
  ybar = colMeans(Y); p = ncol(Y); n = nrow(Y)
  THETA = matrix(nrow = S, ncol = p)
  SIGMA = array(dim = c(p, p, S))
  sigma = cov(Y) # Start with sigma sample
  # Gibbs sampling
  for (s in 1:S) {
    # Update theta
    lambda_n = solve(solve(lambda0) + n * solve(sigma))
    mu_n = lambda_n %*% (solve(lambda0) %*% mu0 + n * solve(sigma) %*% ybar)
    theta = mvrnorm(n = 1, mu_n, lambda_n)
    # Update sigma
    resid = t(Y) - c(theta)
    s_theta = resid %*% t(resid)
    s_n = s0 + s_theta
    sigma = solve(rWishart(1, nu0 + n,solve(s_n))[, , 1])

    THETA[s, ] = theta
    SIGMA[, , s] = sigma
  }
  list(theta = THETA, sigma = SIGMA)
}
prior_mcmc = mcmc(Y.marr, mu0, lambda0, s0, nu0)
THETA = prior_mcmc$theta
SIGMA = prior_mcmc$sigma

print_quantiles = function(THETA, SIGMA) {
  print("Husband")
  print(quantile(THETA[, 1], probs = c(0.025, 0.5, 0.975))) # Husband
  print("Wife")
  print(quantile(THETA[, 2], probs = c(0.025, 0.5, 0.975))) # Wife
  cors = apply(SIGMA, MARGIN = 3, FUN = function(covmat) {
    covmat[1, 2] / (sqrt(covmat[1, 1] * covmat[2, 2]))
  })
  print("Correlation")
  print(quantile(cors, probs = c(0.025, 0.5, 0.975)))
}
print_quantiles(THETA, SIGMA)
```

```
## [1] "Husband"
##     2.5%      50%    97.5%
## 41.67368 44.34450 46.96521
## [1] "Wife"
##     2.5%      50%    97.5%
## 38.33041 40.86515 43.37179
## [1] "Correlation"
##      2.5%       50%     97.5%
## 0.8622936 0.9044451 0.9343839
```

d) Obtain 95% posterior confidence intervals for θ_h, θ_w and the correlation coefficient using the following prior distributions:

i. Jeffreys' prior in Exercise 7.1;

```
THETA = matrix(nrow = S, ncol = p)
SIGMA = array(dim = c(p, p, S))
sigma = cov(Y)# Start with sigma sample
# Gibbs sampling
for (s in 1:S) {
  # Update theta
  theta = mvrnorm(n = 1, ybar, sigma/n)
  # Update sigma
  resid = t(Y) - c(theta)
  s_theta = resid %*% t(resid)
  sigma = solve(rWishart(1, n + 1, solve(s_theta))[, , 1])
  THETA[s, ] = theta
  SIGMA[, , s] = sigma
}
print_quantiles(THETA, SIGMA)
```

```
## [1] "Husband"
##     2.5%      50%    97.5%
## 41.65875 44.43980 47.20583
## [1] "Wife"
##     2.5%      50%    97.5%
## 38.32534 40.92770 43.48841
## [1] "Correlation"
##      2.5%       50%     97.5%
## 0.8605211 0.9042151 0.9345824
```

iii. a "diffuse prior" with $\mu_0 = \mathbf{0}, \Lambda_0 = 10^5 \times \mathbf{I}, \mathbf{S_0} = 1000 \times \mathbf{I}$ and $\nu_0 = 3$.

```
mu0 = rep(0, p)
lambda0 = 10^5 * diag(p)
s0 = 1000 * diag(p)
nu0 = 3
diffuse_mcmc = mcmc(Y.marr, mu0, lambda0, s0, nu0)
print_quantiles(diffuse_mcmc$theta, diffuse_mcmc$sigma)
```

```
## [1] "Husband"
##     2.5%      50%    97.5%
## 41.67544 44.40246 47.15921
## [1] "Wife"
##     2.5%      50%    97.5%
## 38.27102 40.89394 43.45358
## [1] "Correlation"
##      2.5%       50%     97.5%
## 0.7925982 0.8549079 0.9002584
```

e) Compare the confidence intervals from d) to those obtained in c).

Discusswhetherornotyouthinkthatyourpriorinformationishelpful in estimating θ and Σ, or if you think one of the alternatives in d) is preferable. What about if the sample size were much smaller, say $n = 25$?

- My prior

```
mu0 = rep(40, p)
lambda0 = s0 = rbind(c(100,90), c(90,100))
nu0 = p + 2
# nu0 = p + 2 + 10
prior_mcmc_short = mcmc(Y.marr[1:25,], mu0, lambda0, s0, nu0)
print_quantiles(prior_mcmc_short$theta, prior_mcmc_short$sigma)
```

```
## [1] "Husband"
##     2.5%      50%    97.5%
## 39.77854 44.85202 49.86623
## [1] "Wife"
##     2.5%      50%    97.5%
## 37.28127 42.63775 47.83414
## [1] "Correlation"
##      2.5%       50%     97.5%
## 0.8369135 0.9201213 0.9621092
```

- Diffuse prior

```
mu0 = rep(0, p)
lambda0 = 10^5 * diag(p)
s0 = 1000 * diag(p)
nu0 = 3
diffuse_mcmc_short = mcmc(Y.marr[1:25,], mu0, lambda0, s0, nu0)
print_quantiles(diffuse_mcmc_short$theta, diffuse_mcmc_short$sigma)
```

```
## [1] "Husband"
##     2.5%      50%    97.5%
## 39.17441 45.20190 51.03090
## [1] "Wife"
##     2.5%      50%    97.5%
## 36.58430 42.80811 49.14339
## [1] "Correlation"
##      2.5%       50%     97.5%
## 0.5416490 0.7609687 0.8847605
```

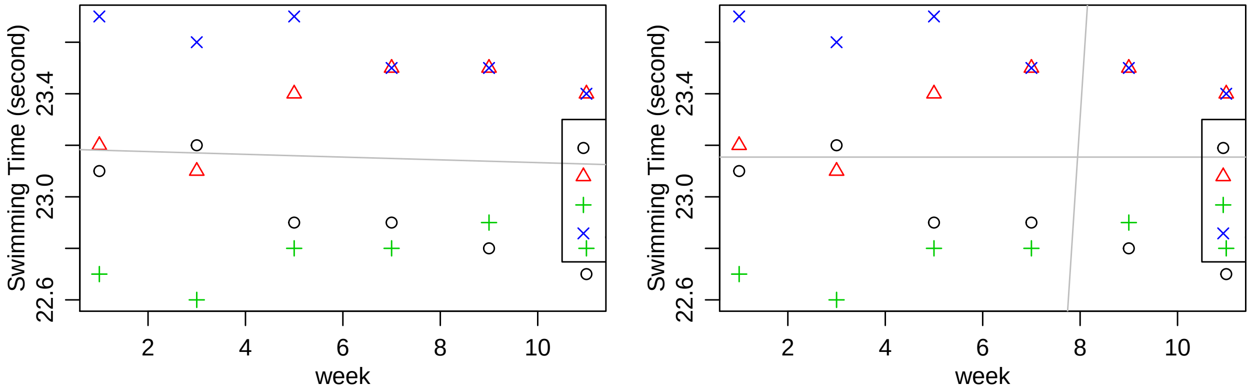
4. Exercise 9.1 (pg 242) Extrapolation

The file swim.dat contains data on the amount of time, in seconds, it takes each of four high school swimmers to swim 50 yards. Each swimmer has six times, taken on a biweekly basis.

a) Perform the following data analysis for each swimmer separately:

i. Fit a linear regression model of swimming time as the response and week as the explanatory variable. To formulate your prior, use the information that competitive times for this age group generally range from 22 to 24 seconds.

```
##
## Call:
## lm(formula = time ~ week, data = data.frame(Y.swim))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.5702 -0.3301 -0.0256  0.3512  0.5405
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 23.186310   0.149065 155.545   <2e-16 ***
## week        -0.005357   0.021591  -0.248    0.806
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3613 on 22 degrees of freedom
## Multiple R-squared:  0.002791,	Adjusted R-squared:  -0.04254
## F-statistic: 0.06156 on 1 and 22 DF,  p-value: 0.8063
```



The results show that the intercept (average estimate swimming time) is $\hat{\beta}_0 = 23.186$ second at 95% significent level.

The number of weeks has a tiny nagetive effect ($\hat{\beta}_1 = -0.00536$) on the swimming time but not significent at 95% levels.

Use the information that competitive times for this age group generally range from 22 to 24 seconds, we choose 23 as the prior expectation of swimming time. To start we need $\beta_0^{(0)} \pm 2\sigma^{(0)} = (22, 24)$ with high prob, get

$$\mathrm{E}(\beta_1) = 23 \quad \text{and} \quad \mathrm{E}(\beta_2) = 0, \quad \text{and}$$
$$\mathrm{var}(\beta_1) = (1/2)^2 \quad \text{and} \quad \mathrm{var}(\beta_2) = 0^2.$$

Assuming that $\sigma^2 \sim \mathrm{IG}(\nu_0/2, \nu_0\sigma_0^2/2)$, let's set

$$\nu_0 = 1 \quad \text{and} \quad \sigma_0^2 = (1/2)^2.$$

ii. For each swimmer j, obtain a posterior predictive distribution for $Y_j^\star$, their time if they were to swim two weeks from the last recorded time.

Given initial values $\{\beta^{(0)}, \sigma^{2(0)}\}$, new values can be generated by

1. updating β:

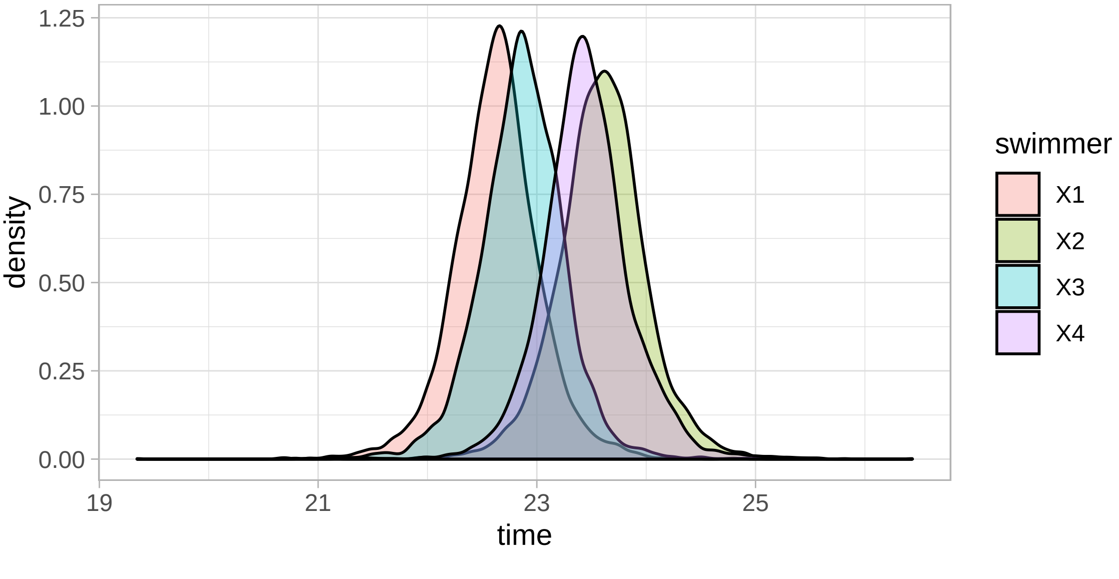
a) compute $V = Var[\beta|y, X, \sigma^{2(s)}]$ and $m = E[\beta|y, X, \sigma^{2(s)}]$

b) sample $\beta^{2(s+1)} \sim$ multivariate normal (m, V)

2. updating σ^2:

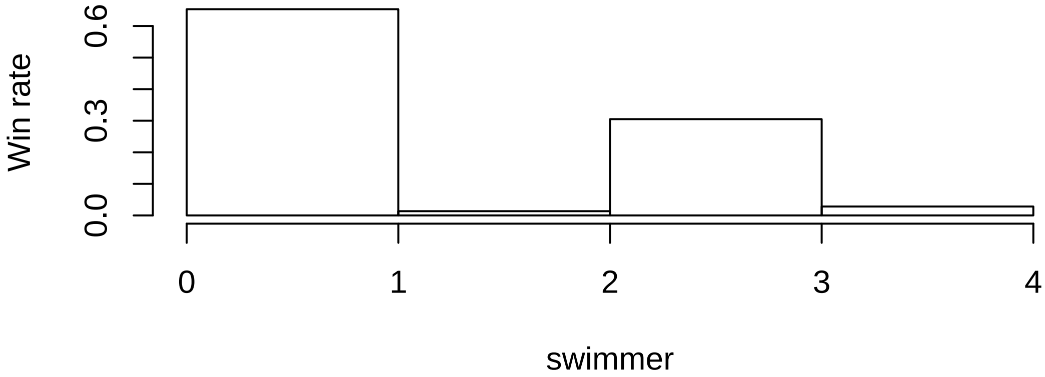
a) compute $SSR(\beta^{2(s)})$;

b) sample $\sigma^{2(s+1)} \sim$ inverse-gamma $([\nu_0 + n]/2, [\nu_0\sigma_0^2 + SSR(\beta^{2(s+1)})]/2)$.



b) The coach of the team has to decide which of the four swimmers will compete

in a swimming meet in two weeks. Using your predictive distributions, compute $Pr(Y_j^\star = \max\{Y_1^\star, ..., Y_4^\star\}|Y))$ for each swimmer j , and based on this make a recommendation to the coach.



```
## fastest_times
##      1      2      3      4
## 0.6534 0.0134 0.3050 0.0282
```

In posterior predictive dataset, swimmer 1 is the fastest about 65% of the time by week 13, so we recommend that swimmer 1 race.

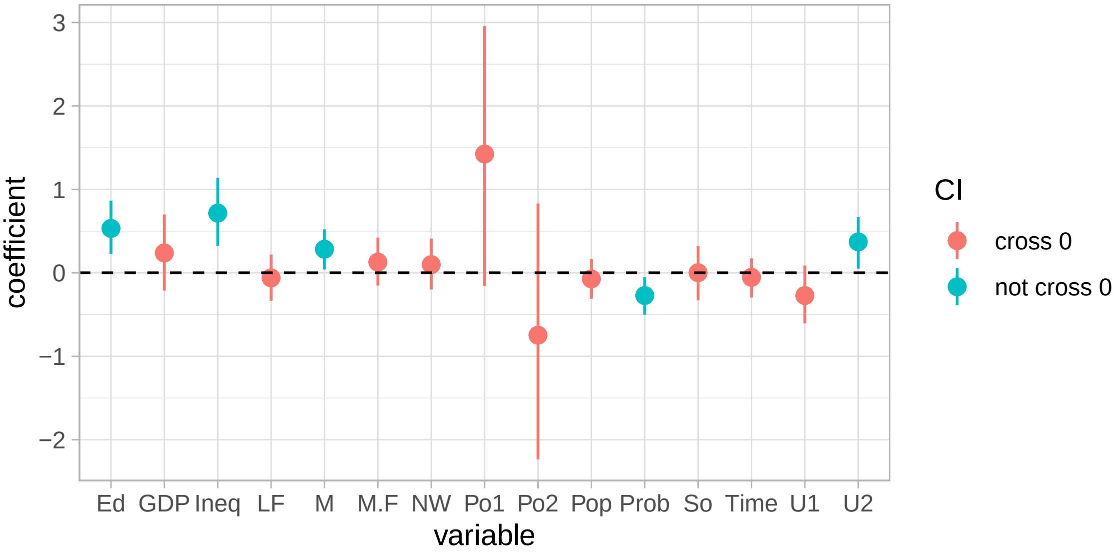
5. Exercise 9.3 (pg 243) Crime

The file crime.dat contains crime rates and data on 15 explanatory variables for 47 U.S. states, in which both the crime rates and the explanatory variables have been centered and scaled to have variance 1. A description of the variables can be obtained by typing library(MASS);?UScrime in R.

```
data("UScrime",package="MASS")
namvars <- names(UScrime)
```

a) Fit a regression model

$y = X\beta + \epsilon$ using the g-prior with $g = n, \nu_0 = 2$ and $\sigma_0^2 = 1$. Obtain marginal posterior means and 95% confidence intervals for β, and compare to the least squares estimates. Describe the relationships between crime and the explanatory variables. Which variables seem strongly predictive of crime rates?



The results show that Ed (mean years of schooling), Ineq (Income inequality), M (percentage of males aged 14-24), Prob (probability of imprisonment), and U2 (unemployment rate of urban males 35-39).

b) Lets see how well regression models can predict crime rates based on the

X-variables. Randomly divide the crime roughly in half, into a training set $\{y_{tr}, X_{tr}\}$ and a test set $\{y_{te}, X_{te}\}$

i. Using only the training set, obtain least squares regression coefficients $\hat{\beta}_{ols}$.

Obtain predicted values for the test data by computing $\hat{y}_{ols} = \mathbf{X}_{te}\hat{\beta}_{ols}$. Plot $\hat{y}_{ols}$ versus y_{te} and compute the prediction error $\frac{1}{n_{te}}\sum(y_{i,te} - \hat{y}_{i,ols})^2$.

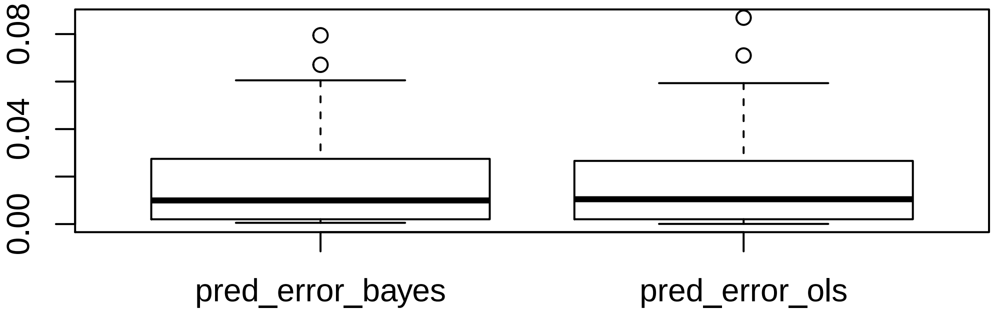
Table 1: Table continues below

M	So	Ed	Po1	Po2	LF	M.F	Pop
0.1942	0.2062	0.6433	0.3059	0.4473	-0.09199	0.03538	0.0844

NW	U1	U2	GDP	Ineq	Prob	Time
-0.01228	-0.03279	0.1549	0.1046	0.7669	-0.2582	0.05938

ii. Now obtain the posterior mean $\beta_{Bayes} = E[\beta|y_{tr}]$ using the g-prior described above and the training data only.

Obtain predictions for the test set $\hat{y}_{Bayes} = \mathbf{X}_{test}\hat{\beta}_{Bayes}$. Plot versus the test data, compute the prediction error, and compare to the OLS prediction error. Explain the results.



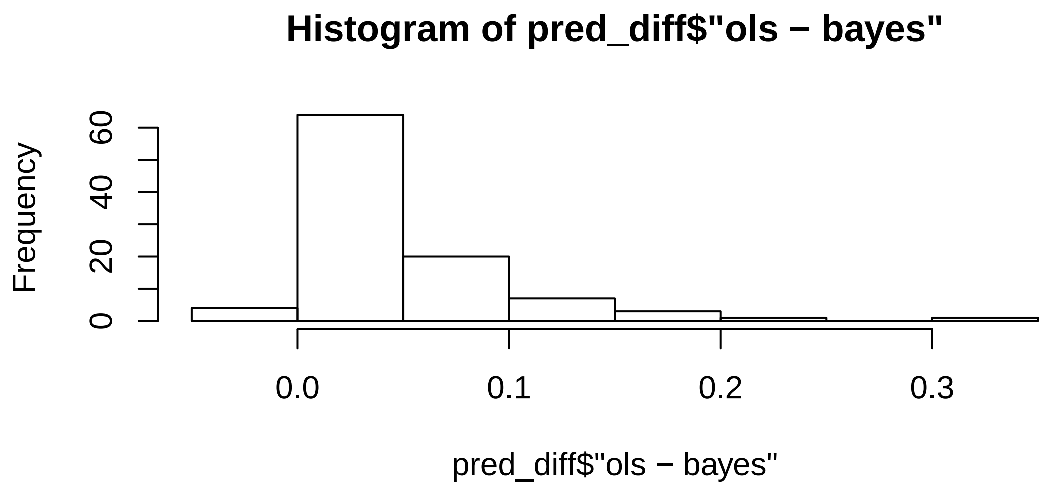
The result shows that there isn't significent difference between the two methods.

c) Repeat the procedures in b) many times with different randomly generated

test and training sets. Compute the average prediction error for both the OLS and Bayesian methods.

```
N = 100
set.seed(1)
pred_errors = t(sapply(1:N, function(i) {
  y = crime$y
  X =as.matrix(crime[,-1])
  train_i = sample.int(length(y), size = round(length(y) / 2), replace = FALSE)
  ytr = y[train_i]
  Xtr = X[train_i, ]
  yte = y[-train_i]
  Xte = X[-train_i, ]
  # OLS
  beta_ols = inv(t(Xtr) %*% Xtr) %*% t(Xtr) %*% ytr
  beta_ols
  y_ols = Xte %*% beta_ols
  pred_error_ols = sum((yte - y_ols)^2) / length(yte)
  # Bayes
  y = ytr
  X = Xtr
  n = dim(X)[1]
  p = dim(X)[2]
  g = n
  nu0 = 2
  s20 = 1
  S = 1000
  Hg = (g / (g + 1)) * X %*% inv(t(X) %*% X) %*% t(X)
  SSRg = t(y) %*% (diag(1, nrow = n) - Hg) %*% y
  s2 = 1 / rgamma(S, (nu0 + n) / 2, (nu0 * s20 + SSRg) / 2)
  Vb = g * inv(t(X) %*% X) / (g + 1)
  Eb = Vb %*% t(X) %*% y
  E = matrix(rnorm(S * p, 0, sqrt(s2)), S, p)
  beta = t(t(E %*% chol(Vb)) + c(Eb))
  beta_bayes = as.matrix(colMeans(beta))
  y_bayes = Xte %*% beta_bayes
  pred_error_bayes = sum((yte - y_bayes)^2) / length(yte)
  c(pred_error_ols, pred_error_bayes)
})) %>% as.data.frame
colnames(pred_errors) = c('ols', 'bayes')
```

```
pred_diff = pred_errors %>% transmute(`ols - bayes` = ols - bayes)
hist(pred_diff$'ols - bayes')
```



```
mean(pred_errors$bayes < pred_errors$ols)
```

```
## [1] 0.96
```

The result of $\text{err}_{\text{Bayes}} - \text{err}_{\text{ols}}$ shows that the Bayes estimator did better than the OLS estimator For 100 samples, 96% of the time, the predictive error using the Bayes estimators is less than the predictive error using the OLS estimators.