SDS 383D: Modeling II **Spring 2017**

Section 2: Bayesian inference in Gaussian models

2.1 Bayesian inference in a simple Gaussian model

Let's start with a simple, one-dimensional Gaussian example, where

$$y_i|\mu,\sigma^2 \sim \mathrm{N}(\mu,\sigma^2).$$

We will assume that μ and σ are unknown, and will put conjugate priors on them both, so that

$$\begin{aligned}\sigma^2 \sim&\text{Inv-Gamma}(\alpha_0,\beta_0)\\ \mu|\sigma^2 \sim&\text{Normal}\left(\mu_0,\frac{\sigma^2}{\kappa_0}\right)\end{aligned}$$

or, equivalently,

$$\begin{aligned}y_i|\mu,\omega \sim&\mathrm{N}(\mu,1/\omega)\\ \omega \sim&\text{Gamma}(\alpha_0,\beta_0)\\ \mu|\omega \sim&\text{Normal}\left(\mu_0,\frac{1}{\omega\kappa_0}\right)\end{aligned}$$

We refer to this as a normal/inverse gamma prior on μ and σ^2 (or a normal/gamma prior on μ and ω). We will now explore the posterior distributions on μ and $\omega(/\sigma^2)$ – much of this will involve similar results to those obtained in the first set of exercises.

Exercise 2.1 *Derive the conditional posterior distributions $p(\mu,\omega|y_1,\ldots,y_n)$ (or $p(\mu,\sigma^2|y_1,\ldots,y_n)$) and show that it is in the same family as $p(\mu,\omega)$. What are the updated parameters α_n,β_n,μ_n and κ_n?*

Solution Begin:

$$p(\mu,\omega|\underline{y}) \propto p(\underline{y}|\mu,\omega)\cdot p(\omega)$$

$$p(\mu,\omega) = \frac{\sqrt{\omega\kappa_0}}{\sqrt{2\pi}}\exp\left(-\frac{\omega\kappa_0}{2}(\mu-\mu_0)^2\right)\cdot\frac{1}{\Gamma(\alpha_0)}\beta_0^{\alpha_0}\omega^{\alpha_0-1}\exp(-\beta_0\omega)$$

This is a normal-gamma core

$$p(\mu,\omega|\underline{y}) \propto \omega^{\alpha_0+\frac{n}{2}+\frac{1}{2}}\exp(-\omega(\beta_0+\frac{\kappa}{2}(\mu-\mu_0)^2+\frac{1}{2}\sum_{i=1}^{n}(y_i-\mu)^2)$$

Thus, $\alpha_n = \alpha_0+\frac{n}{2}$
Next to find κ_n,β_n and μ_n Need to put $I = (\beta_0+\frac{\kappa}{2}(\mu-\mu_0)^2+\frac{1}{2}\sum_{i=1}^{n}(y_i-\mu)^2$ into form $\beta_n+\frac{\kappa_n}{2}(\mu-\mu_n)^2$

$$I = \beta_0+\frac{\kappa_0+n}{2}\left[\left(\mu-\frac{\mu_0+\sum y_i}{\kappa_0+n}\right)^2+\frac{\mu_0^2+\sum y_i^2}{\kappa_0+n}-\left(\frac{\mu_0+\sum y}{\kappa_0+n}\right)^2\right]$$

Therefore,

$$\mu_n = \frac{\kappa_0\mu_0 + \sum y_i}{\kappa_0 + n}$$

$$\kappa_n = \kappa_0 + n$$

$$\beta_n = \beta_0 - \frac{\left(\kappa_0\mu_0 + \sum y_i^2\right)^2}{\kappa + n} + \frac{\kappa_0\mu_0^2 + \sum y_i^2}{2}$$

Solution End

Exercise 2.2 *Derive the conditional posterior distribution $p(\mu|\omega, y_1, \ldots, y_n)$ and $p(\omega|y_1, \ldots, y_n)$ (or if you'd prefer, $p(\mu|\sigma^2, y_1, \ldots, y_n)$ and $p(\sigma^2|y_1, \ldots, y_n)$). Based on this and the previous exercise, what are reasonable interpretations for the parameters $\mu_0, \kappa_0, \alpha_0$ and β_0?*

Solution Begin:

$$p(\mu|\omega, \underline{y}) \propto p(\underline{y}|\mu, \omega) \cdot p(\mu|\omega)$$

$$\propto exp\left(-\frac{\omega}{2}\left(\sum y_i^2 - 2\mu\sum y_i + n\mu^2 + \kappa_0\mu^2 - 2\kappa_0\mu\mu_0 + \kappa_0\mu_0^2\right)\right)$$

$$\propto exp\left\{-\frac{\omega(n+\kappa_0)}{2}\left(\mu - \frac{\sum y_i + \kappa_0\mu_0}{n+\kappa_0}\right)^2\right\}$$

This is normal distribution core, with mean $\frac{\sum y_i + \kappa_0\mu_0}{n+\kappa_0}$ and precision $-\omega(n+\kappa_0)$

$$p(\omega|\underline{y}) \propto p(\underline{y}|\omega, \mu)p(\omega)$$

$$\propto \omega^{\alpha_0 - 1 + \frac{n}{2}} exp(\frac{w}{2}\sum(y_i - \mu)^2 - \beta_0\omega)$$

$$\propto \omega^{\alpha_0 - 1 + \frac{n}{2}} exp\left(-\omega\left(\beta_0 + \frac{\sum(y_i - \mu)^2}{2}\right)\right)$$

This is kernel of Gamma distribution, with hyper-parameters $(\alpha_0 + \frac{n}{2}, \beta_0 + \frac{\sum(y_i-\mu)^2}{2}))$)
Solution End

Exercise 2.3 *Show that the marginal distribution over μ is a centered, scaled t-distribution (note we showed something very similar in the last set of exercises!), i.e.*

$$p(\mu) \propto \left(1 + \frac{1}{\nu}\frac{(\mu - m)^2}{s^2}\right)^{-\frac{\nu+1}{2}}$$

What are the location parameter m, scale parameter s, and degree of freedom ν?

Solution Begin:

$$p(\mu) = \int p(\mu, \omega)d\omega = \frac{\sqrt{\kappa_0}\beta_0^{\alpha_0}}{\Gamma(\alpha_0)\sqrt{2\pi}\left[\beta_0 + \frac{\kappa_0}{2}(\mu - \mu_0)^2\right]} \cdot \int_\omega f_{gamma}(\omega)d\omega$$

$$= \int p(\mu, \omega)d\omega = \frac{\sqrt{\kappa_0}\beta_0^{\alpha_0}}{\Gamma(\alpha_0)\sqrt{2\pi}\left[\beta_0 + \frac{\kappa_0}{2}(\mu - \mu_0)^2\right]}$$

$$\propto \left(1 + \frac{1}{\nu}\frac{(\mu - m)^2}{s^2}\right)$$

where

$$\nu = 2\alpha_0, m = \mu_0, s = \sqrt{\frac{\beta_0}{2\kappa_0\alpha_0^2}}$$

Solution End

Exercise 2.4 *The marginal posterior $p(\mu|y_1, \ldots, y_n)$ is also a centered, scaled t-distribution. Find the updated location, scale and degrees of freedom.*

Solution Begin: According to **Exercise 2.1**, take $\alpha_n = \alpha_0 + \frac{n}{2}$, $\mu_n = \frac{\kappa_0\mu_0 + \sum y_i}{\kappa_0 + n}$, $\kappa_n = \kappa_0 + n$, $\beta_n = \beta_0 - \frac{\left(\kappa_0\mu_0 + \sum y_i^2\right)^2}{\kappa + n} + \frac{\kappa_0\mu_0^2 + \sum y_i^2}{2}$
Then,

$$\nu = 2\alpha_n, m = \mu_n, s = \sqrt{\frac{\beta_n}{2\kappa_n\alpha_n^2}}$$

Solution End

Exercise 2.5 *Derive the posterior predictive distribution $p(y_{n+1}, \ldots, y_{n+m}|y_1, \ldots, y_m)$.*

Begin Solution:

$$p(y|\underline{y}) = \int\int p(y|\mu, \omega)p(\mu, \omega|\underline{y})d\mu d\omega$$

$$= \int\int \left(\frac{w}{2\pi}\right)^{\frac{1}{2}} exp\left\{-\frac{\omega}{2}(y_i - \mu)^2\right\} \frac{\sqrt{\kappa_n}}{\Gamma(\alpha_n)\sqrt{2\pi}}\beta_n^{\alpha_n} \exp\left\{-\omega(\beta_n + \frac{\kappa_n}{2}(\mu - \mu_n)^2)\right\} \omega^{\alpha_n - \frac{1}{2}} d\mu d\omega$$

$$\propto \left(\beta_n + \frac{1}{2}(\mu - y)^2\right)^{-\alpha_n - \frac{1}{2}} \cdot \int\int f_{NG}(\omega, \mu)d\omega d\mu$$

$$= \left(\beta_n + \frac{1}{2}(\mu_n - y)^2\right)^{-\alpha_n - \frac{1}{2}}$$

Put into Student's t-distribution form $\left(1 + \frac{1}{\nu}\frac{(\mu - m)^2}{s^2}\right)^{-\frac{\nu+1}{2}}$
where

$$v = 2\alpha_n, m = \mu_n, s = \sqrt{\frac{\beta_0}{\alpha_n}}$$

End Solution

Exercise 2.6 *Derive the marginal distribution over $y_1, \ldots, y_n$.*

2.2 Bayesian inference in a multivariate Gaussian model

Let's now assume that each y_i is a d-dimensional vector, such that

$$y_i \sim \mathrm{N}(\mu, \Sigma)$$

for d-dimensional mean vector μ and $d \times d$ covariance matrix Σ.

We will put an *inverse Wishart* prior on Σ. The inverse Wishart distribution is a distribution over positive-definite matrices parametrized by $\nu_0 > d-1$ degrees of freedom and positive definite matrix Λ_0^{-1}, with pdf

$$p(\Sigma|\nu_0, \Lambda_0^{-1}) = \frac{|\Lambda|^{\nu_0/2}}{2^{(\nu_0+d)/2}\Gamma_d(\nu_0/2)}|\Sigma|^{-\frac{\nu_0+d+1}{2}}e^{-\frac{1}{2}\text{tr}(Lambda\Sigma^{-1})}$$

where $\Gamma_d(x) = \pi^{d(d-1)/4}\prod_{i=1}^{d}\Gamma\left(x - \frac{j-1}{2}\right)$.

Exercise 2.7 *Show that in the univariate case, the inverse Wishart distribution reduces to the inverse gamma distribution.*

Begin Solution: for $d = 1$,

$$\Gamma_d(x) = \pi^0 \cdot \Gamma(x) = \Gamma(x)$$

$$P(\Sigma|\nu_0, \Lambda_0^{-1}) = \frac{\Lambda^{\nu_0/2}}{2^{\nu_0/2}\Gamma(\nu_0/2)} \cdot \Sigma^{\nu_0/2-1}e^{-\frac{1}{2}\frac{\Lambda}{2}}$$

Let $\alpha = \frac{\nu_0}{2}, \beta = \frac{1}{2}\Lambda$
Inverse gamma pdf $f(x;\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\Sigma^{\frac{\nu_0}{2}-1}\exp(-\frac{\frac{1}{2}\Lambda}{\Sigma})$

End Solution

Exercise 2.8 *Let $\Sigma \sim$ Inv-Wishart(ν_0, Λ_0^{-1}) and $\mu|\Sigma \sim N(\mu_0, \Sigma/\kappa_0)$, so that*

$$p(\mu, \Sigma) \propto |\Sigma|^{-\frac{\nu_0+d+1}{2}}e^{-\frac{1}{2}tr(\Lambda_0\Sigma^{-1})+\frac{\kappa_0}{2}(\mu-\mu_0)^T\Sigma^{-1}(\mu-\mu_0)}$$

and let

$$y_i \sim N(\mu, \Sigma)$$

Show that $p(\mu, \Sigma|y_1, \ldots, y_n)$ is also normal-inverse Wishart distributed, and give the form of the updated parameters μ_n, κ_n, ν_n and Λ_n. It will be helpful to note that

$$\begin{aligned}\sum_{i=1}^{n}(y_i-\mu)^T\Sigma^{-1}(y_i-\mu) &= \sum_{i=1}^{n}\sum_{j=1}^{d}\sum_{k=1}^{d}(x_{ij}-\mu_j)(\Sigma^{-1})_{jk}(x_{ik}-\mu_k) \\ &= \sum_{j=1}^{d}\sum_{k=1}^{d}(\Sigma^{-1})_{ab}\sum_{i=1}^{n}(x_{ij}-\mu_j)(x_{ik}-\mu_k) \\ &= tr\left(\Sigma^{-1}\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^T\right)\end{aligned}$$

Based on this, give interpretations for the prior parameters.

Begin Solution:

$$P(\mu, \Sigma|y) \propto p(y|\mu, \Sigma) \cdot p(\mu, \Sigma)$$

$$\propto |\Sigma|^{\frac{\nu_0+d+1+n}{2}} exp(-\frac{1}{2}tr(\Lambda\Sigma^{-1}) + \frac{\kappa_0}{2}(\mu-\mu_0)^T\Sigma^{-1}(\mu-\mu_0) - \frac{1}{2}\sum_i (y_i-\mu)\Sigma^{-1}(y_i-\mu)^T)$$

$$\propto |\Sigma|^{\frac{\nu_0+d+1+n}{2}} exp\left\{\frac{1}{2}tr\left(\Sigma^{-1}(\mu\mu^T(n+\kappa_0) - \mu(n\bar{x}^T + \kappa_0\mu_0^T + \sum_i x_i x_i^T))\right) - \frac{1}{2}tr(\Lambda_0\Sigma^{-1})\right\}$$

Put into the form

$$|\Sigma|^{\frac{\nu_0+d+1+n}{2}} exp\left\{-\frac{\kappa_0+n}{2}tr\left(\Sigma^{-1}(\mu-\mu_n)(\mu-\mu_n)^T\right) - \frac{1}{2}tr\left(\left[\Lambda_0 + \sum_i x_i x_i^T + \kappa_0\mu_0\mu_0^T - \kappa_n\mu_n\mu_n^T\right]\Sigma^{-1}\right)\right\}$$

Thus,

$$\nu_n = \nu_0 + n, \qquad \mu_n = \frac{n\bar{x} + \kappa_0\mu_0}{n+\kappa_0}$$

$$\kappa_n = \kappa_0 + n, \qquad \Lambda_n = \Lambda_0 + \sum_i (x_i-\bar{x})(x_i-\bar{x})^T + \frac{\kappa_0 n}{\kappa_0+n}(x-\bar{x})(x-\bar{x})^T$$

End Solution

2.3 A Gaussian linear model

Lets now add in covariates, so that

$$\mathbf{y}|\beta, X \sim \text{Normal}(X\beta, (\omega\Lambda)^{-1})$$

where $\mathbf{y}$ is a vector of n responses; X is a $n \times d$ matrix of covariates; and Λ is a known positive definite matrix. Let's assume $\beta \sim \text{Normal}(\mu, (\omega K)^{-1})$ and $\omega \sim \text{Gamma}(a, b)$, where K is assumed fixed.

Exercise 2.9 *Derive the conditional posterior* $p(\beta|\omega, y_1, \ldots, y_n)$

Begin Solution:

$$p(\beta|\omega, y) \propto p(y|\omega, \beta) \cdot p(\beta|\omega)$$
$$\propto exp\left\{\frac{1}{2}\left[-2y^T\omega\Lambda\beta + \beta^T X^T\omega\Lambda X\beta + \beta^T\omega k\beta - 2\beta^T\omega k\mu\right]\right\}$$
$$\propto exp\left\{\frac{1}{2}(X^T\omega\Lambda X + \omega k)\left[\beta^T\beta - 2(y^T\Lambda X + y^T k)(X^T X + y^T k)^{-1}\right]\right\}$$

Thus,

$$\Sigma_n = (X^T\omega\Lambda X + \omega K)^{-1}, \qquad \mu_n = (X^T\Lambda X + k)^{-1}(X^T\Lambda y + K\mu)$$

End Solution

Exercise 2.10 *Derive the marginal posterior* $p(\omega|y_1, \ldots, y_n)$

Begin Solution:

$$p(\omega, \beta|y) \propto p(y|\omega, \beta) \cdot p(\beta|\omega) \cdot p(\omega)$$
$$p(\omega|y) = \int p(\omega, \beta|y)d\beta$$

$p(y|\omega, \beta) \cdot p(\beta\omega)$ follows normal distribution, and integrates to 1. Thus,

$$p(\omega, \beta|y) \propto \omega^{\frac{n+d}{2}+a-1} exp\left\{-\omega\left(b + \frac{1}{2}(y^T\Lambda y + \mu^T K\mu - \mu_n^T(X^T\Lambda X + K)\mu_n)\right)\right\}$$

This is Gamma distribution with

$$a_n = a + \frac{n+d}{2} \qquad b_n = b + \frac{1}{2}(y^T\Lambda y + mu^T K\mu - \mu_n^T(X^T\Lambda X + K)\mu_n)$$

End Solution

Exercise 2.11 *Derive the marginal posterior,* $p(\beta|y_1, \ldots, y_n)$

Begin Solution:

$$p(\beta|y) = \int p(\omega, \beta|y)d\omega$$

$$= det\left(\frac{X^T\Lambda X + K}{2\pi}\right)^{\frac{1}{2}} \frac{b_n^{a_n}}{\Gamma(a_n)} \int \omega^{a_n+\frac{n}{2}-1} exp\left\{-\omega\left[b_n + \frac{1}{2}(\beta - \mu_n)^T(X^T\Lambda X + K)(\beta - \mu_n)\right]\right\} d\omega$$

Integrating out the normal-gamma distribution part

$$p(\beta|y) = \frac{\Gamma\left(\frac{\mu+n}{2}\right)|X^T\Lambda X + K|^{\frac{1}{2}}}{\Gamma\left(\frac{\nu}{2}\right)(\pi\nu)^{\frac{n}{2}}}\left(1 + \frac{1}{\nu}(\beta-\mu_n)^T(X^T\Lambda X + K)(\beta-\mu_n)\right)$$

This is student's t-distribution with $\mu - \mu_n$, $\Sigma = (X^T\Lambda X + K)^{-1}$. Degree of freedom is ν
End Solution

Exercise 2.12 *Download the dataset dental.csv from Github. This dataset measures a dental distance (specifically, the distance between the center of the pituitary to the pterygomaxillary fissure) in 27 children. Add a column of ones to correspond to the intercept. Fit the above Bayesian model to the dataset, using $\Lambda = I$ and $K = I$, and picking vague priors for the hyperparameters, and plot the resulting fit. How does it compare to the frequentist LS and ridge regression results?*

Begin Solution: For details please refer to the R script file on Github. Some summary information is provided here. In total three models are attempted, with the first model considering both males and females, and the second model considering only males, and the third model considering only females. For the model that considers all data, the gender information is coded as a dummy variable, with $male- > 0$ and $female- > 0$.
The RMSE for these models are summarized below.

	All data included	Male only	Female only
Bayesian Model	2.31	2.45	2.45

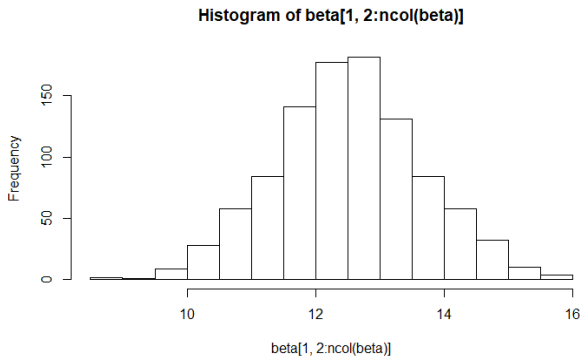

Figure 2.1: histogram of the posterior distribution of β when all data are included

[H]

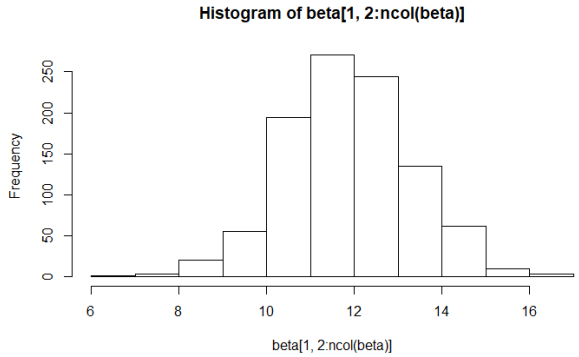

Figure 2.2: histogram of the posterior distribution of β when only male data are included

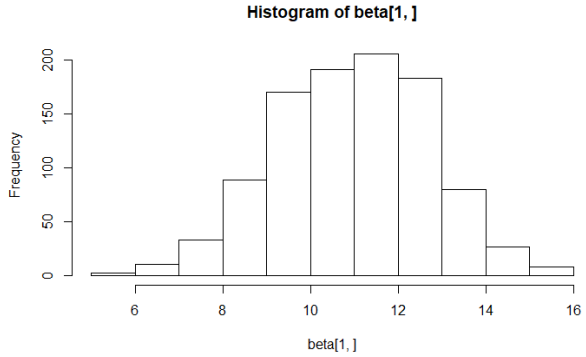

Figure 2.3: histogram of the posterior distribution of β when only female data are included

End Solution

2.4 A hierarchical Gaussian linear model

The dental dataset has heavier tailed residuals than we would expect under a Gaussian model. We've seen previously that we can model a scaled t-distribution using a scale mixture of Gaussians; let's put that into effect here. Concretely, let

$$\begin{aligned}
\mathbf{y}|\beta, \omega, \Lambda \sim & \mathrm{N}(X\beta, (\omega\Lambda)^{-1}) \\
\Lambda = & \mathrm{diag}(\lambda_1, \ldots, \lambda_n) \\
\lambda_i \overset{iid}{\sim} & \mathrm{Gamma}(\tau, \tau) \\
\beta|\omega \sim & \mathrm{N}(\mu, (\omega K)^{-1}) \\
\omega \sim & \mathrm{Gamma}(a, b)
\end{aligned}$$

Exercise 2.13 *What is the conditional posterior, $p(\lambda_i|\mathbf{y}, \beta, \omega)$?*

Begin Solution:

$$p(\lambda_i|\mathbf{y}, \beta, \omega) \propto p(\mathbf{y}|\lambda_i, \beta, \omega) \cdot p(\lambda_i|\beta, \omega)$$

$$= \sqrt{\frac{\omega\lambda_i}{2\pi}} exp\left\{-\frac{\omega\lambda_i}{2}(y_i - x_i^T\beta)^2\right\} \frac{\tau^\tau}{\Gamma(\tau)} \lambda_i^{\tau-1} exp(-\tau\lambda_i)$$

$$\propto \lambda_i^{\tau+\frac{1}{2}-1} exp\left\{-\lambda_i\left(\tau + \frac{\omega}{2}(y_i - x_i^T\beta)^2\right)\right\}$$

This is $Gamma(\tau + \frac{1}{2}, \tau + \frac{\omega}{2}(y_i - x_i^T\beta)^2)$
End Solution

Exercise 2.14 *Write a Gibbs sampler that alternates between sampling from the conditional posteriors of λ_i, β and ω, and run it for a couple of thousand samplers to fit the model to the dental dataset.*

Begin Solution: Similar to 2.12, three models are built. The MCMC takes 3,000 iterations, with the first 500 samples regarded as burn in. The histograms of *beta* are plotted below. The RMSE for these models are summarized below.

	All data included	Male only	Female only
Bayesian Model	2.43	2.37	2.32

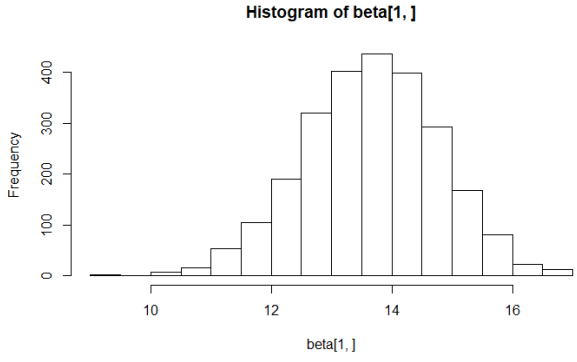

Figure 2.4: histogram of the posterior distribution of β when all data are included

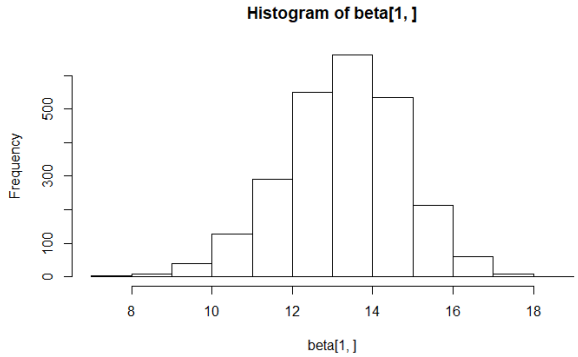

Figure 2.5: histogram of the posterior distribution of β when male data are included

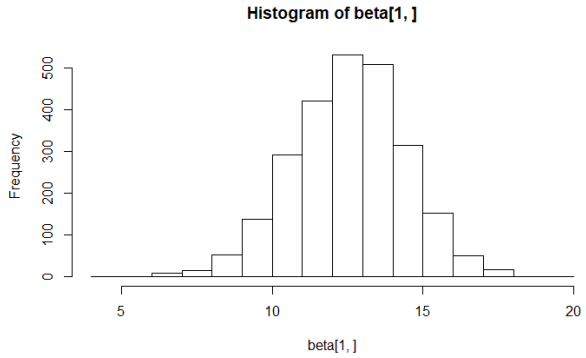

Figure 2.6: histogram of the posterior distribution of β when female data are included

End Solution

Exercise 2.15 *Compare the two fits. Does the new fit capture everything we would like? What assumptions is it making? In particular, look at the fit for just male and just female subjects. Suggest ways in which we could modify the model, and for at least one of the suggestions, write an updated Gibbs sampler and run it on your model.*

Begin Solution: The new fit is an improvement over the one that treats Λ as constant. However improvements can be made by considering the "subject" feature in the dataset. One way to improve the model is by building separate models for each individual.
End Solution