Bayesian Hierarchical Model

BIOS 525: Longitudinal and Multi-level Data Analysis

Jacob Englert
Department of Biostatistics & Bioinformatics
Rollins School of Public Health
Emory University

31 October 2022

Outline

Bayesian Inference Review

1. $\mathbf{Y}$ = observed data
2. $\boldsymbol{\theta}$ = unknown parameter vector
3. Data likelihood: $[\,\mathbf{Y}\,|\,\boldsymbol{\theta}\,]$
4. Prior distribution: $[\,\boldsymbol{\theta}\,]$

We wish to obtain the posterior distribution:

$$[\,\boldsymbol{\theta}\,|\,\mathbf{Y}\,] = \frac{[\,\mathbf{Y}\,|\,\boldsymbol{\theta}\,] \times [\,\boldsymbol{\theta}\,]}{[\,\mathbf{Y}\,]} \propto [\,\mathbf{Y}\,|\,\boldsymbol{\theta}\,] \times [\,\boldsymbol{\theta}\,]$$

where $[\,\mathbf{Y}\,]$ is the marginal distribution given by

$$[\,\mathbf{Y}\,] = \int [\,\mathbf{Y}\,|\,\boldsymbol{\theta}\,]\,[\,\boldsymbol{\theta}\,]\,d\boldsymbol{\theta}$$

Inference for Normal Mean

Suppose $y_1, y_2, \ldots, y_n \overset{iid}{\sim} \text{Normal}(\mu, \sigma^2)$. We will assume σ^2 **is known** for now.

Because μ is unconstrained, let's start with a Normal prior:

$$\mu \sim \text{Normal}(\mu_0, \tau^2)$$

where μ_0 is our prior belief in the central location of μ and τ^2 reflects our uncertainty (i.e. larger $\tau^2 \rightarrow$ less information).

The the data likelihood and prior distributions are given by:

$$[\,\mathbf{y}|\mu\,] = [y_1, y_2, \ldots, y_n|\mu] \propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right)$$

$$[\mu] \propto \exp\left(-\frac{1}{2\tau^2}(\mu - \mu_o)^2\right)$$

Inference for Normal Mean

After some simplification, we obtain the following result for the posterior:

$$\mu|\mathbf{y} \propto \exp\left\{-\frac{1}{2}\left[\mu^2\left(\frac{n}{\sigma^2}+\frac{1}{\tau^2}\right)-2\mu\left(\frac{n}{\sigma^2}\bar{\mathbf{y}}+\frac{1}{\tau^2}\mu_0\right)\right]\right\}$$

We can find constants that can help re-express the above as a normal density. After completing the squares we have:

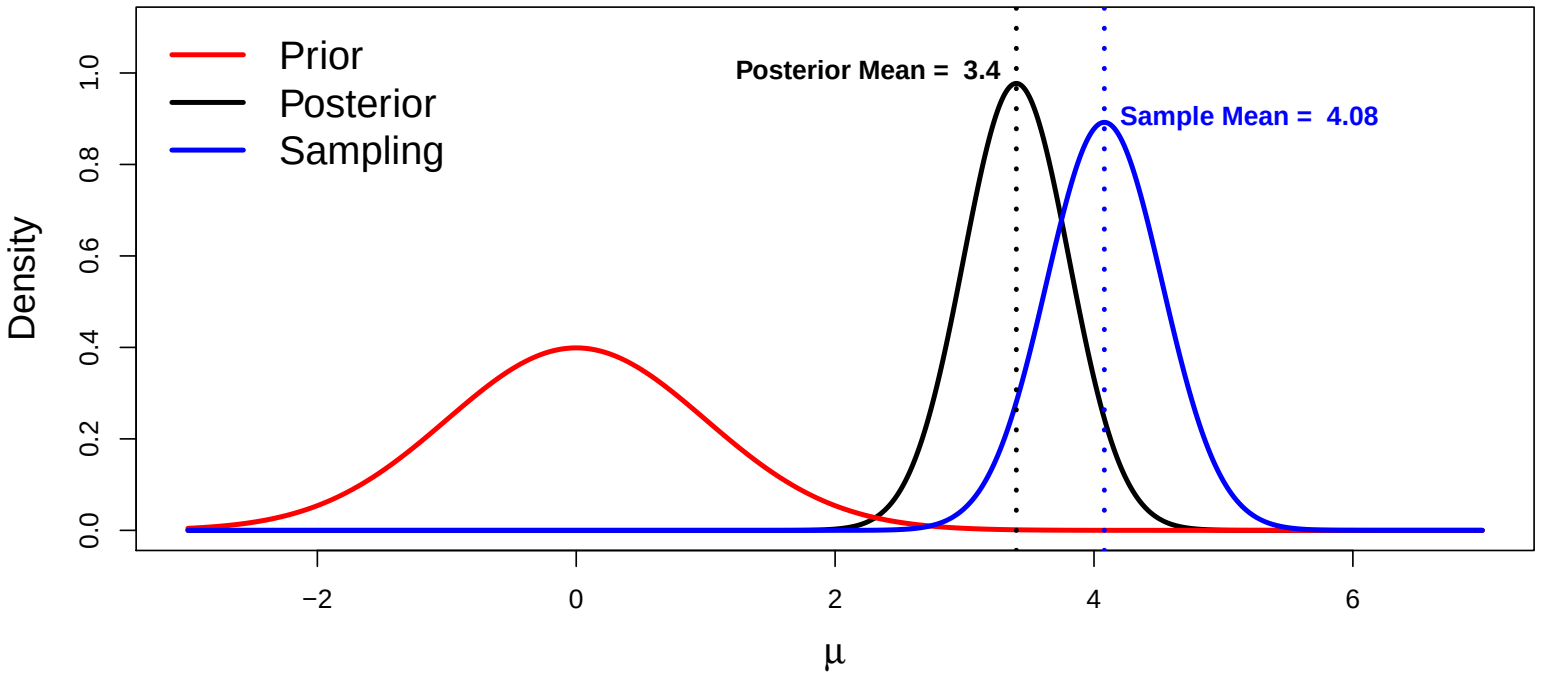
$$\mu \,|\, \mathbf{y} \sim \text{Normal}\left(\bar{\mu}, \bar{\tau}^2\right)$$

$$\bar{\mu} = \left[\frac{n}{\sigma^2}+\frac{1}{\tau^2}\right]^{-1}\left[\frac{n}{\sigma^2}\bar{\mathbf{y}}+\frac{1}{\tau^2}\mu_0\right] \qquad \bar{\tau}^2 = \left[\frac{n}{\sigma^2}+\frac{1}{\tau^2}\right]^{-1}$$

where $\bar{\mathbf{y}}$ is the sample mean $\frac{1}{n}\sum_i y_i$.

When the posterior distribution belongs to the same family as the prior distribution, we say the prior is conjugate for the likelihood.

Normal Mean Example



- Posterior mean is between the sample mean and the prior mean
- Posterior standard deviation is smaller than the sampling variation
- Posterior mean is *biased*

Posterior Distribution

Let's first examine the posterior mean:

$$\begin{aligned}\bar{\mu} &= \left[\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right]^{-1}\left[\frac{n}{\sigma^2}\bar{\mathbf{y}} + \frac{1}{\tau^2}\mu_0\right] \\ &= \frac{n\tau^2}{n\tau^2+\sigma^2}\bar{\mathbf{y}} + \frac{\sigma^2}{n\tau^2+\sigma^2}\mu_0\end{aligned}$$

- This is a weighted average of the prior mean μ_0 and the sample mean $\bar{\mathbf{y}}$
- $\bar{\mu}$ is a biased estimate of μ!
- The weights are inverse of the variance associated with the prior and data-likelihood estimates
  - weight for μ_0 is $1/\tau^2$
  - weight for $\bar{\mathbf{y}}$ is $1/(\sigma^2/n)$
- The posterior mean is closer to $\bar{\mathbf{y}}$ for large n and small σ^2
- The posterior mean is closer to μ_0 for small τ^2
- If $\tau^2 \to \infty$ (increasing prior uncertainty), $\bar{\mu} \to$ the sample mean $\bar{\mathbf{y}}$

Posterior Distribution

Let's look at the posterior variance now:

$$\bar{\tau}^2 = \left[\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right]^{-1} = \frac{\sigma^2\tau^2}{n\tau^2 + \sigma^2}$$

- For a non-zero τ^2, note $\bar{\tau}^2 = \frac{\sigma^2}{n}\left(\frac{\tau^2}{\tau^2+\sigma^2/n}\right) < \frac{\sigma^2}{n}$. So using prior on μ always leads to smaller standard error compared to that from the sample mean.
- This effect is reduced when:
  - n is large (i.e. data outweigh prior information)
  - τ^2 is large (i.e. prior information is not helpful)

Comparison to Frequentist Inference

We know that the sample mean $\bar{\mathbf{y}}$ has distribution:

$$\bar{\mathbf{y}} \sim \text{Normal}\left(\mu, \frac{\sigma^2}{n}\right) \tag{1}$$

Under our Bayesian analysis, if we assume the variance of the prior for μ is ∞, then

$$\mu \sim \text{Normal}\left(\bar{\mathbf{y}}, \frac{\sigma^2}{n}\right) \tag{2}$$

Note the difference in how $\bar{\mathbf{y}}$ and μ are swapped in the above two distributions.

- Equation (1) allows confidence statements on how $\bar{\mathbf{y}}$ behaves with different repeated experiments
- Equation (2) allows probabilistic statement on μ given the observed data

The Inverse-Gamma Distribution

Suppose $y_1, y_2, \ldots, y_n \overset{iid}{\sim} \text{Normal}(\mu, \sigma^2)$. We now will assume μ **is known** and we wish to estimate σ^2.

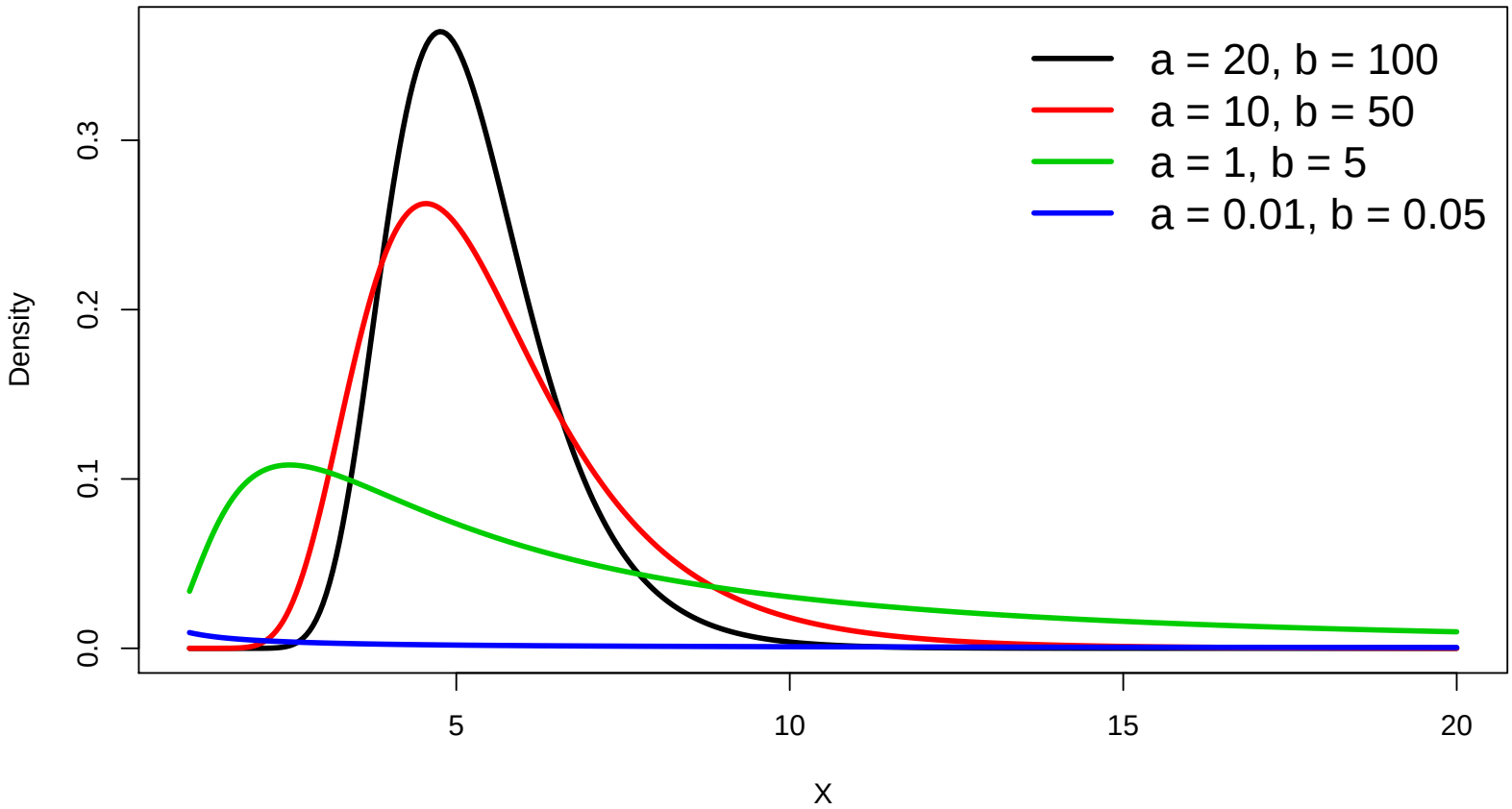
The most common distribution to model variance components is the Inverse-Gamma distribution. The Inv-Gamma distribution has two parameters $\alpha > 0$ and $\beta > 0$. For a random variable y with support $y > 0$, it has density

$$f(y; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} y^{-(\alpha+1)} e^{-\frac{\beta}{y}}$$

Inv-Gamma is a skewed distribution for positive random variables. It has

$$E[Y] = \frac{\beta}{\alpha - 1} \qquad \text{and} \qquad Var[Y] = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$$

Inverse-Gamma (a, b) Density



Bayesian Inference for Population Variance

Let's assume $\sigma^2 \sim$ Inv-Gamma(α, β). Then the posterior distribution is given by

$$
\begin{aligned}
[\sigma^2|\mathbf{y}] &\propto [\mathbf{y}|\sigma^2] \times [\sigma^2] \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(y_i-\mu)^2\right] \times \frac{\beta^\alpha}{\Gamma(\alpha)}(\sigma^2)^{-(\alpha+1)} e^{-\frac{\beta}{\sigma^2}} \\
&\propto (\sigma^2)^{-n/2} \exp\left[-\sum_{i=1}^{n} \frac{1}{2\sigma^2}(y_i-\mu)^2\right] (\sigma^2)^{-(\alpha+1)} e^{-\frac{\beta}{\sigma^2}} \\
&= (\sigma^2)^{-n/2-(\alpha+1)} \exp\left[-\sum_{i=1}^{n} \frac{1}{2\sigma^2}(y_i-\mu)^2 - \frac{\beta}{\sigma^2}\right] \\
&= (\sigma^2)^{-[(n/2+\alpha)+1)]} \exp\left[-\frac{1}{\sigma^2}\left(\frac{1}{2}\sum_{i=1}^{n}(y_i-\mu)^2 + \beta\right)\right]
\end{aligned}
$$

Treating σ^2 as the random variable, notice that

$$
[\sigma^2|\mathbf{y}] \sim \text{Inv-gamma}\left(\frac{n}{2}+\alpha,\ \frac{1}{2}\sum_{i=1}^{n}(y_i-\mu)^2 + \beta\right)
$$

Bayesian Inference for Population Variance

The Inverse-Gamma prior distribution on σ^2 has mean $\frac{\beta}{\alpha-1}$. The posterior Inverse-Gamma distribution has mean

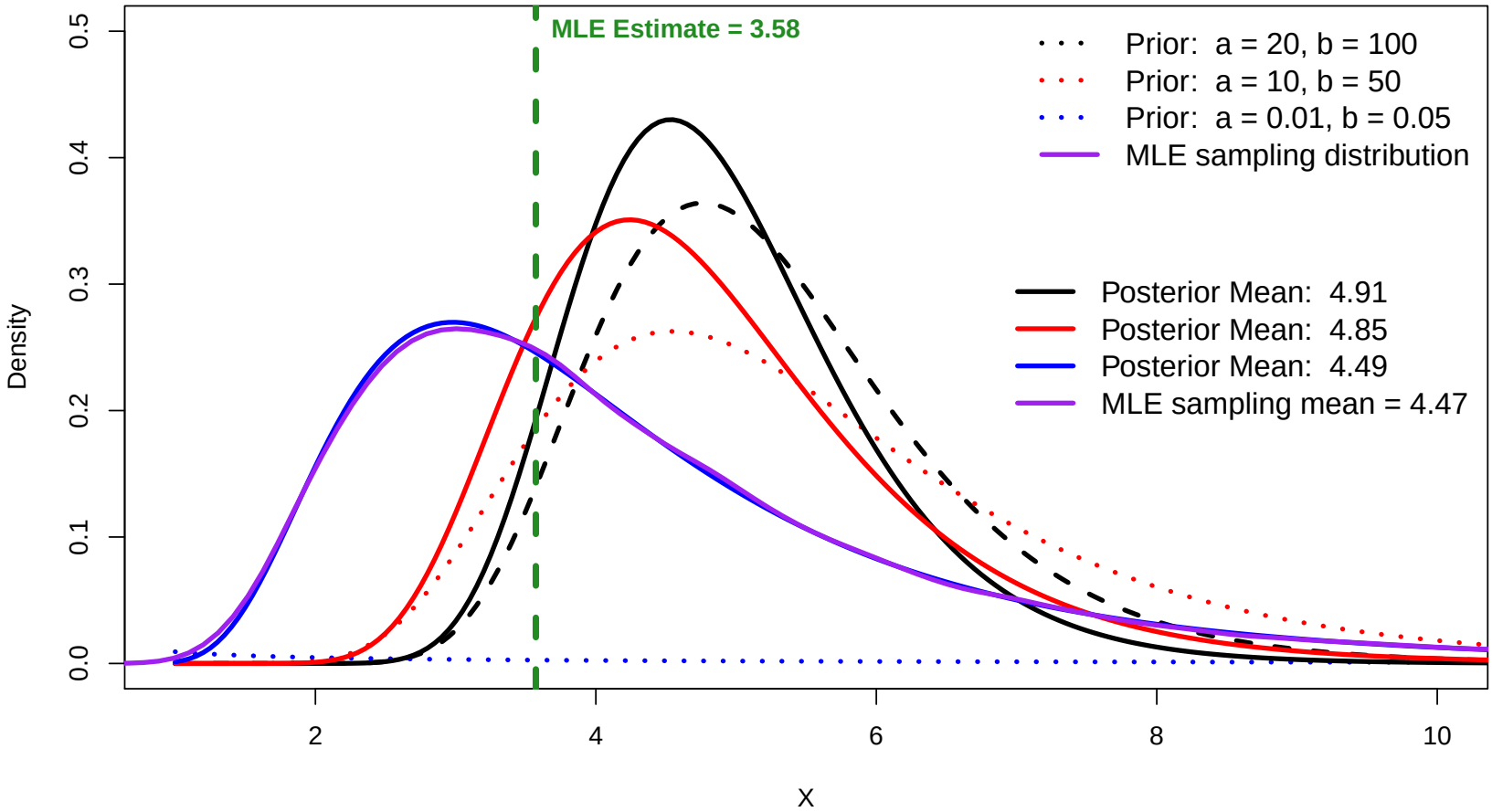
$$E[\sigma^2|\mathbf{y}] = \frac{\frac{1}{2}\sum_{i=1}^{n}(y_i-\mu)^2+\beta}{n/2+\alpha-1}$$

The values of α and β have the interpretation of a pilot study that estimated σ^2 with 2α sample size, and sum of squares 2β

Assume the prior mean $\frac{\beta}{\alpha-1}$ to be fixed. Note that when $\alpha, \beta \to 0$ (less prior information),

$$E[\sigma^2|\mathbf{y}] = \frac{\frac{1}{2}\sum_{i=1}^{n}(y_i-\mu)^2}{n/2-1} \approx \sigma^2_{mle} \quad \text{for large n}$$

Effects of Inverse-Gamma Prior



Prior and Posterior Distributions

- For estimating population mean and variance, we used prior distributions that are the same as their corresponding posterior distributions. These priors are known as conjugate priors. They are convenient because the posterior density is available in closed-form.
- We can often find uninformative/vague priors such that the posterior distributions are very close to that from a Frequentist inference.
- For vague priors and large sample size, we also have Bayesian central limit theorem, where the posterior distribution can be approximated by a normal distribution.
- In practice, data analyses usually utilize uninformative (proper or improper) priors.
- Prior information is particularly useful when the sample size is small.
- Posterior distributions can be viewed as the prior distributions being updated by the data.

Multi-Parameter Model

Often the data likelihood is determined by more than one parameter:

$$[\,\mathbf{y}\,|\,\theta_1, \theta_2, ..., \theta_P]$$

Example:

- Inference for Normal means with unknown variances
- Regression model with multiple regression coefficients

In a Bayesian framework, we need to assign a prior to each unknown parameter. We often assume they are independent:

$$[\theta_1, \theta_2, \ldots, \theta_P] = \prod_{k=1}^{P} [\,\theta_k\,]$$

Our goal is to obtain the joint posterior distribution:

$$[\,\theta_1, \theta_2, ..., \theta_P\,|\,\mathbf{y}]$$

Multi-Parameter Model

The joint posterior distribution not only contains information on the unknown parameters, but also their dependence-structure. Note that the joint posterior $[\theta_1, \theta_2, ..., \theta_P \,|\, \mathbf{y}]$ is different from:

1. the marginal posterior distribution:

$$[\theta_1 \,|\, \mathbf{y}] = \int [\theta_1, \theta_2, ..., \theta_P \,|\, \mathbf{y}] \, d\theta_2, \ldots, d\theta_P$$

   This describes the information on θ_1 after accounting for uncertainties in the other parameters (nuisance parameters).

2. the full conditional distribution:

$$[\theta_1 \,|\, \mathbf{y}, \theta_2, ..., \theta_P] \propto [\mathbf{y} \,|\, \theta_1, \theta_2, ..., \theta_P] \, [\theta_1]$$

   This describes the posterior distribution of θ_1 given the other parameter values are known.

Bayesian Computation

Consider a model with multiple parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_P)$. Here $\boldsymbol{\theta}$ may include regression coefficients and residual errors. To obtain the marginal posterior distributions, it becomes increasingly difficult to perform integration over $\boldsymbol{\theta}$. For example, to make inference on θ_1 we need

$$[\theta_1|\mathbf{y}] = \int \cdots \int [\theta_1, \theta_2, \ldots, \theta_p \,|\, \mathbf{y}]\, d\theta_2 d\theta_3 \ldots d\theta_p$$

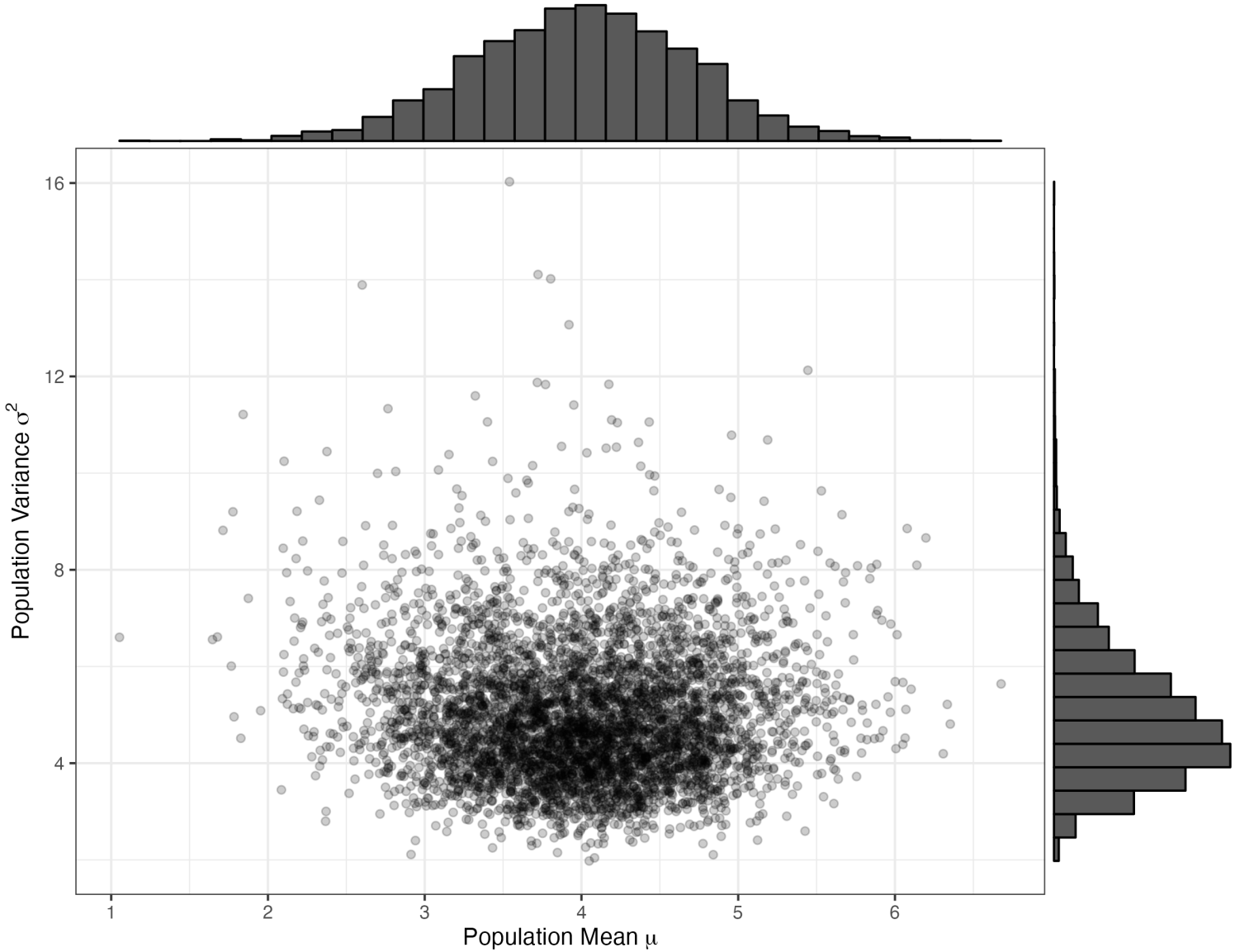
Often, the desired marginal distribution is also not a standard probability density.

One approach is to perform numerical integration. Specifically, we will generate samples/realizations from the joint distribution.

$$\boldsymbol{\theta}^{(k)} \sim [\theta_1, \theta_2, \ldots, \theta_p \,|\, \mathbf{y}]$$

We can then make inference on by examining the marginal distribution of the sampled values.

Joint Distribution of σ^2 and μ



Bayesian Inference via Monte Carlo Simulations

Given K samples of $\mu^{(k)}$, we can compute several useful statistics for inference.

- Posterior mean:

$$E[\mu \,|\, \mathbf{y}] = \int \mu \, [\mu \,|\, \mathbf{y}] \, d\mu \approx \frac{1}{K} \sum_{k=1}^{K} \mu^{(k)}$$

- Posterior variance:

$$Var[\mu \,|\, \mathbf{y}] = \int (\mu - E[\mu \,|\, \mathbf{y}])^2 \, [\mu \,|\, \mathbf{y}] \, d\mu \approx \frac{1}{K} \sum_{k=1}^{K} (\mu^{(k)})^2 - \left[\frac{1}{K} \sum_{k=1}^{K} \mu^{(k)} \right]^2$$

- Exceedance probability:

$$P(\mu > 0 \,|\, \mathbf{y}) = \int_0^{\infty} [\mu \,|\, \mathbf{y}] \approx \frac{1}{K} \sum_{k=1}^{K} I_{\mu^{(k)} > 0}$$

Markov Chain Monte Carlo (Gibbs Sampler)

Markov Chain Monte Carlo (MCMC) is an algorithm to obtain realizations from a high-dimensional probability density. Assume we have three parameters θ_1, θ_2, and θ_3.

1. Set initial values of θ_1, θ_2, and θ_3
2. For the $k+1^{\text{th}}$ iteration
   1. Draw (update) a new $\theta_1^{(k+1)}$ from it's full conditional distribution:
   $$\theta_1^{(k+1)} \sim [\theta_1 | \theta_2^{(k)}, \theta_3^{(k)}, \mathbf{y}]$$
   2. Draw (update)
   $$\theta_2^{(k+1)} \sim [\theta_2 | \theta_1^{(k+1)}, \theta_3^{(k)}, \mathbf{y}]$$
   3. Draw (update)
   $$\theta_3^{(k+1)} \sim [\theta_3 | \theta_1^{(k+1)}, \theta_2^{(k+1)}, \mathbf{y}]$$
3. Repeat step 2 until convergence.
4. Discard the first B-many samples as burn-in (pre-convergence) samples.

Note that the generated samples $\theta_1^{(1)}, \theta_1^{(2)}, \theta_1^{(3)}, \ldots$ are dependent.

Gibbs Sampler for Normal Mean and Variance

Gibbs algorithm to sample μ and σ^2

1. Set initial values (e.g. $\mu = 0$, $\sigma^2 = 1$)
2. For the $k + 1^{\text{th}}$ iteration
   1. Draw $\mu^{(k+1)}$ from

$$\text{Normal}\left(\left[\frac{n}{\sigma^{2,(k)}} + \frac{1}{\tau^2}\right]^{-1}\left[\frac{n}{\sigma^{2,(k)}}\bar{Y} + \frac{1}{\tau^2}\mu_0\right], \left[\frac{n}{\sigma^{2,(k)}} + \frac{1}{\tau^2}\right]^{-1}\right)$$

   2. Draw $\sigma^{2,(k+1)}$ from

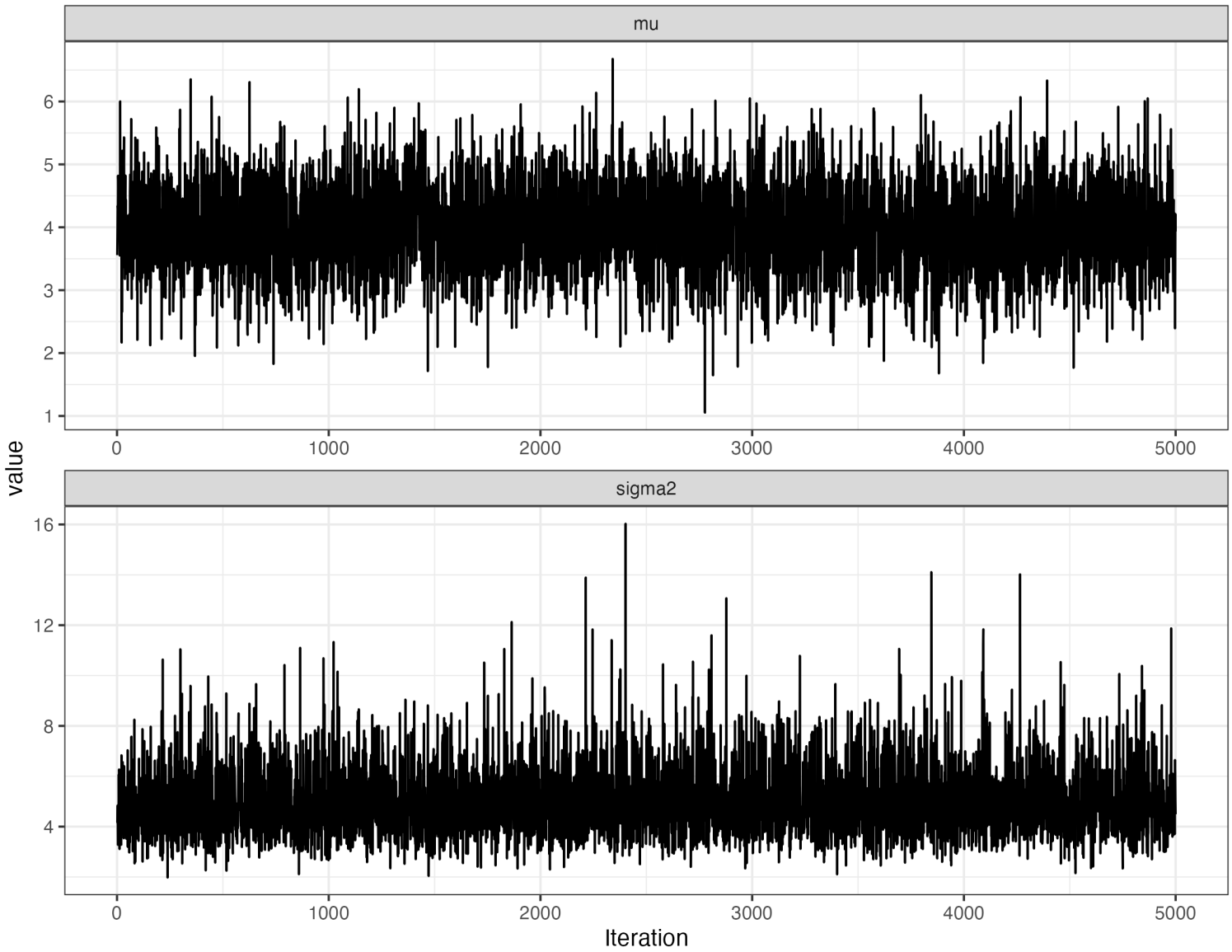
$$\text{Inv-Gamma}\left(\frac{n}{2} + \alpha, \frac{1}{2}\sum_{i=1}^{n}(y_i - \mu^{(k+1)})^2 + \beta\right)$$

3. Repeat step 2 until convergence
4. Discard the first B-many samples as burn-in (pre-convergence) samples

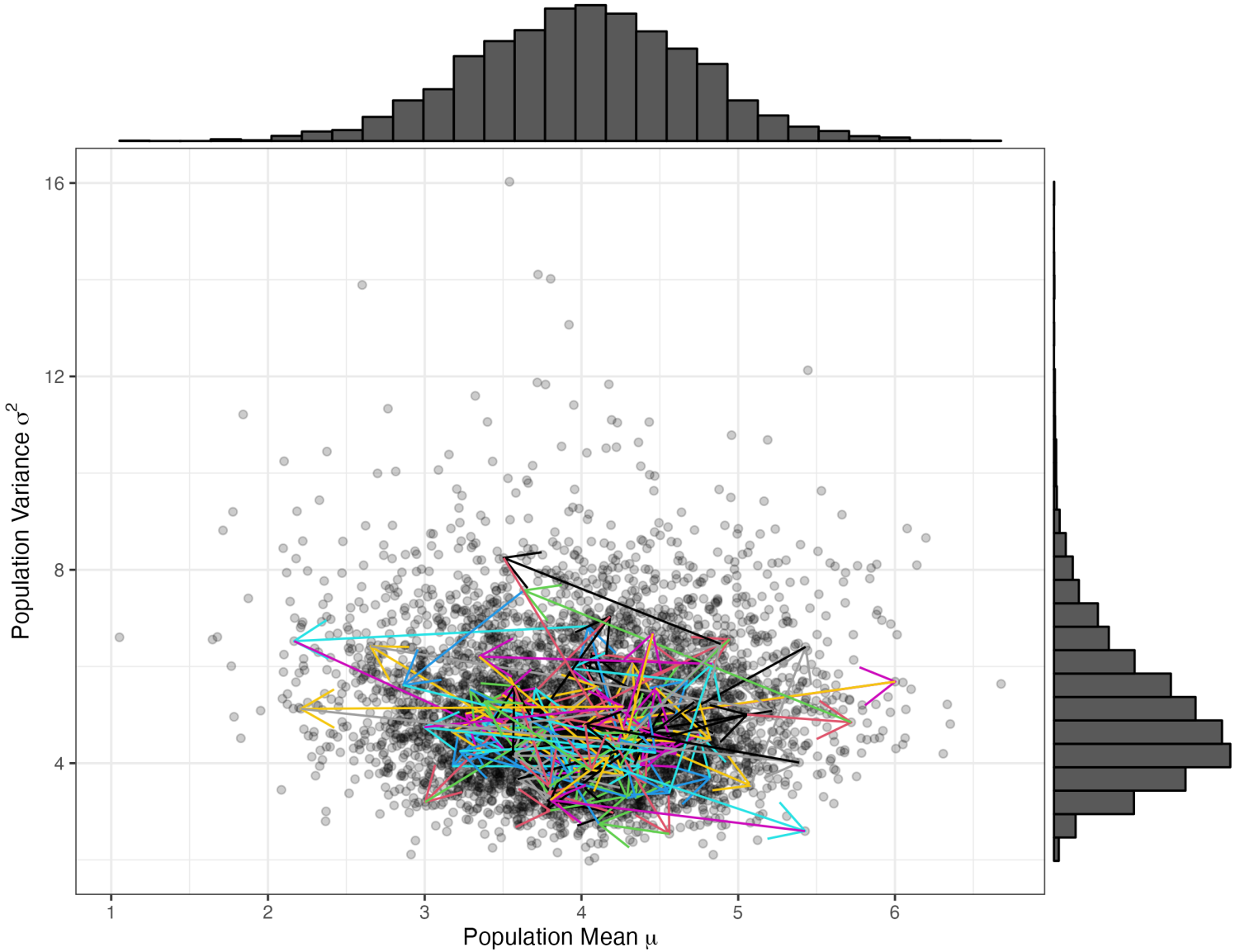
Example R Code

```
y <- c(4.2, 6.6, 5.1, 2.0, 2.8, 3.2, 4.7, 4.1, 7.3, 0.8) # Data
n <- length(y)  # Sample size
mu <- 0         # Initial values for mu
sigma2 <- 1     # Initial values for sigma2
mu0 <- 0        # Prior mean for mu
tau2 <- 5^2     # Prior variance for mu
alpha <- 10     # Prior alpha value for sigma2
beta <- 50      # Prior beta value for sigma2
n.iter <- 5000  # Number of MCMC samples

mu.save <- numeric(n.iter)
sigma2.save <- numeric(n.iter)

# Run Gibbs Sampler!
for(i in 1:n.iter){

  # Step 1) Update mu
  mu.V <- 1 / (n/sigma2 + 1/tau2)
  mu.M <- mu.V * (n/sigma2*mean(y) + mu0 / tau2)
  mu <- rnorm(1, mean = mu.M, sd = sqrt(mu.V))

  # Step 2) Update sigma2
  sigma2 <- 1 / rgamma(1, shape = n/2 + alpha, rate = sum((y - mu)^2)/2 + beta)

  # Store updates
  mu.save[i] <- mu
  sigma2.save[i] <- sigma2
}
```

Trace Plots


mu
sigma2
value
Iteration
1
2
3
4
5
6
4
8
12
16
0
1000
2000
3000
4000
5000

Gibbs Sampler Paths



Other Sampling Approaches

The Gibbs sampler example above is possible because we have the full conditional distributions of μ and σ^2 in closed-form that are easy to sample from.

Additional samplers have been developed to sample from a density that we only know up to a proportional constant.

- Adaptive rejection sampling
- Slice sampler
- Random-walk Metropolis-Hastings
- Hamiltonian Metropolis-Hastings

These algorithms require tuning that determines how to move (jump) efficiently in the multi-dimensional posterior sample space. BUGS, JAGS, and STAN will employ these automatically when Gibbs sampler is not available.

Linear Regression Model

We will first consider the following multiple linear regression model. For $i = 1, \ldots, n$, assume

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} \text{Normal}(0, \sigma^2)$$

Here we have $(p+1)$ unknown parameters $(\beta_1, \ldots, \beta_p, \sigma^2)$.

The above model can be written in matrix form and the observed data vector $\mathbf{y}$ has distribution

$$\mathbf{y} \sim \text{Normal}\left(\mathbf{X}\boldsymbol{\beta},\ \sigma^2 \mathbf{I}_{n\times n}\right)$$

with likelihood

$$[\mathbf{y}|\boldsymbol{\beta}, \sigma^2] = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right]$$

To carry out Bayesian inference, we need to assign prior distributions for $\boldsymbol{\beta}$ and σ^2.

Priors for Linear Regression Model

The standard priors for linear regression parameters are non-informative multivariate normal distribution for $\boldsymbol{\beta}$ and inverse-gamma for σ^2:

$$\boldsymbol{\beta} \sim \text{Normal}\left(\boldsymbol{\mu}_0,\, \tau^2 \mathbf{I}_{p\times p}\right) \qquad \text{and} \qquad \sigma^2 \sim \text{Inv-Gamma}(c_0, d_0)$$

where τ^2 is set to be large (e.g. 10000), c_0 and d_0 are set to be small (e.g. 0.0001), and $\boldsymbol{\mu}_0$ is often set to be $\mathbf{0}$.

The above also assume that the regression coefficients are *a priori* independent of each other. These hyper-parameters choices attempt to reflect a lack of prior information for the parameters.

Estimation is usually carried out using a Gibbs sampler that updates $\boldsymbol{\beta}$ and σ^2 iteratively. We will need to calculate the full conditional distribution of

$$[\boldsymbol{\beta} \,|\, \sigma^2, \mathbf{y}] \qquad \text{and} \qquad [\sigma^2 \,|\, \boldsymbol{\beta}, \mathbf{y}]$$

It terms out that the above two distributions have closed-form solutions.

Full Conditional Distribution for Linear Regression

- Full conditional distribution of $\boldsymbol{\beta}$

$$[\boldsymbol{\beta} \,|\, \sigma^2, \mathbf{y}] \sim \text{Normal}\left(\tilde{\mu}, \tilde{\mathbf{V}}\right)$$

where

$$\tilde{\mu} = \left[\sigma^{-2}(\mathbf{X}'\mathbf{X}) + \tau^{-2}\mathbf{I}_{p\times p}\right]^{-1}\left[\sigma^{-2}(\mathbf{X}'\mathbf{y}) + \tau^{-2}\boldsymbol{\mu}_0\right]$$

$$\tilde{\mathbf{V}} = \left[\sigma^{-2}(\mathbf{X}'\mathbf{X}) + \tau^{-2}\mathbf{I}_{p\times p}\right]^{-1}$$

- Full conditional distribution of σ^2

$$[\sigma^2 \,|\, \boldsymbol{\beta}, \mathbf{y}] \sim \text{Inv-Gamma}\left(\tilde{c}, \tilde{d}\right)$$

where

$$\tilde{c} = \frac{n}{2} + c_0 \qquad \text{and} \qquad \tilde{d} = \frac{\sum_{i=1}^{n}(y_i - \mathbf{x}_i\boldsymbol{\beta})^2}{2} + d_0$$

Note the similarities between the above two expressions and what we saw for the univariate Normal mean and variance.

Bayesian Random Intercept Model

A Bayesian version of the random intercept model is complete with prior distributions assigned to all the unknown parameters.

$$y_{ij} = \alpha_i + \mathbf{X}_{ij}^T \boldsymbol{\beta} + \epsilon_{ij}$$

$$\alpha_i \sim \text{Normal}(\alpha_0, \tau^2) \qquad \epsilon_{ij} \sim \text{Normal}(0, \sigma^2)$$

Priors:

$$\begin{aligned}
\boldsymbol{\beta} &\sim \text{Normal}(\mathbf{0}, v_\beta^2 \mathbf{I}_{p\times p}) \\
\alpha_0 &\sim \text{Normal}(0, v_\alpha^2) \\
\sigma^2 &\sim \text{Inv-Gamma}(\, c_\sigma,\, d_\sigma\,) \\
\tau^2 &\sim \text{Inv-Gamma}(\, c_\tau,\, d_\tau\,)
\end{aligned}$$

To reflect uninformative priors, we often set v_β^2 and v_α^2 to be large (e.g. 1000^2) and $c_\sigma, d_\sigma, c_\tau, d_\tau$ to be small.

Viewing $\alpha_i \sim \text{Normal}(\alpha_0, \tau^2)$ as another prior distribution, here we have 3 levels of hierarchy!

Bayesian Random Slope Model

Similarly, for random intercept and a random slope on covariate x_{ij}:

$$y_{ij} = \alpha_{0i} + \alpha_{1i} x_{ij} + \mathbf{X}_{i,-j}^T \boldsymbol{\beta} + \epsilon_{ij}$$

$$\begin{bmatrix} \alpha_{0i} \\ \alpha_{1i} \end{bmatrix} \sim \text{Normal}\left(\boldsymbol{\alpha} = \begin{bmatrix} \alpha_0 \\ \alpha_1 \end{bmatrix}, \boldsymbol{\Sigma}_{2\times 2} \right) \qquad \epsilon_{ij} \sim \text{Normal}(0, \sigma^2)$$

Priors:

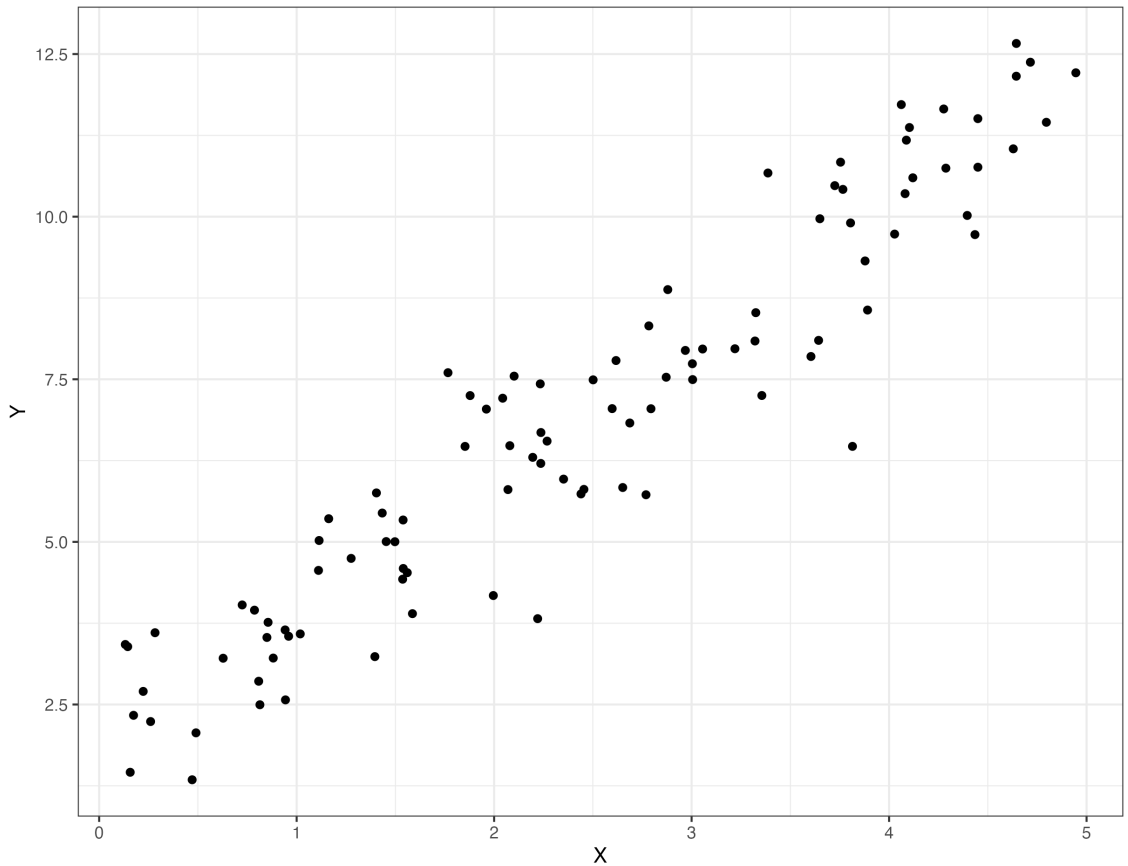
$$\begin{aligned} \boldsymbol{\beta} &\sim \text{Normal}(\, \mathbf{0}, v_\beta^2 \mathbf{I}_{p\times p}\,) \\ \boldsymbol{\alpha} &\sim \text{Normal}(\, \mathbf{0},\, v_\alpha^2 \mathbf{I}_{2\times 2}\,) \\ \sigma^2 &\sim \text{Inv-Gamma}(\, c_\sigma,\, d_\sigma\,) \\ \boldsymbol{\Sigma} &\sim \text{Bivariate Inv-Gamma}(\, r,\, \mathbf{R}\,) \end{aligned}$$

We treat the random effects as realizations from a bivariate Normal distribution with population mean $\boldsymbol{\alpha}$ and population covariance $\boldsymbol{\Sigma}$.

Linear Regression Example

Consider a simple linear regression analysis. We wish to fit the model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} \text{Normal}(0, \sigma^2)$$



Installing JAGS

We will use JAGS (Just Another Gibbs Sampler) to fit Bayesian models. JAGS is a cross-platform engine for the BUGS language.

Some housekeeping:

- Install JAGS here (make sure to select the correct version for your OS)
- Install the R2jags R package

Other R packages such as brms, rstanarm, MCMCpack, NIMBLE, and rethinking may be used to fit Bayesian models as well.

Working with JAGS in R

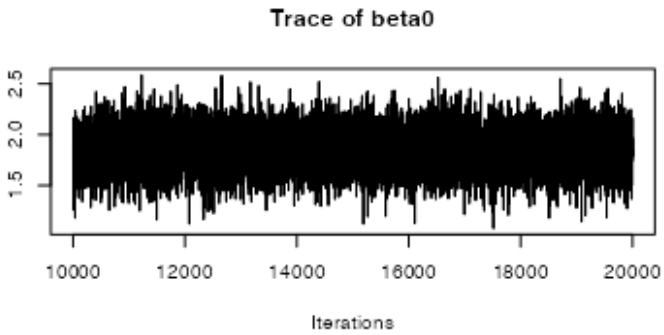
```
> library(R2jags)
> jags
function (data, inits, parameters.to.save, model.file = "model.bug",
    n.chains = 3, n.iter = 2000, n.burnin = floor(n.iter/2),
    n.thin = max(1, floor((n.iter - n.burnin)/1000)), DIC = TRUE, jags.seed = 123, ...)
```

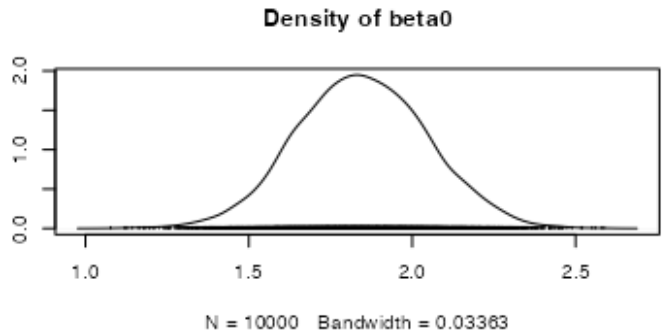
- data: a vector or list of the data objects used by the model
- inits: a list of starting values
- parameters.to.save: character vector of the parameters to save
- model.file: file containing the model written in BUGS code
- n.chains: number of Markov chains (deafult: 3)
- n.iter: number of total iterations including burn-in
- n.burnin: number of burn-in iterations
- n.thin: thinning rate
- DIC: logical; if TRUE (default) compute deviance, pD, and DIC
- jags.seed: random seed for JAGS. Same seed = same posterior samples

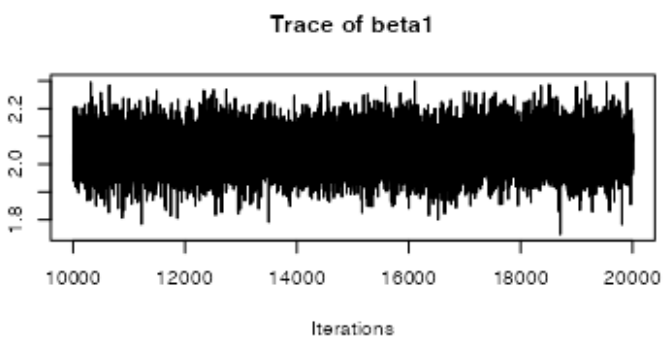
Linear Regression Example - JAGS Model

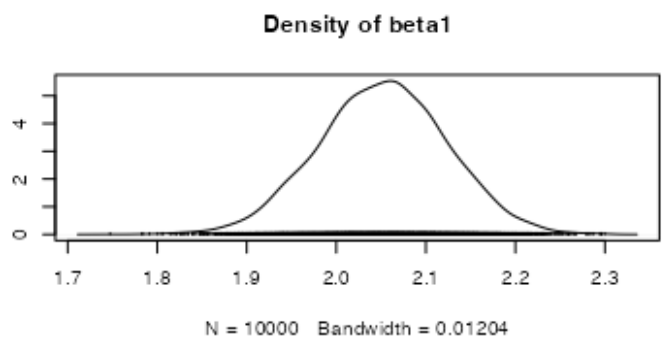
```
# Simulate linear regression data
n <- 100
X <- runif(n, 0, 5)
Y <- 2 + 2*X + rnorm(n)

# Specify JAGS model
jags.SLR <- function(){

  # Data Likelihood
  for(i in 1:n){
    Y[i] ~ dnorm(mu[i], sigma2.inv)
    mu[i] = beta0 + beta1 * X[i]
  }

  # Prior distribution
  beta0 ~ dnorm(0,  1e-06)
  beta1 ~ dnorm(0,  1e-06)
  sigma2.inv ~ dgamma(0.0001, 0.0001)
  sigma2 = 1 / sigma2.inv
}

# Fit JAGS model
SLR.fit <- jags(data = list("Y", "X", "n"), model.file = jags.SLR,
                parameters.to.save = c("beta0", "beta1", "sigma2"), DIC = FALSE,
                n.chain = 1, n.iter = 20000, n.burnin = 10000, n.thin = 1)
SLR.mcmc <- as.mcmc(SLR.fit)
```

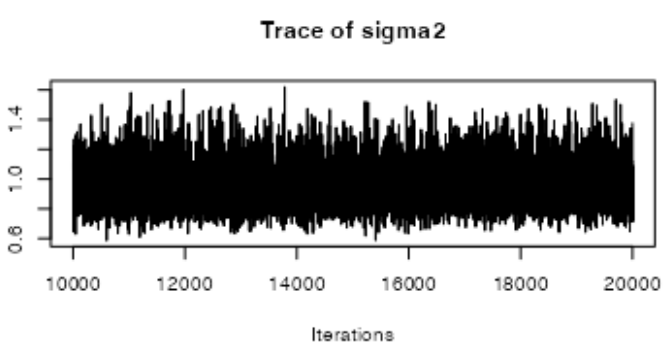
Trace Plots and Marginal Posterior Densities

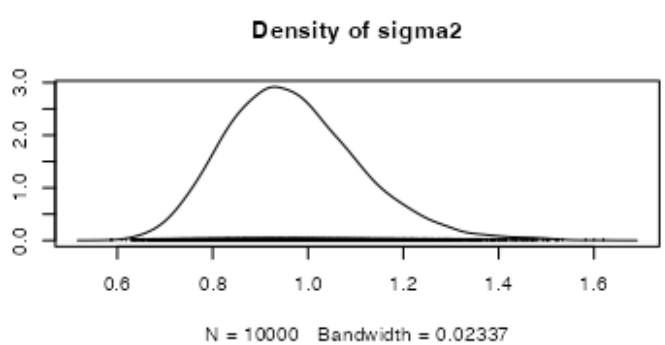












Posterior Summary Statistics

```
1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

         Mean      SD  Naive SE Time-series SE
beta0  1.8459 0.20015 0.0020015      0.0020762
beta1  2.0516 0.07183 0.0007183      0.0007339
sigma2 0.9681 0.14187 0.0014187      0.0014620

2. Quantiles for each variable:

         2.5%    25%    50%   75% 97.5%
beta0  1.4575 1.7095 1.8438 1.981 2.240
beta1  1.9106 2.0038 2.0527 2.100 2.190
sigma2 0.7292 0.8677 0.9557 1.054 1.285
```

- Mean: mean of the posterior distribution (point estimate)
- SD: std. deviation of the posterior distribution (uncertainty measure)
- Naive SE: standard error of the posterior mean ($\sqrt{\text{SD}/\text{iterations}}$). This is a measure of Monte Carlo error for the point estimate
- Time-series SE: same as *Naive SE* but corrects for dependent samples. You want this to be similar to *Naive SE*
- Quantiles: quantile values of the posterior distribution (interval estimate)

Posterior Summary Statistics

```
1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

          Mean      SD  Naive SE Time-series SE
beta0  1.8459 0.20015 0.0020015      0.0020762
beta1  2.0516 0.07183 0.0007183      0.0007339
sigma2 0.9681 0.14187 0.0014187      0.0014620

2. Quantiles for each variable:

          2.5%    25%    50%   75% 97.5%
beta0  1.4575 1.7095 1.8438 1.981 2.240
beta1  1.9106 2.0038 2.0527 2.100 2.190
sigma2 0.7292 0.8677 0.9557 1.054 1.285
```

- A 95% posterior interval for β_1 is:
  - $2.05 \pm 1.96 \times 0.072$ (assuming normality of the posterior)
  - $(1.91, 2.19)$ (without normality assumption)
- The posterior medians of the regression coefficients are close to the posterior means. So the posterior distributions appear to be symmetric
- Because the posterior distribution of σ^2 is skewed to the right, an appropriate 95% posterior interval is $(0.73, 1.29)$

Compare with Frequentist Results

```
> summary(lm(Y ~ X))

Residuals:
    Min      1Q  Median      3Q     Max
-3.2036 -0.5318  0.0283  0.7694  2.1301

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.84819    0.19934   9.272 4.61e-15 ***
X            2.05103    0.07158  28.654  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9735 on 98 degrees of freedom
Multiple R-squared:  0.8934,Adjusted R-squared:  0.8923
F-statistic:   821 on 1 and 98 DF,  p-value: < 2.2e-16
```

The regression coefficients are nearly identical to the Bayesian analysis, and $\hat{\sigma}^2 = 0.9735^2 = 0.9477$ is close to before. This is because we used non-informative priors.

Toenail Dataset

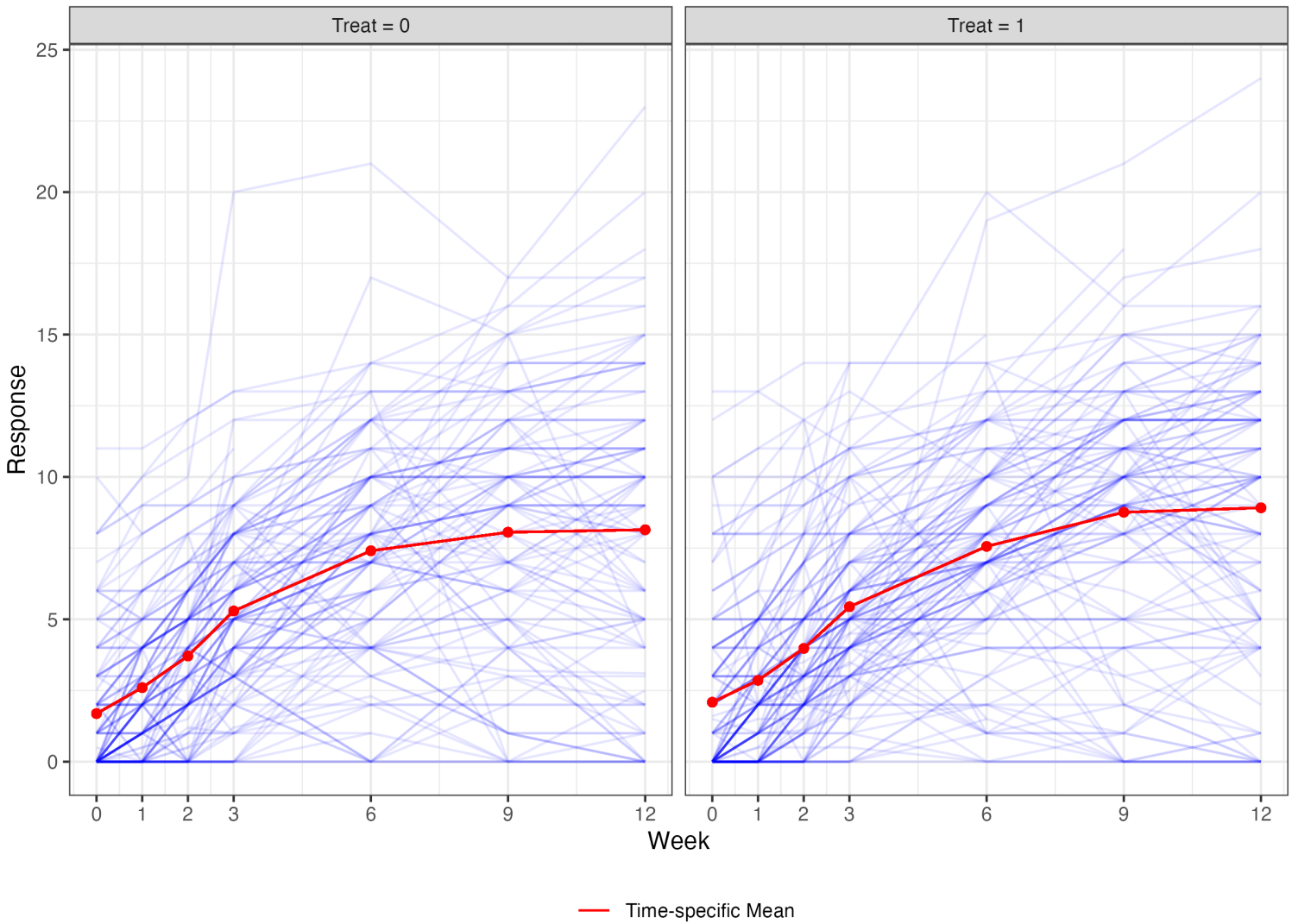
A multi-center randomized controlled trial with 298 subjects to compare the effects of two drugs on toenail infection. Response to treatment was evaluated at week 0, 1, 2, 3, 6, 9, and 12.

Variables:

- *id*: subject ID
- *response*: length of unaffected nail length for the big toenail (mm)
- *treat*: 0 for Itraconazol and 1 for Laisil
- *time*: number of weeks

We have longitudinal data with measurements nested within subjects.

Toenail Response Trajectory



Interaction Model for the Toenail Data

Consider a model with interaction between treatment and time:

$$y_{ij} = \alpha_i + \beta_1 treat_i + \beta_2 time_{ij} + \beta_3 * treat_i \times time_{ij} + \epsilon_{ij}$$

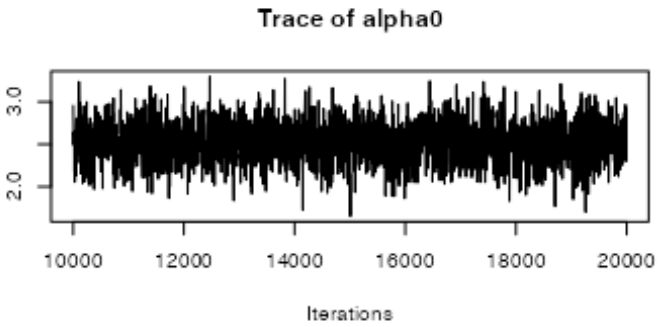
$$\alpha_i \sim N(\alpha_0, \tau^2) \qquad \epsilon_{ij} \sim N(0, \sigma^2)$$

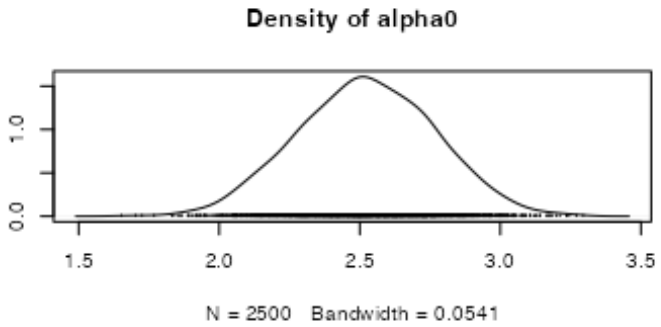
$$\begin{aligned}
\alpha_0 &\sim \text{Normal}(0, 100,000^2) \\
\boldsymbol{\beta} &\sim \text{Normal}(\mathbf{0}, 100,000^2 \mathbf{I}_{p\times p}) \\
\sigma^2 &\sim \text{Inv-Gamma}(\, 0.0001,\, 0.0001\,) \\
\tau^2 &\sim \text{Inv-Gamma}(\, 0.0001,\, 0.0001\,)
\end{aligned}$$

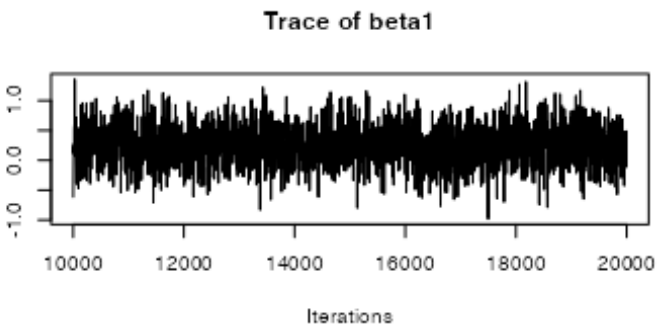
Random Intercept Model in JAGS

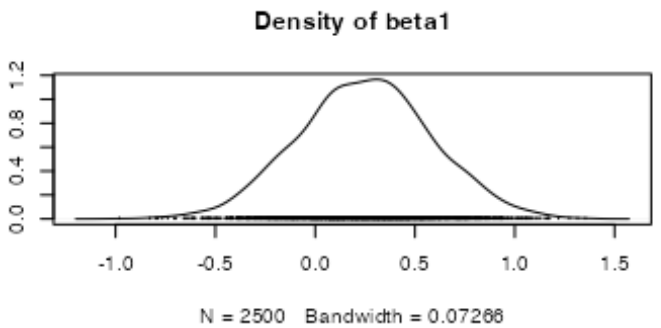
```
# Random Intercept Model
toe.jags1 <- function(){

  # Data Likelihood
  for(i in 1:n.obs){
    response[i] ~ dnorm(mu[i], prec1)
    mu[i] = alpha[id[i]] + beta1*treat[i] + beta2*time[i]+ beta3*time[i]*treat[i]
  }

  for(j in 1:n.id){ alpha[j] ~ dnorm(alpha0, prec2) }

  # Priors
  alpha0 ~ dnorm(0, 1E-10)
  beta1 ~ dnorm(0, 1E-10)
  beta2 ~ dnorm(0, 1E-10)
  beta3 ~ dnorm(0, 1E-10)
  prec1 ~ dgamma(0.0001, 0.0001)
  prec2 ~ dgamma(0.0001, 0.0001)
  sigma2 = 1/prec1
  tau2 = 1/prec2
}

toe.fit1 <- jags(data = list("response", "treat", "time", "id", "n.id", "n.obs"),
                 model.file = toe.jags1, DIC = FALSE, jags.seed = 1031,
                 parameters.to.save = c("alpha0","beta1","beta2","beta3","sigma2","tau2"),
                 n.chain = 1, n.iter = 20000, n.burnin = 10000, n.thin = 4)
toe.mcmc0 <- as.mcmc(toe.fit1)
```

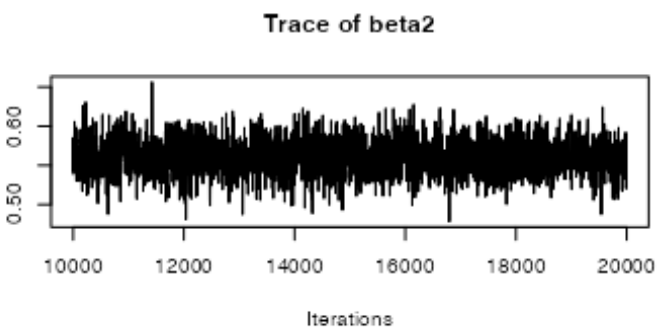
Trace Plots and Marginal Posterior Densities

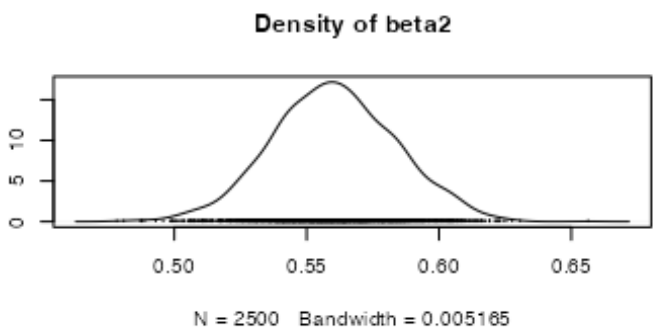




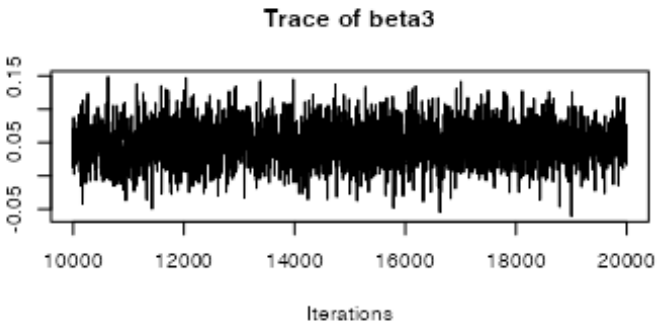


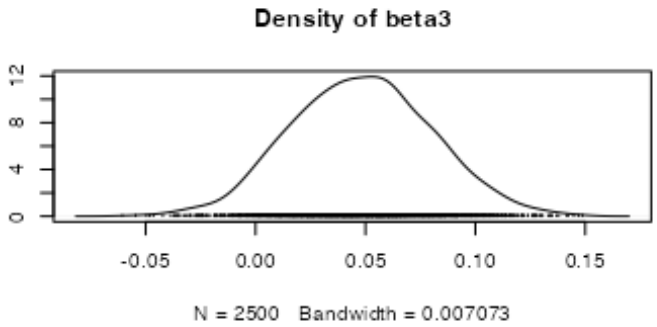


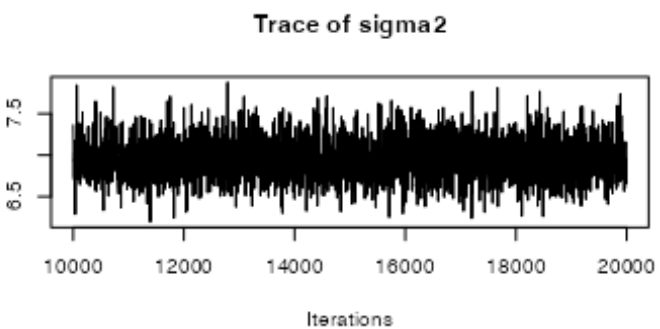


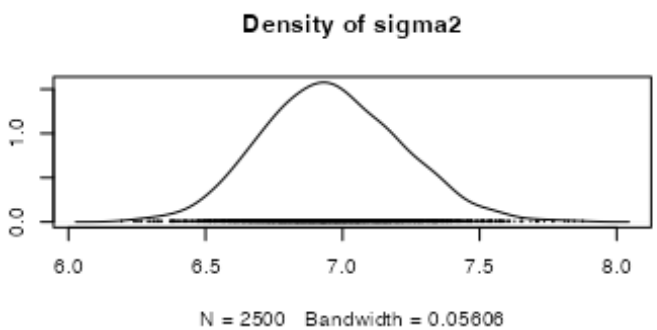


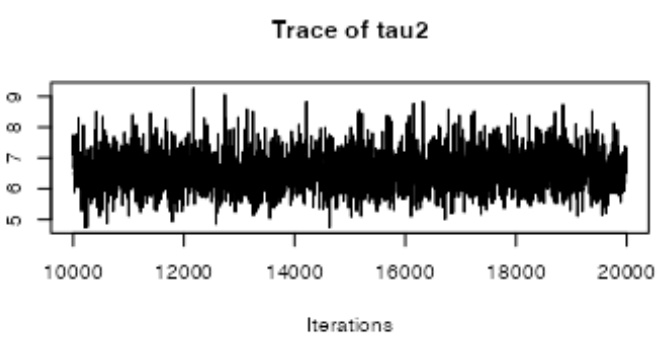
Trace Plots and Marginal Posterior Densities

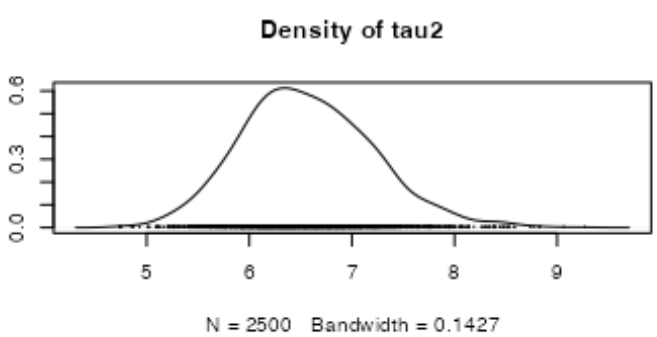




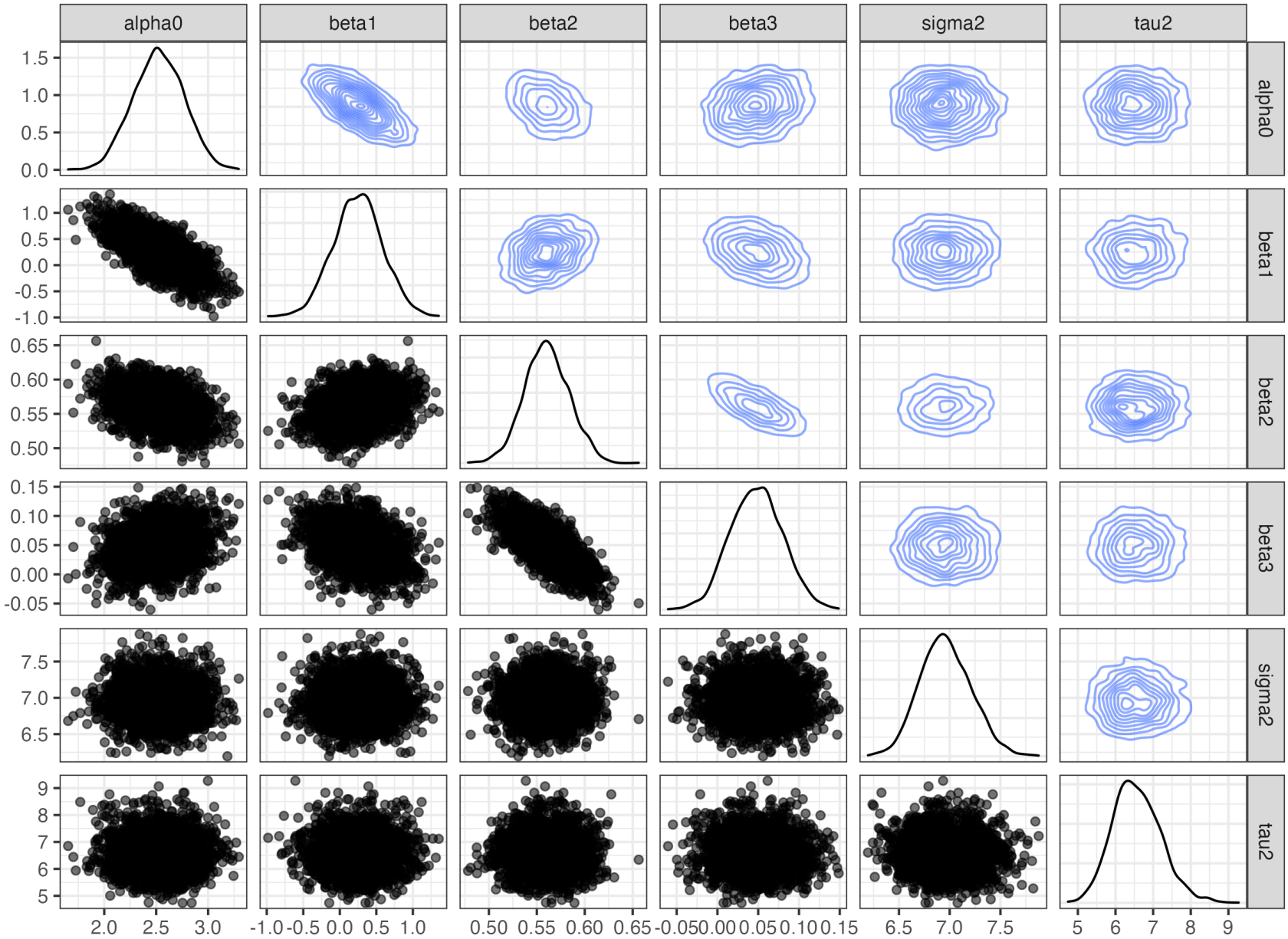








Pairwise Joint Posteriors



Making Posterior Inference

Given our fitted model, what kind of questions can we answer?

- Point estimates and intervals for regression coefficients
- Point estimates and intervals for variance components
- Point estimates and intervals for functions of parameters
- Predictions for in- and out-of-sample subjects
- Probability statements about parameters
- Comparison of subject-specific random effects

Posterior Inference: Parameter Estimates

Parameter	Post Mean	95% Post Int
α_0	2.52	(2.02, 3.02)
β_1	0.24	(-0.47, 0.95)
β_2	0.56	(0.52, 0.61)
β_3	0.05	(-0.01, 0.11)
σ^2	6.96	(6.47, 7.48)
τ^2	6.58	(5.40, 7.96)

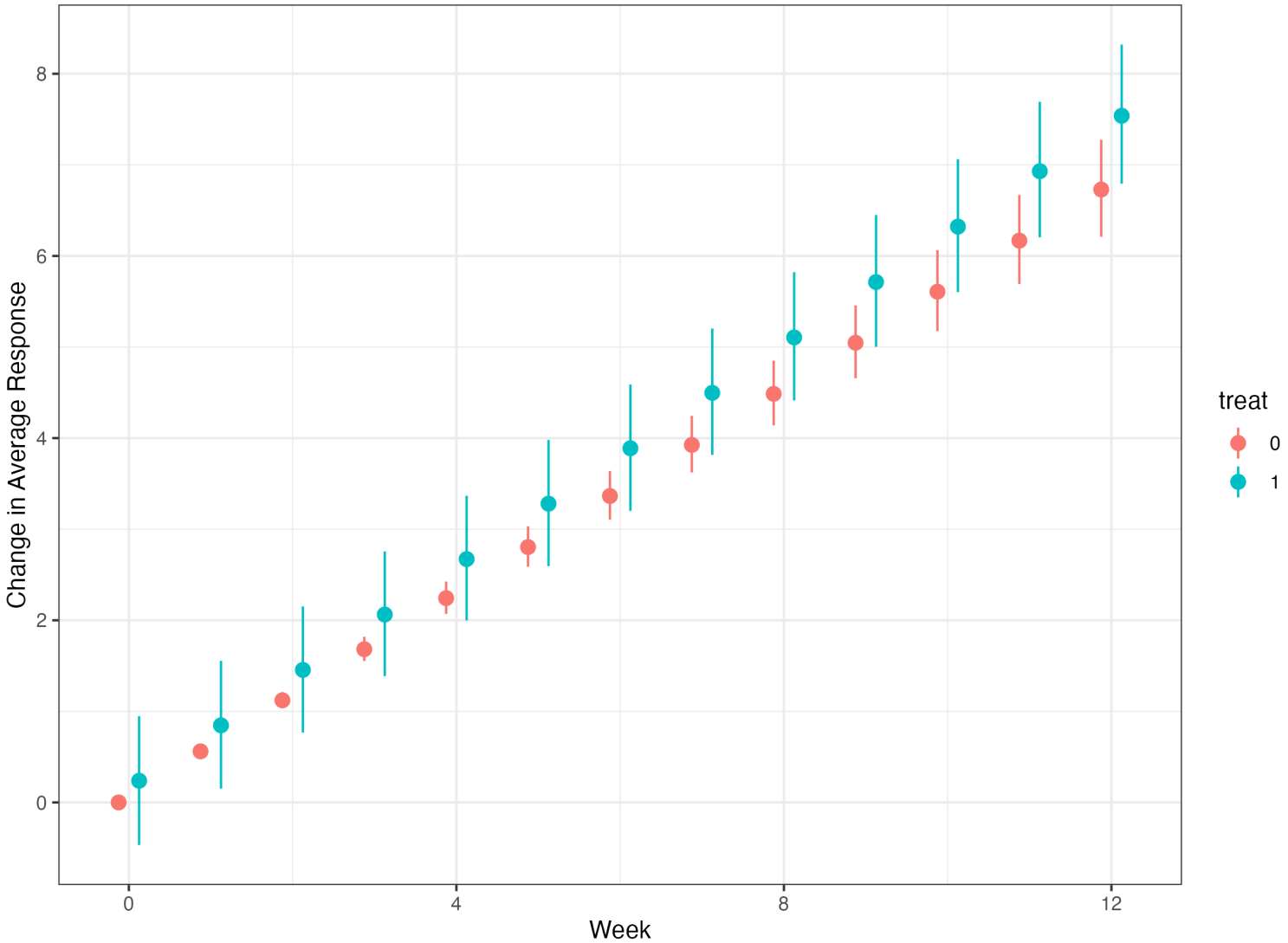
Compared to LMM results:

Parameter	Estimate	95% CI
α_0	2.52	(2.03, 3.00)
β_1	0.25	(-0.43, 0.93)
β_2	0.56	(0.52, 0.61)
β_3	0.05	(-0.01, 0.11)
σ^2	6.95	
τ^2	6.54	

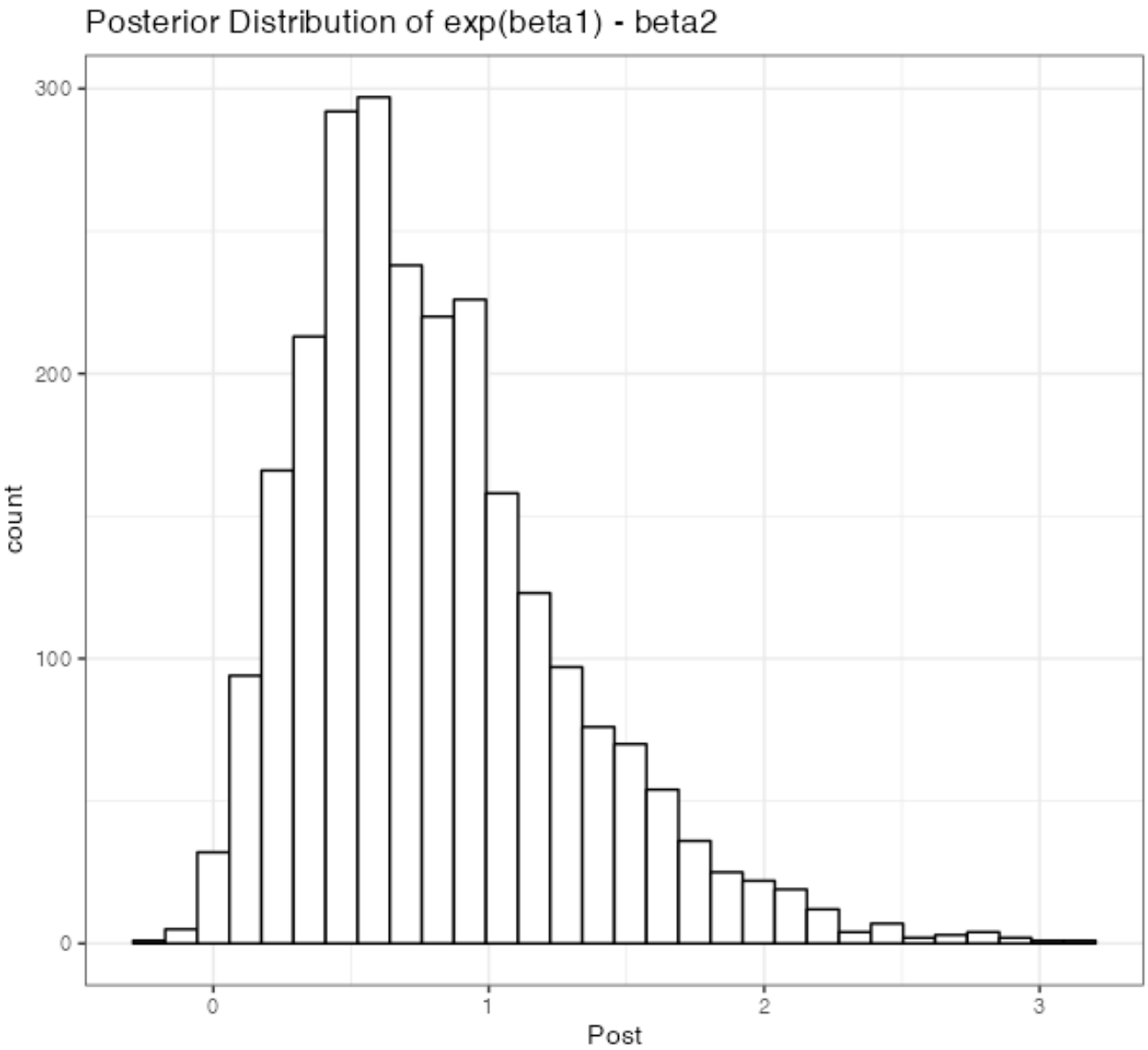
There is some evidence that treatment effect increases with time ($\beta_3 > 0$).

Changes in Average Response by Week

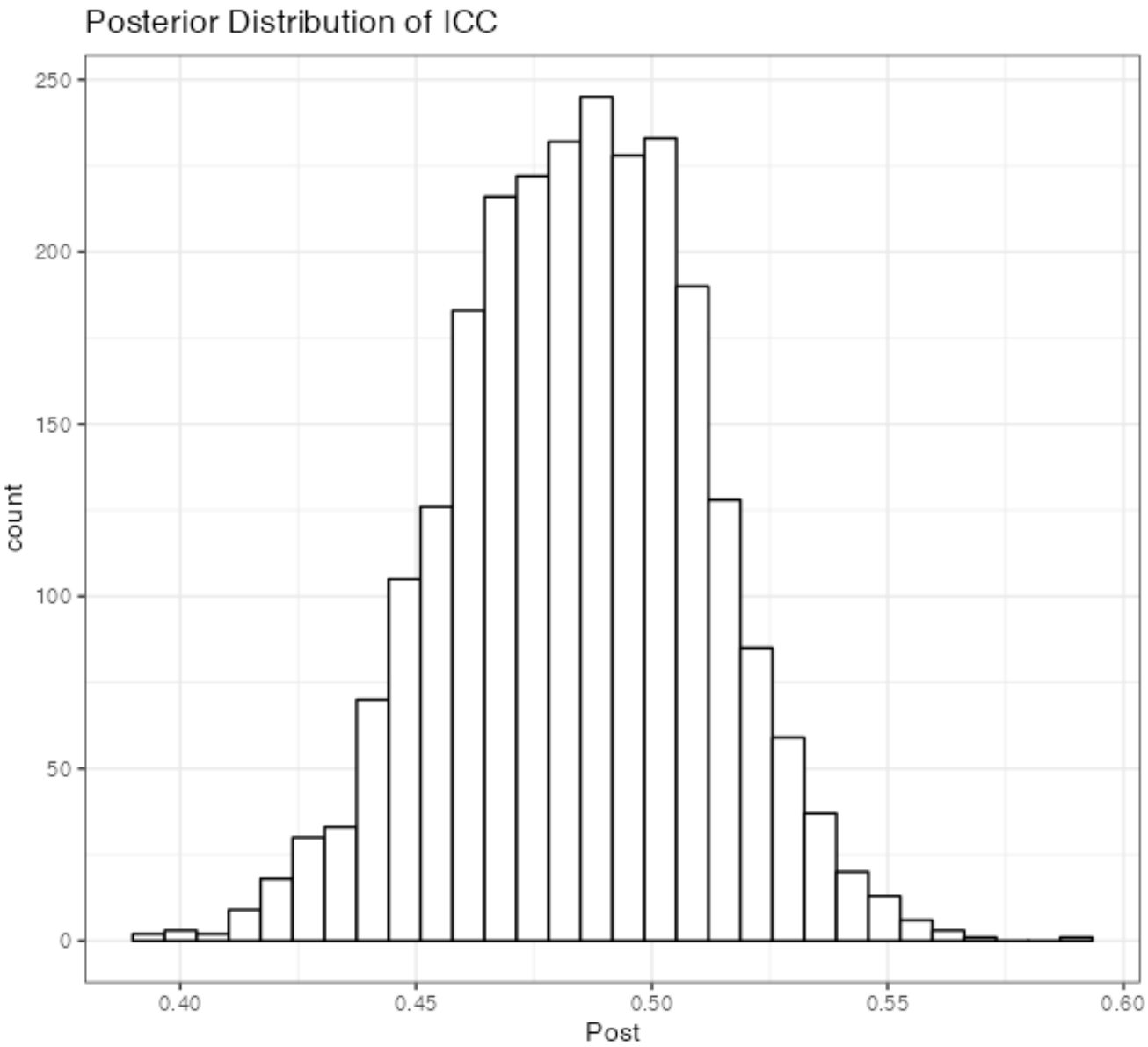
Posterior Mean and 95% Credible Interval by week and treatment.



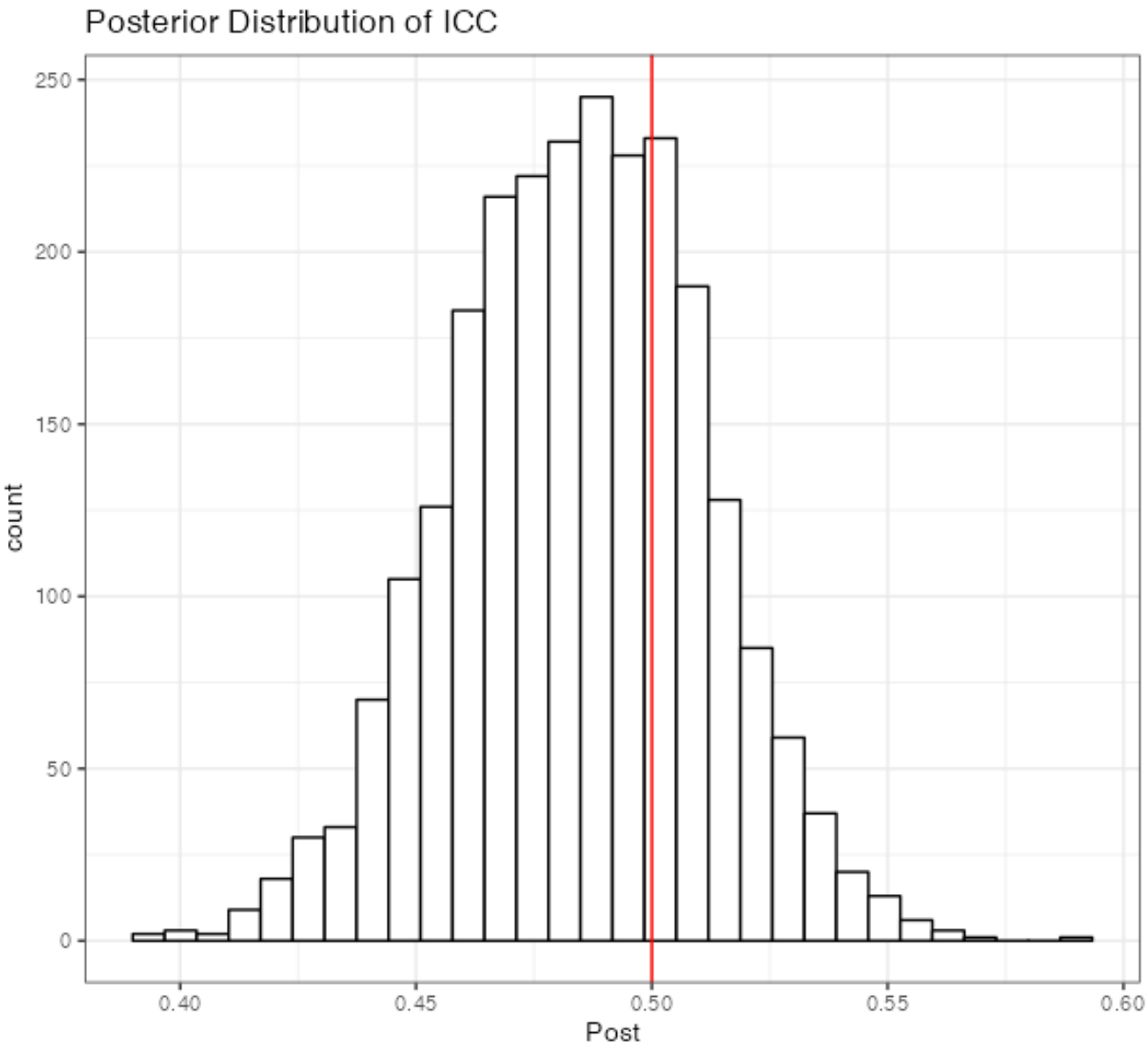
Posterior Inference: $\exp(\beta_1) - \beta_2$



Posterior Inference: ICC

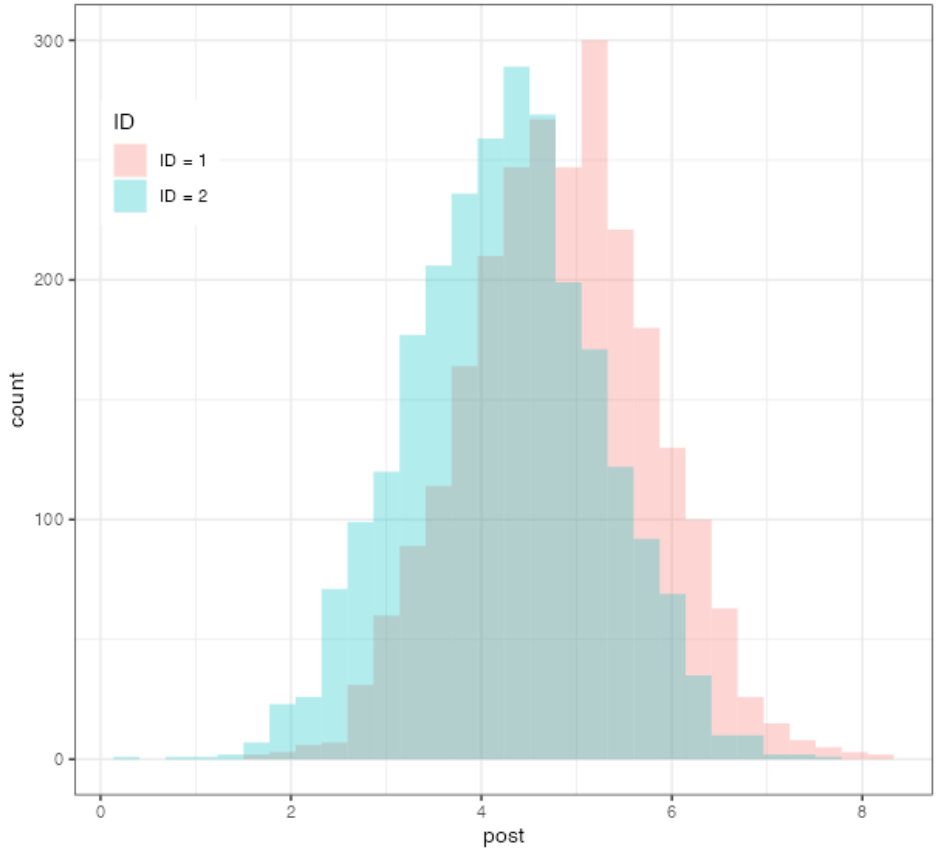


Posterior Inference: Pr(ICC > 0.5)



Posterior Inference: Random Effects

We also have posterior samples of every random effect α_i.



Evidence that the random effect for subject 1 is greater than for subject 2?

```
> mean(toe.post0$alpha.1. > toe.post0$alpha.2.)
[1] 0.6652
```

Posterior Inference: Predictions

Given a hierarchical model:

$$y_{ij} = \alpha_i + \mathbf{X}_{ij}^T \boldsymbol{\beta} + \epsilon_{ij}$$

$$\alpha_i \sim N(\alpha_0, \tau^2) \qquad \epsilon_{ij} \sim N(0, \sigma^2)$$

we have two types of predictions.

1. New observation for an observed cluster (in-sample):

$$y_{ij^*} \sim \text{Normal}\left(\alpha_i + \mathbf{X}_{ij^*}^T \boldsymbol{\beta},\ \sigma^2\right)$$

2. New observation for a new cluster (out-of-sample:

$$y_{i^*j^*} \sim \text{Normal}\left(\alpha_0 + \mathbf{X}_{i^*j^*}^T \boldsymbol{\beta},\ \sigma^2 + \tau^2\right)$$

Note that y_{ij^*} is conditional on the group-specific intercept α_i, which we can estimate. Out-of-sample predictions need to consider all possible values of α_i.

In-Sample Prediction

Let $\boldsymbol{\theta} = (\alpha_{i^*}, \alpha_0, \boldsymbol{\beta}, \sigma^2, \tau^2)$ denote all the model parameters.

$$
\begin{aligned}
[y_{ij^*} \,|\, \mathbf{y}] &= \int [y_{ij^*}, \alpha_i, \alpha_0, \boldsymbol{\beta}, \sigma^2, \tau^2 \,|\, \mathbf{y}] \, d\boldsymbol{\theta} \\
&= \int [y_{ij^*} \,|\, \alpha_i, \alpha_0, \boldsymbol{\beta}, \sigma^2, \tau^2, \mathbf{y}] \times [\alpha_i, \alpha_0, \boldsymbol{\beta}, \sigma^2, \tau^2 \,|\, \mathbf{y}] \, d\boldsymbol{\theta} \\
&= \int [y_{ij^*} \,|\, \alpha_i, \boldsymbol{\beta}, \sigma^2] \times [\alpha_i, \alpha_0, \boldsymbol{\beta}, \sigma^2, \tau^2 \,|\, \mathbf{y}] \, d\boldsymbol{\theta} \\
&= \int [y_{ij^*} \,|\, \alpha_i, \boldsymbol{\beta}, \sigma^2] \times [\alpha_i, \boldsymbol{\beta}, \sigma^2 \,|\, \mathbf{y}] \, d\boldsymbol{\theta}
\end{aligned}
$$

Numerically, to obtain posterior predictive distribution of y_{ij^*}, this means

1. Simulate $\alpha_i, \boldsymbol{\beta}, \sigma^2$ from $[\alpha_i, \boldsymbol{\beta}, \sigma^2 \,|\, \mathbf{y}]$
2. Simulate y_{ij^*} from $[y_{ij^*} \,|\, \alpha_i, \boldsymbol{\beta}, \sigma^2]$

Out-of-Sample Prediction

Let $\boldsymbol{\theta} = (\alpha_{i^*}, \alpha_0, \boldsymbol{\beta}, \sigma^2, \tau^2)$ denote all the model parameters.

$$
\begin{aligned}
[y_{i^*j^*} \,|\, \mathbf{y}] &= \int [y_{i^*j^*}, \alpha_{i^*}, \alpha_0, \boldsymbol{\beta}, \sigma^2, \tau^2 \,|\, \mathbf{y}] \, d\boldsymbol{\theta} \\
&= \int [y_{i^*j^*} \,|\, \alpha_{i^*}, \alpha_0, \boldsymbol{\beta}, \sigma^2, \tau^2, \mathbf{y}] \times [\alpha_{i^*}, \alpha_0, \boldsymbol{\beta}, \sigma^2, \tau^2 \,|\, \mathbf{y}] \, d\boldsymbol{\theta} \\
&= \int [y_{i^*j^*} \,|\, \alpha_{i^*}, \boldsymbol{\beta}, \sigma^2] \times [\alpha_{i^*}, \alpha_0, \boldsymbol{\beta}, \sigma^2, \tau^2 \,|\, \mathbf{y}] \, d\boldsymbol{\theta} \\
&= \int [y_{i^*j^*} \,|\, \alpha_{i^*}, \boldsymbol{\beta}, \sigma^2] \times [\alpha_{i^*} | \alpha_0, \boldsymbol{\beta}, \sigma^2, \tau^2, \mathbf{y}] \times [\alpha_0, \boldsymbol{\beta}, \sigma^2, \tau^2 \,|\, \mathbf{y}] \, d\boldsymbol{\theta} \\
&= \int [y_{i^*j^*} \,|\, \alpha_{i^*}, \boldsymbol{\beta}, \sigma^2] \times [\alpha_{i^*} | \alpha_0, \tau^2] \times [\alpha_0, \boldsymbol{\beta}, \sigma^2, \tau^2 \,|\, \mathbf{y}] \, d\boldsymbol{\theta}.
\end{aligned}
$$

Note that we only need the posterior distribution $[\alpha_0, \boldsymbol{\beta}, \sigma^2, \tau^2 \,|\, \mathbf{y}]$. We then average overall possible values of α_{i^*} based on $[\alpha_{i^*} | \alpha_0, \tau^2]$.

Posterior Inference: Prediction Example

Using the random intercept model for the toenail data...

- Predict the response for subject i when $treat_i = 1$ and $time_{ij} = 15$

```
for(k in 1:n.id){
    pred.mu1[k] = alpha[k] + beta1 + beta2*15 + beta3*15
    pred1[k] ~ dnorm(pred.mu1[k], prec1)
}
```

- Predict the response for a new subject when $treat_i = 1$ and $time_{ij} = 15$

```
pred.alpha ~ dnorm(alpha0, prec2)
pred.mu2 = pred.alpha + beta1 + beta2*15 + beta3*15
pred2 ~ dnorm(pred.mu2, prec1)
```

Question: Did these lines need to be included in the JAGS model?

Posterior Inference: Prediction Example

	Post. Mean	Post. SD	95% P.I.
Within-sample (Subject 1)	14.13	2.81	(8.57, 19.87)
Within-sample (Subject 2)	13.70	2.78	(8.01, 19.05)
Within-sample (Subject 3)	10.00	2.88	(4.36, 15.51)
Out-of-sample (typical subject)	11.91	3.59	(4.77, 18.80)

- Within-sample predictions have smaller posterior intervals because we have data to estimate the random effects
- Posterior standard deviation may vary across subjects because each random effect has different uncertainty depending on how much shrinkage there is

Random Intercept and Slope Model for the Toenail Data

Consider adding a random slope to the time variable

$$y_{ij} = \alpha_{0i} + \beta_1 treat_i + \alpha_{1i} time_{ij} + \beta_2 * treat_i \times time_{ij} + \epsilon_{ij}$$

$$\begin{bmatrix} \alpha_{0i} \\ \alpha_{1i} \end{bmatrix} \sim \text{Normal}\left(\boldsymbol{\alpha} = \begin{bmatrix} \alpha_0 \\ \alpha_1 \end{bmatrix}, \boldsymbol{\Sigma}_{2\times 2} \right) \qquad \epsilon_{ij} \sim \text{Normal}(0, \sigma^2)$$

Priors:

$$\begin{aligned} \boldsymbol{\beta} &\sim \text{Normal}(\, \mathbf{0}, v_\beta^2 \mathbf{I}_{p\times p}\,) \\ \boldsymbol{\alpha} &\sim \text{Normal}(\, \mathbf{0},\, v_\alpha^2 \mathbf{I}_{2\times 2}\,) \\ \sigma^2 &\sim \text{Inv-Gamma}(\, c_\sigma,\, d_\sigma\,) \\ \boldsymbol{\Sigma} &\sim \text{Bivariate Inv-Gamma}(\, r,\, \mathbf{R}\,) \end{aligned}$$

Prior for Random Effects Covariance Matrix

In JAGS, a multivariate normal distribution is specified by

$$\sim \text{dmnorm}\ (\boldsymbol{\mu}, \Omega)$$

where Ω is the precision matrix (inverse of the covariance matrix).

The Inverse-Wishart distribution:

- is a multivariate generalization of the inverse-Gamma distribution
- is a probability distribution for random covariance matrices! This is tricky because a covariance matrix $\boldsymbol{\Sigma}$ needs to be
  - symmetric
  - positive definite: $a'\Sigma a > 0$ for all vector $a \neq 0$
- if $\boldsymbol{\Sigma} \sim \text{Inv-Wishart}(r, r\mathbf{R})$, then

$$E[\,\boldsymbol{\Sigma}\,] = \frac{r\mathbf{R}}{r - p - 1}$$

where $\mathbf{R}$ is a $p \times p$ positive definite matrix and $r > p - 1$ is a real number.

The Inverse-Wishart Distribution

The Inverse-Wishart distribution is useful for making inference about a population covariance matrix. Specifically, let $\boldsymbol{\theta}_i = (\theta_{0i}, \theta_{1i})$, for i in $1, \ldots, J$. Let's assume

$$\boldsymbol{\theta}_i \sim N(0, \boldsymbol{\Sigma})$$

$$\boldsymbol{\Sigma} \sim \text{Inv-Wishart}(r, \mathbf{R}).$$

Then let $\mathbf{A}$ be the the sample covariance matrix. We can show that the posterior distribution of $[\boldsymbol{\Sigma}|\text{data}]$ is

$$[\boldsymbol{\Sigma}|\text{data}] \sim \text{Inv-Wishart}(r + J, r\mathbf{R} + J\mathbf{A}),$$

which has mean

$$E[\boldsymbol{\Sigma}|\text{data}] = \frac{r\mathbf{R} + J\mathbf{A}}{r + J}.$$

Note that for small r the above is approximately $\frac{JA}{J} = \mathbf{A}$.

Therefore, for prior $\boldsymbol{\Sigma} \sim \text{Inv-Wishart}(r, \mathbf{R})$, we can think of

- $\mathbf{R}$ as the prior assumption on the random effect covariance matrix, and
- r as the sample size that $\mathbf{R}$ is based on

The Inverse-Wishart Distribution: Examples

Some realizations from Inv-Wishart $\left(r = 2,\ r\mathbf{R} = 2 \times \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right)$.

```
> riwish (2, 2*diag(2))
         [,1]     [,2]
[1,] 7.993478 1.056120
[2,] 1.056120 1.184883
```

```
> riwish (2, 2*diag(2))
          [,1]     [,2]
[1,] 0.9417932 1.434751
[2,] 1.4347506 6.704089
```

```
> riwish (2, 2*diag(2))
           [,1]      [,2]
[1,] 153.62756 -68.26434
[2,] -68.26434  30.74045
```

Some realizations from Inv-Wishart $\left(r = 20,\ r\mathbf{R} = 20 \times \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right)$.

```
> riwish (20, 20*diag(2))
           [,1]       [,2]
[1,]  1.0477453 -0.4198504
[2,] -0.4198504  0.8217627
```

```
> riwish (20, 20*diag(2))
          [,1]      [,2]
[1,] 0.7991031 0.1561055
[2,] 0.1561055 1.1227496
```

```
> riwish (20, 20*diag(2))
            [,1]        [,2]
[1,]  1.09554075 -0.09640348
[2,] -0.09640348  0.94267139
```

Some realizations from Inv-Wishart $\left(r = 200,\ r\mathbf{R} = 200 \times \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right)$.

```
> riwish (200, 200*diag(2))
            [,1]        [,2]
[1,]  0.92784643 -0.04097586
[2,] -0.04097586  0.90093856
```

```
> riwish (200, 200*diag(2))
           [,1]       [,2]
[1,] 0.97532764 0.05100576
[2,] 0.05100576 1.06659492
```

```
> riwish (200, 200*diag(2))
            [,1]        [,2]
[1,]  0.99540455 -0.05506508
[2,] -0.05506508  1.10315592
```

Random Intercept and Slope Model in JAGS

```
# Random Intercept and Slope Model
toe.jags3 <- function(){

  # Data Likelihood
  for(i in 1:n.obs){
    response[i] ~ dnorm(mu[i], prec1)
    mu[i] = alpha[ id[i],1] + beta1*treat[i] + alpha[id[i],2]*time[i]+ beta2*time[i]*treat[i]
  }

  for (j in 1:n.id){ alpha[j,1:2] ~ dmnorm(alpha_mu, Omega) }

  # Priors
  alpha_mu[1] ~ dnorm(0, 1E-10)
  alpha_mu[2] ~ dnorm(0, 1E-10)
  beta1 ~ dnorm(0, 1E-10)
  beta2 ~ dnorm(0, 1E-10)
  prec1 ~ dgamma(0.0001, 0.0001)
  Omega[1:2, 1:2] ~ dwish(R, 2)
  sigma = 1 / prec1
  BigSigma = inverse(Omega)
}

# Get estimate for R
toe.lmer2 <- lmer(response ~ (1 + time|id) + treat + time + treat*time, data = toe)
summary(toe.lmer2)
R <- 2 * matrix(c(7.37, -0.39, -0.39, 0.23), ncol = 2)
```

Random Intercept and Slope Model Results

Parameter	Post Mean	95% Post Int
α_0	2.45	(1.99, 2.94)
α_1	0.59	(0.50, 0.69)
β_1	0.28	(-0.39, 0.94)
β_2	0.03	(-0.09, 0.15)
σ^2	3.16	(2.91, 3.40)
$\mathbf{\Sigma}_{11}$	7.45	(6.14 8.87)
$\mathbf{\Sigma}_{22}$	0.23	(0.19, 0.28)
$Cor(\alpha_{0i}, \alpha_{1i})$	-0.39	(-0.50, -0.27)

Ohio Lung Cancer Example

Recall the final model fit last week for the Ohio lung cancer example:

$$y_{stk} \sim \text{Poisson}(\lambda_{stk} \times pop_{stk})$$

$$\log(\lambda_{stk}) = \theta_s + \beta_0 + \beta_1 sex_k + \beta_2 race_k + \beta_3 sex_k \times race_k + \beta_4 year_t$$

$$\theta_s \sim \text{Normal}(0, \tau^2)$$

Ohio Lung Cancer Example: JAGS code

```
# JAGS Model
lung.jags <- function(){

  # Data Likelihood
  for(i in 1:n.obs){
    death[i] ~ dpois(lambda[i]*pop[i])
    log(lambda[i]) = beta0 + theta[cty[i]] + beta1*sex[i] + beta2*race[i]
                                          + beta3*sex[i]*race[i] + beta4*year[i]
  }

  for(s in 1:n.cty){theta[s] ~ dnorm(0, tau2.inv)}

  # Priors
  beta0 ~ dnorm(0, 1e-10)
  beta1 ~ dnorm(0, 1e-10)
  beta2 ~ dnorm(0, 1e-10)
  beta3 ~ dnorm(0, 1e-10)
  beta4 ~ dnorm(0, 1e-10)
  tau2.inv ~ dgamma(0.0001, 0.0001)
  tau2 = 1 / tau2.inv
}
```

Note: JAGS might not take factors well, so code your own indicator variables.

Ohio Lung Cancer Example: Parameter Estimates

Let's compare the frequentist parameter estimates obtained using *glmer()* with $nAGQ = 100$ to the MCMC results.

Covariate	GLMM	JAGS
Intercept β_0	-7.333 (0.024)	-7.331 (0.025)
Female β_1	-0.885 (0.010)	-0.885 (0.010)
Non-white β_2	-0.029 (0.017)	-0.029 (0.016)
Female $\times$ Non-white β_3	-0.219 (0.031)	-0.220 (0.031)
Year β_4	0.023 (0.002)	0.023 (0.002)
τ^2	0.198^2	0.201^2 (0.007)

The results are nearly identical.

Crossover Trial Example

Recall the final model fit last week for the crossover trial example:

$$y_{ij} \sim \text{Binomial}(p_{ij})$$

$$\text{logit}\ (p_{ij}) = \theta_i + \beta_0 + \beta_1 trt_{ij} + \beta_2 period_{ij} + \beta_3 trt_{ij} \times period_{ij}$$

$$\theta_i \sim \text{Normal}(0, \tau^2)$$

Crossover Trial Example: JAGS code

```
# JAGS Model
cross.jags <- function(){

  # Data Likelihood
  for(i in 1:n.obs){
    outcome[i] ~ dbern(p[i])
    logit(p[i]) = beta0 + theta[ID[i]] + beta1*trt[i] + beta2*period[i]
                                                     + beta3*trt[i]*period[i]
  }

  for(s in 1:n.ID){theta[s] ~ dnorm(0, tau2.inv)}

  # Priors
  beta0 ~ dnorm(0, 1e-10)
  beta1 ~ dnorm(0, 1e-10)
  beta2 ~ dnorm(0, 1e-10)
  beta3 ~ dnorm(0, 1e-10)
  tau2.inv ~ dgamma(0.0001, 0.0001)
  tau2 = 1 / tau2.inv
}
```

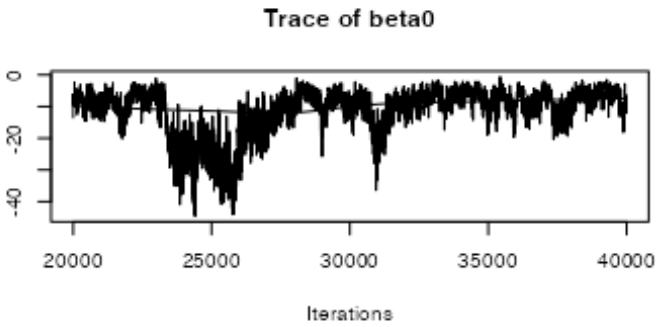
Crossover Trial Example: Parameter Estimates

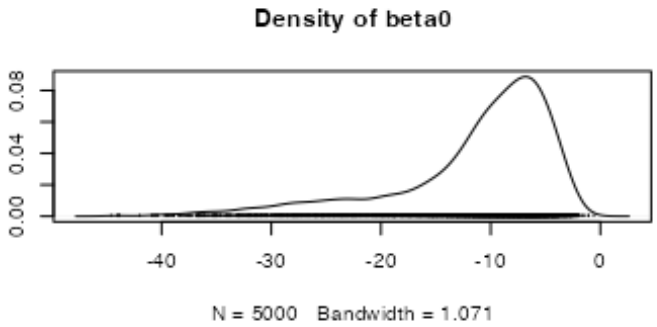
Let's compare the frequentist parameter estimates obtained using *glmer()* with $nAGQ = 100$ to the MCMC results.

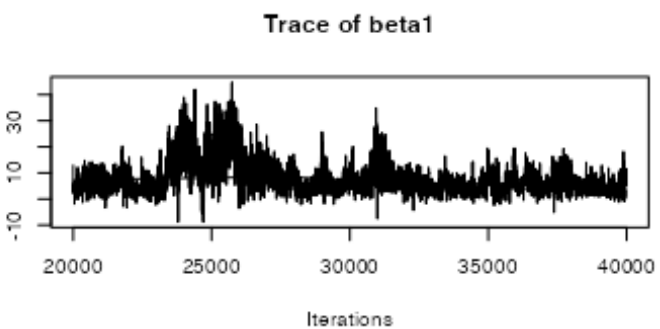
Covariate	GLMM	JAGS
Intercept β_0	-4.98 (2.12)	-11.45 (7.42)
Treatment β_1	3.58 (2.11)	8.17 (6.35)
Period β_2	2.77 (2.02)	6.89 (6.16)
Treatment $\times$ Period β_3	-3.32 (3.27)	-7.73 (9.23)
τ^2	4.91^2	13.55^2 (232.88)

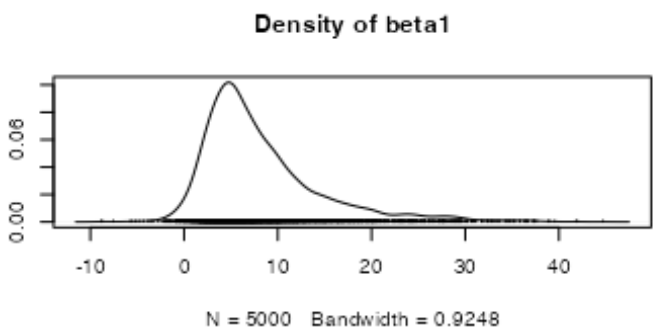
Not so great... Why?

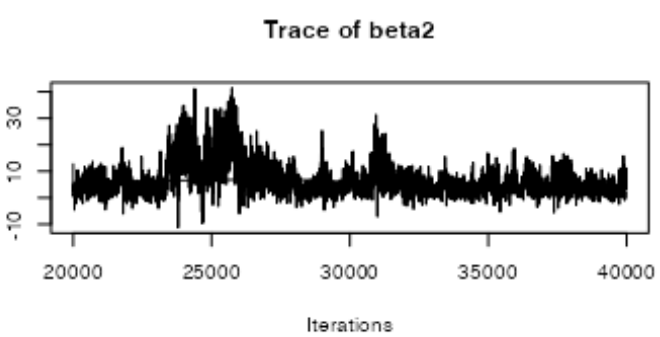
Crossover Trial Example: Trace Plots

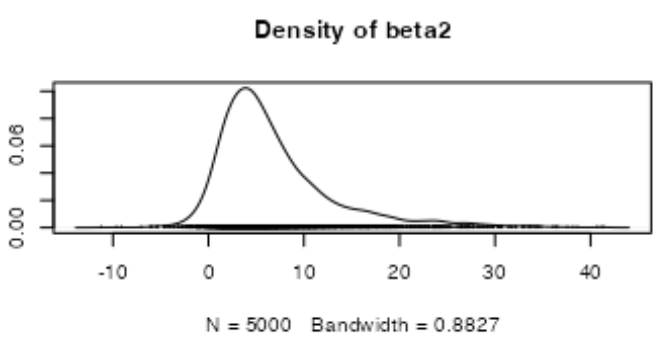




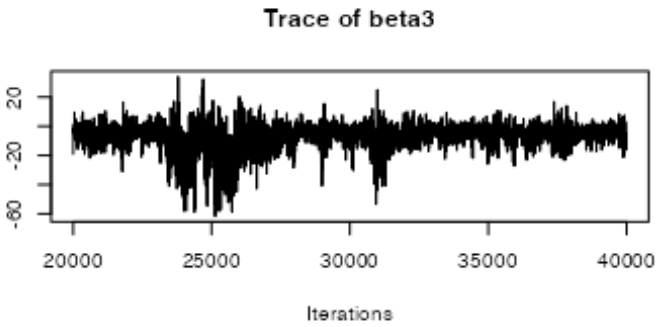


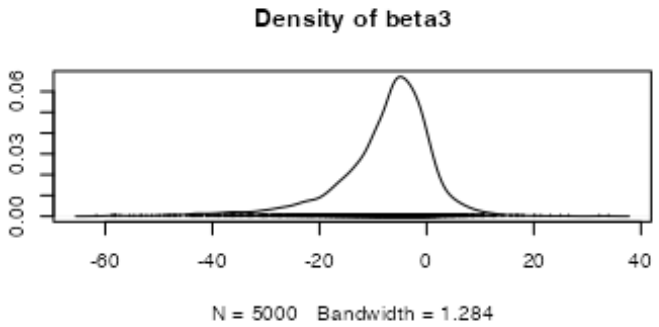


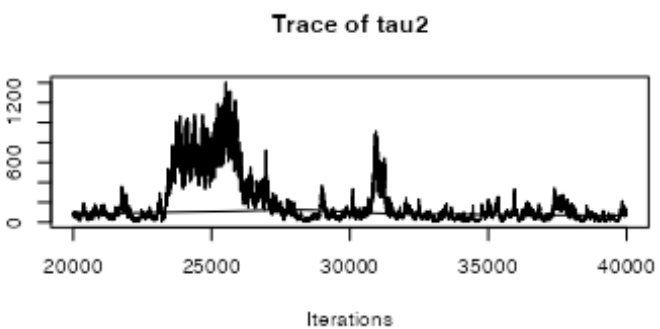


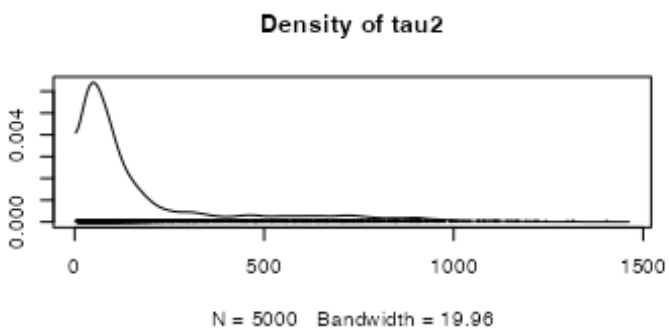


Crossover Trial Example: Trace Plots









Bayesian Hierarchical Model Summary

For linear mixed models, results from Frequentist and Bayesian analysis are typically similar. Consider a Bayesian analysis when you:

- Have only a small sample size
- Want to make inference on random effect estimates and the heterogeneity parameters
- Want to quantify and reflect all sources of uncertainties
- Need to relax model assumptions
- Want to incorporate *a priori* information through prior distributions

Some limitations for Bayesian inference

- A lot more computationally intensive
- Sometimes priors need to be carefully chosen, especially for complex models
- Important to evaluate sensitivity of model results to prior specification