# Introduction to Bayesian methods and implementation via R/JAGS

Ben Saville, PhD

Berry Consultants, Vanderbilt University

September 13, 2018

# Outline

# Example

- Consider a study evaluating the effect of a treatment drug on the severity of depression
- Depression severity is measured with the Hamilton Rating Scale for Depression (HAMD-17)
- Suppose we fit a linear regression model, where $\beta$ is the parameter corresponding to the treatment
- Suppose a linear regression model gives $\hat{\beta} = -0.4$ with 95% CI of (-.9, 0.1)
- $H_0 : \beta = 0$ yields a p-value of 0.10

# Interpretation



- Confidence interval:
  - We are 95% confident the true value of $\beta$ is between -0.9 and 0.4
  - If we replicate this study 100 times and compute a confidence interval for each data set, 95 out of 100 confidence intervals will capture the unknown parameter $\beta$.
- P-value:
  - The probability of observing a result as extreme or more extreme than our data given the null hypothesis is true is 0.10
  - We fail to reject $H_0$, meaning we fail to find evidence to contradict the null hypothesis

# What we would prefer



- The probability that the treatment is effective
- The interval that contains the true parameter $\beta$ with probability 0.95

# Another example



- Suppose there are two therapies, A and B
- The true rate of success using therapy A is $\theta_A$ and using therapy B is $\theta_B$
- The observed p-value is 0.10
- A classical conclusion: "If $\theta_A = \theta_B$ then there is a 10% probability of observing a result as or more extreme than the one observed in this trial."
- What we would prefer: "Given the results of this trial, the probability that $\theta_A$ is greater than $\theta_B$ is 85%."

# Two general branches of statistics



- Frequentist (traditional) methods
  - All the information is contained in the data
  - Base inference on point estimates (CIs) and p-values
- Bayesian methods
  - Assign a prior probability distribution $\pi(\beta)$ to $\beta$, reflecting uncertainty in $\beta$
  - Inference is based on the posterior distribution of $\beta$ given the observed data $y$

# How do Bayesian methods work?



- A Bayesian writes down a prior guess for $\beta$, $\pi(\beta)$, then combines this with information from the data $y$ to obtain the posterior distribution of $\beta$, written as $p(\beta|y)$
- posterior information $\geq$ prior information $\geq 0$

# Example: Cancer rates



**County-level breast cancer rates per 10,000 women**

| | | | | |
|---|---|---|---|---|
| 79 | 87 | 83 | 80 | 78 |
| 90 | 89 | 92 | 99 | 95 |
| 96 | 100 | **x** | 110 | 115 |
| 101 | 109 | 105 | 108 | 112 |
| 96 | 104 | 92 | 101 | 96 |

- With no direct data for **x**, what estimate would you use?
- Is 200 reasonable?

# Example: Cancer rates



- Now suppose data becomes available for county **x**, with 100 women at risk and 2 cancer cases

$$\text{rate} = \frac{2}{100} \times 10,000 = 200$$

- Would you use 200 as a reliable estimate?
- What about a compromise between 200 and the neighboring counties?
- How would your estimate change if you observed data 20/1000, 200/10000, etc.?

# History



- Bayesian conceptual framework was developed by the Reverend Thomas Bayes (1702-1761), and published posthumously in 1764
- Frequentist philosophy formalized in early 20th century (Neyman, Pearson, Fisher et al.) and quickly became dominant
- Revival of Bayesian statistics in late 20th century due largely to computational advances

# Bayesian Estimation



- Bayesian estimation is based on the posterior distribution of $\boldsymbol{\theta}$ given the observed data $\mathbf{y}$

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\boldsymbol{\theta}, \mathbf{y})}{p(\mathbf{y})} = \frac{p(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{p(\mathbf{y})} = \frac{p(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int p(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}$$

  - $p(\mathbf{y}) = \int p(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$ is the marginal likelihood
- Instead of a point estimate and CI, Bayesians can plot the entire (posterior) distribution of the parameters

# Hypothesis Testing: The Bayes factor



- Using Bayes theorem, we can write the posterior odds of one hypothesis versus another as

$$\frac{p(H_1|\mathbf{y})}{p(H_0|\mathbf{y})} = \frac{p(\mathbf{y}|H_1)p(H_1)}{p(\mathbf{y}|H_0)p(H_0)}$$

- The Bayes factor is

$$B_{10} = \frac{p(\mathbf{y}|H_1)}{p(\mathbf{y}|H_0)}$$

  or the posterior odds of one hypothesis versus another when the prior probabilities of the two hypotheses are equal

# The Bayes factor



**Table 1. Grades of Evidence of Bayes Factors**

| Bayes factor | Interpretation |
|---|---|
| $B_{10} < 1/10$ | Strong evidence for $H_0$ |
| $1/10 < B_{10} < 1/3$ | Moderate evidence for $H_0$ |
| $1/3 < B_{10} < 1$ | Weak evidence for $H_0$ |
| $1 < B_{10} < 3$ | Weak evidence for $H_1$ |
| $3 < B_{10} < 10$ | Moderate evidence for $H_1$ |
| $B_{10} > 10$ | Strong evidence for $H_1$ |

# Objective Bayesian analysis



- In many situations there is no prior information, or the researcher does not want to assume any prior information
- Objective Bayesian analysis incorporates non-informative prior distributions
- Well-spread priors; the data dominate the posterior distribution
- Works well for inference; problematic for Bayes factors

# Revisit the Depression Study



- We want to evaluate the effect of a treatment (lamotrigine) on the severity of depression

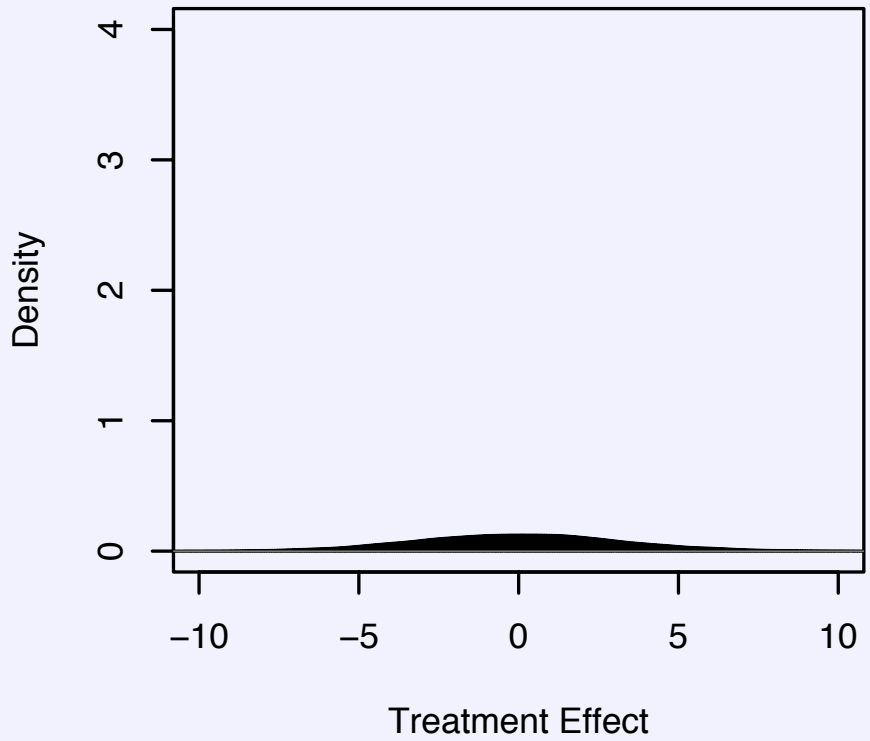



Figure: Prior distribution for treatment effect

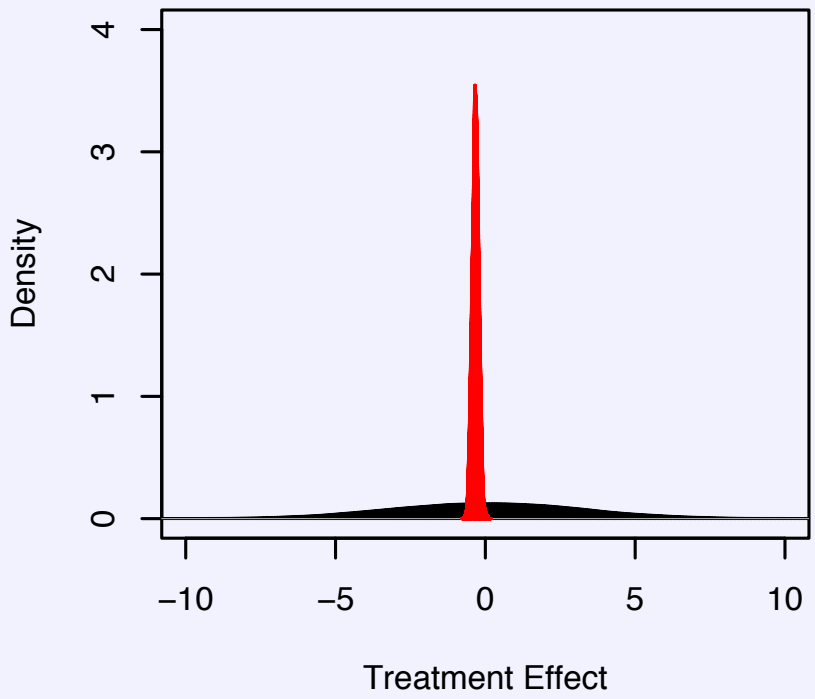



Figure: Prior & Posterior distribution for treatment effect

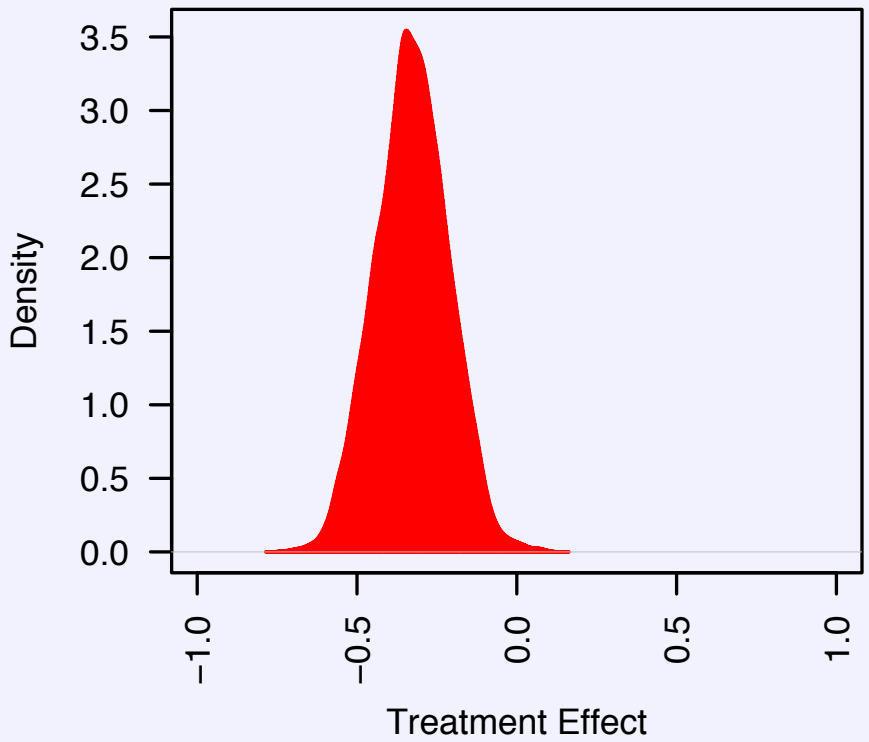



Figure: Posterior distribution for treatment effect

# Gibbs Sampling



- For many models the posterior distributions are not available in closed form (cannot directly solve using calculus)
- Especially true for complex models with many parameters
- Gibbs sampling (MCMC, or Markov Chain Monte Carlo)
  - We draw a random sample of one parameter at a time assuming that the current values of all the other parameters are correct
  - Then we go on to the next parameter and cycle through all the parameters many times
  - This generates samples from the posterior distribution of all the parameters

# Gibbs Sampling



- For example, consider a simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \ \varepsilon_i \sim N(0, \sigma^2)$$

- Two general options for priors:
  - Conjugate model:

$$\beta_0 \sim N(0, \sigma^2 1000)$$
$$\beta_1 \sim N(0, \sigma^2 1000)$$
$$\tau = \frac{1}{\sigma^2} \sim Gam(0.01, 0.01)$$

  - Non Conjugate model:

$$\beta_0 \sim N(0, 1000)$$
$$\beta_1 \sim N(0, 1000)$$
$$\tau = \frac{1}{\sigma^2} \sim Gam(0.01, 0.01)$$

# Gibbs Sampling



- We need the joint posterior distribution $p(\beta_0, \beta_1, \sigma^2|\mathbf{Y})$
- We iteratively sample from $p(\beta_0|\sigma^2, \beta_1, \mathbf{Y})$, $p(\beta_1|\sigma^2, \beta_0, \mathbf{Y})$, $p(\sigma^2|\beta_0, \beta_1, \mathbf{Y})$ until they converge to the joint distribution

# Methods for fitting models



- Very limited applications/functions in SAS or R
- Can use calculus and program using software (e.g., R)
- OpenBUGS (Bayesian Inference Using Gibbs Sampling)
  - WinBUGS introduced in 1997 and is no longer being developed
  - OpenBUGS (open-source) started in 2004 and is an active project: Windows, Linux
  - Can also run OpenBUGS from R (BRugs), but doesn't (?) support Linux
- JAGS (Just Another Gibbs Sampler)
  - Similar syntax to OpenBugs, runs on Linux and Mac
  - Perhaps more potential than OpenBugs
  - Easy interface with R package runjags
- Others: STAN, NIMBLE

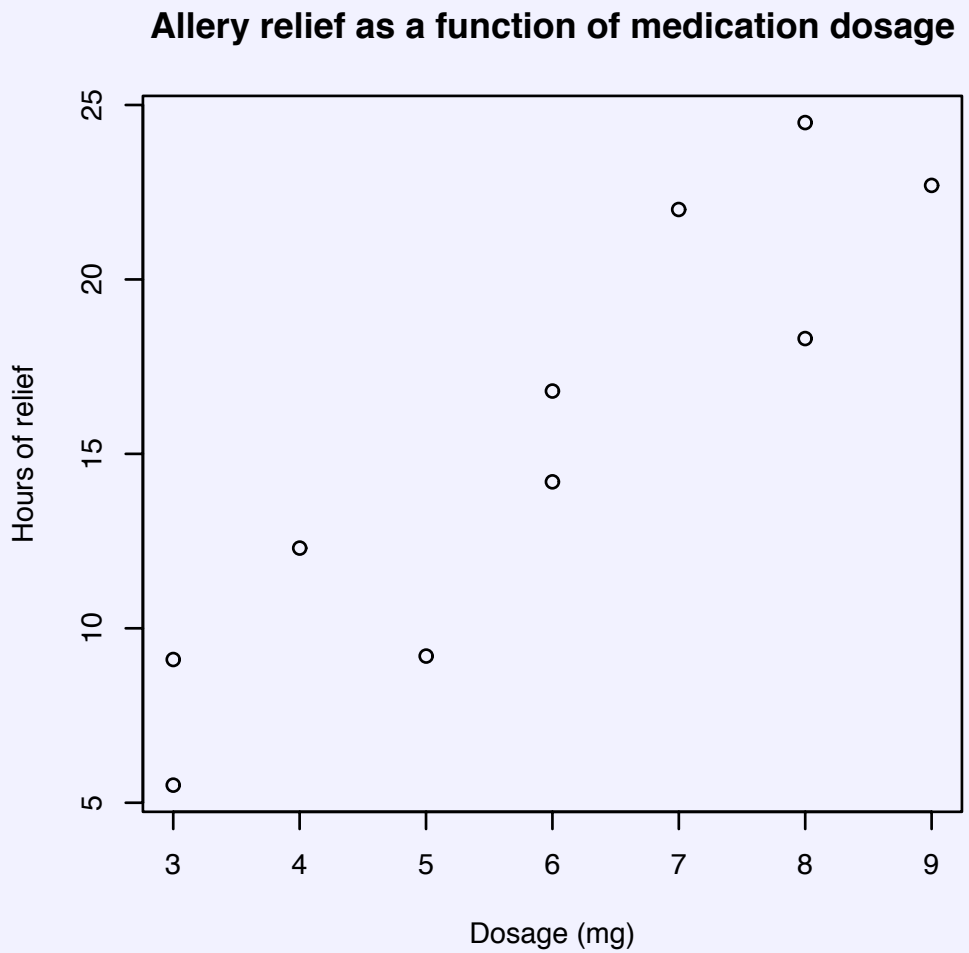
# Example: Allergy Medication



| Dose of medication (mg) | Hours of Relief |
|---|---|
| 3 | 9.1 |
| 3 | 5.5 |
| 4 | 12.3 |
| 5 | 9.2 |
| 6 | 14.2 |
| 6 | 16.8 |
| 7 | 22.0 |
| 8 | 18.3 |
| 8 | 24.5 |
| 9 | 22.7 |

# Example: Allergy Medication





# Three main inputs to JAGS



1. The model
2. The data
3. Initial values for MCMC sampling

# Variance vs. precision



- JAGS parameterizes normal distributions as means and precisions, not variances
- Precision is the reciprocal of a variance

# Model



```
model{
  for(i in 1:N)  {
    y[i] ~ dnorm(mu[i], tau);
    mu[i] <- beta0 + beta1*x[i];
    }
  beta0 ~ dnorm(0, 0.001);
  beta1 ~ dnorm(0, 0.001);
  tau ~ dgamma(0.01, 0.01);
}
```

If tau $\sim$ Gamma, then $\sigma^2$ $\sim$InvGamma
dgamma(v,w): mean $=$ v/w, var $=$ v/w$^2$

# Data



- The best way to supply a data file to JAGS is in "R" list format

```
list(
  y = c(9.1, 5.5, 12.3, 9.2, 14.2, 16.8, 22, 18.3,
        24.5, 22.7),
  x = c(3, 3, 4, 5, 6, 6, 7, 8, 8, 9),
  N = 10
)
```

# Initial Values



- In this example, there are three parameters involved in MCMC: beta0, beta1, and tau
- Two options for generating initial values
  1. Let JAGS generate initial values for your (from the prior distribution). This can be unreliable
  2. Supply your own initial values

# Initial Values



Each single line below are in separate files

```
list( beta0 = 3.332, beta1 = -1.544, tau = 0.155 )
list( beta0 = 0.087, beta1 = -3.914, tau = 1.976 )
list( beta0 = 0.568, beta1 = 2.440, tau = 0.841 )
```

# Steps to fitting a model in JAGS



- Load the model, data, and initial values. More than one chain can be specified.
- Specify a burn-in phase to reach convergence
- Specify which parameters you would like to store MCMC values
- Update the MCMC sampling with more samples, this time storing values for monitored parameters
- Look at trace plots, posterior distributions, quantiles, etc.
- Check MCMC diagnostics
- Also straightforward to obtain predictive posterior distributions