Bayesian Approach

Objectives

- Bayes' Theorem: prior, likelihood & posterior
- Bayesian Approach to Linear Models
- Something for the future

Bayes' Theorem

Ex 1. Suppose we have 100 (DVD) movies and 50 books. Suppose also there are 3 different movie types: Action, Sci-fi, Romance; and 2 different book types: Sci-fi, Romance. Furthermore,

Of the 100 movies	Of the 50 books
28 = Action	
31 = Sci-fi	17 = Sci-fi
41 = Romance	33 = Romance

So, when you were given an unspefied object:

- The probability that it's a movie is $\frac{100}{150}$, and $\frac{50}{150}$ for book.
- The probability that it's an Action type is $\frac{28}{150}$, $\frac{48}{150}$ for Sci-fi, and $\frac{74}{150}$ for Romance.
- If we already know it's a movie, then the probability that it's an Action is $\frac{28}{100}$, $\frac{31}{100}$ for Sci-fi, and $\frac{41}{100}$ for Romance.
- If we already know it's a book, then that probability that it's a Sci-fi is $\frac{17}{50}$, $\frac{33}{50}$ for Romance.

Right now, we want to know that given an object which is Sci-fi, the probability that it's a movie.

Answer:

$$P(\text{movie}) = \frac{100}{150} = \frac{2}{3}, \quad P(\text{book}) = \frac{50}{150} = \frac{1}{3}$$
$$P(A) = \frac{28}{150} = 0.1867, \quad P(S) = \frac{48}{150} = 0.32, \quad P(R) = \frac{74}{150} = 0.4933$$
$$P(S \,|\, \text{movie}) = \frac{31}{100}, \quad P(S \,|\, \text{book}) = \frac{17}{50}$$

Now, the answer:

$$P(\text{movie} \,|\, S) = \frac{P(S \,|\, \text{movie}) \cdot P(\text{movie})}{P(S)} = \frac{P(S \,|\, \text{movie}) \cdot P(\text{movie})}{P(S \,|\, \text{movie}) \cdot P(\text{movie}) + P(S \,|\, \text{book}) \cdot P(\text{book})}$$
$$= \frac{\frac{31}{100} \cdot \frac{2}{3}}{\frac{31}{100} \cdot \frac{2}{3} + \frac{17}{50} \cdot \frac{1}{3}} = \frac{62/300}{(62/300) + (34/300)} = \frac{62}{96} = \frac{31}{48} = 0.6458$$

In the above example:

- $P(\text{movie})$ is called **prior**.
- $P(S \mid \text{movie})$ is called **likelihood**.
- $P(\text{movie} \mid S)$ is called **posterior**.

Theorem 1. ***Bayes' Theorem***
Let $S = B_1 \cup B_2 \cup \cdots \cup B_m$ and $B_i \cap B_j = \emptyset, i \neq j$, (i.e., B_i's are mutually exclusive and exhaustive partition of the sample space S), then

$$P(B_k|A) = \frac{P(B_k)P(A|B_k)}{\sum_{i=1}^{m} P(B_i)P(A|B_i)}, \quad k = 1, 2, \ldots, m.$$

Proof.

□

Consider the general problem of inferring a distribution for a parameter θ given some data x. From Bayes' theorem, the posterior distribution $p(\theta \mid x)$ is equal to the product of the prior $p(\theta)$ and the likelihood function $p(x \mid \theta)$, i.e.,

$$p(\theta \mid x) = \frac{p(x \mid \theta) \cdot p(\theta)}{\int p(x \mid \theta') \cdot p(\theta') \; d\theta'} \quad \propto \quad p(\theta) \cdot p(x \mid \theta)$$

Bayesian thoughts are summarized by this statement:

$$\text{posterior (probability)} \propto \text{prior (probability)} \times \text{likelihood}$$

Two more classic examples are shown below.

Ex 2. Machines I, II, and III are all producing springs of the same length. Of their production, machines I, II, and III produce 2%, 1%, and 3% defective springs, respectively. Of the total production of springs in the factory, machine I produces 35%, machine II produces 25%, and machine III produces 40%. (a) If one spring is selected at random from the total produced in a day, what is the probability that it's defective? (b) If the selected spring is defective, what is the conditional probability that it was produced by machine III?

Answer:

$$\begin{aligned} P(D) &= P(I)P(D|I) + P(II)P(D|II) + P(III)P(D|III) \\ &= (0.35)(0.02) + (0.25)(0.01) + (0.4)(0.03) = 0.0215 \end{aligned}$$

$$P(III|D) = \frac{P(III)P(D|III)}{P(D)} = \frac{(0.4)(0.03)}{0.0215} = 0.5581$$

Ex 3. Let T^+ = test positive, T^- = test negative; C^+ = has cancer, and C^- = does not have cancer. A pap smear test is used to detect cervical cancer and is known to show about 16% *false negatives*, and about 19% *false positives*, i.e., $P(T^-|C^+) = 0.16$ and $P(T^+|C^-) = 0.19$. In the US, there are about 8 women in 100,000 with this cancer. Find the probability that a randomly selected woman who was just tested positive actually has the cancer.
Answer:

$$\begin{aligned} P(C^+|T^+) &= \frac{P(C^+ \cap T^+)}{P(T^+)} \\ &= \frac{P(T^+|C^+)P(C^+)}{P(T^+|C^+)P(C^+) + P(T^+|C^-)P(C^-)} \\ &= \frac{(0.84)(0.00008)}{(0.84)(0.00008) + (0.19)(0.99992)} \\ &= 0.000354 \end{aligned}$$

Bayesian statistics involve the following steps:

1. Define the prior distribution that incorporates your subjective beliefs about a parameter. The prior can be uninformative or informative.
2. Gather data.
3. Update your prior distribution with the data using Bayes' theorem to obtain a posterior distribution. The posterior distribution is a probability distribution that represents your updated beliefs about the parameter after having seen the data.
4. Analyze the posterior distribution and summarize it (mean, median, sd, quantiles, ...).

Furthermore,

- If the prior is uninformative, the posterior is very much determined by the data (the posterior is data-driven).
- If the prior is informative, the posterior is a mixture of the prior and the data.
- The more informative the prior, the more data you need to "change" your beliefs, so to speak because the posterior is very much driven by the prior information.
- If you have a lot of data, the data will dominate the posterior distribution (they will overwhelm the prior).

Ex 4. {β prior − binomial likelihood − β posterior}: In a group of students, there are 2 out of 18 that are left-handed. Find the posterior distribution of left-handed students in the population assuming uninformative prior.

Answer: First, some distribution properties:

- Suppose $X \sim \text{Beta}\,(\alpha, \beta)$, then, pdf $f(x) = \dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \quad 0 < x < 1.$
- $X \sim \text{Beta}\,(\alpha, \beta)$, then $\mu = \dfrac{\alpha}{\alpha+\beta}, \quad \sigma^2 = \dfrac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}$
- *Uniform*$(0, 1)$ is a special case of Beta (α, β) with $\alpha = 1 \,\&\, \beta = 1$.

Here, the likelihood is binomial (left-handed vs right-handed). Also, since we don't know about the distribution of the true rate of "left-handed" people, let's assume a beta distribution. The reason for this is: when the prior has a beta distribution and the likelihood has a binomial distribution, then the posterior also has a beta distribution (as sown below). We say that the beta distribution is the conjugate family for the binomial likelihood. Analysis like this example is convenient but rarely occur in real-world problems. In most cases, the posterior distribution has to be found numerically via MCMC (using WinBUGS, OpenBUGS, STAN, JAGS or some other programs), for example.

$$\textbf{posterior} \propto \textbf{(prior)} \times \textbf{(likelihood)}$$

Detailed proof is shown below.

$$\begin{aligned}
p(\theta \,|\, x) &= \frac{p(x \,|\, \theta) \cdot p(\theta)}{\int p(x \,|\, \theta') \cdot p(\theta') \; d\theta'} \\
&= \frac{\binom{n}{x}\theta^x(1-\theta)^{n-x} \cdot \left\{ \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} \right\}}{\int_{y=0}^{1} \left\{ \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1} \right\} \cdot \left\{ \binom{n}{x} y^x(1-y)^{n-x} \right\} dy} \\
&= \frac{\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \binom{n}{x} \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}}{\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \binom{n}{x} \int_{y=0}^{1} y^{x+\alpha-1}(1-y)^{n-x+\beta-1} dy} \\
&= \frac{\theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}}{\int_{y=0}^{1} y^{x+\alpha-1}(1-y)^{n-x+\beta-1} dy}, \\
&= \frac{\theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}}{\text{Beta}\,(x+\alpha, n-x+\beta)}, \\
&= \frac{\Gamma(n+\alpha+\beta)}{\Gamma(x+\alpha)\Gamma(n-x+\beta)} \theta^{\alpha+x-1}(1-\theta)^{n-x+\beta-1} \\
&\sim \text{Beta}\,(x+\alpha, n-x+\beta)
\end{aligned}$$

That is, the posterior has mean $\dfrac{x+\alpha}{n+\alpha+\beta}$ and variance $\dfrac{(x+\alpha)(n-x+\beta)}{(n+\alpha+\beta+1)(n+\alpha+\beta)^2}$. Here, we have $x = 2$, $n = 18$, $\alpha = \beta = 1$, the "posterior" distribution is Beta $(3, 17)$ and the "posterior" mean

$\hat{\theta} = \dfrac{2+1}{18+1+1} = \dfrac{3}{20} = 0.15$ and the variance $\dfrac{(2+1)(18-2+1)}{(18+1+1+1)(18+1+1)^2} = 0.00607143$, i.e., SD $= 0.07792$.

For certain choices of the prior, the posterior has the same algebraic form as the prior (generally with different parameter values). Such a choice is called a *conjugate prior*.

A conjugate prior is an algebraic convenience, giving a closed-form expression for the posterior; otherwise numerical integration may be necessary. Further, conjugate priors may give intuition, by more transparently showing how a likelihood function updates a prior distribution.

All members of the exponential family have conjugate priors. For a complete list of conjugate priors, search for "conjugate prior" at Wikipedia.

```
> x <- seq(0,1,length=200)
> lbinom <- function(p,x,n) dbinom(x,n,p)
>
> par(mfrow=c(1,3))
>
> plot(x,dbeta(x,1,1),type="l",xlab="p",col="blue",ylab="",
+  main=c("Prior is beta(1,1)"))
>
> plot(x,unlist(lapply(x,lbinom,x=2,n=18)),type="l",xlab="p",col="gray",ylab="",
+  main=c("Likelihood is binomial (x=2, n=18)"))
>
> postbetbin <- function(p, Y, N, alpha, beta) {
+        return(dbinom(sum(Y),N,p)*dbeta(p,alpha,beta))
+    }
>
> plot(x,unlist(lapply(x,postbetbin,2,18,alpha=1,beta=1)),type="l",xlab="p",col="red",ylab="",
+  main=c("Posterior is beta(3,17)"))
```

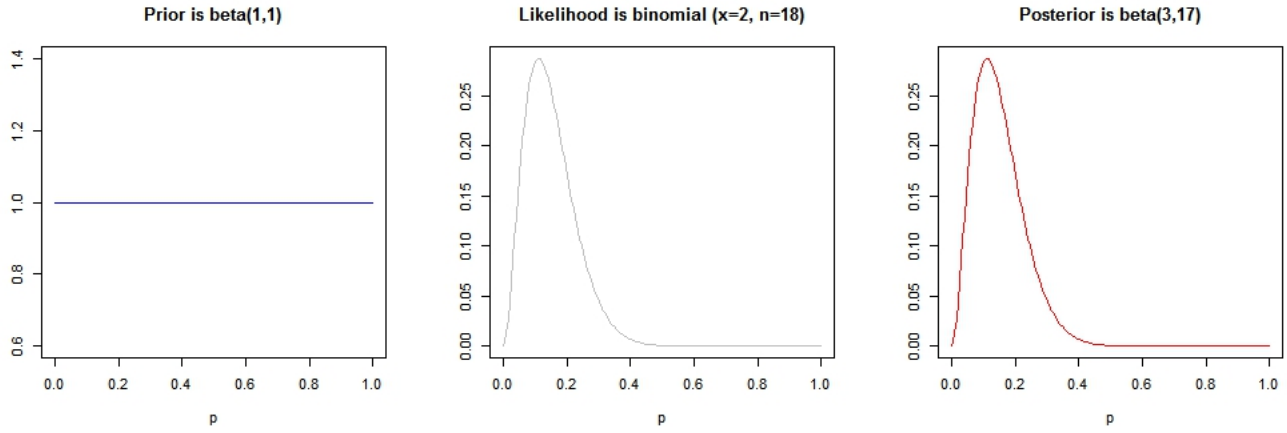

Figure 1: Beta prior, Binomial likelihood, and Beta posterior

Ex 5. {Γ prior – Poisson likelihood – Γ posterior}

- Suppose $X \sim \Gamma(\alpha, \beta)$, then, pdf $f(x) = \dfrac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} e^{-\beta/x}, \quad 0 < x < \infty.$
- $X \sim \Gamma(\alpha, \beta)$, then $\mu = \alpha\beta, \quad \sigma^2 = \alpha\beta^2$
- *Poisson*(θ) has a pdf $f(x) = \dfrac{e^{-\theta}\theta^x}{x!}, \quad x = 0, 1, 2, \ldots.$

Proof:

$$\begin{aligned} p(\theta \mid x) &\propto p(\theta) \cdot p(x \mid \theta) \\ &\propto \theta^{\alpha-1} e^{-\theta/\beta} e^{-n\theta} \theta^{\Sigma x_i} \\ &= \theta^{\Sigma x_i + \alpha - 1} e^{-\left(n + \beta^{-1}\right)\theta} \\ &\sim \Gamma\left\{\Sigma x_i + \alpha, \left(n + \beta^{-1}\right)^{-1}\right\} \end{aligned}$$

Ex 6. {Γ prior – Exponential likelihood – Γ posterior}

- *Exponential*$(1/\theta)$ has a pdf $f(x) = \theta e^{-\theta x}, \quad 0 < x < \infty.$

Proof:

$$\begin{aligned} p(\theta \mid x) &\propto p(\theta) \cdot p(x \mid \theta) \\ &\propto \theta^{\alpha-1} e^{-\theta/\beta} e^{-\theta \Sigma x_i} \theta^{n} \\ &= \theta^{n+\alpha-1} e^{-\left(\Sigma x_i + \beta^{-1}\right)\theta} \\ &\sim \Gamma\left\{n + \alpha, \left(\Sigma x_i + \beta^{-1}\right)^{-1}\right\} \end{aligned}$$

Ex 7. Normal prior – Normal likelihood – Normal posterior

- Suppose the prior of $\theta \sim N\left(\mu_0, \sigma_0^2\right)$.
- Suppose $X \sim N\left(\theta, \sigma^2\right)$, where $\sigma^2 =$ known.
- Then the posterior distribution of $\theta \sim N\left\{\mu(x), \sigma^2(x)\right\}$, where

$$\mu(x) = \frac{\frac{n\bar{x}}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}}{\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}}, \quad \sigma^2(x) = \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}}$$

Bayesian Approach to Liner Models

Making use of Bayesian ideas comes down to intelligently choosing the priors. Without too many ideas, we often begin with non-informative (i.e., data-dominated) priors, which allows us to avoid

the need to do tedious searching of previous evidence or expert elicitation.

Ex 1. Constructing posterior densities of β's

In case of the simple linear regression model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, where $\varepsilon_i \sim \text{iid}\, N(0, \sigma^2)$. For now, let's suppose σ is known. For a Bayesian analysis we first need a prior distribution for the two parameters, β_0 and β_1, and then then compute the posterior distribution. For now we use the following prior and later you will see why we chose this prior and examine its consequences:

$$\begin{aligned}
&\beta_0 \sim N\left(\mu_0, \sigma_0^2\right), \\
&\beta_1 \sim N\left(\mu_1, \sigma_1^2\right), \text{ and} \\
&\beta_0 \text{ and } \beta_1 \text{ are independent.}
\end{aligned}$$

This is written as:

$$\begin{pmatrix}\beta_0 \\ \beta_1\end{pmatrix} \sim N\left\{\begin{pmatrix}\mu_0 \\ \mu_1\end{pmatrix}, \begin{pmatrix}\sigma_0^2 & 0 \\ 0 & \sigma_1^2\end{pmatrix}\right\}$$

for a certain choice of $(\mu_0, \mu_1, \sigma_0^2, \sigma_1^2)$. The likelihood function is

$$\begin{aligned}
L\left(\beta_0, \beta_1\right) &= \prod_{i=1}^{n} f\left(y_i \,|\, \beta_0, \beta_1\right) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left\{\frac{y_i - (\beta_0 + \beta_1 x_i)}{\sigma}\right\}^2}
\end{aligned}$$

To find the posterior density we will use matrix notation: $\boldsymbol{\beta} = \begin{pmatrix}\beta_0 \\ \beta_1\end{pmatrix}$, $\boldsymbol{\mu} = \begin{pmatrix}\mu_0 \\ \mu_1\end{pmatrix}$, $\boldsymbol{\Sigma} = \begin{pmatrix}\sigma_0^2 & 0 \\ 0 & \sigma_1^2\end{pmatrix}$, $\boldsymbol{Y} = (y_1, y_2, \ldots, y_n)'$ and $\boldsymbol{X} = \begin{pmatrix}1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n\end{pmatrix}$.

Conditional on $\boldsymbol{\beta}$, $\boldsymbol{Y}$ is an n-dimensional normal distribution with mean $\boldsymbol{X\beta}$ and covariance matrix $\sigma^2 I_n$. Since the posterior density is proportional to the prior times the likelihood,

$$p(\boldsymbol{\beta} \,|\, \boldsymbol{Y}) \propto e^{-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu}) - \frac{1}{2}(\boldsymbol{Y}-\boldsymbol{X\beta})'(\boldsymbol{Y}-\boldsymbol{X\beta})/\sigma^2}$$

Note that the exponent is a quadratic form in $\boldsymbol{\beta}$. Therefore, the posterior density will be a two-dimensional normal distribution for $\boldsymbol{\beta}$ and we just need to complete the square to find the mean

vector and covariance matrix. The exponent can be simplified to

$$(\boldsymbol{\beta} - \boldsymbol{\mu^*})'\Sigma^{*-1}(\boldsymbol{\beta} - \boldsymbol{\mu^*}) + \cdots,$$

where $\Sigma^* = \left(\Sigma^{-1} + \boldsymbol{X'X}/\sigma^2\right)^{-1}$ and $\mu^* = \Sigma^*\left(\Sigma^{-1}\boldsymbol{\mu} + \boldsymbol{X'Y}/\sigma^2\right)$.

That is, the posterior distribution of $\boldsymbol{\beta}$ given $\boldsymbol{Y}$ is normal with mean $\boldsymbol{\mu^*}$ and covariance matrix $\boldsymbol{\Sigma^*}$. It is worth noting (1) that the posterior precision matrix $\boldsymbol{(\Sigma^*)^{-1}}$ is the sum of the prior precision matrix $\boldsymbol{\Sigma^{-1}}$ and a part that comes from the data, $\boldsymbol{X'X}/\sigma^2$ and (2) that the posterior mean is $\mu^* = \Sigma^*\left(\Sigma^{-1}\boldsymbol{\mu} + \boldsymbol{X'Y}/\sigma^2\right) = \Sigma^*\left\{\Sigma^{-1}\boldsymbol{\mu} + \left(\boldsymbol{X'X}/\sigma^2\right)\left(\boldsymbol{X'X}\right)^{-1}\boldsymbol{X'Y}\right\}$, a weighted average of the prior mean $\boldsymbol{\mu}$ and the least-squares estimate $\boldsymbol{(X'X)^{-1}X'Y}$ where the weights are the two precisions $\boldsymbol{\Sigma^{-1}}$and $\boldsymbol{X'X}/\sigma^2$.

The posterior distribution does not depend on any particular choice of $\boldsymbol{(\mu, \Sigma)}$, but it does depend on the fact that the prior distribution is normal because that's what gives us the quadratic form in the exponent. That's one reason why we take the prior distribution for $\boldsymbol{\beta}$ to be normal, which makes the derivation easy.

Now let's look at the posterior more closely, see what it implies for (β_0, β_1), and see how sensitive the conclusions are to the choice of $\boldsymbol{(\mu, \Sigma)}$. We're also assuming σ as known.

We begin with the choice

$$\boldsymbol{\mu} = \begin{pmatrix}\mu_0\\ \mu_1\end{pmatrix} = \begin{pmatrix}0\\ 0\end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix}\sigma_0^2 & 0\\ 0 & \sigma_1^2\end{pmatrix} = \begin{pmatrix}10^6 & 0\\ 0 & 10^6\end{pmatrix}, \quad \text{and} \quad \sigma = 0.05$$

The large diagonal entries in $\boldsymbol{\Sigma}$ say that we have very little *a priori* knowledge of $\boldsymbol{\beta}$. We use R to calculate the posterior mean and covariance.

```
> library(faraway)
> attach(stat500)
> dim(stat500)
[1] 55  4
> head(stat500)
  midterm final   hw total
1    24.5  26.0 28.5  79.0
2    22.5  24.5 28.2  75.2
3    23.5  26.5 28.3  78.3
4    23.5  34.5 29.2  87.2
5    22.5  30.5 27.3  80.3
6    16.0  31.0 27.5  74.5
> model1 <- lm(final~midterm)
> coef(model1)
```

```
(Intercept)     midterm
 15.0461727   0.5632756
> MU <- c(0,0)
> SIGMA <- diag(rep(10^6,2))
> SIGMA
      [,1]  [,2]
[1,] 1e+06 0e+00
[2,] 0e+00 1e+06
> sigma <- 0.05
> y <- final
> x <- cbind(1, midterm)
> x[1:5,]
        midterm
 [1,] 1    24.5
 [2,] 1    22.5
 [3,] 1    23.5
 [4,] 1    23.5
 [5,] 1    22.5
> SIGMAstar <- solve( solve(SIGMA) + t(x) %*% x / (sigma^2))
> MUstar <- SIGMAstar %*% (solve(SIGMA) %*% MU + t(x) %*% y / (sigma^2))
> MUstar
              [,1]
        15.0461727
midterm  0.5632756
> SIGMAstar
                            midterm
         8.766427e-04 -4.090859e-05
midterm -4.090859e-05  2.013398e-06
```

The result is

$$\boldsymbol{\mu}^* = \begin{pmatrix} 15.0462 \\ 0.5633 \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Sigma}^* = \begin{pmatrix} 8.7664 \times 10^{-4} & -4.0909 \times 10^{-5} \\ -4.0909 \times 10^{-5} & 2.0134 \times 10^{-6} \end{pmatrix}.$$

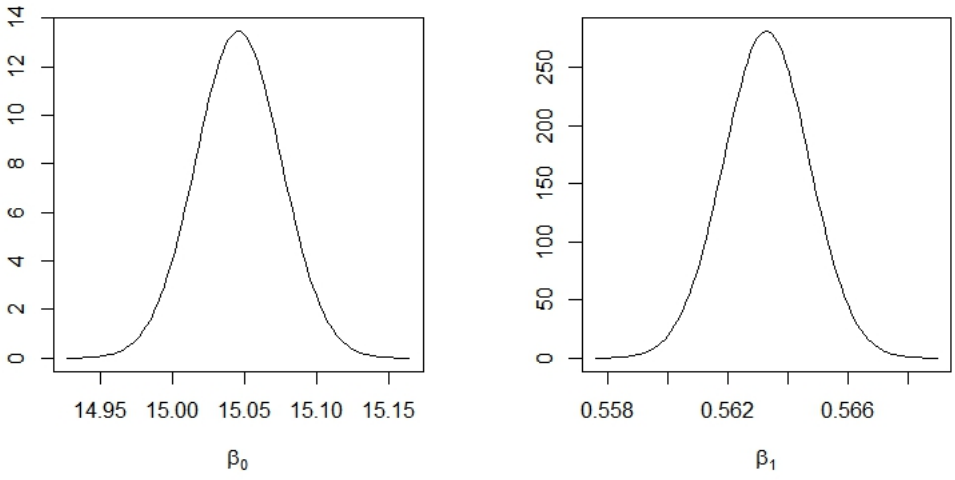

Figure 2: Posterior densities of β_0 and β_1 using the `stat500` sample data

Fig. 2 shows the posterior densities. The posterior density of β_0 is not very meaningful because it pertains to the final exam score when the midterm score is 0. Since our data was collected at midterm scores between 8 and 30, extrapolating to scores around 0 would be dangerous. And because β_0 is not meaningful, neither is the joint density of (β_0, β_1). On the other hand, our inference for β_1 is more meaningful. It says that final exam score goes up about .56 (± 0.001 or so) points for every point increase in midterm score.

R codes that created Fig. 2 are shown below.

```
> m <- MUstar[1]
> s <- sqrt(SIGMAstar[1,1])
> t <- seq(m-4*s, m+4*s, length=100)
> par(mfrow=c(1,2))
> plot(t, dnorm(t,m,s),type="l",xlab=expression(beta[0]),ylab="")
>
> m <- MUstar[2]
> s <- sqrt(SIGMAstar[2,2])
> t <- seq(m-4*s, m+4*s, length=100)
> plot(t, dnorm(t,m,s),type="l",xlab=expression(beta[1]),ylab="")
```

Now let's investigate the role of the prior density. We notice that the prior SD of β_1 was 10^3 while the posterior SD is $\sqrt{2.0134 \times 10^{-6}} \approx 0.00142$. In other words, the data has reduced the uncertainty by a huge amount. Because there's so much information in the data, we expect the prior to have little influence. We can verify this by considering priors with different SD's and comparing their posteriors. To that end, consider different priors

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_0 \\ \mu_1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_0^2 & 0 \\ 0 & \sigma_1^2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \text{and} \quad \sigma = 0.05$$

With this prior, the posterior becomes

$$\boldsymbol{\mu}^* = \begin{pmatrix} 15.0330 \\ 0.5639 \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Sigma}^* = \begin{pmatrix} 8.7587 \times 10^{-4} & -4.0873 \times 10^{-5} \\ -4.0873 \times 10^{-5} & 2.0117 \times 10^{-6} \end{pmatrix},$$

nearly identical to the previous posterior. That's because the data contain much more information than the prior, so the prior plays a negligible role in determining the posterior distribution. The posterior precision matrix (inverse of the covariance matrix) is $\Sigma^{-1} + \boldsymbol{X}'\boldsymbol{X}/\sigma^2$, and $\boldsymbol{X}'\boldsymbol{X}/\sigma^2$ is much bigger than those in Σ^{-1} whichever prior we use.

Even if we did a careful job of assessing our prior, it would be influential only if our prior precision matrix had entries of the same order of magnitude as $\boldsymbol{X}'\boldsymbol{X}/\sigma^2$. Since that's unlikely – our true *a priori* variances are probably not as small as 1/30 – there's little to be gained by choosing the prior carefully and much effort can be saved by using an arbitrary prior, as long as it has reasonably large

variances

Ex 2. Another Bayesian Approach to Linear Regression

```
## Annette Dobson (1990) "An Introduction to Generalized Linear Models", p 9: Plant Weight Data.
ctl <- c(4.17,5.58,5.18,6.11,4.50,4.61,5.17,4.53,5.33,5.14)
trt <- c(4.81,4.17,4.41,3.59,5.87,3.83,6.03,4.89,4.32,4.69)
group <- gl(2, 10, 20, labels = c("Ctl","Trt"))
weight <- c(ctl, trt)
model1 <- lm(weight~group)
```

The standard non-informative prior for the linear regression analysis example takes an improper (uniform) prior on the coefficients of the regression (β: the intercept and the effects of the "Trt" variable) and the logarithm of the residual variance σ^2. With these priors, the posterior distribution of $\hat{\beta}$ conditional on $\hat{\sigma}^2$ and the response variable y is: $\hat{\beta} \,|\, \hat{\sigma}^2, y \sim N(\beta, V_\beta \sigma^2)$.

The marginal posterior distribution for $\hat{\sigma}^2$ is a scaled inverse χ^2 distribution with scale s and $n-p$ degrees of freedom, where n is the number of data points and p the number of predictor variables+1. In our example $n = 20, k = 2$, while s^2 is the standard frequentist estimate of the residual variance.

The quantities $\hat{\beta}, s^2$ are directly available from the information returned by R's `lm`, while V_β can be computed from the `qr` element of the `lm` object as shown below:

```
QR <- model1$qr
df.residual <- model1$df.residual
R <- qr.R(QR) ## R component
coef <- model1$coef
Vb <- chol2inv(R) ## variance(unscaled)
s2 <- (t(lmfit$residuals)%*%lmfit$residuals)
s2 <- s2[1,1]/df.residual
```

To compute the marginal distribution of $\hat{\beta} \,|\, y$ we can use a simple Monte Carlo algorithm, first drawing σ^2 from the posterior, and then $\hat{\beta} \,|\, \hat{\sigma}^2, y$. The following function will do this; it accepts as arguments an `lm` object and the desired number of Monte Carlo samples and returns everything in a dataframe for further processing:

```
## function to compute the Bayesian analog of the lmfit
## using non-informative priors and Monte Carlo scheme
## based on N samples

Bayesfit <- function(lmfit,N){
    QR <- lmfit$qr
    df.residual <- lmfit$df.residual
    R <- qr.R(QR) ## R component
    coef <- lmfit$coef
    Vb <- chol2inv(R) ## variance(unscaled)
    s2 <- (t(lmfit$residuals)%*%lmfit$residuals)
    s2 <- s2[1,1]/df.residual

    ## now to sample residual variance
    sigma <- df.residual*s2/rchisq(N,df.residual)
    coef.sim <- sapply(sigma,function(x) mvrnorm(1,coef,Vb*x))
```

```
    ret <- data.frame(t(coef.sim))
    names(ret) <- names(lmfit$coef)
    ret$sigma <- sqrt(sigma)
    ret
}

Bayes.sum<-function(x)
    {
    c("mean"=mean(x),"se"=sd(x),"t"=mean(x)/sd(x),
      "median"=median(x),"CrI"=quantile(x,prob=0.025),"CrI"=quantile(x,prob=0.975))
    }
```

To use these functions and contrast Bayesian and frequentist estimates, one simply needs to fit the regression model with `lm`, call the `Bayesfit` function to run the Bayesian analysis and pass the results to `Bayes.sum` as shown below:

```
> set.seed(1234)  ## To make this reproducible
> library(MASS)
> model1 <- lm(weight~group)
> bf <- Bayesfit(model1,10000)
> t(apply(bf,2,Bayes.sum))
                  mean        se         t     median   CrI.2.5% CrI.97.5%
(Intercept)  5.0332172 0.2336049 21.545857  5.0332222  4.5651327 5.4902380
groupTrt    -0.3720698 0.3335826 -1.115375 -0.3707408 -1.0385601 0.2895787
sigma        0.7262434 0.1275949  5.691789  0.7086832  0.5277051 1.0274083
> summary(model1)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.0320     0.2202  22.850 9.55e-15 ***
groupTrt     -0.3710     0.3114  -1.191    0.249
---
Residual standard error: 0.6964 on 18 degrees of freedom
Multiple R-squared:  0.07308,   Adjusted R-squared:  0.02158
F-statistic: 1.419 on 1 and 18 DF,  p-value: 0.249
```

It can be seen that the Bayesian estimates are almost identical to the frequentist ones (up to 2 significant digits, which is the limit of precision of the Monte Carlo run based on 10000 samples), but uncertainty in terms of these estimates (the standard deviation) and the residual variance is larger. This conservativeness is an inherent feature of Bayesian analysis which guards against too many false positives.

Metropolis, Metropolis-Hastings, and Gibbs Sampling

In "Markov chain Monte Carlo", the term "Monte Carlo" refers to evaluating an integral by using many random draws from a distribution. To fix ideas, suppose we want to evaluate the posterior distribution. Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$. If we could generate many samples $\boldsymbol{\theta_1}, \ldots, \boldsymbol{\theta_M}$ of $\boldsymbol{\theta}$ (where $\boldsymbol{\theta_i} = (\theta_{i,1}, \ldots, \theta_{i,k})$) from its posterior distribution then we could approximate the posterior distribution by

1. discarding $\theta_{i,2}, \ldots, \theta_{i,k}$ from each iteration,
2. retaining $\theta_{1,1}, \ldots, \theta_{M,1}$,
3. using $\theta_{1,1}, \ldots, \theta_{M,1}$ and standard density estimation techniques to estimate $p(\theta_1 \mid y)$, or
4. for any set A, using
$$\frac{\text{number of } \theta_{i,1}\text{'s in } A}{M}$$
as an estimate of $P(\theta_1 \in A \mid y)$.

That's the idea behind Monte Carlo integration.

The term "Markov chain" refers to how the samples $\boldsymbol{\theta_1, \ldots, \theta_M}$ are produced. In a Markov chain there is a *transition density* or *transition kernel* $k(\theta_i \mid \theta_{i-1})$ which is a density for generating $\boldsymbol{\theta_i}$ given $\boldsymbol{\theta_{i-1}}$. We first choose $\boldsymbol{\theta_1}$ almost arbitrarily, then generate $(\boldsymbol{\theta_2} \mid \boldsymbol{\theta_1})$, $(\boldsymbol{\theta_3} \mid \boldsymbol{\theta_2})$, and so on, in succession, for as many steps as we like. Each $\boldsymbol{\theta_i}$ has a density $p_i = p(\boldsymbol{\theta_i})$ that depends on $\boldsymbol{\theta_1}$ and the transition kernel. But,

1. under some fairly benign conditions, the sequence $p_1, p_2, \ldots$ converges to a limit p, the *stationary distribution*, that does not depend on $\boldsymbol{\theta_1}$;
2. the transition density $k(\theta_i \mid \theta_{i-1})$ can be chosen so that the stationary distribution p is equal to $p(\boldsymbol{\theta} \mid y)$;
3. we can find an m such that $i > m \Rightarrow p_i \approx p = p(\boldsymbol{\theta} \mid y)$;
4. then $\boldsymbol{\theta_{m+1}, \ldots, \theta_M}$ are, approximately, a sample from $p(\boldsymbol{\theta} \mid y)$.

The Metropolis-Hastings algorithm [Metropolis et al., 1953; Hastings, 1970] is one way to construct an MCMC algorithm whose stationary distribution is $p(\boldsymbol{\theta} \mid y)$. It works according to the following steps.

1. Choose a proposal density $g(\boldsymbol{\theta^*} \mid \boldsymbol{\theta})$.
2. Choose $\boldsymbol{\theta_1}$.
3. For $i = 2, 3, \ldots$
   - Generate a proposal $\boldsymbol{\theta^*}$ from $g(\boldsymbol{\theta^*} \mid \boldsymbol{\theta_{i-1}})$.
   - Set
   $$r = min\left\{1, \frac{p(\boldsymbol{\theta^*} \mid y) \cdot g(\boldsymbol{\theta_{i-1}} \mid \boldsymbol{\theta^*})}{p(\boldsymbol{\theta_{i-1}} \mid y) \cdot g(\boldsymbol{\theta^*} \mid \boldsymbol{\theta_{i-1}})}\right\}.$$
   - Set
   $$\theta_i = \begin{cases} \boldsymbol{\theta^*} & \text{with probability } r, \\ \boldsymbol{\theta_{i-1}} & \text{with probability } 1 - r. \end{cases}$$

Step 3 defines the transition kernel k. In many MCMC chains, the acceptance probability r may be strictly less than one, so the kernel k is a mixture of two parts: one that generates a new value of $\boldsymbol{\theta_{i+1}} \neq \boldsymbol{\theta_i}$ and one that sets $\boldsymbol{\theta_{i+1}} = \boldsymbol{\theta_i}$.

To illustrate MCMC, suppose we want to generate a sample $\theta_1, \ldots, \theta_{10,000}$ from the Beta(5, 2) distribution. We arbitrarily choose a proposal density $g(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}) = \text{unif}(\theta - 0.1, \theta + 0.1)$ and arbitrarily choose $\theta_1 = 0.5$. The following R code draws the sample.

Ex 1. 10,000 MCMC samples of the Beta(5,2) density

```
> samp <- rep ( NA, 10000 )
> samp[1] <- 0.5
> for ( i in 2:10000 ) {
+     prev <- samp[i-1]
+     theta_star <- runif ( 1, prev - .1, prev + .1 )
+     r <- min ( 1, dbeta(theta_star,5,2) / dbeta(prev,5,2) )
+     if ( rbinom ( 1, 1, r ) == 1 )
+        new <- theta_star
+     else
+        new <- prev
+        samp[i] <- new
+ }
> par(mfrow=c(3,1))
>
> hist (samp[-(1:1000)], prob=TRUE, xlab=expression(theta),ylab="", main="" )
> x <- seq(0,1,length=200)
> lines (x, dbeta(x,5,2))
> plot (samp, pch=".", col=2, ylab=expression(theta) )
> plot (dbeta(samp,5,2), pch=".", col=4, ylab=expression(p(theta)) )
```

The code `samp[-(1:1000)]` discards the first 1,000 draws in the hope that the sampler will have converged to its stationary distribution after 1,000 iterations. The top panel of Fig. 3 shows the result. The solid curve is the Beta(5, 2) density and the histogram is made from the Metropolis-Hastings samples. They match closely, showing that the algorithm performed well.

Assuming that convergence conditions have been met and that the algorithm is well constructed, MCMC chains are guaranteed eventually to converge and deliver samples from the desired distribution.The middle panel of Fig. 3 shows that the chain spends most of its iterations in values of θ between about 0.6 and 0.9 but makes occasional excursions down to 0.4 or 0.2 or so. After each excursion it comes back to the mode around 0.8. The chain has taken many excursions, so it has explored the space well.

The bottom panel shows that the chain spent most of its time near the mode where $p(\theta) \approx 2.4$ but made multiple excursions down to places where $p(\theta)$ is around 0.5, or even less.

Running Bayesian Analysis in R

JAGS stands for "Just Another Gibbs Sampler" and it's a program for analysis of Bayesian models using MCMC simulation. It's very much like BUGS (Bayesian inference Using Gibbs Sampling)

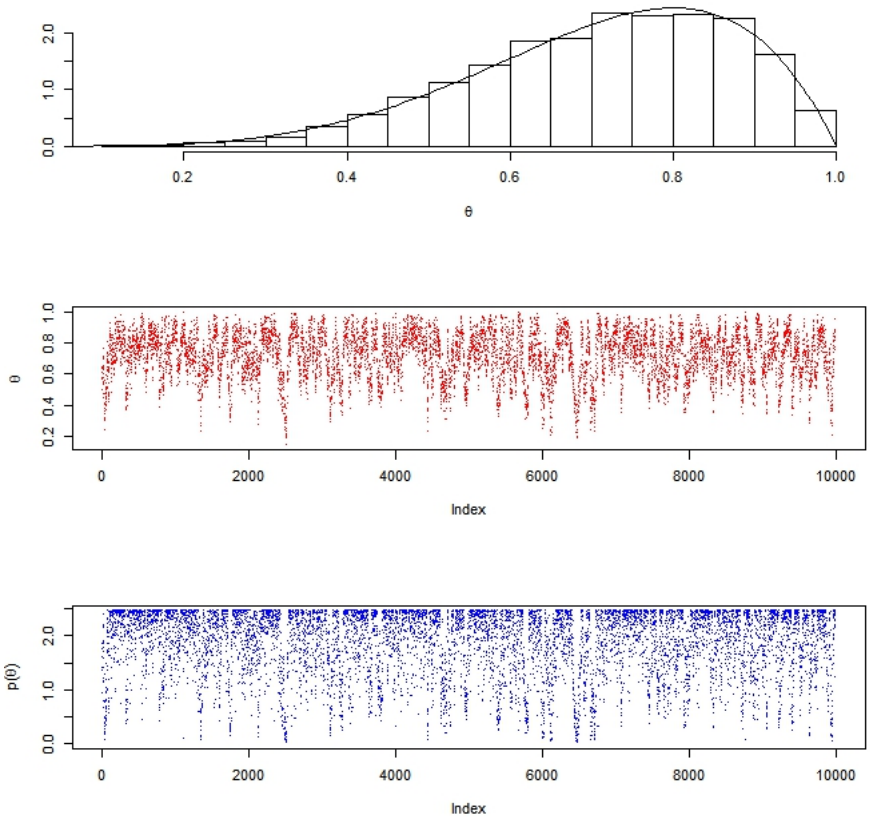

Figure 3: 10,000 MCMC samples of the Beta(5, 2) density. **Top:** histogram of samples from the Metropolis-Hastings algorithm and the Beta(5, 2) density. **Middle:** θ_i plotted against i. **Bottom:** $p(\theta_i)$ plotted against i.

in spirit and allows users to write their own functions, distributions and samplers. JAGS is licensed under the GNU General Public License. First, install JAGS on your computer by visiting `http://mcmc-jags.sourceforge.net/`. Click to "Download JAGS" on your computer.

Next, install the `R2jags` package that allows R to communicate with JAGS and vice versa. Once you load it, you're ready to start Bayesian modeling. You use the BUGS language to write down your model of the likelihood and the priors, incorporating all of the details about hierarchical structure, pseudoreplication, and so forth. After sketching out the mdoel, you write it in a *text* editor and save it as an ASCII file (i.e., `.txt` file). I will show this using the `stat500` sample dataset of the `faraway` library.

Ex 2. MCMC for a simple linear regression using JAGS

First, I typed the following codes and saved it as `myregression.bugs.txt` at `STAT510` sbdirectory of `U:` drive.

```
model {
for(i in 1:N) {
    final[i] ~ dnorm(mu[i], tau)
    mu[i] <- a + b*midterm[i]
    }
a ~ dnorm(0.0,1.0E-4)
b ~ dnorm(0.0,1.0E-4)
sigma <- 1.0/sqrt(tau)
tau ~ dgamma(1.0E-3,1.0E-3)
}
```

Next, go to R and do the following:

```
> N = 55
> data.jags <- list("final","midterm","N")
> library(R2jags)
> model <- jags(data=data.jags,parameters.to.save=c("a","b","tau"),
+ n.iter=100000,model.file="U:\\STAT510\\myregression.bugs.txt",n.chains=3)
> model
Inference for Bugs model at "U:\STAT510\myregression.bugs.txt", fit using jags,
 3 chains, each with 1e+05 iterations (first 50000 discarded), n.thin = 50
 n.sims = 3000 iterations saved
         mu.vect sd.vect    2.5%     25%     50%     75%   97.5%  Rhat n.eff
a         14.994   2.492   9.946  13.384  14.957  16.697  19.915 1.001  3000
b          0.564   0.120   0.327   0.486   0.565   0.643   0.806 1.001  3000
tau        0.057   0.011   0.037   0.049   0.056   0.064   0.080 1.001  3000
deviance 314.763   2.508 311.879 312.956 314.103 315.845 321.268 1.002  2600

For each parameter, n.eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor (at convergence, Rhat=1).

DIC info (using the rule, pD = var(deviance)/2)
pD = 3.1 and DIC = 317.9
DIC is an estimate of expected predictive error (lower deviance is better).

> plot(model)
> model.mcmc <- as.mcmc(model)
> plot(model.mcmc)
> model.mcmc
[[1]]
Markov Chain Monte Carlo (MCMC) output:
Start = 50001
End = 99951
Thinning interval = 50
                a         b deviance        tau
   [1,] 15.700031 0.5289675 311.8174 0.06105786
   [2,] 15.608993 0.5179451 312.3209 0.05565197
   [3,] 13.841507 0.6438927 313.5529 0.06826426
   ..................
<<< Omitted to save the world. >>>
> library(lattice)
> densityplot(model.mcmc)
```

To compare the result with those obtained by the linear model, here is the printout of the usual linear model fit.

```
> model1 <- lm(final~midterm)
> summary(model1)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  15.0462     2.4822   6.062 1.44e-07 ***
midterm       0.5633     0.1190   4.735 1.67e-05 ***
---
Residual standard error: 4.192 on 53 degrees of freedom
Multiple R-squared:  0.2973,    Adjusted R-squared:  0.284
F-statistic: 22.42 on 1 and 53 DF,  p-value: 1.675e-05
```

As you can see, the parameter estimates are quite close to those obtained by the linear model (intercept = 14.994±2.492 rather than 15.0462±2.4822; slope = 0.564±0.120 rather than 0.5633±0.1190). The deviance (i.e., Residual SS) is substantially smaller (314.763 rather than 931.29).

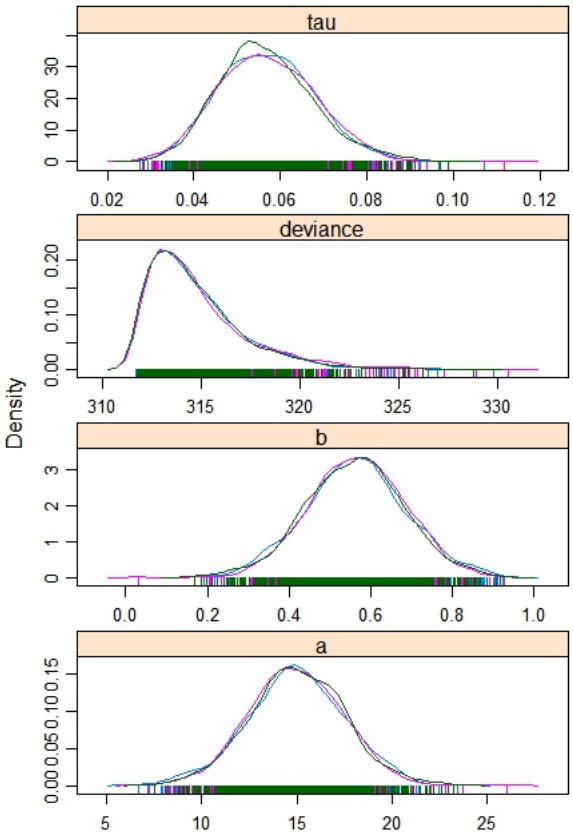

Figure 4: Density plots of the estimates

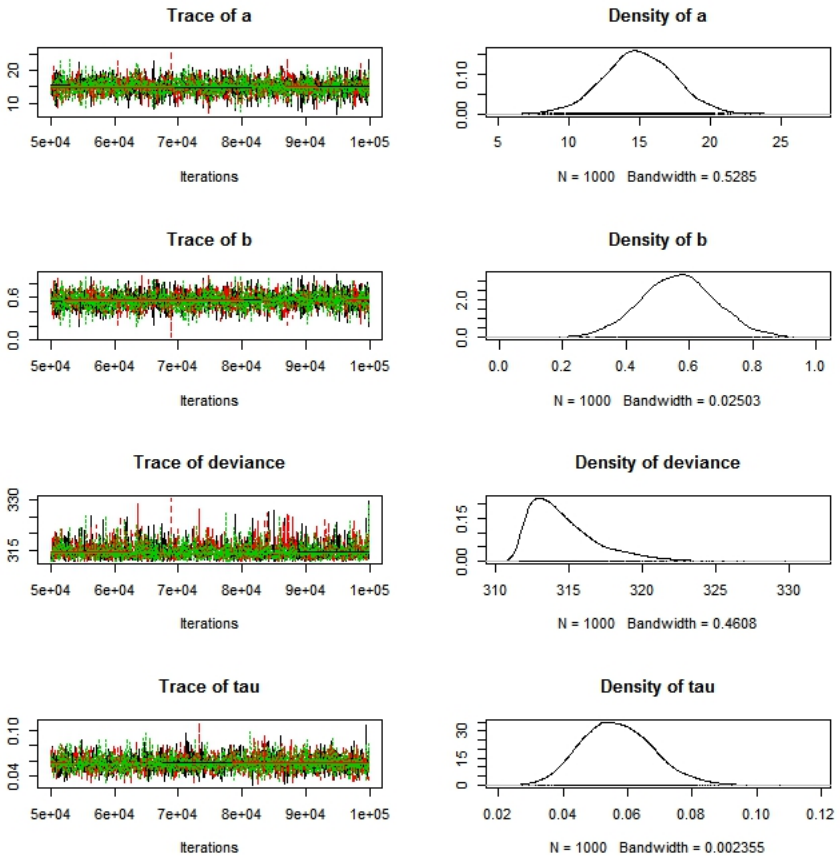

Figure 5: Another kinds of density plots of the estimates

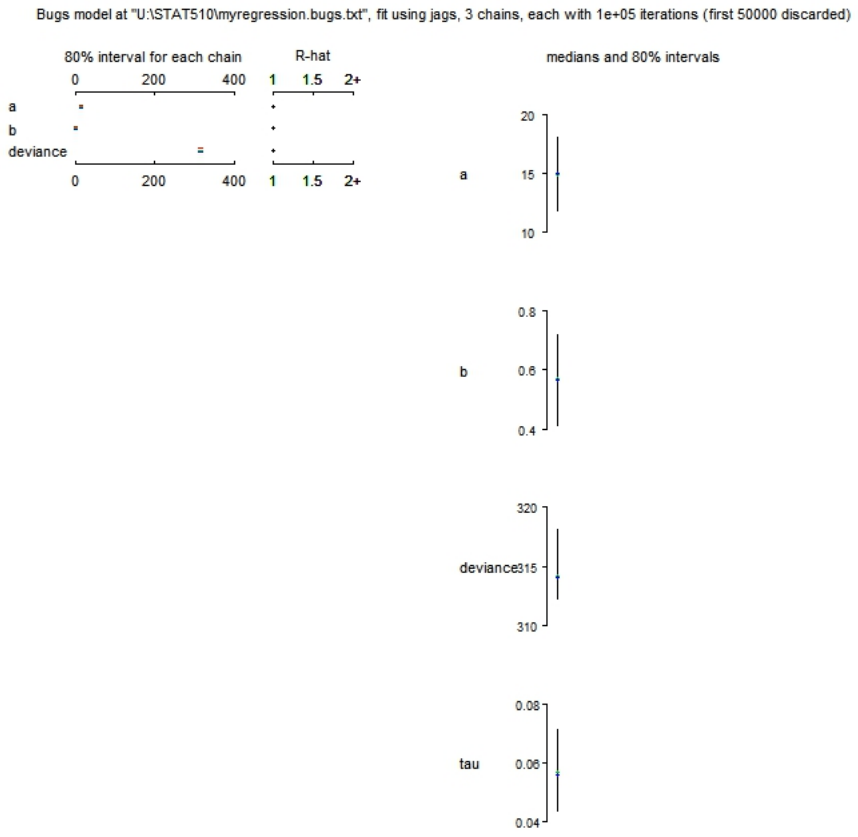

Figure 6: Strip diagrams with "credible" interval bars for the parameters, the deviance and tau (the reciprocal of the error variance).