Chapter 9.7b Using JAGS

Jim Albert and Monika Hu

Chapter 9 Simulation by Markov Chain Monte Carlo

Posterior predictive checking

- Basic idea is to simulate a number of replicated datasets from the posterior predictive distribution.
- See how the observed sample compares to the replications.
- If the observed data does resemble the replications, one says that the observed data is consistent with predicted data from the Bayesian model.

Snowfall Example

- Suppose one wishes to simulate a replicated sample from the posterior predictive distribution.
- Simulate a sample of values $\tilde{y}_1, \ldots, \tilde{y}_{20}$ from the posterior predictive distribution as follows.
- Draw a set of parameter values, say μ^*, σ^* from the posterior distribution of (μ, σ).
- Given these parameter values, we simulate $\tilde{y}_1, \ldots, \tilde{y}_{20}$ from the Normal sampling density with mean μ^* and standard deviation σ^*.

R Function

- Recall that the simulated posterior values are stored in the matrix `post`. We write a function `postpred_sim()` to simulate one sample from the predictive distribution.

```
post <- data.frame(posterior$mcmc[[1]])
postpred_sim <- function(j){
  rnorm(20, mean = post[j, "mu"],
        sd = post[j, "sigma"])
}
print(postpred_sim(1), digits = 3)
 [1]   5.37  10.91  40.87  15.94  16.93  43.49  22.48
 [8]  -6.43   3.26   7.30  35.27  20.79  21.47  16.62
[15]   5.45  44.69  23.10 -18.18  26.51   6.84
```

Repeat Process

- If this process is repeated for each draw from the posterior distribution, then one obtains 5000 samples of size 20 drawn from the predictive distribution.
- The function `sapply()` is used together with `postpred_sim()` to simulate 5000 samples that are stored in the matrix `ypred`.

```
ypred <- t(sapply(1:5000, postpred_sim))
```

Graph

- Figure on next slide displays histograms of the predicted snowfalls from eight of these simulated samples and the observed snowfall measurements are displayed in the lower right panel.
- The center and spread of the observed snowfalls appear to be similar in appearance to the eight predicted snowfall samples from the fitted model.
- One concern is that we observed an "outlying" snowfall of 65.1 inches in our sample and none of our eight samples had a snowfall this large. Perhaps there is an outlier in our sample that is not consistent with predictions from our model.

Graph of Repliciated Datasets

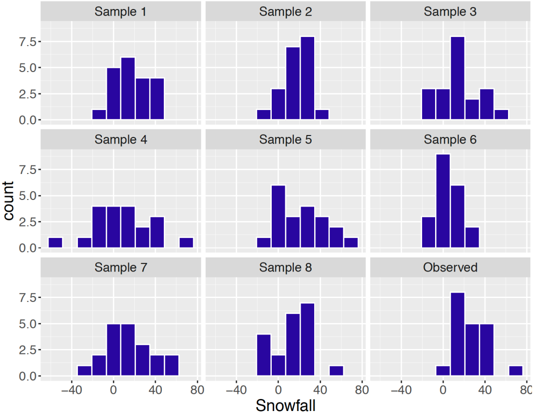

Figure 1: Histograms of eight simulated predictive samples and the observed sample for the snowfall example.

Checking Function

- When one notices a possible discrepancy between the observed sample and simulated prediction samples, one thinks of a checking function $T()$ that will distinguish the two types of samples.
- Here our observation suggests that we use $T(y) = \max y$ as a checking function.
- One simulates the posterior predictive distribution of $T(\tilde{y})$ by evaluating the function $T()$ on each simulated sample from the predictive distribution.
- In R, this is conveniently done using the `apply()` function.

```
postpred_max <- apply(ypred, 1, max)
```

Interpretation

- If the checking function evaluated at the observed sample $T(y)$ is not consistent with the distribution of $T(\tilde{y})$, predictions from the model are not similar to the observed data.
- Figure on the next slide displays a histogram of the predictive distribution of $T(y)$ in our example where $T()$ is `max()`, and the observed maximum snowfall is shown by a vertical line.
- Here the observed maximum is in the right tail of the distribution – the interpretation is that this largest snowfall of 65.1 inches is not predicted from the model.
- Maybe the data follow a distribution with flatter tails than the Normal.

Histogram of post. pred. distribution of checking function

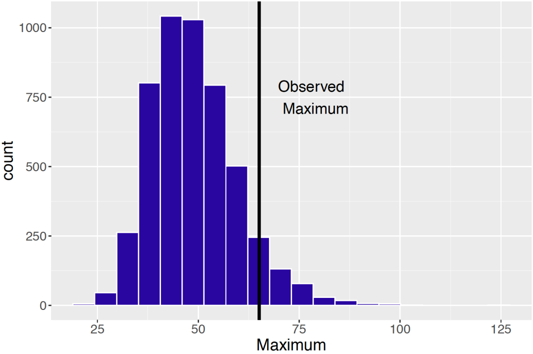

Figure 2: Histogram of the posterior predictive distribution of T(y) where T() is the maximum function. The vertical line shows the location of the observed value T(y).

Comparing two proportions

- Consider a problem comparing two proportions from independent samples.
- To better understand the behavior of Facebook users, a survey was administered to 244 students. Each student was asked their gender and the average number of times they visited Facebook in a day. The number of daily visits is "high" if the number of visits is 5 or more; otherwise it is "low". One obtains the two by two table of counts as shown this table.

	High	Low
Male	y_M	$n_M - y_M$
Female	y_F	$n_F - y_F$

Sampling Model

- The random variable Y_M represents the number of males who have a high number of Facebook visits in a sample of n_M, and Y_F and n_M are the analogous count and sample size for women.
- Reasonable to assume that Y_M and Y_F are independent with Y_M distributed Binomial with parameters n_M and p_M, and Y_F is Binomial with parameters n_F and p_F.

	High	Low
Male	p_M	$1 - p_M$
Female	p_F	$1 - p_F$

Learning About Association

- One is interested in the association between gender and Facebook visits.
- The odds of "high" for the men and odds of 'high" for the women are defined by

$$\frac{p_M}{1-p_M}, \frac{p_F}{1-p_F}$$

- The odds ratio

$$\alpha = \frac{p_M/(1-p_M)}{p_F/(1-p_F)},$$

  is a measure of association in this two-way table.
- If $\alpha = 1$, this means that $p_M = p_L$ – this says that tendency to have high visits to Facebook does not depend on gender.

Log Odds

- Can express association on a log scale – the log odds ratio λ is written as

$$\lambda = \log \alpha = \log \left(\frac{p_M}{1 - p_M} \right) - \log \left(\frac{p_F}{1 - p_F} \right).$$

- If gender is independent of Facebook visits, then $\lambda = 0$.

Prior

- One's prior beliefs about association in the two-way table is expressed in terms of the log odds ratio.
- If one believes that gender and Facebook visits are independent, then the log odds ratio is assigned a Normal prior with mean 0 and standard deviation σ.
- The mean of 0 reflects the prior guess of independence and σ indicates the strength of the belief in independence.
- Define the mean of the logits

$$\theta = \frac{\mathrm{logit}(p_M) + \mathrm{logit}(\mathrm{p_F})}{2}$$

and assume that θ has a Normal prior with mean θ_0 and standard deviation σ_0 (precision ϕ_0).

Using JAGS

- Write a script defining the model.

```
modelString = "
model{
## sampling
yF ~ dbin(pF, nF)
yM ~ dbin(pM, nM)
logit(pF) <- theta - lambda / 2
logit(pM) <- theta + lambda / 2
## priors
theta ~ dnorm(mu0, phi0)
lambda ~ dnorm(0, phi)
}
"
```

Comments

- The two first lines define the Binomial sampling models, and the logits of the probabilities are defined in terms of the log odds ratio `lambda` and the mean of the logits `theta`.
- In the priors part of the script, note that `theta` is assigned a Normal prior with mean `mu0` and precision `phi0`, and `lambda` is assigned a Normal prior with mean 0 and precision `phi`.

Data

- One observes 75 of the 151 female students are high visitors of Facebook, and 39 of the 93 male students are high visitors.
- Enter data and the values of the prior parameters are entered into R by use of a list. Note that `phi = 2` indicating some belief that gender is independent of Facebook visits, and `mu0 = 0` and `phi0 = 0.001` reflecting little knowledge about the location of the logit proportions.

```
the_data <- list("yF" = 75, "nF" = 151,
                 "yM" = 39, "nM" = 93,
                 "mu0" = 0, "phi0" = 0.001, "phi" = 2)
```

Run JAGS

Using the `run.jags()` function, we take an adapt period of 1000, burn-in period of 5000 iterations and collect 5000 iterations, storing values of `pF`, `pM` and the log odds ratio `lambda`.

```
posterior <- run.jags(modelString,
                      data = the_data,
                      n.chains = 1,
                      monitor = c("pF", "pM", "lambda"),
                      adapt = 1000,
                      burnin = 5000,
                      sample = 5000)
```

Posterior Inference

- Figure on the next slide displays a density estimate of the posterior draws of the log odds ratio λ.
- A reference line at $\lambda = 0$ is drawn on the graph which corresponds to the case where $p_M = p_L$.
- The probability women are more likely than men to have high visits in Facebook is answered by computing the posterior probability $Prob(\lambda < 0 \mid data)$ that is computed to be 0.874.
- Based on this computation, one concludes that it is very probable that women have a higher tendency than men to have high visits on Facebook.

Graph of posterior of log odds ratio

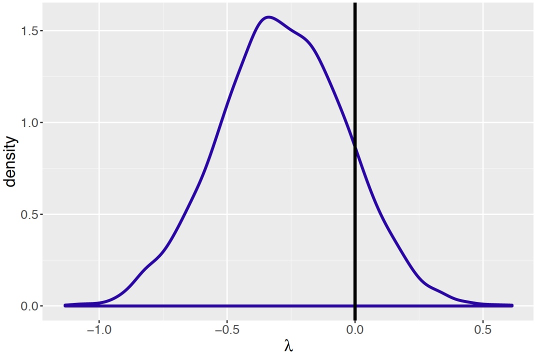

Figure 3: Posterior density estimate of simulated draws of log odds ratio for visits to Facebook example. A vertical line is drawn at the value 0 corresponding to no association between gender and visits to Facebook.