Lab 6 – Posterior predictive check and model selection

Posterior predictive check

All statistical inference is based on some type of statistical model. A truly fundamental requirement for reliable inference is that the statistical model is capable of giving rise to the data. This motivates the need for model checking. Models that fail checks should be discarded at the outset.

The Bayesian approach provides a method, simple to implement, that allows you to check if your model is capable of producing the data. It is called a *posterior predictive check*. The algorithm goes like this:

1. Simulate a new data set at each iteration of the MCMC. This sounds formidable, but it is really no more than drawing a random variable from the likelihood. So, for example if your likelihood is

```
y[i] ~ dnorm(mu[i], tau)
```

you can simualte a new data set by embedding

```
y.sim[i] ~ dnorm(mu[i], tau)
```

in the same for loop.

2. Calculate a test statistic based on the real and the simulated data. The test statistic could be a mean, standard deviation, coefficient of variation, discrepancy, minimum, maximum – really any function that helps you compare the simulated and real data.

3. We are interested in calculating a Bayesian p value, the probability that the test statistic computed from the simulated data is more extreme than the test statistic computed from the real data. There is evidence of lack of fit – the model cannot give rise to the data – if the Bayesian p value is large or small. We want values between, say, .10 and .90, ideally close to .5. To obtain this the Bayesian p we use the JAGS step(x) function that returns 0 if x is less 0 and 1 otherwise. So, presume our test statistic for was the standard deviation. Consider the following pseudo-code:

```
for(i in 1:length(y)){
  mu[i] <- prediction from model
  y[i] ~ dnorm(mu[i], tau)
  y.sim[i] ~ dnorm(mu[i], tau)
}
sd.data<-sd(y[])
sd.sim <-sd(y.sim[])
```

```
p.sd <- step(sd.sim - sd.data)
```

That is all there is to it. You then include p.sd in your jags or coda object.

Problem

Return to the pooled model you developed in the first problem of multi-level modeling exercise. do posterior predictive checks using the mean, standard deviation, minimum, and discrepancy as test statistics. The discrepancy statistic is $\sum_{i=1}^{n}(y_i - \mu_i)^2$ where μ_i is the ith prediction of your model. Overlay the posterior distribution of the simulated data on the histogram of the read data (density on y axis, not frequency). What do you conclude? Is there any indication of lack of fit? Enough to discard the model?