LAB 8. MODEL EVALUATION

Purpose. Making models requires making decisions. We must decide about what sort of equations to use to represent ecological process. We must choose statistical distributions to represent the uncertainties that arise in our model of the process and the way we observe it. If we make good decisions, then the stochastic model offers a reasonable approximation of the data. If we make poor ones, the model and the data differ in important ways. The question arises, how do we know whether our model approximates the data in a reasonable way? The answer to this question requires that we assess the "goodness of fit" of the model. Posterior predictive checks are tools for estimating goodness of fit.

Approach. The idea is this. Once we have our model in hand, we can use it to simulate data. This can be done in R by drawing random values from the distributions of the estimates of the parameters and calculating a set of predictions from multiple draws. If we store enough of those predictions, then we can plot the probability density of the simulated data and overlay it on the real data. We seek a good match between a normalized histogram of the data and the probability density of the real data.

Methods. The problem with comparing histograms and density plots is that it is subjective, and ideally we would like to objectively test for goodness of fit. There are a number of objective ways to assess goodness of fit by using the following steps.

1. Using statistics from the data

1. Calculate a summary statistic using the real data. This statistic might be a mean, minimum, maximum standard deviation, quantile etc. The choice of the statistic really depends on the problem at hand. For example, if it is really important that your model be able to accurately estimate extreme values, then you should use a statistic like the .975% quantile. It is good practice to include some statistic representing the central tendency as well as some statistic representing the dispersion of the data.

2. Simulate n_{sim} data sets using your model. Calculate a summary statistic from each replication of the simulated data for comparison to the same statistic calculated from the observed data. (Note: The sample variance is not a good candidate for the summary statistic because the posterior distribution will automatically be centered on the observed value so it's preferable to use the coefficient of variation. See Gelman et al. (1995). Say we have a model that predicts tree growth for n trees as a function of diameter. The JAGS code is for fitting the data is:

```
for(i in 1:n){
        mu[i] <- b0 + b1*diam[i] # fitted data
        y[i] ~ dnorm(mu[i],tau)
```

and we could simulate the posterior distribution of n y.sim unobserved trees by adding the following line of code:

```
y.sim[i] ~ dnorm(mu[i],tau) } # simulated data from same distribution
```

3. If the fit of the model is good (but obviously not perfect) then we expect roughly half of the y_{sim} summary statistics calculated from the simulated data to exceed the summary statistic calculated from the from the real data (the y's) and half to be less than the summary statistic calculated from the real data. What this means, really, is that the distribution of the summary statistic from the simulated data is centered on the value from the real data, indicating a good fit.

4. If the fit is poor then we expect a disproportionate number of the statistics calculated from the simulated data to exceed or fall below the statistics calculated from the real data. In this case the summary statistics from the data simulated from the model consistently over or underestimate the

summary of the real data, which is to say the distribution of the summary statistic from the simulated data is not centered on the value from the real data. This, of course, indicates a poor fit.

To implement this approach, we calculate some statistics (mean and CV) for the data and the y.sim's:

```
cv.y <- sd(y[ ])/mean(y[ ])
cv.y.sim <- sd(y.sim[])/mean(y.sim[ ])
mean.y <-mean(y[])
mean.y.sim <-mean(y.sim[])
```

The bracket notation averages across all values stored in the vector y or y.sim.

5. A Bayesian P value (P_B) calculates the proportion of statistics from y_{rep} data sets that have a statistic exceeding the summary statistic of the real data. As a rule of thumb, if the value $P_B \leq 0.10$ or $P_B \geq 0.90$, then we need to be concerned about lack of fit in our model. In JAGS code, we can calculate the p values by using the step function [if argument inside step> 0, step=1, else step= 0], as follows:

```
pvalue.mean <-step(mean.y.sim - mean.y)
pvalue.cv <- step(cv.y.sim-cv.y)
```

2. The math

We can think of our summary statistic from the real data as being analogous to the test statistic that you have grown to love in your frequentist training. As usual, θ is a parameter or vector of parameters in our model. Let T(y, θ) be the test statistic (i.e. one of the summary statistics above, mean, standard deviation, etc). Let $T(y^{rep}, \theta)$ be the value of the test statistic calculated from the simulated y_{sim}. The Bayesian P value, P_B, is defined as the probability that the simulated data could be more extreme than the observed data (Gelman et al. 1995) as measured by the quantity:

$$P_B = Pr\,[T(y^{rep}, \theta) \geq T(y, \theta)|y] \qquad \text{(Eqn. 1)}$$

Equation 1 is a two tailed probability, which means that values that are very high or very low indicate lack of fit. Thus, a model is suspect if the tail area probability is close to 0 or 1, which indicates that the pattern would be improbable in replications of the data if the model were true (Gelman et al. 1995).

Omnibus tests

There is some very real value in using the approach above, where we examine goodness of fit of specific aspects of the model. Anything that the model predicts and that can be calculated from the data can be used to form tests of goodness of fit. Good examples can be found in Gelman et al. (1995, Table 6.1) and Gelman et al. (2009, example 24.2). The reason that this is a good approach is that is allows us to focus in on specific flaws if they exist. For example, we might discover that the variance in a Poisson model always is consistently less than the variance in the data, which might motivate us to use a distribution with more flexible variance, for example, a negative binomial.

However, we may also desire an approach that seeks to summarize goodness of fit (or its absence) in a single quantity. In this case the test statistic (a.k.a summary statistic from the data) is some measure of discrepancy between the model and the data, e.g., sums of squared deviations or an X^2 statistic. Here is how it works, illustrated with sums of squares.

1. For each data point in the data set, estimate the squared residual as the squared difference between the model prediction and the data.
2. Generate a new data point using the model and appropriate sources of uncertainty.
3. Estimate the squared difference between the model prediction and the new, simulated data point. Call this quantity a new residual.

4. Sum the residuals (from part 1).
5. Sum the new residuals (from part 3).
6. Repeat this many times. From these repetitions, estimate the proportion of the sums of the new residuals that exceeds the sum of the observed residuals (from 4).

These are the steps in JAGS for the tree growth model:

```
for(j in 1:n){
        sq[j] <- (y[j]-mu[j])^2
        sq.new[j] <- (y.sim[j]-mu[j])^2
        }
fit <- sum(sq[])
fit.new <- sum(sq.new[])
pvalue.fit <- step(fit.new-fit)
```

Exercise 1. Let's return to the invasive species problem from last week. Now, reformulate the invasive species problem so that the model does not include a site effect (i.e., number of invasives only depends on site disturbance and the intercept is the same across all sites). Run posterior predictive checks on this model. Generate some simulated data, and go through the steps of a posterior predictive check:

(a) Calculate the mean and coefficient of variation of the observed and simulated data
(b) Plot histograms of observed and simulated data.
(c) Calculate p-values for mean and CV's as described above.
(d) Calculate p values for Sums of Squares.

Exercise 2. Now, use the original hierarchical version of the invasives model (the one that includes a roads effect). Run posterior predictive checks on this model. Fill in table below. Is it a better fit? Calculate DIC for both models.

Post. Pred Checks	Observed Data	No regional effect	Regional effect	p-value obs-sim (no regional effect)	p-value obs-sim (regional effect)
Mean					
CV					
SSQ					

References

Gelman, A. and J. Hill, 2009. *Data analysis using regression and multilievel / hierarchical modeling.* Cambridge University Press, Cambridge, UK.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin. 1995. *Bayesian data analysis.* Chapman and Hall, London.