Model Checking

Bayesian Models for Ecologists

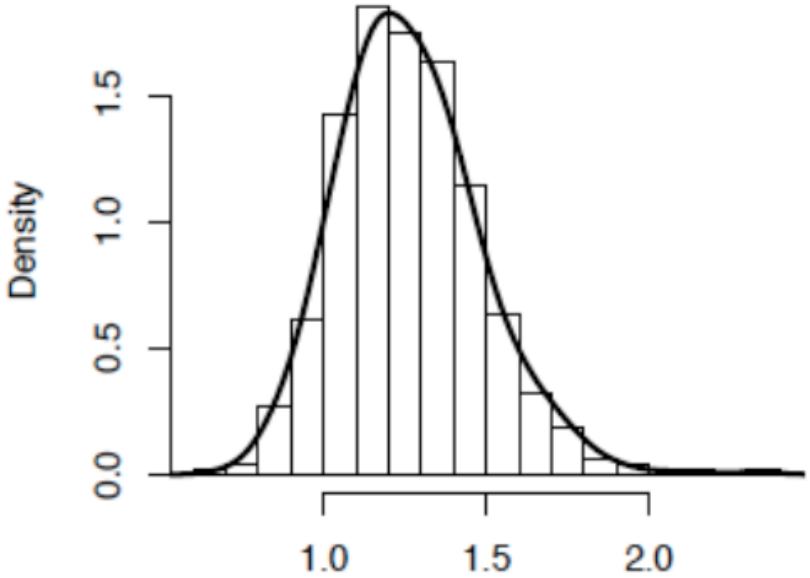
Alison Ketz (slides adapted from Mevin Hooten and Tom Hobbs)

June 03, 2024

DEB 2042028

What are the first questions you should ask after fitting a model?

- Does this model make sense? (sniff test)
- *Are the results (i.e. predictions) of the model consistent with the data?*
- Does the model adequately represent the process?
- Have you made the right choices for distributions?
- Does the model represent inherent stochasticity and uncertainty?



What is model checking?

Model-based inference depends on whether your model could give rise to the data.

- *Model checking* is the process of evaluating whether this assumption is true.
- We use predictions of new data for model checking.
- Bayesian prediction uses the joint distribution of new data and parameters given observed data

Marginal distributions - recap

The joint PDF $[a, b]$ is the density of continuous random variables *a and b* together.

If we want the density of only one variable at a time:

- $[a] = \int [a, b] db$ is the marginal probability of a.
- $[b] = \int [a, b] da$ is the marginal probability of b.
- This idea applies to any number of jointly distributed random variables: We integrate out all but one.

Posterior predictive checks

- Posterior predictive checks help us assess how different our *predictions* of new "data", y^{new}, are from our observed data $\mathbf{y}$.
- The *posterior predictive distribution* of new, unobserved data is

$$[y^{\text{new}} \mid \mathbf{y}] = \underbrace{\int [y^{\text{new}} \mid \boldsymbol{\theta}][\boldsymbol{\theta} \mid \mathbf{y}] d\boldsymbol{\theta}}_{\textbf{Posterior Predictive Distribution}}$$

- This is a marginal distribution because we are integrating out $\boldsymbol{\theta}$.

Consider this model

$$
\begin{aligned}
y_i &\sim \text{normal}(\mu_i, \sigma^2) \\
\mu_i &= g(\theta_1, \theta_2, \theta_3, \mathbf{x}_i) \\
\boldsymbol{\theta} &\sim [\boldsymbol{\theta}] \\
\sigma^2 &\sim [\sigma^2]
\end{aligned}
$$

with PPD

$$
[y^{\text{new}} \mid \mathbf{y}] = \int \int [y^{\text{new}} \mid \boldsymbol{\theta}, \sigma^2][\boldsymbol{\theta}, \sigma^2 \mid \mathbf{y}] d\boldsymbol{\theta} d\sigma^2
$$

Obtain Posterior Predictive Realizations

- At every MCMC iteration ($k = 1, \ldots, K$), make a draw of new data, $y^{\text{new}(k)}$, based on each value from a set of covariates $\mathbf{x}$ at the present values of the parameters $\boldsymbol{\theta}^{(k)}$ such that

$$[y^{new} \mid g(\theta^{(k)}, x), \sigma^{2(k)}]$$

- We can use these draws to summarize the posterior predictive distribution

$$\hat{y}^{\text{new}} = \mathrm{E}(y^{\text{new}} | \mathbf{y}) \approx \frac{\sum_{k=1}^{K} y^{\text{new}(k)}}{K}$$

- Covariates $\mathbf{x}_i$ for $i = 1, \ldots, n$ are our data set, we obtain $\mathbf{y}^{\text{new}} = (y_1^{\text{new}}, \ldots, y_n^{\text{new}})'$

$$g(b_0, b_1, x_i) = b_0 + b_1 x_i$$

$$[b_0, b_1, \tau \mid \mathbf{y}] \propto \prod_{i=1}^{n} \text{normal}\left(y_i \mid g(b_0, b_1, x_i)_i, \tau\right) \times$$

$$\text{normal}\left(b_0 \mid 0.0001\right) \text{normal}(b_1 \mid 0, .0001) \text{gamma}\left(\tau \mid .01 .01\right)$$

```
model{
b0 ~ dnorm(0,.0001)
b1 ~ dnorm(0,.0001)
tau ~ dgamma(.01,.01)
sigma<-1/sqrt(tau)
for(i in 1:length(y)){
  mu[i] <- b0 + b1*x[i]
  y[i] ~ dnorm(mu[i],tau)
  #posterior predictive distribution of y.new[i]
  y.new[i] ~ dnorm(mu[i],tau)
  }
}
```

Posterior Predictive Checking

- $T(\mathbf{y}, \boldsymbol{\theta})$ is a test statistic calculated from the observed data.
- $T(\mathbf{y}^{new}, \boldsymbol{\theta})$ is the corresponding statistic based on the posterior predictive realizations.
- Calculate:

$$p_b = \Pr(\, T(\mathbf{y}^{new}, \boldsymbol{\theta}) \geq T(\mathbf{y}, \theta) \mid \mathbf{y})$$

- If p_b is very large or very small (i.e., close to 1 or 0), **it indicates lack of fit**.

Candidates for test statistics

- mean
- variance
- coefficient of variation
- quantiles
- maximum, minimum
- chi-square
- deviance

Tick example

We seek to know the average number of ticks on sheep.

- We round up 60 sheep and count ticks on each one.
- Does a Poisson distribution fit the distribution of the data?

$$[\lambda \mid \mathbf{y}] \propto \prod_{i=1}^{60} \text{Poisson}(y_i \mid \lambda)[\lambda]$$

- For each value of λ in the MCMC chain, we generate a new data set, $\mathbf{y}^{\text{new}(k)}$, by sampling

$$y_i^{\text{new}(k)} \sim \text{Poisson}(\lambda^{(k)})$$

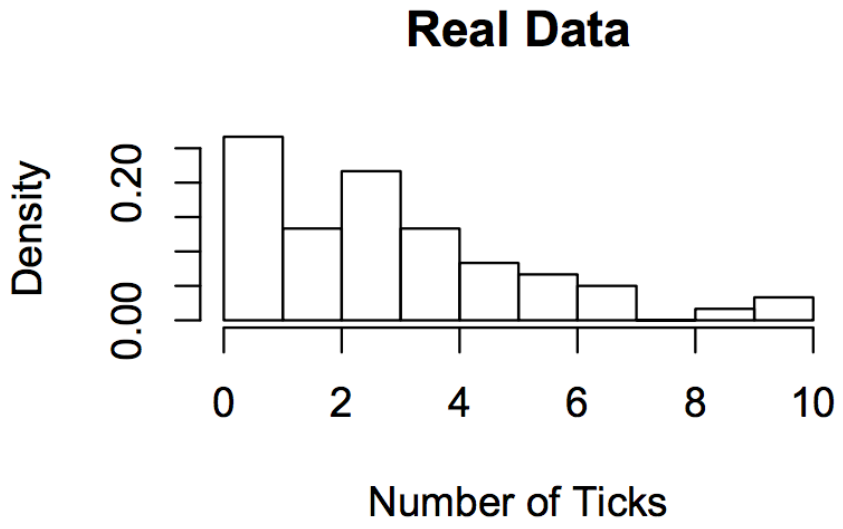
for $i = 1, \ldots, n$.

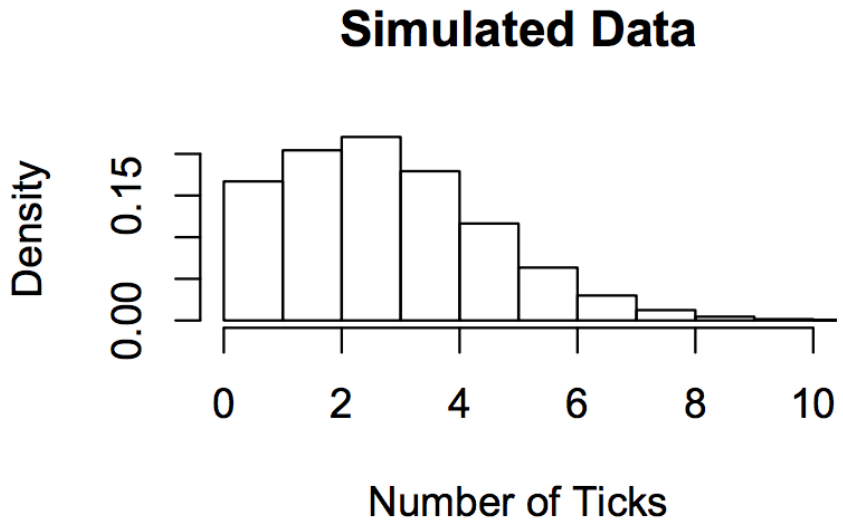
Side note: What assumptions are we making with this model?

Code

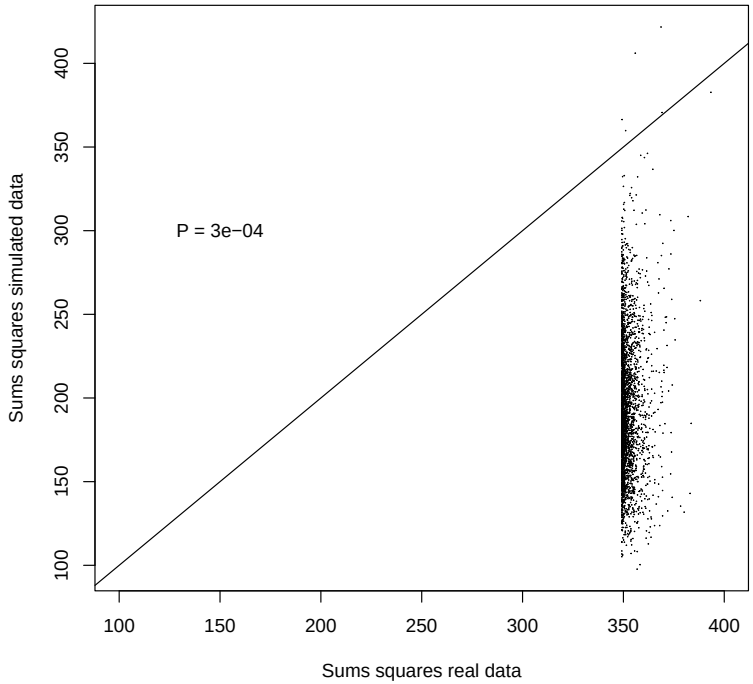
```
model{
lambda ~ dgamma(0.001,0.001)
for(i in 1:60){
     y[i] ~ dpois(lambda)
     y.new[i] ~ dpois(lambda)  #simulate a new data set of 60 points
}
cv.y <- sd(y[ ])/mean(y[ ])
cv.y.new <- sd(y.new[])/mean(y.new[ ])
pvalue.cv <- step(cv.y.new-cv.y)  # find Bayesian P value--the mean of
many 0's and 1's returned by the step function, one for each iteration in
the chain. The function step(z) returns a 1 if z > 0, returns 0
otherwise.
mean.y <-mean(y[])
mean.y.new <-mean(y.new[])
pvalue.mean <-step(mean.y.new - mean.y)
for(j in 1:60){
     sq[j] <- (y[j]-lambda)^2
     sq.new[j] <- (y.new[j]-lambda)^2
}
fit <- sum(sq[])
fit.new <- sum(sq.new[])
pvalue.fit <- step(fit.new-fit)
} #end of model
```

Simple Model





Posterior Predictive Check



- p-value for CV $= 0.0013$
- p-value for mean $= 0.51$
- *Values close to 0 or 1 indicate lack of fit.*

How could you modify this model to allow "extra" variance?

- Draw a Bayesian network and write out the posterior and joint distributions.

Hierarchical model

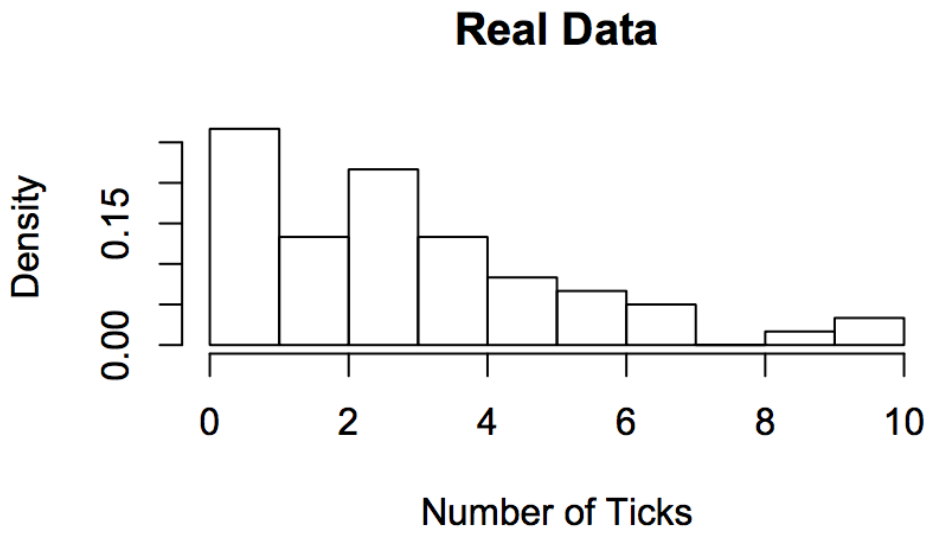
$$[a, b, \lambda \mid \mathbf{y}] \propto \prod_{i=1}^{60} [y_i \mid \lambda_i][\lambda_i \mid a, b][a][b]$$

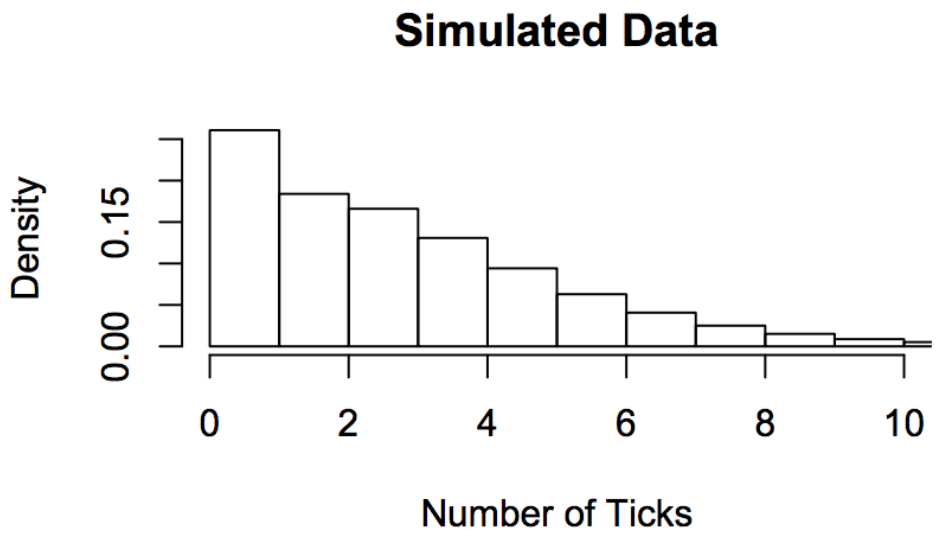
```
model{
a~ dgamma(.001,.001)
b~ dgamma(.001,.001)
for(i in 1:60){
    lambda[i] ~ dgamma(a,b)
    y[i] ~ dpois(lambda[i])
    y.sim[i] ~ dpois(lambda[i])
}
cv.y <- sd(y[ ])/mean(y[ ])
cv.y.sim <- sd(y.sim[])/mean(y.sim[ ])
pvalue.cv <- step(cv.y.sim-cv.y)  # find Bayesian P
 value--the mean of many 0's and 1's returned by
 the step function, one for each step in the chain
mean.y <-mean(y[])
mean.y.sim <-mean(y.sim[])
pvalue.mean <-step(mean.y.sim - mean.y)
for(j in 1:60){
    sq[j] <- (y[j]-lambda[j])^2
    sq.new[j] <- (y.sim[j]-lambda[j])^2
 }
 fit <- sum(sq[])
 fit.new <- sum(sq.new[])
 pvalue.fit <- step(fit.new-fit)
} #end of model
```

Hierarchical model

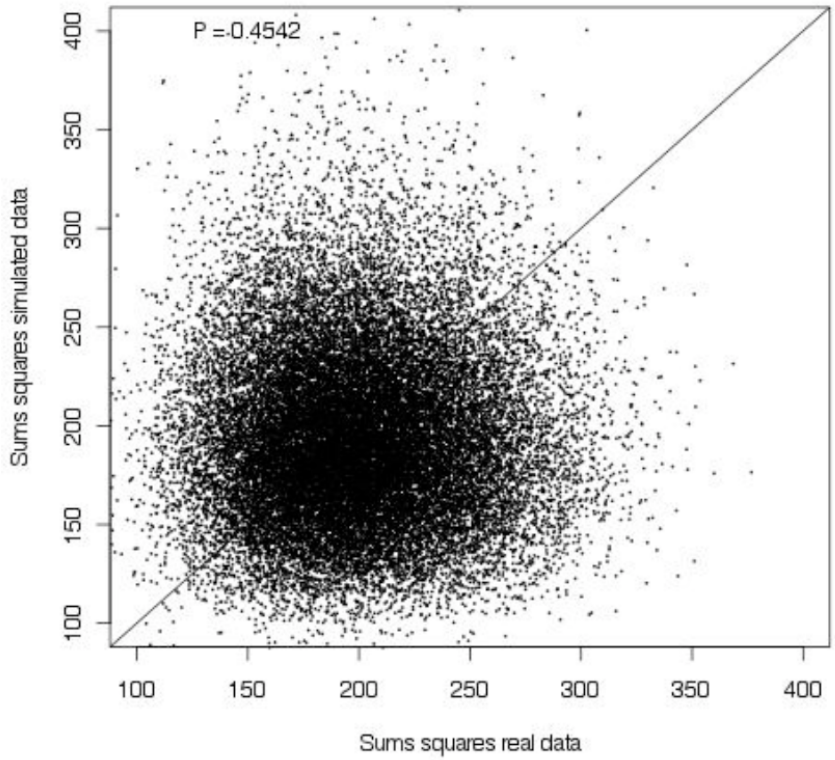
$$[a, b, \boldsymbol{\lambda}|\mathbf{y}] \propto \prod_{i=1}^{60} [y_i|\lambda_i]\,[\lambda_i|a, b]\,[a]\,[b]$$

Include pvalue.fit in variable names list for `coda.samples` or `jags.samples`. Report the mean of pvalue.fit


Real Data
Density
0.00
0.15
0
2
4
6
8
10
Number of Ticks


Simulated Data
Density
0.00
0.15
0
2
4
6
8
10
Number of Ticks

Posterior Predictive Checks



- p-value for CV $= 0.45$
- p-value for mean $= 0.5$

Reporting your posterior predictive checks

Posterior predictive checks revealed little evidence of lack of fit between model estimates and data for five data sets (Table 4). Bayesian p-values were between 0.12 and 0.88 for 14 out of 15 test statistics for each of the three models. There was some evidence of poor fit of data simulated from the model to observations of the mean of yearling serology for all three models. (Hobbs et al. 2015)

TABLE 4. Bayesian P values for lack of fit between data simulated from posterior predictive distributions and observations for five data sets.

Model and data set	Discrepancy	Mean	SD
Frequency dependent			
Total population size	0.51	0.5	0.51
Proportion juvenile	0.57	0.64	0.93
Juvenile serology	0.8	0.75	0.81
Yearling serology	0.13	0.058	0.19
Adult serology	0.59	0.69	0.54
Density dependent			
Total population size	0.51	0.5	0.48
Proportion juvenile	0.57	0.69	0.95
Juvenile serology	0.88	0.84	0.88
Yearling serology	0.19	0.084	0.28
Adult serology	0.59	0.64	0.56
Combined			
Total population size	0.51	0.5	0.51
Proportion juvenile	0.57	0.64	0.93
Juvenile serology	0.8	0.75	0.81
Yearling serology	0.13	0.058	0.19
Adult serology	0.59	0.69	0.54

Notes: Bayesian P values, P_B, are defined as the probability that the test statistic calculated from simulated data is more extreme than the test statistic calculated from observed data. Lack of fit is indicated by values near 1 or 0. Test statistics were the mean of observations and simulated data, the standard deviation, and the discrepancy, calculated as $\sum_{i=1}^{n}(y_i - \mu_i)^2$ where y_i is an observation, μ_i is the model prediction of the observation, and n is the number of observations in the data set.

Additional sources

- A. Gelman and J. Hill. Data Analysis Using Regression and Multilievel / Hierarchical Modeling. Cambridge University Press, Cambridge, UK, 2009 Chapter 8
- P. B. Conn, D. S. Johnson, P. J. Williams, S. R. Melin, and M. B. Hooten. A guide to Bayesian model checking for ecologists. Ecological Monographs, 88(4):526–542, 2018.