ST 740: Posterior Predictive Checking and Sensitivity Analysis

Alyson Wilson

Department of Statistics
North Carolina State University

November 18, 2013

Key Questions

1. How can I tell if my model is providing an adequate fit to the data? (*goodness of fit*)
2. What aspects of reality are not captured by my model? Are these important given the use I would like to make of the model? (*model checking*)
3. How can I tell if any of the modeling choices I have made are having an undue impact on my results? (*sensitivity analysis*)
4. Which model (or models) should I ultimately choose for the final presentation of my results? (*model selection*)

Model Checking

All models are wrong, or at best an approximation.

Model building involves three components: estimation, criticism, revision.

- Estimation: Estimate (or infer) parameters conditioned on the truth of the model.
- Criticism "involves a confrontation of [the model] with available data (old as well as newly acquired) and asks whther [the model] is consonant with [it]" (Box 1980).
- Revision decides how to change the model based on the criticisms (and available time/computing/knowledge resources).

Model Checking

We've spent a lot of the class discussing how to do estimation. Today we are focusing on *criticism*.

- Is my model good enough?
- In what ways is it not good enough?

Notice that this is different from *model comparison*, as we are only thinking about one model right now.

Conditional Predictive Ordinate

$$\begin{aligned} f(y_i \mid \mathbf{y}_{-i}) &= \int f(y_i \mid \theta, \mathbf{y}_{-i}) \pi(\theta \mid \mathbf{y}_{-i}) d\theta \\ &= \int f(y_i \mid \theta) \pi(\theta \mid \mathbf{y}_{-i}) d\theta \end{aligned}$$

Define the *conditional predictive ordinate* as the conditional predictive distribution evaluated at the observed value y_i. Very low CPO values suggest that y_i is an outlier with regard to the model.

Notice that we can use the weighted bootstrap to change a sample from $\pi(\theta \mid \mathbf{y})$ to one from $\pi(\theta \mid \mathbf{y}_{-i})$.

Posterior Predictive Checking

Gelman, Meng, Stern 1996

General idea: Is the model consistent with the data? If the model fits, then replicated data generated under the model should look similar to observed data. More specifically, the observed data should look plausible under the posterior predictive distribution.

Basic technique: Draw simulated values of replicated data from the posterior predictive distribution and compare these samples to the observed data. Any systematic differences between the simulations and the data indicate potential failings of the model.

Posterior Predictive Checking Examples

Suppose that we have a model

$$\begin{aligned} Y_i &\sim \text{Bernoulli}(\theta) \\ \theta &\sim \text{Uniform}(0,1) = \text{Beta}(1,1) \end{aligned}$$

This implies that the posterior distribution for θ is $\pi(\theta \mid \mathbf{y}) \propto \theta^s(1-\theta)^{n-s}$, where s is the number of 1s observed and $n-s$ is the number of 0s. This is, of course, $\theta \sim \text{Beta}(s+1, n-s+1)$.

Suppose that our data are, in order, 1, 1, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0.

The observed autocorrelation (i.e., the long strings of 0s and 1s) is suggestive that the model is flawed. Can we quantify the evidence?

Posterior Predictive Checking: Bernoulli Trials

Consider the statistic $T(\mathbf{y}, \theta) =$ number of switches between 0 and 1 in the sequence. The observed value is $T(\mathbf{y}) = 3$. Notice that this particular statistic doesn't depend on θ.

Now we need to simulate some data.

1. First, we need draws from the posterior distribution, with $s = 7$ and $n - s = 13$. Call these $\theta^{(j)}$ for $j = 1, \ldots, 1000$.
2. Next, we need to generate "replicate" data sets, which are 20 observations from a Bernoulli($\theta^{(j)}$). `rbinom(20,1,theta)`
3. Next, we need to calculate how many switches there are between 0 and 1 in the replicate data set. This is $T(\mathbf{y}_{rep})^{(j)}$. `sum(abs(diff(rbinom(20,1,theta))))`
4. Next, we approximate $P(T(\mathbf{y}_{rep}) \geq T(\mathbf{y}) \,|\, \mathbf{y})$ by counting what proportion of the time $T(\mathbf{y}_{rep}) \geq T(\mathbf{y}) = 3$ in our simulation.

Bayesian P-value

This is an example of a *Bayesian p-value*, where

$$\begin{aligned} p_B &= \mathbf{P}(T(\mathbf{y}^{rep}, \theta) \geq T(\mathbf{y}, \theta) \,|\, \mathbf{y}) \\ &= \int \int I_{T(\mathbf{y}^{rep},\theta) \geq T(\mathbf{y},\theta)} f(\mathbf{y}^{rep} \,|\, \theta) \pi(\theta \,|\, \mathbf{y}) d\mathbf{y}^{rep} d\theta \end{aligned}$$

Define a classical p-value based on a test statistic $T(y)$ as

$$p_c = \mathbf{P}(T(\mathbf{y}^{rep}) \geq T(\mathbf{y}) \,|\, \theta)$$

For a given test statistic, we can think about the Bayesian p-value as the classical p-value averaged over the posterior distribution of θ.

$$p_B = \int \mathbf{P}(T(\mathbf{y}^{rep}) \geq T(\mathbf{y}) \,|\, \theta) \pi(\theta \,|\, \mathbf{y}) d\theta$$

Posterior Predictive Checking

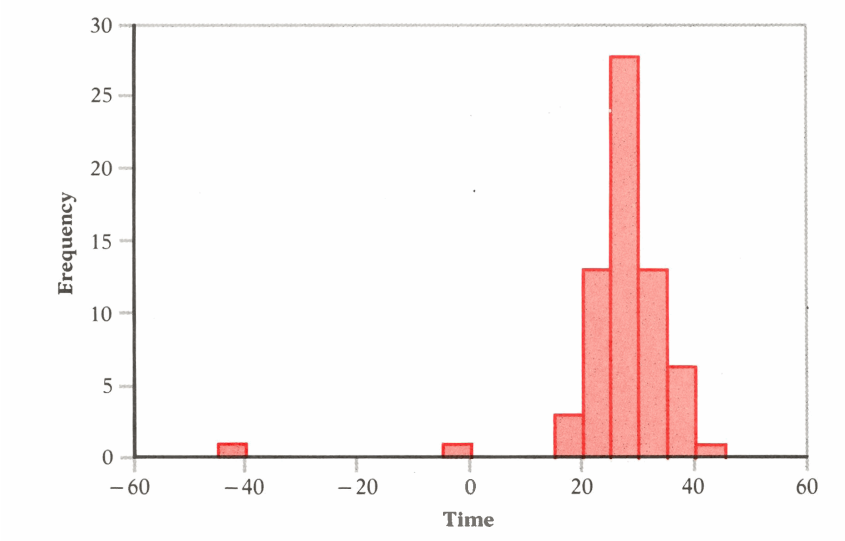
Newcomb's Speed of Light Data

Simon Newcomb set up an experiment in 1882 to measure the speed of light. Newcomb measured the amount of time required for light to travel a distance of 7442 meters (the distance from his laboratory on the Potomac River to the Washington Monument and back). The data are recorded as deviations from 24,800 nanoseconds. (Based on the currently accepted speed of light, the "true" value is 33.)

Newcomb's Speed of Light Data

TABLE 1.1 Newcomb's measurements of the passage time of light

28	22	36	26	28	28
26	24	32	30	27	24
33	21	36	32	31	25
24	25	28	36	27	32
34	30	25	26	26	25
−44	23	21	30	33	29
27	29	28	22	26	27
16	31	29	36	32	28
40	19	37	23	32	29
−2	24	25	27	24	16
29	20	28	27	39	23



Newcomb's Speed of Light Data

Suppose that we model this data as coming from a normal distribution with a noninformative prior $\pi(\mu, \sigma^2) \propto \frac{1}{\sigma^2}$.

The lowest of these measurements looks like an outlier compared with the rest of the data. Could the extreme measurement have come from a normal distribution?

Normal Model

Assume that we have $Y_i \,|\, \mu, \sigma^2 \sim \text{Normal}(\mu, \sigma^2)$. Both μ and σ^2 are unknown.

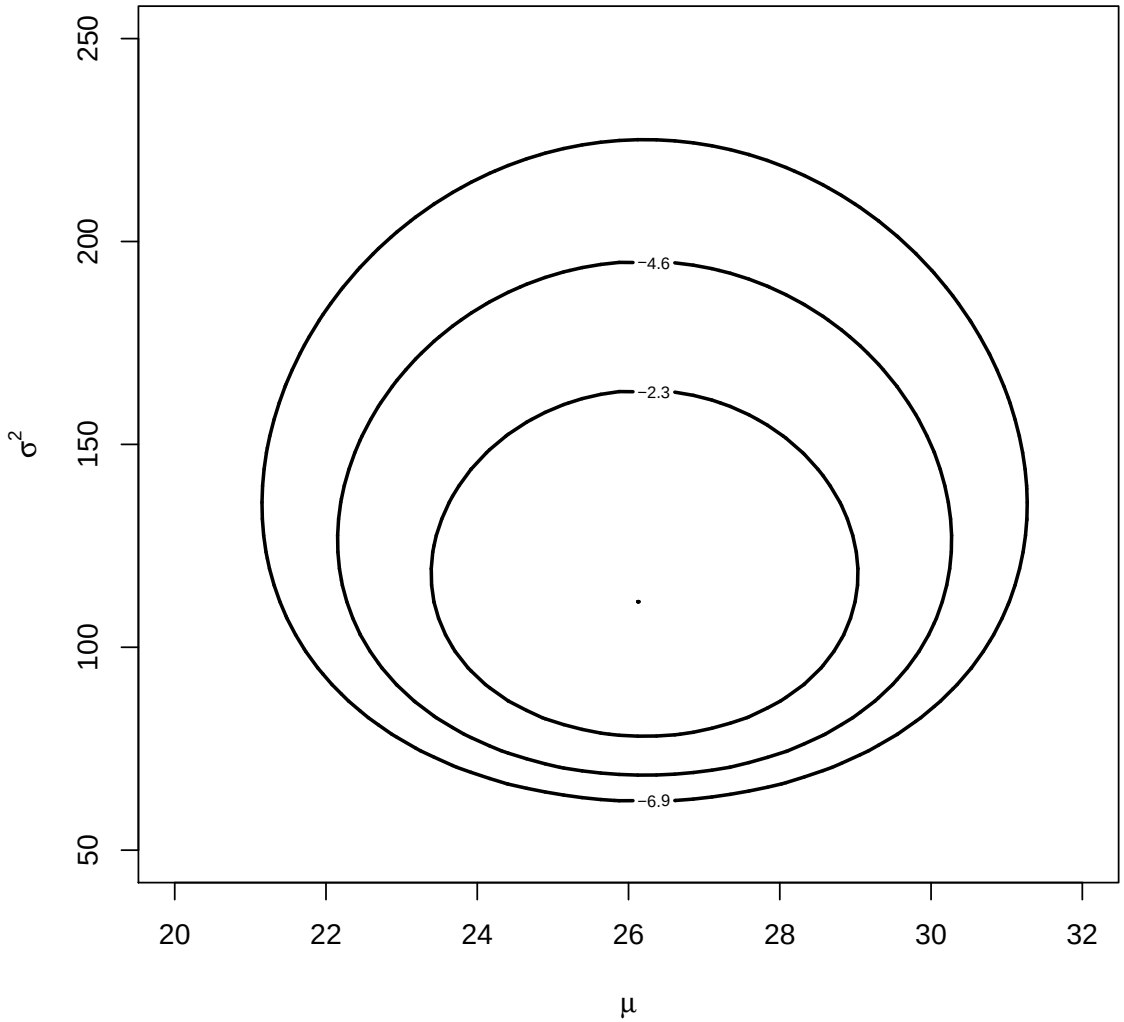
The non-informative prior for (μ, σ^2) assuming prior independence is

$$\pi(\mu, \sigma^2) \propto 1 \times \frac{1}{\sigma^2}$$

The joint posterior distribution is

$$\begin{aligned} \pi(\mu, \sigma^2 \,|\, \mathbf{y}) &\propto \pi(\mu, \sigma^2) f(\mathbf{y} \,|\, \mu, \sigma^2) \\ &\propto \sigma^{-n-2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2\right) \end{aligned}$$

Contour Plot of Posterior Distribution



Random Sample from Posterior Distribution

We can get a random sample from the posterior distribution in two ways.

1.
   - Sample $(\sigma^2)^{(i)}$ from $\left[\sigma^2 \,|\, \mathbf{y}\right] \sim \text{InverseGamma}(\frac{n-1}{2}, \frac{(n-1)s^2}{2})$.
   - Then sample $\mu^{(i)}$ from $\left[\mu \,|\, \sigma^2, \mathbf{y}\right] \sim \text{Normal}(\bar{y}, (\sigma^2)^{(i)}/n)$.
2.
   - Sample $\mu^{(i)}$ from $[\mu \,|\, \mathbf{y}] \sim \text{t}(n-1, \bar{y}, \frac{s^2}{n})$.
   - Then sample $(\sigma^2)^{(i)}$ from $\left[\sigma^2 \,|\, \mu, \mathbf{y}\right] \sim \text{InverseGamma}(\frac{n}{2}, \frac{(n-1)s^2 + n(\bar{y}-\mu^{(i)})^2}{2})$.

Posterior Predictive Checking

Remember that our goal is to "Draw simulated values of replicated data from the posterior predictive distribution and compare these samples to the observed data."

There are typically two strategies used to make the comparisons: graphical and numerical.

For the graphical checks, we generate replicated data sets and compare plots of the data, residuals, parameters, ...

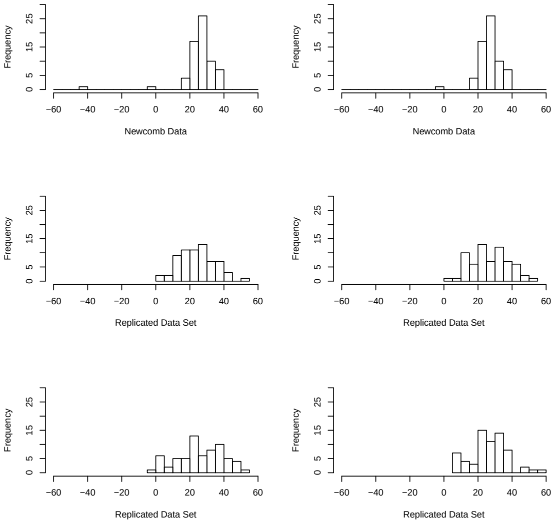
For the numerical checks, we choose a test quantity (sometimes called a *discrepancy*) and compare its values for the replicated data with the value from the observed data.

Posterior Predictive Checking

We will use y^{rep} to denote *replicated data*. Think of y^{rep} as the data we *would* see tomorrow if the experiment that produced y today were replicated with the same model (including all of the hyperparameters) and the same value of θ that produced the observed data.

$$f(y^{rep} \mid \mathbf{y}) = \int f(y^{rep} \mid \theta)\pi(\theta \mid y)d\theta$$

Replicated Data Sets



Newcomb's Speed of Light Data

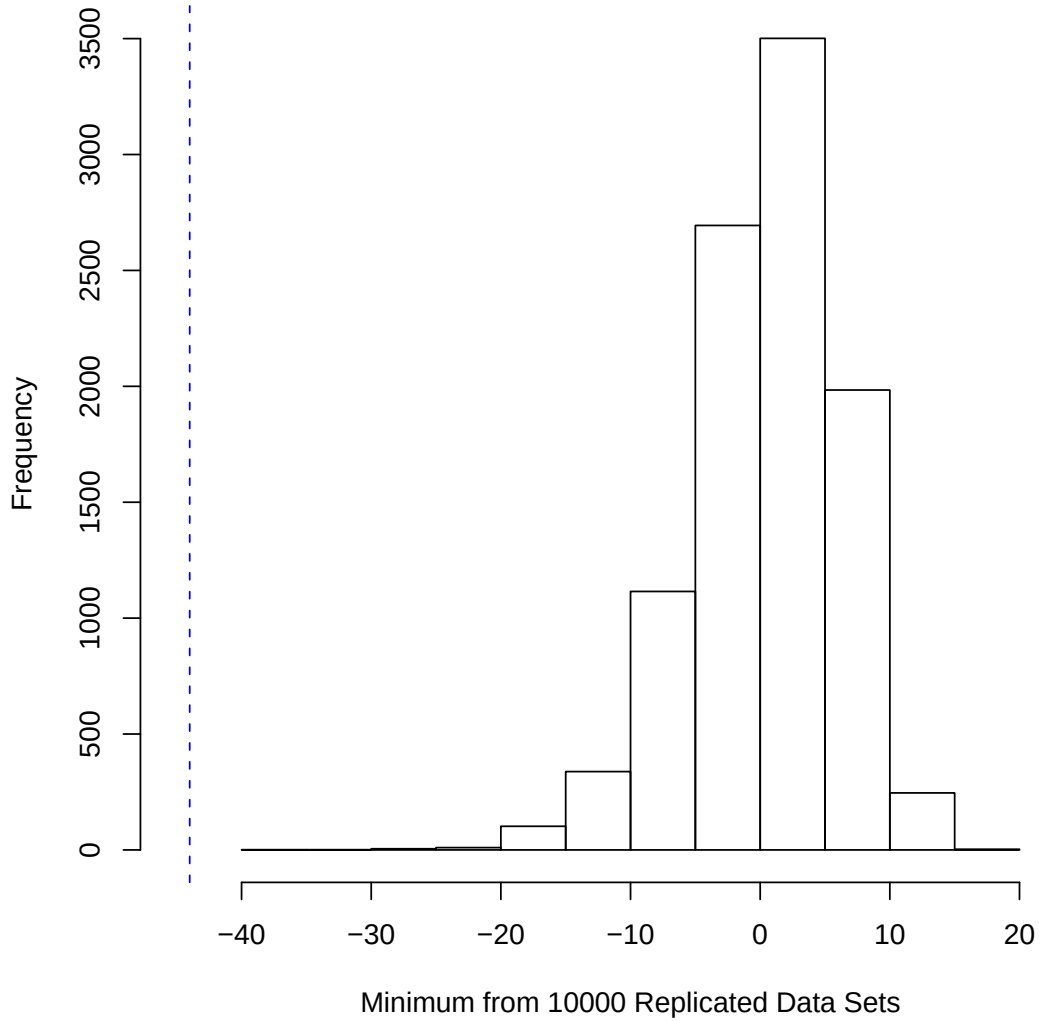
We measure the discrepancy between the model and the data by defining *test quantities* that capture some aspect of the data that we wish to check. A *test quantity* (or *discrepancy*), $T(y, \theta)$, is a scalar summary of parameters and data that is used as a standard when comparing data to predictive simulations.

Test quantities play the role in Bayesian model checking that test statistics play in classical testing.

We use the notation $T(y)$ for a *test statistic*, which is a test quantity that depends only on the data; in the Bayesian context, we can generalize test statistics to allow dependence on the model parameters under their posterior distribution.

Newcomb's Speed of Light Data

Consider the test statistic $T(\mathbf{y}, \theta) = \min(y_i)$.



Newcomb's Speed of Light Data

We've discovered the model is inadequate for some purposes (like estimating extreme tail values).

Suppose that we want instead to examine whether there is asymmetry around the center of the distribution.

Newcomb's Speed of Light Data

Consider

$$T(y, \mu, \sigma^2) = |y_{61} - \mu| - |y_6 - \mu|$$

The 61st and 6th order statistics are approximately the 90% and 10% points of the distribution.

Posterior Predictive Checking

In practice, if we have M simulations from the posterior distribution of θ, we just draw one $\mathbf{y}^{rep}$ for each simulated θ and we calculate $T(\mathbf{y}^{rep}, \theta)$. After calculating this for each draw, we evaluate p_B. We are looking for extreme values (say below 0.05 or above 0.95) that would indicate lack of fit.

As an omnibus goodness-of-fit measure, consider the χ^2 discrepancy:

$$T(\mathbf{y}, \theta) = \sum_i \frac{(y_i - E(Y_i \mid \theta))^2}{\text{Var}(Y_i \mid \theta)}$$

Another omnibus goodness-of-fit measure that is sometimes used is the *deviance*:

$$T(\mathbf{y}, \theta) = -2 \log(f(\mathbf{y} \mid \theta))$$

Sensitivity Analysis

It is typcially the case that more than one reasonable model can provide an adequate fit to the data in a scientific problem. The basic question of a *sensitivity analysis* is: how much do posterior inferences change when other reasonable models are used in place of the present model?

Operationally, we make *reasonable* modifications to our assumptions, recompute the posterior quantities of interest, and see whether they have changed in a way that has practical impact on interpretations or decisions.

If so, we may wish to communicate the sensitivity, think more carefully about it, collect more data, or all of these.

Sensitivity Analysis

From a Bayesian perspective, when we're talking about sensitivity analysis, we're almost always thinking about prior distributions. Changes to the likelihood, covariates, etc. are usually considered as part of *model selection*.

- Example Reese et al. (2001, Bowhead Whales). Constructed a 2^{16-10} fractional factorial design (64 runs), no main effects confounded with other main effects or two-way interactions. Used reasonable "low" and "high" values of hyperparameters: in this case, took the current value, divided by 2 and multiplied by 2.
- Or, consider other functional forms for the prior distribution.

Sensitivity Analysis

How do you do this in practice?

- Run lots more MCMC.
- Use importance sampling or a weighted bootstrap.

Weighted Bootstrap for Sensitivity Analysis

We have a sample from the posterior distribution $\pi(\theta \mid y)$. Let's use a weighted bootstrap to convert this to a sample from a posterior distribution calculated with a different prior.

Recall the stategy for the weighted bootstrap.

- We have a sample from a "proposal density" (the current posterior).
- We want a sample from a "density of interest" (a posterior calculated using a new prior).

Weighted Bootstrap

The algorithm to follow is

- Simulate $\theta^{(i)}$ from $\pi_{old}(\theta \mid \mathbf{y})$, $i = 1, \ldots, m$
- Calculate $w_i = \frac{\pi_{new}(\theta^{(i)} \mid \mathbf{y})}{\pi_{old}(\theta^{(i)} \mid \mathbf{y})}$
- Normalize the weights, $q_i = \frac{w_i}{\sum_{i=1}^{m} w_i}$
- Resample (with replacement) the $\theta^{(i)}$ with weights q_i

Be sure to check your weights and make sure that you don't have one or two large ones and the rest tiny. This depends on how well $\pi_{old}(\theta \mid \mathbf{y})$ approximates $\pi_{new}(\theta \mid \mathbf{y})$.

External Validation

Do inferences from the model make sense?

Use the model to make predictions about future data. Posterior means should be correct on average, 50% intervals should contain the true values half of the time, etc.