CHAPTER 6

Model checking and improvement

6.1 The place of model checking in applied Bayesian statistics

Once we have accomplished the first two steps of a Bayesian analysis—constructing a probability model and computing (typically using simulation) the posterior distribution of all estimands—we should not ignore the relatively easy step of assessing the fit of the model to the data and to our substantive knowledge. It is difficult to include in a probability distribution all of one's knowledge about a problem, and so it is wise to investigate what aspects of reality are *not* captured by the model.

Checking the model is crucial to statistical analysis. Bayesian prior-to-posterior inferences assume the whole structure of a probability model and can yield misleading inferences when the model is poor. A good Bayesian analysis, therefore, should include at least some check of the adequacy of the fit of the model to the data and the plausibility of the model for the purposes for which the model will be used. This is sometimes discussed as a problem of sensitivity to the prior distribution, but in practice the likelihood model is typically just as suspect; throughout, we use 'model' to encompass the sampling distribution, the prior distribution, any hierarchical structure, and issues such as which explanatory variables have been included in a regression.

Sensitivity analysis and model improvement

It is typically the case that more than one reasonable probability 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 probability models are used in place of the present model? Other reasonable models may differ substantially from the present model in the prior specification, the sampling distribution, or in what information is included (for example, predictor variables in a regression). It is possible that the present model provides an adequate fit to the data, but that posterior inferences differ under plausible alternative models.

In theory, both model checking and sensitivity analysis can be incorporated into the usual prior-to-posterior analysis. Under this perspective, model checking is done by setting up a comprehensive joint distribution, such that any data that might be observed are plausible outcomes under the joint distribution. That is, this joint distribution is a mixture of all possible 'true' models or realities, incorporating all known substantive information. The prior dis-

tribution in such a case incorporates prior beliefs about the likelihood of the competing realities and about the parameters of the constituent models. The posterior distribution of such an *exhaustive* probability model automatically incorporates all 'sensitivity analysis' but is still predicated on the truth of some member of the larger class of models.

In practice, however, setting up such a super-model to include all possibilities and all substantive knowledge is both conceptually impossible and computationally infeasible in all but the simplest problems. It is thus necessary for us to examine our models in other ways to see how they fail to fit reality and how sensitive the resulting posterior distributions are to arbitrary specifications.

Judging model flaws by their practical implications

We do not like to ask, 'Is our model true or false?', since probability models in most data analyses will not be perfectly true. Even the coin tosses and die rolls ubiquitous in probability theory texts are not truly exchangeable in reality. The more relevant question is, 'Do the model's deficiencies have a noticeable effect on the substantive inferences?'

In the examples of Chapter 5, the beta population distribution for the tumor rates and the normal distribution for the eight school effects are both chosen partly for convenience. In these examples, making convenient distributional assumptions turns out not to matter, in terms of the impact on the inferences of most interest. How to judge when assumptions of convenience can be made safely is a central task of Bayesian sensitivity analysis. Failures in the model lead to practical problems by creating clearly false inferences about estimands of interest.

6.2 Do the inferences from the model make sense?

In any applied problem, there will be knowledge that is not included formally in either the prior distribution or the likelihood, for reasons of convenience or objectivity. If the additional information suggests that posterior inferences of interest are false, then this suggests a potential for creating a more accurate probability model for the parameters and data collection process. We illustrate with an example of a hierarchical regression model.

Example. Evaluating election predictions by comparing to substantive political knowledge

Figure 6.1 displays a forecast, made in early October, 1992, of the probability that Bill Clinton would win each state in the November, 1992, U.S. Presidential election. The estimates are posterior probabilities based on a hierarchical linear regression model. For each state, the height of the shaded part of the box represents the estimated probability that Clinton would win the state. Even before the election occurred, the forecasts for some of the states looked wrong; for example, from state polls, Clinton was known in October to be much weaker in Texas and Florida than shown in the map. This does not mean that the forecast

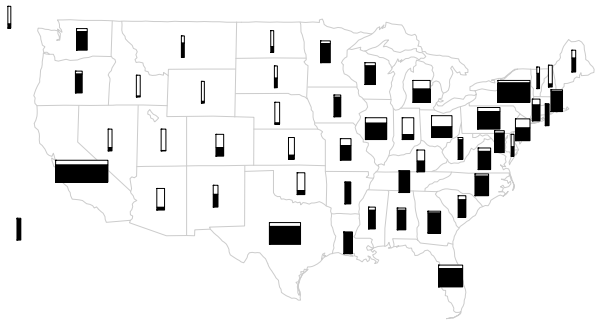

Figure 6.1 *Summary of a forecast of the 1992 U.S. Presidential election performed one month before the election. For each state, the proportion of the box that is shaded represents the estimated probability of Clinton winning the state; the width of the box is proportional to the number of electoral votes for the state.*

is useless, but it is good to know where the weak points are. Certainly, after the election, we can do an even better job of criticizing the model and understanding its weaknesses. We return to this election forecasting example in Section 15.2 as an example of a hierarchical linear model.

More formally, we can check a model by *external validation* using the model to make predictions about future data, and then collecting those data and comparing to their predictions. Posterior means should be correct on average, 50% intervals should contain the true values half the time, and so forth. We used external validation to check the empirical probability estimates in the record-linkage example in Section 1.7, and we apply the idea again to check a toxicology model in Section 20.3. In the latter example, the external validation (see Figure 20.10 on page 514) reveals a generally reasonable fit but with some notable discrepancies between predictions and external data.

6.3 Is the model consistent with data? Posterior predictive checking

If the model fits, then replicated data generated under the model should look similar to observed data. To put it another way, the observed data should look plausible under the posterior predictive distribution. This is really a self-consistency check: an observed discrepancy can be due to model misfit or chance.

Our basic technique for checking the fit of a model to data is to draw simulated values from the posterior predictive distribution of replicated data and compare these samples to the observed data. Any systematic differences between the simulations and the data indicate potential failings of the model. We introduce posterior predictive checking with a simple example of an obviously poorly fitting model, and then in the rest of this section we lay out

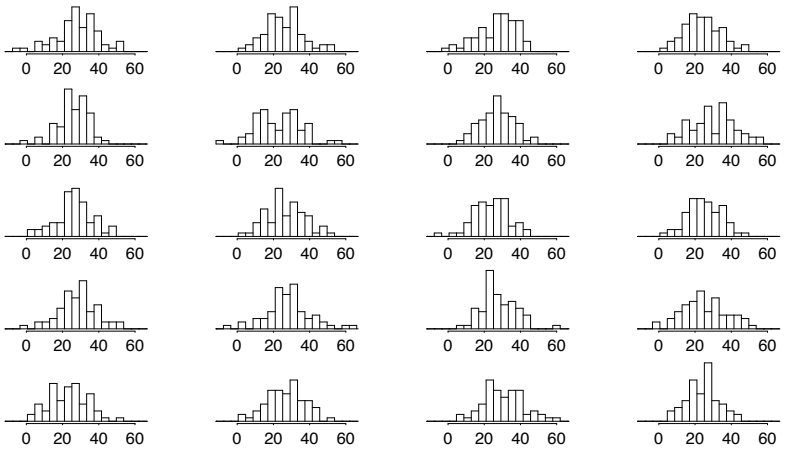

Figure 6.2 *Twenty replications, y^{rep}, of the speed of light data from the posterior predictive distribution, $p(y^{\text{rep}}|y)$; compare to observed data, y, in Figure 3.1. Each histogram displays the result of drawing 66 independent values $\tilde{y}_i$ from a common normal distribution with mean and variance (μ, σ^2) drawn from the posterior distribution, $p(\mu, \sigma^2|y)$, under the normal model.*

the key choices involved in posterior predictive checking. Sections 6.4 and 6.5 discuss graphical and numerical predictive checks in more detail.

Example. Comparing Newcomb's speed of light measurements to the posterior predictive distribution

Simon Newcomb's 66 measurements on the speed of light are presented in Section 3.2. In the absence of other information, in Section 3.2 we modeled the measurements as $\text{N}(\mu, \sigma^2)$, with a noninformative uniform prior distribution on $(\mu, \log \sigma)$. However, the lowest of Newcomb's measurements look like outliers compared to the rest of the data.

Could the extreme measurements have reasonably come from a normal distribution? We address this question by comparing the observed data to what we expect to be observed under our posterior distribution. Figure 6.2 displays twenty histograms, each of which represents a single draw from the posterior predictive distribution of the values in Newcomb's experiment, obtained by first drawing (μ, σ^2) from their joint posterior distribution, then drawing 66 values from a normal distribution with this mean and variance. All these histograms look quite different from the histogram of actual data in Figure 3.1 on page 78. One way to measure the discrepancy is to compare the smallest value in each hypothetical replicated dataset to Newcomb's smallest observation, -44. The histogram in Figure 6.3 shows the smallest observation in each of the 20 hypothetical replications; all are much larger than Newcomb's smallest observation, which is indicated by a vertical line on the graph. The normal model clearly does not capture the variation that Newcomb observed. A revised model might use an asymmetric contaminated normal distribution or a symmetric long-tailed distribution in place of the normal measurement model.

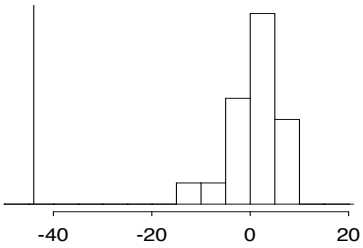

Figure 6.3 *Smallest observation of Newcomb's speed of light data (the vertical line at the left of the graph), compared to the smallest observations from each of the 20 posterior predictive simulated datasets displayed in Figure 6.2.*

Many other examples of posterior predictive checks appear throughout the book, including the educational testing example in Section 6.8, linear regressions example in Sections 14.3 and 15.2, and a hierarchical mixture model in Section 18.4.

For many problems, it is useful to examine graphical comparisons of summaries of the data to summaries from posterior predictive simulations, as in Figure 6.3. In cases with less blatant discrepancies than the outliers in the speed of light data, it is often also useful to measure the 'statistical significance' of the lack of fit, a notion we formalize here.

Notation for replications

Let y be the observed data and θ be the vector of parameters (including all the hyperparameters if the model is hierarchical). To avoid confusion with the observed data, y, we define y^{rep} as the *replicated* data that *could have been* observed, or, to think predictively, as the data we *would* see tomorrow if the experiment that produced y today were replicated with the same model and the same value of θ that produced the observed data.

We distinguish between y^{rep} and $\tilde{y}$, our general notation for predictive outcomes: $\tilde{y}$ is any future observable value or vector of observable quantities, whereas y^{rep} is specifically a replication just like y. For example, if the model has explanatory variables, x, they will be identical for y and y^{rep}, but $\tilde{y}$ may have its own explanatory variables, $\tilde{x}$.

We will work with the distribution of y^{rep} given the current state of knowledge, that is, with the posterior predictive distribution

$$p(y^{\text{rep}}|y) = \int p(y^{\text{rep}}|\theta)p(\theta|y)d\theta. \tag{6.1}$$

Test quantities

We measure the discrepancy between model and data by defining *test quantities*, the aspects of the data we wish to check. A test quantity, or *discrepancy*

measure, $T(y, \theta)$, 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 data; in the Bayesian context, we can generalize test statistics to allow dependence on the model parameters under their posterior distribution. This can be useful in directly summarizing discrepancies between model and data. We discuss options for graphical test quantities in Section 6.4 and numerical summaries in Section 6.5.

Tail-area probabilities

Lack of fit of the data with respect to the posterior predictive distribution can be measured by the tail-area probability, or p-value, of the test quantity, and computed using posterior simulations of (θ, y^{rep}). We define the p-value mathematically, first for the familiar classical test and then in the Bayesian context.

Classical p-values. The classical p-value for the test statistic $T(y)$ is

$$p_C = \Pr(T(y^{\text{rep}}) \geq T(y)|\theta) \tag{6.2}$$

where the probability is taken over the distribution of y^{rep} with θ fixed. (The distribution of y^{rep} given y and θ is the same as its distribution given θ alone.) Test statistics are classically derived in a variety of ways but generally represent a summary measure of discrepancy between the observed data and what would be expected under a model with a particular value of θ. This value may be a 'null' value, corresponding to a 'null hypothesis,' or a point estimate such as the maximum likelihood value. A point estimate for θ must be substituted to compute a p-value in classical statistics.

Posterior predictive p-values. To evaluate the fit of the posterior distribution of a Bayesian model, we can compare the observed data to the posterior predictive distribution. In the Bayesian approach, test quantities can be functions of the unknown parameters as well as data because the test quantity is evaluated over draws from the posterior distribution of the unknown parameters. The Bayesian p-value is defined as the probability that the replicated data could be more extreme than the observed data, as measured by the test quantity:

$$p_B = \Pr(T(y^{\text{rep}}, \theta) \geq T(y, \theta)|y),$$

where the probability is taken over the posterior distribution of θ and the posterior predictive distribution of y^{rep} (that is, the joint distribution, $p(\theta, y^{\text{rep}}|y)$):

$$p_B = \int\int I_{T(y^{\text{rep}},\theta) \geq T(y,\theta)} p(y^{\text{rep}}|\theta) p(\theta|y) dy^{\text{rep}}\, d\theta,$$

where I is the indicator function. In this formula, we have used the property of the predictive distribution that $p(y^{\text{rep}}|\theta, y) = p(y^{\text{rep}}|\theta)$.

In practice, we usually compute the posterior predictive distribution using

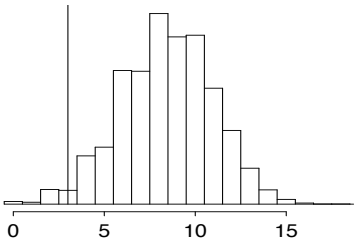

Figure 6.4 *Observed number of switches (vertical line at $T(y) = 3$), compared to* 10,000 *simulations from the posterior predictive distribution of the number of switches, $T(y^{\text{rep}})$.*

simulation. If we already have L simulations from the posterior density of θ, we just draw one y^{rep} from the predictive distribution for each simulated θ; we now have L draws from the joint posterior distribution, $p(y^{\text{rep}}, \theta|y)$. The posterior predictive check is the comparison between the realized test quantities, $T(y, \theta^l)$, and the predictive test quantities, $T(y^{\text{rep}\,l}, \theta^l)$. The estimated p-value is just the proportion of these L simulations for which the test quantity equals or exceeds its realized value; that is, for which $T(y^{\text{rep}\,l}, \theta^l) \geq T(y, \theta^l), l = 1, \ldots, L$.

In contrast to the classical approach, Bayesian model checking does not require special methods to handle 'nuisance parameters'; by using posterior simulations, we implicitly average over all the parameters in the model.

Example. Checking the assumption of independence in binomial trials We illustrate posterior predictive model checking with a simple hypothetical example. Consider a sequence of binary outcomes, $y_1, \ldots, y_n$, modeled as a specified number of iid Bernoulli trials with a uniform prior distribution on the probability of success, θ. As discussed in Chapter 2, the posterior density under the model is $p(\theta|y) \propto \theta^s(1-\theta)^{n-s}$, which depends on the data only through the sufficient statistic, $s = \sum y_i$. Now suppose the observed 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 is evidence that the model is flawed. To quantify the evidence, we can perform a posterior predictive test using the test quantity T = number of switches between 0 and 1 in the sequence. The observed value is $T(y) = 3$, and we can determine the posterior predictive distribution of $T(y^{\text{rep}})$ by simulation. To simulate y^{rep} under the model, we first draw θ from its Beta$(8, 14)$ posterior distribution, then draw $y^{\text{rep}} = (y_1^{\text{rep}}, \ldots, y_{20}^{\text{rep}})$ as independent Bernoulli variables with probability θ. Figure 6.4 displays a histogram of the values of $T(y^{\text{rep}\,l})$ for simulation draws $l = 1, \ldots, 10000$, with the observed value, $T(y) = 3$, shown by a vertical line. The observed number of switches is about one-third as many as would be expected from the model under the posterior predictive distribution, and the discrepancy cannot easily be explained by chance, as indicated by the computed p-value of $\frac{9838}{10000}$. To convert to a p-value near zero, we can change the sign of the test statistic, which amounts to computing $\Pr(T(y^{\text{rep}}, \theta) \leq T(y, \theta)|y)$, which is 0.028 in this

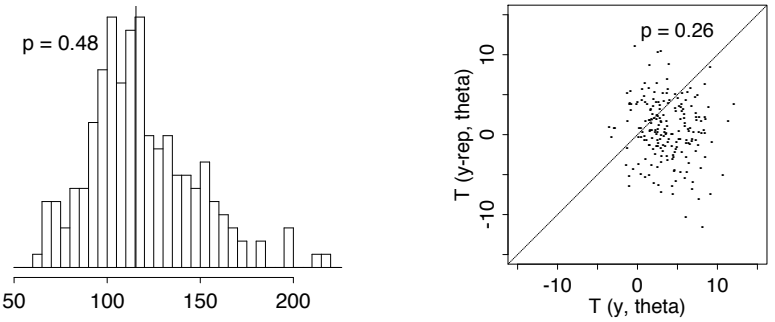

Figure 6.5 *Realized vs. posterior predictive distributions for two more test quantities in the speed of light example: (a) Sample variance (vertical line at 115.5), compared to 200 simulations from the posterior predictive distribution of the sample variance. (b) Scatterplot showing prior and posterior simulations of a test quantity:* $T(y,\theta) = |y_{(61)} - \theta| - |y_{(6)} - \theta|$ *(horizontal axis) vs.* $T(y^{\text{rep}},\theta) = |y^{\text{rep}}_{(61)} - \theta| - |y^{\text{rep}}_{(6)} - \theta|$ *(vertical axis) based on 200 simulations from the posterior distribution of* (θ, y^{rep})*. The p-value is computed as the proportion of points in the upper-left half of the plot.*

case. The p-values measured from the two ends have a sum that is greater than 1 because of the discreteness of the distribution of $T(y^{\text{rep}})$.

Example. Speed of light (continued)
In Figure 6.3, we demonstrated the poor fit of the normal model to the speed of light data using $\min(y_i)$ as the test statistic. We continue this example using other test quantities to illustrate how the fit of a model depends on the aspects of the data and parameters being monitored. Figure 6.5a shows the observed sample variance and the distribution of 200 simulated variances from the posterior predictive distribution. The sample variance does not make a good test statistic because it is a sufficient statistic of the model and thus, in the absence of an informative prior distribution, the posterior distribution will automatically be centered near the observed value. We are not at all surprised to find an estimated p-value close to $\frac{1}{2}$.

The model check based on $\min(y_i)$ earlier in the chapter suggests that the normal model is inadequate. To illustrate that a model can be inadequate for some purposes but adequate for others, we assess whether the model is adequate except for the extreme tails by considering a model check based on a test quantity sensitive to asymmetry in the center of the distribution,

$$T(y,\theta) = |y_{(61)} - \theta| - |y_{(6)} - \theta|.$$

The 61st and 6th order statistics are chosen to represent approximately the 90% and 10% points of the distribution. The test quantity should be scattered about zero for a symmetric distribution. The scatterplot in Figure 6.5b shows the test quantity for the observed data and the test quantity evaluated for the simulated data for 200 simulations from the posterior distribution of (θ, σ^2). The estimated p-value is 0.26, implying that any observed asymmetry in the middle of the distribution can easily be explained by sampling variation.

Defining replications

Depending on the aspect of the model one wishes to check, one can define the reference set of replications y^{rep} by conditioning on some or all of the observed data. For example, in checking the normal model for Newcomb's speed of light data, we kept the number of observations, n, fixed at the value in Newcomb's experiment. In Section 6.8, we check the hierarchical normal model for the SAT coaching experiments using posterior predictive simulations of new data on the same eight schools. It would also be possible to examine predictive simulations on new schools drawn from the same population. In analyses of sample surveys and designed experiments, it often makes sense to consider hypothetical replications of the experiment with a new randomization of selection or treatment assignment, by analogy to classical randomization tests.

6.4 Graphical posterior predictive checks

The basic idea of graphical model checking is to display the data alongside simulated data from the fitted model, and to look for systematic discrepancies between real and simulated data. This section gives examples of three kinds of graphical display:

- Direct display of all the data (as in the comparison of the speed-of-light data in Figure 3.1 to the 20 replications in Figure 6.2).
- Display of data summaries or parameter inferences. This can be useful in settings where the dataset is large and we wish to focus on the fit of a particular aspect of the model.
- Graphs of residuals or other measures of discrepancy between model and data.

Direct data display

Figure 6.6 shows another example of model checking by displaying all the data. The left column of the figure displays a three-way array of binary data—for each of 6 persons, a possible 'yes' or 'no' to each of 15 possible reactions (displayed as rows) to 23 situations (columns)—from an experiment in psychology. The three-way array is displayed as 6 slices, one for each person. Before displaying, the reactions, situations, and persons have been ordered in increasing average response. We can thus think of the test statistic $T(y)$ as being this graphical display, complete with the ordering applied to the data y.

The right columns of Figure 6.6 display seven independently-simulated replications y^{rep} from a fitted logistic regression model (with the rows, columns, and persons for each dataset arranged in increasing order before display, so that we are displaying $T(y^{\text{rep}})$ in each case). Here, the replicated datasets look fuzzy and 'random' compared to the observed data, which have strong rectilinear structures that are clearly not captured in the model. If the data

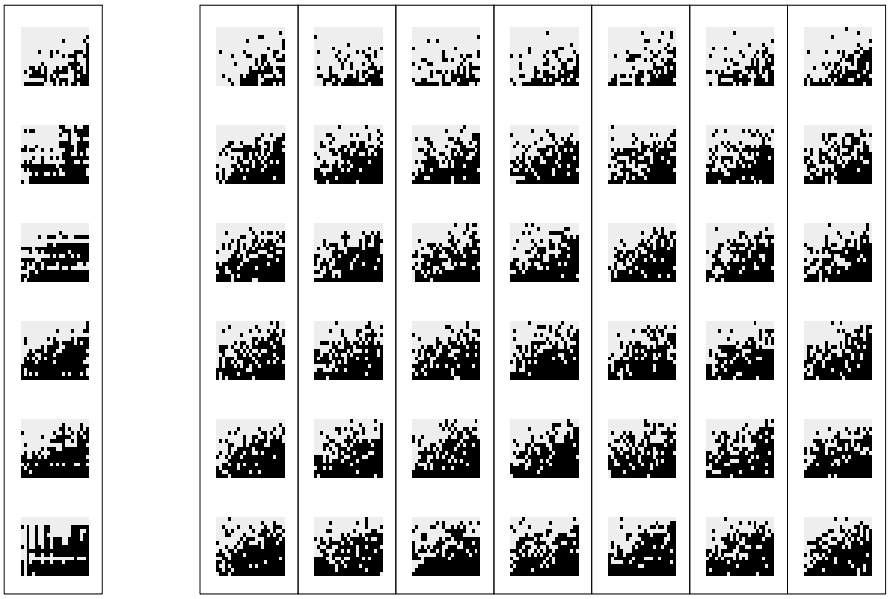

Figure 6.6 *Left column displays observed data y (a 15×23 array of binary responses from each of 6 persons); right columns display seven replicated datasets y^{rep} from a fitted logistic regression model. A misfit of model to data is apparent: the data show strong row and column patterns for individual persons (for example, the nearly white row near the middle of the last person's data) that do not appear in the replicates. (To make such patterns clearer, the indexes of the observed and each replicated dataset have been arranged in increasing order of average response.)*

were actually generated from the model, the observed data on the left would fit right in with the simulated datasets on the right.

Interestingly, these data have enough internal replication that the model misfit would be clear in comparison to a single simulated dataset from the model. But, to be safe, it is good to compare to several replications to see if the patterns in the observed data could be expected to occur by chance under the model.

Displaying data is not simply a matter of dumping a set of numbers on a page (or a screen). For example, we took care to align the graphs in Figure 6.6 to display the three-dimensional dataset and seven replications at once without confusion. Even more important, the arrangement of the rows, columns, and persons in increasing order is crucial to seeing the patterns in the data over and above the model. To see this, consider Figure 6.7, which presents the same information as in Figure 6.6 but without the ordering. Here, the discrepancies between data and model are not clear at all.

Figure 6.7 *Redisplay of Figure 6.6* without *ordering the rows, columns, and persons in order of increasing response. Once again, the left column shows the observed data and the right columns show replicated datasets from the model. Without the ordering, it is very difficult to notice the discrepancies between data and model, which are easily apparent in Figure 6.6.*

Displaying summary statistics or inferences

A key principle of exploratory data analysis is to exploit regular structure to display data more effectively. The analogy in modeling is hierarchical or multilevel modeling, in which batches of parameters capture variation at different levels. When checking model fit, hierarchical structure can allow us to compare batches of parameters to their reference distribution. In this scenario, the replications correspond to new draws of a batch of parameters.

We illustrate with inference from a hierarchical model from psychology. This was a fairly elaborate model, whose details we do not describe here; all we need to know for this example is that the model included two vectors of parameters, $\phi_1, \ldots, \phi_{90}$, and $\psi_1, \ldots, \psi_{69}$, corresponding to patients and psychological symptoms, and that each of these 159 parameters were assigned independent Beta$(2, 2)$ prior distributions. Each of these parameters represented a probability that a given patient or symptom is associated with a particular psychological syndrome.

Data were collected (measurements of which symptoms appeared in which patients) and the full Bayesian model was fitted, yielding posterior simulations for all these parameters. If the model were true, we would expect any single simulation draw of the vectors of patient parameters ϕ and symptom

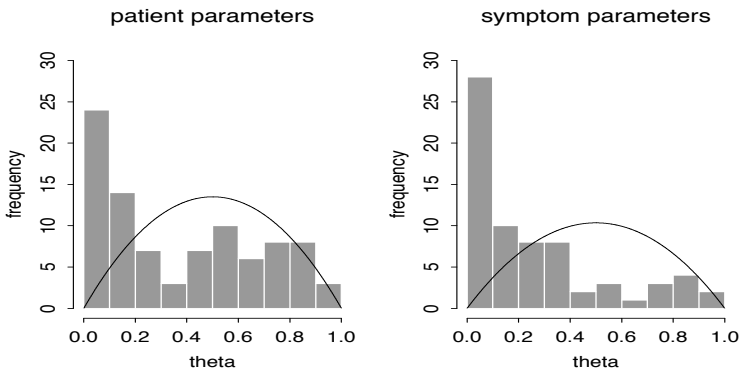

Figure 6.8 *Histograms of (a) 90 patient parameters and (b) 69 symptom parameters, from a single draw from the posterior distribution of a psychometric model. These histograms of posterior estimates contradict the assumed Beta*$(2, 2)$ *prior densities (overlain on the histograms) for each batch of parameters, and motivated us to switch to mixture prior distributions. This implicit comparison to the values under the prior distribution can be viewed as a posterior predictive check in which a new set of patients and a new set of symptoms are simulated.*

parameters ψ to look like independent draws from the Beta$(2, 2)$ distribution. We know this because of the following reasoning:

- If the model were indeed true, we could think of the observed data vector y and the vector θ of the true values of all the parameters (including ϕ and ψ) as a random draw from their joint distribution, $p(y, \theta)$. Thus, y comes from the marginal distribution, the prior predictive distribution, $p(y)$.
- A single draw θ^l from the posterior inference comes from $p(\theta^l|y)$. Since $y \sim p(y)$, this means that y, θ^l come from the model's joint distribution of y, θ, and so the marginal distribution of θ^l is the same as that of θ.
- That is, y, θ, θ^l have a combined joint distribution in which θ and θ^l have the same marginal distributions (and the same joint distributions with y).

Thus, as a model check we can plot a histogram of a single simulation of the vector of parameters ϕ or ψ and compare to the prior distribution. This corresponds to a posterior predictive check in which the inference from the observed data is compared to what would be expected if the model were applied to a new set of patients and a new set of symptoms.

Figure 6.8 shows histograms of a single simulation draw for each of ϕ and ψ as fitted to our dataset. The lines show the Beta$(2, 2)$ prior distribution, which clearly does not fit. For both ϕ and ψ, there are too many cases near zero, corresponding to patients and symptoms that almost certainly are not associated with a particular syndrome.

Our next step was to replace the offending Beta$(2, 2)$ prior distributions by mixtures of two beta distributions—one distribution with a spike near zero, and another that is uniform between 0 and 1—with different models for the

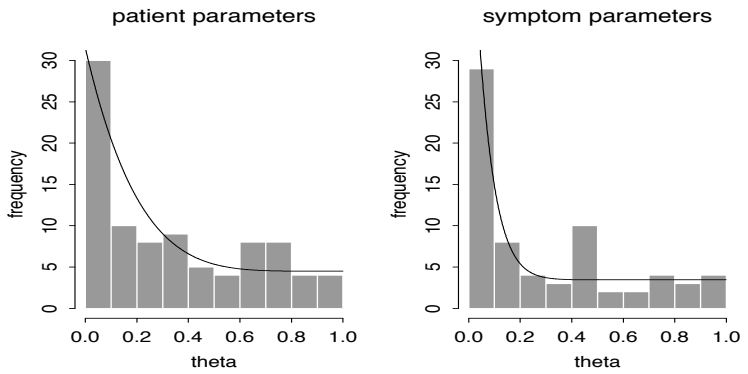

Figure 6.9 *Histograms of (a) 90 patient parameters and (b) 69 symptom parameters, as estimated from an expanded psychometric model. The mixture prior densities (overlain on the histograms) are not perfect, but they approximate the corresponding histograms much better than the Beta*$(2, 2)$ *densities in Figure 6.8.*

ϕ's and the ψ's. The exact model is,

$$\begin{aligned} p(\phi_j) &= 0.5\text{Beta}(\phi_j|1, 6) + 0.5\text{Beta}(\phi_j|1, 1) \\ p(\psi_j) &= 0.5\text{Beta}(\psi_j|1, 16) + 0.5\text{Beta}(\psi_j|1, 1). \end{aligned}$$

We set the parameters of the mixture distributions to fixed values based on our understanding of the model. It was reasonable for these data to suppose that any given symptom appeared only about half the time; however, labeling of the symptoms is subjective, so we used beta distributions peaked near zero but with some probability of taking small positive values. We assigned the Beta$(1, 1)$ (that is, uniform) distributions for the patient and symptom parameters that were not near zero—given the estimates in Figure 6.8, these seemed to fit the data better than the original Beta$(2, 2)$ models. (In fact, the original reason for using Beta$(2, 2)$ rather than uniform prior distributions was so that maximum likelihood estimates would be in the interior of the interval $[0, 1]$, a concern that disappeared when we moved to Bayesian inference; see Exercise 8.4.)

Some might object to revising the prior distribution based on the fit of the model to the data. It is, however, consistent with common statistical practice, in which a model is iteratively altered to provide a better fit to data. The natural next step would be to add a hierarchical structure, with hyperparameters for the mixture distributions for the patient and symptom parameters. This would require additional computational steps and potential new modeling difficulties (for example, instability in the estimated hyperparameters). Our main concern in this problem was to reasonably model the individual ϕ_j and ψ_j parameters without the prior distributions inappropriately interfering (which appears to be happening in Figure 6.8).

We refitted the model with the new prior distribution and repeated the

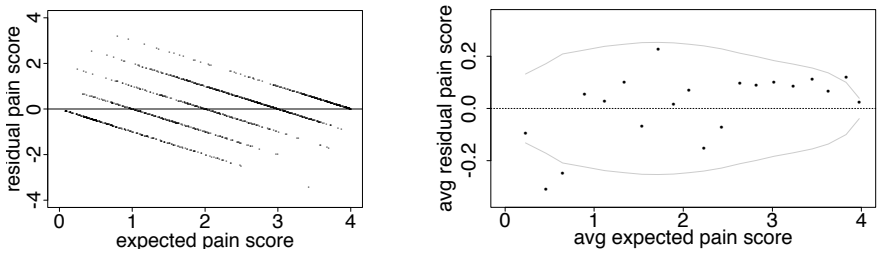

Figure 6.10 *(a) Residuals (observed − expected) vs. expected values for a model of pain relief scores (0 = no pain relief, ..., 5 = complete pain relief) (b) Average residuals vs. expected pain scores, with measurements divided into 20 equally-sized bins defined by ranges of expected pain scores. The average prediction errors are relatively small (note the scale of the y-axis), but with a consistent pattern that low predictions are too low and high predictions are too high. Dotted lines show 95% bounds under the model.*

model check, which is displayed in Figure 6.9. The fit of the prior distribution to the inferences is not perfect but is much better than before.

Residual plots and binned residual plots

Bayesian residuals. Linear and nonlinear regression models, which are the core tools of applied statistics, are characterized by a function $g(x, \theta) = \mathrm{E}(y|x, \theta)$, where x is a vector of predictors. Then, given the unknown parameters θ and the predictors x_i for a data point y_i, the *predicted value* is $g(x_i, \theta)$ and the *residual* is $y_i - g(x_i, \theta)$. This is sometimes called a 'realized' residual in contrast to the classical or estimated residual, $y_i - g(x_i, \hat{\theta})$, which is based on a point estimate $\hat{\theta}$ of the parameters.

A Bayesian residual graph plots a single realization of the residuals (based on a single random draw of θ). An example appears on page 513. However, classical residual plots can be thought of as approximations to the Bayesian version, ignoring posterior uncertainty in θ.

Binned residuals for discrete data. Unfortunately, for discrete data, plots of residuals can be difficult to interpret because, for any particular value of $\mathrm{E}(y_i|X, \theta)$, the residual r_i can only take on certain discrete values; thus, even if the model is correct, the residuals will not generally be expected to be independent of predicted values or covariates in the model. Figure 6.10 illustrates with data and then residuals plotted vs. fitted values, for a model of pain relief scores, which were discretely reported as 0, 1, 2, 3, or 4. The residuals have a distracting striped pattern because predicted values plus residuals equal discrete observed data values.

A standard way to make discrete residual plots more interpretable is to work with binned or smoothed residuals, which should be closer to symmetric about zero if enough residuals are included in each bin or smoothing category

(since the expectation of each residual is by definition zero, the central limit theorem ensures that the distribution of averages of many residuals will be approximately symmetric). In particular, suppose we would like to plot the vector of residuals r vs. some vector $w = (w_1, \ldots, w_n)$ that can in general be a function of X, θ, and perhaps y. We can bin the predictors and residuals by ordering the n values of w_i and sorting them into bins $k = 1, \ldots, K$, with approximately equal numbers of points n_k in each bin. For each bin, we then compute $\bar{w}_k$ and $\bar{r}_k$, the average values of w_i and r_i, respectively, for points i in bin k. The *binned residual plot* is the plot of the points $\bar{r}_k$ vs. $\bar{w}_k$, which actually must be represented by several plots (which perhaps can be overlain) representing variability due to uncertainty of θ in the posterior distribution.

Since we are viewing the plot as a test variable, it must be compared to the distribution of plots of $\bar{r}_k^{\text{rep}}$ vs. $\bar{w}_k^{\text{rep}}$, where, for each simulation draw, the values of $\bar{r}_k^{\text{rep}}$ are computed by averaging the replicated residuals $r_i^{\text{rep}} = y_i^{\text{rep}} - \text{E}(y_i|X, \theta)$ for points i in bin k. In general, the values of w_i can depend on y, and so the bins and the values of $\bar{w}_k^{\text{rep}}$ can vary among the replicated data sets.

Because we can compare to the distribution of simulated replications, the question arises: why do the binning at all? We do so because we want to understand the model misfits that we detect. Because of the discreteness of the data, the individual residuals r_i have asymmetric discrete distributions. As expected, the binned residuals are approximately symmetrically distributed. In general it is desirable for the posterior predictive reference distribution of a discrepancy variable to exhibit some simple features (in this case, independence and approximate normality of the $\bar{r}_k$'s) so that there is a clear interpretation of a misfit. This is, in fact, the same reason that one plots residuals, rather than data, vs. predicted values: it is easier to compare to an expected horizontal line than to an expected 45° line.

Under the model, the residuals are independent and, if enough are in each bin, the mean residuals $\bar{r}_k$ are approximately normally distributed. We can then display the reference distribution as 95% error bounds, as in Figure 6.10b. We never actually have to display the replicated data; the replication distribution is implicit, given our knowledge that the binned residuals are independent, approximately normally distributed, and with expected variation as shown by the error bounds.

General interpretation of graphs as model checks

More generally, we can compare any data display to replications under the model—not necessarily as an explicit model check but more to understand what the display 'should' look like if the model were true. For example, the maps and scatterplots of high and low cancer rates (Figures 2.7–2.9) show strong patterns, but these are not particularly informative if the same patterns would be expected of replications under the model. The erroneous initial interpretation of Figure 2.7—as evidence of a pattern of high cancer rates in

the sparsely-populated areas in the center-west of the country—can be thought of as an erroneous model check, in which the data display was compared to a random pattern rather than to the pattern expected under a reasonable model of variation in cancer occurrences.

6.5 Numerical posterior predictive checks

Choosing test quantities

The procedure for carrying out a posterior predictive model check requires specifying a test quantity, $T(y)$ or $T(y, \theta)$, and an appropriate predictive distribution for the replications y^{rep} (which involves deciding which if any aspects of the data to condition on, as discussed at the end of Section 6.3). If $T(y)$ does not appear to be consistent with the set of values $T(y^{\text{rep}\,1}), \ldots, T(y^{\text{rep}\,\text{L}})$, then the model is making predictions that do not fit the data. The discrepancy between $T(y)$ and the distribution of $T(y^{\text{rep}})$ can be summarized by a p-value (as discussed in Section 6.3) but we prefer to look at the magnitude of the discrepancy as well as its p-value. For example, in Figure 6.4 on page 163 we see that the observed number of switches is only about one-third what would be expected from the model, *and* the p-value of 0.028 indicates that it would be highly unlikely to see such a discrepancy in replicated data.

For many problems, a function of data and parameters can directly address a particular aspect of a model in a way that would be difficult or awkward using a function of data alone. If the test quantity depends on θ as well as y, then the test quantity $T(y, \theta)$ as well as its replication $T(y^{\text{rep}}, \theta)$ are unknowns and are represented by L simulations, and the comparison can be displayed either as a scatterplot of the values $T(y, \theta^l)$ vs. $T(y^{\text{rep}\,l}, \theta^l)$ or a histogram of the differences, $T(y, \theta^l) - T(y^{\text{rep}\,l}, \theta^l)$. Under the model, the scatterplot should be symmetric about the 45° line and the histogram should include 0.

Because a probability model can fail to reflect the process that generated the data in any number of ways, posterior predictive p-values can be computed for a variety of test quantities in order to evaluate more than one possible model failure. Ideally, the test quantities T will be chosen to reflect aspects of the model that are relevant to the scientific purposes to which the inference will be applied. Test quantities are commonly chosen to measure a feature of the data not directly addressed by the probability model; for example, ranks of the sample, or correlation of model residuals with some possible explanatory variable.

Example. Checking the fit of hierarchical regression models for adolescent smoking

We illustrate with a model fitted to a longitudinal data set of about 2000 Australian adolescents whose smoking patterns were recorded every six months (via questionnaire) for a period of three years. Interest lay in the extent to which smoking behavior could be predicted based on parental smoking and other background variables, and the extent to which boys and girls picked up the habit of smoking during their teenage years. Figure 6.11 illustrates the overall rate of

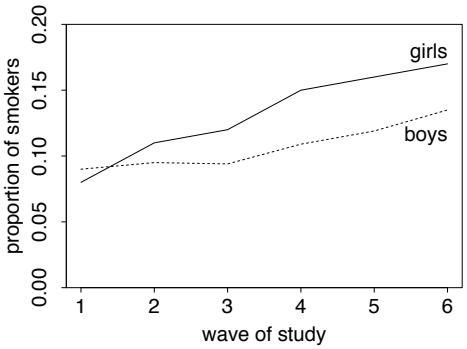

Figure 6.11 *Prevalence of regular (daily) smoking among participants responding at each wave in the study of Australian adolescents (who were on average 15 years old at wave 1).*

smoking among survey participants, who had an average age of 14.9 years at the beginning of the study.

We fit two models to these data. Our first model is a hierarchical logistic regression, in which the probability of smoking depends on sex, parental smoking, the wave of the study, and an individual parameter for the person. For person j at wave t, we model the probability of smoking as,

$$\Pr(y_{jt} = 1) = \text{logit}^{-1}(\beta_0 + \beta_1 X_{j1} + \beta_2 X_{j2} + \beta_3(1 - X_{j2})t + \beta_4 X_{j2} t + \alpha_j), \quad (6.3)$$

where X_{j1} is an indicator for parental smoking and X_{j2} is an indicator for females, so that β_3 and β_4 represent the time trends for males and females, respectively. The individual effects α_j are assigned a $\text{N}(0, \tau^2)$ distribution, with a noninformative uniform prior distribution on β, τ. (See Chapter 18 for more on hierarchical generalized linear models.)

The second model is an expansion of the first, in which each person j has an unobserved 'susceptibility' status S_j that equals 1 if the person might possibly smoke or 0 if he or she is 'immune' from smoking (that is, has no chance of becoming a smoker). This model is an oversimplification but captures the separation in the data between adolescents who often or occasionally smoke and those who never smoke at all. In this mixture model, the smoking status y_{jt} is automatically 0 at all times for non-susceptible persons. For those persons with $S_j = 1$, we use the model (6.3), understanding that these probabilities now refer to the probability of smoking, conditional on being susceptible. The model is completed with a logistic regression for susceptibility status given the individual-level predictors: $\Pr(S_j = 1) = \text{logit}^{-1}(\gamma_0 + \gamma_1 X_{j1} + \gamma_2 X_{j2})$, and a uniform prior distribution on these coefficients γ.

Table 6.1 shows the results for posterior predictive checks of the two fitted models using three different test statistics $T(y)$:

- The percentage of adolescents in the sample who never smoked.
- The percentage in the sample who smoked during all waves.
- The percentage of 'incident smokers': adolescents who began the study as non-

Test variable	$T(y)$	Model 1 95% int. for $T(y^{\text{rep}})$	Model 1 p-value	Model 2 95% int. for $T(y^{\text{rep}})$	Model 2 p-value
% never-smokers	77.3	$[75.5, 78.2]$	0.27	$[74.8, 79.9]$	0.53
% always-smokers	5.1	$[5.0, 6.5]$	0.95	$[3.8, 6.3]$	0.44
% incident smokers	8.4	$[5.3, 7.9]$	0.005	$[4.9, 7.8]$	0.004

Table 6.1 *Summary of posterior predictive checks for three test statistics for two models fit to the adolescent smoking data: (1) hierarchical logistic regression, and (2) hierarchical logistic regression with a mixture component for never-smokers. The second model better fits the percentages of never- and always-smokers, but still has a problem with the percentage of 'incident smokers,' who are defined as persons who report incidents of non-smoking followed by incidents of smoking.*

smokers, switched to smoking during the study period, and did not switch back.

From the first column of Table 6.1, we see that 77% of the sample never smoked, 5% always smoked, and 8% were incident smokers. The table then displays the posterior predictive distribution of each test statistic under each of the two fitted models. Both models accurately capture the percentage of never-smokers, but the second model better fits the percentage of always-smokers. It makes sense that the second model should fit this aspect of the data better, since its mixture form separates smokers from non-smokers. Finally, both models underpredict the proportion of incident smokers, which suggests that they are not completely fitting the variation of smoking behavior within individuals.

Posterior predictive checking is a useful direct way of assessing the fit of the model to these various aspects of the data. Our goal here is not to compare or choose among the models (a topic we discuss in Section 6.7) but rather to explore the ways in which either or both models might be lacking.

Numerical test quantities can also be constructed from patterns noticed visually (as in the test statistics chosen for the speed-of-light example in Section 6.3). This can be useful to quantify a pattern of potential interest, or to summarize a model check that will be performed repeatedly (for example, in checking the fit of a model that is applied to several different datasets).

Multiple comparisons

More generally, one might worry about interpreting the significance levels of multiple tests, or for test statistics chosen by inspection of the data. For example, we looked at three different test variables in checking the adolescent smoking models, so perhaps it is less surprising than it might seem at first that the worst-fitting test statistic had a p-value of 0.005. A 'multiple comparisons' adjustment would calculate the probability that the most extreme p-value would be as low as 0.005, which would perhaps yield an adjusted p-value somewhere near 0.015.

We do not make this adjustment, because we use the posterior predictive checks to see how particular aspects of the data would be expected to appear in replications. If we examine several test variables, we would not be surprised for some of them not to be fitted by the model—but if we are planning to apply the model, we might be interested in those aspects of the data that do not appear typical. We are not concerned with 'Type I error' rate—that is, the probability of rejecting a hypothesis conditional on it being true—because we use the checks not to accept or reject a model but rather to understand the limits of its applicability in realistic replications.

Omnibus tests such as χ^2

In addition to focused tests, it is often useful to check a model with a summary measure of fit such as the χ^2 discrepancy quantity, written here in terms of univariate responses y_i:

$$\chi^2 \text{ discrepancy:} \quad T(y,\theta) = \sum_i \frac{(y_i - E(y_i|\theta))^2}{\text{var}(y_i|\theta)}, \tag{6.4}$$

where the summation is over the sample observations. When θ is known, this test quantity resembles the classical χ^2 goodness-of-fit measure. A related option is the *deviance*, defined as $T(y,\theta) = -2\log p(y|\theta)$.

Classical χ^2 tests are based on test statistics such as $T(y) = T(y,\theta_{\text{null}})$, or $T(y) = \min_\theta T(y,\theta)$ or perhaps $T(y) = T(y,\theta_{\text{mle}})$, where θ_{null} is a value of θ under a null hypothesis of interest and θ_{mle} is the maximum likelihood estimate. The χ^2 reference distribution in these cases is based on large-sample approximations to the posterior distribution. The same test statistics can be used in a posterior predictive model check to produce a valid p-value with no restriction on the sample size. However, in the Bayesian context it can make more sense to simply work with $T(y,\theta)$ and avoid the minimization or other additional computational effort required in computing a purely data-based summary, $T(y)$. For a Bayesian χ^2 test, as in any other posterior predictive check, the reference distribution is automatically calculated from the posterior predictive simulations.

Interpreting posterior predictive p-values

A model is suspect if a discrepancy is of practical importance and its observed value has a tail-area probability that is close to 0 or 1, thereby indicating that the observed pattern would be unlikely to be seen in replications of the data if the model were true. An extreme p-value implies that the model cannot be expected to capture this aspect of the data. A p-value is a posterior probability and can therefore be interpreted directly—although *not* as Pr(model is true|data). Major failures of the model, typically corresponding to extreme tail-area probabilities (less than 0.01 or more than 0.99), can be addressed by expanding the model in an appropriate way. Lesser failures might

also suggest model improvements or might be ignored in the short term if the failure appears not to affect the main inferences. In some cases, even extreme p-values may be ignored if the misfit of the model is substantively small compared to variation within the model. We will often evaluate a model with respect to several test quantities, and we should be sensitive to the implications of this practice.

If a p-value is close to 0 or 1, it is not so important exactly how extreme it is. A p-value of 0.00001 is virtually no stronger, in practice, than 0.001; in either case, the aspect of the data measured by the test quantity is inconsistent with the model. A slight improvement in the model (or correction of a data coding error!) could bring either p-value to a reasonable range (between 0.05 and 0.95, say). The p-value measures 'statistical significance,' not 'practical significance.' The latter is determined by how different the observed data are from the reference distribution on a scale of substantive interest and depends on the goal of the study; an example in which a discrepancy is statistically but not practically significant appears at the end of Section 14.3.

The relevant goal is not to answer the question, 'Do the data come from the assumed model?' (to which the answer is almost always no), but to quantify the discrepancies between data and model, and assess whether they could have arisen by chance, under the model's own assumptions.

Limitations of posterior tests. Finding an extreme p-value and thus 'rejecting' a model is never the end of an analysis; the departures of the test quantity in question from its posterior predictive distribution will often suggest improvements of the model or places to check the data, as in the speed of light example. Conversely, even when the current model seems appropriate for drawing inferences (in that no unusual deviations between the model and the data are found), the next scientific step will often be a more rigorous experiment incorporating additional factors, thereby providing better data. For instance, in the educational testing example of Section 5.5, the data do not allow rejection of the model that all the θ_j's are equal, but that assumption is clearly unrealistic, and some of the substantive conclusions are greatly changed when the parameter τ is not restricted to be zero.

Finally, the discrepancies found by posterior predictive checking should be considered in their applied context. A model can be demonstrably wrong but can still work for some purposes, as we illustrate in a linear regression example in Section 14.3.

Relation to classical statistical tests

Bayesian posterior predictive checks are generalizations of classical tests in that they average over the posterior distribution of the unknown parameter vector θ rather than fixing it at some estimate $\hat{\theta}$. The Bayesian tests do not rely on the clever construction of pivotal quantities or on asymptotic results, and are therefore applicable to any probability model. This is not to suggest that the tests are automatic; the choice of test quantity and appropriate predictive

distribution requires careful consideration of the type of inferences required for the problem being considered.

6.6 Model expansion

Sensitivity analysis

In general, the posterior distribution of the model parameters can either overestimate or underestimate different aspects of 'true' posterior uncertainty. The posterior distribution typically overestimates uncertainty in the sense that one does not, in general, include all of one's substantive knowledge in the model; hence the utility of checking the model against one's substantive knowledge. On the other hand, the posterior distribution underestimates uncertainty in two senses: first, the assumed model is almost certainly wrong—hence the need for posterior model checking against the observed data—and second, other reasonable models could have fit the observed data equally well, hence the need for sensitivity analysis. We have already addressed model checking. In this section, we consider the uncertainty in posterior inferences due to the existence of reasonable alternative models and discuss how to expand the model to account for this uncertainty. Alternative models can differ in the specification of the prior distribution, in the specification of the likelihood, or both. Model checking and sensitivity analysis go together: when conducting sensitivity analysis, it is only necessary to consider models that fit substantive knowledge and observed data in relevant ways.

The basic method of sensitivity analysis is to fit several probability models to the same problem. It is often possible to avoid surprises in sensitivity analyses by replacing improper prior distributions with proper distributions that represent substantive prior knowledge. In addition, different questions are differently affected by model changes. Naturally, posterior inferences concerning medians of posterior distributions are generally less sensitive to changes in the model than inferences about means or extreme quantiles. Similarly, predictive inferences about quantities that are most like the observed data are most reliable; for example, in a regression model, interpolation is typically less sensitive to linearity assumptions than extrapolation. It is sometimes possible to perform a sensitivity analysis by using 'robust' models, which ensure that unusual observations (or larger units of analysis in a hierarchical model) do not exert an undue influence on inferences. The typical example is the use of the Student-t distribution in place of the normal (either for the sampling or the population distribution). Such models can be quite useful but require more computational effort. We consider robust models in Chapter 17.

Adding parameters to a model

There are several possible reasons to expand a model:

1. If the model does not fit the data or prior knowledge in some important

way, it should be altered in some way, possibly by adding enough new parameters to allow a better fit.

2. If a modeling assumption is questionable or has no real justification, one can broaden the class of models (for example, replacing a normal by a Student-t, as we do in Section 17.4 for the SAT coaching example).
3. If two different models, $p_1(y,\theta)$ and $p_2(y,\theta)$, are under consideration, they can be combined into a larger model using a continuous parameterization that includes the original models as special cases. For example, the hierarchical model for SAT coaching in Chapter 5 is a continuous generalization of the complete-pooling ($\tau = 0$) and no-pooling ($\tau = \infty$) models.
4. A model can be expanded to include new data; for example, an experiment previously analyzed on its own can be inserted into a hierarchical population model. Another common example is expanding a regression model of $y|x$ to a multivariate model of (x,y) in order to model missing data in x (see Chapter 21).

All these applications of model expansion have the same mathematical structure: the old model, $p(y,\theta)$, is embedded in or replaced by a new model, $p(y,\theta,\phi)$ or, more generally, $p(y,y^*,\theta,\phi)$, where y^* represents the added data.

The joint posterior distribution of the new parameters, ϕ, and the parameters θ of the old model is,

$$p(\theta,\phi|y,y^*) \propto p(\phi)p(\theta|\phi)p(y,y^*|\theta,\phi).$$

The conditional prior distribution, $p(\theta|\phi)$, and the likelihood, $p(y,y^*|\theta,\phi)$, are determined by the expanded family. The marginal distribution of ϕ is obtained by averaging over θ:

$$p(\phi|y,y^*) \propto p(\phi)\int p(\theta|\phi)p(y,y^*|\theta,\phi)d\theta. \tag{6.5}$$

In any expansion of a Bayesian model, one must specify a set of prior distributions, $p(\theta|\phi)$, to replace the old $p(\theta)$, and also a hyperprior distribution $p(\phi)$ on the hyperparameters. Both tasks typically require thought, especially with noninformative prior distributions (see Exercises 6.12 and 6.13). For example, Section 14.7 discusses a model for unequal variances that includes unweighted and weighted linear regression as extreme cases. In Section 17.4, we illustrate the task of expanding the normal model for the SAT coaching example of Section 5.5 to a Student-t model by including the degrees of freedom of the t distribution as an additional hyperparameter. Another detailed example of model expansion appears in Section 18.4, for a hierarchical mixture model applied to data from an experiment in psychology.

Practical advice for model checking and expansion

It is difficult to give appropriate general advice for model choice; as with model building, scientific judgment is required, and approaches must vary with context.

Our recommended approach, for both model checking and sensitivity analysis, is to examine posterior distributions of substantively important parameters and predicted quantities. Then we compare posterior distributions and posterior predictions with substantive knowledge, including the observed data, and note where the predictions fail. Discrepancies should be used to suggest possible expansions of the model, perhaps as simple as putting real prior information into the prior distribution or adding a parameter such as a nonlinear term in a regression, or perhaps requiring some substantive rethinking, as for the poor prediction of the Southern states in the Presidential election model as displayed in Figure 6.1 on page 159.

Sometimes a model has stronger assumptions than are immediately apparent. For example, a regression with many predictors and a flat prior distribution on the coefficients will tend to overestimate the variation among the coefficients, just as the independent estimates for the eight schools were more spread than appropriate. If we find that the model does not fit for its intended purposes, we are obliged to search for a new model that fits; an analysis is rarely, if ever, complete with simply a rejection of some model.

If a sensitivity analysis reveals problems, the basic solution is to include the other plausible models in the prior specification, thereby forming a posterior inference that reflects uncertainty in the model specification, or simply to report sensitivity to assumptions untestable by the data at hand. Of course, one must sometimes conclude that, for practical purposes, available data cannot effectively answer some questions. In other cases, it is possible to add information to constrain the model enough to allow useful inferences; Section 9.3 presents an example in the context of a simple random sample from a nonnormal population, in which the quantity of interest is the population total.

6.7 Model comparison

There are generally many options in setting up a model for any applied problem. Our usual approach is to start with a simple model that uses only some of the available information—for example, not using some possible predictors in a regression, fitting a normal model to discrete data, or ignoring evidence of unequal variances and fitting a simple equal-variance model. Once we have successfully fitted a simple model, we can check its fit to data (as discussed in Sections 6.3–6.5) and then expand it (as discussed in Section 6.6).

There are two typical scenarios in which models are compared. First, when a model is expanded, it is natural to compare the smaller to the larger model and assess what has been gained by expanding the model (or, conversely, if a model is simplified, to assess what was lost). This generalizes into the problem of comparing a set of nested models and judging how much complexity is necessary to fit the data.

In comparing nested models, the larger model typically has the advantage of making more sense and fitting the data better but the disadvantage of

being more difficult to understand and compute. The key questions of model comparison are typically: (1) is the improvement in fit large enough to justify the additional difficulty in fitting, and (2) is the prior distribution on the additional parameters reasonable?

The statistical theory of hypothesis testing is associated with methods for checking whether an improvement in fit is statistically significant—that is, whether it could be expected to occur by chance, even if the smaller model were correct. Bayes factors (see page 184) are sometimes used to make these model comparisons, but we find them generally to be irrelevant because they compute the relative probabilities of the models conditional on one of them being true. We prefer approaches that measure the distance of the data to each of the approximate models. Let θ be the vector of parameters in the smaller model and ψ be the additional parameters in the expanded model. Then we are comparing the two posterior distributions, $p(\theta|y)$ and $p(\theta, \psi|y)$, along with their predictive distributions for replicated data y^{rep}.

The second scenario of model comparison is between two or more nonnested models—neither model generalizes the other. In a regression context, one might compare regressions with completely different sets of predictors, for example, modeling political behavior using information based on past voting results or on demographics. In these settings, we are typically not interested in *choosing* one of the models—it would be better, both in substantive and predictive terms, to construct a larger model that includes both as special cases, including both sets of predictors and also potential interactions in a larger regression, possibly with an informative prior distribution if needed to control the estimation of all the extra parameters. However, it can be useful to *compare* the fit of the different models, to see how either set of predictors performs when considered alone.

Expected deviance as a measure of predictive accuracy

We first introduce some measures of prediction error and then discuss how they can be used to compare the performance of different models. We have already discussed discrepancy measures in Section 6.5 applied to checking whether data fit as well as could be expected under the model. Here, however, we are comparing the data to two (or more) different models and seeing which predicts with more accuracy. Even if none (or all) of the models fit the data, it can be informative to compare their relative fit.

Model fit can be summarized numerically by a measure such as weighted mean squared error as in (6.4): $T(y, \theta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \text{E}(y_i|\theta))^2/\text{var}(y_i|\theta)$. A more general option is to work with the 'deviance,' which is defined as -2 times the log-likelihood:

$$\text{deviance:} \quad D(y, \theta) = -2 \log p(y|\theta), \tag{6.6}$$

and is proportional to the mean squared error if the model is normal with constant variance. The deviance has an important role in statistical model com-

parison because of its connection to the Kullback-Leibler information measure $H(\theta)$ defined in (B.1) on page 586. The expected deviance—computed by averaging (6.6) over the true sampling distribution $f(y)$—equals 2 times the Kullback-Leibler information, up to a fixed constant, $\int f(y) \log f(y) dy$, that does not depend on θ. As discussed in Appendix B, in the limit of large sample sizes, the model with the lowest Kullback-Leibler information—and thus, the lowest expected deviance—will have the highest posterior probability. Thus, it seems reasonable to estimate expected deviance as a measure of overall model fit. More generally, we could measure lack of fit using any discrepancy measure D, but the deviance is a standard summary.

The discrepancy between data and model depends in general on θ as well as y. To get a summary that depends only on y, one could define

$$D_{\hat{\theta}}(y) = D(y, \hat{\theta}(y)) \tag{6.7}$$

and use a point estimate for θ such as the mean of the posterior simulations. From a Bayesian perspective, it is perhaps more appealing to average the discrepancy itself over the posterior distribution:

$$D_{\text{avg}}(y) = \text{E}(D(y, \theta)|y), \tag{6.8}$$

which would be estimated using posterior simulations θ_l:

$$\widehat{D}_{\text{avg}}(y) = \frac{1}{L} \sum_{l=1}^{L} D(y, \theta^l). \tag{6.9}$$

The estimated average discrepancy (6.9) is a better summary of model error than the discrepancy (6.7) of the point estimate; the point estimate generally makes the model fit better than it really does, and $\widehat{D}_{\text{avg}}$ averages over the range of possible parameter values. This can be thought of as a special case of the numerical posterior predictive checks discussed in Section 6.5, but with a focus on comparing discrepancies between models rather than checking model fit.

Counting parameters and model complexity

The difference between the posterior mean deviance (6.9) and the deviance at $\hat{\theta}$ (6.7),

$$p_D^{(1)} = \widehat{D}_{\text{avg}}(y) - D_{\hat{\theta}}(y), \tag{6.10}$$

represents the effect of model fitting and has been used as a measure of the *effective number of parameters* of a Bayesian model. For a normal linear model with unconstrained parameters (or, equivalently, for a fixed model with large sample size; see Chapter 4), $p_D^{(1)}$ is equal to the number of parameters in the model.

A related way to measure model complexity is as half the posterior variance

of the deviance, which can be estimated from the posterior simulations:

$$p_D^{(2)} = \frac{1}{2}\widehat{\text{var}}(D(y,\theta)|y) = \frac{1}{2}\frac{1}{L-1}\sum_{l=1}^{L}(D(y,\theta^l) - \widehat{D}_{\text{avg}}(y))^2.$$

Both measures $p_D^{(1)}$ and $p_D^{(2)}$ can be derived from the asymptotic χ^2 distribution of the deviance relative to its minimum, with a mean equal to the number of parameters estimated in the model and variance equal to twice the mean.

More generally, p_D can be thought of as the number of 'unconstrained' parameters in the model, where a parameter counts as: 1 if it is estimated with no constraints or prior information; 0 if it is fully constrained or if all the information about the parameter comes from the prior distribution; or an intermediate value if both the data and prior distributions are informative.

It makes sense that both definitions of p_D depend on the data, y. For a simple example, consider the model $y \sim \text{N}(\theta, 1)$, with $\theta \sim U(0, \infty)$. That is, θ is constrained to be positive but otherwise has a noninformative uniform prior distribution. How many parameters are being estimated by this model? If the measurement y is close to zero, then the effective number of parameters p is approximately 1/2, since roughly half the information in the posterior distribution is coming from the data and half from the prior constraint of positivity. However, if y is large, then the constraint is essentially irrelevant, and the effective number of parameters is approximately 1.

For hierarchical models, the effective number of parameters strongly depends on the variance of the group-level parameters. We shall illustrate with the example from Chapter 5 of the educational testing experiments in 8 schools. Under the hierarchical model, the effective number of parameters falls between 8 (one for each school) and 1 (for the mean of all the schools).

Estimated predictive discrepancy and the deviance information criterion

So far, we have defined the mean deviance $D_{\text{avg}}(y)$ of the fitted model to the data, along with p_D, the effective number of parameters being fitted. From (6.10), p_D represents the decrease in the deviance (that is, the expected improvement in the fit) expected from estimating the parameters in the model.

A related approach is to estimate the error that would be expected when applying the fitted model to future data, for example the expected mean squared predictive error, $D_{\text{avg}}^{\text{pred}}(y) = \text{E}\left[\frac{1}{n}\sum_{i=1}^{n}(y_i^{\text{rep}} - \text{E}(y_i^{\text{rep}}|y))^2\right]$, where the expectation averages over the posterior predictive distribution of replicated data y^{rep}. More generally, one can compute the expected deviance for replicated data:

$$D_{\text{avg}}^{\text{pred}}(y) = \text{E}[D(y^{\text{rep}}, \hat{\theta}(y))], \tag{6.11}$$

where $D(y^{\text{rep}}, \theta) = -2\log p(y^{\text{rep}}|\theta)$, the expectation averages over the distribution of y_{rep} under the unknown true sampling model, and $\hat{\theta}$ is a parameter estimate such as the posterior mean (or, more generally, the estimate that minimizes the expected deviance for replicated data). This use of a point estimate

Model	$D_{\hat{\theta}}$	$\widehat{D}_{\text{avg}}$	p_D	DIC
no pooling ($\tau = \infty$)	54.9	62.6	7.7	70.3
complete pooling ($\tau = 0$)	59.5	60.5	1.0	61.5
hierarchical (τ unknown)	57.8	60.6	2.8	63.4

Table 6.2 *Point-estimate and average deviances, estimated number of parameters, and DIC for each of three models fitted to the SAT coaching experiments in Section 5.5. Lower values of deviance represent better fit to data. The no-pooling model has the best-fitting point estimate $\hat{\theta}$, the complete-pooling and hierarchical models have the best average fit to data D_{avg}, and the complete-pooling model has lowest estimated expected predictive error DIC.*

$\hat{\theta}$ departs from our usual practice of averaging over the posterior distribution. In general, the expected predictive deviance (6.11) will be higher than the expected deviance $\widehat{D}_{\text{avg}}$ defined in (6.9) because the predicted data y^{rep} are being compared to a model estimated from data y.

The expected predictive deviance (6.11) has been suggested as a criterion of model fit when the goal is to pick a model with best out-of-sample predictive power. It can be approximately estimated by an expression called the *deviance information criterion* (DIC):

$$\text{DIC} = \widehat{D}^{\text{pred}}_{\text{avg}}(y) = 2\widehat{D}_{\text{avg}}(y) - D_{\hat{\theta}}(y), \tag{6.12}$$

with $D_{\hat{\theta}}(y)$ and $\widehat{D}_{\text{avg}}(y)$ as defined in (6.7) and (6.9). Expression (6.12) can be derived for normal models or in the limit of large sample sizes as in Chapter 4; see Exercise 6.8.

Example. Deviance for models of the educational testing experiments Table 6.2 illustrates the use of deviance to compare the three models—no pooling, complete pooling, and hierarchical—fitted to the SAT coaching data in Section 5.5. For this model, the deviance is simply

$$\begin{aligned} D(y,\theta) &= -2\sum_{j=1}^{J} \log\left[\text{N}(y_j|\theta_j,\sigma_j^2)\right] \\ &= \log(2\pi J\sigma^2) + \sum_{j=1}^{J}((y_j-\theta_j)/\sigma_j)^2. \end{aligned}$$

The first column, with $D_{\hat{\theta}}$, compares the best fits from 200 simulations for each model. The no-pooling has the best fit based on the point estimate $\hat{\theta}$, which makes sense since it allows all 8 parameters θ to be independently fitted by the data. The average discrepancy $\widehat{D}_{\text{avg}}$ of the data to the parameters as summarized by the posterior distribution of the no-pooling model is 62.6. There is no easy interpretation of this level, but the difference between the average discrepancy and that of the point estimate is the expected number of parameters, p_D, which is estimated at 7.7. This differs from the exact value of 8 because of simulation variability. Finally, DIC= 62.6+7.7 = 70.3 is the estimated expected discrepancy of the fitted unpooled model to replicated data.

The complete-pooling model, which assumes identical effects for all 8 schools, behaves much differently. The deviance of the point estimate is 59.5—quite a bit higher than for the no-pooling model, which makes sense, since the complete pooling model has only one parameter with which to fit the data. The average discrepancy D_{avg}, however, is lower for the complete pooling model, because there is little uncertainty in the posterior distribution. The estimated number of parameters p_D is 1.0, which makes perfect sense.

Finally, the point estimate from the hierarchical model has a deviance that is higher than the no-pooling model and lower than complete pooling, which makes sense since the hierarchical model has 8 parameters with which to fit the data, but these parameters are constrained by the prior distribution. The expected discrepancy D_{avg} is about the same as under complete pooling, and the estimated number of parameters is 2.8—closer to 1 than to 8, which makes sense, considering that for these data, τ is estimated to be low and the parameter estimates θ_j are shrunk strongly toward their common mean (see Figure 5.6). Finally, DIC for this model is 63.4, which is better than the 70.3 for the no-pooling model but worse than the 61.5 for complete pooling as applied to these data. Thus, we would expect the complete-pooling model to do the best job—in terms of log-likelihood—of predicting future data.

We would probably still prefer to use the hierarchical model, for the reasons discussed in Section 5.5—basically, we do not want to make the strong assumption that all coaching programs are identical—but for this particular dataset, the variance inherent in estimating the eight θ_j values, with a uniform hyperprior distribution on τ, does seem to add appreciable noise to predictions (of results for new schools).

In general, the predictive accuracy measures are useful in parallel with posterior predictive checks to see if there are important patterns in the data that are not captured by each model. As with predictive checking, DIC can be computed in different ways for a hierarchical model depending on whether the parameter estimates $\hat{\theta}$ and replications y^{rep} correspond to estimates and replications of new data from the existing groups (as we have performed the calculations in the above example) or new groups (additional schools from the $\text{N}(\mu, \tau^2)$ distribution in the above example).

Comparing a discrete set of models using Bayes factors

In a problem in which a discrete set of competing models is proposed, the term *Bayes factor* is sometimes used for the ratio of the marginal likelihood under one model to the marginal likelihood under a second model. If we label two competing models as H_1 and H_2, then the ratio of their posterior probabilities is

$$\frac{p(H_2|y)}{p(H_1|y)} = \frac{p(H_2)}{p(H_1)} \times \text{Bayes factor}(H_2; H_1),$$

where

$$\text{Bayes factor}(H_2; H_1) = \frac{p(y|H_2)}{p(y|H_1)} = \frac{\int p(\theta_2|H_2)p(y|\theta_2, H_2)d\theta_2}{\int p(\theta_1|H_1)p(y|\theta_1, H_1)d\theta_1}. \qquad (6.13)$$

In many cases, the competing models have a common set of parameters, but this is not necessary; hence the notation θ_i for the parameters in model H_i. As expression (6.13) makes clear, the Bayes factor is only defined when the marginal density of y under each model is proper.

The goal when using Bayes factors is to choose a single model H_i or average over a discrete set using their posterior distributions, $p(H_i|y)$. As we show by our examples in this book, we generally prefer to replace a discrete set of models with an expanded continuous family of models. To illustrate this, we consider two examples: one in which the Bayes factor is helpful and one in which it is not. The bibliographic note at the end of the chapter provides pointers to more extensive treatments of Bayes factors.

Example. An example in which Bayes factors are helpful
The Bayesian inference for the genetics example in Section 1.4 can be fruitfully reintroduced in terms of Bayes factors, with the two competing 'models' being H_1: the woman is affected, and H_2: the woman is unaffected, that is, $\theta = 1$ and $\theta = 0$ in the notation of Section 1.4. The prior odds are $p(H_2)/p(H_1) = 1$, and the Bayes factor of the data that the woman has two unaffected sons is $p(y|H_2)/p(y|H_1) = 1.0/0.25$. The posterior odds are thus $p(H_2|y)/p(H_1|y) = 4$. Computation by multiplying odds ratios makes the accumulation of evidence clear.

This example has two features that allow Bayes factors to be helpful. First, each of the discrete alternatives makes scientific sense, and there are no obvious scientific models in between. Second, the marginal distribution of the data under each model, $p(y|H_i)$, is proper.

Example. An example in which Bayes factors are a distraction
We now consider an example in which discrete model comparisons and Bayes factors are a distraction from scientific inference. Suppose we analyzed the SAT coaching experiments in Section 5.5 using Bayes factors for the discrete collection of previously proposed standard models, no pooling (H_1) and complete pooling (H_2):

$$H_1 : p(y|\theta_1, \ldots, \theta_J) = \prod_{j=1}^{J} \mathrm{N}(y_j|\theta_j, \sigma_j^2),\ p(\theta_1, \ldots, \theta_J) \propto 1$$

$$H_2 : p(y|\theta_1, \ldots, \theta_J) = \prod_{j=1}^{J} \mathrm{N}(y_j|\theta_j, \sigma_j^2),\ \theta_1 = \cdots = \theta_J = \theta, p(\theta) \propto 1.$$

(Recall that the standard deviations σ_j are assumed known in this example.)

If we use Bayes factors to choose or average among these models, we are immediately confronted with the fact that the Bayes factor—the ratio $p(y|H_1)/p(y|H_2)$—is not defined; because the prior distributions are improper, the ratio of density functions is 0/0. Consequently, if we wish to continue with the approach of assigning posterior probabilities to these two discrete models, we must consider (1) proper prior distributions, or (2) improper prior distributions that are carefully constructed as limits of proper distributions. In either case, we shall see that the results are unsatisfactory.

More explicitly, suppose we replace the flat prior distributions in H_1 and H_2 by

independent normal prior distributions, $\mathrm{N}(0, A^2)$, for some large A. The resulting posterior distribution for the effect in school j is

$$p(\theta_j|y) = (1-\lambda)p(\theta_j|y, H_1) + \lambda p(\theta_j|y, H_2),$$

where the two conditional posterior distributions are normal centered near y_j and $\overline{y}$, respectively, and λ is proportional to the prior odds times the Bayes factor, which is a function of the data and A (see Exercise 6.9). The Bayes factor for this problem is highly sensitive to the prior variance, A^2; as A increases (with fixed data and fixed prior odds, $p(H_2)/p(H_1)$) the posterior distribution becomes more and more concentrated on H_2, the complete pooling model. Therefore, the Bayes factor cannot be reasonably applied to the original models with noninformative prior densities, even if they are carefully defined as limits of proper prior distributions.

Yet another problem with the Bayes factor for this example is revealed by considering its behavior as the number of schools being fitted to the model increases. The posterior distribution for θ_j under the mixture of H_1 and H_2 turns out to be sensitive to the dimensionality of the problem, as very different inferences would be obtained if, for example, the model were applied to similar data on 80 schools (see Exercise 6.9). It makes no scientific sense for the posterior distribution to be highly sensitive to aspects of the prior distributions and problem structure that are scientifically incidental.

Thus, if we were to use a Bayes factor for this problem, we would find a problem in the model-checking stage (a discrepancy between posterior distribution and substantive knowledge), and we would be moved toward setting up a smoother, continuous family of models to bridge the gap between the two extremes. A reasonable continuous family of models is $y_j \sim \mathrm{N}(\theta_j, \sigma_j^2)$, $\theta_j \sim \mathrm{N}(\mu, \tau^2)$, with a flat prior distribution on μ, and τ in the range $[0, \infty)$; this, of course, is the model we used in Section 5.5. Once the continuous expanded model is fitted, there is no reason to assign discrete positive probabilities to the values $\tau = 0$ and $\tau = \infty$, considering that neither makes scientific sense.

6.8 Model checking for the educational testing example

We illustrate the ideas discussed in this chapter with the SAT coaching example introduced in Section 5.5.

Assumptions of the model

The posterior inference presented for the educational testing example is based on several model assumptions: (1) the normality of the estimates y_j given θ_j and σ_j, where the values σ_j are assumed known; (2) the exchangeability of the prior distribution of the θ_j's; (3) the normality of the prior distribution of each θ_j given μ and τ; and (4) the uniformity of the hyperprior distribution of (μ, τ).

The assumption of normality with a known variance is made routinely when a study is summarized by its estimated effect and standard error. The design (randomization, reasonably large sample sizes, adjustment for scores on earlier

tests) and analysis (for example, the raw data of individual test scores were checked for outliers in an earlier analysis) were such that the assumptions seem justifiable in this case.

The second modeling assumption deserves commentary. The real-world interpretation of the mathematical assumption of exchangeability of the θ_j's is that before seeing the results of the experiments, there is no desire to include in the model features such as a belief that (a) the effect in school A is probably larger than in school B or (b) the effects in schools A and B are more similar than in schools A and C. In other words, the exchangeability assumption means that we will let the data tell us about the relative ordering and similarity of effects in the schools. Such a prior stance seems reasonable when the results of eight parallel experiments are being scientifically summarized for general presentation. Of course, generally accepted information concerning the effectiveness of the programs or differences among the schools might suggest a nonexchangeable prior distribution if, for example, schools B and C have similar students and schools A, D, E, F, G, H have similar students. Unusual types of detailed prior knowledge (for example, two schools are very similar but we do not know which schools they are) can suggest an exchangeable prior distribution that is not a mixture of iid components. In the absence of any such information, the exchangeability assumption implies that the prior distribution of the θ_j's can be considered as independent samples from a population whose distribution is indexed by some hyperparameters—in our model, (μ, τ)—that have their own hyperprior distribution.

The third and fourth modeling assumptions are harder to justify *a priori* than the first two. Why should the school effects be normally distributed rather than say, Cauchy distributed, or even asymmetrically distributed, and why should the location and scale parameters of this prior distribution be uniformly distributed? Mathematical tractability is one reason for the choice of models, but if the family of probability models is inappropriate, Bayesian answers can be quite misleading.

Comparing posterior inferences to substantive knowledge

Inference about the parameters in the model. When checking the model assumptions, our first step is to compare the posterior distribution of effects to our knowledge of educational testing. The estimated treatment effects (the posterior means) for the eight schools range from 5 to 10 points, which are plausible values. (The SAT-V is scored on a scale from 200 to 800.) The effect in school A could be as high as 31 points or as low as -2 points (a 95% posterior interval). Either of these extremes seems plausible. We could look at other summaries as well, but it seems clear that the posterior estimates of the parameters do not violate our common sense or our limited substantive knowledge about SAT preparation courses.

Inference about predicted values. Next, we simulate the posterior predictive distribution of a hypothetical replication of the experiments. Computationally,

drawing from the posterior predictive distribution is nearly effortless given all that we have done so far: from each of the 200 simulations from the posterior distribution of (θ, μ, τ), we simulate a hypothetical replicated dataset, $y^{\text{rep}} = (y_1^{\text{rep}}, \ldots, y_8^{\text{rep}})$, by drawing each y_j^{rep} from a normal distribution with mean θ_j and standard deviation σ_j. The resulting set of 200 vectors y^{rep} summarizes the posterior predictive distribution. (Recall from Section 5.5 that we are treating y—the eight separate estimates—as the 'raw data' from the eight experiments.)

The model-generated values for each school in each of the 200 replications are all plausible outcomes of experiments on coaching. The smallest hypothetical observation generated was -48, and the largest was 63; because both values are possible for estimated effects from studies of coaching for the SAT-V, all estimated values generated by the model are credible.

Posterior predictive checking

But does the model fit the data? If not, we may have cause to doubt the inferences obtained under the model such as displayed in Figure 5.8 and Table 5.3. For instance, is the largest observed outcome, 28 points, consistent with the posterior predictive distribution under the model? Suppose we perform 200 posterior predictive simulations of the SAT coaching experiments and compute the largest observed outcome, $\max_j y_j^{\text{rep}}$, for each. If all 200 of these simulations lie below 28 points, then the model does not fit this important aspect of the data, and we might suspect that the normal-based inference in Section 5.5 shrinks the effect in School A too far.

In order to test the fit of the model to the observed data, we examine the posterior predictive distribution of the following test statistics: the largest of the eight observed outcomes, $\max_j y_j$, the smallest, $\min_j y_j$, the average, $\text{mean}(y_j)$, and the sample standard deviation, $\text{sd}(y_j)$. We approximate the posterior predictive distribution of each test statistic by the histogram of the values from the 200 simulations of the parameters and predictive data, and we compare each distribution to the observed value of the test statistic and our substantive knowledge of SAT coaching programs. The results are displayed in Figure 6.12.

The summaries suggest that the model generates predicted results similar to the observed data in the study; that is, the actual observations are typical of the predicted observations generated by the model.

Of course, there are many other functions of the posterior predictive distribution that could be examined, such as the differences between individual values of y_j^{rep}. Or, if we had a particular skewed nonnormal prior distribution in mind for the effects θ_j, we could construct a test quantity based on the skewness or asymmetry of the simulated predictive data as a check on whether the normal model is adequate. Often in practice we can obtain diagnostically useful displays directly from intuitively interesting quantities without having to supply a specific alternative model.

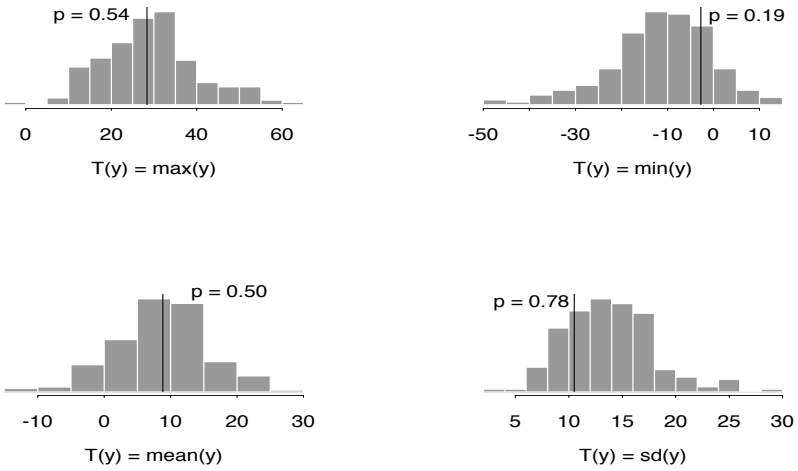

Figure 6.12 *Posterior predictive distribution, observed result, and p-value for each of four test statistics for the educational testing example.*

Sensitivity analysis

The model checks seem to support the posterior inferences for the SAT coaching example. Although we may feel confident that the data do not contradict the model, this is not enough to inspire complete confidence in our general substantive conclusions, because other reasonable models might provide just as good a fit but lead to different conclusions. Sensitivity analysis can then be used to assess the effect of alternative analyses on the posterior inferences.

The uniform prior distribution for τ. To assess the sensitivity to the prior distribution for τ we consider Figure 5.5, the graph of the marginal posterior density, $p(\tau|y)$, obtained under the assumption of a uniform prior density for τ on the positive half of the real line. One can obtain the posterior density for τ given other choices of the prior distribution by multiplying the density displayed in Figure 5.5 by the prior density. There will be little change in the posterior inferences as long as the prior density is not sharply peaked and does not put a great deal of probability mass on values of τ greater than 10.

The normal population distribution for the school effects. The normal distribution assumption on the θ_j's is made for computational convenience, as is often the case. A natural sensitivity analysis is to consider longer-tailed alternatives, such as the Student-t, as a check on robustness. We defer the details of this analysis to Section 17.4, after the required computational techniques have been presented. Any alternative model must be examined to ensure that the predictive distributions are restricted to realistic SAT improvements.

The normal likelihood. As discussed earlier, the assumption of normal data conditional on the means and standard deviations need not and cannot be

seriously challenged in this example. The justification is based on the central limit theorem and the designs of the studies. Assessing the validity of this assumption would require access to the original data from the eight experiments, not just the estimates and standard errors given in Table 5.2.

6.9 Bibliographic note

The posterior predictive approach to model checking described here was presented in Rubin (1981, 1984). Gelman, Meng, and Stern (1996) discuss the use of test quantities that depend on parameters as well as data; related ideas appear in Zellner (1976) and Tsui and Weerahandi (1989). Rubin and Stern (1994) and Raghunathan (1994) provide further applied examples. The examples in Section 6.4 appear in Meulders et al. (1998) and Gelman (2003). The antisymmetric discrepancy measures discussed in Section 6.5 appear in Berkhof, Van Mechelen, and Gelman (2003a). The adolescent smoking example appears in Carlin et al. (2001). Sinharay and Stern (2003) discuss posterior predictive checks for hierarchical models, focusing on the SAT coaching example. Johnson (2002) discusses Bayesian χ^2 tests as well as the idea of using predictive checks as a debugging tool, as discussed in Section 10.3.

Model checking using simulation has a long history in statistics; for example, Bush and Mosteller (1955, p. 252) check the fit of a model by comparing observed data to a set of simulated data. Their method differs from posterior predictive checking only in that their model parameters were fixed at point estimates for the simulations rather than being drawn from a posterior distribution. Ripley (1988) applies this idea repeatedly to examine the fits of models for spatial data. Early theoretical papers featuring ideas related to Bayesian posterior predictive checks include Guttman (1967) and Dempster (1971). Bernardo and Smith (1994) discuss methods of comparing models based on predictive errors.

A related approach to model checking is *cross-validation*, in which observed data are partitioned, with each part of the data compared to its predictions conditional on the model and the rest of the data. Some references to Bayesian approaches to cross-validation include Stone (1974), Geisser and Eddy (1979), and Gelfand, Dey, and Chang (1992). Geisser (1986) discusses predictive inference and model checking in general, and Barbieri and Berger (2002) discuss Bayesian predictive model selection.

Box (1980, 1983) has contributed a wide-ranging discussion of model checking ('model criticism' in his terminology), including a consideration of why it is needed in addition to model expansion and averaging. Box proposed checking models by comparing data to the *prior predictive distribution*; in the notation of our Section 6.3, defining replications with distribution $p(y^{\text{rep}}) = \int p(y^{\text{rep}}|\theta)p(\theta)d\theta$. This approach has quite different implications for model checking; for example, with an improper prior distribution on θ, the prior predictive distribution is itself improper and thus the check is not generally defined, even if the posterior distribution is proper (see Exercise 6.7).

Box was also an early contributor to the literature on sensitivity analysis and robustness in standard models based on normal distributions: see Box and Tiao (1962, 1973).

Various theoretical studies have been performed on Bayesian robustness and sensitivity analysis examining the question of how posterior inferences are affected by prior assumptions; see Leamer (1978b), McCulloch (1989), Wasserman (1992), and the references at the end of Chapter 17. Kass and coworkers have developed methods based on Laplace's approximation for approximate sensitivity analysis: for example, see Kass, Tierney, and Kadane (1989) and Kass and Vaidyanathan (1992).

Nelder and Wedderburn (1972) explore the deviance as a measure of model fit, Akaike (1973) introduce the expected predictive deviance and the AIC, and Mallows (1973) derives the related C_p measure. Bayesian treatments of expected predictive errors for model comparison include Dempster (1974), Laud and Ibrahim (1995), and Gelfand and Ghosh (1998). Hansen and Yu (2001) review related ideas from an information-theoretic perspective.

The deviance information criterion (DIC) and its calculation using posterior simulations are described by Spiegelhalter et al. (2002) and is implemented in the software package Bugs; see Spiegelhalter et al. (1994, 2003). Burnham and Anderson (2002) discuss and motivate the use of the Kullback-Leibler information for model comparison, which relates to the log-likelihood deviance function used in determining DIC. The topic of counting parameters in nonlinear, constrained, and hierarchical models is discussed by Hastie and Tibshirani (1990), Gelman, Meng, and Stern (1996), Hodges and Sargent (2001), and Vaida and Blanchard (2002). The last paper discusses the different ways that information criteria can be computed in hierarchical models.

A comprehensive overview of the use of Bayes factors for comparing models and testing scientific hypotheses is given by Kass and Raftery (1995), which contains many further references in this area. Carlin and Chib (1993) discuss the problem of averaging over models that have incompatible parameterizations. Chib (1995) and Chib and Jeliazkov (2001) describe approaches for calculating the marginal likelihoods required for Bayes factors from iterative simulation output (as produced by the methods described in Chapter 11). Pauler, Wakefield, and Kass (1999) discuss Bayes factors for hierarchical models. Weiss (1996) considers the use of Bayes factors for sensitivity analysis.

Bayes factors are not defined for models with improper prior distributions, but there have been several attempts to define analogous quantities; see Spiegelhalter and Smith (1982) and Kass and Raftery (1995). A related proposal is to treat Bayes factors as posterior probabilities and then average over competing models—see Raftery (1996) for a theoretical treatment, Rosenkranz and Raftery (1994) for an application, and Hoeting et al. (1999) and Chipman, George, and McCulloch (2001) for reviews.

A variety of views on model selection and averaging appear in the articles by Draper (1995) and O'Hagan (1995) and the accompanying discussions. We refer the reader to these articles and their references for further discussion

and examples of these methods. Because we emphasize continuous *families* of models rather than discrete *choices*, Bayes factors are rarely relevant in our approach to Bayesian statistics; see Raftery (1995) and Gelman and Rubin (1995) for two contrasting views on this point.

There are many examples of applied Bayesian analyses in which sensitivity to the model has been examined, for example Racine et al. (1986), Weiss (1994), and Smith, Spiegelhalter, and Thomas (1995).

Finally, many model checking methods in common practical use, including tests for outliers, plots of residuals, and normal plots, can be interpreted as Bayesian posterior predictive checks, where the practitioner is looking for discrepancies from the expected results under the assumed model (see Gelman, 2003, for an extended discussion of this point). Many non-Bayesian treatments of graphical model checking appear in the statistical literature, for example, Atkinson (1985). Tukey (1977) presents a graphical approach to data analysis that is, in our opinion, fundamentally based on model checking (see Gelman, 2003). The books by Cleveland (1985, 1993) and Tufte (1983, 1990) present many useful ideas for displaying data graphically; these ideas are fundamental to the graphical model checks described in Section 6.4.

Calvin and Sedransk (1991) provide an interesting example comparing various Bayesian and non-Bayesian methods of model checking and expansion.

6.10 Exercises

1. Posterior predictive checking:
   (a) On page 140, the data from the SAT coaching experiments were checked against the model that assumed identical effects in all eight schools: the expected order statistics of the effect sizes were (26, 19, 14, 10, 6, 2, −3, −9), compared to observed data of (28, 18, 12, 7, 3, 1, −1, −3). Express this comparison formally as a posterior predictive check comparing this model to the data. Does the model fit the aspect of the data tested here?
   (b) Explain why, even though the identical-schools model fits under this test, it is still unacceptable for some practical purposes.
2. Model checking: in Exercise 2.13, the counts of airline fatalities in 1976–1985 were fitted to four different Poisson models.
   (a) For each of the models, set up posterior predictive test quantities to check the following assumptions: (1) independent Poisson distributions, (2) no trend over time.
   (b) For each of the models, use simulations from the posterior predictive distributions to measure the discrepancies. Display the discrepancies graphically and give p-values.
   (c) Do the results of the posterior predictive checks agree with your answers in Exercise 2.13(e)?
3. Model improvement:

(a) Use the solution to the previous problem and your substantive knowledge to construct an improved model for airline fatalities.

(b) Fit the new model to the airline fatality data.

(c) Use your new model to forecast the airline fatalities in 1986. How does this differ from the forecasts from the previous models?

(d) Check the new model using the same posterior predictive checks as you used in the previous models. Does the new model fit better?

4. Model checking and sensitivity analysis: find a published Bayesian data analysis from the statistical literature.

(a) Compare the data to posterior predictive replications of the data.

(b) Perform a sensitivity analysis by computing posterior inferences under plausible alternative models.

5. Hypothesis testing: discuss the statement, 'Null hypotheses of no difference are usually known to be false before the data are collected; when they are, their rejection or acceptance simply reflects the size of the sample and the power of the test, and is not a contribution to science' (Savage, 1957, quoted in Kish, 1965). If you agree with this statement, what does this say about the model checking discussed in this chapter?

6. Variety of predictive reference sets: in the example of binary outcomes on page 163, it is assumed that the number of measurements, n, is fixed in advance, and so the hypothetical replications under the binomial model are performed with $n = 20$. Suppose instead that the protocol for measurement is to stop once 13 zeros have appeared.

(a) Explain why the posterior distribution of the parameter θ under the assumed model does not change.

(b) Perform a posterior predictive check, using the same test quantity, $T =$ number of switches, but simulating the replications y^{rep} under the new measurement protocol. Display the predictive simulations, $T(y^{\text{rep}})$, and discuss how they differ from Figure 6.4.

7. Prior vs. posterior predictive checks (from Gelman, Meng, and Stern, 1996): consider 100 observations, $y_1, \ldots, y_n$, modeled as independent samples from a $N(\theta, 1)$ distribution with a diffuse prior distribution, say, $p(\theta) = \frac{1}{2A}$ for $\theta \in [-A, A]$ with some extremely large value of A, such as 10^5. We wish to check the model using, as a test statistic, $T(y) = \max_i |y_i|$: is the maximum absolute observed value consistent with the normal model? Consider a dataset in which $\overline{y} = 5.1$ and $T(y) = 8.1$.

(a) What is the posterior predictive distribution for y^{rep}? Make a histogram for the posterior predictive distribution of $T(y^{\text{rep}})$ and give the posterior predictive p-value for the observation $T(y) = 8.1$.

(b) The prior predictive distribution is $p(y^{\text{rep}}) = \int p(y^{\text{rep}}|\theta)p(\theta)d\theta$. (Compare to equation (6.1).) What is the prior predictive distribution for y^{rep} in this example? Roughly sketch the prior predictive distribution of $T(y^{\text{rep}})$ and give the approximate prior predictive p-value for the observation $T(y) = 8.1$.

(c) Your answers for (a) and (b) should show that the data are consistent with the posterior predictive but not the prior predictive distribution. Does this make sense? Explain.

8. Deviance information criterion: show that expression (6.12) is appropriate for normal models or in the asymptotic limit of large sample sizes (see Spiegelhalter et al., 2002, p. 604).

9. Prior and posterior predictive checks when the prior distribution is improper: on page 185, we discuss Bayes factors for comparing two extreme models for the SAT coaching example.

(a) Derive the Bayes factor, $p(H_2|y)/p(H_1|y)$, as a function of $y_1, \ldots, y_J$, $\sigma_1^2, \ldots, \sigma_J^2$, and A, for the models with $\text{N}(0, A^2)$ prior distributions.

(b) Evaluate the Bayes factor in the limit $A \to \infty$.

(c) For fixed A, evaluate the Bayes factor as the number of schools, J, increases. Assume for simplicity that $\sigma_1^2 = \ldots = \sigma_J^2 = \sigma^2$, and that the sample mean and variance of the y_j's do not change.

10. Variety of posterior predictive distributions: for the educational testing example in Section 6.8, we considered a reference set for the posterior predictive simulations in which $\theta = (\theta_1, \ldots, \theta_8)$ was fixed. This corresponds to a replication of the study with the same eight coaching programs.

(a) Consider an alternative reference set, in which (μ, τ) are fixed but θ is allowed to vary. Define a posterior predictive distribution for y^{rep} under this replication, by analogy to (6.1). What is the experimental replication that corresponds to this reference set?

(b) Consider switching from the analysis of Section 6.8 to an analysis using this alternative reference set. Would you expect the posterior predictive p-values to be less extreme, more extreme, or stay about the same? Why?

(c) Reproduce the model checks of Section 6.8 based on this posterior predictive distribution. Compare to your speculations in part (b).

11. Cross-validation and posterior predictive checks:

(a) Discuss the relation of cross-validation (see page 190) to Bayesian posterior predictive checking. Is there a Bayesian version of cross-validation?

(b) Compare the two approaches with one of the examples considered so far in the book.

12. Model expansion: consider the Student-t model, $y_i|\mu, \sigma^2, \nu \sim t_\nu(\mu, \sigma^2)$, as a generalization of the normal. Suppose that, conditional on ν, you are willing to assign a noninformative uniform prior density on $(\mu, \log \sigma)$. Construct

what you consider a noninformative joint prior density on $(\mu, \log\sigma, \nu)$, for the range $\nu \in [1, \infty)$. Address the issues raised in setting up a prior distribution for the power-transformed normal model in Exercise 6.13 below.

13. Power-transformed normal models: A natural expansion of the family of normal distributions, for all-positive data, is through power transformations, which are used in various contexts, including regression models. For simplicity, consider univariate data $y = (y_1, \ldots, y_n)$, that we wish to model as iid normal after transformation.

    Box and Cox (1964) propose the model, $y_i^{(\phi)} \sim \mathrm{N}(\mu, \sigma^2)$, where

    $$y_i^{(\phi)} = \begin{cases} (y_i^\phi - 1)/\phi & \text{for } \phi \neq 0 \\ \log y_i & \text{for } \phi = 0. \end{cases} \tag{6.14}$$

    The parameterization in terms of $y_i^{(\phi)}$ allows a continuous family of power transformations that includes the logarithm as a special case. To perform Bayesian inference, one must set up a prior distribution for the parameters, (μ, σ, ϕ).

    (a) It seems natural to apply a prior distribution of the form $p(\mu, \log\sigma, \phi) \propto p(\phi)$, where $p(\phi)$ is a prior distribution (perhaps uniform) on ϕ alone. Unfortunately, this prior distribution leads to unreasonable results. Set up a numerical example to show why. (Hint: consider what happens when all the data points y_i are multiplied by a constant factor.)

    (b) Box and Cox (1964) propose a prior distribution that has the form $p(\mu, \sigma, \phi) \propto \dot{y}^{1-\phi} p(\phi)$, where $\dot{y} = (\prod_{i=1}^n y_i)^{1/n}$. Show that this prior distribution eliminates the problem in (a).

    (c) Write the marginal posterior density, $p(\phi|y)$, for the model in (b).

    (d) Discuss the implications of the fact that the prior distribution in (b) depends on the data.

    (e) The power transformation model is used with the understanding that negative values of $y_i^{(\phi)}$ are not possible. Discuss the effect of the implicit truncation on the model.

    See Pericchi (1981) and Hinkley and Runger (1984) for further discussion of Bayesian analysis of power transformations.

14. Fitting a power-transformed normal model: Table 6.3 gives short-term radon measurements for a sample of houses in three counties in Minnesota (see Section 22.4 for more on this example). For this problem, ignore the first-floor measurements (those indicated with asterisks in the table).

    (a) Fit the power-transformed normal model from Exercise 6.13(b) to the basement measurements in Blue Earth County.

    (b) Fit the power-transformed normal model to the basement measurements in all three counties, holding the parameter ϕ equal for all three counties but allowing the mean and variance of the normal distribution to vary.

    (c) Check the fit of the model using posterior predictive simulations.

County	Radon measurements (pCi/L)
Blue Earth	5.0, 13.0, 7.2, 6.8, 12.8, 5.8*, 9.5, 6.0, 3.8, 14.3*, 1.8, 6.9, 4.7, 9.5
Clay	0.9*, 12.9, 2.6, 3.5*, 26.6, 1.5, 13.0, 8.8, 19.5, 2.5*, 9.0, 13.1, 3.6, 6.9*
Goodhue	14.3, 6.9*, 7.6, 9.8*, 2.6, 43.5, 4.9, 3.5, 4.8, 5.6, 3.5, 3.9, 6.7

Table 6.3 *Short-term measurements of radon concentration (in picoCuries/liter) in a sample of houses in three counties in Minnesota. All measurements were recorded on the basement level of the houses, except for those indicated with asterisks, which were recorded on the first floor.*

(d) Discuss whether it would be appropriate to simply fit a lognormal model to these data.

15. Model checking: check the assumed model fitted to the rat tumor data in Section 5.3. Define some test quantities that might be of scientific interest, and compare them to their posterior predictive distributions.

16. Checking the assumption of equal variance: Figures 1.1 and 1.2 display data on point spreads x and score differentials y of a set of professional football games. (The data are available at the website for this book.) In Section 1.6, a model is fit of the form, $y \sim \mathrm{N}(x, 14^2)$. However, Figure 1.2a seems to show a pattern of decreasing variance of $y - x$ as a function of x.

(a) Simulate several replicated data sets y^{rep} under the model and, for each, create graphs like Figure 1.1 and 1.2. Display several graphs per page, and compare these to the corresponding graphs of the actual data. This is a graphical posterior predictive check as described in Section 6.4.

(b) Create a numerical summary $T(x, y)$ to capture the apparent decrease in variance of $y - x$ as a function of x. Compare this to the distribution of simulated test statistics, $T(x, y^{\mathrm{rep}})$ and compute the p-value for this posterior predictive check.