Applied Regression Analysis and Generalized Linear Models, Third Edition

Supplement: Chapter 25 (Draft)

Bayesian Estimation of Regression Models

John Fox

Last Modified: 2023-03-20

Contents

Chapter 25

Bayesian Estimation of Regression Models

There has been a recent proliferation of applications of Bayesian statistical inference in the social sciences and more generally. Progress in applied Bayesian statistics has several sources, including the development, dating to the 1950s, of methods for sampling from complex probability distributions that defy analytic solutions (see Section 25.2 on Markov-Chain Monte-Carlo simulation); great advances in the speed and memory capacity of computing hardware; and the evolution of improved computational algorithms and statistical software to render Bayesian inference flexible, convenient, and practical.

This chapter develops Bayesian statistical inference and illustrates its application to a variety of regression models, including linear models, generalized linear models, and mixed-effects models. With a couple of exceptions (Section 20.4 on Bayesian multiple imputation of missing data, and some material in Chapter 22 on model selection and model averaging), Bayesian ideas are essentially absent from preceding chapters, where statistical inference in conducted via "classical frequentist" parameter estimation, hypothesis tests, confidence intervals, and confidence regions, which are assumed to be familiar from a basic introduction to statistics. I don't assume, however, that Bayesian statistical inference is familiar, and so I introduce the topic here from first principles.

25.1 Introduction to Bayesian Statistical Inference

This section introduces Bayesian statistical inference as an alternative to classical frequentist methods. The treatment is brief, presenting and illustrating the principal ideas of Bayesian inference but not developing the topic in detail.[1]

[1] Nevertheless, this section substantially expands the introduction to Bayesian inference in Section D.7 of the on-line Appendix to the book.

25.1.1 Bayes's Theorem

Recall (from Section D.1 of on-line Appendix D on probability and estimation[2]) the definition of *conditional probability*: The probability of an event A given that another event B is known to have occurred is

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)} \tag{25.1}$$

where $\Pr(A \cap B)$ is the *joint probability* that both A and B occur, and $\Pr(B)$ is the *marginal* or *unconditional probability* of B. Likewise, the conditional probability of B given A is

$$\Pr(B|A) = \frac{\Pr(A \cap B)}{\Pr(A)} \tag{25.2}$$

Solving Equation 25.2 for the joint probability of A and B produces

$$\Pr(A \cap B) = \Pr(B|A)\Pr(A)$$

and substituting this result into Equation 25.1 yields *Bayes's theorem*:

$$\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)} \tag{25.3}$$

Bayes's theorem is named after its discoverer, the Reverend Thomas Bayes, an 18th-century English mathematician.

Bayesian statistical inference is based on the following interpretation of Equation 25.3: Let A represent some uncertain proposition (a "hypothesis") whose truth or falsity we wish to establish—for example, the proposition that a parameter is equal to a particular value. Let B represent observed data that are relevant to the truth of the proposition. We interpret the unconditional probability $\Pr(A)$, called the *prior probability* of A, as our strength of belief in the truth of A prior to collecting data, and $\Pr(B|A)$ as the probability of obtaining the observed data assuming the truth of A—that is, the *likelihood* of the data given A (see on-line Appendix Section D.6.1). The *unconditional* probability of the data B is[3]

$$\Pr(B) = \Pr(B|A)\Pr(A) + \Pr(B|\overline{A})\Pr(\overline{A})$$

where $\overline{A}$ is the event not-A (the *complement* of A). Then $\Pr(A|B)$, given by Equation 25.3 and called the *posterior probability* of A, represents our revised strength of belief in A in light of the data B.

[2]Indeed, this would be a good time to review Appendix D!

[3]This is an application of the *law of total probability*: Given an event B and a set of k disjoint events $A_1, \ldots, A_k$ for which $\sum_{i=1}^{k} \Pr(A_i) = 1$ (i.e., the events A_i partition the sample space S),

$$\Pr(B) = \sum_{i=1}^{k} \Pr(B|A_i)\Pr(A_i)$$

In the current context, $S = A \cup \overline{A}$, that is, the sample space is the union of A and $\overline{A}$.

Bayesian inference is therefore a rational procedure for updating one's beliefs on the basis of evidence. This *subjectivist* interpretation of probabilities as relative strength of belief contrasts with the *objective* or *frequentist* interpretation of probabilities as long-run proportions. Bayes's theorem follows from elementary probability theory *whether or not* one accepts its subjectivist interpretation, but it is the latter that gives rise to common procedures of Bayesian statistical inference.[4]

Named after the Reverend Thomas Bayes, an 18th-century English mathematician, Bayes's theorem, which follows from elementary probability theory, states that for events A and B,

$$\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)}$$

where $\Pr(B) = \Pr(B|A)\Pr(A) + \Pr(B|\overline{A})\Pr(\overline{A})$ is the unconditional probability of B (and $\overline{A}$ is the event not-A).
Bayesian statistical inference is based on the following interpretation of Bayes's theorem:

- Let A represent an uncertain proposition (a "hypothesis"), and let B represent observed data that are relevant to the truth of the proposition.
- $\Pr(A)$ is the *prior probability* of A, our strength of belief in A prior to collecting data.
- $\Pr(B|A)$ is the probability of obtaining the observed data assuming the truth of A—the *likelihood* of the data given A.
- $\Pr(A|B)$, the *posterior probability* of A, represents our revised strength of belief in A in light of the data B.

Bayesian inference is therefore a rational procedure for updating one's beliefs on the basis of evidence.

An Application of Bayes's Theorem

I conclude this section with a simple application of Bayes's theorem that I present as an exercise for the reader,[5] an example that nicely reinforces the

[4]The identification of "classical" statistical inference with the frequentist interpretation of probability and of Bayesian inference with subjective probability is a simplification that glosses over differences in both camps. Such subtleties are well beyond the scope of this presentation, and the shorthand—Bayesian inference versus classical or frequentist inference—is convenient.

[5]See Exercise 25.1.

point that Bayes's theorem is just a consequence of basic probability theory. The application is well known, but most people find the result surprising (and it is topical as I write this during the COVID-19 pandemic):

- Suppose that 10% of the population of a country have been infected by a disease-causing virus and have developed antibodies to it. Let A represent the event that a person selected at random from the population has *antibodies* to the virus, so $\Pr(A) = .1$ and $\Pr(\overline{A}) = .9$.

- Imagine that a test for antibodies has been developed that never produces a false negative. Let P represent the event that a person tests *positive* for antibodies. The conditional probability that a person with antibodies correctly tests positive, called the *sensitivity* of the test, is then $\Pr(P|A) = 1$.

- Imagine further that the test has a false positive rate of 10%—that is, 10% of people who don't have antibodies to the virus nevertheless test positive (perhaps because they've been infected by a similar virus). The conditional probability that a person who doesn't have antibodies incorrectly tests positive is therefore $\Pr(P|\overline{A}) = .1$.[6]

- Imagine, finally, that *you* test positive. What is the probability that you actually have antibodies to the virus—that is, what is $\Pr(A|P)$?[7] [Hint: $\Pr(A|P)$ is much smaller than $\Pr(P|A) = 1$.]

25.1.2 A Preliminary Example of Bayesian Inference

Consider the following simple (if contrived) situation: Suppose that you are given a gift of two "biased" coins, one of which produces heads with probability $\Pr(H) = .3$ and the other with $\Pr(H) = .8$. Each of these coins comes in a box marked with its bias, but you carelessly misplace the boxes and put the coins in a drawer; a year later, you forget which coin is which. To try to distinguish the coins, you pick one arbitrarily and flip it 10 times, obtaining the data $HHTHHHTTHH$ —that is, a particular sequence of 7 heads and 3 tails. (These are the "data" used in a preliminary example of maximum-likelihood estimation in on-line Appendix Section D.6.1.)

Let A represent the event that the selected coin has $\Pr(H) = .3$; then $\overline{A}$ is the event that the coin has $\Pr(H) = .8$. Under these circumstances, it seems

[6] The probability that a person who doesn't have antibodies correctly tests negative, called the *specificity* of the test, is $\Pr(\overline{P}|\overline{A}) = 1 - .1 = .9$, but this probability isn't needed for the problem.

[7] A strict frequentist would object to referring to the probability that a specific individual, like you, has antibodies to the virus because, after all, either you have antibodies or you don't. $\Pr(A|P)$ is therefore a subjective probability, reflecting your ignorance of the true state of affairs. $\Pr(A|P)$ can be given an objective frequentist interpretation as the long-run *proportion* (i.e., the relative frequency) of individuals testing positive who are actually positive.

reasonable to take as prior probabilities $\Pr(A) = \Pr(\overline{A}) = .5$. Calling the data B, the likelihood of the data under A and $\overline{A}$ is

$$\Pr(B|A) = .3^7(1-.3)^3 = .0000750$$
$$\Pr(B|\overline{A}) = .8^7(1-.8)^3 = .0016777$$

As is typically the case, the likelihood of the observed data is small in both cases, but the data are much more likely under $\overline{A}$ than under A. (The likelihood of these data for *any* value of $\Pr(H)$ between 0 and 1 appears in on-line Appendix Figure D.18.)

Using Bayes's theorem (Equation 25.3), you find the posterior probabilities

$$\Pr(A|B) = \frac{.0000750 \times .5}{.0000750 \times .5 + .0016777 \times .5} = .0428$$
$$\Pr(\overline{A}|B) = \frac{.0016777 \times .5}{.0000750 \times .5 + .0016777 \times .5} = .9572$$

suggesting that it is much more probable that the selected coin has $\Pr(H) = .8$ than $\Pr(H) = .3$.

25.1.3 Extending Bayes's Theorem

Bayes's theorem extends readily to situations in which there are more than two hypotheses A and $\overline{A}$: Let the various hypotheses be represented by $H_1, H_2, \ldots, H_k$, with prior probabilities $\Pr(H_i)$, $i = 1, \ldots, k$, that sum to 1;[8] and let D represent the observed data, with likelihood $\Pr(D|H_i)$ under hypothesis H_i. Then the posterior probability of hypothesis H_i is

$$\Pr(H_i|D) = \frac{\Pr(D|H_i)\Pr(H_i)}{\sum_{j=1}^{k}\Pr(D|H_j)\Pr(H_j)} \tag{25.4}$$

The denominator in Equation 25.4 insures that the posterior probabilities for the various hypotheses sum to 1. It is sometimes convenient to omit this normalization, simply noting that

$$\Pr(H_i|D) \propto \Pr(D|H_i)\Pr(H_i)$$

that is, that the posterior probability of a hypothesis is proportional to the product of the likelihood $\Pr(D|H_i)$ under the hypothesis and its prior probability $\Pr(H_i)$. If necessary, we can always divide by $\sum \Pr(D|H_i)\Pr(H_i)$ to recover the posterior probabilities.

Bayes's theorem is also applicable to random variables: Let α represent a parameter of interest, with prior probability distribution or density $p(\alpha)$, and let $L(\alpha) \equiv p(D|\alpha)$ represent the likelihood function for the parameter α. Then

$$p(\alpha|D) = \frac{L(\alpha)p(\alpha)}{\sum_{\text{all } \alpha'} L(\alpha')p(\alpha')}$$

[8]To employ Bayesian inference, your prior beliefs must be consistent with probability theory, and so the prior probabilities must sum to 1.

when the parameter α is discrete, or

$$p(\alpha|D) = \frac{L(\alpha)p(\alpha)}{\int_A L(\alpha')p(\alpha')\,d\alpha'}$$

when, as is more common, α is continuous (and where A represents the set of all values of α).[9] In either case,

$$p(\alpha|D) \propto L(\alpha)p(\alpha)$$

That is, the posterior probability or density is proportional to the product of the likelihood and the prior probability or density. As before, we can if necessary divide by $\sum L(\alpha)p(\alpha)$ or $\int L(\alpha)p(\alpha)\,d\alpha$ to recover the posterior probabilities or densities.[10]

Two points are noteworthy:

- We require a prior distribution $p(\alpha)$ over the possible values of the parameter α (the *parameter space*) to set the machinery of Bayesian inference in motion.

- In contrast to a frequentist statistician, a Bayesian treats the parameter α as a *random variable* rather than as an unknown *constant*. I retain Greek letters for parameters, however, because unlike the data, parameters are never known with certainty—even after collecting data.

[9] I've resisted starring this material because parameters are almost always continuous. If you're unfamiliar with calculus, just think of the integral symbol $\int$ (which is an elongated "S") as the continuous analog of a Sum $\sum$—indeed, the analogy is very close—here the area under the density function of a continuous random, which represents a probability, an idea that is surely familiar from a basic statistics course.

[10] The statement is glib, in that it may not be easy in the continuous case to evaluate the integral $\int L(\alpha)p(\alpha)\,d\alpha$. This potential difficulty motivates the use of conjugate priors, discussed immediately below, and the more generally applicable Markov-chain Monte-Carlo methods described later in the chapter (Section 25.2).

Bayes's theorem can be extended to several hypotheses $H_1, H_2, \ldots, H_k$, with prior probabilities $\Pr(H_i)$, $i = 1, \ldots, k$, that sum to 1, and observed data D with likelihood $\Pr(D|H_i)$ under hypothesis H_i; the posterior probability of hypothesis H_i is

$$\Pr(H_i|D) = \frac{\Pr(D|H_i)\Pr(H_i)}{\sum_{j=1}^{k}\Pr(D|H_j)\Pr(H_j)}$$

Similarly, Bayes's theorem is applicable to random variables such as a parameter α, with prior probability distribution or density $p(\alpha)$, and likelihood $L(\alpha) \equiv p(D|\alpha)$ for the data D. Then

$$p(\alpha|D) = \frac{L(\alpha)p(\alpha)}{\sum_{\text{all } \alpha'} L(\alpha')p(\alpha')}$$

when the parameter α is discrete, or

$$p(\alpha|D) = \frac{L(\alpha)p(\alpha)}{\int_A L(\alpha')p(\alpha')\, d\alpha'}$$

when α is continuous (and where A represents the set of all values of α).

25.1.4 Conjugate and Other Priors

The mathematics of Bayesian inference is especially simple when the prior distribution is selected so that the likelihood and prior combine to produce a posterior distribution that is in the same family as the prior (see the example in the following section). In this case, we say that the prior distribution is a *conjugate prior*.

At one time, Bayesian inference was only practical when conjugate priors were employed, radically limiting its scope of application. Advances in computer software and hardware, however, make it practical to approximate mathematically intractable posterior distributions by simulated random sampling. Such *Markov-chain Monte-Carlo* ("*MCMC*") methods (which are described in Section 25.2) have produced a flowering of Bayesian applied statistics. Nevertheless, the choice of prior distribution can be an important one.

25.1.5 Generalizing the Example of Bayesian Inference

Continuing the previous example, suppose more realistically that you are given a coin and wish to estimate the probability π that the coin turns up heads, but cannot restrict π in advance to a small number of discrete values; rather, π could, in principle, be any number between 0 and 1. To estimate π, you plan to gather data by independently flipping the coin 10 times. We know from our previous work that the likelihood (called the *Bernoulli* likelihood, after the 17th

Century Swiss mathematician Jacob Bernoulli) for this experiment is[11]

$$L(\pi) = \pi^h(1-\pi)^{10-h} \tag{25.5}$$

where h is the observed number of heads. You conduct the experiment, obtaining the data $HHTHHHTTHH$, and thus $h = 7$.

The conjugate prior for the Bernoulli likelihood in Equation 25.5 is the beta distribution (see on-line Appendix Section D.3.8),

$$\begin{aligned} p(\pi) &= \text{Beta}(a, b) \\ &= \frac{\pi^{a-1}(1-\pi)^{b-1}}{B(a,b)} \text{ for } 0 \leq \pi \leq 1 \text{ and } a, b \geq 0 \end{aligned}$$

(where the function $B(a, b)$ in the denominator is given in Appendix Section D.3.8). When you multiply the beta prior by the likelihood, you get a posterior density of the form

$$p(\pi|D) \propto \pi^{h+a-1}(1-\pi)^{10-h+b-1} = \pi^{6+a}(1-\pi)^{2+b}$$

that is, a beta distribution with shape parameters $a' = h + a = 7 + a$ and $b' = 10 - h + b = 3 + b$. Put another way, the prior in effect adds a heads and b tails to the likelihood.

How should you select a and b? One approach would be to reflect your subjective assessment of the plausible values of π. For example, you might confirm that the coin has both a head and a tail, and that it seems to be reasonably well balanced, suggesting that π is probably close to .5. Picking $a = b = 16$ would in effect confine your estimate of π to the range between .3 and .7 (see Figure 25.1, which conveys the flexibility of the beta family of prior distributions). If you are uncomfortable with this restriction, then you could select smaller values of a and b: When $a = b = 1$, all values of π are equally likely (i.e., have equal probability density)—a so-called *flat* or *uninformative prior distribution*, reflecting complete ignorance about the value of π.[12]

[11] The *Bernoulli distribution* isn't introduced in on-line Appendix D on probability and estimation, but it is the simplest case of the binomial distribution, discussed in Section D.2.1, where there is only a single binomial trial. The probability mass function of the Bernoulli distribution is therefore $p(x) = \pi^x(1-\pi)^{1-x}$, where π is the probability of "success" (i..e, a head in the coin-flipping example), and $x = 1$ for a success and 0 for a failure.

The Bernoulli likelihood function follows immediately by multiplying the probabilities for the 10 (more generally, n) independent trials. We *could* work directly with the binomial likelihood and obtain the same result, taking the data as h, the number of heads in n binomial trials, because h is a sufficient statistic for the probability π: See on-line Appendix D.5.4 for an explanation of sufficiency.

That is, given h heads in n independent flips of a coin with probability π of a head on an individual flip, the Bernoulli likelihood function is $L_{\text{Bern}}(\pi) = \pi^h(1-\pi)^{n-h}$, while the binomial likelihood is $L_{\text{binom}}(\pi) = \binom{h}{n}\pi^h(1-\pi)^{n-h}$, where $\binom{h}{n} \equiv \frac{n!}{h!(n-h)!}$ is the binomial coefficient. Because n and h are both constants after the data are observed, the binomial likelihood is proportional to the Bernoulli likelihood. The Bernoulli likelihood takes the *order* of the h heads and $n - h$ tails into account while the binomial likelihood ignores the order and only attends to the *number* of heads; the order is immaterial to estimating the value of π because the flips are independent.

[12] But see the discussion of uninformative priors in the following section.

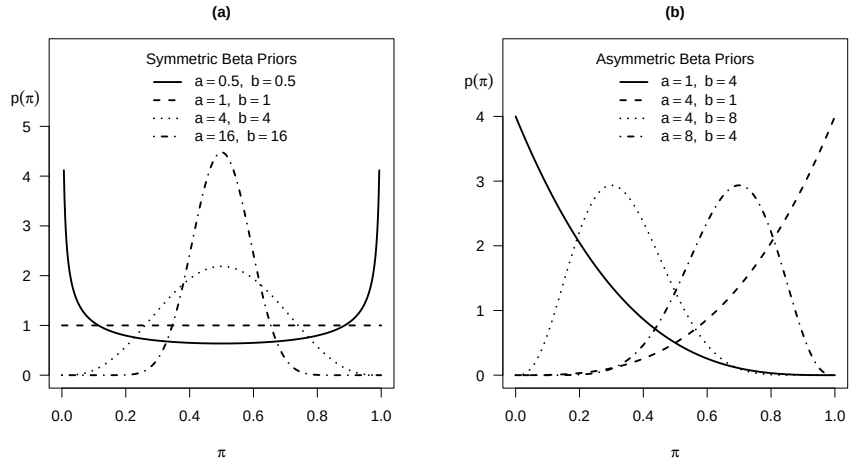

Figure 25.1 Beta priors for various choices of a and b: (a) symmetric priors, for which $a = b$; (b) asymmetric priors, for which $a \neq b$. Beta(0.5, 0.5) is the Jeffreys prior, discussed in Section 25.1.6.

Figure 25.2 shows the posterior distributions for π under these two priors. Under the flat prior, the posterior is proportional to the likelihood, and therefore if you take the mode of the posterior as your estimate of π, you get the maximum-likelihood estimate $\widehat{\pi} = .7$.[13] The posterior for the *informative prior* $a = b = 16$, in contrast, has a mode at $\pi \approx .55$, which is much closer to the mode of the prior distribution $\pi = .5$.

It may be disconcerting that the conclusion should depend so crucially on the prior distribution, but this result is a consequence of the very small sample (10 coin flips) in the example: Recall that using a beta prior in this case is like adding $a + b$ observations to the data. As the sample size grows, the likelihood comes to dominate the posterior distribution, and the influence of the prior distribution fades.[14] In the current example, if the coin is flipped n times, then the posterior distribution takes the form

$$p(\pi|D) \propto \pi^{h+a-1}(1-\pi)^{n-h+b-1}$$

and the numbers of heads h and tails $n - h$ will grow with the number of flips. It is intuitively sensible that your prior beliefs should carry greater weight when the sample is small than when it is large.

[13] An alternative is to take the mean or median of the posterior distribution as a point estimate of π. In most cases, however, the posterior distribution will approach a normal distribution as the sample size increases, and the posterior mean, median, and mode will therefore be approximately equal if the sample size is sufficiently large.

[14] An exception to this rule occurs when the prior distribution assigns 0 density to some values of the parameter; such values will necessarily have posterior densities of 0 regardless of the data.

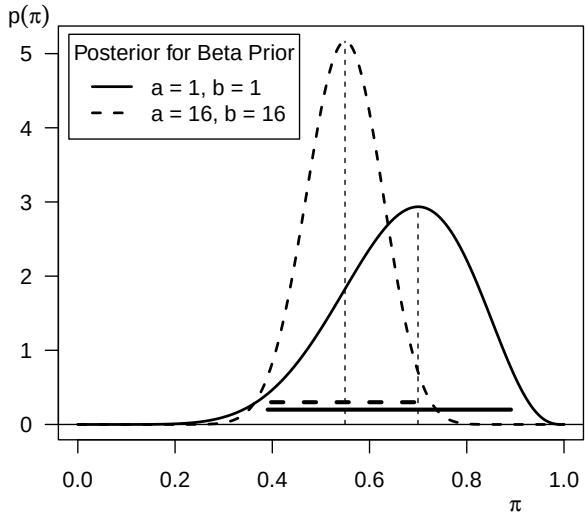

Figure 25.2 Posterior distributions for the probability of a head π under two prior distributions: the flat beta prior with $a = 1$, $b = 1$ (the posterior for which is shown as a solid curve), and the informative beta prior with $a = 16, b = 16$ (the broken curve). The data contain 7 heads in 10 flips of a coin. The two horizontal lines near the bottom of the graph show 95% central posterior intervals (described in Section 25.1.7) corresponding to the two priors.

Bayesian inference is simplest with a *conjugate prior distribution*, which combines with the likelihood to produce a posterior distribution in the same family as the prior. For example, if h counts the number of heads in n independent flips of a coin with probability π of obtaining on a head on an individual flip, then the Bernoulli likelihood for the data is $L(\pi) = \pi^h(1-\pi)^{n-h}$. Combining this likelihood with the prior distribution $p(\pi) = \text{Beta}(a, b)$ produces a posterior distribution in the same family as the prior, $p(\pi|D) = \text{Beta}(h + a, n - h + b)$.

25.1.6 More on Uninformative Prior Distributions

In estimating a probability π from a sequence of Bernoulli trials, as in the previous section, the flat prior is the rectangular density function $p(\pi) = 1$, with the parameter π bounded between 0 and 1. In other cases, such as estimating the mean μ of a normal distribution, which is unbounded, a flat prior of the form $p(\mu) = c$ (for any positive constant c) over $-\infty < \mu < \infty$ does not enclose a finite probability, and hence cannot represent a density function. When combined with the likelihood, such an *improper prior* can nevertheless lead to a *proper posterior distribution*—that is, to a posterior density that integrates to 1.

A more subtle point is that a flat prior for one parametrization of a probability model for the data need not be flat for an alternative parametrization. For example, suppose that rather than the probability π of a head in the preceding example, you take the odds $\omega \equiv \pi/(1-\pi)$ as the parameter of interest, or the logit (i.e., log-odds) $\Lambda \equiv \log_e[\pi/(1-\pi)]$; a flat prior for ω or for Λ is not flat for π (see Figure 25.3), and so will lead to a different posterior estimate of π. This lack of invariance contrasts with inference based purely on the likelihood, where, within broad conditions, the maximum-likelihood estimate $\widehat{f(\theta)}$ of a transformation $f(\theta)$ of the parameter θ is simply $f(\widehat{\theta})$, where $\widehat{\theta}$ is the MLE of θ.

A related question is whether there exists a prior distribution that *is* invariant with respect to transformation of the parameter of interest. This question was answered in the affirmative by Sir Harold Jeffreys (1946). In the case of estimating a probability π, the *Jeffreys prior* takes the form $p_J(\pi) = 1/\left[\Pi\sqrt{\pi(1-\pi)}\right]$ (see Figure 25.3).[15] The notation is a bit awkward here: π is the probability of success, and so I use an uppercase $\Pi \approx 3.14159$ for the mathematical constant, whose role here is to ensure that the prior density integrates to 1. The Jeffreys prior $p_J(\pi)$ is a beta distribution with $a = b = 0.5$ (shown in Figure 25.1) and so is a conjugate prior to the Bernoulli likelihood.

[15]The general solution for the Jeffreys prior turns out to be remarkably simple, and is explored in Exercise 25.4. Although I don't take it up here, the Jeffreys prior can be extended to several parameters. There are, as well, other general uninformative priors based on various principles, such as *reference priors*, introduced by Bernardo (1979).

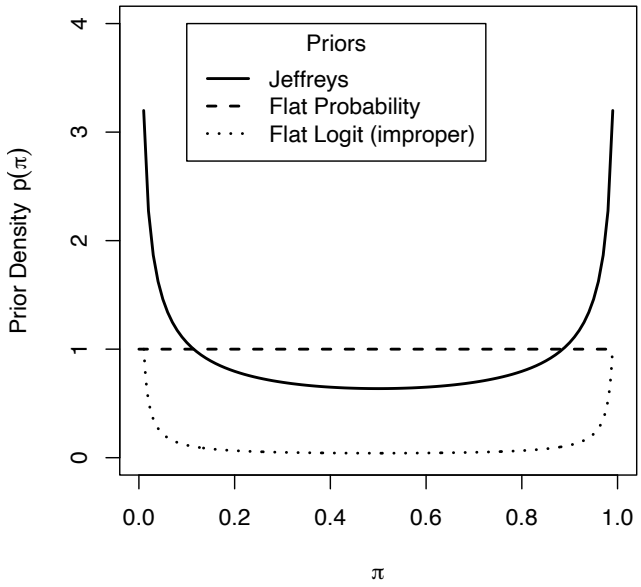

Figure 25.3 Three putatively uninformative priors expressed on the probability scale π: The flat prior on the probability scale (broken line); the flat prior on the logit scale (dotted line); and the Jeffreys prior (solid line). Because the flat prior on the logit scale is improper, its height in the figure is essentially arbitrary. A question for the reader: Are you comfortable with the flat logit prior or the Jeffreys prior as an expression of weak prior belief?

Finally a note on terminology: I haven't differentiated clearly between "flat" and "uninformative" priors, using the terms essentially interchangeably, but the latter is really a more general idea than the former, and, as we've seen, the notion of a flat prior doesn't entirely hold up under close scrutiny. Synonyms, or near synonyms, for uninformative priors, are *non-informative priors*, *vague priors*, *weak priors* and *diffuse priors*. So-called *weakly informative priors* are often employed in practice, and are selected to place broad plausible constraints on the value of a parameter.

There are several kinds of *uninformative prior distributions*. The *flat prior* asigns equal probability density to all values of a parameter; if the parameter is unbounded, then the flat prior is *improper*, in that it doesn't integrate to 1, and a flat prior for a parameter is not in general flat for a transformation of the parameter. The *Jeffreys prior* for a parameter (introduced by Sir Harold Jeffreys), in contrast, is invariant with respect to transformation of the parameter. *Weakly informative priors* are often employed in practice, and are selected to place broad plausible constraints on the value of a parameter.

25.1.7 Bayesian Interval Estimates

As in classical frequentist statistical inference, it is desirable not only to provide a point estimate of a parameter but also to quantify uncertainty in the estimate. The posterior distribution of the parameter displays statistical uncertainty in a direct form, and the standard deviation of the posterior distribution is a Bayesian analog of the frequentist standard error of an estimate. One can also compute various kinds of Bayesian interval estimates (termed *credible intervals* and analogous to frequentist confidence intervals) from the posterior distribution.

A very simple choice of Bayesian interval estimate is the *central posterior interval*: The $100a\%$ central posterior interval runs from the $(1 - a)/2$ to the $(1 + a)/2$ quantile of the posterior distribution; for example, the 95% central posterior interval runs from the .025 to the .975 quantile of the posterior. Unlike a classical confidence interval, however, the interpretation of which is famously convoluted (to the confusion of innumerable students of basic statistics), a Bayesian credible interval has a simple interpretation as a probability statement: The probability is .95 that the parameter is in the 95% posterior interval. This difference reflects the Bayesian treatment of a parameter as a random variable, with the posterior distribution expressing subjective uncertainty about the value of the parameter after observing the data.

Ninety-five percent central posterior intervals for the example are shown for the two posterior distributions in Figure 25.2 (page 10).[16]

> Bayesian interval estimates, termed *credible intervals* (analogous to frequentist confidence intervals), are computed from the posterior distribution of a parameter. The $100a\%$ central posterior interval runs from the $(1-a)/2$ to the $(1+a)/2$ quantile of the posterior distribution. A Bayesian credible interval has a simple interpretation as a probability statement: For example, the probability is .95 that the parameter is in the 95% posterior interval.

25.1.8 Bayesian Inference for Several Parameters

Bayesian inference extends straightforwardly to the simultaneous estimation of several parameters $\boldsymbol{\alpha} \equiv [\alpha_1, \alpha_2, \ldots, \alpha_k]'$.[17] In this case, it is necessary to specify a *joint prior distribution* for the parameters, $p(\boldsymbol{\alpha})$,[18] along with the *joint likelihood*, $L(\boldsymbol{\alpha})$. Then, as in the case of one parameter, the *joint posterior distribution* is proportional to the product of the prior distribution and the likelihood,

$$p(\boldsymbol{\alpha}|D) \propto p(\boldsymbol{\alpha})L(\boldsymbol{\alpha}) \tag{25.6}$$

or

$$p(\boldsymbol{\alpha}|D) = \frac{p(\boldsymbol{\alpha})L(\boldsymbol{\alpha})}{\int_{\mathbf{A}} p(\boldsymbol{\alpha}^*)L(\boldsymbol{\alpha}^*)d^k\boldsymbol{\alpha}^*} \tag{25.7}$$

where $\mathbf{A}$ is the set of all values of the parameter vector $\boldsymbol{\alpha}$ (i.e., $\mathbf{A}$ is the multidimensional parameter space). Inference typically focuses on the *marginal posterior distribution* of each parameter, $p(\alpha_i|D)$, or possibly on the distributions of scalar functions of the parameters, such as predicted values in regression, which are functions of regression coefficients, or the ratio or difference of two parameters.[19]

---

[16] Page references in this chapter may be to pages within the chapter or to pages in the printed text (i.e., Chapters 1 through 24). The references should, however, be clear from the context: In this case, for example, Figure 25.2 is obviously in the current chapter.

[17] If you're not familiar with vector notation, just think of $\boldsymbol{\alpha} \equiv [\alpha_1, \alpha_2, \ldots, \alpha_k]'$ as a collection of several parameters.

[18] It's frequently the case in practice that prior distributions are specified *independently* for the various parameters, so that the joint prior is the product of the separate (marginal) priors.

[19] Although there may be several parameters, a *scalar function* of the parameters returns a single number.

Bayesian inference extends to the simultaneous estimation of several parameters $\boldsymbol{\alpha} \equiv [\alpha_1, \alpha_2, \ldots, \alpha_k]'$. Given the joint prior distribution for the parameters $p(\boldsymbol{\alpha})$ along with the joint likelihood $L(\boldsymbol{\alpha})$ based on data D, the posterior distribution of $\boldsymbol{\alpha}$ is

$$p(\boldsymbol{\alpha}|D) = \frac{p(\boldsymbol{\alpha})L(\boldsymbol{\alpha})}{\int_{\mathbf{A}} p(\boldsymbol{\alpha}^*)L(\boldsymbol{\alpha}^*)d^k\boldsymbol{\alpha}^*}$$

where $\mathbf{A}$ is the set of all values of the parameter vector $\boldsymbol{\alpha}$ (i.e., the multidimensional parameter space).

25.2 Markov-Chain Monte Carlo Simulation

To find $p(\boldsymbol{\alpha}|D)$ explicitly is simple for a conjugate prior. More generally, however, we must integrate over all values $\mathbf{A}$ of $\boldsymbol{\alpha}$, and the integral in the denominator of Equation 25.7 is usually intractable analytically. *Markov-chain Monte Carlo* (*MCMC*) is a set of methods for drawing random samples from—and hence approximating—the posterior distribution $p(\boldsymbol{\alpha}|D)$ without having explicitly to evaluate the denominator of Equation 25.7. MCMC methods, coupled with the increasing power of computer hardware, have rendered Bayesian inference practical for a broad range of statistical problems.

There are three common (and related) MCMC methods: the Metropolis-Hastings algorithm, the Gibbs sampler, and Hamiltonian Monte Carlo:

- What's come to be called the *Metropolis-Hastings algorithm* was originally formulated by Metropolis et al. (1953) and subsequently generalized by Hastings (1970). I'll explain the more general version of the algorithm, but will use the original, simpler version in an initial example and in an application to Bayesian inference.

- The *Gibbs sampler* is an MCMC algorithm developed for applications in image processing by Geman and Geman (1984), who named it after the American physicist Josiah Gibbs (1839–1903). Gelfand and Smith (1990) pointed out the applicability of the Gibbs sampler to statistical problems. The Gibbs sampler is based on the observation that the joint distribution of an n-dimensional vector random variable $\mathbf{x}$ can be composed from the conditional distribution of each of its elements given the others, that is $p(X_j|\mathbf{x}_{-j})$ for $j = 1, 2, \ldots, n$ (where $\mathbf{x}_{-j} = [X_1, X_2, \ldots, X_{j-1}, X_{j+1}, \ldots, X_n]'$ is $\mathbf{x}$ with the jth element removed). Although it was developed independently, in this basic form the Gibbs sampler turns out to be a special case of the general Metropolis-Hastings algorithm (see Gelman et al., 2013, page 281). The popular Bayesian statistical software BUGS (Lunn et al., 2009) is based on the Gibbs sampler, and indeed its name is an acronym for Bayesian inference Using Gibbs Sampling.

- *Hamiltonian Monte Carlo* (*HMC*), introduced to statistics by Neal (1996), is an improvement to the Metropolis-Hastings algorithm that, when properly tuned, provides more efficient sampling from a target distribution. Hamiltonian Monte Carlo is named after the Irish mathematician and physicist William Rowan Hamilton (1805–1865), who reformulated the mathematics of classical Newtonian mechanics. HMC exploits an analogy between exploring the surface of a probability density function and the motion of an object along a frictionless surface, propelled by its initial momentum and gravity. HMC is considered the best current method of MCMC for sampling from continuous distributions, and is the basis for the state-of-the-art Stan Bayesian software (Carpenter et al., 2017).[20]

Markov-chain Monte Carlo (*MCMC*) is a set of methods for drawing random samples from—and hence approximating—the posterior distribution

$$p(\boldsymbol{\alpha}|D) = \frac{p(\boldsymbol{\alpha})L(\boldsymbol{\alpha})}{\int_{\mathbf{A}} p(\boldsymbol{\alpha}^*)L(\boldsymbol{\alpha}^*)d^k\boldsymbol{\alpha}^*}$$

without having explicitly to evaluate the integral in the denominator, which is typically intractible analytically. MCMC methods have therefore rendered Bayesian inference practical for a broad range of statistical problems.
There are three common (and related) MCMC methods: the *Metropolis-Hastings algorithm*, the *Gibbs sampler*, and *Hamiltonian Monte Carlo* (*HMC*). HMC is considered the best current method of MCMC for sampling from continuous distributions.

25.2.1 The Metropolis-Hastings Algorithm*

Here's the problem that the Metropolis-Hastings algorithm addresses: We have a continuous vector random variable $\mathbf{x}$ with n elements and with density function $p(\mathbf{x})$. We don't know how to compute $p(\mathbf{x})$, but we do have a function proportional to it, $p^*(\mathbf{x}) = c \times p(\mathbf{x})$, where $c = \int_{\mathbf{X}} p^*(\mathbf{x})d^n\mathbf{x}$. We don't know the *normalizing constant* c, which makes $p(\mathbf{x})$ integrate to 1, or we'd know $p(\mathbf{x})$. We nevertheless want to draw a random sample from the *target distribution* $p(\mathbf{x})$. One way this situation might arise is in Bayesian inference, where $p^*(\cdot)$ could be the *unnormalized posterior*, computed (as in Expression 25.6 on page 14) as the product of the prior density and the likelihood.

The Metropolis-Hastings algorithm starts with an arbitrary value $\mathbf{x}_0$ of $\mathbf{x}$, and proceeds to generate a sequence of m realized values $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{i-1}, \mathbf{x}_i, \ldots, \mathbf{x}_m$.

[20] Stan is named after Stanislaw Ulam, a mathematician and physicist who invented Monte-Carlo simulation in the 1940s.

Each subsequent realization is selected randomly based on a *candidate* or *proposal distribution*, with conditional density function $f(\mathbf{x}_i|\mathbf{x}_{i-1})$, from which we know how to sample. The proposal distribution $f(\cdot)$ is generally distinct from the target distribution $p(\cdot)$.

As the notation implies, the proposal distribution employed depends only on the immediately preceding value of $\mathbf{x}$. The next value of $\mathbf{x}$ sampled from the proposal distribution may be accepted or rejected, hence the term "proposal" or "candidate." If the proposed value of $\mathbf{x}_i$ is rejected, then the preceding value is retained; that is, $\mathbf{x}_i$ is set to $\mathbf{x}_{i-1}$. This procedure—where the probability of transition from one "state" to another (i.e., one value of $\mathbf{x}$ to the next) depends only on the previous state—defines a *Markov process*,[21] yielding a *Markov chain* of sampled values.

Within broad conditions, the choice of proposal distribution is arbitrary. For example, it's necessary that the proposal distribution and initial value $\mathbf{x}_0$ lead to a Markov process capable of visiting the complete *support* of $\mathbf{x}$—that is, all values of $\mathbf{x}$ for which the density $p(\mathbf{x})$ in nonzero. And different choices of proposal distributions may be differentially desirable, for example, in the sense that they are more or less efficient—that is, tend to require respectively fewer or more generated values to cover the support of $\mathbf{x}$ thoroughly.

With this background, the Metropolis-Hastings algorithm proceeds as follows. For each $i = 1, 2, \ldots, m$:

1. Sample a candidate value $\mathbf{x}^*$ from the proposal distribution $f(\mathbf{x}_i|\mathbf{x}_{i-1})$.

2. Compute the *acceptance ratio*

$$\begin{aligned} a &= \frac{p(\mathbf{x}^*)f(\mathbf{x}^*|\mathbf{x}_{i-1})}{p(\mathbf{x}_{i-1})f(\mathbf{x}_{i-1}|\mathbf{x}^*)} \\ &= \frac{p^*(\mathbf{x}^*)f(\mathbf{x}^*|\mathbf{x}_{i-1})}{p^*(\mathbf{x}_{i-1})f(\mathbf{x}_{i-1}|\mathbf{x}^*)} \end{aligned} \tag{25.8}$$

   The substitution of $p^*(\cdot)$ for $p(\cdot)$ in the second line of Equation 25.8 is justified because the unknown normalizing constant c (recall, the number that makes the density integrate to 1) cancels in the numerator and denominator, making the *ratio* in the equation computable even though the numerator and denominator in the first line of the equation are not separately computable. Calculate $a' = \min(a, 1)$.

3. Generate a uniform random number u on the unit interval, $U \sim \text{Unif}(0, 1)$.[22] If $u \leq a'$, set the ith value in the chain to the proposal, $\mathbf{x}_i = \mathbf{x}^*$; otherwise retain the previous value, $\mathbf{x}_i = \mathbf{x}_{i-1}$. In effect, the proposal is accepted with certainty if it is "at least as probable" as the preceding value, taking

[21] Named after the Russian mathematician Andrey Andreevich Markov (1856–1922).

[22] The *uniform distribution* in introduced in on-line Appendix Section D.1.2 as the *rectangular distribution*. The notation $X \sim \text{Unif}(a, b)$ means that the random variable X is uniformly (or rectangularly) distributed with a minimum value of a and a maximum of b. Then $p(x) = 1/(b - a)$ for $a \leq x \leq b$ and 0 otherwise.

into account the possible bias in the direction of movement of the proposal function from the preceding value. If the proposal is less probable than the preceding value, then the probability of accepting the proposal declines with the ratio a, but isn't 0. Thus, the chain will tend to visit higher-density regions of the target distribution with greater frequency but will still explore the entire target distribution. It can be shown (e.g., Chib and Greenberg, 1995) that the *limiting distribution* of the Markov chain (the distribution to which the sample tends as $m \to \infty$) is indeed the target distribution, and so the algorithm should work if m is big enough.

The Metropolis-Hastings algorithm is simpler when the proposal distribution is symmetric, in the sense that $f(\mathbf{x}_i|\mathbf{x}_{i-1}) = f(\mathbf{x}_{i-1}|\mathbf{x}_i)$. This is true, for example, when the proposal distribution is multivariate-normal (see on-line Appendix Section D.3.5) with mean vector $\mathbf{x}_{i-1}$ and some specified covariance matrix $\mathbf{S}$:

$$\begin{aligned} f(\mathbf{x}_i|\mathbf{x}_{i-1}) &= \frac{1}{(2\pi)^{n/2}\sqrt{\det \mathbf{S}}} \times \exp\left[-\frac{1}{2}(\mathbf{x}_i - \mathbf{x}_{i-1})'\mathbf{S}^{-1}(\mathbf{x}_i - \mathbf{x}_{i-1})\right] \\ &= f(\mathbf{x}_{i-1}|\mathbf{x}_i) \end{aligned} \tag{25.9}$$

Then, a in Equation 25.8 becomes

$$a = \frac{p^*(\mathbf{x}^*)}{p^*(\mathbf{x}_{i-1})} \tag{25.10}$$

which (again, because the missing normalizing constant c cancels) is equivalent to the ratio of the target density at the proposed and preceding values of $\mathbf{x}$. This simplified version of the Metropolis-Hastings algorithm, based on a symmetric proposal distribution, is the version originally introduced by Metropolis et al. (1953).

By construction, the Metropolis-Hastings algorithm generates statistically *dependent* successive values of $\mathbf{x}$. If an approximately independent sample is desired, then the sequence of sampled values can be *thinned* by discarding a sufficient number of intermediate values of $\mathbf{x}$, retaining only every kth value. Additionally, because of an unfortunately selected initial value $\mathbf{x}_0$, it may take some time for the sampled sequence to approach its limiting distribution—that is, the target distribution. It may therefore be advantageous to discard a number of values at the beginning of the sequence, termed the *burn-in* or *warm-up period*.

Example: Sampling from the Bivariate-Normal Distribution

I'll demonstrate the Metropolis algorithm by sampling from a bivariate-normal distribution (introduced in on-line Appendix Section D.3.5) with the following (arbitrary) mean vector and covariance matrix:

$$\begin{aligned} \boldsymbol{\mu} &= [1, 2]' \\ \boldsymbol{\Sigma} &= \begin{bmatrix} 1 & 1 \\ 1 & 4 \end{bmatrix} \end{aligned} \tag{25.11}$$

It's not necessary to use MCMC in this case, because it's easy to approximate bivariate-normal probabilities or to draw samples from the distribution directly, but the bivariate-normal distribution provides a simple setting in which to demonstrate the Metropolis algorithm, and for pedagogical purposes it helps to know the right answer in advance—that is, the example is selected for its transparency.

Let's pretend that we know the bivariate-normal distribution only up to a constant of proportionality. To this end, I omit the normalizing constant, which for this simple example works out to $2\pi \times \sqrt{3}$ (see on-line Appendix Section D.3.5).

To illustrate that the proposal distribution and the target distribution are distinct, I use a bivariate-rectangular proposal distribution centered at the preceding value $\mathbf{x}_{i-1}$ with half-extent $\delta_1 = 2$ in the direction of the coordinate x_1 and $\delta_2 = 4$ in the direction of x_2, reflecting the relative sizes of the standard deviations of the two variables. This proposal distribution is symmetric, as required by the simpler Metropolis algorithm. Clearly, because it has finite support, the rectangular proposal distribution doesn't cover the entire support of the bivariate-normal target distribution, which extends infinitely in each direction, but because the proposal distribution "travels" (i.e., moves in the $\{x_1, x_2\}$ plane) with $\mathbf{x}_i$, it can generate a valid Markov chain.

I arbitrarily set $\mathbf{x}_0 = [0, 0]'$, and sampled $m = 10^5$ values of $\mathbf{x}$. As it turned out, 41.7% of proposals were accepted. To get a sense of how Metropolis sampling proceeds, I show the first 50 accepted proposals, along with the duplicated points corresponding to rejected proposals (of which there are 75), in Figure 25.4. The 95% concentration ellipse for the bivariate-normal distribution is also shown on the graph.[23]

How well does the Metropolis algorithm approximate the bivariate-normal distribution? Here are the mean vector and covariance matrix of the sampled points, which are quite close to the corresponding parameters in Equations 25.11:

$$\begin{aligned} \widehat{\boldsymbol{\mu}} &= [1.003, 1.987]' \\ \widehat{\boldsymbol{\Sigma}} &= \begin{bmatrix} 0.989 & 0.972 \\ 0.972 & 3.963 \end{bmatrix} \end{aligned}$$

Figure 25.5 shows all of the 10^5 sampled points together with several theoretical elliptical contours of constant density and corresponding empirical density contours.[24] Clearly, the Metropolis algorithm does a good job of recovering the target bivariate-normal density.

As mentioned, the Metropolis algorithm doesn't produce an independent random sample from the target distribution. One way to measure the dependence among the sampled values is to compute their *autocorrelations*. Focus, for

[23] As explained in Section 9.4.4, density contours of the multivariate-normal distribution are ellipsoidal—that is, ellipses in the bivariate case.

[24] The empirical density contours are computed by a two-dimensional *kernel-density estimator*; see Silverman (1986, Chap. 4). One-dimensional kernel-density estimation is introduced in Section 3.1.2.

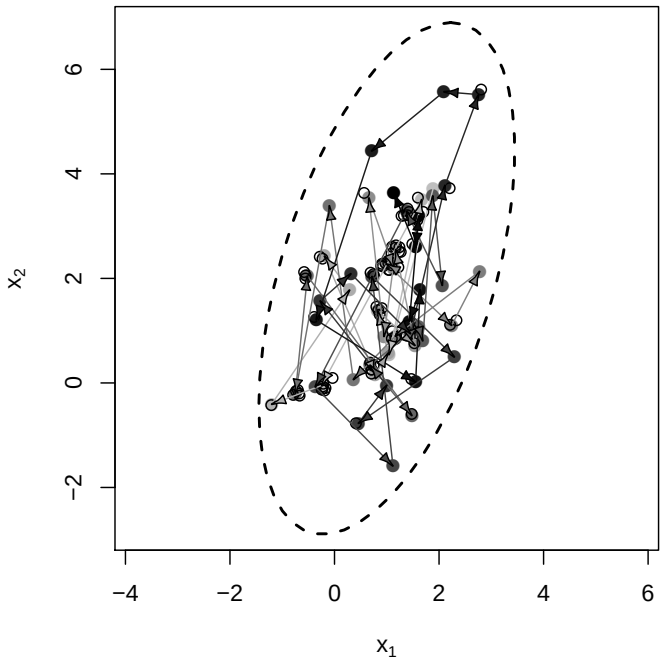

Figure 25.4 First 50 accepted proposals and duplicated points representing rejected proposals, sampling from the bivariate-normal distribution in Equations 25.11, which is represented by its 95% concentration ellipse. The solid dots represent the 50 distinct points, corresponding to accepted proposals, starting out as light gray and getting progressively darker, with the arrows showing the transitions. Duplicated points corresponding to rejected proposals, shown as hollow dots, are slightly offset so that they don't precisely over-plot the accepted proposals.

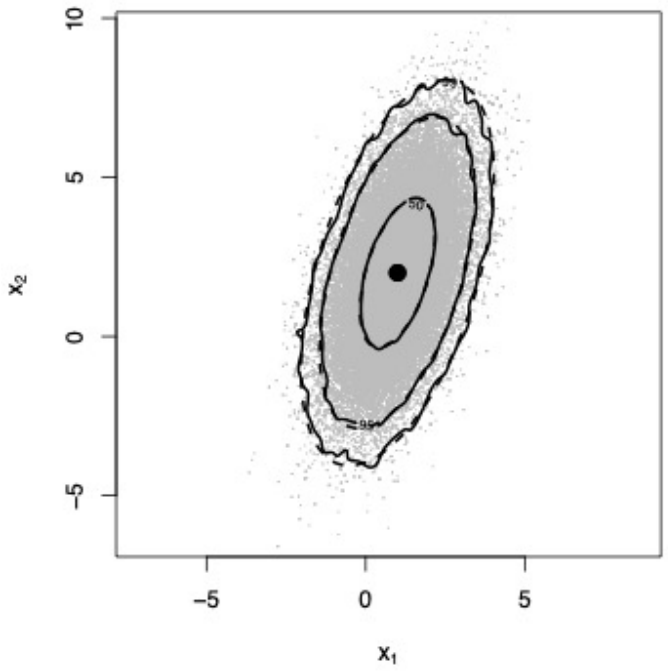

Figure 25.5 The gray dots show $m = 10^5$ values sampled by the Metropolis algorithm from the bivariate-normal distribution in Equations 25.11. The slightly irregular solid lines represent estimated density contours enclosing 50%, 95%, and 99% of the sampled points. The broken lines are the corresponding elliptical density contours of the bivariate-normal target distribution.

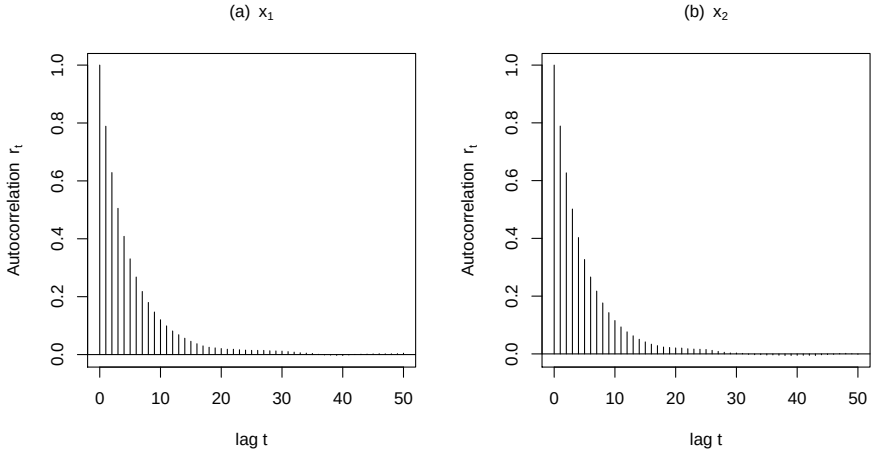

Figure 25.6 Autocorrelations of the sampled values of (a) x_1 and (b) x_2 produced by the Metropolis algorithm applied to the bivariate-normal distribution in Equations 25.11.

example, on the vector of jth sampled coordinates, say $\mathbf{x}_j = [x_{1j}, x_{2j}, \ldots, x_{mj}]'$, with mean $\bar{x}_j$. The sample autocorrelation at *lag* t is defined as[25]

$$r_{tj} = \frac{\sum_{i=t+1}^{m}(x_{ij} - \bar{x}_j)(x_{i-t,j} - \bar{x}_j)}{\sum_{i=1}^{m}(x_{ij} - \bar{x}_j)^2} \tag{25.12}$$

Figure 25.6 shows autocorrelations at lags $t = 0, 1, \ldots, 50$, for the coordinates x_1 and x_2 in the example (where the autocorrelation at lag 0, r_0, is necessarily 1). The autocorrelations are large at small lags, but decay to near 0 by around lag 25, which suggests thinning by selecting, say, every 25th value to produce an approximately independent sample

A Simple Application to Bayesian Inference

I'll illustrate the application of the Metropolis algorithm to Bayesian inference by considering a simple and familiar single-parameter problem: estimating a probability (or population proportion) π, a problem that we previously encountered in this chapter (in Section 25.1.5).

To recapitulate briefly, the likelihood for this problem comes from the Bernoulli distribution. As before, I'll use a prior distribution from the beta family, the conjugate prior to the Bernoulli likelihood. The posterior distribution is also beta, and it's therefore not necessary to approximate it by MCMC. Doing so,

[25] See Chapter 16 on time-series regression for more on autocorrelations.

however, allows us to compare the results of MCMC with the known right answer.

Recall our coin-flipping experiment, which produced $h = 7$ heads in $n = 10$ independent flips of a coin, with Bernoulli likelihood $L(\pi|h = 7) = \pi^h(1 - \pi)^{n-h} = \pi^7(1 - \pi)^3$. As in Section 25.1.5, I'll consider two prior distributions: a flat prior, in which the parameters of the Beta(a, b) distribution (see on-line Appendix Section D.3.8) are set to $a = b = 1$, and an informative prior centered on the population proportion $\pi = 0.5$ (representing a "fair" coin) in which $a = b = 16$. In the first case, the posterior is Beta$(8, 4)$, and in the second case, it is Beta$(23, 19)$. These posteriors appear in Figure 25.2 (page 10).

For the first simulation—with a flat prior—I set the standard deviation of the normal proposal distribution $N(\pi_{i-1}, s^2)$ in the Metropolis algorithm to $s = 0.1$, starting arbitrarily from $\pi_0 = 0.5$ and sampling $m = 10^5$ values from the posterior distribution of π, with an acceptance rate of 77.5%. An estimate of the resulting posterior density function is shown in panel (a) of Figure 25.7, along with the true Beta$(8, 4)$ posterior density; panel(b) shows a *quantile-comparison* (*QQ*) *plot* of the sampled values versus the Beta$(8, 4)$ distribution: If the values were sampled from Beta$(8, 4)$, then the points would lie approximately on a 45° straight line (shown on the QQ plot), within the bounds of sampling error.[26]

The agreement between the approximate posterior produced by the Metropolis algorithm and the true posterior distribution is very good, except at the extreme left of the distribution, where the sampled values are slightly shorter-tailed than the Beta$(8, 4)$ distribution. The results for the second simulation, employing the informative Beta$(16, 16)$ prior, for which the true posterior is Beta$(23, 19)$ (shown in panels (c) and (d) of Figure 25.7), are similarly encouraging. The acceptance rate for the Metropolis algorithm in the second simulation was 63.2%. In both cases (but particularly with the flat prior), the Metropolis samples of π are highly autocorrelated and would require thinning to produce an approximately independent sample; see Figure 25.8.

In the first case, using the flat Beta$(1, 1)$ prior, an estimate of π based on the median of the true Beta$(8, 4)$ posterior distribution is $\widehat{\pi} = 0.676$, and the 95% Bayesian credible interval for π from the 0.025 and 0.975 quantiles of the posterior is $0.390 < \pi < 0.891$. In comparison, using the median and 0.025 and 0.975 quantiles of the Metropolis sample, we have $\widehat{\pi} = 0.677$ and $0.392 < \pi < 0.891$.

The analogous results for the second case, with the informative Beta$(16, 16)$ prior, are $\widehat{\pi} = 0.548$ and $0.397 < \pi < 0.693$ based on the true Beta$(23, 19)$ posterior, and $\widehat{\pi} = 0.549$ and $0.396 < \pi < 0.692$ based on the Metropolis sample.

Finally, returning to the flat prior, Figure 25.9 demonstrates how making the standard deviation of the proposal distribution larger (it's set initially, recall, to 0.1) can, up to a point, decrease the autocorrelation of the sampled values, reducing the need for thinning. In this example, setting the standard deviation of the proposal distribution to 0.4 produces lower autocorrelation in the sampled

[26] See Section 3.1.3 for an explanation of QQ plots.

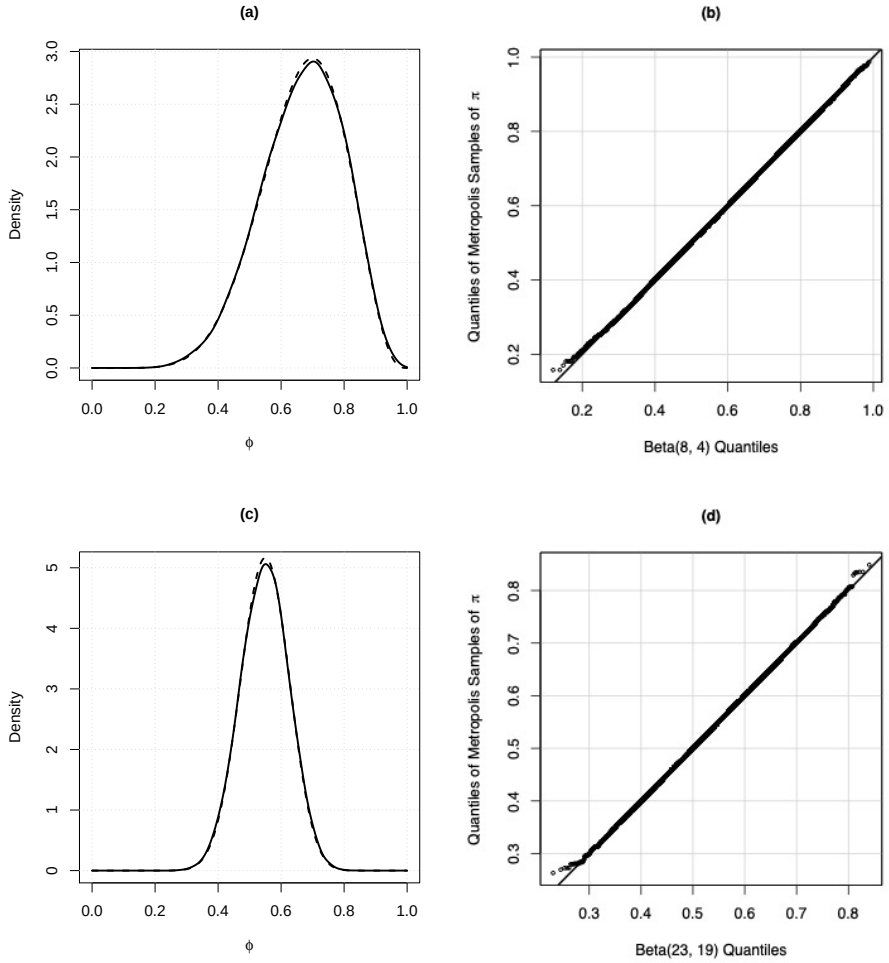

Figure 25.7 Comparing the results produced by the Metropolis algorithm to the true posterior distribution of the population proportion of heads π, based on an independent sample of size $n = 10$ with $h = 7$ heads, and a prior distribution in the conjugate beta family. For panels (a) and (b), the flat prior $\text{Beta}(1, 1)$ was used, producing the true posterior $\text{Beta}(8, 4)$; in panels (c) and (d), the informative prior $\text{Beta}(16, 16)$ was used, producing the true posterior $\text{Beta}(23, 19)$. Panels (a) and (c) show nonparametric density estimates (solid lines) for the Metropolis samples, comparing these to the true posterior densities (broken lines); panels (b) and (d) are quantile-comparison plots for the Metropolis samples versus the true posterior distributions.

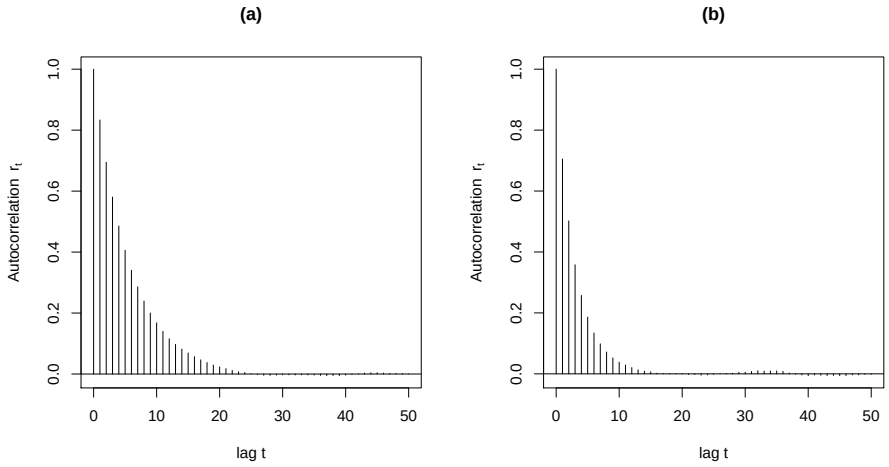

Figure 25.8 Autocorrelations of the Metropolis samples from the two posterior distributions: (a) based on the flat Beta(1, 1) prior, and (b) based on the informative Beta(16, 16) prior.

values than do standard deviations of 0.1 or 0.7.

The acceptance rate of proposals in this example declines, however, as the standard deviation of the proposal distribution increases:

SD of Proposal Distribution	Acceptance Ratio
0.1	0.775
0.4	0.373
0.7	0.230

Because the Metropolis algorithm spends more time in high-density regions of the target distribution than in low-density regions, taking larger steps away from current parameter values tends to decrease the density at proposed values, consequently decreasing the probability of acceptance.

25.2.2 The Gibbs Sampler*

As I mentioned, the simple Gibbs sampler described in this section is based on the observation that the joint distribution of an n-dimensional vector random variable $\mathbf{x}$ can be composed from the conditional distribution of each of its elements given the others, that is $p(X_j|\mathbf{x}_{-j})$ for $j = 1, 2, \ldots, n$ (where $\mathbf{x}_{-j} = [X_1, X_2, \ldots, X_{j-1}, X_{j+1}, \ldots, X_n]'$ is $\mathbf{x}$ with the jth element removed). There are many variations on the Gibbs sampler, such as its application to subsets of $\mathbf{x}$ (some of which are of size greater than 1) that partition $\mathbf{x}$: That is, with suitable ordering of its elements, $\mathbf{x} = [\mathbf{x}_\mathbf{1}', \mathbf{x}_\mathbf{2}', \ldots, \mathbf{x}_\mathbf{q}']'$. The corresponding (generally multivariate) conditional distributions are $p(\mathbf{x_j}|\mathbf{x_{-j}})$. Conditional

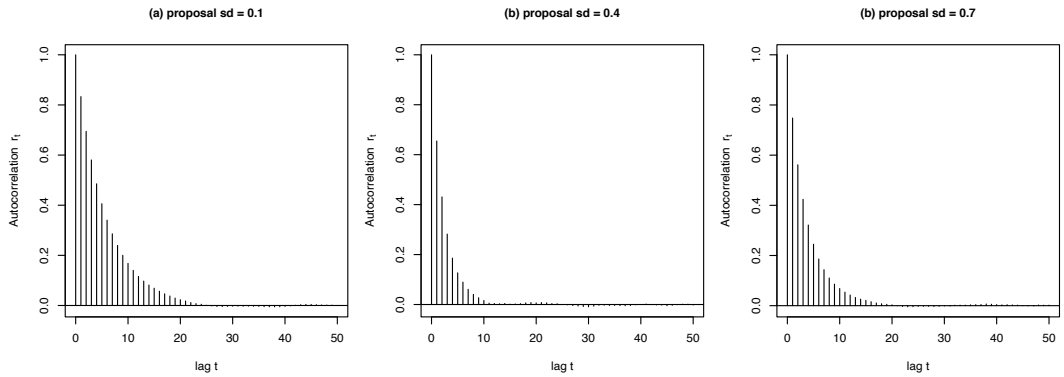

Figure 25.9 Autocorrelations of the sampled values of π produced by the Metropolis algorithm using a flat prior, for various standard deviations of the normal proposal distribution.

distributions of this form can arise naturally in the process of specifying hierarchical Bayesian statistical models in circumstances where it is difficult to derive the joint distribution $p(\mathbf{x})$ analytically.

The basic Gibbs sampler is simple to describe and proceeds as follows:

1. Pick an arbitrary set of initial values $\mathbf{x} = \mathbf{x}_0$.
2. Then for each of m iterations, sample in succession each element of $\mathbf{x}$ from its conditional distribution, conditioning on the most recent values of the other elements. That is for $i = 1, 2, \ldots, m$:

   Sample $x_1^{(i)}$ from $p(x_1|X_2 = x_2^{(i-1)}, \ldots, X_n = x_n^{(i-1)})$.
   Sample $x_2^{(i)}$ from $p(x_2|X_1 = x_1^{(i)}, X_3 = x_3^{(i-1)}, \ldots, X_n = x_n^{(i-1)})$.
   $\vdots$
   Sample $x_n^{(i)}$ from $p(x_n|X_1 = x_1^{(i)}, \ldots, X_{n-1} = x_{n-1}^{(i)})$.
   Save $\mathbf{x}_i = [x_1^{(i)}, x_2^{(i)}, \ldots, x_n^{(i)}]'$.

Using the Gibbs Sampler to Sample From a Bivariate-Normal Distribution

As I did for the Metropolis algorithm in Section 25.2.1, I'll illustrate the Gibbs sampler by drawing values from the bivariate-normal distribution with mean vector and covariance matrix

$$\begin{aligned} \boldsymbol{\mu} &= [1, 2]' \\ \boldsymbol{\Sigma} &= \begin{bmatrix} 1 & 1 \\ 1 & 4 \end{bmatrix} \end{aligned} \tag{25.13}$$

As previously mentioned, this example is artificial because it's easy to sample directly from the bivariate-normal distribution.

To apply the Gibbs sampler, we need the conditional distributions $p(x_1|x_2)$ and $p(x_2|x_1)$. In the bivariate-normal case, the conditional distributions are normal, with means and standard deviations provided by the population linear regression of each variable on the other: The regression of X_1 on X_2 is

$$E(X_1|x_2) = \alpha_{12} + \beta_{12} x_2 \tag{25.14}$$

where

$$\begin{aligned} \beta_{12} &= \frac{\sigma_{12}}{\sigma_2^2} \\ \alpha_{12} &= \mu_1 - \beta_{12}\mu_2 \end{aligned} \tag{25.15}$$

with constant error (i.e., conditional) variance

$$\sigma^2_{1|2} = \sigma_1^2 \left(1 - \frac{\sigma_{12}^2}{\sigma_1^2 \sigma_2^2}\right) \tag{25.16}$$

In these equations, μ_1 and μ_2 are the unconditional means of X_1 and X_2, σ_1^2 and σ_2^2 are their variances, and σ_{12} is their covariance. Thus

$$X_1|x_2 \sim N\left(\alpha_{12} + \beta_{12}x_2, \sigma^2_{1|2}\right) \tag{25.17}$$

The results for the conditional distribution of X_2 given $X_1 = x_1$ are entirely analogous.

I sampled $m = 10^5$ values of $\mathbf{x}$, producing the following estimated means and covariances:

$$\begin{aligned} \widehat{\boldsymbol{\mu}} &= [1.007, 2.017]' \\ \widehat{\boldsymbol{\Sigma}} &= \begin{bmatrix} 0.997 & 1.002 \\ 1.002 & 4.022 \end{bmatrix} \end{aligned} \tag{25.18}$$

The sampled values are shown in Figure 25.10 along with a nonparametric density estimate and the corresponding bivariate-normal density contours. Evidently, the Gibbs sampler, like the Metropolis algorithm, accurately recovers the means, variances, covariance, and shape of the bivariate-normal target distribution.

The values of X_1 and X_2 drawn by the Gibbs sampler are autocorrelated, but less so than those produced for this example by the Metropolis algorithm: See Figure 25.11 (and cf., Figure 25.6 on page 22). Consequently, we can get an approximately independent sample with less thinning (taking every third or fourth value).

25.2.3 Hamiltonian Monte Carlo*

As I mentioned, Hamiltonian Monte Carlo is named after the 19th Century physicist William Rowen Hamilton, who reformulated the mathematics of classical Newtonian mechanics.[27] HMC exploits an analogy between exploring the surface of a probability density function and the motion of an object along a frictionless surface, propelled by its initial momentum and gravity.

Hamiltonian Dynamics

Extending an example suggested by Neal (2011), think of a hockey puck (Neal's paper and this chapter were written in Canada after all) given an initial push in a particular direction on a frictionless and completely flat horizontal ice surface: The puck will continue to travel indefinitely in a straight line and with constant velocity in the direction in which it's pushed.[28]

Now imagine a surface that isn't flat, such as the surface in Figure 25.12: The graphs in this figure record the final results of two 100-step simulations, in which

[27] Although it's slightly off-topic, Susskind and Hrabovsky (2013) provide an especially lucid account of classical mechanics, including an explanation of Hamilton's contribution to the subject. Monte-Carlo simulation methods were coincidentally originally developed to solve problems in physics, and these methods later acquired prominent statistical applications.

[28] This is Newton's first law of motion: the law of inertia.

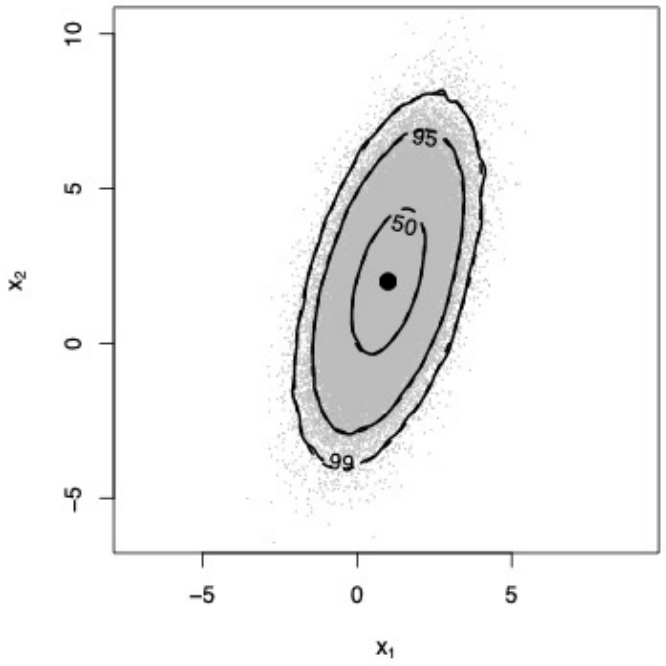

Figure 25.10 The gray dots show $m = 10^5$ values drawn by the Gibbs sampler from the bivariate-normal distribution with $\mu_1 = 1$, $\mu_2 = 2$, $\sigma_1^2 = 1$, $\sigma_2^2 = 4$, and $\sigma_{12} = 1$. The slightly irregular solid lines represent estimated density contours enclosing 50%, 95%, and 99% of the sampled points. The broken lines are the corresponding elliptical density contours of the bivariate-normal target distribution.

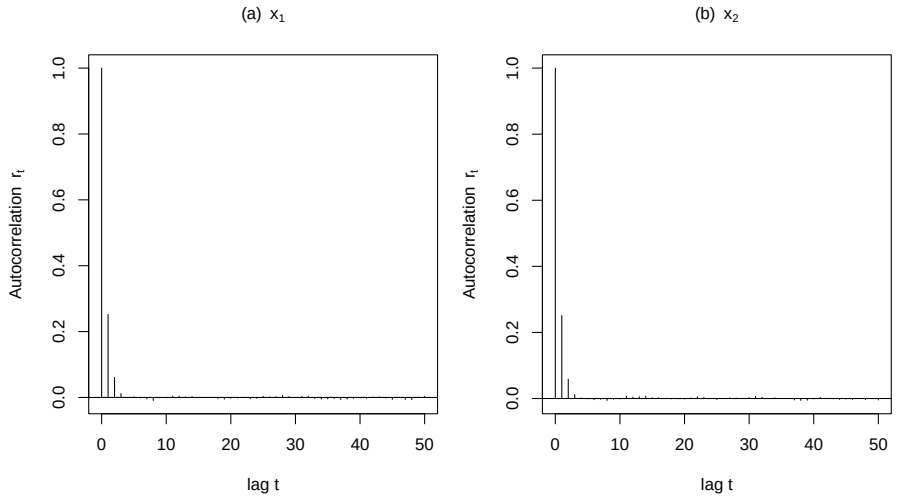

Figure 25.11 Autocorrelations of the values of (a) x_1 and (b) x_2 drawn by the Gibbs sampler applied to the bivariate-normal distribution with $\mu_1 = 1$, $\mu_2 = 2$, $\sigma_1^2 = 1$, $\sigma_2^2 = 4$, and $\sigma_{12} = 1$.

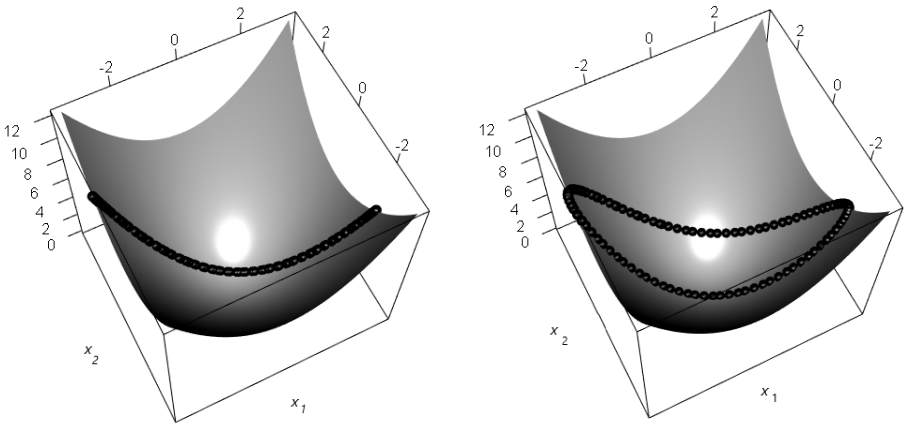

Figure 25.12 Trajectories described by two pucks released on a frictionless surface. The puck at the left is released with 0 initial momentum; the puck at the right with small momentum in the directions of x_1 and x_2. The surface was generated by the bivariate-normal distribution with $\mu_1 = \mu_2 = 0, \sigma_1^2 = \sigma_2^2 = 1$, and $\sigma_{12} = 0$, as described in the text.

two pucks are released on the surface depicted above the point $\mathbf{x}_0 = (-3.5, 2)'$, where x_1 and x_2 are the horizontal axes of the 3D space:

- The puck at the left is released with 0 momentum, and is therefore initially subject only to the force of gravity; the puck oscillates between the release point and an equally high point opposite to it on the surface.

- The puck at the right is released with small momentum (0.5 and 1, respectively) in the positive directions of the x_1 and x_2 axes; this puck describes a more complex looping trajectory over the surface but also returns to its starting point.

Assuming, without any real loss of generality, that the puck has mass equal to 1,[29] at any instant, the potential energy of the puck due to gravity is equal to its height, while its kinetic energy is equal to the sum of the squared values of its momentum (mass $\times$ velocity $= 1 \times$ velocity) in the directions of x_1 and x_2. Because there is no loss of energy due to friction, conservation of energy dictates that the sum of potential energy and kinetic energy remains the same as the puck moves. When the puck moves downwards on the surface, its velocity increases, and hence its momentum and kinetic energy increase, while its height and hence its potential energy decrease; when the momentum of the puck carries it upwards on the surface, against gravity, the opposite is the case: Momentum and kinetic energy decrease, height and potential energy increase.

By repeatedly introducing randomness into the momentum of the puck, HMC is able to visit the surface more generally, favoring lower regions of the surface. In a statistical application, the surface in question is the negative of the log of a multivariate probability density function (up to an additive constant on the log-density scale—i.e., a multiplicative constant on the density scale), and thus low regions correspond to regions of high probability. In the context of exploring a probability-density surface, the position variables X_1 and X_2 correspond to the random variables to be sampled, while the momentum variables are purely artificial, though necessary for the physical Hamiltonian analogy.

Metropolis proposals in HMC are more adapted to the probability surface to be sampled than in the traditional Metropolis or Metropolis-Hastings algorithms. By adjusting factors (discussed below) that affect the trajectory, it's possible to increase the proportion of accepted proposals and to decrease the autocorrelation of successively sampled values.

The surface shown in Figure 25.12 is for a bivariate-normal distribution with mean vector $\boldsymbol{\mu} = (0, 0)'$ and covariance matrix $\boldsymbol{\Sigma} = \mathbf{I}_2$ (the order-two identity matrix)—that is, two uncorrelated standard-normal random variables. The graphed surface is the negative log-density of this bivariate-normal distribution,

[29] The motion of the puck doesn't depend on its mass—recall Galileo's famous, if apocryphal, experiment in which he simultaneously dropped objects of different weight from the Leaning Tower of Pisa and observed that they hit the ground at the same time—and setting mass to 1 simplifies the formulas given below. Thought of another way, the units of mass, say kg, are purely conventional and arbitrary, so we're entitled to pick units that make the mass of the puck equal to 1.

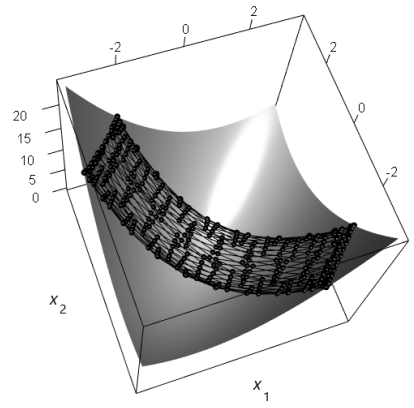

Figure 25.13 Trajectory described by a puck released on a frictionless surface with 0 initial momentum; 400 steps are shown. The surface was generated by the bivariate-normal distribution with $\mu_1 = \mu_2 = 0, \sigma_1^2 = \sigma_2^2 = 1$, and $\sigma_{12} = 0.5$, as described in the text.

omitting the normalizing constant. Thus, the negative log-density surface differs from the graphed surface only by a constant difference in elevation, producing essentially the same dynamics for pucks sliding along both surfaces. This surface is a simple "bowl," whose vertical slices are parabolic and whose horizontal slices are circular, yielding the very simple dynamics illustrated in Figure 25.12.

For a partly contrasting example, consider the surface shown in Figure 25.13, generated by the bivariate normal distribution with mean vector $\boldsymbol{\mu} = (0, 0)'$ and covariance matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix} \tag{25.19}$$

That is, X_1 and X_2 are standard-normal random variables with correlation $\rho = 0.5$. As in the first example, a puck is released with 0 momentum above the point $\mathbf{x} = (-3.5, 2)'$, and now 400 steps are shown, tracing out a much more elaborate trajectory than before. In this case, the "bowl" representing the negative log-density function is still parabolic in vertical cross-sections, but it is now elliptical in horizontal cross-sections, producing more complex dynamics.[30]

In the general case with which we'll eventually be concerned, the value of the surface giving the "elevation" of the puck is a function $g(\mathbf{x})$ of n "position" variables, comprising the vector $\mathbf{x}$. In a statistical application of HMC, the elements of $\mathbf{x}$ are the random variables that we want to sample, and the height of the surface is the negative of the log of a function that may differ from the target

[30] Because the dynamics are unchanged by rotation of the surface around the vertical axis, the same effect can be achieved with uncorrelated Xs that have different standard deviations, stretching the bowl in the direction of the larger standard deviation.

density by a multiplicative constant. Thus, in the notation of Sections 25.2.1 and 25.2.2, $g(\mathbf{x}) = -\log_e p^*(\mathbf{x})$.

There are also n momentum variables, in the vector $\mathbf{m}$, one for each of the coordinates of $\mathbf{x}$.[31] The *Hamiltonian*, $H(\mathbf{x}, \mathbf{m})$, is a function of $\mathbf{x}$ and $\mathbf{m}$, and, for the cases that I'll consider, is composed of two functions, which, in their physical interpretation, represent respectively potential and kinetic energy. Potential energy is equal to the elevation of the surface at the current position, $g(\mathbf{x})$, while kinetic energy, $k(\mathbf{m})$, is purely a function of momentum. The total energy of the puck is conserved as it moves, and so at any point in time t,

$$E = H[\mathbf{x}(t), \mathbf{m}(t)] = g[\mathbf{x}(t)] + k[\mathbf{m}(t)] \tag{25.20}$$

where

$$k(\mathbf{m}) = \tfrac{1}{2}\mathbf{m}'\mathbf{m} = \tfrac{1}{2}\sum_{i=1}^{n} m_i^2 \tag{25.21}$$

All this assumes, recall, that the mass of the puck is 1, which slightly simplifies the results (e.g., equating momentum to velocity).

The trajectory of the puck over "time," t, is given by Hamilton's equations

$$\begin{aligned} \frac{d\mathbf{m}}{dt} &= -\frac{\partial H(\mathbf{x}, \mathbf{m})}{\partial \mathbf{x}} = -\frac{\partial g(\mathbf{x})}{\partial \mathbf{x}} \\ \frac{d\mathbf{x}}{dt} &= \frac{\partial H(\mathbf{x}, \mathbf{m})}{\partial \mathbf{m}} = \mathbf{m} \end{aligned} \tag{25.22}$$

The Leapfrog Method

The time trajectory implied by Hamilton's differential equations (25.22) isn't in general solvable analytically, but it can be accurately approximated by discretizing time in small steps, ε, and applying the following algorithm (adapted from Neal, 2011), called the leapfrog method:[32]

- Start with values of $\mathbf{x}(0)$ and $\mathbf{m}(0)$ at time $t = 0$, and take a half-step for the momentum variables $\mathbf{m}$,

$$\mathbf{m}(\varepsilon/2) = \mathbf{m}(0) - \frac{\varepsilon}{2} \times \frac{\partial g(\mathbf{x})}{\partial \mathbf{x}}\Big|_{\mathbf{x}=\mathbf{x}(0)} \tag{25.23}$$

- Then for $t = \varepsilon, 2 \times \varepsilon, \ldots, s \times \varepsilon$ (where s is the number of steps), serially update the position and momentum variables, using the most recent value of each:

$$\begin{aligned} \mathbf{x}(t) &= \mathbf{x}(t - \varepsilon) + \varepsilon \times \mathbf{m}(t - \varepsilon/2) \\ \mathbf{m}(t + \varepsilon/2) &= \mathbf{m}(t - \varepsilon/2) - \varepsilon \times \frac{\partial g(\mathbf{x})}{\partial \mathbf{x}}\Big|_{\mathbf{x}=\mathbf{x}(t)} \end{aligned} \tag{25.24}$$

  The time-length of the trajectory is thus $s \times \varepsilon$.

[31] My notation here is adapted to the statistical application of Hamiltonian Monte Carlo. In physics, it's conventional to represent the position variables by $\mathbf{q}$ and the momentum variables by $\mathbf{p}$.

[32] How good the approximation is depends on the step size and length of the trajectory.

To apply the leapfrog method to the illustrative bivariate normal distributions, we need the negative of the log density (ignoring the normalizing constant[33]) and its partial derivatives (the gradient), which are simply

$$\begin{aligned} g(\mathbf{x}) &= \tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \\ \frac{\partial g(\mathbf{x})}{\partial \mathbf{x}} &= \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \end{aligned} \tag{25.25}$$

HMC Sampling

To implement HMC sampling, the leapfrog method can be used to generate proposals to the Metropolis algorithm (Section 25.2.1).[34] To generate a proposal $\mathbf{x}^*$, start at the current values $\mathbf{x}_i$ of the variables to be sampled, and randomly select the starting momentum in each direction—here, the momentum values are sampled independently from the standard-normal distribution, $N(0,1)$. The Metropolis acceptance ratio (Equation 25.10 on page 18) becomes

$$a = \exp[H(\mathbf{x}_i, \mathbf{m}_i) - H(\mathbf{x}^*, \mathbf{m}^*)] \tag{25.26}$$

The acceptance ratio a depends on both the momentum variables $\mathbf{m}$ and the position variables $\mathbf{x}$, because both are necessary to characterize the current state of the system, even though only the position variables are of real interest.

If the leapfrog method were exact, then (because of conservation of energy) the energy for the proposal at the end of the path, $H(\mathbf{x}^*, \mathbf{m}^*)$, would be exactly equal to the energy at the beginning of the path, $H(\mathbf{x}_i, \mathbf{m}_i)$, in which case the acceptance ratio $a = \exp(0) = 1$, and the proposal would *always* be accepted. The acceptance ratio can only depart from 1 due to discretization error in the leapfrog method.[35] If we "tune" the step size and number of steps well, therefore, we should expect a high acceptance rate for proposals. To achieve both a high rate of acceptance and nearly independent draws from the target distribution, tuning is more critical for HMC than for simpler Metropolis sampling.

I proceed to apply HMC to the bivariate-normal target distribution with

$$\begin{aligned} \boldsymbol{\mu} &= [1, 2]' \\ \boldsymbol{\Sigma} &= \begin{bmatrix} 1 & 1 \\ 1 & 4 \end{bmatrix} \end{aligned} \tag{25.27}$$

[33] For the bivariate-normal distribution, the (ignored) multiplicative normalizing constant is $\left(2\pi\sqrt{\det\boldsymbol{\Sigma}}\right)^{-1}$.

[34] The leapfrog method is symmetric (i.e., reversible), and so we can use the simpler Metropolis algorithm instead of Metropolis-Hastings. We can also permit different "step" sizes for the various elements of $\mathbf{x}$, effectively multiplying the time-increment ε by scaling factors for the position variables; this can be a useful approach when the variables have different scales (e.g., standard deviations). The resulting path along the surface isn't a true Hamiltonian trajectory, but it still provides legitimate update candidates to the Metropolis algorithm. A common alternative is to complicate the Hamiltonian equations by introducing a diagonal $n \times n$ "mass matrix" $\mathbf{M}$, the diagonal entries of which reflect the scales of the variables.

[35] Because, however, the leapfrog method is reversible, an error in approximating the Hamiltonian doesn't invalidate the method.

used previously in Sections 25.2.1 and 25.2.2 to illustrate the Hastings and Gibbs algorithms, producing the following estimates, based on $m = 10^5$ sampled values:

$$\begin{aligned} \widehat{\boldsymbol{\mu}} &= [1.001, 2.001]' \\ \widehat{\boldsymbol{\Sigma}} &= \begin{bmatrix} 0.997 & 1.001 \\ 1.001 & 3.956 \end{bmatrix} \end{aligned} \tag{25.28}$$

The percentage of accepted proposals, 98.7%, is much higher than for the standard Metropolis version of this example in Section 25.2.1. The density estimate in Figure 25.14 shows that the HMC sample closely reproduces the target distribution, and the sampled values are much less autocorrelated (Figure 25.15) than the values produced by the traditional Metropolis algorithm or by the Gibbs sampler (cf., Figures 25.6 and 25.11, respectively on pages 22 and 30).

To achieve these desirable results, I had to select by trial and error suitable values for the step sizes and number of steps. In particular, step size was set twice as large in the direction of X_2 (which has standard deviation $\sigma_2 = 2$) than in the direction of X_1 (which has standard deviation $\sigma_1 = 1$.)

25.2.4 Convergence of MCMC Sampling to the Target Distribution

In theory, Markov chains produced by MCMC sampling converge to the target distribution as the number of simulated draws goes to infinity, but in practice several sorts of problems can occur in chains of finite length. The high-dimensional posterior distributions associated with complex statistical models can be much more challenging to sample effectively than a well-behaved low-dimensional distribution like the bivariate normal or the beta. Although there is no way to *guarantee* that MCMC samples have converged to the target distribution, a variety of methods has been developed to diagnose non-convergence. In this section, I'll describe a simple graphical diagnostic, called a *trace plot*, along with two numeric diagnostics suggested by Gelman et al. (2013).

A trace plot is simply a line graph of each sampled quantity—typically a parameter or a function of parameters in an application of MCMC to Bayesian inference—versus the simulation index. If the MCMC samples have converged to the target distribution, then the center and spread of the trace plot shouldn't change on average with the index.

Some prototypical examples of problematic trace plots are shown in Figure 25.16, for an imagined parameter θ:

- In (a), simulations start far from the high-density region of the target distribution, and so a long burn-in period is required before convergence occurs.
- In (b), successive sampled values are highly autocorrelated, and they therefore contain much less information about the target distribution than an

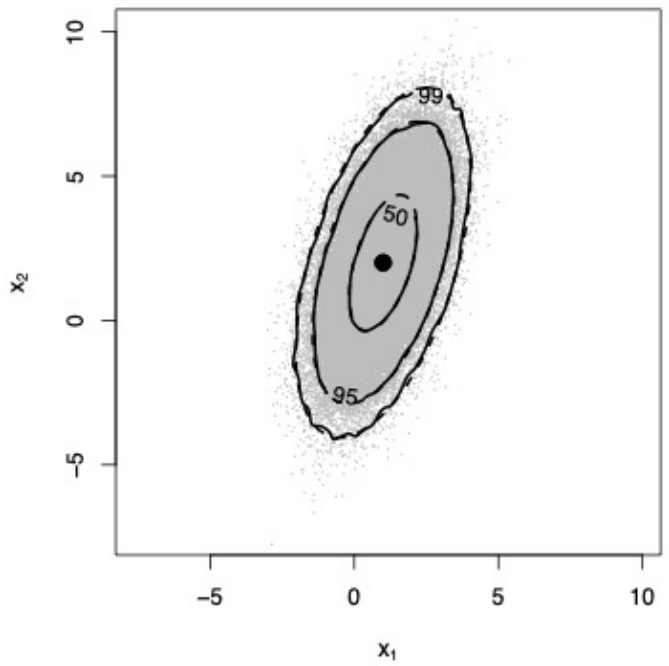

Figure 25.14 The gray dots show $m = 10^5$ values drawn by Hamiltonian Monte-Carlo from the bivariate-normal distribution with $\mu_1 = 1$, $\mu_2 = 2$, $\sigma_1^2 = 1$, $\sigma_2^2 = 4$, and $\sigma_{12} = 1$. The slightly irregular solid lines represent estimated density contours enclosing 50%, 95%, and 99% of the sampled points. The broken lines are the corresponding elliptical density contours of the bivariate-normal target distribution.

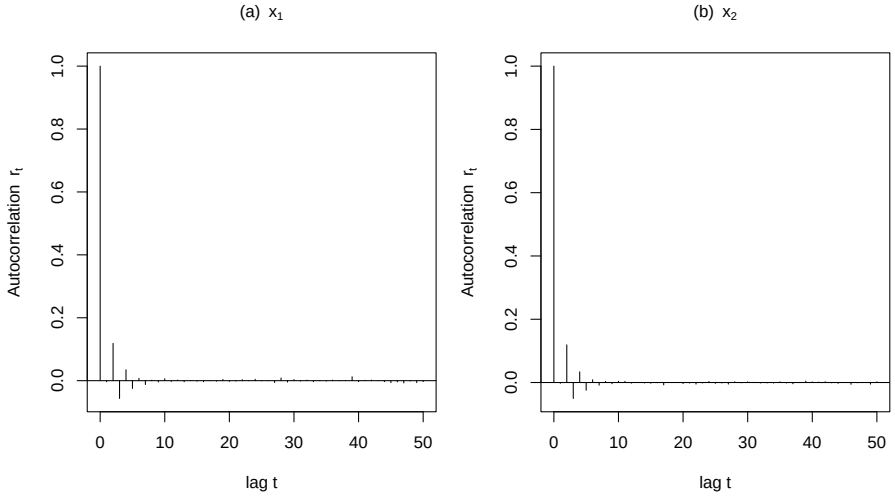

Figure 25.15 Autocorrelations of the values of (a) x_1 and (b) x_2 drawn by HMC applied to the bivariate-normal distribution with $\mu_1 = 1$, $\mu_2 = 2$, $\sigma_1^2 = 1$, $\sigma_2^2 = 4$, and $\sigma_{12} = 1$.

equal number of independently sampled values. As previously mentioned, we can thin the sampled values to produce an approximately independent sample. We can also use all of the sampled values, with the recognition that they are not as informative as they would be were they independent. I briefly address how to measure the amount of information in an autocorrelated Markov chain later in this section.

- In (c), there is a trend in the average sampled value, and convergence to the target distribution never occurs in the course of the simulation.
- Panel (d) shows trace plots for two different Markov chains, following a suggestion by Gelman et al. (2013, Section 11.4) to compare two or more chains starting in different places. If the two chains both converge to the target distribution, then their trace plots should overlap—should "mix" well, in the jargon of MCMC; their vertical separation in the example suggests to the contrary that the two chains are visiting different regions of the parameter space—perhaps two separated high-density regions—and that neither simulation has converged to the target distribution.

Figure 25.17 shows trace plots for x_1 and x_2 in the 10^5 values that I sampled by HMC from the bivariate normal distribution with $\mu_1 = 1$, $\mu_2 = 2$, $\sigma_1^2 = 1$, $\sigma_2^2 = 4$, and $\sigma_{12} = 1$. As I explained, this is a simple distribution from which to sample, and so it should be no surprise that the trace plots look well-behaved. The broken horizontal lines are drawn at the average values of x_1 and x_2. The solid lines are local regressions with spans of 0.01 (see Section 18.1); the central line in each panel is computed for all of the sampled values, and the lower and

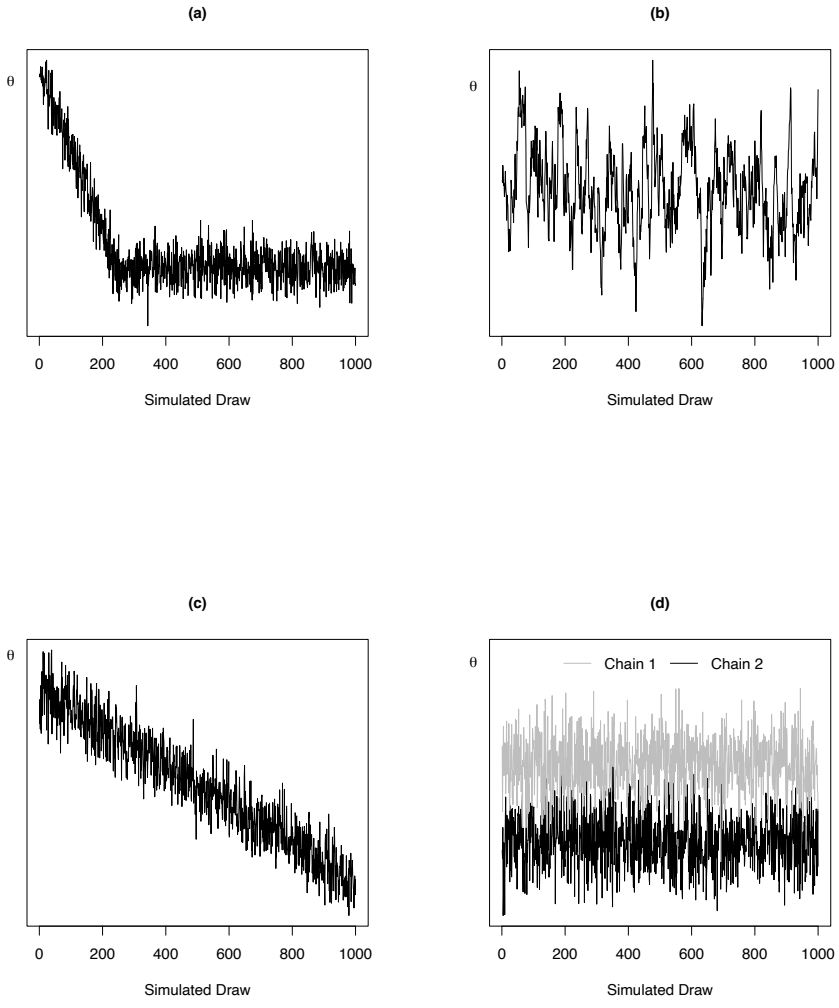

Figure 25.16 Potential issues in MCMC trace plots for a parameter θ: (a) burn-in required; (b) highly autocorrelated values; (c) lack of convergence; and (d) two independent chains that have *apparently* converged to different values.

upper lines are computed respectively for the values below and above the mean. Both the center and spread of the simulated values appear to be constant over the 100,000 simulated draws.

Gelman et al. (2013, Section 11.4) also suggest the following numeric procedure for assessing convergence:

1. Sample several Markov chains, starting in different places (speaking to the potential problem in Figure 25.16 (d)).

2. Discard the first half of each chain (that is, use the first half of the sampled values as a burn-in period, to avoid the problem in Figure 25.16 (a)).

3. Split the retained values from each chain in half, treating each half as a separate chain (to be able to detect the problem illustrated in Figure 25.16 (c)).

This procedure produces p chains, each with m sampled values. If the chains converge to the target distribution, then they should be similar to one-another. Gelman et al. (2013) compare the chains by a kind of analysis of variance. Suppose that the parameter of interest is θ with sampled values $\theta_{ij}, i = 1, \ldots, m, j = 1, \ldots, p$. Define

$$\begin{aligned}
\bar{\theta}_{\cdot j} &\equiv \frac{1}{m}\sum_{i=1}^{m}\theta_{ij} \\
\bar{\theta}_{\cdot\cdot} &\equiv \frac{1}{p}\sum_{j=1}^{p}\bar{\theta}_{\cdot j} = \frac{\sum_{j=1}^{p}\sum_{i=1}^{m}\theta_{ij}}{p \times m} \\
s_j^2 &\equiv \frac{1}{m-1}\sum_{i=1}^{m}\left(\theta_{ij} - \bar{\theta}_{\cdot j}\right)^2
\end{aligned}$$

Between-chain and within-chain variation are measured respectively by

$$\begin{aligned}
B &\equiv \frac{m}{p-1}\sum_{j=1}^{p}\left(\bar{\theta}_{\cdot j} - \bar{\theta}_{\cdot\cdot}\right)^2 \\
W &\equiv \frac{1}{p}\sum_{j=1}^{p}s_j^2
\end{aligned}$$

An estimate of the posterior variance of θ is then

$$\widehat{V}(\theta) = \frac{m-1}{m}W + \frac{1}{m}B$$

Finally, compute the diagnostic statistic

$$\widehat{R} = \sqrt{\frac{\widehat{V}(\theta)}{W}}$$

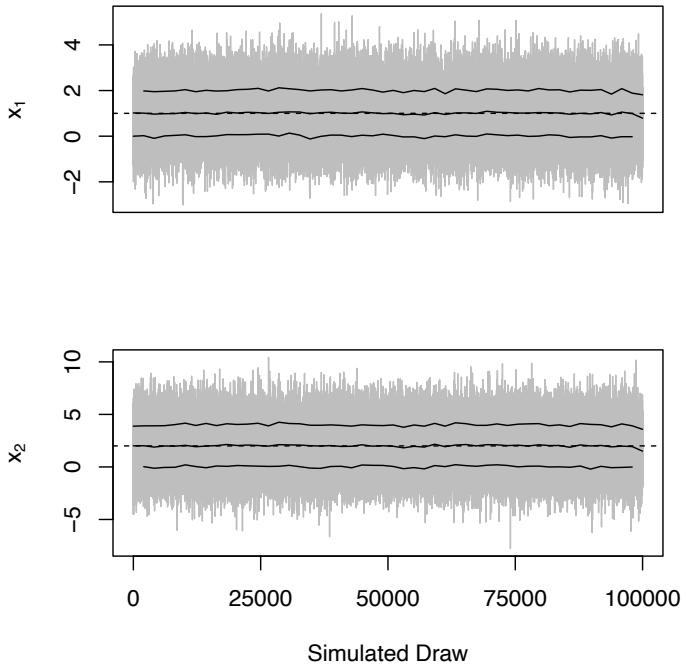

Figure 25.17 Trace plots of x_1 and x_2 for the 10^5 HMC samples drawn from the bivariate-normal distribution with $\mu_1 = 1$, $\mu_2 = 2$, $\sigma_1^2 = 1$, $\sigma_2^2 = 4$, and $\sigma_{12} = 1$. The broken horizontal lines are drawn at the average values of x_1 and x_2. The central solid line in each graph is from a nonparametric regression of the x-values versus index, and the lower and upper solid lines are from nonparametric regressions for the values below and above the mean, respectively.

called the *potential scale-reduction factor* because it estimates how much the dispersion of the posterior distribution of θ would decline if there were an infinite number of MCMC samples. $\widehat{R}$ substantially larger than 1 (say, exceeding 1.1) suggests that the several chains are heterogeneous, indicative of a convergence problem.[36]

I applied Gelman et al.'s procedure to the HMC samples drawn from the bivariate-normal distribution with $\mu_1 = 1$, $\mu_2 = 2$, $\sigma_1^2 = 1$, $\sigma_2^2 = 4$, and $\sigma_{12} = 1$, sampling a second Markov chain of 10^5 values, and splitting each chain in half; I didn't bother with a burn-in period for this example. Thus, $p = 4$ and $m = 5 \times 10^4$. The values of $\widehat{R}$ for x_1 and x_2 were both equal to 1 to several places to the right of the decimal point, consistent with convergence of the chains to the target distribution.

Gelman et al. (2013, Section 11.5) also suggest a measure of the *effective sample size* for a set of autocorrelated Markov chains. Consider p chains for a parameter θ, each of size m, thus comprising $p \times m$ non-independent draws from the target distribution, where θ_{ij} is the ith draw in the jth chain. Adapting Equation 25.12 (on page 22) to pool the sampled values across the several chains, the estimated autocorrelations are

$$r_t = \frac{\sum_{j=1}^{p} \sum_{i=t+1}^{m} (\theta_{ij} - \bar{\theta}_{..})(\theta_{i-t,j} - \bar{\theta}_{..})}{\sum_{j=1}^{p} \sum_{i=1}^{m} (\theta_{ij} - \bar{\theta}_{..})^2}, t = 1, 2, \ldots$$

where (as above)

$$\bar{\theta}_{..} = \frac{\sum_{j=1}^{p} \sum_{i=1}^{m} \theta_{ij}}{p \times m}$$

Then (simplifying slightly), the effective sample size is an inverse function of the autocorrelations,

$$m_{\text{eff}} \approx \frac{p \times m}{1 + 2 \sum_{t=1}^{t'} r_t}$$

where t' is picked so that the r_t for $t > t'$ are negligibly small.

Continuing with the HMC bivariate-normal example, the autocorrelations computed from the $p = 4$ pooled chains are all small (cf., Figure 25.15 on page 37)—for example, $r_1 = -0.0018$ and $r_2 = 0.1235$ for both x_1 and x_2—and so the effective sample sizes based (arbitrarily) on the first 10 autocorrelations are, respectively, $m_{\text{eff}} \approx 173,160$ for x_1 and $m_{\text{eff}} \approx 171,633$ for x_2, not much less than the $p \times m = 4 \times 50,000 = 200,000$ draws from the target distribution.

[36] Brooks and Gelman (1997) introduce an improved estimate of R, but the details need not concern us here.

In theory, Markov chains produced by MCMC sampling converge to the target distribution as the number of simulated draws goes to infinity, but in practice several sorts of problems can occur in chains of finite length. Convergence diagnostics help to determine whether MCMC samples adequately characterize a target distribution. In formulating these diagnostics, it helps to sample and compare two or more independent Markov chains and to discard the initial samples of each (e.g., the first half) as a "burn-in period."

- A *trace plot* is a line graph of a sampled quantity—typically a parameter or a function of parameters—versus the simulation index. If the MCMC samples have converged to the target distribution, then the center and spread of the trace plot shouldn't change on average with the index, and trace plots for independent Markov chains should be similar.
- The *potential scale-reduction factor* $\widehat{R}$ measures the similarity of two or more Markov chains for a sampled quantity such as a parameter. If the chains have converged, then $\widehat{R}$ should be close to 1.
- The *effective sample size* $m_{\text{eff}} \approx pm/\left(1 + 2\sum_{t=1}^{t'} r_t\right)$ measures the amount of information about a sampled quantity contained in the MCMC samples, where p is the number of independent chains employed, m is the number of samples retained from each chain, r_t is the estimated autocorrelation of the sampled values at lag t, and t' is selected so that r_t is negligible for $t > t'$.

25.3 Bayesian Estimation of Linear and Generalized Linear Models

As I have explained, Bayesian statistical inference requires a probability model for the data, from which the likelihood for the observed data is derived, and a prior distribution for the parameters of the model. Combining the likelihood and the prior produces the posterior distribution of the parameters, on which inference is based. Both the normal linear model, described in Part II, and traditional generalized linear models, such as the logit model and the Poisson-regression model, described in Part IV, provide clearly defined probability models for the data, and so the principal open task in applying the machinery of Bayesian inference to these models is to specify the prior distribution of the parameters of the model.

25.3.1 The Normal Linear Model

Previously (e.g., in Section 6.2.1), I wrote the normal linear model as

$$Y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i$$

where $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$ and $\varepsilon_i, \varepsilon_{i'}$ are independent for $i \neq i'$. For our current purposes, I'll equivalently specify the conditional distribution of the data Y directly as

$$\begin{aligned} &Y_i | x_{i1}, x_{i2}, \ldots, x_{ik} \sim N(\alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}, \sigma_\varepsilon^2) \\ &Y_i, Y_{i'} \text{ are independent for } i \neq i' \end{aligned} \tag{25.29}$$

Writing statistical models in this form will be convenient in more complex applications, such as the mixed-effects models considered in Section 25.4.

With the likelihood of the data in hand, we next require the joint prior distribution of the $k+2$ parameters of the model, $\alpha, \beta_1, \beta_2, \ldots, \beta_k, \sigma_\varepsilon^2$, which is typically addressed by specifying *independent* prior distributions for the various parameters.[37]

For concreteness, let's focus on Duncan's occupational prestige regression (introduced in Section 5.2 and discussed a several points in the text). Recall that the response variable in Duncan's regression is the prestige of each of 45 U. S. occupations, circa 1950, measured as the percentage of ratings of "good' or better in a national survey. There are two explanatory variables in Duncan's regression—the educational and income levels of the occupations, measured respectively as the percentage of high-school graduates in the occupation and the percentage of individuals in the occupation earning $3500 or more, both from the 1950 U. S. Census.

I previously fit Duncan's model by least-squares regression (see Sections 5.2.1 and 6.2.2). Table 25.1 compares the least-squares estimates of the regression coefficients, their standard errors,[38] and the error standard deviation to two sets of Bayesian estimates: one produced by specifying flat independent prior distributions for all of the parameters of the model, and the other produced by specifying vaguely informative independent prior distributions for the parameters, both to be explained presently. The table also shows the $\widehat{R}$ convergence diagnostic for each Bayesian estimate, all of which round to 1.000, along with the effective sample size, m_{eff}. Trace plots for the two sets of Bayesian parameter estimates were unremarkable and aren't shown.[39]

[37] We know (see, e.g., Section 9.4.4) that when the regressors in a linear model are correlated, so are the estimated regression coefficients, and consequently independent prior distributions for regression coefficients aren't generally credible. It's not obvious, however, how one would go about specifying non-independent priors for the parameters of the model. See Section 25.5 and footnote 40 for further discussion of this point.

[38] In the case of the Bayesian estimates, the "standard errors" are the standard deviations of the the posterior distributions of the parameters—"standard error" is strictly speaking a familiar frequentist term for the estimated standard deviation of the sampling distribution of a statistic.

[39] These and all of the Bayesian estimates in the remainder of this chapter were computed

Table 25.1 Least-Squares and Bayesian Estimates for Duncan's Occupational Prestige Regression

	Estimated Parameter[a] *(Standard Error*[b]*)*			
Method	α/α^{*c}	β_1	β_2	σ_ε
Least-Squares	−6.06	0.599	0.546	13.4
	(4.27)	(0.120)	(0.098)	
Bayes, Flat Priors	−6.10	0.599	0.547	13.6
	(4.37)	(0.123)	(0.101)	
$\widehat{R}$	1.000	1.000	1.000	1.000
m_{eff}	11,773	10,213	9824	11,709
Bayes, Vague Priors	47.71	0.595	0.546	13.4
	(1.98)	(0.121)	(0.099)	
$\widehat{R}$	1.000	1.000	1.000	1.000
m_{eff}	18,513	12,118	11,969	16,722

[a,b] For Bayesian estimates, posterior mean and standard deviation.

[c] α^* with centered xs for the Bayes estimates with vague priors, otherwise α.

The flat improper priors employed for the first set of Bayesian estimates are straightforward for the three regression coefficients:

$$\begin{aligned}\alpha &\sim \text{Unif}(-\infty, \infty)\\ \beta_1 &\sim \text{Unif}(-\infty, \infty)\\ \beta_2 &\sim \text{Unif}(-\infty, \infty)\end{aligned}$$

The error standard deviation σ_ε is non-negative, and so I used a flat prior for its log: $\log_e \sigma_\varepsilon \sim \text{Unif}(-\infty, \infty)$.

The second set of Bayesian estimates, employing vaguely informative priors, requires more explanation:

- I centered the explanatory variables, income and education, to 0 means, $x^*_{i1} = x_{i1} - \bar{x}_1$, $x^*_{i2} = x_2 - \bar{x}_{i2}$, so that the regression model becomes $Y_i | x^*_{i1}, x^*_{i2} \sim N(\alpha^* + \beta_1 x^*_{i1} + \beta_2 x^*_{i2}, \sigma^2_\varepsilon)$. That's helpful for thinking about the prior distribution of the intercept, because now the intercept α^* is interpretable as the expected value of Y when the xs are equal to their means (and in least-squares regression $\widehat{\alpha}^*$ would simply be the unconditional mean of Y).[40]

using Stan. For each model, I sampled four independent chains, each of length 10,000, with the first 5000 iterations used for burn-in and no thinning of the remaining 5000 iterations; thus, $4 \times 5000 = 20,000$ sampled values of the parameters were retained in each case. I generally omit reporting $\widehat{R}$, m_{eff}, and trace plots in the subsequent examples in this chapter, because, unless indicated to the contrary, all of the diagnostics are satisfactory.

[40] A careful reader may notice that centering the explanatory variables at their sample

I then used the prior distribution $\alpha^* \sim N(50, 15^2)$, effectively confining α^* (which is a percentage) between 5 and 95—that is, ± 3 prior standard deviations around 50. Recall that almost all—99.7%—of the density of a normal distribution is within 3 standard deviations of its mean, and that 95% and 68% of the density lie respectively within 2 and 1 standard deviations of the mean.[41]

- The explanatory variables x_1 and x_2 are also percentages, and I set the prior distributions of the corresponding regression coefficients to $\beta_1 \sim N(0, 1)$ and $\beta_2 \sim N(0, 1)$, effectively restricting these coefficients to the range between -3 and 3. A 3-percent change in prestige for a 1-percent increase in income or education would be a very large effect.

- Finally, I specified $\log_e \sigma_\varepsilon \sim N(0, 1.5^2)$, which constrains the error standard deviation to lie approximately between $e^{-3\times 1.5} \approx 0.01$ and $e^{3\times 1.5} \approx 90$. Recall that σ_ε is on the same percentage scale as Y.

Here is a table summarizing the normal priors:

Prior SDs	-3	-2	-1	1	2	3
Centered Intercept (α^*)	5	20	35	65	80	95
Income Coefficient (β_1)	-3	-2	-1	1	2	3
Education Coefficient (β_2)	-3	-2	-1	1	2	3
Error Standard Deviation[a] (σ_ε)	0.01	0.05	0.22	4.5	20.9	90

[a] Showing $e^{Z\times 1.5}$, where Z is $-3, -2, -1, 1, 2, 3$.

Alternatively, we can simply graph the prior distributions as in Figure 25.18: Panel (a) shows the prior for the intercept α^* of the centered model; panel (b) shows the common prior for the regression coefficients β_1 and β_2 of income and education; and panels (c) and (d) show the prior for the error standard deviation σ_ε—on the $\log_e \sigma_\varepsilon$ scale in (c) (with the corresponding σ_ε scale at the top of the graph) and the σ_ε scale in (d).

Some comments about the prior distributions:

- Vaguely informative priors like these are common in applied Bayesian statistical modeling, but while the priors that I specified are reasonable, they don't reflect my honest prior beliefs about the regression coefficients.

means requires looking at the data to help construct a prior distribution for the intercept, a procedure that I've generally tried to avoid. I could instead center the xs at fixed values, such as 50%. One defense of centering at the sample means is that in this example, and typically, we're not terribly interested in the intercept and simply want to develop a reasonable prior for it. More deeply, however, we might find it legitimate to condition on the distribution of the xs, as we do in classical inference for regression. That raises the possibility of using the joint distribution of the xs for other purposes, such as to construct priors for the correlations among the regression coefficients.

[41] Because the normal distribution is unbounded below and above, it's *possible*—but very unlikely—to sample a value outside the range 0 to 100. Moreover, the normal linear model, whether estimated by least-squares or by Bayesian methods, doesn't constrain Y to lie between 0 and 100.

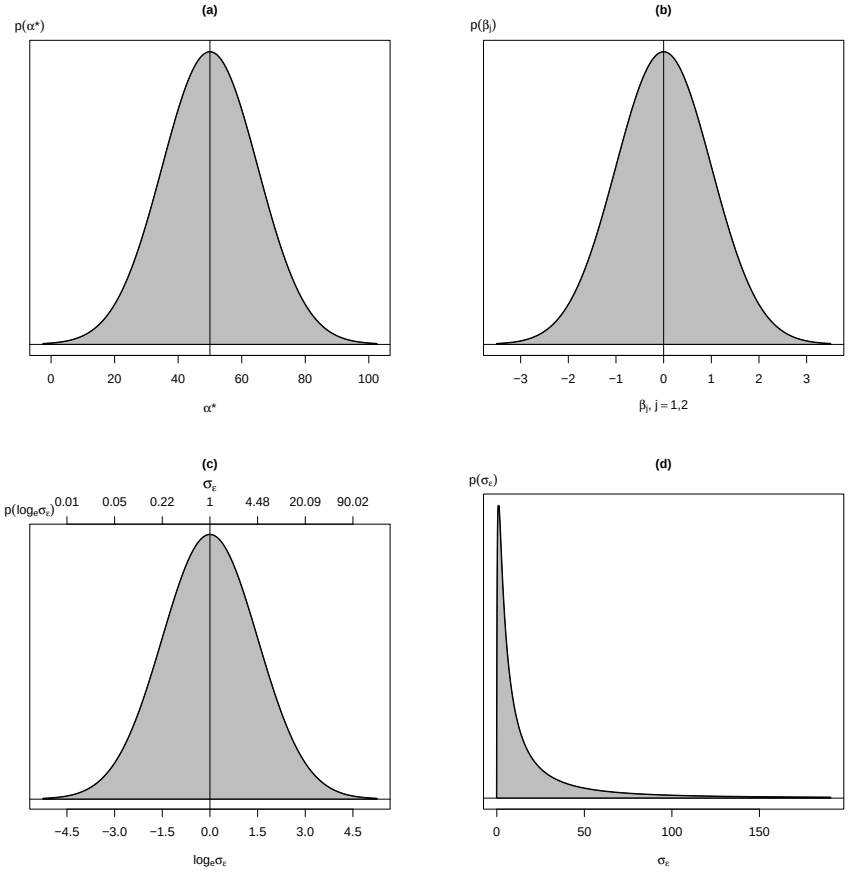

Figure 25.18 Vaguely informative priors for the parameters of Duncan's occupational prestige regression model: (a) prior for the intercept α^* of the centered model; (b) common prior for the regression coefficients β_1 and β_2 of income and education; (c) and (d) prior for the error standard deviation σ_ε.

- In particular, before examining the data (which is in itself an exercise in imagination because I'm familiar with Duncan's data and regression), I'd strongly expect both the income coefficient and the education coefficient to be positive, yet, as is the norm in applications, the priors for these coefficients are centered at 0. The result is that the posterior estimates of the regression coefficients tend to "shrink" towards 0 (as in ridge regression, discussed in Section 13.2.3).

- I believe that this argument applies more generally to vaguely informative priors for regression coefficients: When researchers include an explanatory variable in a regression equation it's generally *because* they expect its coefficient to be nonzero and, typically, expect the coefficient to have a particular sign (positive or negative). To specify a prior distribution for the coefficient centered at 0 contradicts genuine prior belief and is therefore fundamentally non-Bayesian—at least from the researcher's subjective point of view. Perhaps we can rescue a Bayesian interpretation of vaguely informative priors centered at 0 by ascribing them to a hypothetical critic who is skeptical of the researcher's hypothesis but who is uncertain in his or her skepticism.[42]

- Ironically, the justification for vague priors like the ones employed here often appeals to frequentist properties of the resulting estimators, such as robustness—the vague priors are not very restrictive but they do rule out wild estimates of the regression coefficients—and by shrinking regression coefficients towards 0 may improve the mean-squared error of estimates, a process termed *regularization*.

- I used—and in the remainder of this chapter, I will continue to use—normal priors simply because of their familiarity: That is, I expect that the reader will have developed intuition about the family of normal distributions. Other distributional choices are certainly possible, and arguably more reasonable.[43]

- As expected, the flat priors produce Bayesian estimates similar to the least-squares estimates,[44] but in this case, and despite the small sample size of Duncan's data set ($n = 45$), the same is true of the Bayesian estimates produced by the vaguely informative priors.

[42] This last point of interpretation was suggested to me by Georges Monette.

[43] For example, an exponential prior could be used for the error standard deviation. See the references given at the end of the chapter for extensive discussions of prior distributions.

[44] Bayesian estimates with flat priors can differ from the MLEs because of simulation error—that is, the posterior distribution is only *approximated* by MCMC—but even if the posterior produced by MCMC were exact, the MLE would correspond to the posterior mode, not to the posterior mean, and the latter is used for the Bayesian point estimates reported here.

The normal linear regression model provides a probability model for the data from which the likelihood can be calculated:

$$Y_i|x_{i1}, x_{i2}, \ldots, x_{ik} \sim N(\alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}, \sigma_\varepsilon^2)$$
$$Y_i, Y_{i'} \text{ are independent for } i \neq i'$$

To obtain Bayesian estimates of the parameters of the regression model, $\alpha, \beta_1, \ldots, \beta_k$, and σ_ε, we require prior distributions for the parameters. One approach is to use vaguely informative normal priors, remembering that almost all—99.7%—of the density of a normal distribution is within 3 standard deviations of its mean, and that 95% and 68% of the density lie respectively within 2 and 1 standard deviations of the mean. In this approach, the priors for the various parameters are specified separately and are treated as independent.

The standard deviation of the errors, σ_ε, can't be negative, and so we can use a normal prior for its log. Because the intercept α is often far from the observed data, in specifying a prior distribution for α, it often helps first to center the xs at their means or at other meaningful values.

Once the regression model and the priors are specified, the joint posterior distribution of the parameters is approximated by Markov-chain Monte Carlo.

25.3.2 Generalized Linear Models

Bayesian estimation of generalized linear models is largely similar to Bayesian estimation of linear models. As in the preceding section, it's helpful to focus on an example: Cowles and Davis's logistic regression relating volunteering for a psychological experiment to sex, extraversion, and neuroticism (introduced in Section 17.1; see in particular Table 17.1 on page 505). To recapitulate briefly, Cowles and Davis (1987) asked 1421 students in an introductory psychology class whether they were willing to volunteer for an experiment. Sex was coded as either female or male, and the two personality dimensions, extraversion and neuroticism, were measured on 0-to-24 scales.

To facilitate specification of vaguely informative prior distributions for the logistic-regression coefficients, I'll write the logistic-regression model as

$$Y_i|x_{i1}, x_{i2}, x_{i3} \sim \text{Bernoulli}(\pi_i)$$
$$Y_i, Y_{i'} \text{ independent for } i \neq i'$$
$$\pi_i = \frac{1}{1 + e^{-\eta_i}}$$
$$\eta_i = \alpha + \beta_1 x_{i1} + \beta_2 (x_{i2} - 12) + \beta_3 (x_{i3} - 12) + \beta_4 (x_{i2} - 12)(x_{i3} - 12)$$

where

- the response Y is coded 1 for those who volunteered and 0 otherwise;
- $\pi = \Pr(Y = 1)$ is the probability of volunteering;
- x_1 is coded $-\frac{1}{2}$ for males and $\frac{1}{2}$ for females, so that β_1 represents the difference in volunteering between females and males (on the logit scale) at fixed levels of extraversion and neuroticism;
- x_2, extraversion, and x_3, neuroticism, are each centered at 12—the midpoint of their scales;
- $(x_2 - 12)(x_3 - 12)$ is an interaction regressor; and
- the intercept α consequently represents the level of volunteering (on the logit scale) averaged across females and males when extraversion and neuroticism are at the centers of their scales.

When I fit a logistic regression by maximum likelihood to the Cowles and Davis volunteering data, as reported in Table 17.1, extraversion and neuroticism weren't centered, and sex was represented by a dummy regressor, coded 0 for females and 1 for males. As I mentioned, centering the numeric explanatory variables makes it easier to specify prior distributions for the regression coefficients (as I'll show presently). Coding the regressor for sex as $-\frac{1}{2}, +\frac{1}{2}$ has a different, and more fundamental motivation.

When a generalized linear model is fit by maximum likelihood, we know that decisions such as the selection of a baseline category for a set of dummy regressors, or centering a numeric explanatory variable at its mean or another value, are inessential, in the sense that the regression surface fit to the data doesn't depend on these arbitrary choices. The situation can be more complex, however, when we specify prior distributions for regression coefficients to obtain Bayesian estimates.

Suppose, for example, that I were to set the regressor x_1 for sex to 0 for females and to 1 for males, and were to specify prior distributions of the form $\alpha \sim (a, s_0^2)$ and $\beta_1 \sim N(0, s_1^2)$ for the intercept and sex coefficient, where a, s_0^2, and s_1^2 are specific numbers, representing the prior mean of α and the prior variances of α and β_1. Then the prior distribution of the intercept for the baseline female group would be $\alpha \sim N(a, s_0^2)$, while that for the male group would be $(\alpha+\beta_1) \sim N(a, s_0^2+s_1^2)$. That is, the symmetry we expect in maximum-likelihood estimation induced by the arbitrary choice of baseline level for the sex dummy regressor is lost in Bayesian estimation, where the prior variance of the intercept for males is necessarily larger than that for female.

In contrast, coding the regressor for sex as $-\frac{1}{2}, +\frac{1}{2}$ preserves symmetry in the prior distributions of the intercepts for the two groups: Using the same priors as before for the two parameters, $\alpha \sim N(a, s_0^2)$ and $\beta_1 \sim N(0, s_1^2)$, the prior distributions of the intercepts for females and males now both have the same variance; the priors for the intercepts are $(\alpha - \frac{1}{2} \times \beta_1) \sim N(a, s_0^2 + \frac{1}{4} \times s_1^2)$ for females and $(\alpha + \frac{1}{2} \times \beta_1) \sim N(a, s_0^2 + \frac{1}{4} \times s_1^2)$ for males. I'll return to this issue in Section 25.4, where I'll address it in more detail.

Table 25.2 Maximum-Likelihood and Bayesian Estimates for Cowles and Davis's Logistic Regression.

	Estimated Parameter[a] *(Standard Error*[b]*)*				
Method	α	β_1	β_2	β_3	β_4
Maximum Likelihood	−0.382	0.247	0.0642	0.0081	−0.00855
	(0.056)	(0.112)	(0.0144)	(0.0115)	(0.00293)
Bayes, Flat Priors	−0.385	0.247	0.0645	0.0081	−0.00859
	(0.057)	(0.112)	(0.0143)	(0.0117)	(0.00294)
Bayes, Vague Priors	−0.380	0.245	0.0597	0.0074	−0.00858
	(0.056)	(0.111)	(0.0138)	(0.0115)	(0.00295)

[a,b]For Bayesian estimates, posterior mean and standard deviation.

Table 25.2 shows three sets of estimated parameters and standard errors (or posterior standard deviations) for the Cowles and Davis logistic regression: maximum-likelihood estimates, Bayesian estimates with flat priors for the regression coefficients, and Bayesian estimates with vaguely informative priors. As expected, the ML estimates and the Bayesian estimates with flat priors are virtually identical; the results for the Bayesian estimates with vague priors are also very similar to the ML estimates. An effect plot for Cowles and Davis's logistic regression estimated by maximum likelihood appears in Figure 17.2 (on page 506).

As elsewhere in this chapter, the vague priors used here are normal distributions centered at 0, and producing them for Cowles and Davis's logistic regression requires some thought:

- Setting $\alpha \sim N(0, 1)$ effectively confines the logit of volunteering at central values of the personality dimensions to lie between −3 and 3, which translate respectively to probabilities of .05 and .95, a large range but certainly suitable for a probability.[45] Also see Table 25.3 for the implied probabilities of volunteering at $\pm 1, 2$, and 3 prior standard deviations, and Figure 25.19 (a) for a graph of the prior for α.

- Recall that an exponentiated logistic-regression coefficient is interpretable as a multiplicative effect on the odds. As explained, β_1 is the difference in volunteering on the logit scale between females and males, and so e^{β_1} is the multiplicative sex effect. Adopting the prior distribution $\beta_1 \sim N(0, 1)$ corresponds to multiplicative effects of $\exp(-3 \times 1) \approx 0.05$ at 3 standard deviations below the prior mean of 0 and $\exp(3 \times 1) \approx 20$ at 3 standard deviations above the prior mean, which seem reasonable, if broad, prior constraints. Again, more detail is given in Table 25.3 and Figure 25.19 (b).

[45]As it turned out, 42% of the students volunteered to participate in an experiment.

Table 25.3 Summaries of Vaguely Informative Prior Distributions for the Parameters of Cowles and Davis's Logistic Regression

	Prior SDs, Z					
	-3	-2	-1	1	2	3
$\pi = 1/[1 + e^{-Z \times \text{SD}(\alpha)}]$	.047	.12,	.27	.73	.88	.95
$e^{Z \times \text{SD}(\beta_1)}$	0.050	0.13	0.37	2.7	7.4	20
$e^{24 \times Z \times \text{SD}(\beta_j)}, j = 2, 3, 4$	0.027	0.091	0.30	3.3	11	37

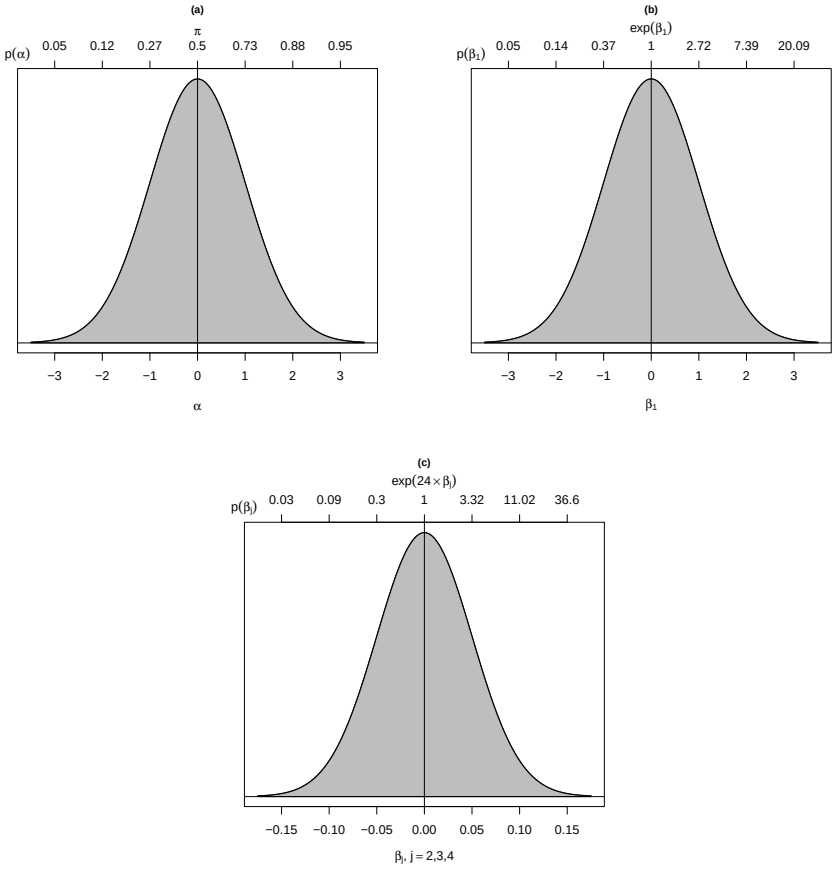

Figure 25.19 Vaguely informative priors for the coefficients of the logistic-regression model fit to Cowles and Davis's data on volunteering for a psychological experiment.

- The coefficients of x_2 (extraversion) and x_3 (neuroticism) are harder to address. Each variable has a range of 24 units.

  Let's consider the extraversion coefficient, β_2, and, for the moment, ignore the interaction. Because extraversion and neuroticism are centered at 12, β_2 is the extraversion logit slope when neuroticism is at the middle of its scale. Employing the normal prior $\beta_2 \sim N(0, 0.05^2)$ implies that the multiplicative effect of extraversion over the 24 units of its scale is effectively constrained to lie between $\exp(24 \times -3 \times 0.05) \approx 0.03$ and $\exp(24 \times 3 \times 0.05) \approx 37$, which once more seem reasonably vague constraints.

  Similar reasoning applies to neuroticism, and so $\beta_3 \sim N(0, 0.05^2)$. As before, more information about these priors appears in in Table 25.3 and Figure 25.19 (c).

- Finally, the coefficient β_4 of the interaction regressor $(x_2 - 12)(x_3 - 12)$ is the change in the slope of x_1 (on the logit scale) associated with a 1-unit increase in x_2 (or the change in the slope of x_2 for a 1-unit increase in x_1). Over the whole 24-unit range of x_1, the multiplicative change in its slope is $\exp(24 \times \beta_4)$, and so setting the prior distribution to $\beta_4 \sim N(0, 0.05^2)$ restricts the multiplicative interaction to the range 0.03 to 37 (also see Table 25.3 and Figure 25.19 (c)).

Bayesian estimation of generalized linear models is very similar to Bayesian estimation of linear models: The GLM provides a probability model for the data (see Section 15.1):

$$\begin{aligned}\eta_i &= \alpha + \beta_1 x_{i1} + \cdots \beta_k x_{ik} \\ \mu_i &= g^{-1}(\eta_i) \\ Y_i | x_{i1}, \ldots, x_{ik} &\sim p(\mu_i, \phi)\end{aligned}$$

where the conditional distribution $p(\mu_i, \phi)$ of the response Y_i is a member of an exponential family, with expectation μ_i and dispersion parameter ϕ (which recall is set to 1 in the binomial and Poisson families).
Once prior distributions for the parameters of the model are specified, their posterior distribution can be approximated by MCMC.

Bias Reduction in Logistic Regression

An interesting application of Bayesian ideas is to bias reduction in the estimation of generalized linear models, specifically in logistic regression (another context

Table 25.4 Maximum-Likelihood and Bayesian Estimates of a Logistic Regression Model with Complete Separation of the Data (see Figure 25.20).

	Estimated Parameter (Standard Error[a])		
Method	α	β_1	β_2
Maximum Likelihood	0	$-\infty$	$-\infty$
	(—)	(—)	(—)
Firth (Bayes, Jeffreys Prior)	0.343	−31.0	−31.8
	(0.869)	(14.0)	(13.9)
Bayes, Diffuse Normal Prior	0.328	−30.1	−31.7
	(0.852)	(8.5)	(8.7)

[a]For Bayesian estimates, posterior standard deviation.

in which a frequentist criterion—bias—is used to justify a Bayesian estimator). Maximum-likelihood estimators are asymptotically unbiased (see, e.g., on-line Appendix Section D.6.2), but they can be biased in small samples. In logistic regression, small-sample bias can be acute when the probability of the response is close to 0 or 1.

An especially problematic data pattern for logistic regression is *complete separation*, where a linear function of the regressors in the model partitions the data into disjoint regions of 0s (i.e., "failures") and 1s ("successes"), as illustrated with artificial data for two xs in Figure 25.20.[46] For the case depicted in the graph, the maximum-likelihood estimates of β_1 and β_2 in the logistic regression of Y on the two xs are both $-\infty$, and if you try to estimate the logistic regression model with statistical software, the computation will fail to converge.

Firth (1993) (also see Kosmidis and Firth, 2009, 2021) showed that substantial bias reduction in estimating the logistic regression model under difficult circumstances can be achieved by employing the Jeffreys prior, and he suggested an efficient method for computing the resulting estimates. Firth's estimates for the data in Figure 25.20 are shown in Table 25.4, along with similar estimates (but smaller posterior standard deviations, labeled "standard errors" in the table) produced by using diffuse normal priors for the logistic-regression coefficients:

$$\alpha \sim N(0, 20^2)$$
$$\beta_1 \sim N(0, 20^2)$$
$$\beta_2 \sim N(0, 20^2)$$

[46]There are $n = 100$ points in this artificial data set, with each of x_1 and x_2 generated by sampling independently from the uniform distribution Unif$(-1, 1)$. Then $Y = 1$ for $x_1 + x_2 < 0$ (i.e., below the $-45°$ line in the graph) and $Y = 0$ otherwise.

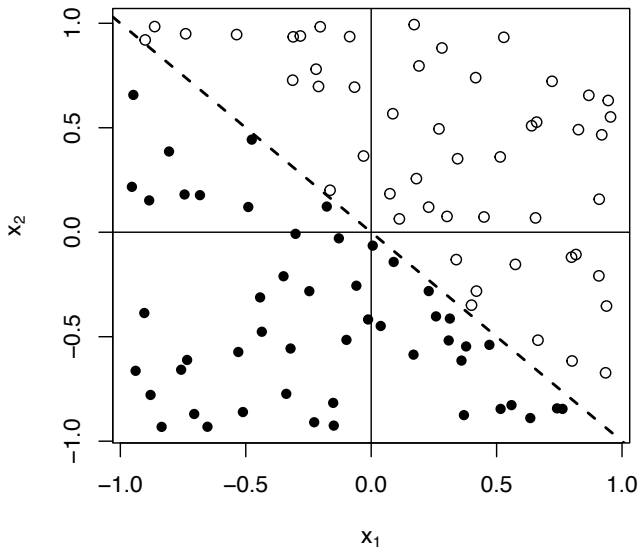

Figure 25.20 Complete separation for a binary response variable; the filled points represent $Y = 1$ ("successes") and the hollow points $Y = 0$ ("failures").

Firth (1993) showed that substantial bias reduction in estimating the logistic regression model under difficult circumstances can be achieved by employing the Jeffreys prior. An especially problematic data pattern for logistic regression to which Firth's estimator is applicable is complete separation, where a linear function of the regressors in the model partitions the data into disjoint regions of 0s and 1s. In this case, maximum-likelihood estimation of the logistic regression coefficients produces one or more infinite estimates, while Firth's method yields finite estimates.

25.4 Mixed-Effects Models

A common application of Bayesian estimation is to mixed-effects models.[47] As in the preceding section on linear and generalized-linear models, I'll focus on a representative application introduced earlier in the text, Davis et al.'s (2005) study of the relationship between the developmental trajectory of exercise and eating disorders in female adolescents (discussed in Section 23.4). Recall that Davis et al. collected retrospective data on weekly hours of exercise, at intervals of 2 years starting at age 8 and ending at the age of hospitalization, for 138 teen-aged female patients who were treated for eating disorders; parallel data were collected for 93 control subjects who did not suffer from eating disorders but were otherwise generally similar to the patients.

The object of the research was to compare the typical exercise trajectories of the patients and the controls. To that end, I fit several mixed-effects models of varying complexity to Davis et al.'s data. An initial model, with potentially different fixed-effects intercepts and age slopes for the two groups (patients and controls), along with random effects for person-specific intercepts and slopes, appears in equation and tabular form on page 721, with estimates by restricted maximum likelihood (REML). The model is parametrized so that the intercept represents average exercise in the control group at age 8 (i.e., the start of the study). The response variable in the model, hours of weekly exercise, is log-transformed using logs to the base 2, after adding 5 minutes to weekly exercise to avoid taking the log of 0. The fixed-effects part of the model—averaging over the random effects and adjusting the notation slightly to conform to usage in

[47] A Bayesian might object to the term "mixed effects," which is rooted in the frequentist distinction between fixed effects, which are conceptualized as parameters with fixed though unknown values, and random effects (e.g., individual-specific regression coefficients), which are random variables whose variance and covariance components are parameters. Bayesians treat *all* unknown quantities as random variables. A Bayesian, therefore, might emphasize the application, and term such a model "hierarchical," "multilevel," or "longitudinal" rather than calling it a "mixed-effects" model. The key point, however, is that the random effects are expressed in terms of more fundamental parameters, the variance and covariance components, and so I continue to find the distinction between fixed and random effects useful.

the current chapter—is (adapting Equation 23.11)

$$E(Y_{ij}) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2ij} + \beta_3 x_{1i} x_{2ij} \tag{25.30}$$

where Y_{ij} is log-exercise for individual i on the jth measurement occasion; x_{1i} is a dummy regressor coded 0 for individuals i in the control group and 1 for those in the patient group; and x_{2ij} is individual i's age on measurement occasion j, less 8 years.

When we specify and interpret linear models fit by least squares or generalized linear models fit by maximum likelihood, we rely on certain symmetries and invariances. For example, the baseline level for a dummy regressor or set of dummy regressors is an inessential choice, in that the regression surface fit to the data doesn't depend on the baseline level selected. Moreover, as long as we use a complete set of contrasts to represent a factor in a linear or generalized-linear model, the details of the coding employed are inessential, and so we can use deviation regressors or arbitrary contrasts in lieu of dummy regressors (see Sections 8.2.4, 8.5, and 9.1.2). Similarly, when we fit a linear or generalized linear model in which there are higher-order terms, such as interactions, details of model specification such as contrast coding for factors and centering numeric explanatory variables are inessential as long as the model conforms to the principle of marginality(discussed in Section 7.3.2). These observations extend to linear and generalized-linear mixed models fit by ML or REML, and so, in the current example, it doesn't matter that the control group rather than the patient group is selected as the baseline: The model implies two regression lines, one for the control group and the other for patients, potentially with different intercepts and slopes. The lines fit don't depend on which group is selected as the baseline for the group dummy regressor, nor on centering age at 8 rather than at some other value.

Once we specify prior distributions for regression parameters, however, and proceed to estimate a model by Bayesian methods, these symmetries and invariances generally no longer hold. We're already familiar with parametrizing a Bayesian regression model to facilitate the specification of priors. For example, in specifying a normal linear model for Duncan's occupational prestige regression in Section 25.3.1, I centered the numeric explanatory variables income and education to 0 means to simplify the specification of a prior distribution for the intercept in the regression model.

The issue raised here is more general, however. I made a similar, if less general, point in Section 25.3.2, in connection with selecting the baseline level for a dummy regressor in Bayesian estimation of a logistic-regression model where the dummy regressor entered the regression equation additively. Consider, now, the fixed effects in Equation 25.30, and suppose that, as described, the baseline level for the dummy regressor x_1 is the control group. We proceed to specify

independent priors for the regression coefficients, say

$$\begin{aligned}\beta_0 &\sim N(0, s_0^2)\\ \beta_1 &\sim N(0, s_1^2)\\ \beta_2 &\sim N(0, s_2^2)\\ \beta_3 &\sim N(0, s_3^2)\end{aligned}$$

where the s_j^2s are specific numbers representing prior variances. Then the intercepts and slopes of the regression lines for the two groups have different prior variances. In particular, for the control group, the prior variance of the intercept is $V(\beta_0) = s_0^2$ and the prior variance of the age slope is $V(\beta_2) = s_2^2$, while the corresponding variances for the patient group are necessarily larger, $V(\beta_0 + \beta_1) = s_0^2 + s_1^2$ and $V(\beta_2 + \beta_3) = s_2^2 + s_3^2$, respectively. Notice that this issue isn't specific to mixed-effects models and would apply to linear and generalized-linear models as well.

As far as I'm aware, this isn't a commonly discussed problem, even though it must occur frequently in applications of Bayesian regression modeling. McElreath (2020, Section 8.1) raises the problem (which is how I became aware of it), but his solution is, I believe, insufficiently general: McElreath suggests parametrizing the model so that distinct intercepts and slopes for the two groups appear directly in the model, in which case $E(Y_{ij}) = \beta_{0g} + \beta_{1g}x_{1ij}$, where $g = 1$ for individuals in the control group and 2 for those in the patient group, and x_1 is now age. We would proceed to specify priors for β_{01}, β_{02}, β_{11}, and β_{12} (and would probably make the prior variances of the two intercepts the same, and the prior variances of the two slopes the same). This solution is perfectly fine for simple models, but breaks down, for example, when there are higher-order interactions.

A general solution to this problem is complex (see the starred paragraph below), but the current example suggests how we might proceed. Instead of using 0/1 dummy-regressor coding suppose we code the contrast $x_{i1} = -\frac{1}{2}$ for individuals i in the control group and $+\frac{1}{2}$ for those in the patient group. We retain the specification $E(Y_{ij}) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2ij} + \beta_3 x_{1i} x_{2ij}$, and so β_1 still captures the difference in intercepts between the patient and control groups, while β_3 captures the difference in slopes. Now, however, $E(Y_{ij}) = (\beta_0 - \frac{1}{2}\times\beta_1) + (\beta_2 - \frac{1}{2}\times\beta_3)x_{2ij}$ for individuals in the control group; $E(Y_{ij}) = (\beta_0 + \frac{1}{2}\times\beta_1) + (\beta_2 + \frac{1}{2}\times\beta_3)x_{2ij}$ for those in the patient group; and the corresponding prior variances of the intercepts and slopes in the two groups are the same: $V(\beta_0 - \frac{1}{2}\times\beta_1) = V(\beta_0 + \frac{1}{2}\times\beta_1) = s_0^2 + \frac{1}{4}\times s_1^2$ for the intercepts; and $V(\beta_2 - \frac{1}{2}\times\beta_3) = V(\beta_2 + \frac{1}{2}\times\beta_3) = s_2^2 + \frac{1}{4}\times s_3^2$ for the slopes.

*More generally, if a factor has $m \geq 2$ levels, we require $m - 1$ regressors to represent it in a linear or generalized linear model. As I have explained, choice of the specific set of regressors employed is inessential in least-squares or maximum-likelihood estimation, but not in Bayesian estimation. To achieve the symmetry and invariance that we require in Bayesian estimation, we may use a set of $m - 1$ regressors that are (1) orthogonal to the constant regressor, and

(2) orthonormal in their row-basis. For $m = 2$ levels, this produces a regressor coded $-\sqrt{\frac{1}{2}}, \sqrt{\frac{1}{2}}$ (or its reflection, $\sqrt{\frac{1}{2}}, -\sqrt{\frac{1}{2}}$), which is simply a multiple of the solution that I employ in the current application, that is, $-\frac{1}{2}, +\frac{1}{2}$. For $m = 3$, the row-basis of the regressors takes the form

$$\begin{bmatrix} \sqrt{\frac{1}{2}} & -\sqrt{\frac{1}{6}} \\ 0 & 2\sqrt{\frac{1}{6}} \\ -\sqrt{\frac{1}{2}} & -\sqrt{\frac{1}{6}} \end{bmatrix}$$

up to a permutation of the rows and reflections of the columns. In the general case, we can construct a suitable $(m-1)$-column row-basis for the contrast matrix by generating an arbitrary set of $m-1$ orthogonal contrasts, and then normalizing the columns to unit lengths (or to other fixed common lengths). This approach is particularly compelling for independent normal priors, which are completely specified by their means and variances.

With these preliminaries out of the way, let's formulate a Bayesian linear mixed-effect model for the Davis et. al data. As usual, Bayesian analysis begins with specification of a probability model for the observed data, for which I'll retain the notation of the Laird-Ware form of the linear mixed-effects model (introduced in Section 23.2):[48]

$$\begin{aligned} &Y_{ij}|x_{i1}, x_{i2} \sim N(\mu_{ij}, \sigma_\varepsilon^2) \\ &Y_{ij}, Y_{i'j'} \text{ independent for } i \neq i' \text{ or } j \neq j' \\ &\mu_{ij} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2ij} + \beta_3 x_{1i} x_{2ij} + B_{0i} + B_{1i} x_{2ij} \\ &\mathbf{b}_i \equiv [B_{0i}, B_{1i}]' \sim N_2\left([0,0]', \underset{2\times 2}{\mathbf{\Psi}}\right) \\ &\mathbf{b}_i, \mathbf{b}_{i'} \text{ independent for } i \neq i' \end{aligned} \tag{25.31}$$

- the Y_{ij} are the values of the response, log-exercise, for individual i on occasion j, and are normally and independently distributed with expectation μ_{ij} and common error variance σ_ε^2;
- the βs represent fixed effects;
- x_{1i} is a contrast regressor coded $\frac{1}{2}$ for patients and $-\frac{1}{2}$ for control subjects;
- x_{2ij} is patient i's age in years, starting at age 8, on measurement occasion j (i.e., $\text{age}_{ij} - 8$);
- $x_{1i}x_{2ij}$ is therefore an interaction regressor, permitting different fixed-effect age slopes in the patient and control groups;

[48] If you're unfamiliar with matrix notation, just think of $\mathbf{b}_i$ as a list of the two person-specific regression coefficients for individual i, and $\mathbf{\Psi}$ as a table of the variance-covariance components (see Equation 25.32).

- B_{0i} is the deviation of the intercept for individual i from the fixed-effects intercept $\beta_0 - \frac{1}{2} \times \beta_1$, for the control group, or from $\beta_0 + \frac{1}{2} \times \beta_1$, for the patient group; and

- B_{1i} is the deviation of the age slope for individual i from the fixed-effects slope $\beta_2 - \frac{1}{2} \times \beta_3$, for the control group, or from $\beta_2 + \frac{1}{2} \times \beta_3$, for the patient group.

The covariance matrix $\mathbf{\Psi}$ of the random effects $\mathbf{b}_i$ depends on three fundamental parameters, which I'll take as $\text{SD}(B_{0i}) \equiv \psi_0$, $\text{SD}(B_{1i}) \equiv \psi_1$, and $\text{cor}(B_{0i}, B_{1i}) \equiv \rho_{01}$, so that

$$\mathbf{\Psi} = \begin{bmatrix} \psi_0^2 & \rho_{01}\psi_0\psi_1 \\ \rho_{01}\psi_1\psi_0 & \psi_1^2 \end{bmatrix} \tag{25.32}$$

Table 25.5 shows three sets of estimates for this model: Maximum-likelihood estimates, based only on the data;[49] Bayesian estimates based on flat priors; and Bayesian estimates based on the following weakly informative priors (deferring, for the moment, the prior for the random-effects correlation parameter ρ_{01}):

$$\begin{aligned} \beta_j &\sim N(0, 0.5^2) \text{ for } j = 0, 1, 2 \\ \beta_3 &\sim N(0, 1) \\ \log_e \psi_j &\sim N(0, 0.4^2) \text{ for } j = 0, 1 \\ \log_e \sigma_\varepsilon &\sim N(0, 0.4^2) \end{aligned}$$

As usual, I specified normal priors centered on 0. The prior standard deviations take account of the scale of the response, which is log-base-2 hours per week of exercise: An increase of 1 on the log-2 scale, for example, doubles exercise, while a decrease of 1 halves exercise.

- Appealing to the $\pm 3 \times \text{SD}$ rule as defining the effective prior limits on a parameter, and setting the prior standard deviation $\text{SD}(\beta_0) = 0.5$ constrains the general intercept (i.e., average exercise at age 8) to lie between $2^{-3\times 0.5} \approx \frac{1}{3}$rd of an hour and $2^{3\times 0.5} \approx 3$ hours (less, in each case, the start of 5 minutes used to avoid taking the log of 0). Checking the prior limits more generally,

Prior SDs	-3	-2	-1	1	2	3
Hours of exercise	0.353	0.5	0.707	1.41	2	2.83

- Setting $\text{SD}(\beta_1) = 0.5$ constrains the intercept for the patient group to differ from that of the control by a multiplicative factor of from $2^{-3\times 0.5} \approx \frac{1}{3}$ to $2^{3\times 0.5} \approx 3$.

[49] The ML estimates in Table 25.5 are similar, but not identical, to the REML estimates given in the table on page 721; note as well that the latter shows the estimated *covariance* ψ_{01} of the random-effect intercepts and slopes rather than their *correlation* ρ_{01}. Standard errors of variance and covariance components are of limited usefulness and are not shown.

- Similarly, setting $\mathrm{SD}(\beta_2) = 0.5$ constrains the general age slope to a decline or increase in exercise by a factor of at most 3 per year. The table for β_0, shown above, is also applicable to β_1 and β_2, both of which set the prior standard deviation to 0.5, treating the numbers given in the table as multiplicative effects.

- Finally, setting $\mathrm{SD}(\beta_3) = 1$ allows the age slope in the patient group to be smaller or larger than that in the control group by a factor of at most $2^3 = 8$; more generally, checking at ± 1, 2, and 3 prior SDs,

Prior SDs	-3	-2	-1	1	2	3
Multiplicative factor	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{2}$	2	4	8

The standard deviations of the normal priors for the random-effect log standard deviations, $\log_e \psi_0$ and $\log_e \psi_1$, and for the error log standard deviation, $\log_e \sigma_\varepsilon$, are harder to understand because the response is itself on the log-base-2 scale. I've taken these prior standard deviations as 0.4, which implies the following multiplicative constraint on the original hours-per week scale of the response: from $2^{\exp(-3\times 0.4)} \approx 1.2$ to $2^{\exp(3\times 0.4)} \approx 10$, which is quite broad—ranging over nearly an order of magnitude. More generally,

Prior SDs	-3	-2	-1	1	2	3
Multiplicative factor	1.23	1.37	1.59	2.81	4.68	9.99

The correlation ρ between the random-effect intercept B_0 and slope B_1 also merits special treatment because, as a correlation, it's bounded between -1 and 1, and thus isn't naturally modeled with a normal prior. R. A. Fisher's z-transformation of the correlation (introduced in Fisher, 1915),

$$z(\rho) \equiv \tfrac{1}{2} \log_e \frac{1+\rho}{1-\rho} = \operatorname{arctanh}(\rho)$$

maps ρ to $(-\infty, \infty)$ and serves to normalize its distribution. I then used the prior $z(\rho_{01}) \sim N(0, 0.6)$ which effectively constrains ρ_{01} between $\tanh(-3 \times 0.6) \approx -.95$ and $\tanh(3 \times 0.6) \approx .95$.[50] Checking more generally at ± 1, 2, and 3 prior standard deviations,

Prior SDs	-3	-2	-1	1	2	3
ρ_{01}	$-.947$	$-.834$	$-.537$	.537	.834	947

[50]*This trick works because there are only two sets of random-effect coefficients in the model, the B_{0i}s and the B_{1i}s. More generally, the random-effect covariance matrix $\boldsymbol{\Psi}$ can be larger than 2×2 but must be positive-definite. It consequently requires a parametrization and prior distribution that preserve positive-definiteness. See the references at the end of the chapter for further discussion, in particular, the LKJ prior for the *correlation* matrix of the random effects described by Gelman et al. (2013, pages 578, 584).

You may be unfamiliar with the *hyperbolic tangent function*, tanh, and its inverse, the arctanh function, which is used for Fisher's z-transformation. The hyperbolic tangent of x is defined as

$$\tanh(x) \equiv \frac{e^{2x} - 1}{e^{2x} + 1}$$

and $\operatorname{arctanh}(y) = x$ when $\tanh(x) = y$.

The prior distributions for the fixed effects, error standard deviation, random-effect standard deviations, and random-effect correlation are graphed in Figure 25.21.

The ML estimates and the Bayesian estimates using flat priors are similar to one-another (see Table 25.5), as is to be expected. The Bayesian estimates produced by the vaguely informative priors just discussed are also similar to the other two sets of estimates.

I could take this example in several directions. In Section 23.4, for example, I considered different structures for the random effects in the model along with the possibility of serially correlated intra-individual errors. Here I'll pursue another issue in the Davis et al. data: the relatively large number of 0 responses (about 12% overall). To this point, I've dealt with the 0s by adding 5 minutes to each exercise value prior to taking logs to reduce the positive skew in the distribution of the response. I'll instead now specify a so-called *hurdle model*, which takes the 0s explicitly into account, and which is similar in spirit to the zero-inflated Poisson and negative-binomial models for overdispersed count data discussed in Section 15.2.1.[51]

The mixed-effects hurdle model consists of two components—a logistic-regression component for modeling the 0s, and a linear-regression component for modeling the nonzero measurements:[52]

$$
\begin{aligned}
&N_{ij}|x_{i1}, x_{i2} \sim \text{Bernoulli}(\pi_{ij}) \\
&N_{ij}, N_{i'j'} \text{ independent for } i \neq i' \text{ or } j \neq j' \\
&\eta_{ij} = \xi_0 + \xi_1 x_{1i} + \xi_2 x_{2ij} + \xi_3 x_{1i} x_{2ij} + Z_{0i} + Z_{1i} x_{2ij} \\
&\pi_{ij} = \frac{1}{1 + e^{-\eta_{ij}}} \\
&\mathbf{z}_i \equiv [Z_{0i}, Z_{1i}]' \sim N_2\left([0, 0]', \underset{2\times 2}{\mathbf{\Psi_z}}\right) \\
&\mathbf{z}_i, \mathbf{z}_{i'} \text{ independent for } i \neq i' \\
\log_2 Y_{ij}|(Y_{ij} > 0, x_{i1}, x_{i2}) &\sim N(\mu_{ij}, \sigma^2_\varepsilon) \\
&Y_{ij}, Y_{i'j'} \text{ independent for } i \neq i' \text{ or } j \neq j' \\
&\mu_{ij} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2ij} + \beta_3 x_{1i} x_{2ij} + B_{0i} + B_{1i} x_{2ij} \\
&\mathbf{b}_i \equiv [B_{0i}, B_{1i}]' \sim N_2\left([0, 0]', \underset{2\times 2}{\mathbf{\Psi_b}}\right) \\
&\mathbf{b}_i, \mathbf{b}_{i'} \text{ independent for } i \neq i'
\end{aligned}
$$

where

[51] I don't want to imply, however, that a hurdle model *requires* a Bayesian approach—it's possible, with proper software (and such software exists), to fit a mixed-effects hurdle model by maximum likelihood.

[52] As before, if you're unfamiliar with matrix notation, just think of $\mathbf{z}_i$ and $\mathbf{b}_i$ as lists of individual-specific regression coefficients and $\mathbf{\Psi_z}$ and $\mathbf{\Psi_b}$ as tables of the variance and covariance components for the corresponding random effects.

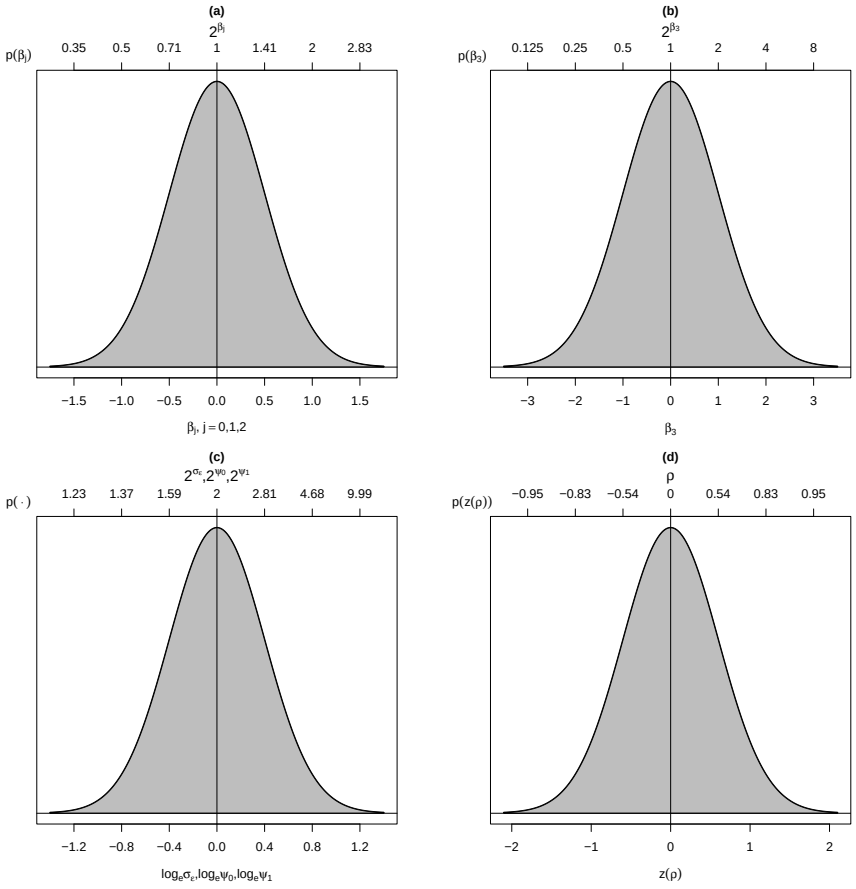

Figure 25.21 Vaguely informative priors for the parameters of the simple mixed-effects model fit to Davis et al.'s data on exercise and eating orders.

Table 25.5 Maximum-Likelihood and Bayesian Estimates for a Simple Mixed-Effects Model Fit to Davis et al.'s Exercise Data.

	Estimated Parameter (Standard Error[a]*)*							
Method	β_0	β_1	β_2	β_3	ψ_1	ψ_2	ρ_{12}	σ_ε
Maximum Likelihood[b]	−0.285	−0.226	0.161	0.201	0.992	0.131	−0.0617	0.879
	(0.082)	(0.163)	(0.015)	(0.029)				
Bayes, Flat Priors	−0.286	−0.226	0.161	0.201	0.993	0.129	−0.0239	0.886
	(0.082)	(0.165)	(0.015)	(0.029)				
Bayes, Vague Priors	−0.284	−0.227	0.161	0.201	1.002	0.132	−0.0424	0.884
	(0.083)	(0.167)	(0.015)	(0.029)				

[a] For Bayesian estimates, posterior standard deviation.

[b] Cf., the REML estimates on page 721.

- the dichotomous response $N_{ij} = 0$ if exercise is 0 for individual i on occasion j and $N_{ij} = 1$ if exercise is nonzero;
- η_{ij} is the linear predictor for the logit part of the model;
- the ξs are fixed effects in the logit part of the model;
- the Zs are random-effects coefficients in the logit part of the model;
- the elements of $\mathbf{\Psi_z}$ are variance and covariance components for the random effects in the logit part of the model; and
- Y_{ij}, μ_{ij}, the βs, the Bs, and $\mathbf{\Psi_b}$ (replacing $\mathbf{\Psi}$) are defined as in Equations 25.31 (on page 58), but only for nonzero exercise, where Y_{ij} is now hours of exercise *without* the necessity of adding a start to the exercise values to avoid the log of 0.

I'll use the same priors as before for the linear part of the model, along with the following vague priors for the logit part of the model:

- ξ_0 is the average logit of nonzero exercise at age 8. I know that overall about $\frac{1}{8}$ of exercise measurements are 0s and thus about $\frac{7}{8}$ are nonzero, but the proportion of 0s is likely greater at age 8 than overall.[53] To use the data to help specify the prior is problematic from a Bayesian perspective,[54] however, and so I'll disregard this information, and simply specify the normal prior $\xi_0 \sim N(0, 1)$. Then three prior standard deviations correspond to logits between -3 and 3, which translate to probabilities of nonzero exercise between approximately .05 and .95, a broad prior range. More generally, we have the following probabilities of nonzero exercise at $\pm 1, 2$, and 3 prior SDs:

Prior SDs	-3	-2	-1	1	2	3
π	.047	.119	.269	.731	.881	.953

- ξ_1 is the average difference in the logit of nonzero exercise between eating-disordered and control subjects at age 8. I don't expect a large difference at such an early age, but taking $\xi_1 \sim N(0, 0.5^2)$ constrains the relative odds of nonzero exercise for the two groups to lie between $\exp(-3 \times 0.5) = 0.22$ and $\exp(3 \times 0.5) = 4.5$, that is, to differ by a factor of no more than about 5, which is certainly not very restrictive. More generally,

Prior SDs	-3	-2	-1	1	2	3
relative odds	0.22	0.37	0.61	1.65	2.7	4.5

[53] Of course, we don't have to speculate because we have the data. I invite the reader to find the percentage of 0s at age 8.

[54] For more on this point, see Section 25.5.

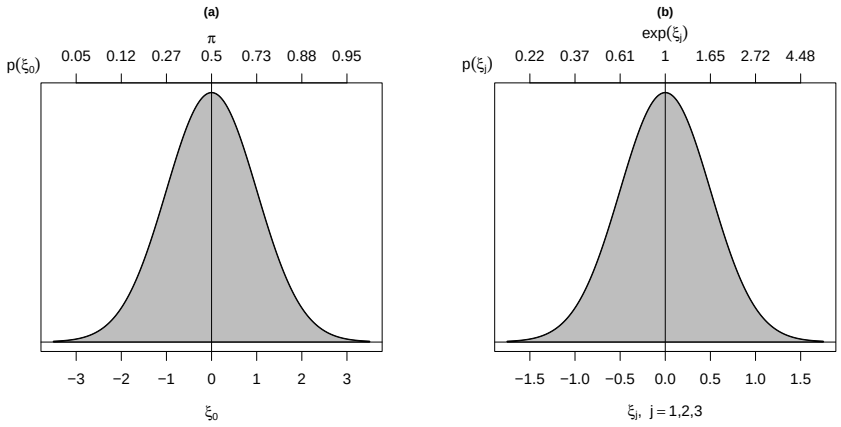

Figure 25.22 Vaguely informative priors for the fixed-effect intercept (a) ξ_0 and (b) coefficients ξ_1, ξ_2, and ξ_3 in the logit part of the hurdle model fit to Davis et al.'s data on exercise and eating orders.

- ξ_2 is the average age slope of nonzero exercise on the logit scale. Using the prior $\xi_2 \sim N(0, 0.5^2)$ therefore permits the odds of nonzero exercise to either increase or decrease by at most a factor of 5 per year (and the relative odds at ± 1, 2, and 3 prior SDs are as in the table for ξ_1 immediately above).
- ξ_3 is the average difference in the age slope of nonzero exercise between eating-disordered and control subjects on the logit scale. Once more using the prior $\xi_2 \sim N(0, 0.5^2)$ constrains the odds of nonzero exercise to increase or decrease in eating-disordered subjects up to 5 times more slowly or rapidly in comparison to control subjects.[55] Again, the relative odds at ± 1, 2, and 3 prior SDs are as in the table for ξ_1.
- I'll use the same prior distributions for the variance-covariance components of the random effects in the logit part of the model as in the linear part of the model, thus taking $\log_e \psi_{Z_j} \sim N(0, 0.4^2)$ for $j = 0, 1$ and $z(\rho_{Z_0 Z_1}) \sim N(0, 0.6^2)$.

The priors for the fixed-effect coefficients in the logit part of the hurdle model are shown in Figure 25.22.

As usual, I estimated the posterior distribution of the parameters by Hamiltonian Monte Carlo, using four independent Markov chains of 10,000 iterations each, the first half of which were discarded as warm-up, leaving $m = 20,000$ sampled values of each parameter. The hurdle model, as specified, proved difficult

[55] The researchers' expectation was that exercise would increase more rapidly with age in the eating-disordered group.

Table 25.6 Posterior Means and Standard Deviations for the Mixed-Effects Hurdle Model fit to the Davis et al. Data.

Parameter	Posterior Mean	Posterior SD
β_0	−0.304	0.107
β_1	−0.186	0.199
β_2	0.209	0.018
β_3	0.210	0.035
ξ_0	1.782	0.161
ξ_1	−0.491	0.274
ξ_2	0.296	0.085
ξ_3	0.163	0.105
σ_ε	1.022	
ψ_{B_0}	1.352	
ψ_{B_1}	0.151	
$\rho_{B_0B_1}$	−0.478	
ψ_{Z_1}	0.394	

to estimate, with two of the parameters having unacceptably small estimated effective sample sizes, $\widehat{m}_{\text{eff}} = 465$ for ψ_{Z_0} and $\widehat{m}_{\text{eff}} = 236$ for $\rho_{Z_0Z_1}$.

I encountered a similar problem in Section 23.4, where I couldn't fit a mixed model to the Davis et al. data that had random intercepts, random slopes, *and* serially correlated errors. After all, there are few measurements per individual (about 4, on average) to estimate the variance and covariance components, along with parameters modeling serial dependence in the errors.

In the current context, tightening the priors for the variance-covariance components of the logit part of the model might make it possible to estimate the hurdle model.[56] I decided instead to simplify the random-effects structure by eliminating random intercepts from the logit part of the model, retaining random slopes; that reduces the variance-covariance components for this part of the model to the single standard deviation ψ_{Z_1}. The resulting parameter estimates (posterior means and standard deviations) are shown in Table 25.6.

[56] Fiddling with a prior in light of an unsatisfactory posterior could also be regarded from a Bayesian point of view as cheating; see Section 25.5 for further discussion, and Exercise 25.7.

A common application of Bayesian estimation is to mixed-effects models of various kinds. The linear, generalized-linear, and nonlinear mixed models discussed in Chapters 23 and 24 all provide probability models for data, and these models can be extended in various ways, as illustrated by the hurdle model fit in the current section. With suitable priors for the regression coefficients and variance-covariance components, Bayesian estimates for mixed-effects models can by obtained by MCMC methods. The prior distribution for the variance-covariance components of a mixed model must be suitably parametrized to produce a positive-definite covariance matrix for the random effects.

A convenient by-product of estimating the model by MCMC is that we can calculate quantities derived from the parameters for each sample of the parameter values, providing estimated posterior distributions for these derived quantities. Figure 25.23, for example, is an effect plot for the estimated hurdle model. It is constructed by calculating the linear predictors η and μ for the logit and linear parts of the hurdle model in each Monte-Carlo sample, setting age to 8 through 18 at intervals of 2 years, in combination with the two levels of group (patient and control), producing $6 \times 2 = 12$ values of η and μ.

This computation is repeated for each of the $m = 20,000$ retained HMC samples. The corresponding fitted value of hours of exercise per week in each sample is then the probability that exercise is nonzero—that is, $1/(1 + e^{\eta})$—times estimated hours of exercise per week conditional on nonzero exercise, 2^{μ}. The 12 means across the 20,000 samples are the points in the effect plot in Figure 25.23, while their .025 and .975 quantiles provide 95% credible intervals for the effects. Compare this graph to Figure 23.10 (page 725), which is based on a simpler mixed-effects model for $\log_2(\text{exercise} + 5 \text{ minutes})$ fit by REML (and which evaluates the fitted model between ages 8 and 16 rather than between 8 and 18).

A convenient by-product of estimating a regression model by MCMC is that we can calculate quantities derived from the parameters for each sample of the parameter values, providing estimated posterior distributions of the derived quantities. One application of this idea is to the construction of effect plots.

25.5 Concluding Remarks

First, a couple of caveats: Although I have an interest in the foundations of statistical inference, I'm far from an expert on the subject. You should take these remarks as an expression of my opinion, although, as you'll see, I don't

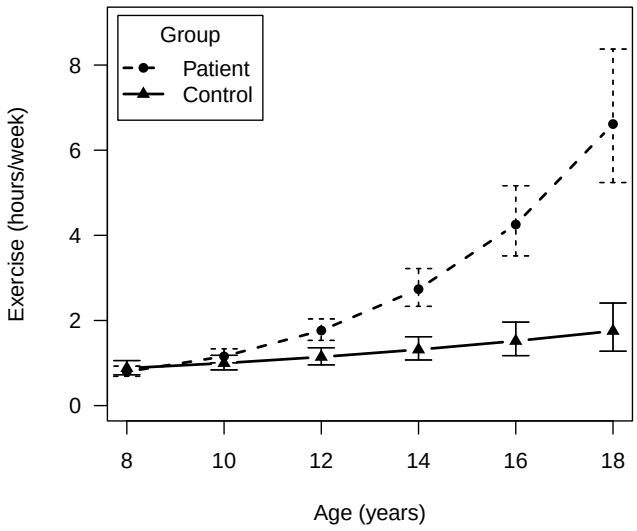

Figure 25.23 Average exercise (displayed as points) as a function of age and group, based on the mixed-effects hurdle model fit to the Davis et al. data. The bars around the points show 95% central credible intervals.

hold strong opinions on the relative merits of the various approaches to statistical inference, and I think that both frequentist and Bayesian inference can be reasonably used in applications.[57]

There is a lot to recommend Bayesian approaches to statistical inference. Quantifying uncertainty in terms of probabilities makes the interpretation of results much more straightforward than in classical inference, where we have to be careful, for example, *not* to treat a p-value as the probability that the null hypothesis is wrong, and *not* to interpret a confidence level as the probability that an unknown parameter lies in a confidence interval. Indeed, in some circumstances, failure to make what appear to be pedantic distinctions can produce grossly distorted applications of classical inference. When you genuinely hold subjective prior beliefs that are consistent with the axioms of probability theory, or can construct a prior distribution based on existing research, Bayesian inference provides clear and correct answers to questions of statistical inference.

The problem that I see with almost all applications of Bayesian inference, however, is that the priors that are used *don't* have compelling subjective justification or support in existing research. Instead, priors are typically justified by vague, if generally plausible, arguments, such as those employed in the examples in this chapter,[58] or are selected automatically. In the latter instance, the prior may use aspects of the data, begging the question of whether it is "prior" at all.[59] I've even encountered an application of Bayesian inference, by a prominent Bayesian (who will remain anonymous), in which the prior distribution of a parameter was altered because the resulting posterior distribution appeared unreasonable, turning Bayesian inference on its head.

Prior distributions for regression coefficients and other parameters (such as the error variance) are typically specified individually and are taken to be independent, so that the joint prior distribution of the parameters of the model is the product of the marginal priors. That is the approach that I used for the examples in this chapter. We know, however, that regression coefficients are almost always correlated, rendering this procedure generally unreasonable even when the marginal priors have strong justification, either subjectively or in existing research. If the joint prior is based on existing research, we probably

[57]This section was strongly influenced by discussions with Georges Monette of York University in Toronto, whose knowledge of the foundations of statistical inference vastly exceeds mine. Of course, Georges isn't responsible for my remarks here.

[58]Indeed, it's my impression that often less thought is given to the formulation of vaguely informative priors than in the examples in this chapter.

[59]There is, however, a principled approach, termed *empirical Bayes estimation*, that uses the data to construct a prior distribution. Empirical Bayes methods are potentially applicable to situations in which parameters are hierarchically related, as is the case in the mixed effects models considered in Section 25.4, where, for example, the distributions of individual-level regression coefficients depend on variance-covariance components, which are termed *hyperparameters*. The justification of empirical Bayes methods typically appeals to reduced mean-squared error of estimation in comparison to classical estimators, a frequentist criterion (see on-line Appendix Section D5.2). Empirical Bayes estimators may be thought of as approximations to Bayesian estimators. Efron and Morris (1977) and Casella (1985) provide nice introductions to empirical Bayes methods. Also see the discussion in footnote 40 (page 44) of conditioning on the sample distribution of the explanatory variables.

also have a basis for addressing dependencies among parameters, but if the marginal priors are subjective, it's difficult to see how to specify a joint prior directly.

That's not to say that Bayesian estimation should be ruled out when there's not a compelling prior distribution, because it may in many instances be justified on other—including frequentist—grounds, such as regularization: That is, the constraints imposed by a suitably formulated, if only weakly justified, prior prevent wild estimates, such as those produced by highly unusual data. In these instances, however, the argument for Bayesian inference is not as obviously right.

Bayesians often correctly point out that the importance of the prior fades as the quantity of data grows. But the other side of this coin is that the prior can be important when data are sparse: Tautologically, when the prior matters, it matters what prior you choose. As I explained, if you have a strong justification for a prior, either subjective or based on experience, then all is well. If specification of a prior is ambiguous, however, and the posterior distribution changes substantially with the prior that's selected, it's not clear what has been gained from Bayesian inference. You could show how results vary by choice of prior, but that hardly provides a satisfactory conclusion. You might as well report a wide confidence interval based on the data alone.

I believe that the flexibility of current software for Bayesian estimation partly explains the increasing popularity of the Bayesian approach. That is, software such as Stan (Carpenter et al., 2017) allows one to tailor a statistical model to the data. This advantage of Bayesian software is somewhat ironic, in that until relatively recently the *inability* to obtain estimates except in the case of conjugate priors rendered Bayesian estimation impractical in most applications. Nor is this kind of flexibility intrinsic to Bayesian estimation—Stan, for example, can be used to obtain approximate maximum-likelihood estimates.

Most of my comments in this section concern the choice of prior distributions in Bayesian inference about regression coefficients. None of the points that I raise are original, and Bayesians have various responses to the reservations about priors that I express here. That the choice of priors in Bayesian inference is a matter of controversy, even among Bayesians, however, and that the subject can quickly get esoteric, suggest that these matters are far from settled.

Exercises

Please find data analysis exercises and data sets for this chapter on the website for the book.

Exercise 25.1. Return to the application of Bayes's theorem described at the end of Section 25.1.1 (on page 3). On the basis of the information supplied, calculate the probability that you have antibodies to the virus given a positive test, $\Pr(A|P)$. Compare this probability to the probability of a positive test given antibodies $\Pr(P|A)$. Are you surprised by the difference between these two probabilities? Can you explain the source of the difference in simple terms?

In developing your explanation, it might help to consider what $\Pr(A|P)$ would be in a population in which *no one* had antibodies—that is, for which $\Pr(A) = 0$.

Exercise 25.2. The preliminary example of Bayesian inference in Section 25.1.2 describes a situation in which two biased coins, one with the probability of a head $\Pr(H) = .3$ and the other with $\Pr(H) = .8$, are loose in a drawer and you don't know which is which. Suppose instead that these two biased coins are mixed in with eight fair coins for which $\Pr(H) = .5$. As in Section 25.1.2, you choose one coin from the drawer and flip it 10 times, observing 7 heads in the 10 flips. Let H_1 represent the hypothesis that you picked the coin with $\Pr(H) = .3$, H_2 the hypothesis that you picked the coin with $\Pr(H) = .8$, and H_3 the hypothesis that you picked one of the coins with $\Pr(H) = .5$. What is a reasonable set of prior probabilities for these hypotheses? Find the likelihood of the data under each hypothesis, and then compute the posterior probabilities for the three hypotheses. What do you conclude?

Exercise 25.3. Figure 25.2 (page 10) summarizes the example in Section 25.1.5, where a coin is flipped $n = 10$ times, producing $h = 7$ heads, and the resulting Bernoulli likelihood is combined with a beta prior to produce a beta posterior distribution. Figure 25.2 shows the posteriors for π, the probability of a head on an individual flip, produced by the flat Beta$(a = 1, b = 1)$ prior and the informative Beta$(a = 16, b = 16)$ prior.

(a) What happens to the posterior distribution of π for both of these priors as the sample size grows, successively taking on the values $n = 10, 100$, and 1000, while the observed proportion of heads remains at .7? For each of the two priors, graph the resulting posterior distributions, showing the posterior modes along with 95% central posterior intervals for π. Comment on the results.

(b) Now let $n = 10$, as in the text, but vary the observed number of heads $h = 0, 1, 2, \ldots, 10$; use the informative Beta$(a = 16, b = 16)$ prior. How does the posterior mode of π change with the number of heads?

Exercise 25.4. *The Jeffreys prior is invariant with respect to transformation of a parameter. The Jeffreys prior for estimating a probability (population proportion) π is $p_J(\pi) = 1/\left[\Pi\sqrt{\pi(1-\pi)}\right]$, as given in Section 25.1.6. Here $\Pi \approx 3.14159$ is the mathematical constant, capitalized to differentiate it from the probability π.

More generally, suppose that the parameter α is transformed to $\beta = f(\alpha)$, where $f(\cdot)$ is a smooth, continuous function. Then, for the prior density $p_\alpha(\alpha)$ to be invariant with respect to transformation, it must be the case that

$$p_\beta(\beta) = p_\alpha(\alpha)\left|\frac{d\alpha}{d\beta}\right|$$

where $p_\beta(\beta)$ is the prior density for β and $\left|\frac{d\alpha}{d\beta}\right|$ is the Jacobian of the transformation from α to β.[60] It turns out that the Fisher information,[61]

$$\mathcal{I}_\alpha(\alpha) = -E\left[\frac{d^2 \log_e L(\alpha)}{d\alpha^2}\right]$$

(where $L(\alpha)$ is the likelihood function) has the property that

$$\mathcal{I}_\beta(\beta) = \mathcal{I}_\alpha(\alpha)\left(\frac{d\alpha}{d\beta}\right)^2$$

and so the Jeffreys prior is proportional to the square-root of the Fisher information: $p_\alpha(\alpha) \propto \sqrt{\mathcal{I}_\alpha(\alpha)}$.

A Bernoulli random variable takes on the value 1 with probability π and the value 0 with probability $1-\pi$, and so the likelihood for the sum h (for "heads," invoking our example of flipping a coin repeatedly and counting the number of heads) of n independent Bernoulli random variables is $L(\pi) = \pi^h(1-\pi)^{n-h}$. Show that the corresponding Jeffreys prior is then proportional to $n/\sqrt{\pi(1-\pi)}$ and thus to $1/\sqrt{\pi(1-\pi)}$ (because n is a constant).[62]

Exercise 25.5. *Bayesian estimation of the mean of a normal distribution with known variance:* Although this is an unrealistic problem—if the mean μ of $Y \sim N(\mu, \sigma^2)$ is unknown, then it would be very unusual to know the variance σ^2—the solution is instructive.

If Y is normally distributed with mean μ and variance σ^2, then the density function for a sample of n independent observations drawn from the distribution is[63]

$$\begin{aligned} p_{\text{data}}(\text{data} = \{y_1, y_2, \ldots, y_n\}) &= \prod_{i=1}^{n} \frac{1}{(2\pi\sigma)^{\frac{1}{2}}} \exp\left[-\frac{(y_i-\mu)^2}{2\sigma^2}\right] \\ &= \frac{1}{(2\pi\sigma)^{\frac{n}{2}}} \exp\left[-\frac{\sum_{i=1}^{n}(y_i-\mu)^2}{2\sigma^2}\right] \end{aligned} \tag{25.33}$$

Thinking of Equation 25.33 as a function of μ given σ and the data produces the likelihood function $L(\mu|\sigma, \text{data})$. It turns out that the conjugate prior for μ is also a normal distribution, say $\mu \sim N(\mu_0, \sigma_0^2)$ and so the prior density is

$$p_\mu(\mu) = \frac{1}{(2\pi\sigma_0)^{\frac{1}{2}}} \exp\left[-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right]$$

[60] See on-line Appendix Section D.1.3 for an explanation of the the Jacobian of a transformation of a random variable. The formula in the appendix is for a vector random variable but simplifies to the result given here when the random variable, α, is a scalar.

[61] See on-line Appendix Section D.6.2.

[62] The result in the text includes the normalizing constant $1/\Pi$.

[63] See on-line Appendix Section D.3.1 for the formula of the density function of the normal distribution.

Because $p_\mu(\mu) = N(\mu_0, \sigma_0^2)$ is a conjugate prior, the posterior is also normal, and takes the form

$$p_{\mu|\text{data}}(\mu|\text{data}) = \frac{1}{(2\pi\sigma_1^2)^{\frac{1}{2}}} \exp\left[-\frac{(\mu-\mu_1)^2}{2\sigma_1^2}\right]$$

where the posterior mean μ_1 and variance σ_1^2 of μ are

$$\mu_1 = \frac{\dfrac{\bar{y}}{\sigma^2/n} + \dfrac{\mu_0}{\sigma_0^2}}{\dfrac{1}{\sigma^2/n} + \dfrac{1}{\sigma_0^2}} = \frac{\dfrac{n\bar{y}}{\sigma^2} + \dfrac{\mu_0}{\sigma_0^2}}{\dfrac{n}{\sigma^2} + \dfrac{1}{\sigma_0^2}}$$

$$\sigma_1^2 = \frac{1}{\dfrac{1}{\sigma^2/n} + \dfrac{1}{\sigma_0^2}} = \frac{1}{\dfrac{n}{\sigma^2} + \dfrac{1}{\sigma_0^2}}$$

Here, $\bar{y} = \frac{1}{n}\sum y_i$ is the familiar sample mean of y.[64]

(a) The posterior mean μ_1 is therefore a weighted average of the sample mean $\bar{y}$ and the prior mean μ_0 with respective weights $\dfrac{1}{\sigma^2/n} = n/\sigma^2$ and $1/\sigma_0^2$. Do you recognize the quantity σ^2/n in the denominator of the weight for the sample mean? What is it?

(b) The weights associated with the sample mean and the prior mean are called their *precision*. Thus, $\dfrac{1}{\sigma^2/n} = n/\sigma^2$ is the precision of $\bar{y}$ and $1/\sigma_0^2$ is the precision of μ_0. What is the relationship between precision and variance? Why is the term "precision" used? Why does it make intuitive sense to weight each term by its precision?

(c) The posterior variance of μ is the inverse of the sum of the precision of the sample mean and the precision of the prior mean. Is that a sensible result?

Exercise 25.6. *More on complete separation in logistic regression:* Anderson's (1935) iris data set has been a staple of the statistical literature since R. A. Fisher (1936) used it to introduce linear discriminant analysis (a method of classification). Logistic regression can also be used to classify objects into two or more classes based on characteristics of the objects.

Anderson's data consist of 150 specimens of irises collected in the Gaspé Peninsula of Québec, Canada, 50 each of the species setosa, versicolor, and

[64] Although setting up this result is straightforward—just multiply the formulas for $p_\mu(\mu)$ and $p_{\mu|\text{data}}(\mu|\text{data})$ and note that multiplying two exponentials is equivalent to the exponential of the sum of their exponents—putting the result in the form given here requires nontrivial algebra. See, for example, Box and Tiao (1973, Appendix A1.1).

virginica. Four measurements were made for each flower: sepal length, sepal width, petal length, and petal width, all in cm.

(a) Draw a scatterplot matrix for the four measured variables, marking the points by iris species. What do you conclude about the ability of the four variables to distinguish among the species?

(b) Perform a binary logistic regression of species setosa versus the other two species as the response on the four measured variables as predictors, fitting the model by maximum likelihood. Are you able to fit the model? Try again using Bayesian logistic regression, either via Firth's bias-reduced logistic regression or directly specifying a vague prior. What do you conclude?

(c) Dichtomizing iris species is artificial, and we can instead perform a polytomous logistic regression (see Section 14.2.1) with the three species as the response. Try to fit the polytomous logit model by maximum likelihood. Are you successful? Try again using Bayesian polytomous logistic regression with a vague prior. Alternatively, Kosmidis and Firth (2011) extend Firth's bias-reduced binary logistic regression to polytomous logistic regression; use their method to fit the polytomous logit model to the iris data.

Exercise 25.7. Return to the mixed-effects hurdle model fit to Davis et al.'s data on eating disorders and exercise, the results of which are summarized in Table 25.6 (page 66). I encountered an obstacle in fitting this model that led me to remove the random-effect intercepts from the model, thereby eliminating the parameters ψ_{Z_0} and $\rho_{Z_0 Z_1}$. As I indicated (see footnote 56 on page 66), a conceptually problematic alternative to simplifying the random effects might be to tighten the priors for some of the parameters. Experiment with this approach. Are you able to obtain useful estimates? If so, do the estimates of the *fixed* effects differ substantially from those reported in Table 25.6?

Summary

- Named after the Reverend Thomas Bayes, an 18th-century English mathematician, Bayes's theorem, which follows from elementary probability theory, states that for events A and B,

$$\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)}$$

  where $\Pr(B) = \Pr(B|A)\Pr(A) + \Pr(B|\overline{A})\Pr(\overline{A})$ is the unconditional probability of B (and $\overline{A}$ is the event not-A).

  Bayesian statistical inference is based on the following interpretation of Bayes's theorem:

  - Let A represent an uncertain proposition (a "hypothesis"), and let B represent observed data that are relevant to the truth of the proposition.
  - $\Pr(A)$ is the *prior probability* of A, our strength of belief in A prior to collecting data.
  - $\Pr(B|A)$ is the probability of obtaining the observed data assuming the truth of A—the *likelihood* of the data given A.
  - $\Pr(A|B)$, the *posterior probability* of A, represents our revised strength of belief in A in light of the data B.

  Bayesian inference is therefore a rational procedure for updating one's beliefs on the basis of evidence.

- Bayes's theorem can be extended to several hypotheses $H_1, H_2, \ldots, H_k$, with prior probabilities $\Pr(H_i)$, $i = 1, \ldots, k$, that sum to 1, and observed data D with likelihood $\Pr(D|H_i)$ under hypothesis H_i; the posterior probability of hypothesis H_i is

$$\Pr(H_i|D) = \frac{\Pr(D|H_i)\Pr(H_i)}{\sum_{j=1}^{k}\Pr(D|H_j)\Pr(H_j)}$$

  Similarly, Bayes's theorem is applicable to random variables such as a parameter α, with prior probability distribution or density $p(\alpha)$, and likelihood $L(\alpha) \equiv p(D|\alpha)$ for the data D. Then

$$p(\alpha|D) = \frac{L(\alpha)p(\alpha)}{\sum_{\text{all } \alpha'} L(\alpha')p(\alpha')}$$

  when the parameter α is discrete, or

$$p(\alpha|D) = \frac{L(\alpha)p(\alpha)}{\int_A L(\alpha')p(\alpha')\,d\alpha'}$$

  when α is continuous (and where A represents the set of all values of α)

- Bayesian inference is simplest with a *conjugate prior distribution*, which combines with the likelihood to produce a posterior distribution in the same family as the prior. For example, if h counts the number of heads in n independent flips of a coin with probability π of obtaining on a head on an individual flip, then the Bernoulli likelihood for the data is $L(\pi) = \pi^h(1-\pi)^{n-h}$. Combining this likelihood with the prior distribution $p(\pi) = \text{Beta}(a, b)$ produces a posterior distribution in the same family as the prior, $p(\pi|D) = \text{Beta}(h + a, n - h + b)$.

- There are several kinds of *uninformative prior distributions*. The *flat prior* assigns equal probability density to all values of a parameter; if the parameter is unbounded, then the flat prior is *improper*, in that it doesn't

integrate to 1, and a flat prior for a parameter is not in general flat for a transformation of the parameter. The *Jeffreys prior* for a parameter (introduced by Sir Harold Jeffreys), in contrast, is invariant with respect to transformation of the parameter. *Weakly informative priors* are often employed in practice, and are selected to place broad plausible constraints on the value of a parameter.

- Bayesian interval estimates, termed *credible intervals* (analogous to frequentist confidence intervals), are computed from the posterior distribution of a parameter. The $100a\%$ central posterior interval runs from the $(1-a)/2$ to the $(1+a)/2$ quantile of the posterior distribution. A Bayesian credible interval has a simple interpretation as a probability statement: For example, the probability is .95 that the parameter is in the 95% posterior interval.

- Bayesian inference extends to the simultaneous estimation of several parameters $\boldsymbol{\alpha} \equiv [\alpha_1, \alpha_2, \ldots, \alpha_k]'$. Given the joint prior distribution for the parameters $p(\boldsymbol{\alpha})$ along with the joint likelihood $L(\boldsymbol{\alpha})$ based on data D, the posterior distribution of $\boldsymbol{\alpha}$ is

$$p(\boldsymbol{\alpha}|D) = \frac{p(\boldsymbol{\alpha})L(\boldsymbol{\alpha})}{\int_{\mathbf{A}} p(\boldsymbol{\alpha}^*)L(\boldsymbol{\alpha}^*)d^k\boldsymbol{\alpha}^*}$$

  where $\mathbf{A}$ is the set of all values of the parameter vector $\boldsymbol{\alpha}$ (i.e., the multidimensional parameter space).

- *Markov-chain Monte Carlo* (*MCMC*) is a set of methods for drawing random samples from—and hence approximating—the posterior distribution $p(\boldsymbol{\alpha}|D)$ without having explicitly to evaluate the integral in the denominator, which is typically intractible analytically. MCMC methods have therefore rendered Bayesian inference practical for a broad range of statistical problems.

  There are three common (and related) MCMC methods: the *Metropolis-Hastings algorithm*, the *Gibbs sampler*, and *Hamiltonian Monte Carlo* (*HMC*). HMC is considered the best current method of MCMC for sampling from continuous distributions.

- In theory, Markov chains produced by MCMC sampling converge to the target distribution as the number of simulated draws goes to infinity, but in practice several sorts of problems can occur in chains of finite length. Convergence diagnostics help to determine whether MCMC samples adequately characterize a target distribution. In formulating these diagnostics, it helps to sample and compare two or more independent Markov chains and to discard the initial samples of each (e.g., the first half) as a "burn-in period."

  - A *trace plot* is a line graph of a sampled quantity—typically a parameter or a function of parameters—versus the simulation index. If the

MCMC samples have converged to the target distribution, then the center and spread of the trace plot shouldn't change on average with the index, and trace plots for independent Markov chains should be similar.

  - The *potential scale-reduction factor* $\widehat{R}$ measures the similarity of two or more Markov chains for a sampled quantity such as a parameter. If the chains have converged, then $\widehat{R}$ should be close to 1.
  - The *effective sample size* $m_{\text{eff}} \approx pm/\left(1 + 2\sum_{t=1}^{t'} r_t\right)$ measures the amount of information about a sampled quantity contained in the MCMC samples, where p is the number of independent chains employed, m is the number of samples retained from each chain, r_t is the estimated autocorrelation of the sampled values at lag t, and t' is selected so that r_t is negligible for $t > t'$.

- The normal linear regression model provides a probability model for the data from which the likelihood can be calculated:

$$
\begin{aligned}
&Y_i|x_{i1}, x_{i2}, \ldots, x_{ik} \sim N(\alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}, \sigma_\varepsilon^2) \\
&Y_i, Y_{i'} \text{ are independent for } i \neq i'
\end{aligned}
$$

  To obtain Bayesian estimates of the parameters of the regression model, $\alpha, \beta_1, \ldots, \beta_k$, and σ_ε, we require prior distributions for the parameters. One approach is to use vaguely informative normal priors, remembering that almost all—99.7%—of the density of a normal distribution is within 3 standard deviations of its mean, and that 95% and 68% of the density lie respectively within 2 and 1 standard deviations of the mean. In this approach, the priors for the various parameters are specified separately and are treated as independent.

  The standard deviation of the errors, σ_ε, can't be negative, and so we can use a normal prior for its log. Because the intercept α is often far from the observed data, in specifying a prior distribution for α, it often helps first to center the xs at their means or at other meaningful values.

  Once the regression model and the priors are specified, the joint posterior distribution of the parameters is approximated by Markov-chain Monte Carlo.

- Bayesian estimation of generalized linear models is very similar to Bayesian estimation of linear models: The GLM provides a probability model for the data (see Section 15.1):

$$
\begin{aligned}
\eta_i &= \alpha + \beta_1 x_{i1} + \cdots \beta_k x_{ik} \\
\mu_i &= g^{-1}(\eta_i) \\
Y_i|x_{i1}, \ldots, x_{ik} &\sim p(\mu_i, \phi)
\end{aligned}
$$

where the conditional distribution $p(\mu_i, \phi)$ of the response Y_i is a member of an exponential family, with expectation μ_i and dispersion parameter ϕ (which recall is set to 1 in the binomial and Poisson families).

Once prior distributions for the parameters of the model are specified, their posterior distribution can be approximated by MCMC.

- Firth (1993) showed that substantial bias reduction in estimating the logistic regression model under difficult circumstances can be achieved by employing the Jeffreys prior. An especially problematic data pattern for logistic regression to which Firth's estimator is applicable is complete separation, where a linear function of the regressors in the model partitions the data into disjoint regions of 0s and 1s. In this case, maximum-likelihood estimation of the logistic regression coefficients produces one or more infinite estimates, while Firth's method yields finite estimates.

- A common application of Bayesian estimation is to mixed-effects models of various kinds. The linear, generalized-linear, and nonlinear mixed models discussed in Chapters 23 and 24 all provide probability models for data, and these models can be extended in various ways, as illustrated by the hurdle model fit in Section 25.4. With suitable priors for the regression coefficients and variance-covariance components, Bayesian estimates for mixed-effects models can by obtained by MCMC methods. The prior distribution for the variance-covariance components of a mixed model must be suitably parametrized to produce a positive-definite covariance matrix for the random effects.

- A convenient by-product of estimating a regression model by MCMC is that we can calculate quantities derived from the parameters for each sample of the parameter values, providing estimated posterior distributions of the derived quantities. One application of this idea is to the construction of effect plots.

Recommended Reading

Books and articles on Bayesian statistics are typically written by committed Bayesians, and so I suggest that you maintain an open mind with respect to the case that Bayesians often press against classical approaches to statistical inference. There is a very large literature on modern Bayesian methods; the few sources given here are therefore highly selected.

- The treatise on Bayesian data analysis by Gelman et al. (2013) is a wide-ranging and through treatment of the subject by some of the leading figures in contemporary Bayesian statistics. The book covers in greater detail all of the topics included in this chapter (basics of Bayesian inference, MCMC methods, linear models, generalized linear models, and mixed-effects models) and more, and almost all of the book should be accessible to readers of the starred sections of the current text.

- McElreath (2020) presents a much gentler introduction to Bayesian methods, but, like Gelman et al. (2013), takes up the topics in this chapter, usually more thoroughly than I do, along with others, such as causal inference. Depending on your taste, you may find the author's style either engaging or irritating, but his exposition is almost always very clear and carefully thought-out. The computing in the text tends towards the idiosyncratic, with a compelling justification: Rather than using standard black-box Bayesian software such as Stan (see below), McElreath guides the reader through the nuts and bolts of Bayesian computations, rendering the process transparent.

- Gelman and Hill (2007), and its partial successor Gelman et al. (2021), are intended for a second course in statistics. Both books treat linear and generalized linear models from a Bayesian perspective, while Gelman and Hill (2007) (as the title of their book implies) additionally cover hierarchical (i.e., mixed-effects) regression models. A projected second updated volume supplementing Gelman et al. (2021) will also focus on hierarchical models. The older text by Gelman and Hill (2007) primarily uses R and BUGS for computing, while Gelman et al. (2021) use R and Stan.

- The documentation for Stan (Carpenter et al., 2017), available at `https://mc-stan.org/users/documentation/` and including a user's guide, language reference manual, and more, provides valuable information not only on using the Stan software but also on current practices in Bayesian data analysis more generally.

- Thomas and Tu (2021), which appeared after this chapter was written, is a generally accessible introduction to Hamilton Monte Carlo. In addition to explaining how the Metropolis-Hastings and HMC algorithms work, the authors develop applications to linear regression, logistic regression, and Poisson regression with random intercepts. The article is associated with the **hmclearn** R package for exploring HMC. The implementation of various regression models in this package, and in the article, is weakened by cavalier treatment of prior distributions, specifying independent normal priors with a common prior standard deviation for regression coefficients.

- For what Bayesian statistics was like in the era prior to the use of MCMC, see Box and Tiao (1973), who present a lucid treatment of the basics of Bayesian inference along with a range of applications, including to regression models. Although they are Bayesians, the authors don't hesitate to use the term "random effect models."

References for Chapter 25

E. Anderson. The irises of the Gaspé Peninsula. *Bulletin of the American Iris Society*, 59:2–5, 1935.

J. M. Bernardo. Reference posterior distributions for Bayesian inference. *Journal of the Royal Statistical Society, Series B*, 41:113–147, 1979.

G. E. P. Box and G. C. Tiao. *Bayesian Inference in Statistical Analysis*. Addison-Wesley, Reading MA, 1973.

S. P. Brooks and A. Gelman. General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7: 343–455, 1997.

B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1):1–32, 2017.

G. Casella. An introduction to empirical Bayes data analysis. *The American Statistician*, 19:83–87, 1985.

S. Chib and E. Greenberg. Understanding the Metropolis-Hastings algorithm. *The American Statistician*, 49:327–335, 1995.

M. Cowles and C. Davis. The subject matter of psychology: volunteers. *British Journal of Social Psychology*, 26:97–102, 1987.

C. Davis, E. Blackmore, D. K. Katzman, and J. Fox. Female adolescents with anorexia nervosa and their parents: A case-control study of exercise attitudes and behaviours. *Psychological Medicine*, 35:377–386, 2005.

B. Efron and C. Morris. Stein's paradox in statistics. *Scientific American*, 236 (5):119–127, 1977.

D. Firth. Bias reduction of maximum likelihood estimates. *Biometrika*, 80: 27–38, 1993.

R. A. Fisher. Frequency distribution of the values of the correlation coefficient in samples of an indefinitely large population. *Biometrika*, 10:507–521, 1915.

R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188, 1936.

A. E. Gelfand and A. F. M. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85:398–409, 1990.

A. Gelman and J. Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, Cambridge UK, 2007.

A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC Press, Boca Raton FL, third edition, 2013.

A. Gelman, J. Hill, and A. Vehtari. *Regression and Other Stories*. Cambridge University Press, Cambridge UK, 2021.

S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–742, 1984.

W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109, 1970.

H. Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 186:453–461, 1946.

I. Kosmidis and D. Firth. Bias reduction in exponential family nonlinear models. *Biometrika*, 96:793–804, 2009.

I. Kosmidis and D. Firth. Multinomial logit bias reduction via the Poisson log-linear model. *Biometrika*, 98:755–759, 2011.

I. Kosmidis and D. Firth. Jeffreys-prior penalty, finiteness and shrinkage in binomial-response generalized linear models. *Biometrika*, 108:71–82, 2021.

D. Lunn, D. Spiegelhalter, A. Thomas, and N. Best. The BUGS project: Evolution, critique and future directions (with discussion). *Statistics in Medicine*, 28:3049–3082, 2009.

R. McElreath. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Chapman & Hall/CRC Press, Boca Raton FL, second edition, 2020.

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087–1092, 1953.

R. M. Neal. *Bayesian Learning for Neural Networks*. Springer, New York, 1996.

R. M. Neal. MCMC using Hamiltonian dynamics. In S. Brooks, A. Gelman, G. L. Jones, and X.-L. Meng, editors, *Handbook of Markov Chain Monte Carlo*, chapter 5, pages 113–162. Chapman & Hall/CRC Press, Boca Raton FL, 2011.

B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman & Hall/CRC, Boca Raton FL, 1986.

L. Susskind and G. Hrabovsky. *The Theoretical Minimum: What You Need to Know to Start Doing Physics*. Basic Books, New York, 2013.

S. Thomas and M. Tu. Learning Hamiltonian Monte Carlo in R. *The American Statistician*, 75:403–413, 2021.