Running head: BAYESIAN STATISTICAL METHODS

Bayesian Statistical Methods

David Kaplan
Sarah Depaoli
Department of Educational Psychology
University of Wisconsin – Madison

To appear in Little, T. D. (ed.) *Oxford Handbook of Quantitative Methods*

Abstract

This chapter provides a general overview of Bayesian statistical methods. Topics include the notion of probability from a Bayesian perspective, Bayesian inference and hypothesis testing, and Bayesian computation. Three examples are provided to demonstrate the utility of Bayesian methods: simple linear regression, multilevel regression, and confirmatory factor analysis. Throughout the chapter, references are made to the epistemological differences between Bayesian theory and classical (frequentist) theory.

Keywords: Bayesian statistical methods, Bayesian inference and hypothesis testing, Bayesian computation

Bayesian Statistical Methods

Bayesian statistics has long been overlooked in the quantitative methods training of social scientists. Typically, the only introduction that a student might have to Bayesian ideas is a brief overview of Bayes' theorem while studying probability in an introductory statistics class. There are two reasons for this. First, until recently, it was not feasible to conduct statistical modeling from a Bayesian perspective owing to its complexity and lack of available software. Second, Bayesian statistics challenges many of the assumptions underlying frequentist (classical) statistics, and is therefore, controversial. We will use the term *frequentist* to describe the paradigm of statistics commonly used today, and which represents the counterpart to the Bayesian paradigm of statistics. Historically, however, Bayesian statistics predates frequentist statistics by about 150 years.

Recently, however, there has been extraordinary growth in the development and application of Bayesian statistical methods, due mostly to developments of powerful statistical software tools that render the specification and estimation of complex models feasible from a Bayesian perspective. As a result, there have been scores of books written over the last 10 years, and at a variety of technical levels, that lead students and researchers through Bayesian theory and computation. For a technical treatment of Bayesian statistics see e.g. Gelman, Carlin, Stern, and Rubin (2003). For a less technical treatment see e.g Hoff (2009).

The scope of this chapter is, by necessity, limited because the field of Bayesian inference is remarkably wide ranging and space limitations preclude a full development of Bayesian theory. Thus, the goal of the chapter will be to lay out the fundamental issues that separate Bayesian statistics from its frequentist counterpart and to provide a taste of its applications through specific examples.

The organization of this chapter will cover (1) Bayesian probability, (2) Bayesian inference and hypothesis testing, (3) Bayesian computation, and (4) simple empirical examples of Bayesian linear regression, Bayesian multilevel modeling, and Bayesian confirmatory factor analysis. To support the pedagogical features of this chapter, the software code for each example is provided.

Bayesian Probability

Most students in the social and behavioral sciences were introduced to the axioms of probability by studying the properties of the coin toss or the dice roll. These studies address questions such as (1) What is the probability that the flip of a fair coin will return heads?; (2) What is the probability that the roll of two fair die will return a value of seven? To answer these questions requires enumerating the possible outcomes and then counting the number of times the event could occur. The probabilities of interest are obtained by dividing the number of times the event occurred by the number of possible outcomes. But what of more complex situations – such as the famous "Monty Hall" problem? In this problem, named after the host of a popular old game show, a contestant is shown three doors, one of which has a desirable prize, while the other two have quite undesirable prizes. The contestant picks a door, but before Monty opens the door, he shows the contestant another door with an undesirable prize and asks the contestant whether she wants to stay with the chosen door or switch. To address this situation requires an understanding of the Kolmogorov axioms of probability (Kolmogorov, 1956) and the Renyi axioms of conditional probability (Renyi, 1970). These sets of axioms, though appearing longer after Bayes' work, provide the theoretical foundation for Bayes' theorem.

The Kolmogorov Axioms of Probability

Before motivating Bayes' theorem, it is useful to remind ourselves of the axioms of probability that have formed the basis of frequentist statistics. These axioms of probability can be attributed to the work of Kolmogorov (1956). This particular set of axioms relate the notion of probability to the frequency of events over a large number of trials. These axioms form the basis of the frequentist paradigm of statistics.

Consider two events denoted A and B. To keep the example simple, consider these both to be the flip of a fair coin. Then the following are the axioms of probability - namely

1. $p(A) \geq 0$
2. The probability of the sample space is 1.0
3. Countable additivity: If A and B are mutually exclusive, then

$p(A, B) = p(A) + p(B)$. Or, more generally,

$$p\left\{\bigcup_{j=1}^{\infty} A_j\right\} = \sum_{j=1}^{\infty} p(A_j), \tag{1}$$

which states that the probability of the union of mutually exclusive events is simply the sum of their individual probabilities.

A number of other axioms of probability can be derived from these three basic axioms. Nevertheless, these three can be used to deal with the relatively easy case of the coin flipping example mentioned above. For example, if we toss a fair coin an infinite number of times, we expect it to land heads 50% of the time. Interestingly, this expectation is not based on having actually tossed the coin an infinite number times. Rather, this expectation is a prior belief. Arguably, this is one example of how Bayesian thinking is automatically embedded in frequentist logic. This probability, and others like it, satisfy the first axiom that probabilities are greater than or equal to zero. Secondly, over an infinite number of coin flips, the sum of all possible outcomes (in this case, heads and tails) is equal to one. Indeed, the number of possible outcomes represents the *sample space* and the sum of probabilities over the sample space is one. Finally, assuming that one outcome precludes the occurrence of another outcome (e.g., rolling a 1 precludes the occurrence of rolling a 2) then the probability of the joint event $p(A, B)$ is the product of the separate probabilties - that is $p(A, B) = p(A) + p(B)$. We may wish to add to these axioms the notion of *independent events*. If two events are independent, then the occurrence of one event does not influence the probability of another event. For example, with two coins A and B, the probability of A resulting in "heads", does not influence the result of a flip of B. Formally, we define independence as $p(A, B) = p(A)p(B)$.

The Renyi Axioms of Probability

In the previous paragraph, we discussed quite simple cases – in particular the case of independent events. Consider the case of non-independent events. In this situation, the Kolmogorov axioms do not take into account how probabilities might be affected by conditioning on the dependency of events. An extension of the Kolmogorov system that accounts for conditioning was put forth by Renyi (1970). As a motivating example, consider the case of observing the presence or absence of coronary heart disease (C) and the behavior of smoking or not smoking (S). We

may be able to argue on the basis of prior experience and medical research that C is not independent of S – that is, the joint probability $p(C, S) \neq p(C)p(S)$. To handle this problem, we define the *conditional probability* of C "given" S (i.e. $p(C|S)$) as

$$p(C|S) = \frac{p(C, S)}{p(S)}. \tag{2}$$

The denominator on the right hand side of Equation 2 shows that the sample space associated with $p(C, S)$ is reduced by knowing S. Notice that if C and S were independent, then

$$\begin{aligned} p(C|S) &= \frac{p(C, S)}{p(S)}, \\ &= \frac{p(C)p(S)}{p(S)}, \\ &= p(C) \end{aligned} \tag{3}$$

which states that knowing S tells us nothing about C.

Following Press (2003), Renyi's axioms can be defined, with respect to our coronary heart disease example, as follows:

1. For any events, A, B, we have $P(A|B) \geq 0$ *and* $p(B|B) = 1$.
2. For disjoint events A_j and some event B

$$p\left\{\bigcup_{j=1}^{\infty} A_j|B\right\} = \sum_{j=1}^{\infty} p(A_j|B)$$

3. For every collection of events (A, B, C), with B a subset of C (i.e. $B \subseteq C$), and $0 < p(B|C)$, we have

$$p(A|B) = \frac{p(A \cap B|C)}{p(B|C)}.$$

Renyi's third axiom allows one to obtain the conditional probability of A given B, while conditioning on yet a third variable C.

An important feature of Renyi's axioms is that it covers the Kolmogorov axioms as a special case. Moreover, it is general enough to encompass both frequentist interpretations of probability as well as personal belief interpretations of probability (Ramsey, 1926; Savage, 1954; de Finetti, 1974). The personal belief interpretation of probability is central to the subjectivist view of probability

embedded in Bayesian statistics. See Press (2003) for a more detailed discussion.

Bayes' Theorem

An interesting feature of Equation 2 underpins Bayes' theorem. Specifically, joint probabilities are symmetric – namely, $p(C, S) = p(S, C)$. Therefore, we can also express the conditional probability of smoking, S, given observing coronary heart disease, C as

$$p(S|C) = \frac{p(S, C)}{p(C)}. \tag{4}$$

Because of the symmetry of the joint probabilities, we obtain

$$p(C|S)p(S) = p(S|C)p(C). \tag{5}$$

Therefore,

$$p(C|S) = \frac{p(S|C)p(C)}{p(S)}. \tag{6}$$

Equation 6 is Bayes' theorem. In words, Bayes' theorem states that the conditional probability of an individual having coronary heart disease given that he smokes is equal to the probability that he smokes given that he has coronary heart disease times the probability of having coronary heart disease. The denominator of Equation 6, $p(S)$, is the marginal probability of smoking. This can be considered the probability of smoking across individuals with and without coronary heart disease, which we write as $p(S) = p(S|C) + p(S|\neg C)$.[1] Because this marginal probability is obtained over all possible outcomes of coronary heart disease it does not carry information relevant to the conditional probability. In fact, $p(S)$ can be considered a *normalizing factor*, that ensures that the probability sums to one. Thus, it is not uncommon to see Bayes' theorem written as

$$p(C|S) \propto p(S|C)p(C). \tag{7}$$

Equation 7 states that the probability of observing coronary heart disease given smoking is proportional to the probability of smoking given coronary heart disease times the marginal probability of coronary heart disease.

Let's return to the Monty Hall problem in order to demonstrate the complexities of conditional probability and how a Bayesian perspective can be

helpful. At the start of the game, it is assumed that there is one desirable prize and that the probability that the desirable prize is behind any of the three doors is 1/3. Once a door is picked, Monty Hall shows the contestant a door with an undesirable prize and asks the contestant if she would like to switch from the door she originally chose. It is important to note that Monty will not show the contestant the door with the desirable prize. Also, we assume that because the remaining doors have undesirable prizes, which door Monty opens is basically random. Given that there are two doors remaining in this three-door problem, the probability is 1/2. Thus, Monty's knowledge of where the prize is located plays a crucial role in this problem. With the following information in hand, we can obtain the necessary probabilities to apply Bayes' theorem. Assume the contestant picks door A. Then, the necessary conditional probabilities are

1. p(Monty opens door B|prize is behind A) $= \frac{1}{2}$
2. p(Monty opens door B|prize is behind B) $= 0$
3. p(Monty opens door B|prize is behind C) $= 1$.

The final probability is due to the fact that there is only one door for Monty to choose given that the contestant chose door A and the prize is behind door B.

Let M represent Monty opening door B. Then, the joint probabilities can be obtained follows.

$$\begin{aligned} p(M, A) &= p(M|A)p(A) = \frac{1}{2} \times \frac{1}{3} = \frac{1}{6}, \\ p(M, B) &= p(M|B)p(B) = 0 \times \frac{1}{3} = 0, \\ p(M, C) &= p(M|C)p(C) = 1 \times \frac{1}{3} = \frac{1}{3}. \end{aligned}$$

Before applying Bayes' theorem, note that we have to obtain the marginal distribution of Monty opening door B. This is

$$\begin{aligned} p(M) &= p(M, A) + p(M, B) + p(M, C) \\ &= \frac{1}{6} + 0 + \frac{1}{3} = \frac{1}{2} \end{aligned}$$

Finally, we can now apply Bayes' theorem to obtain the probabilities of the prize lying behind door A or door C.

$$p(A|M) = \frac{p(M|A)p(A)}{p(M)} = \frac{\frac{1}{2} \times \frac{1}{3}}{\frac{1}{2}} = \frac{1}{3}$$
$$p(C|M) = \frac{p(M|C)p(C)}{p(M)} = 1 \times \frac{\frac{1}{3}}{\frac{1}{2}} = \frac{2}{3}$$

Thus, from Bayes' theorem, the best strategy on the part of the contestant is to switch doors.

Bayesian Statistical Inference

The material presented thus far has concerned Bayesian probability. The goal of this chapter is to present the role of Bayes' theorem as it pertains to statistical inference. Setting the foundations of Bayesian statistical inference provides the framework for application to a variety of statistical models commonly employed in social and behavioral science research.

To begin, denote by Y a random variable that takes on a realized value y. For example, a person's socio-economic status could be considered a random variable taking on a very large set of possible values. Once the person identifies his/her socioeconomic status, the random variable Y is now realized as y. In a sense, Y is unobserved – it is the probability model that we wish to understand from the actual data values y.

Next, denote by θ a parameter that we believe characterizes the probability model of interest. The parameter θ can be a scalar (i.e. a single parameter), such as the mean or the variance of a distribution, or it can be vector-valued (i.e. a collection of parameters), such as the parameters of a factor analysis model. To avoid too much notational complexity, for now we will use θ to represent either scalar or vector valued parameters where the difference will be revealed by the context. Of importance to this chapter, θ could represent the parameters of an underlying hypothesized model – such as a regression model or structural equation model.

We are concerned with determining the probability of observing y given the unknown parameters θ, which we write as $p(y|\theta)$. Equivalently, we are concerned with obtaining estimates of the population parameters given the data expressed as the "likelihood" and formally denoted as $L(\theta|y)$. Often we work with the

log-likelihood written as $l(\theta|y)$.

The key difference between Bayesian statistical inference and frequentist statistical inference concerns the nature of the unknown parameters θ. In the frequentist tradition, the assumption is that θ is unknown but fixed. In Bayesian statistical inference, θ is considered random, possessing a probability distribution that reflects our uncertainty about the true value of θ. Because both the observed data y and the parameters θ are assumed random, we can model the joint probability of the parameters and the data as a function of the conditional density of the data given the parameters, and the prior distribution of the parameters. More formally,

$$p(\theta, y) = p(y|\theta)p(\theta). \tag{8}$$

Following Bayes' theorem described earlier, we obtain the following,

$$p(\theta|y) = \frac{p(\theta, y)}{p(y)} = \frac{p(y|\theta)p(\theta)}{p(y)}, \tag{9}$$

where $p(\theta|y)$ is referred to as the *posterior distribution* of the parameters θ given the observed data y. Thus, from Equation 9, the posterior distribution of θ given y is equal to the data distribution $p(y|\theta)$ times the prior distribution of the parameters $p(\theta)$ normalized by $p(y)$ so that the posterior distribution sums (or integrates) to one. For discrete variables

$$p(y) = \sum_{\theta} p(y|\theta)p(\theta), \tag{10}$$

and for continuous variables

$$p(y) = \int_{\theta} p(y|\theta)p(\theta)d\theta. \tag{11}$$

Note that the denominator in Equation 9 does not involve model parameters, so we can omit the term and obtain the *unnormalized posterior density*

$$p(\theta|y) \propto p(y|\theta)p(\theta). \tag{12}$$

Consider the data density $p(y|\theta)$ on the right hand side of Equation 12. When expressed in terms of the unknown parameters θ for fixed values of y, this term is the *likelihood* $L(\theta|y)$, which we defined earlier. Thus, Equation 12 can be re-written as

$$p(\theta|y) \propto L(\theta|y)p(\theta). \tag{13}$$

Equation 12 (or Equation 13) represents the core of Bayesian statistical inference and is what separates Bayesian statistics from frequentist statistics. Specifically, Equation 13 states that our uncertainty regarding the parameters of our model, as expressed by the prior density $p(\theta)$, is *weighted* by the actual data $p(y|\theta)$ (or equivalently, $L(\theta|y)$), yielding an updated estimate of our uncertainty, as expressed in the posterior density $p(\theta|y)$.

The Nature of the Likelihood

Equation 13 states that Bayes' theorem can be written as the product of the likelihood of the unknown parameters for fixed values of the data and the prior distribution of the model parameters. In this section, we consider two common statistical distributions and their likelihoods before moving on to discuss prior distributions. Specifically, we will consider the binomial distribution and normal distribution. Before beginning, however, it is necessary to discuss the assumption of *exchangeability*.

Exchangeability arises from de Finetti's Theorem (de Finetti, 1974) and implies that the subscripts of a vector of data, e.g. $y_1, y_2, \ldots y_n$ do not carry information that is relevant to describing the probability distribution of the data. In other words, the joint distribution of the data, $f(y_1, y_2, \ldots y_n)$ is invariant to permutations of the subscripts.[2]

As a simple example of exchangeability consider a vector of responses to a ten item test where a correct response is coded "1" and an incorrect response is coded "0". Exchangeability implies that only the total number of correct responses matter, not the location of those correct responses in the vector. Exchangeability is a subtle assumption insofar as it means that we believe that there is a parameter θ that generates the observed data via a statistical model and that we can describe that parameter without reference to the particular data at hand (Jackman, 2009). As an example, consider the observed responses on an IQ test. The fundamental idea behind statistical inference generally is that the observed responses on an IQ test are assumed to be generated from a population distribution (e.g. the normal distribution) characterized by a parameter θ (e.g. the population mean). As Jackman (2009) points out, the fact that we can describe θ without reference to a

particular set of IQ data is, in fact, what is implied by the idea of a prior distribution. In fact, as Jackman notes, "the existence of a prior distribution over a parameter is a *result* of de Finetti's Representation Theorem, rather than an assumption" (pg. 40, italics Jackman's).

It is important to note that exchangeability is weaker than the statistical assumption of independence. In the case of two events - say A and B, independence implies that $p(A|B) = p(A)$. If these two events are independent then they are exchangeable - however, exchangeability does not imply independence.

Example 1: The Binomial Probability Model

First, consider the number of correct answers on a test of length n. Each item on the test represents a "Bernoulli trial", with y outcomes 0 = wrong and 1 = right. The natural probability model for data arising from n Bernoulli sequences is the binomial sampling model. Under the assumption of exchangeability - meaning the indexes 1 ... n provide no relevant information, we can summarize the total number of successes by n. Letting θ be the proportion of correct responses in the population, the binomial sampling model can be written as

$$p(y|\theta) = \text{Bin}(y|n, \theta) = \binom{n}{y}\theta^y(1-\theta)^{(n-y)}. \tag{14}$$

where $\binom{n}{y}$ is read as "n choose y" and refers to the number of successes y in a sequence of "right/wrong" Bernoulli trials that can be obtained from an n item test. The symbol Bin is shorthand for the binomial density function.

Example 2: The Normal Sampling Model

The likelihood function for the parameters of the simple normal distribution can be written as

$$f(y|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right). \tag{15}$$

Under the assumption of independent observations, we can write Equation 15 as

$$\begin{aligned} f(y_1, y_2, \ldots, y_n|\mu, \sigma^2) &= \prod_i^n f(y_i|\mu, \sigma^2), \\ &= \left(\frac{1}{\sqrt{2\pi}\sigma^2}\right)^{n/2}\exp\left(-\frac{\sum_i(y_i-\mu)^2}{2\sigma^2}\right), \\ &= L(\theta|y), \end{aligned} \tag{16}$$

where $\theta = (\mu, \sigma)$.

The Nature of the Prior Distribution

It is useful to remind ourselves of the reason why we specify a prior distribution on the parameters. The key philosophical reason concerns our view that progress in science generally comes about by learning from previous research findings and incorporating information from these findings into our present studies. Upon reflection, it seems obvious that no study is conducted in the complete absence of previous research. From experimental designs to path diagrams, the information gleaned from previous research is almost always incorporated in our choice of designs, variables to be measured, or conceptual diagrams to be drawn. Researchers who postulate a directional hypothesis for an effect are almost certainly using prior information about the direction that an estimate must take. Bayesian statistical inference, therefore, simply requires that our prior beliefs be made explicit, but then moderates our prior beliefs by the actual data in hand. Moderation of our prior beliefs by the data in hand is the key meaning behind Equation 12.

But how do we choose a prior? The general approach to considering the choice of a prior is based on how much information we believe we have prior to the data collection and how accurate we believe that information to be (Lynch, 2007). This issue has also been discussed by Leamer (1983), who orders priors on the basis of degree of confidence. Leamer's hierarchy of confidence is as follow: truths (e.g. axioms) > facts (data) > opinions (e.g. expert judgement) > conventions (e.g. pre-set alpha levels).

An interesting feature of this hierarchy, as noted by Leamer, concerns the inherent lack of "objectivity" in such choices as pre-set alpha levels, or any of a number of assumptions made in linear regression based models. In describing the "whimisical" nature of statistical inference, Leamer goes on to argue that the problem should be to articulate exactly where a given investigation is located on this hierarchy. The strength of Bayesian inference lies precisely in its ability to incorporate existing knowledge into statistical specifications.

Objective Priors

A very important discussion regarding general types of prior distributions can be found in Press (2003). In his book, Press distinguishes between *objective* versus *subjective* prior distributions. The notion of an objective prior relates to having very

little information regarding the process that generated the data prior to the data being collected.

Public policy prior. One type of objective prior discussed by Press (2003) is the *public policy prior*. The public policy prior concerns reporting the results of an experiment or study to the public that contains a minimal amount of the researcher's subjective judgements as possible.

To take an example from education, suppose one is interested in a policy to reduce class size because it is viewed as being related to academic achievement – lower class sizes being associated with higher academic achievement, particularly for low income students. Assume for this example that based on previous research, the investigator has a sense of how much student achievement will increase (based on a standardized test) for a given reduction in class size. From the standpoint of educational policy, the results reported to stakeholders should not depend on the prior beliefs of an individual researcher. In this case, the researcher may decide to use a *vague* prior reflecting an unwillingness to report an effect of reduced class size that is based on a specific prior belief.[3]

Non-informative prior. In some cases we may not be in possession of enough prior information to aid in drawing posterior inferences. From a Bayesian perspective, this lack of information is still important to consider and incorporate into our statistical specifications. In other words, it is equally important to quantify our ignorance as it is to quantify our cumulative understanding of a problem at hand.

The standard approach to quantifying our ignorance is to incorporate a non-informative prior into our specification. Non-informative priors are also referred to as *vague* or *diffuse* priors. Perhaps the most sensible non-informative prior distribution to use in this case is the uniform distribution over some sensible range of values. Care must be taken in the choice of the range of values over the uniform distribution. Specifically, a Uniform[$-\infty, \infty$] is an *improper* prior distribution insofar as it does not integrate to 1.0 as required of probability distributions.

Jeffreys' Prior. A problem with the uniform prior distribution is that it is not invariant to simple transformations. In fact a transformation of a uniform prior can result in a prior that is not uniform and will end up favoring some values more than others. As pointed out by Gill (2002), the invariance problem associated with

uniform priors, and indeed the use of uniform priors specifically, had been greeted with extreme skepticsm by many early statisticians and used as the foundation of major critiques of Bayesian statistics generally.

Despite the many criticisms against the uniform prior, its use dominates applied Bayesian work. Justification for the use of the uniform prior has been given in Bauwens, Lubrano, and Richard (2003) who point out that (1) the effect of the uniform prior tends to diminish with increasing sample size, (2) the uniform prior is useful when models contain nuisance parameters, such as the variance of the normal distribution when the mean is of interest, as they will be integrated out anyway, and (3) the uniform distribution is the limit of certain conjugate distributions. In Bayesian statistics, conjugate distributions are those that when multiplied by the likelihood via Bayes' theorem yield posterior distributions in the same distributional family as the prior distribution.

In specifically addressing the invariance problem associated with the uniform distribution, Jeffreys (1961) proposed a general approach that yields a prior that is invariant under transformations. The central idea is that the subjective beliefs contained in the specification of the prior distribution of a parameter θ should not be lost when there is a one-to-one transformation from θ to another parameter, say ϕ. More specifically, using transformation-of-variables calculus, the prior distribution $p(\phi)$ will be equivalent to $p(\theta)$ when obtained as

$$p(\phi) = p(\theta) \left| \frac{d\theta}{d\phi} \right| . \tag{17}$$

On the basis of the relationship in Equation 17, Jeffreys (1961) developed a non-informative prior distribution that is invariant under transformations, written as

$$p(\theta) \propto [I(\theta)]^{1/2} , \tag{18}$$

where $I(\theta)$ is the *Fisher information matrix* for θ.

Jeffreys' prior is obtained as follows. Following Gelman et al. (2003), let $f(x|\theta)$ be the likelihood for θ and write its associated Fisher information matrix as

$$I(\theta) = \left[-E_{x|\theta} \left(\frac{\partial^2 (\log f(x|\theta))}{\partial \theta^2} \right) \right]^{\frac{1}{2}} . \tag{19}$$

Next, we write the Fisher information matrix for ϕ as

$$I(\phi) = \left[-E_{x|\phi} \left(\frac{\partial^2 (\log f(x|\phi))}{\partial \phi^2} \right) \right]^{\frac{1}{2}}. \tag{20}$$

From the change of variables expression in Equation 17, we can rewrite Equation 20 as

$$\begin{aligned} I(\phi) &= \left[-E_{x|\theta} \left(\frac{\partial^2 (\log f(x|\theta))}{\partial \theta^2} \times \left| \frac{d\theta}{d\phi} \right| \right) \right]^{\frac{1}{2}}, \\ &= I(\theta) \left| \frac{d\theta}{d\phi} \right|^2. \end{aligned} \tag{21}$$

Therefore,

$$I^{1/2}(\phi) = I^{1/2}(\theta) \times \left| \frac{d\theta}{d\phi} \right|, \tag{22}$$

from which we obtain the relationship to Equation 18. The Jeffreys prior can also be extended to a vector of model parameters and thus is applicable to regression models and their extensions (see Gill, 2002).

Press (2003) then goes on to weigh the advantages and disadvantages of objective priors. Following Press (2003), in terms of advantages,

1. Objective priors can be used as benchmarks against which choices of other priors can be compared,

2. Objective priors reflect the view that little information is available about the process that generated the data,

3. There are cases in which the results of a Bayesian analysis with an objective prior provides equivalent results to those based on a frequentist analysis – though there are philosophical differences in interpretation that we allude to later in the chapter,

4. Objective priors are sensible public policy priors.

In terms of disadvantages, Press (2003) notes that

1. Objective priors can lead to improper results when the domain of the parameters lie on the real number line,

2. Parameters with objective priors are often independent of one another, whereas in most multi-parameter statistical models, parameters are correlated. The problem of correlated model parameters is of extreme importance for methods such

as structural equation modeling (see e.g. Kaplan & Wenger, 1993),

3. Expressing complete ignorance about a parameter via an objective prior leads to incorrect inferences about functions of the parameter.

Subjective Priors

To motivate the use of subjective priors, consider again the class size reduction example. In this case, we may have a considerable amount of prior information regarding the increase in achievement arising from previous investigations. It may be that previous investigations used different tests of academic achievement but when examined together, it has been found that reducing class size to approximately 17 children per classroom results in one-quarter of a standard deviation increase (say about 8 points) in academic achievement. In addition to a prior estimate of the average achievement gain due to reduction in class size, we may also wish to quantify our uncertainty about the exact value of θ by specifying a probability distribution around the prior estimate of the average. Perhaps a sensible prior distribution would be a normal distribution centered at $\theta = 8$. However, let us imagine that previous research has shown that achievement gains due to class size reduction has almost never been less than 5 points, and almost never more than 14 points (almost a full standard deviation). Taking this range of uncertainty into account, we might propose a prior distribution on θ that is $N(8, 1)$. The parameters of this prior distribution $\theta = N(8, 1)$ are referred to as *hyperparameters*.

The careful reader may have wondered if setting hyperparameters to fixed values violates the essence of Bayesian philosophy. To address that concern, note first that the Bayesian approach treats the hyperparameters as elicited quantities that are *known and fixed*. The Bayesian approach is to be contrasted with the frequentist approach that treats parameters as *unknown and fixed*. Second, it is not necessary to set hyperparameters to known and fixed quantities. In a fully hierarchical Bayesian model, it is possible to specify a probability distribution on the hyperparameters – referred to as a *hyperprior*.

Informative-Conjugate Priors

In the previous section, we considered the situation in which there may not be much prior information that can be brought to bear on a problem. In that situation we focussed on objective priors. Alternatively, it may be the case that some information can be brought to bear on a problem and be systematically incorporated into the prior distribution. Such subjective priors are deemed

informative. One type of informative prior is based on the notion of a conjugate distribution. As noted earlier, a conjugate prior distribution is one that, when combined with the likelihood function yields a posterior that is in the same distributional family as the prior distribution. Conjugacy is a very important and convenient feature because if a prior is not conjugate, the resulting posterior distribution may have a form that is not analytically simple to solve. Arguably, the existence of numerical simulation methods for Bayesian inference, such as Markov Chain Monte Carlo (MCMC) estimation may render conjugacy less of a problem. We focus on conjugate priors in this section.

Example 3: The Beta Prior

As an example of a conjugate prior, consider estimating the number of correct responses y on a test of length n. Let θ be the proportion of correct responses. We first assume that the responses are independent of one another. The binomial sampling model was given in Equation 14 and reproduced here

$$p(y|\theta) = \text{Bin}(y|n, \theta) = \binom{n}{y}\theta^y(1-\theta)^{(n-y)}. \tag{23}$$

One choice of a prior distribution for θ is the *beta*(a,b) distribution. The beta distribution is a continuous distribution appropriate for variables that range from zero to one. The terms a and b are referred to as *hyperparameters* and characterize the distribution of the parameters, which for the beta distribution are the scale and shape parameters, respectively.[4] The form of the *beta*(a,b) distribution is

$$p(\theta; a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}, \tag{24}$$

where Γ is the *gamma*(a,b) distribution. Ignoring terms that don't involve model parameters, we obtain the posterior distribution

$$p(\theta|y) = \frac{\Gamma(n+a+b)}{\Gamma(y+a)\Gamma(n-y+b)}\theta^{y+a-1}(1-\theta)^{n-y+b-1}, \tag{25}$$

which is *beta* distribution with parameters $a' = a + y$ and $b' = b + n - y$. Thus, the beta prior for the binomial sampling model is conjugate.

Example 4: The Normal Prior

This next example explores the normal prior for the normal sampling model.

Let y denote a data vector of size n. We assume that y follows a normal distribution shown in Equation 15 and reproduced here

$$f(y|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right). \tag{26}$$

Consider that our prior distribution on the mean and variance is also normal with hyperparameters, κ and τ^2 which for this example are known. The prior distribution can be written as

$$f(\mu|\kappa,\tau^2) = \frac{1}{\sqrt{2\pi}\tau^2}\exp\left(-\frac{(\mu-\kappa)^2}{2\tau^2}\right). \tag{27}$$

After some algebra, the posterior distribution can be obtained as

$$f(\mu|y) \sim N\left[\frac{\frac{\kappa}{\tau^2}+\frac{n\bar{x}}{\sigma^2}}{\frac{1}{\tau^2}+\frac{n}{\sigma^2}}, \frac{\tau^2\sigma^2}{\sigma^2+n\tau^2}\right], \tag{28}$$

and so we see that the normal prior is conjugate for the normal likelihood.

The posterior distribution in Equation 28 reveals some interesting features regarding the relationship between the data and the prior. To begin, we see that μ is only dependent on $\bar{x}$, the sample mean – hence $\bar{x}$ is sufficient for μ. Second, we see that as the sample size increases, the data (here $\bar{x}$) become more important than the prior. Indeed, as the sample size approaches infinity, there is no information in the prior distribution that is of relevance to estimating the moments of the posterior distribution. To see this, we compute the asymptotic posterior mean as

$$\begin{aligned}\lim_{n\to\infty}\hat{\mu} &= \lim_{n\to\infty}\frac{\frac{\kappa}{\tau^2}+\frac{n\bar{x}}{\sigma^2}}{\frac{1}{\tau^2}+\frac{n}{\sigma^2}},\\ &= \lim_{n\to\infty}\frac{\frac{\kappa\sigma^2}{n\tau^2}+\bar{x}}{\frac{\sigma^2}{n\tau^2}+1} = \bar{x}.\end{aligned} \tag{29}$$

Finally, we introduce the terms $1/\tau^2$ and n/σ^2 to refer to the *prior precision* and *data precision*, respectively. The role of these two measures of precision can be seen by once again examining the variance term for the normal distribution in Equation

28. Specifically,

$$\lim_{n\to\infty} \hat{\sigma^2} = \lim_{n\to\infty} \frac{1}{\frac{1}{\tau^2} + \frac{n}{\sigma^2}},$$
$$= \lim_{n\to\infty} \frac{\sigma^2}{\frac{\sigma^2}{\tau^2} + n} = \frac{\sigma^2}{n}. \tag{30}$$

A similar result emerges if we consider the case where we have very little information regarding the prior precision. That is, choosing a very large value for τ^2 gives the same result.

Example 5: The Inverse-Gamma prior

In most practical applications, the variance in the normal sampling model is unknown. Thus, we need to derive the joint prior density $p(\mu, \sigma^2)$. Derivation of the joint prior density is accomplished by factoring the joint prior density into the product of the conditional density and marginal density, that is,

$$p(\mu, \sigma^2) = p(\mu|\sigma^2)p(\sigma^2) \tag{31}$$

where in this example

$$\mu|\sigma^2 \sim N(\mu_0, \sigma^2/n) \tag{32}$$
$$\sigma^2 \sim \text{inverse-Gamma}(\nu_0/2, \nu\sigma^2/2), \tag{33}$$

where $\nu_0 > 0$ is a "degree-of-freedom" parameter.

Another important feature of the inverse-Gamma distribution is that if the random variable $x \sim$ inverse-Gamma(a, b) then $1/X \sim$ Gamma(a, b). The relationship between the inverse-Gamma and Gamma distributions is important because $1/\sigma^2$ is the precision parameter. Thus, in the case of the normal model, an inverse-Gamma prior can be placed on σ^2 or a Gamma prior can be place on $1/\sigma^2$

Bayesian Hypothesis Testing

Bayes' theorem shows that the posterior distribution is composed of encoded prior information weighted by the data. With the posterior distribution in hand, it is of interest to obtain summaries of its moments – such as the mean and variance. In addition, interval summaries of the posterior distribution can be obtained. Summarizing the posterior distribution provides the necessary ingredients for

Bayesian hypothesis testing.

Before covering summaries of the posterior distribution and their role in Bayesian hypothesis testing, it may be useful to place the Bayesian approach to a hypothesis testing in contrast to the more common frequentist approach. Clearly, a critically important component of applied statistical modeling is hypothesis testing. Indeed, a considerable amount of time is spent in introductory statistics courses laying the foundation for the frequentist perspective on hypothesis testing, beginning with Fisher (1941/1925) and culminating in the Neyman-Pearson approach which is now the standard in the social and behavioral sciences (Neyman & Pearson, 1928). An interesting aspect of the Neyman-Pearson approach to hypothesis testing is that students (as well as many seasoned researchers) appear to have a very difficult time grasping its principles. In a review of the problem of hypothesis testing in the social and behavioral sciences Gigerenzer, Krauss, and Vitouch (2004) argued that much of the problem lies in the conflation of Fisherian hypothesis testing and the Neyman-Pearson approach to hypothesis testing. For interesting discussions on this problem, see e.g. Cohen (1994), Gigerenzer et al. (2004), and the volume by Harlow, Mulaik, and Steiger (1997).

Briefly, Fisher's approach to hypothesis testing specifies only the null hypothesis. A conventional significance level is chosen (usually the 5% level). Once the test is conducted, the result is either significant ($p < .05$) or it is not ($p > .05$). If the resulting test is significant, then the null hypothesis is rejected. However, if the resulting test is not significant, then no conclusion can be drawn. As Gigerenzer et al. (2004) has pointed out, Fisher developed a later version of his ideas wherein one only reports the exact significance level arising from the test and does not place a "significant" or "non-significant" value label to the result. In other words, one reports, say, $p = .045$, but does not label the result as "significant" (Gigerenzer et al., 2004, pg. 399).

In contrast to Fisher's ideas, the approach advocated by Neyman and Pearson requires that two hypotheses be specified – the null and alternative hypothesis. By specifying two hypotheses, one can compute a desired tradeoff between two types of errors: Type I errors (the probability of rejecting the null when it is true, denoted as α) and Type II errors (the probability of not rejecting the null when it is false, denoted as β).

The conflation of Fisherian and Neyman-Pearson hypothesis testing lies in the

use and interpretation of the p-value. In Fisher's paradigm, the p-value is a matter of convention, with the resulting outcome being based on the data. However, in the Neyman-Pearson paradigm, α and β are determined prior to the experiment being conducted and refer to a consideration of the cost of making one or the other error. In other words the p-value and α are not the same thing. The confusion between these two concepts is made worse by the fact that statistical software packages often report a number of p-values that a researcher can choose after having conducted the analysis (e.g., .001, .01, .05). This can lead a researcher to set α ahead of time, as per the Neyman-Pearson school, but then communicate a different level of "significance" after running the test.

Misunderstandings of the Fisherian approach or the Neyman-Pearson approach to hypothesis testing is not a criticism of these methods per se. However, from the frequentist point of view a criticism often leveled at the Bayesian approach to statistical inference is that it is "subjective", while the frequentist approach is "objective". The objection to "subjectivism" is somewhat perplexing insofar as frequentist hypothesis testing also rests on assumptions that do not involve data. The simplest and most ubiquitous example is the test of a null hypothesis against an alternative hypothesis, characteristic of the Neyman-Pearson paradigm. In cases where the value of the null hypothesis is stated (e.g., something other than zero), the question that is immediately raised is where that value came from. Presumably, a (non-null) value of the null hypothesis must be credible, thus restricting the values that the parameters could sensibly take on. A key difference between Bayesian and frequentist approaches to hypothesis testing, is that the Bayesian approach makes this prior information explicit, and does not find the idea that parameters possess probability distributions contrary to a coherent scheme of hypothesis testing.

Point Estimates of the Posterior Distribution

For frequentist and Bayesian statistics alike, hypothesis testing proceeds after obtaining summaries of relevant distributions. For example, in testing for the differences between two groups (e.g. a treatment group and a control group), we first summarize the data, obtaining the means and standard errors for both groups and then perform the relevant statistical tests. These summary statistics are considered "sufficient" summaries of the data – in a sense, they stand in for data. The difference between Bayesian and frequentist statistics is that with Bayesian

statistics we wish to obtain summaries of the posterior distribution. The expressions for the mean and variance of the posterior distribution come from expressions for the mean and variance of conditional distributions generally. Specifically, for the continuous case, the mean of the posterior distribution can be written as

$$E(\theta|y) = \int_{-\infty}^{+\infty} \theta p(\theta|y) d\theta. \tag{34}$$

Thus, the posterior mean is obtained by averaging over the marginal distribution of θ. Similarly, the variance of θ can be obtained as

$$\begin{aligned} var(\theta|y) &= E[(\theta - E[(\theta|y])^2|y), \\ &= \int_{-\infty}^{+\infty} (\theta - E[\theta|y])^2 p(\theta|y) d\theta, \\ &= \int_{-\infty}^{+\infty} (\theta^2 - 2\theta E[\theta|y]) + E[\theta|y]^2) p(\theta|y) d\theta, \\ &= E[\theta^2|y] - E[\theta|y]^2. \end{aligned} \tag{35}$$

The mean and variance of the posterior distribution provide two simple summary values of the posterior distribution. Another summary measure would be the mode of the posterior distribution - referred to as the *maximum a posteriori* (MAP) estimate. Those measures, along with the quantiles of the posterior distribution, provide a complete description of the distribution.

Interval Summaries of the Posterior Distribution

In addition to these measures, we are often interested in obtaining intervals for, say, the mean of the posterior distribution. There are two general approaches to obtaining interval summaries of the posterior distribution. The first is the so-called *credibility interval* also referred to as the *posterior probability interval*, and the second is the *highest posterior density (HPD) interval*.

Credibility Intervals

One important consequence of viewing parameters probabilistically concerns

the interpretation of *confidence intervals*. Recall that the frequentist confidence interval requires that we imagine a fixed parameter, say the population mean μ. Then, we imagine an infinite number of repeated samples from the population characterized by μ.[5] For any given sample, we obtain the sample mean $\bar{x}$ and form a $100(1-\alpha)\%$ confidence interval. The correct frequentist interpretation is that $100(1-\alpha)\%$ of the confidence intervals formed this way capture the true parameter μ under the null hypothesis. Notice that from this perspective, the probability that the parameter is in the interval is either zero or one.

In contrast, the Bayesian perspective forms a *credibility interval* (also known as a *posterior probability interval*). The credibility interval is obtained directly from the quantiles of the posterior distribution of the model parameters. From the quantiles, we can directly obtain the probability that a parameter lies within a particular interval. Therefore, a $100(1-\alpha)\%$ credible interval means that the probability that the parameter lies in the interval is $100(1-\alpha)\%$. Again, notice that this is entirely different from the frequentist interpretation, and arguably aligns with common sense.

In formal terms, a $100(1-\alpha)\%$ credibility interval for a particular subset of the parameter space θ is defined as

$$1-\alpha = \int_C p(\theta|y)d\theta. \tag{36}$$

The credibility interval will be demonstrated through the examples given later in this chapter.

Highest Posterior Density

The simplicity of the credibility interval notwithstanding, it is not the only way to provide an interval estimate of a parameter. Following the argument set down by Box and Tiao (1973), when considering the posterior distribution of a parameter θ, there is a substantial part of the region of that distribution where the density is quite small. It may be reasonable, therefore, to construct an interval in which every point inside the interval has a higher probability than any point outside the interval. Such a construction is referred to as the *highest probability density (HPD)* interval. More formally,

Definition 1 *Let $p(\theta|y)$ be the posterior density function. A region R of the parameter space θ is called the highest probability density region of the interval $1-\alpha$*

if

1. $pr(\theta \in R|y) = 1 - \alpha$
2. *For* $\theta_1 \in R$ *and* $\theta_2 \notin R$, $pr(\theta_1|y) \geq pr(\theta_2|y)$

Note that for unimodal and symmetric distributions, such as the uniform distribution or the normal distribution, the HPD is formed by choosing tails of equal density. The advantage of the HPD arises when densities are not symmetric and/or are not unimodal. In fact, this is an important property of the HPD and sets it apart from standard credibility intervals. Following Box and Tiao (1973) if $p(\theta|y)$ is not uniform over every region in θ, then the HPD region $1 - \alpha$ is unique. Also if $p(\theta_1|y) = p(\theta_2|y)$ then these points are included (or excluded) by a $1 - \alpha$ HPD region. The opposite is true as well, namely if $p(\theta_1|y) \neq p(\theta_2|y)$ then a $1 - \alpha$ HPD region includes one point but not the other (Box & Tiao, 1973, pg 123).

Bayesian Model Evaluation and Comparison

In many respects, the frequentist and Bayesian goals of model building are the same. First, an initial model is specified relying on a lesser or greater degree of prior theoretical knowledge. In fact, at this first stage, a number of different models may be specified according to different theories, with the goal being to choose the "best" model, in some sense of the word. Second, these models will be fit to data obtained from a sample from some relevant population. Third, an evaluation of the quality of the models will be undertaken, examining where each model might deviate from the data, as well as assessing any possible model violations. At this point, model respecification may come into play. Finally, depending on the goals of the research, the "best model" will be chosen for some purpose.

Despite the similarities between the two approaches with regard to the broad goals of model building, there are important differences. A major difference between the Bayesian and frequentist goals of model building lie in the model specification stage. In particular, because the Bayesian perspective views parameters as possessing probability distributions, the first phase of modeling building will require the specification of a full probability model for the data and the parameters. The probability model for the data is encoded in the likelihood and the probability model for the parameters is encoded in the prior distribution. Thus, the notion of model fit implies that the full probability model fits the data, in some sense, and lack of model fit may well be due to incorrect specification of the prior distribution.

Arguably, another difference between the Bayesian and frequentist goals of model building relate to the justification for choosing a particular model among a set of competing models. Specifically, model building and model choice in the frequentist domain is based primarily on choosing the model that best fits the data. Model fit has certainly been the key motivation for model building, respecification, and model choice in the context of structural equation modeling (see Kaplan, 2009).

In this section, we examine the notion of model building and model fit and discuss a number of commonly used Bayesian approaches. We will first introduce Bayes factors as a very general means of choosing from a set of competing models. This will be followed by a special case of the Bayes factor, referred to as the *Bayesian Information Criterion*. Then, we will consider the *Deviance Information Criterion*. Finally, we will consider the idea of borrowing strength from a number of competing models in the form of Bayesian model averaging.

Bayes Factors

A very simple and intuitive approach to model building and model choice uses so-called *Bayes factors* (Kass & Raftery, 1995). In essence, the Bayes factor provides a way to quantify the odds that the data favor one hypothesis over another. A key benefit of Bayes factors is that models do not have to be nested.

Following Raftery (1995), consider two competing models, denoted as M_1 and M_2, that could be nested within a larger space of alternative models, or possibly obtained from distinct parameter spaces. Further, let θ_1 and θ_2 be two parameter vectors. From Bayes' theorem, the posterior probability that, say M_1, is the model preferred by the data can be written as

$$p(M_1|y) = \frac{p(y|M_1)p(M_1)}{p(y|M_1)p(M_1) + p(y|M_2)p(M_2)} \tag{37}$$

where

$$p(y|M_1) = \int p(y|\theta_1, M_1)p(\theta_1|M_1)d\theta_1 \tag{38}$$

is referred to as the marginal probability or predictive probability of the data given M_1. From here, the posterior odds for M_1 over M_2 can be written as

$$\frac{p(M_1|y)}{p(M_2|y)} = \left[\frac{p(y|M_1)}{p(y|M_2)}\right] \times \left[\frac{p(M_1)}{p(M_2)}\right] \tag{39}$$

where the first term on the right hand side of Equation 39 is the Bayes factor (BF), defined as

$$\begin{aligned} BF &= \frac{p(y|M_1)}{p(y|M_2)} \\ &= \frac{\int p(y|\theta_1, M_1)p(\theta_1|M_1)d\theta_1}{\int p(y|\theta_2, M_2)p(\theta_2|M_2)d\theta_2} \end{aligned} \tag{40}$$

In words, the quantity on the left-hand-side of Equation 39 is the posterior probability of the data favoring M_1 over M_2. This posterior probability is related to the prior odds $p(M_1)/p(M_2)$ of the data favoring M_1 over M_2 weighted by the marginal likelihoods $p(y|M_1)/p(y|M_2)$ as seen in Equation 40. Notice that assuming neutral prior odds, i.e. $p(M_1) = p(M_2) = 1/2$, the Bayes factor is equivalent to the posterior odds.

Rules of thumb have been developed to assess the quality of the evidence favoring one hypothesis over another using Bayes factors. Following Kass and Raftery (1995, pg. 777) and using M_1 as the reference model,

$2log_e(BF_{12})$	BF_{12}	Evidence against M_2
0 to 2	1 to 3	Not worth more than a bare mention
2 to 6	3 to 20	Positive
6 to 10	20 to 150	Strong
> 10	> 150	Very strong

The Bayesian Information Criterion

A difficulty with using Bayes factors for hypothesis testing is the requirement that priors be specified. An alternative that does not require the introduction of prior densities can be obtained using the *Bayesian information criterion* (BIC), also referred to as the Schwarz criterion (SC). The BIC is defined as

$$BIC = -2\log(\hat{\theta}|y) + p\log(n), \tag{41}$$

where $-2\log\hat{\theta}|y$ describes model fit while $p\log(n)$ is a penalty for model complexity, where p represents the number of variables in the model and n is the sample size.

As with Bayes factors, the BIC is often used for model comparisons. Specifically, the difference between two BIC measures comparing, say M_1 to M_2 can

be written as

$$\begin{aligned}\Delta(BIC_{12}) &= BIC_{(M_1)} - BIC_{(M_2)}, \\ &= \log(\hat{\theta}_1|y) - \log(\hat{\theta}_2|y) - \frac{1}{2}(p_1 - p_2)\log(n).\end{aligned} \tag{42}$$

However, unlike the Bayes factor, there is no existing rule of thumb regarding the size of the difference between the BICs of two competing models that would guide a choice. In other words, among competing models, the one with the smallest BIC value is to be chosen.

The Deviance Information Criterion

Although the BIC is derived from a fundamentally Bayesian perspective, it is often productively used for model comparison in the frequentist domain. Recently, however, an explicitly Bayesian approach to model comparison was developed by (Spiegelhalter, Best, Carlin, & Linde, 2002) based on the notion of *Bayesian deviance*.

Consider a particular model proposed for a set of data, defined as $p(y|\theta)$. Then, *Bayesian deviance* can be defined as

$$D(\theta) = -2\log[p(y|\theta)] + 2\log[h(y)] \tag{43}$$

where, according to (Spiegelhalter et al., 2002), the term $h(y)$ is a standardizing factor that does not involve model parameters and thus is not involved in model selection. Note that although Equation 43 is similar to the BIC, it is not, as currently defined, an explicit Bayesian measure of model fit. To accomplish this, we use Equation 43 to obtain a posterior mean over θ by defining

$$\overline{D(\theta)} = E_\theta[-2log[p(y|\theta)|y] + 2log[h(y)], \tag{44}$$

and this is referred to as the *deviance information criterion* (DIC). It has been suggested by Lee (2007, pg. 128) that if the difference between the DIC values of two competing models is less than 5.0 *and* the two models give substantively different conclusions, then it may be misleading to choose the model with the lowest DIC value.

Bayesian Model Averaging

As noted earlier, a key characteristic that separates Bayesian statistical inference from frequentist statistical inference is its focus on characterizing uncertainty. Up to this point, we have concentrated on uncertainty in model parameters, addressing that uncertainty through the specification of a prior distribution on the model parameters. In a related, but perhaps more general fashion, the selection of a particular model from a universe of possible models can also be characterized as a problem of uncertainty. This problem was succinctly stated by Hoeting, Madigan, Raftery, and Volinsky (1999) who write

> "Standard statistical practice ignores model uncertainty. Data analysts typically select a model from some class of models and then proceed as if the selected model had generated the data. This approach ignores the uncertainty in model selection, leading to over-confident inferences and decisions that are more risky than one thinks they are."(pg. 382)

An interesting approach to addressing the problem of model uncertainty lies in the method of *Bayesian model averaging* (BMA).

To begin, consider once again a parameter of interest θ (which could be vector valued) and consider a set of competing models M_k, $k = 1, 2, \ldots, K$ that are not necessarily nested. The posterior distribution of θ given data y can be written as

$$p(\theta|y) = \sum_{k=1}^{K} p(\theta|M_k)p(M_k|y), \tag{45}$$

where $p(M_k|y)$ is the posterior probability of model M_k written as

$$p(M_k|y) = \frac{p(y|M_k)p(M_k)}{\sum_{l=1}^{K} p(y|M_l)p(M_l)}, \qquad l \neq k. \tag{46}$$

In words, Equation 46 indicates that one can obtain the posterior probability of a model by multiplying the likelihood of the data given the model, times the prior probability placed on the model. The prior probability $p(M_k)$ can be different for different models. Note that denominator in Equation 46 simply ensures that the probability sums to one. Note also that the term $p(y|M_k)$ can be expressed as an

integrated likelihood

$$p(y|M_k) = \int p(y|\theta_k, M_k)p(\theta_k|M_k)d\theta_k, \tag{47}$$

over the parameters of interest, and where $p(\theta_k|M_k)$ is the prior density of θ_k. Thus, BMA provides an approach for combining models specified by researchers, or perhaps elicited by key stakeholders. The advantage of BMA has been discussed in Madigan and Raftery (1994) who showed that BMA provides better predictive performance than that of a single model.

As pointed out by Hoeting et al. (1999) BMA is difficult to implement. In particular, they note that that the number of terms in Equation 45 can be quite large, the corresponding integrals are hard to compute (though possibly less so with the advent of MCMC), specification of $p(M_k)$ may not be straightforward, and choosing the class of models to average over is also challenging. The problem of reducing the overall number of models that one could incorporate in the summation of Equation 45 has lead to interesting solutions based on the notion of *Occam's window* (Madigan & Raftery, 1994) or the "leaps and bounds" algorithm (Volinsky, Madigan, Raftery, & Kronmal, 1997), discussions of which are beyond the scope of this chapter.

Bayesian Computation

As stated in the introduction, the key reason for the increased popularity of Bayesian methods in the social and behavioral sciences has been the advent of freely available software programs for Bayesian estimation of the parameters of a model. The most common estimation algorithm is based on MCMC sampling. A number of very important papers and books have been written about MCMC sampling (see e.g., Gilks, Richardson, & Spiegelhalter, 1996). The general idea is that instead of attempting to analytically solve a complex integral problem, the MCMC approach instead draws specially constructed samples from the posterior distribution $p(\theta|y)$ of the model parameters. In the interest of space, we will concentrate on one common algorithm for MCMC sampling, referred to as *Gibbs Sampling* (Geman & Geman, 1984). More general treatments of MCMC can be found in Bolstad (2009); Casella and Robert (2003); Gilks et al. (1996)

Gibbs Sampling

The formal algorithm can be specified as follows. Let $\boldsymbol{\theta}$ be a vector of model parameters with elements $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_q\}$. The elements of $\boldsymbol{\theta}$ could be the parameters of a regression model, structural equation model, etc. Note that information regarding $\boldsymbol{\theta}$ is contained in the prior distribution $p(\boldsymbol{\theta})$. A number of algorithms and software programs are available to conduct MCMC sampling. Following the description given in Gilks et al. (1996), the Gibbs sampler begins with an initial set of starting values for the parameters, denoted as $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, \ldots, \theta_q^{(0)})$. Given this starting point, the Gibbs sampler generates $\theta^{(s)}$ from $\theta^{(s-1)}$ as follows:

$$
\begin{aligned}
&1.\ \text{sample} \quad \theta_1^{(s)} \sim p(\theta_1 | \theta_2^{(s-1)}, \theta_3^{(s-1)}, \ldots, \theta_q^{(s-1)}, \mathbf{y}) \\
&2.\ \text{sample} \quad \theta_2^{(s)} \sim p(\theta_2 | \theta_1^{(s)}, \theta_3^{(s-1)}, \ldots, \theta_q^{(s-1)}, \mathbf{y}) \\
&\qquad\qquad \vdots \\
&q.\ \text{sample} \quad \theta_q^{(s)} \sim p(\theta_q | \theta_1^{(s)}, \theta_2^{(s)}, \ldots, \theta_{q-1}^{(s)}, \mathbf{y}).
\end{aligned}
$$

Then, a sequence of dependent vectors are formed

$$
\begin{aligned}
\theta^{(1)} &= \{\theta_1^{(1)}, \ldots \theta_q^{(1)}\} \\
\theta^{(2)} &= \{\theta_1^{(2)}, \ldots \theta_q^{(2)}\} \\
&\vdots \\
\theta^{(S)} &= \{\theta_1^{(S)}, \ldots \theta_q^{(S)}\}.
\end{aligned}
$$

This sequence exhibits the so-called *Markov property* insofar as $\theta^{(s)}$ is conditionally independent of $\{\theta_1^{(0)}, \ldots \theta_q^{(s-2)}\}$ given $\theta^{(s-1)}$. Under some general conditions, the sampling distribution resulting from this sequence will converge to the target distribution as $s \to \infty$. See Gilks et al. (1996) for additional details on the properties of MCMC.

In setting up the Gibbs sampler, a decision must be made regarding the number of Markov chains to be generated, as well as the number of iterations of the sampler. With regard to the number of chains to be generated it is not uncommon to specify multiple chains. Each chain samples from another location of the posterior distribution based on purposefully dispersed starting values. With multiple chains it may be the case that fewer iterations are required, particularly if there is evidence for the chains converging to the same posterior mean for each

parameter. In some cases, the same result can be obtained from one chain, although often requiring a considerably larger number of iterations. Once the chain has stabilized, the iterations prior to the stabilization (referred to as the *burn-in* phase) are discarded. Summary statistics, including the posterior mean, mode, standard deviation and credibility intervals, are calculated on the post-burn-in iterations. Also, convergence diagnostics (discussed next) are obtained on the entire chain or on post-burn-in iterations.

Convergence Diagnostics

Assessing the convergence of parameters within MCMC estimation is a difficult task that has been receiving attention in the literature for many years (see e.g., Mengersen, Robery, & Guihenneuc-Jouyax, 1999; Sinharay, 2004). The difficulty of assessing convergence stems from the very nature of MCMC in that the MCMC algorithm is designed to converge in distribution rather than to a point estimate. Because there is not a single adequate assessment of convergence for this situation, it is common to inspect several different diagnostics that examine varying aspects of convergence conditions.

Perhaps the most common form of assessing MCMC convergence is to examine the convergence (also called history) plots produced for a chain. Typically, a parameter will appear to converge if the sample estimates form a tight horizontal band across this history plot. However, using this method as an assessment for convergence is rather crude since merely viewing a tight plot does not indicate convergence was actually obtained. As a result, this method is more likely to be an indicator of non-convergence (Mengersen et al., 1999). For example, if two chains for the same parameter are sampling from different areas of the target distribution, there is evidence of non-convergence. Likewise, if a plot shows substantial fluctuation or jumps in the chain, it is likely the parameter has not reached convergence. However, because merely viewing history plots may not be sufficient in determining convergence (or non-convergence), it is also common to reference additional diagnostics. Although this list is not exhaustive, we focus on several of the most commonly used diagnostics for single chain situations. All of these diagnostics are available through loading the convergence diagnostic and output analysis (CODA) (Best, Cowles, & Vines, 1996) files (produced by programs such as WinBUGS) into the Bayesian Output Analysis (BOA) program (Smith, 2005)

interface for R (R Development Core Team, 2008a).

The Geweke convergence diagnostic (Geweke, 1992) is used with a single chain to determine whether or not the first part of a chain differs significantly from the last part of a chain. The motivation for this diagnostic is rooted in the dependent nature of an MCMC chain. Specifically, since samples in a chain are not independently and identically distributed., convergence can be difficult to assess due to the inherent dependence between adjacent samples. Stemming from this dilemma, Geweke constructed a diagnostic that aimed at assessing two independent sections of the chain. BOA allows the user to set the proportion of iterations to be assessed at the beginning and the end of the chain. The default for the program mimics the standard suggested by Geweke (1992) which is to compare the first 10% of the chain and the last 50% of the chain. Although the user can modify this default, it is important to note that there should be a sufficient number of iterations between the two samples to ensure the means for the two samples are independent. This method computes a z-statistic where the difference in the two sample means is divided by the asymptotic standard error of their difference. A z-statistic falling in the extreme tail of a standard normal distribution suggests that the sample from the beginning of the chain has not yet converged (Smith, 2005). BOA produces an observed z-statistic and two-sided p-value. It is common to conclude that there is evidence against convergence with a p-value less than 0.05.

The Heidelberger and Welch convergence diagnostic (Heidelberger & Welch, 1983) is a stationarity test that determines whether or not the last part of a Markov chain has stabilized. This test uses the Cramer-von-Mises statistic to assess evidence of non-stationarity. If there is evidence of non-stationarity, the first 10% of the iterations will be discarded and the test will be repeated either until the chain passes the test or more than 50% of the iterations are discarded. If the latter situation occurs, it suffices to conclude there was not a sufficiently long stationarity portion of the chain to properly assess convergence (Heidelberger & Welch, 1983). The results presented in BOA report the number of iterations that were retained as well as the Cramer-von-Mises statistic. Each parameter is given a status of having either passed the test or not passed the test based on the Cramer-von-Mises statistic. If a parameter does not pass this test, this is an indication that the chain needs to run longer before achieving convergence. A second stage of this diagnostic examines the portion of the iterations that pass the stationarity test for accuracy. Specifically, if

the half-width of the estimate confidence interval is less than a preset fraction of the mean, then the test implies the mean was estimated with sufficient accuracy. If a parameter fails under this diagnostic stage (indicating low estimate-accuracy), it may be necessary for a longer run of the MCMC sampler (Smith, 2005).

The Raftery and Lewis convergence diagnostic (Raftery & Lewis, 1992) was originally developed for Gibbs sampling and is used to help determine three of the main features of MCMC: the burn-in length, the total number of iterations, and the thinning interval (described below). A process is carried out that identifies this information for all of the model parameters being estimated. This diagnostic is specified for a particular quantile of interest with a set degree of accuracy within the BOA program. Once the quantile of interest and accuracy are set, the Raftery and Lewis diagnostic will produce the number of iterations needed for a burn-in and a range of necessary post burn-in iterations for a particular parameter to converge. For each of these, a lower-bound value is produced which represents the minimum number of iterations (burn-in or post-burn-in) needed to estimate the specified quantile using independent samples. Note, however, that the minimum value recommended for the burn-in phase can be optimistic and larger values are often required for this phase (Mengersen et al., 1999).

Finally, information is also provided about the thinning interval that should be used for each parameter. Thinning is a process of sampling every s^{th} sequence of the chain for purposes of summarizing the posterior distribution. Thinning is often used when autocorrelations are high, indicating that consecutive draws are dependent. To reach independence between samples, it is common to discard a number of successive estimates between draws that are used for estimation. Thinning involves comparing first-order and second-order Markov chains together for several different thinning intervals. Comparison of first and second order Markov chains is accomplished through computing G^2, a likelihood-ratio test statistic between the Markov models (Raftery & Lewis, 1996). After computing G^2, the BIC can then be computed in order to compare the models directly (Raftery & Lewis, 1996). The most appropriate thinning interval is chosen by adopting the smallest thinning value produced where the first-order Markov chain fits better than the second-order chain.

Although the default in the BOA program is to estimate the 0.025 quantile, the 0.5 quantile is often of more interest in determining the number of iterations needed for convergence because interest typically focuses on the central tendency of

the distribution. It is important to note that using this diagnostic is often an iterative process in that the results from an initial chain may indicate that a longer chain is needed to obtain parameter convergence. A word of caution is that over-dispersed starting values can contribute to the Raftery and Lewis diagnostic requesting a larger number of burn-in and post burn-in iterations. On a related note, Raftery and Lewis (1996) recommend that the maximum number of burn-in and post burn-in iterations produced from the diagnostic be used in the final analysis. However, this may not always be a practical venture when models are complex (e.g., longitudinal mixture models) or starting values are purposefully over-dispersed.

Three Empirical Examples

In this section we provide three simple examples of the application of Bayesian statistical inference: (1) Bayesian multiple regression analysis, (2) Bayesian multilevel modeling, and (3) Bayesian confirmatory factor analysis. The intent of this section is to present three standalone examples that, in part, illustrate how to interpret and report analyses produced through a Bayesian framework. It is not the intention of this section to compare results to those from a frequentist-based analysis. In fact, it is expected in analyses with large samples and non-informative priors that the Bayesian results would be close to those obtained from a frequentist analysis. Differences between the two approaches might appear in comparing credibility intervals to confidence intervals, but the reasons for conducting a Bayesian analysis lie in the philosophical differences underlying the two approaches, which we discuss in the Conclusions and Future Directions section.

Bayesian Multiple Regression Analysis

For this example, we use an unweighted sample of 550 kindergartners from the Early Childhood Longitudinal Study–Kindergarten Class of 1998 (NCES, 2001). Item response theory was used to derive scale scores for a math assessment given in the fall of kindergarten. These scores are used as the dependent variable in this analysis. There are two sets of predictors included in this model. The first set of predictors is comprised of three items that the teacher answered for each student regarding certain social and behavioral issues within the classroom. These three items inquired about each student's approach to learning, self-control, and interpersonal skills. The second set of predictors included three similar items that

the parent answered regarding their child in the home environment. These three items were approaches to learning, self-control, and social interaction. This model includes all six teacher and parent items as predictors of math achievement.

For the purposes of this example, this model was computed through the R environment (R Development Core Team, 2008b) using the *MCMCreg* function within the *MCMCpack* package to carry out the analysis (Martin, Quinn, & Park, 2010). Note, however, that this model can be computed both in other packages within R and also in alternative programs such as WinBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000) and Mplus (Muthén & Muthén, 2010). All of the model parameters were given non-informative prior distributions.

Parameter Convergence

The results obtained through *MCMCpack* were read into the *CODA* package (Best et al., 1996) that provides many different convergence diagnostics discussed earlier. The Geweke convergence diagnostic was computed using the default *CODA* proportions of 0.1 for the beginning of the chain and 0.5 for the end of the chain. None of the parameters produced significant z-scores, indicating there was no evidence against convergence. The Heidelberger and Welch convergence diagnostic indicated that all of the parameters passed the stationarity and half-width tests. Finally, the Raftery and Lewis diagnostic was computed with the following settings: quantile = 0.5, accuracy = 0.05, and probability = 0.95. Results indicated that the burn-in should consist of at least two iterations, the total number of iterations should be at least 3897, and that no thinning interval was necessary. A more conservative analysis with 1,000 burn-in iterations and 10,000 post burn-in iterations was conducted with little computational cost. The results of these diagnostics indicated that the parameters in this model appeared to properly converge.

Model Interpretation

Estimates for the final unstandardized regression analysis can be found in Table 1. The means and standard deviations of the posterior distributions are provided for each model parameter. The Monte Carlo (MC) error is also included in this table. This estimate is of the MC standard error of the mean of the posterior distribution. Finally, the 95% credibility interval is also provided for each parameter. As an example, the unstandardized regression weight for the teacher-reported assessment of a student's approach to learning was 3.81 with a standard deviation of 0.59. The 95% credible interval for this parameter ranges from a lower bound of

2.65 to an upper bound of 4.98. The interpretation of this interval differs from the interpretation of a frequentist confidence interval in that the credibility interval indicates there is a 0.95 probability that the parameter falls in this range of values.

Figure 1 presents convergence plots and posterior density plots for the three teacher predictors and the three parent predictors. The convergence plots exhibit a relatively tight, horizontal band for the predictors, indicating that there was no sign of non-convergence. Non-convergence is typically identified by convergence bands that bounce around in an instable fashion, rather than forming a tight horizontal band. The posterior densities in Figure 1 approximate a normal distribution, which is another indication of parameter convergence. If the density plots exhibit non-normal, or lumpy distributions, this can be a sign that the MCMC chain has not converged properly to the posterior distribution.

Model Comparison

For the purposes of illustrating Bayesian model comparison, two additional regression models have been estimated using the same dependent variable of math achievement but a restricted set of predictors. The first model includes only the teacher-related predictors and results from this analysis can be found in the middle section of Table 1. The second model includes the parent-related predictors and results can be found in the bottom portion of Table 1. Both of these models will be used as a comparison to the original full model containing all of the predictors.

As discussed earlier, the Bayes factor can be used as a tool to quantify the odds of the data favoring one model over another. For the first comparison, the full model with all six predictors will be compared to the model only containing the teacher-related predictors. Using Equation 40, the Bayes factor for this model comparison was computed through the *BayesFactor* function in *MCMCpack* available through R.

The result comparing the full model to the model containing only the teacher-related items yielded a Bayes factor value of 65.00. According to the criteria presented earlier this indicates strong evidence against the restricted model containing only the teacher-related items.

In a similar fashion, the second comparison involves the full model and the model only containing the parent-related predictors. The Bayes factor computed for this comparison was 1.56E+11, indicating very strong evidence against the restricted model containing only the parent-related items. Finally, by comparing the

two restricted models to one another, a Bayes factor value between zero and 1.0 (4.17E-10) was produced. Values less than zero indicate that the model in the denominator (M_2) of the Bayes factor is favored over the model in the numerator (M_1). In this case, there was very strong evidence against the restricted model containing only the teacher-related items.

It is important to point out how this example differs from a frequentist approach to the problem. In particular, the Bayes factor is providing information about the magnitude of support in the data *favoring* one hypothesis over the other. This is in stark contrast to the frequentist view of acceptance versus rejection of a hypothesis given the data.

Bayesian Model Averaging

The full regression model with all parent and teacher predictor variables is used here to demonstrate Bayesian modeling averaging via the *BMA* package (Raftery, Hoeting, Volinsky, Painter, & Yeung, 2009) in R (R Development Core Team, 2008b).[6] The BMA package in R automatically produces the top 5 selected models and these are displayed in Table 2. These models are selected based on posterior model probability values. For each variable in the model, the *posterior effect probability* (POST PROB) gives the effect size of the variable in the metric of posterior probability and is used to draw inferences about the importance of each variable. Specifically, the posterior effect probability is the probability that the regression coefficient is not zero, taking into account model uncertainty.

The Bayesian model averaged coefficients (AVG COEF) are the weighted average of coefficients associated with the specific variable across the top 5 models, weighted by each model's posterior model probability (PMP). For example, the weighed model average coefficient for TEACHER1 is 4.19, with a weighted model averaged standard deviation of 0.53. The posterior effect probability of this coefficient is 1.0, and thus implies that its averaged posterior distribution has 0% of its mass at zero. By contrast, TEACHER2, has a weighted model averaged coefficient of 0.04 with standard deviation of 0.21. The averaged posterior distribution for this coefficient has 94% of its mass at zero, or, in other words, the probability that the TEACHER2 coefficient is not zero is 0.06. As stressed by Hoeting et al. (1999), these parameter estimates and standard deviations account for model uncertainty.

Finally, the model with highest PMP is Model 1 with a probability of 0.77.

This model also produced the lowest BIC value, but it is interesting to note that R^2 values yield inconsistent findings. For future predictive studies, one would use the coefficients shown under AVG COEF, as these have been shown to provide the best predictive performance (see Hoeting et al., 1999). The R syntax for this example is given in Appendix A.

Bayesian Hierarchical Linear Modeling

This example of a two-level hierarchical linear model uses a sample of 110 kindergartners from 39 schools from the ECLS–K database (NCES, 2001). The same math assessment measure from the multiple regression example is used as an outcome here. There are two predictors at Level-1 in this model. The first is a measure assessing the parent's perception of their child's approach to learning. The second predictor is the parent's assessment of their child's self-control. This example was computed through WinBUGS (Lunn et al., 2000), however, there are several packages within the R environment that will estimate this type of model. The WinBUGS syntax is given in Appendix B and all model parameter were given non-informative priors.

Parameter Convergence

An initial model was computed with no burn-in samples and 10,000 total iterations to assess preliminary parameter convergence. This model took about one second to compute. The Geweke diagnostic and the Heidelberger and Welch diagnostic would not compute as a result of substantial divergence within the chains. The Raftery and Lewis diagnostic was computed with the following values: quantile = 0.5, accuracy = 0.05, and probability = 0.95. Results indicated that the longest chain should run for up to 304,168 post burn-in iteration for the 0.5 quantile with a thinning interval up to 193 and a burn-in of 2,509. A final model took these recommendations into consideration and was computed with 20,000 burn-in iterations, 255,000 post burn-in iterations, and no thinning interval. The decision to not include a thinning interval was based on the auto-correlation plots in the initial model as well as the fact that such a large number of post burn-in iterations were being used for the final model. The Geweke convergence diagnostic for this final model indicated that none of the parameters produced significant z-scores. Likewise, the Hiedelberger and Welch diagnostic indicated that all of the parameters passed the stationarity and half-width tests. Based on these diagnostics, all of the

parameters in this model appeared to converge properly. Despite the large number of iterations, this model took less than 2 minutes to run.

Model Interpretation

Estimates for the final hierarchical linear model are presented in Table 3. The means and standard deviations of the posterior distributions are provided for each parameter. Likewise, the MC error and the 95% credibility interval are also provided. The fixed effects for this model are presented in the table first. Results indicated that the intercept for this model was -2.61, representing the expected scaled-math score for a student corresponding to parent-perceptions for the predictors coded as zero. Likewise, the 95% credible interval ranged from -4.12 to -1.10, indicating that there is a 0.95 probability the true parameter value falls in this range. The slope corresponding to the parent-perception of the child's approach to learning was 4.85 and the slope for the parent-perception of the child's self-control was 2.66. Table 3 also presents the correlations between the fixed effects. The two slope parameters have a larger correlation, with an estimate of 0.47. The intercept had lower but comparable correlations between the respective slope parameters.

Figure 2 presents convergence plots, posterior density plots, and auto-correlation plots for all three fixed effects. The convergence plots exhibit a relatively tight, horizontal band for the intercept and the two slopes. The posterior densities approximate a normal distribution, with the intercept exhibiting more variability in the density compared to the two slopes. Finally, the auto-correlation plots all show diminishing dependence within the chain. If auto-correlations were high, this would indicate that the starting values likely had a large impact on the location of the chain. Lower auto-correlations are desirable since the location of the chain should not be depending on the starting values, but rather should be determined by the posterior distribution. Although not presented here, the other parameters in the model showed similar results.

Bayesian Confirmatory Factor Analysis

The data for the Bayesian confirmatory factor analysis example come from the responses of a sample of 3500 public school 10th grade students to survey items in the National Educational Longitudinal Study (NCES, 1988). Students were asked to respond to questions assessing their perceptions of the climate of the school. Questions were placed on a 4-point Likert scale ranging from *strongly agree* to

strongly disagree. A prior exploratory factor analysis using principal axis factoring with promax rotation revealed two correlated factors. The item and factor definitions are given in Table 4. We use the two-factor solution for the Bayesian CFA example. This model was estimated using non-informative priors on the model parameters through WinBUGS; the syntax for this example is given in Appendix C.

Parameter Convergence

An initial model was computed with no burn-in samples and 5,000 total iterations in order to assess preliminary parameter convergence. This model took about 8 minutes to compute. The Geweke convergence diagnostic was computed using the default BOA proportions of 0.1 for the beginning of the chain and 0.5 for the end of the chain. None of the parameters produced significant z-scores, indicating there was no evidence against convergence based on the Geweke diagnostic. Likewise, the Heidelberger and Welch convergence diagnostic yielded results indicating that all of the parameters passed the stationarity and half-width tests. The Raftery and Lewis diagnostic was computed with the following values: quantile = 0.5, accuracy = 0.05, and probability = 0.95. Results indicated that the longest chain should run for up to 5,555 post burn-in iterations for the 0.5 quantile with a thinning interval up to 11 and a burn-in of 44 iterations to converge. A final model was computed based on these recommendations with a burn-in phase of 1,000 and 5,000 post burn-in iterations. Upon inspection of auto-correlation plots for the initial model, it was deemed that no thinning interval was necessary for the final analysis. Based on the diagnostics, all of the parameters in this model appeared to converge properly. This model took approximately 10 minutes to run. The length of time it took to run these models is probably due to the large sample size.

Model Interpretation

Estimates for the final CFA model are presented in Table 4. The means and standard deviations of the posterior distributions are provided for each parameter. The MC error is also included in this table as well as the 95% credibility interval for each parameter. The first factor consisted of postive perceptions of the school climate while the second factor consisted of negative perceptions of the school climate. Note that the first item on each factor was fixed to have a loading of 1.00 in order to set the metric of that factor. However, the flexibility of modeling in a Bayesian framework will allow for any method of scale setting.

The factor assessing positive perceptions of school climate measures had high

(unstandardized) loadings ranging from 0.94 to 1.11. The factor measuring negative perceptions of school climate had slightly lower loadings overall, ranging from 0.80 to 0.97. Notice that all of the 95% credibility intervals are relatively tight for all of the items. For example, the interval for the item measuring the level students get along ranged from 0.95 to 1.03. This indicates that there is a 0.95 probability that the true loading for this item is in this range. Estimates for factor precisions (inverse of the variance), error term variances, and the residual variance/precision are also included in Table 4.

Figure 3 presents convergence plots, posterior density plots, and auto-correlation plots for two of the factor loadings and the corresponding error variances. The convergence plots exhibit a tight, horizontal band for both of the items presented. In conjunction with the convergence diagnostics presented above, this tight band indicates the parameters likely converged properly. The posterior probability densities are approximating a normal distribution, and the auto-correlations are very low indicating sample independence within the chain. Although not shown here, the other parameters included in this model also exhibited proper convergence and low auto-correlations.

Conclusions and Future Directions

This chapter provided a very general overview of Bayesian statistical methods, including elements of Bayesian probability theory, inference, hypothesis testing, and model comparison. We provided very simple examples of Bayesian inference to multiple regression, multilevel modeling, and confirmatory factory analysis in order to motivate the Bayesian approach. It should be pointed out, however, that with the advent of simulation methods for estimating model parameters, virtually all of the common statistical models used in the social and behavioral sciences can be estimated from a Bayesian perspective.

The broad range of models that can be estimated via the Bayesian perspective comes with a price. First, although the MCMC sampling conducted for the examples in this paper took very little time, Bayesian inference via MCMC sampling can take a very long time to run – particularly when compared with maximum likelihood based alternative algorithms such as the expectation-maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977). The issue of extensive computational time is particularly problematic when estimating

models involving finite mixture distributions. Second, there does not currently exist simple "pull down menu" functionality for Bayesian oriented software programs such as WinBUGS or the packages within R. Although it is expected that such functionality will be available in the future, for now, there is a great deal of start-up learning that is required to properly specify and estimate Bayesian models.

Perhaps a more important consideration when embarking on the use of Bayesian inference are the epistemological differences between the Bayesian and frequentist approaches for model building and model selection. As noted earlier, the key epistemological differences between the Bayesian and frequentist perspective include (1) the view that parameters are random and unknown versus fixed and unknown, (2) accepting the validity of the subjective belief framework of probability; quantifying the degree of belief about model parameters in the form of the specification of the prior distribution; and updating that belief in the presence of data, and (3) a shift away from the Fisherian or Neyman and Pearson schools of hypothesis testing and toward an approach based on model selection and posterior predictive accuracy. Thus, although the Bayesian and frequentist results look similar under certain conditions (e.g. large sample sizes and diffuse priors), it does not suggest that they are the same or that they are providing necessarily comparable interpretations. These differences in outlook between the Bayesian approach and the frequentist approach imply that MCMC sampling should not be considered "just another estimator", i.e., no different than, say maximum likelihood or weighted least squares. Rather, if the Bayesian perspective is an appealing approach to data modeling in the social and behavioral sciences, then due consideration must be given as to whether one is comfortable with the epistemological shift that comes from adopting this approach.

We see three important future directions for Bayesian inference in the social and behavioral sciences. First, from a purely practical point of view, it will be difficult to convince social and behavioral science researchers to adopt Bayesian methods unless computational algorithms become both easier to use and considerably faster. Second, it will be important to introduce students to Bayesian methods much earlier in their statistical training, and to articulate the epistemological differences between the Bayesian and frequentist approaches so that students understand precisely the choices they are making. Finally, it will take a slow but steady paradigm shift in the practice of social and behavioral science in

order to move away from conventional hypothesis testing as currently employed and toward the Bayesian perspective.

Appendix A: Glossary

Term	Definition
Bayes factor	A quantity indicating the odds that the data favor one hypothesis over another. With equal prior odds, the Bayes factor is the ratio of the marginal likelihoods.
Bayes' theorem	A theorem originated by the Reverend Thomas Bayes' and popularized by Pierre-Simon Laplace relating conditional probability to its inverse form.
BIC	*Bayesian information criterion.* A statistic used for model selection based on the Bayes factor but not requiring prior distributions.
BMA	*Bayesian model averaging.* A method to account for model uncertainty when specifying and comparing a number of different models.
Burn-in	In MCMC, the iterations prior to the stabilization of the chain.
Conditional probability	The probability of an event given the occurrence or observation of another event.
Credibility interval	Also referred to as the *posterior probability interval.* An interval of the posterior distribution used for interval estimation in Bayesian statistics.
DIC	*Deviance information criterion.* A model selection criterion used to select a model with the best sample predictive performance.
EAP	*Expected a posteriori estimate.* In Bayesian inference, the EAP corresponds to the mean of the posterior distribution.
EM algorithm	An iterative algorithm for finding maximum likelihood estimates of model parameters.
Exchangeability	A sequence of random variables such that future samples behave like earlier samples, meaning that any order of a finite number of samples is equally likely.
Frequentist paradigm	A statistical paradigm based on the view of probability as the limiting quantity in long-run frequency.
HPD	Highest posterior density. An interval in which every point inside the interval has a higher probability than any point outside the interval.
Hyperparameters	The parameters of the prior distribution.
Hyperprior distribution	The prior distribution on the hyperparameters.
Jeffreys' prior	A non-informative prior distribution that is proportional to the square root of the determinant of the Fisher information matrix.
Likelihood	A statistical function of the parameters of a model, assumed to have generated the observed data.
MAP	Maximum a posteriori estimate. The mode of the posterior distribution.
MCMC	*Markov chain Monte Carlo.* In Bayesian statistics, a family of algorithms designed to sample from the posterior probability distribution, in which the equilibrium distribution is the target distribution of interest. Algorithms include the Gibbs sampler and the Metropolis-Hastings algorithm.
Objective prior distribution	A prior distribution in which the specification of the hyperparameters suggest that very little information is conveyed by the distribution. Also referred to as *public policy prior*, *uninformative prior* or *vague prior*.
Post burn-in	In MCMC, the iterations after stabilization of the chain and used for obtaining summaries of the posterior distribution.
Posterior distribution	The distribution of an event after conditioning on relevant prior information.
Precision	The reciprocal of the variance.
Prior distribution	The distribution over the model parameters, characterized by *hyperparameters* that encode beliefs about the model parameters.
Subjective prior distribution	A prior distribution in which the specification of the hyperparameters conveys prior beliefs about the model parameters.
Thinning	A process of sampling every sth sequence of the chain for purposes of summarizing the posterior distribution. Thinning is often used to reduce auto-correlation across chains.

Appendix B

Multiple Regression, CODA, Bayes Factors, and Bayesian Model Averaging R Code

#Multiple Regression Analysis :

```
library(MCMCpack)
datafile <– read.csv("C:/File Path/datafile.csv",header=T)
FullModel <– MCMCregress(math∼teacher1+teacher2+teacher3+parent1+
parent2+parent3,data=datafile,marginal.likelihood="Chib95",mcmc=10000,b0=0,
B0=c(.01,.01,.01))
plot(FullModel) # Produces the convergence plots and the posterior densities
dev.off()
summary(FullModel)
TeacherModel <– MCMCregress(math∼teacher1+teacher2+teacher3,
data=datafile,marginal.likelihood="Chib95",mcmc=10000,b0=0,
B0=c(.01,.01,.01))
plot(TeacherModel)
dev.off()
summary(TeacherModel)
ParentModel <– MCMCregress(math∼parent1+parent2+parent3,
data=datafile,marginal.likelihood="Chib95",mcmc=10000,b0=0,
B0=c(.01,.01,.01))
plot(ParentModel)
dev.off()
summary(ParentModel)
```

#Bayes Factors :

```
bf <– BayesFactor(TeacherModel, FullModel)
print(bf)
bf <– BayesFactor(ParentModel, FullModel)
print(bf)
bf <– BayesFactor(TeacherModel, FullModel) print(bf)
```

#Convergence Diagnostics :

```
library(coda)
geweke.diag(FullModel, frac1=0.1, frac2=0.5) # Geweke convergence diagnostic
heidel.diag(FullModel,eps=0.1,pvalue=0.05) # Heidelberger-Welch convergence diagnostic
raftery.diag(FullModel,q=0.5,r=0.05,s=0.95,converge.eps=0.001) # Raftery-Lewis convergence diagnostic
```

#Bayesian Model Averaging :

```
library(BMA)
setwd("C:/File Path/") # Setting working directory
datafile=read.table("datafile.txt",header=TRUE)
attach(datafile)
bma=bicreg(cbind(teacher1,teacher2,teacher3,parent1,parent2,parent3),math,
strict=FALSE,OR=20)
summary(bma)
plot(bma) # Plots of BMA posterior distributions
imageplot.bma(bma) # The image plot shows which predictors are included in each model
```

Appendix C

Two-Level Hierarchical Linear Modeling in WinBUGS: Two Level-1 Predictors

```
model
#N = number of students, J = number of schools
for (i in 1: N)
Y[i]∼dnorm(mu[i], tau.r[i])
#Regression equation in terms of Level − 2 (schools)
#b[school[i], 1] = intercept
#b[school[i], 2] = slope1
#b[school[i], 3] = slope2
mu[i] <– b[school[i],1] + b[school[i],2]*x[i,1] + b[school[i],3]*x[i,2]
for (j in 1:J) # School-level
b[j,1:3]∼dmnorm(b00[j,],Tau[,]) # Distributions on all 3 regression parameters
for (i in 1:N)
tau.r[i]∼dgamma(3,3) # Distribution on data precision
sigma2.r[i] <– 1/tau.r[i]
for (j in 1:J)
b00[j,1:3]∼dmnorm(B.hat[j,1:3],Tau[,]) # Hyperpriors for the mean on 3 regression parameters
B.hat[j,1]<–g00[1] # Creating intercept fixed effect
B.hat[j,2]<–g00[2] # Creating slope 1 fixed effect
B.hat[j,3]<–g00[3] # Creating slope 2 fixed effect
#Prior specification for fixed effects
g00[1]∼dnorm(0,1) # Distribution on intercept fixed effect
g00[2]∼dnorm(0,1) # Distribution on slope 1 fixed effect
g00[3]∼dnorm(0,1) # Distribution on slope 2 fixed effect
#Setting up fixed effect correlations
Tau[1:3,1:3]∼dwish(R1[1:3,1:3],110) # Precision matrix for all fixed effects
Cov[1:3,1:3]<–inverse(Tau[1:3,1:3])
Sig.intercept<–Cov[1,1]
Sig.slope1<–Cov[2,2]
Sig.slope2<–Cov[3,3]
rho.intercept.slope1<–Cov[1,2]/sqrt(Cov[1,1]*Cov[2,2]) # Correlations for fixed effects
rho.intercept.slope2<–Cov[1,3]/sqrt(Cov[1,1]*Cov[3,3])
rho.slope1.slope2<–Cov[2,3]/sqrt(Cov[2,2]*Cov[3,3])
#Data list(N=110, J=39,R1=structure(.Data=c(1,0,0,0,1,0,0,0,1),.Dim=c(3,3)),
Y=c(23.35,12.3,15.76,...37.43), # Outcome data vector of size N
school=c(1,1,2,2,...38,39,39), # Group-level (schools) data vector of size N
x=structure(.Data=c(3.1,...3.0,3.2), .Dim = c(110, 2))) # (N x 2) matrix of predictors
```

Appendix D

Confirmatory Factor Analysis WinBUGS Code

```
model
for(i in 1:N
#Measurement Equation Model
for(j in 1:P)
y[i,j]~dnorm(mu[i,j],psi[j])
ephat[i,j]<–y[i,j]-mu[i,j]
mu[i,1]<–xi[i,1]+delta[1] # Factor 1
mu[i,2]<–lam[1]*xi[i,1]+delta[2]
mu[i,3]<–lam[2]*xi[i,1]+delta[3]
mu[i,4]<–lam[3]*xi[i,1]+delta[4]
mu[i,5]<–lam[4]*xi[i,1]+delta[5]
mu[i,6]<–lam[5]*xi[i,1]+delta[6]
mu[i,7]<–lam[6]*xi[i,1]+delta[7]
mu[i,8]<–lam[7]*xi[i,1]+delta[8]
mu[i,9]<–xi[i,2]+delta[9] # Factor 2
mu[i,10]<–lam[8]*xi[i,2]+delta[10]
mu[i,11]<–lam[9]*xi[i,2]+delta[11]
mu[i,12]<–lam[10]*xi[i,2]+delta[12]
mu[i,13]<–lam[11]*xi[i,2]+delta[13]
mu[i,14]<–lam[12]*xi[i,2]+delta[14]
mu[i,15]<–lam[13]*xi[i,2]+delta[15]
#Structural Equation Model
xi[i,1:2]~dmnorm(u[1:2],phi[1:2,1:2])
#Priors on Intercepts
for(j in 1:P)delta[j]~dnorm(0.0, 1.0)
#Priors on Loadings
lam[1]~dnorm(0,psi[2])
lam[2]~dnorm(0,psi[3])
lam[3]~dnorm(0,psi[4])
lam[4]~dnorm(0,psi[5])
lam[5]~dnorm(0,psi[6])
lam[6]~dnorm(0,psi[7])
lam[7]~dnorm(0,psi[8])
lam[8]~dnorm(0,psi[10])
lam[9]~dnorm(0,psi[11])
lam[10]~dnorm(0,psi[12])
lam[11]~dnorm(0,psi[13])
lam[12]~dnorm(0,psi[14])
lam[13]~dnorm(0,psi[15])
```

Appendix D cont'd

Confirmatory Factor Analysis WinBUGS Code, cont'd

#Priors on Precisions
for(j in 1:P)
psi[j]∼dgamma(9.0, 4.0) # Error variances
sgm[j]<–1/psi[j]
psd dgamma(9.0, 4.0) # Residual variance
sgd<–1/psd # Residual precision
phi[1:2,1:2]∼dwish(R[1:2,1:2], 5) # Precision matrix

phx[1:2,1:2]<–inverse(phi[1:2,1:2]) # Variance/Covariance matrix
#Data
list(N=3500, P=15, u=c(0,0),y=structure(.Data=c(1, 3,...2, 4),.Dim=c(3500,15)),
R=structure(.Data=c(1,0,0,1),.Dim=c(2,2)))

References

Bauwens, L., Lubrano, M., & Richard, J.-F. (2003). *Bayesian inference in dynamic econometric models*. Oxford: Oxford University Press.

Best, N., Cowles, M. K., & Vines, K. (1996). CODA Convergence diagnosis and output analysis software for Gibbs sampling output-Version 0.30 [Computer software manual].

Bolstad, W. M. (2009). *Understanding computational Bayesian methods*. New York: Wiley.

Box, G., & Tiao, G. (1973). *Bayesian inference in statistical analysis*. New York: Addison-Wesley.

Casella, G., & Robert, C. (2003). *Monte Carlo statistical methods*. New York: Springer.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*, 997–1003.

de Finetti, B. (1974). *Theory of probability, vols. 1 and 2*. New York: John Wiley and Sons.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, *39*, 1-38.

Everitt, B. (2002). *Cambridge dictionary of statistics* (2nd ed.). Cambridge: Cambridge University Press.

Fisher, R. A. (1941/1925). *Statistical methods for research workers* (84th ed.). Edinburgh: Oliver & Boyd.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian data analysis, 2nd edition*. London: Chapman and Hall.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. Pattn. Anal. Mach. Intel.*, *6*, 721–741.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian statistics 4*. Oxford: Oxford University Press.

Gigerenzer, G., Krauss, & Vitouch, O. (2004). The null ritual: What you always wanted to know about significance testing but were afraid to ask. In

D. Kaplan (Ed.), *The Sage handbook of quantitative methodology for the social sciences* (pp. 391–408). Thousand Oaks, CA: Sage Publications.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1996). *Markov chain Monte Carlo in practice*. London: Chapman and Hall.

Gill, J. (2002). *Bayesian methods: A social and behavioral sciences approach*. London: Chapman and Hall/CRC.

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (1997). *What if there were no significance tests?* Mahwah, NJ: Lawrence Erlbaum and Associates.

Heidelberger, P., & Welch, P. (1983). Simulation run length control in the presence of an initial transient. *Operations Research*, *31*, 1109–1144.

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, *14*, 382–417.

Hoff, P. D. (2009). *A first course in Bayesian statistical methods*. New York: Springer.

Jackman, S. (2009). *Bayesian analysis for the social sciences*. New York: John Wiley.

Jeffreys, H. (1961). *Theory of probability* (third ed.). New York: Oxford University Press.

Kaplan, D. (2009). *Structural equation modeling: Foundations and extensions.* (2nd ed.). Newbury Park, CA: Sage Publications.

Kaplan, D., & Wenger, R. N. (1993). Asymptotic independence and separability in covariance structure models: Implications for specification error, power, and model modification. *Multivariate Behavioral Research*, *28*, 483-498.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795.

Kolmogorov, A. N. (1956). *Foundations of the theory of probability* (2nd ed.). New York: Chelsea.

Leamer, E. E. (1983). Model choice and specification analysis. In Z. Griliches & M. Intriligator (Eds.), *Handbook of econometrics, volume 1.* Amsterdam: North Holland.

Lee, S.-Y. (2007). *Structural equation modeling: A Bayesian approach*. New York: Wiley.

Lunn, D., Thomas, A., Best, N., & Spiegelhalter, D. (2000). Winbugs – a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and*

Computing, *10*, 325–337.

Lynch, S. M. (2007). *Introduction to applied Bayesian statistics and estimation for social scientists*. New York: Springer.

Madigan, D., & Raftery, A. E. (1994). Model selection and accounting for model uncertainly in graphical models using Occam's window. *Journal of the American Statistical Association*, *89*, 1535–1546.

Martin, A. D., Quinn, K. M., & Park, J. H. (2010, May 10). *Markov chain Monte Carlo (MCMC) package.* http://mcmcpack.wustl.edu/.

Mengersen, K. L., Robery, C. P., & Guihenneuc-Jouyax, C. (1999). MCMC convergence diagnostics: A review. *Bayesian Statistics*, *6*, 415–440.

Muthén, L. K., & Muthén, B. (2010). *Mplus: Statistical analysis with latent variables*. Los Angeles: Muthén & Muthén.

NCES. (1988). *National educational longitudinal study of 1988*. Washington DC: U.S. Department of Education.

NCES. (2001). *Early childhood longitudinal study: Kindergarten class of 1998-99: Base year public-use data files user's manual* (Tech. Rep. No. NCES 2001-029). U.S. Government Printing Office.

Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika*, *29A, Part I*, 175–240.

Press, S. J. (2003). *Subjective and objective Bayesian statistics: Principles, models, and applications* (2nd ed.). New York: Wiley.

R Development Core Team. (2008a). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Available from `http://www.R-project.org` (ISBN 3-900051-07-0)

R Development Core Team. (2008b). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Available from `http://www.R-project.org` (ISBN 3-900051-07-0)

Raftery, A. E. (1995). Bayesian model selection in social research (with discussion). In P. V. Marsden (Ed.), *Sociological methodology* (Vol. 25, pp. 111–196). New York: Blackwell.

Raftery, A. E., Hoeting, J., Volinsky, C., Painter, I., & Yeung, K. Y. (2009, September 18). *Bayesian model averaging (BMA), version 3.12.* http://www2.research.att.com/ volinsky/bma.html.

Raftery, A. E., & Lewis, S. M. (1992). How many iterations in the Gibbs sampler? In J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian statistics 4* (p. 763-773). Oxford: Oxford University Press.

Raftery, A. E., & Lewis, S. M. (1996). Implementing MCMC. In W. R. Gilks, S. Richardson, & D. J. Spiegelhalter (Eds.), *Markov chain Monte Carlo in practice* (pp. 115–130). New York: Chapman & Hall.

Ramsey, F. P. (1926). Truth and probability. In *The foundations of mathematics and other logical essays.* New York: Humanities Press.

Renyi, A. (1970). *Probability theory.* New York: Elsevier.

Savage, L. J. (1954). *The foundations of statistics.* New York: John Wiley and Sons, Inc.

Sinharay, S. (2004). Experiences with Markov chain Monte Carlo convergence assessment in two psychometric examples. *Journal of Educational and Behavioral Statistics*, *29*, 461–488.

Smith, B. J. (2005, March 23). *Bayesian Output Analysis program (BOA), version 1.1.5.* http://www.public-health.uiowa.edu/boa.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Linde, A. van der. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, *64*, 583–639.

Volinsky, C. T., Madigan, D., Raftery, A. E., & Kronmal, R. A. (1997). Bayesian model averaging in proportional hazzard models: Assessing the risk of a stroke. *Journal of the Royal Statistical Society. Section. C*, *46*, 433-448.

Author Note

The research reported in this paper was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305D110001 to The University of Wisconsin - Madison. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Footnotes

[1]The symbol, ¬, implies "not"

[2]Technically, according to de Finetti (1974), this refers to *finite* exchangeability. Infinite exchangeability is obtained by adding the provision that every finite subset of an infinite sequence is exchangeable.

[3]Press (2003) points out the interesting fact that the uniform prior (a vague prior) was actually used by Bayes' in his investigations.

[4]The scale parameter affects spread of the distribution, in the sense of shrinking or stretching the distribution. The shape parameter, as the term implies, affects the shape of the distribution (Everitt, 2002).

[5]As an aside, the notion of an infinitely large number of repeated samples is no more a conceptual leap than the notion of subjective probability.

[6]The BMA package uses the "leaps and bounds" algorithm to reduce the model space (see e.g. Volinsky et al., 1997, for more details).

Table 1
Bayesian Regression Estimates from R: ECLS–K Database

Node	EAP	SD	MC Error	95% Credibility Interval
Full Model				
Intercept	-4.00	2.79	2.75E-2	-9.46, 1.57
Teacher1: Approaches to Learning	3.81	0.59	5.99E-3	2.65, 4.98
Teacher2: Self-Control	0.41	0.97	8.39E-3	-1.47, 2.32
Teacher3: Interpersonal Skills	0.33	0.95	9.22E-3	-1.57, 2.18
Parent1: Approaches to Learning	2.15	0.77	7.08E-3	0.63, 3.66
Parent2: Self-Control	2.00	0.62	5.37E-3	0.78, 3.23
Parent3: Social Interaction	0.20	0.67	6.57E-3	-1.14, 1.51
Math Achievement Variance	58.52	3.54	3.64E-2	51.92, 65.79
Restricted Model: Teacher-Related Items				
Intercept	5.87	1.76	1.85E-2	2.49, 9.42
Teacher1: Approaches to Learning	4.38	0.59	5.06E-3	3.21, 5.53
Teacher2: Self-Control	0.16	0.97	7.823E-3	-1.77, 2.03
Teacher3: Interpersonal Skills	1.04	0.95	8.14E-3	-0.82, 2.93
Math Achievement Variance	60.90	3.70	3.57E-2	54.03, 68.57
Restricted Model: Parent-Related Items				
Intercept	1.65	2.75	2.89E-2	-3.64, 7.18
Parent1: Approaches to Learning	3.37	0.81	6.80E-3	1.76, 4.93
Parent2: Self-Control	2.94	0.64	5.57E-3	1.65, 4.17
Parent3: Social Interaction	0.62	0.71	7.37E-3	-0.77, 2.01
Math Achievement Variance	65.95	4.01	3.86E-2	58.52, 74.26

Note. Note that these are all unstandardized weights. However, standardized weights are also available through this program. EAP = Expected A Posteriori. SD = Standard Deviation. MC error = Monte Carlo error.

Table 2
Bayesian Model Averaging Results for Five Multiple Regression Models

Node	Post Prob	Avg Coef	SD	Model 1	Model 2	Model 3	Model 4	Model 5
Full Model								
Intercept	1.00	-2.68	3.00	-3.14	2.28	-4.02	-3.80	0.82
Teacher1	1.00	4.19	0.53	4.19	4.56	3.87	3.89	4.48
Teacher2	0.06	0.04	0.22	.	.	0.67	.	.
Teacher3	0.06	0.04	0.21	.	.	.	0.65	.
Parent1	0.93	2.21	0.90	2.35	.	2.33	2.28	2.71
Parent2	0.95	2.02	0.76	2.11	2.41	2.06	2.04	.
Parent3	0.00	0.00	0.00	.	.	.	.	.
R^2				0.20	0.18	0.20	0.20	0.18
BIC				-104.39	-99.47	-99.29	-99.25	-98.91
PMP				0.77	0.07	0.06	0.06	0.05

Note. Post Prob = the posterior probability for each variable in the averaged model, Avg Coef = the average unstandardized coefficient for all variables in the model, SD = the standard deviation for the averaged coefficients, R^2 = percent of variance accounted for by each model, BIC = Bayesian information criteria, and PMP = posterior model probability for each of the five models.

Table 3
WinBugs HLM Estimates: ECLSK Data

Node	EAP	SD	MC Error	95% Credibility Interval
Fixed Effects				
Intercept	-2.61	0.78	3.43E-2	-4.12, -1.10
Approaches to Learning	4.85	0.40	1.72E-2	4.10, 5.63
Self Control	2.66	0.40	1.71E-2	1.88, 3.53
Fixed Effects: Correlations				
Intercept/Learning	0.23	0.15	1.63E-3	-0.07, 0.51
Intercept/Self Control	0.22	0.15	1.68E-3	-0.07, 0.51
Learning/Self Control	0.47	0.15	2.39E-3	0.17, 0.72

Note. EAP = Expected A Posteriori. SD = Standard Deviation. MC error = Monte Carlo error.

Table 4
WinBugs CFA Estimates: NELS:88 Survey

Node	EAP	SD	MC Error	95% Credibility Interval
Loadings: Positive				
Students get along	1.00			
There is school spirit	0.99	0.03	7.05E-4	0.95, 1.03
Discipline is fair	0.99	0.02	7.02E-4	0.95, 1.03
I have friends of other racial groups	0.94	0.02	7.17E-4	0.90, 0.98
Teaching is good	1.08	0.02	7.43E-4	1.04, 1.12
Teachers are interested in students	1.11	0.02	7.40E-4	1.07, 1.15
Teachers praise students	1.02	0.02	7.50E-4	0.98, 1.06
Teachers listen to students	1.04	0.02	7.53E-4	1.01, 1.08
Loadings: Negative				
Students disrupt learning	1.00			
Teachers putdown students	0.84	0.02	8.94E-4	0.80, 0.89
Teachers are strict	0.86	0.02	9.38E-4	0.81, 0.91
Students putdown each other	0.87	0.02	9.91E-4	0.82, 0.92
School is not safe	0.80	0.02	8.79E-4	0.75, 0.84
Disruptions impede my learning	0.93	0.02	9.33E-4	0.89, 0.98
Students get away with bad behavior	0.97	0.02	9.99E-4	0.92, 1.02
Factor Precisions				
Factor 1 Precision	0.59	0.02	8.22E-4	0.55, 0.63
Factor 2 Precision	0.61	0.03	1.22E-3	0.56, 0.66
Factor Covariance Precision	0.43	0.02	5.48E-4	0.40, 0.47
Error Variances				
Students get along	3.66	0.11	2.33E-3	3.45, 3.87
There is school spirit	1.81	0.05	7.36E-4	1.72, 1.90
Discipline is fair	1.61	0.04	8.25E-4	1.52, 1.69
I have friends of other racial groups	1.60	0.04	6.58E-4	1.52, 1.68
Teaching is good	2.58	0.07	1.29E-3	2.44, 2.72
Teachers are interested in students	2.10	0.06	1.09E-3	1.99, 2.22
Teachers praise students	1.99	0.05	1.02E-3	1.88, 2.09
Teachers listen to students	2.35	0.07	1.28E-3	2.23, 2.48
Students disrupt learning	1.86	0.05	1.23E-3	1.76, 1.97
Teachers putdown students	2.02	0.06	1.11E-3	1.91, 2.14
Teachers are strict	1.37	0.04	6.55E-4	1.30, 1.44
Students putdown each other	1.92	0.05	1.19E-3	1.82, 2.03
School is not safe	1.92	0.05	9.15E-4	1.83, 2.03
Disruptions impede my learning	1.56	0.04	7.61E-4	1.48, 1.64
Students get away with bad behavior	1.61	0.04	9.30E-4	1.53, 1.70
Residual Variance and Precision				
Variance	2.24	0.76	9.42E-3	1.01, 3.96
Precision	0.51	0.20	2.47E-3	0.25, 1.00

Note. Note that these are unstandardized factor loadings. However, the program can be specified to produce standardized loadings. EAP = Expected A Posteriori. SD = Standard Deviation. MC error = Monte Carlo error.

Figure Captions

Figure 1. Bayesian Regression: Convergence and Posterior Plots for all Regression Model Predictors.

Figure 2. HLM: Convergence, Posterior Densities, and Auto-Correlations for Fixed Effects.

Figure 3. CFA: Convergence, Posterior Densities, and Auto-Correlations for Select Parameters.

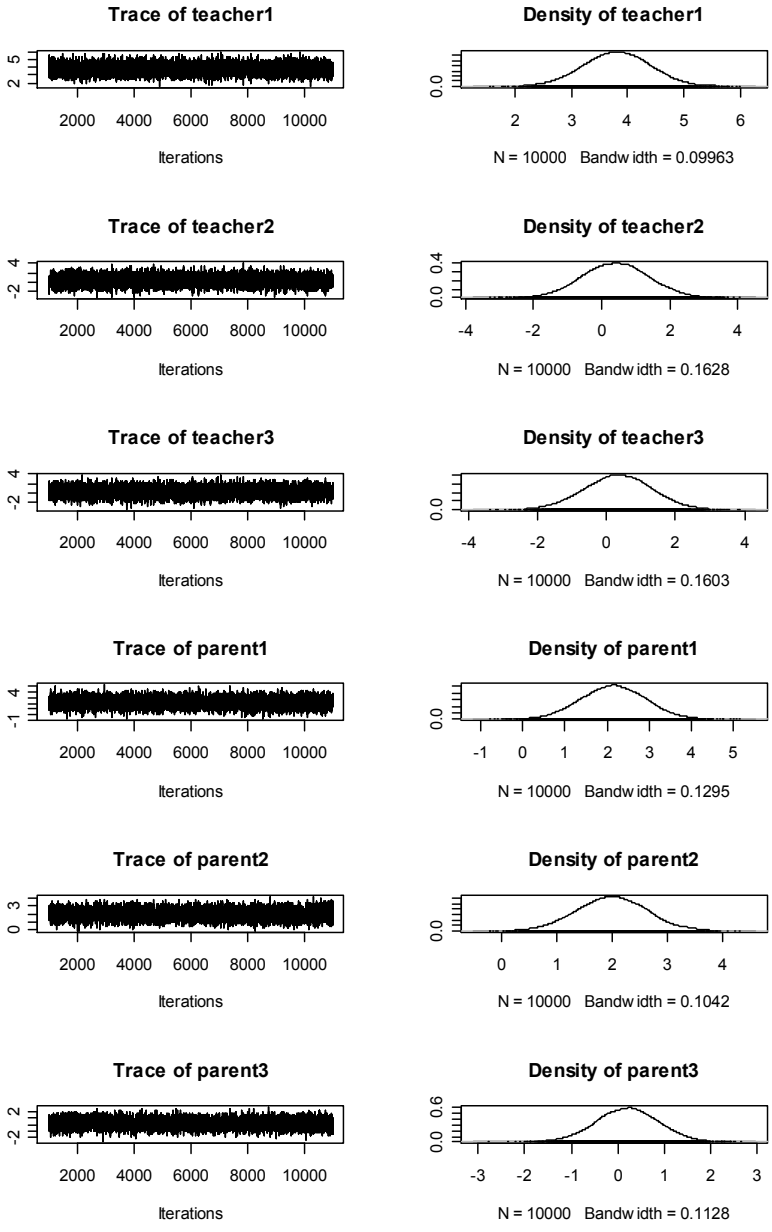

Figure 1. Bayesian Regression: Convergence and Posterior Plots for all Regression Model Predictors.

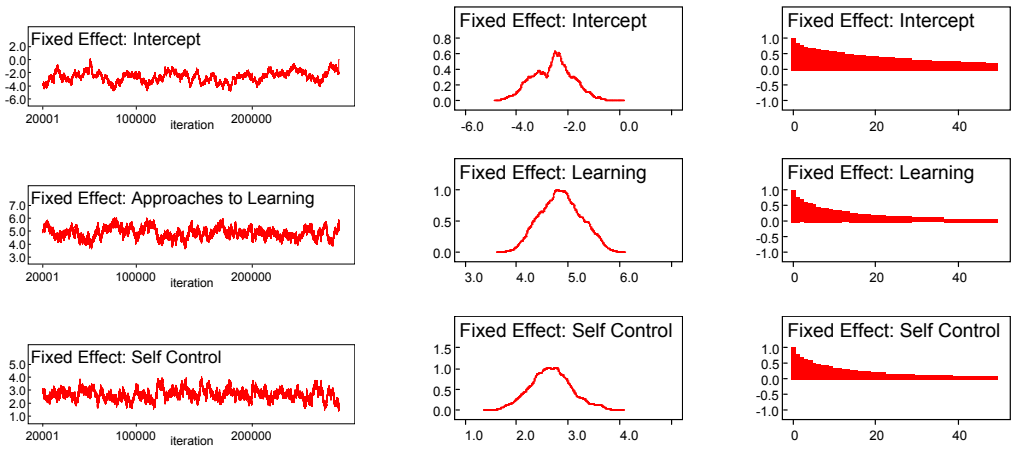

Figure 2. HLM: Convergence, Posterior Densities, and Auto-Correlations for Fixed Effects.

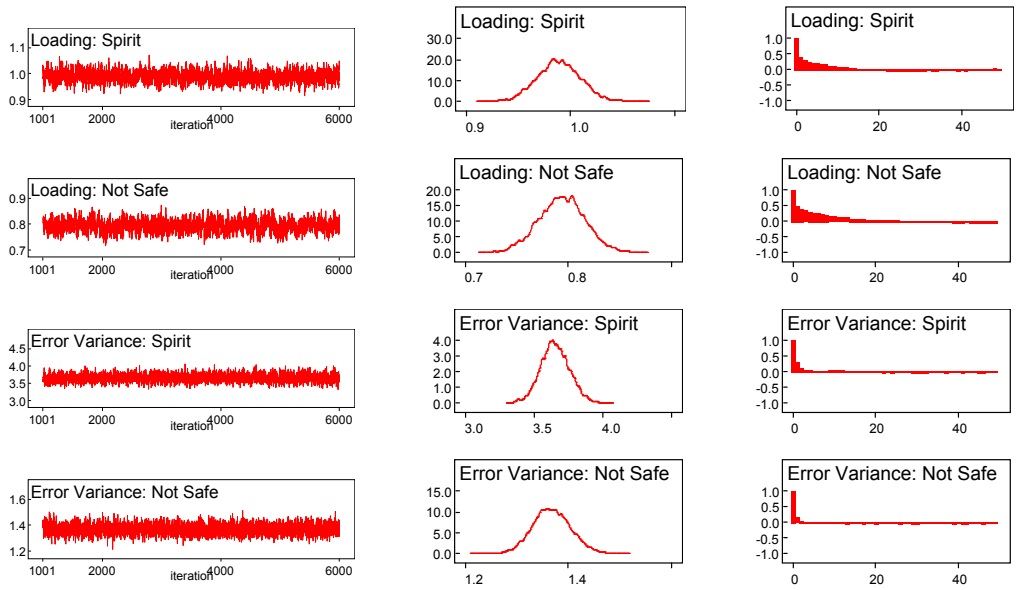

Figure 3. CFA: Convergence, Posterior Densities, and Auto-Correlations for Select Parameters.