Workshop on Bayesian Statistics for Education Research

David Kaplan
Department of Educational Psychology

19 – 21 January, 2015
CEMO, Oslo, Norway
The material for this workshop is drawn from Kaplan, D (2014), *Bayesian Statistics for the Social Sciences*. New York, Guilford Press.

©David Kaplan, 2015
Day 1 AM: Introduction to Workshop and Bayesian Theory

It is clear that it is not possible to think about learning from experience and acting on it without coming to terms with Bayes' theorem. - Jerome Cornfield
It is clear that it is not possible to think about learning from experience and acting on it without coming to terms with Bayes’ theorem. - Jerome Cornfield
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.
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.

Until recently, it was not feasible to conduct statistical modeling from a Bayesian perspective owing to its complexity and lack of availability.
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.

Until recently, it was not feasible to conduct statistical modeling from a Bayesian perspective owing to its complexity and lack of availability.

Bayesian statistics represents a powerful alternative to frequentist (classical) statistics, and is therefore, controversial.
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.

Until recently, it was not feasible to conduct statistical modeling from a Bayesian perspective owing to its complexity and lack of availability.

Bayesian statistics represents a powerful alternative to frequentist (classical) statistics, and is therefore, controversial.

Bayesian statistical methods are now more popular owing to the development of powerful statistical software tools that make the estimation of complex models feasible from a Bayesian perspective.
Day 1

- AM: Major differences between the Bayesian and frequentist paradigms of statistics; Bayes’ theorem; Bayesian hypothesis testing
- PM: MCMC; Bayesian computation with R and rjags

Day 2

- AM: Bayesian model building; evaluation; Bayesian linear regression
- PM: Student analyses – Bayesian regression analysis

Day 3

- AM: Advanced Topics: HLM, Factor analysis (time permitting)
- PM: Wrap-up: Student analyses
Paradigm Differences

- For frequentists, the basic idea is that probability is represented by the model of **long run frequency**.

- Frequentist probability underlies the Fisher and Neyman-Pearson schools of statistics – the conventional methods of statistics we most often use.

- The frequentist formulation rests on the idea of equally probable and stochastically independent events

- The physical representation is the coin toss, which relates to the idea of a very large (actually infinite) number of repeated experiments.
The entire structure of Neyman - Pearson hypothesis testing and Fisherian statistics (together referred to as the **frequentist school**) is based on frequentist probability.

Our conclusions regarding null and alternative hypotheses presuppose the idea that we could conduct the same experiment an infinite number of times.

Our interpretation of confidence intervals also assumes a fixed parameter and CIs that vary over an infinitely large number of identical experiments.
But there is another view of probability as **subjective belief**.

The physical model in this case is that of the “bet”.

Consider the situation of betting on who will win the the World Cup (or the World Series).

Here, probability is not based on an infinite number of repeatable and stochastically independent events, but rather on how much knowledge you have and how much you are willing to bet.

Subjective probability allows one to address questions such as “what is the probability that my team will win the World Cup?” Relative frequency supplies information, but it is not the same as probability and can be quite different.

This notion of subjective probability underlies Bayesian statistics.
Consider the joint probability of two events, $Y$ and $X$, for example observing lung cancer and smoking jointly.

The joint probability can be written as
Consider the joint probability of two events, $Y$ and $X$, for example observing lung cancer and smoking jointly.

The joint probability can be written as

$$p(cancer, smoking) = p(cancer|smoking)p(smoking) \quad (1)$$
Consider the joint probability of two events, $Y$ and $X$, for example observing lung cancer and smoking jointly.

The joint probability can be written as

$$p(cancer, smoking) = p(cancer|smoking)p(smoking) \quad (1)$$

Similarly
Consider the joint probability of two events, $Y$ and $X$, for example observing lung cancer and smoking jointly.

The joint probability can be written as

$$p(cancer, smoking) = p(cancer|smoking)p(smoking) \quad (1)$$

Similarly

$$p(smoking, cancer) = p(smoking|cancer)p(cancer) \quad (2)$$
Because these are symmetric, we can set them equal to each other to obtain the following
Because these are symmetric, we can set them equal to each other to obtain the following

\[ p(cancer|smoking)p(smoking) = p(smoking|cancer)p(cancer) \]  

(3)
Because these are symmetric, we can set them equal to each other to obtain the following

\[ p(\text{cancer} | \text{smoking})p(\text{smoking}) = p(\text{smoking} | \text{cancer})p(\text{cancer}) \quad (3) \]

\[ p(\text{cancer} | \text{smoking}) = \frac{p(\text{smoking} | \text{cancer})p(\text{cancer})}{p(\text{smoking})} \quad (4) \]
Because these are symmetric, we can set them equal to each other to obtain the following

$$p(\text{cancer}|\text{smoking})p(\text{smoking}) = p(\text{smoking}|\text{cancer})p(\text{cancer})$$  \hspace{1cm} (3)

The inverse probability theorem (Bayes’ theorem) states

$$p(\text{cancer}|\text{smoking}) = \frac{p(\text{smoking}|\text{cancer})p(\text{cancer})}{p(\text{smoking})}$$  \hspace{1cm} (4)
Because these are symmetric, we can set them equal to each other to obtain the following

\[ p(\text{cancer} | \text{smoking})p(\text{smoking}) = p(\text{smoking} | \text{cancer})p(\text{cancer}) \]  

(3)

\[ p(\text{cancer} | \text{smoking}) = \frac{p(\text{smoking} | \text{cancer})p(\text{cancer})}{p(\text{smoking})} \]  

(4)

The inverse probability theorem (Bayes’ theorem) states

\[ p(\text{smoking} | \text{cancer}) = \frac{p(\text{cancer} | \text{smoking})p(\text{smoking})}{p(\text{cancer})} \]  

(5)
Why do we care?

Because this is how we go from the probability of a patient having cancer given that the patient smokes, to the probability that the patient smokes given that he/she has cancer.

We simply need the marginal probability of smoking and the marginal probability of cancer ("base rates" or what we will call prior probabilities).
What is the role of Bayes’ theorem for statistical inference?

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 $y$.

Because $Y$ is unobserved and random, we need to specify a probability model to explain how we obtained the actual data values $y$. 

**Statistical Elements of Bayes’ Theorem**
Next, let $\theta$ be a parameter that we believe characterizes the probability model of interest.

The parameter $\theta$ can be a scalar, such as the mean or the variance of a distribution, or it can be vector-valued, such as a set of regression coefficients in regression analysis or factor loadings in factor analysis.

We are concerned with determining the probability of observing $y$ given the unknown parameters $\theta$, which we write as $p(y|\theta)$.

In statistical inference, the goal is to obtain estimates of the unknown parameters given the data.

This is expressed as the likelihood of the parameters given the data, often denoted as $L(\theta|y)$. 
A key difference between Bayesian statistical inference and frequentist statistical inference concerns the nature of the unknown parameters $\theta$.

In the frequentist tradition, the assumption is that $\theta$ is unknown and has a fixed value that we wish to estimate. Our uncertainty about $\theta$ is not taken into account in the frequentist tradition.

In Bayesian statistical inference, $\theta$ is also considered unknown but specify a probability distribution that reflects our uncertainty about the true value of $\theta$.

Because both the observed data $y$ and the parameters $\theta$ are assumed random, we can model the joint probability of the parameters and the data as a function of the conditional distribution of the data given the parameters, and the prior distribution of the parameters.
More formally,
More formally,

$$p(\theta, y) = p(y|\theta)p(\theta).$$  \hspace{1cm} (6)

where $p(\theta, y)$ is the joint distribution of the parameters and the data. Following Bayes’ theorem described earlier, we obtain
More formally,

\[ p(\theta, y) = p(y|\theta)p(\theta). \]  

where \( p(\theta, y) \) is the joint distribution of the parameters and the data. Following Bayes’ theorem described earlier, we obtain

\[ p(\theta|y) = \frac{p(\theta, y)}{p(y)} = \frac{p(y|\theta)p(\theta)}{p(y)}, \]

where \( p(\theta|y) \) is referred to as the posterior distribution of the parameters \( \theta \) given the observed data \( y \).
From equation (7) the posterior distribution of $\theta$ 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
From equation (7) the posterior distribution of $\theta$ give $y$ 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

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

(8)
From equation (7) the posterior distribution of $\theta$ 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

$$p(y) = \sum_{\theta} p(y|\theta)p(\theta), \quad (8)$$

For continuous variables
From equation (7) the posterior distribution of $\theta$ give $y$ 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

$$p(y) = \sum_{\theta} p(y|\theta)p(\theta), \quad (8)$$

For continuous variables

$$p(y) = \int_{\theta} p(y|\theta)p(\theta)d\theta, \quad (9)$$
Notice that \( p(y) \) does not involve model parameters, so we can omit the term and obtain the *unnormalized posterior distribution*.
Notice that $p(y)$ does not involve model parameters, so we can omit the term and obtain the *unnormalized posterior distribution*

$$p(\theta|y) \propto p(y|\theta)p(\theta).$$ (10)
Notice that $p(y)$ does not involve model parameters, so we can omit the term and obtain the \textit{unnormalized posterior distribution}

$$p(\theta|y) \propto p(y|\theta)p(\theta).$$

This can also be expressed in terms of the likelihood
Notice that $p(y)$ does not involve model parameters, so we can omit the term and obtain the *unnormalized posterior distribution*

$$p(\theta|y) \propto p(y|\theta)p(\theta). \quad (10)$$

This can also be expressed in terms of the likelihood

$$p(\theta|y) \propto L(\theta|y)p(\theta). \quad (11)$$
Equations (10) and (11) represents the core of Bayesian statistical inference and is what separates Bayesian statistics from frequentist statistics.

Equation (11) 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 $p(\theta|y)$. 
Exchangeability

- It is common in statistics to invoke the assumption that the data \( y_1, y_2, \ldots y_n \) are independently and identically distributed – often referred to as the \( i.i.d \) assumption.

- Bayesians invoke the deeper notion of exchangeability to produce likelihoods and address the issue of independence.

- Exchangeability arises from de Finetti’s Representation Theorem 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, \( p(y_1, y_2, \ldots y_n) \) is invariant to permutations of the subscripts.
Consider the response that student $i$ ($i = 1, 2, \ldots, 10$) would make to a PIRLS question “I like what I read in school“, where

$$y_i = \begin{cases} 
1, & \text{if student } i \text{ agrees} \\
0, & \text{if student } i \text{ disagrees}
\end{cases}$$

(12)

Next, consider three patterns of responses by 10 randomly selected students

$$p(1, 0, 1, 1, 0, 1, 0, 1, 0, 0)$$  \hspace{1cm} (13a)

$$p(1, 1, 0, 0, 1, 1, 1, 0, 0, 0)$$  \hspace{1cm} (13b)

$$p(1, 0, 0, 0, 0, 1, 1, 1, 1, 1)$$  \hspace{1cm} (13c)

Note that there are actually $2^{10}$ possible response patterns.
• If our task were to assign probabilities to all possible outcomes, this could become prohibitively difficult.

• However, suppose we now assume that student responses are independent of one another.

• Exchangeability implies that only the proportion of agreements matter, not the location of those agreements in the vector.

• Given that the sequences are the same length $n$, we can exchange the response of student $i$ for student $j$, the without changing our belief about the probability model that generated that sequence.
Exchangeability means that we believe that there is a parameter $\theta$ that generates the observed data via a stochastic model and that we can describe that parameter without reference to the particular data at hand.

The fact that we can describe $\theta$ without reference to a particular set of data is, in fact, what is implied by the idea of a prior distribution.
Why do we specify a prior distribution on the parameters?
Why do we specify a prior distribution on the parameters?

Our view is that progress in science generally comes about by learning from previous research findings and incorporating information from these findings into our present studies.
Why do we specify a prior distribution on the parameters?

Our view is that progress in science generally comes about by learning from previous research findings and incorporating information from these findings into our present studies.

Information from previous research is almost always incorporated into our choice of designs, variables to be measured, or conceptual diagrams to be drawn.
Why do we specify a prior distribution on the parameters?

Our view is that progress in science generally comes about by learning from previous research findings and incorporating information from these findings into our present studies.

Information from previous research is almost always incorporated into our choice of designs, variables to be measured, or conceptual diagrams to be drawn.

Bayesian statistical inference requires that our prior beliefs be made explicit, but then moderates our prior beliefs by the actual data in hand.
Why do we specify a prior distribution on the parameters?

- Our view is that progress in science generally comes about by learning from previous research findings and incorporating information from these findings into our present studies.

- Information from previous research is almost always incorporated into our choice of designs, variables to be measured, or conceptual diagrams to be drawn.

- Bayesian statistical inference 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 equations (10) and (11).
But how do we choose a prior?
But how do we choose a prior?

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.
But how do we choose a prior?

- 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.

- Leamer (1983), who orders priors on the basis of degree of confidence.
But how do we choose a prior?

- 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.

- 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).
But how do we choose a prior?

- 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.

- 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).

- The strength of Bayesian inference lies precisely in its ability to incorporate existing knowledge into statistical specifications.
Non-informative priors

- 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 reference, default, or objective priors.
Perhaps the most common non-informative prior distribution is the uniform distribution $U(\alpha, \beta)$ over some sensible range of values from $\alpha$ to $\beta$.

The uniform distribution indicates that we believe that the value of our parameter of interest lies in range $\beta - \alpha$ and that all values have equal probability.

Care must be taken in the choice of the range of values over the uniform distribution. A $U[-\infty, \infty]$ is an improper prior distribution because it does not integrate to 1.0 as required of probability distributions.
Informative (conjugate) priors

- 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 called *informative, personal, or epistemic*.

- One type of informative prior is based on the notion of a conjugate distribution.

- 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.
## Conjugate Priors for Some Common Distributions

<table>
<thead>
<tr>
<th>Data Distribution</th>
<th>Conjugate Prior</th>
</tr>
</thead>
<tbody>
<tr>
<td>The normal distribution</td>
<td>The normal distribution or Uniform Distribution</td>
</tr>
<tr>
<td>The Poisson distribution</td>
<td>The gamma distribution</td>
</tr>
<tr>
<td>The binomial distribution</td>
<td>The Beta Distribution</td>
</tr>
<tr>
<td>The multinomial distribution</td>
<td>The Dirichlet Distribution</td>
</tr>
</tbody>
</table>
Figure 1: Normal distribution, mean unknown/variance known with varying conjugate priors
Figure 2: Poisson distribution with varying gamma-density priors
A critically important component of applied statistics is hypothesis testing.

The approach most widely used in the social and behavioral sciences is the Neyman-Pearson approach.

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.

Gigerenzer (2004) argued that much of the difficulty in grasping frequentist hypothesis testing lies in the conflation of Fisherian hypothesis testing and the Neyman-Pearson approach to hypothesis testing.
Fisher’s early approach to hypothesis testing required specifying 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.

Fisher’s approach was based on looking at how data could inform evidence for a hypothesis.
The Neyman and Pearson approach requires that two hypotheses be specified – the null and alternative hypothesis – and is designed to inform specific sets of actions. It’s about making a choice, not about evidence against a hypothesis.

By specifying two hypotheses, one can compute a desired tradeoff between two types of errors: Type I errors ($\alpha$) and Type II errors ($\beta$)

The goal is not to assess the evidence against a hypothesis (or model) taken as true. Rather, it is whether the data provide support for taking one of two competing sets of actions.

In fact, prior belief as well as interest in “the truth” of the hypothesis is irrelevant – only a decision that has a low probability of being the wrong decision is relevant.
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.

In the Neyman-Pearson paradigm, $\alpha$ and $\beta$ are determined prior to the experiment being conducted and refer to a consideration of the cost of making one or the other error.
In the Neyman-Pearson approach, the problem is one of finding a balance between $\alpha$, power, and sample size.

The Neyman-Pearson approach is best suited for experimental planning. Sometimes, it is used this way, followed by the Fisherian approach for judging evidence. But, these two ideas may be incompatible (Royall, 1997).

However, this balance is virtually always ignored and $\alpha = 0.05$ is used.

The point is that the $p$-value and $\alpha$ 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 from after having conducted the analysis (e.g., .001, .01, .05).

This can lead a researcher to set \( \alpha \) ahead of time, as per the Neyman-Pearson school, but then communicate a different level of “significance” after running the test.

The conventional practice is even worse than described, such as when describing non-significant results as
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 from after having conducted the analysis (e.g., .001, .01, .05).

This can lead a researcher to set \( \alpha \) ahead of time, as per the Neyman-Pearson school, but then communicate a different level of “significance” after running the test.

The conventional practice is even worse than described, such as when describing non-significant results as “trending toward significance”
Misunderstanding the Fisher or Neyman-Pearson framework to hypothesis testing and/or poor methodological practice is not a criticism of the approach 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 perplexing insofar as frequentist hypothesis testing also rests on assumptions that do not involve data.

The simplest and most ubiquitous example is the choice of variables in a regression equation or asymptotic distributions of test statistics.
Hypothesis testing begins first by obtaining summaries of relevant distributions.

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.
For the continuous case, the mean of the posterior distribution of $\theta$ given the data $y$ is referred to as the *expected a posteriori* or EAP estimate and can be written as
For the continuous case, the mean of the posterior distribution of $\theta$ given the data $y$ is referred to as the *expected a posteriori* or EAP estimate and can be written as

$$E(\theta|y) = \int_{-\infty}^{+\infty} \theta p(\theta|y) d\theta. \quad (14)$$
For the continuous case, the mean of the posterior distribution of $\theta$ given the data $y$ is referred to as the \textit{expected a posteriori} or EAP estimate and can be written as

\[ E(\theta|y) = \int_{-\infty}^{+\infty} \theta p(\theta|y) d\theta. \]  \hspace{1cm} (14)

Similarly, the variance of posterior distribution of $\theta$ given $y$ can be obtained as
For the continuous case, the mean of the posterior distribution of $\theta$ given the data $y$ is referred to as the expected a posteriori or EAP estimate and can be written as

$$E(\theta|y) = \int_{-\infty}^{+\infty} \theta p(\theta|y) d\theta.$$  \hfill (14)

Similarly, the variance of posterior distribution of $\theta$ given $y$ can be obtained as

$$var(\theta|y) = E[(\theta - E[\theta|y])^2|y),$$

$$= \int_{-\infty}^{+\infty} (\theta - E[\theta|y])^2 p(\theta|y) d\theta.$$ \hfill (15)
Another common summary measure would be the mode of the posterior distribution – referred to as the maximum a posteriori (MAP) estimate.

The MAP begins with the idea of maximum likelihood estimation. The MAP can be written as
Another common summary measure would be the mode of the posterior distribution – referred to as the maximum a posteriori (MAP) estimate.

The MAP begins with the idea of maximum likelihood estimation. The MAP can be written as

$$\hat{\theta}_{MAP} = \arg \max_{\theta} L(\theta|y)p(\theta).$$  \hspace{1cm} (16)

where \( \arg \max_{\theta} \) stands for “maximizing the value of the argument”.
In addition to point summary measures, it may also be desirable to provide interval summaries of the posterior distribution.

Recall that the frequentist confidence interval requires that we imagine an infinite number of repeated samples from the population characterized by $\mu$.

For any given sample, we can obtain the sample mean $\bar{x}$ and then 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 $\mu$ under the null hypothesis. Notice that the probability that the parameter is in the interval is either zero or one.
Posterior Probability Intervals (cont’d)

- In contrast, the Bayesian framework assumes that a parameter has a probability distribution.

- Sampling from the posterior distribution of the model parameters, we can obtain its quantiles. From the quantiles, we can directly obtain the probability that a parameter lies within a particular interval.

- So, a 95% posterior probability interval would mean that the probability that the parameter lies in the interval is 0.95.

- Notice that this is entirely different from the frequentist interpretation, and arguably aligns with common sense.
Day 1 PM: Bayesian Computation with R and rjags

- The key reason for the increased popularity of Bayesian methods in the social and behavioral sciences has been the (re)-discovery of numerical algorithms for estimating the posterior distribution of the model parameters given the data.

- Prior to these developments, it was virtually impossible to analytically derive summary measures of the posterior distribution, particularly for complex models with many parameters.

- Rather than attempting the impossible task of analytically solving for estimates of a complex posterior distribution, we can instead draw samples from \( p(\theta | y) \) and summarize the distribution formed by those samples. This is referred to as Monte Carlo integration.

- The two most popular methods of MCMC are the Gibbs sampler and the Metropolis-Hastings algorithm.
A decision must be made regarding the number of Markov chains to be generated, as well as the number of iterations of the sampler.

Each chain samples from another location of the posterior distribution based on starting values.

With multiple chains, fewer iterations are required, particularly if there is evidence for the chains converging to the same posterior mean for each parameter.

Once the chain has stabilized, the burn-in samples are discarded.

Summaries of the posterior distribution as well as convergence diagnostics are calculated on the post-burn-in iterations.
R is a programming environment for data analysis and graphics. It is also a language similar to C++.

R is part of the GNU (GNU is Not Unix) Project and associated with the Free Software Foundation.

“Free” means freedom not price.

When contributing a package to R, you agree to allow anyone access to the code, to be copied, modified, etc.
• **Downloading R**

  1. Go to http://www.r-project.org/

  2. Click on CRAN and choose mirror site

  3. Click on mirror site and choose OS, e.g. Mac or Windows

  4. For Mac click on latest version; for Windows, click on BASE

• **Finding and Installing Packages**

  1. Go to http://cran.r-project.org/web/views/

  2. Packages are organized topically. For example, click on MCMCpack.
To install a package for Mac

1. At the prompt > type `install.packages("MCMCpack")`. Do this again for the package "coda" and "BMA".

2. Refresh the list. Your package should then show up. Click on button and the package is installed.

To install a package for the PC

1. Go to Packages – > Set Cran Mirror. Choose a mirror site

2. Go back to Packages – > Install Packages. Choose MCMCpack and then again coda.

3. Packages – > Load Packages. Click on these packages and they will then be installed.

4. You can now open the package and see the various programs and syntax.
Writing Programs in R

- Programs can be written directly into the console or in a script file.

- New Script for PC. New Document for Mac. This is to be preferred.

- The file with your script will be save as a .R file.

- If you are writing commands in the console, you may want to save the session when you're finished. Then, all the commands you typed will reappear again.
Reading in Data

You have to have your data saved as a .txt or .dat file. This can be done from most other statistical software packages as well as from Excel. R will read in other data forms as well.

The general format of reading in data is

```
objectname <- read.table(datalocation, header=T)
```

If you’re using .csv files from Excel, then this would read

```
objectname <- read.csv(datalocation, header=T)
```

The “objectname” is a new name given to the datafile that R then can read.

“datalocation” is the location of the data using DOS or Mac file location rules.
Reading in Data (cont’d)

Perhaps the easiest approach is to convert your data set into a .csv file and type the following

```
objectname <- read.csv(file.choose(), header=T)
```

This will open up a window so you can search for the file.

You can also use the package “foreign“ that will allow you to bring many other system files, such as SPSS or SAS.

For most text files, the top row contains the variable names, so `header=T` tells R that’s the case.
Handling Missing Data in R.

- Don’t have any missing data. 😊

- Assuming this is not feasible, the best thing to do is to convert missing data into some kind of absurd value such as 999. This can be done easily in SPSS.

- Once the data are converted to a .dat file (perhaps via Excel), then the goal is to convert the 999 into the code `NA`, which R recognizes.

- Bring in your data set as usual, for example

  ```r
  mydata <- read.csv(file.choose(), header=T)
  ```
Handling Missing Data in R (cont’d)

- To convert all of the 999s to NAs in one step, type the following

  ```r
  mydata[mydata==999] = NA
  ```

- You can check this by typing “mydata” at the command line. That is

  ```r
  > mydata
  ```

- You should have NAs where the 999s were.

- To remove all missing data (coded as NA) (LISTWISE DELETION from an analysis, you would write, for example

  ```r
  mydata <- na.omit(mydata)
  ```

- If missing is blank then R will convert to NA.

- Most programs will use something like listwise deletion. THIS IS NOT PREFERRED!! R has many good missing data programs.
Manipulating Data in R

A quick way to create dummy codes for categorical variable in R is as follows: For a dichotomous variable, say coding 1 = female, 2 = male, to 0 = male

```r
female <- ifelse(sex==2,0,1)
```

For three categories

```r
newrace1 = ifelse(race==1,1,0)
newrace2 = ifelse(race==2,1,0)
```

Categorical variables such as sex and race should be treated as factors. In R this is handled as follows. For example

```r
race = c(1,2,3)
frace=factors(race)
levels(frace) = c("black", "white", "hispanic")
```
Manipulating Data in R (cont’d)

To select a subset of observations in R, we can write the following:

```r
mydata <- read.csv(file.choose(), header=T)
newdata <- subset(mydata, variable > somevalue)
```

To indicate which columns along with rows to keep, you write

```r
newdata <- subset(mydata, variable > somevalue, select=c(v1,v2))
```

For numerous transformations, you could write

```r
newdata <- within(mydata,
v1 < somevalue,
v2 == somevalue,
v3 = log(v4),
rm(v5,v6))
```
Manipulating Data in R (cont’d)

- To transform data, the simplest approach would be to write

  ```r
  junk <- read.csv(file.choose(), header=T)
  newdata <- transform(mydata, log.v1 = log(v1))
  ```

- Combining the columns of two data sets, for example `data1` and `data2`. To combine them, write

  ```r
  data3 <- cbind(data1, data2)
  ```
Saving Output and plots to a file

“sink” will divert all subsequent output to a file. Place “sink” at the beginning of the file where you want output to go. For example:

```r
sink("~/Desktop/test.txt")
d <- rnorm(1000,0,1)
summary(d)
sink()
```

To send a plot to a file, you can type

```r
pdf("~/Desktop/test.pdf")
d <- rnorm(1000, 0, 1)
plot(density(d))
dev.off()
```

This last command prevents the plot from showing up on the screen.
Some Good R Books


Some Interesting Web Sites for R

3. http://www.r-project.org/
4. http://cran.r-project.org/web/views/
Brief Introduction to rjags

For this workshop we will use the R interface of the program “JAGS” (Just Another Gibbs Sampler; Plummer, 2015).

JAGS is a BUGS like program (Bayes Under Gibbs Sampling; Spiegelhalter, Thomas and Best 2000).

BUGS was the first program (within WinBUGS) that made Bayesian analysis possible.
Steps when using rjags:

1. Bring in data and use R to do any manipulations - e.g. handling missing data, etc.

2. Write model in jags and save as a .bug file.

3. Pass the data and model to JAGS using the jags.model command.

4. Use coda.samples and update to run the Gibbs sample.

5. Use R to inspect results and get plots.
## Simple normal distribution with conjugate prior ##
## Install and load packages ##

```r
install.packages("rjags")
require(rjags)
```

# Generate some random data

```r
N <- 10000
y <- rnorm(N, 0, 5)
```

# JAGS Code

modelstring = "
model {
    for (i in 1:N) {
        y[i] ~ dnorm(mu, tau)
    }
    # Priors
    mu ~ dnorm(0, 0.0001);
    tau ~ dgamma(0.001, 0.001);
    # Transformations
    sigma <- 1.0/sqrt(tau);
}
"
```

# End of JAGS code and back to R
writeLines(modelstring, con="model.bug")
variables <- list(y = y, N=N)
parameters <- c("mu","tau","sigma")
simplemod <- jags.model("model.bug", data=variables, n.chains=2,
        n.adapt=1000)

cat("Burning in the MCMC chain ...
")
update(simplemod, n.iter=5000)
cat("Sampling from the final MCMC chain ...
")
codaSamples1 = coda.samples(simplemod, variable.names=parameters,
        n.iter=100000, thin=10, seed=5555)
## Use coda to obtain diagnostic plots ##

```r
par(mar=c(2,2,2,2))  # Help with plotting margins in RStudio
plot(codaSamples1)
acf(codaSamples1[[1]])
geweke.diag(codaSamples1)
geweke.plot(codaSamples1)
gelman.diag(codaSamples1)
gelman.plot(codaSamples1)
```

## Summarize output ##

```r
options(scipen=999)  # A nice trick to remove scientific notation
summary(codaSamples1[[1]])
```
Figure 3: Trace and density plots for two chains
Figure 4: ACF plots for two chains
Figure 5: Geweke plots
Figure 6: Gelman-Rubin-Brooks plot
Fraction in 1st window = 0.1
Fraction in 2nd window = 0.5

\[
\begin{array}{ccc}
\mu & \sigma & \tau \\
0.003232 & 0.419334 & -0.428941
\end{array}
\]

[[2]]

Fraction in 1st window = 0.1
Fraction in 2nd window = 0.5

\[
\begin{array}{ccc}
\mu & \sigma & \tau \\
-1.0197 & -0.1983 & 0.1819
\end{array}
\]

Potential scale reduction factors:

\[
\begin{array}{ccc}
\text{Point est.} & \text{Upper C.I.} \\
\mu & 1 & 1 \\
\sigma & 1 & 1 \\
\tau & 1 & 1 \\
\end{array}
\]

Multivariate psrf

1

1. Empirical mean and standard deviation for each variable, plus standard error of the mean:

\[
\begin{array}{cccccc}
\text{Mean} & \text{SD} & \text{Naive SE} & \text{Time-series SE} \\
\mu & 0.009206 & 0.0503731 & 0.000503731 \\
\sigma & 5.005702 & 0.0353989 & 0.000353989 \\
\tau & 0.039915 & 0.0005645 & 0.000005645 \\
\end{array}
\]

2. Quantiles for each variable:

\[
\begin{array}{cccccc}
2.5\% & 25\% & 50\% & 75\% & 97.5\% \\
\mu & -0.08950 & -0.02500 & 0.00948 & 0.04332 & 0.10741 \\
\sigma & 4.93647 & 4.98148 & 5.00554 & 5.02996 & 5.07527 \\
\tau & 0.03882 & 0.03952 & 0.03991 & 0.04030 & 0.04104 \\
\end{array}
\]
The frequentist and Bayesian goals of model building are the same.
The frequentist and Bayesian goals of model building are the same.

- Model specification based on prior knowledge
The frequentist and Bayesian goals of model building are the same.

1. Model specification based on prior knowledge
2. Model estimation and fitting
The frequentist and Bayesian goals of model building are the same.

1. Model specification based on prior knowledge
2. Model estimation and fitting
3. Model evaluation and modification
The frequentist and Bayesian goals of model building are the same.

1. Model specification based on prior knowledge
2. Model estimation and fitting
3. Model evaluation and modification
4. Model choice
Despite these similarities there are important differences.

A major difference between the Bayesian and frequentist goals of model building lie in the model specification stage.

Because the Bayesian perspective explicitly incorporates uncertainty regarding model parameters using priors, the first phase of modeling building requires the specification of a full probability model for the data and the parameters of the model.

Model fit implies that the full probability model fits the data. Lack of model fit may be due to incorrect specification of likelihood, the prior distribution, or both.
Another difference between the Bayesian and frequentist goals of model building relates to the justification for choosing a particular model among a set of competing models.

Model building and model choice in the frequentist domain is based primarily on choosing the model that best fits the data.

This has certainly been the key motivation for model building, respecification, and model choice in the context of structural equation modeling (Kaplan 2009).

In the Bayesian domain, the choice among a set of competing models is based on which model provides the best posterior predictions.

That is, the choice among a set of competing models should be based on which model will best predict what actually happened.
A very natural way of evaluating the quality of a model is to examine how well the model fits the actual data.

In the context of Bayesian statistics, the approach to examining how well a model predicts the data is based on the notion of **posterior predictive checks**, and the accompanying **posterior predictive p-value**.

The general idea behind posterior predictive checking is that there should be little, if any, discrepancy between data generated by the model, and the actual data itself.

Posterior predictive checking is a method for assessing the specification quality of the model. Any deviation between the data generated from the model and the actual data implies model misspecification.
In the Bayesian context, the approach to examining model fit and specification utilizes the posterior predictive distribution of replicated data.

Let $y^{rep}$ be data replicated from our current model.
In the Bayesian context, the approach to examining model fit and specification utilizes the posterior predictive distribution of replicated data.

Let $y^{rep}$ be data replicated from our current model.

**Posterior Predictive Distribution**

$$p(y^{rep}|y) = \int p(y^{rep}|\theta)p(\theta|y)d\theta$$

$$= \int p(y^{rep}|\theta)p(y|\theta)p(\theta)d\theta$$  (17)
In the Bayesian context, the approach to examining model fit and specification utilizes the posterior predictive distribution of replicated data.

Let $y^{rep}$ be data replicated from our current model.

**Posterior Predictive Distribution**

\[
p(y^{rep}|y) = \int p(y^{rep}|\theta)p(\theta|y)d\theta
\]

\[
= \int p(y^{rep}|\theta)p(y|\theta)p(\theta)d\theta
\]

Equation (17) states that the distribution of future observations given the present data, $p(y^{rep}|y)$, is equal to the probability distribution of the future observations given the parameters, $p(y^{rep}|\theta)$, weighted by the posterior distribution of the model parameters.
To assess model fit, posterior predictive checking implies that the replicated data should match the observed data quite closely if we are to conclude that the model fits the data.

One approach to model fit in the context of posterior predictive checking is based on Bayesian $p$-values.

Denote by $T(y)$ a test statistic (e.g. $\chi^2$) based on the data, and let $T(y^{rep})$ be the same test statistics for the replicated data (based on MCMC). Then, the Bayesian $p$-value is defined to be
To assess model fit, posterior predictive checking implies that the replicated data should match the observed data quite closely if we are to conclude that the model fits the data.

One approach to model fit in the context of posterior predictive checking is based on Bayesian $p$-values.

Denote by $T(y)$ a test statistic (e.g. $\chi^2$) based on the data, and let $T(y^{rep})$ be the same test statistics for the replicated data (based on MCMC). Then, the Bayesian $p$-value is defined to be

$$p-value = pr[T(y^{rep}, \theta) \geq T(y, \theta)|y].$$

(18)
To assess model fit, posterior predictive checking implies that the replicated data should match the observed data quite closely if we are to conclude that the model fits the data.

One approach to model fit in the context of posterior predictive checking is based on Bayesian $p$-values.

Denote by $T(y)$ a test statistic (e.g. $\chi^2$) based on the data, and let $T(y^{rep})$ be the same test statistics for the replicated data (based on MCMC). Then, the Bayesian $p$-value is defined to be

$$p-value = pr[T(y^{rep}, \theta) \geq T(y, \theta) | y].$$

The $p$-value is the proportion of replicated test values that equal or exceed the observed test value. High (or low if signs are reversed) values indicate poor model fit.
A very simple and intuitive approach to model building and model selection 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.

Consider two competing models, denoted as \( M_1 \) and \( M_2 \), that could be nested within a larger space of alternative models. Let \( \theta_1 \) and \( \theta_2 \) be the two parameter vectors associated with these two models.

These could be two regression models with a different number of variables, or two structural equation models specifying very different directions of mediating effects.
The goal is to develop a quantity that expresses the extent to which the data support $M_1$ over $M_2$. One quantity could be the posterior odds of $M_1$ over $M_2$, expressed as
The goal is to develop a quantity that expresses the extent to which the data support $M_1$ over $M_2$. One quantity could be the posterior odds of $M_1$ over $M_2$, expressed as

$$\frac{p(M_1|y)}{p(M_2|y)} = \frac{p(y|M_1)}{p(y|M_2)} \times \left[ \frac{p(M_1)}{p(M_2)} \right]. \quad (19)$$

Bayes Factors
The goal is to develop a quantity that expresses the extent to which the data support $M_1$ over $M_2$. One quantity could be the posterior odds of $M_1$ over $M_2$, expressed as

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

(19)

Notice that the first term on the right hand side of equation (19) is the ratio of two integrated likelihoods.

This ratio is referred to as the \textit{Bayes factor} for $M_1$ over $M_2$, denoted here as $B_{12}$.

Our prior opinion regarding the odds of $M_1$ over $M_2$, given by $p(M_1)/p(M_2)$ is weighted by our consideration of the data, given by $p(y|M_1)/p(y|M_2)$. 

This weighting gives rise to our updated view of evidence provided by the data for either hypothesis, denoted as 
\[ p(M_1|y)/p(M_2|y). \]

An inspection of equation (19) also suggests that the Bayes factor is the ratio of the posterior odds to the prior odds.

In practice, there may be no prior preference for one model over the other. In this case, the prior odds are neutral and 
\[ p(M_1) = p(M_2) = 1/2. \]

When the prior odds ratio equals 1, then the posterior odds is equal to the Bayes factor.
A popular measure for model selection used in both frequentist and Bayesian applications is based on an approximation of the Bayes factor and is referred to as the *Bayesian information criterion* (BIC), also referred to as the Schwarz criterion.

Consider two models, $M_1$ and $M_2$ with $M_2$ nested in $M_1$. Under conditions where there is little prior information, the BIC can be written as

\[
\text{BIC} = -2 \log(\hat{\theta} | y) + p \log(n)
\]
A popular measure for model selection used in both frequentist and Bayesian applications is based on an approximation of the Bayes factor and is referred to as the *Bayesian information criterion* (BIC), also referred to as the Schwarz criterion.

Consider two models, $M_1$ and $M_2$ with $M_2$ nested in $M_1$. Under conditions where there is little prior information, the BIC can be written as

$$BIC = -2 \log(\hat{\theta}|y) + p \log(n)$$  \hspace{1cm} (20)$$

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.
The BIC is often used for model comparisons. The difference between two BIC measures comparing, say $M_1$ to $M_2$ can be written as

$$\Delta(BIC_{12}) = BIC(M_1) - BIC(M_2)$$

(21)

$$= \log(\hat{\theta}_1 | y) - \log(\hat{\theta}_2 | y) - \frac{1}{2}(p_1 - p_2) \log(n)$$
The BIC is often used for model comparisons. The difference between two BIC measures comparing, say $M_1$ to $M_2$ can be written as

$$\Delta(BIC_{12}) = BIC_{(M_1)} - BIC_{(M_2)}$$

$$= \log(\hat{\theta}_1 | y) - \log(\hat{\theta}_2 | y) - \frac{1}{2}(p_1 - p_2) \log(n)$$ (21)
The BIC is often used for model comparisons. The difference between two BIC measures comparing, say $M_1$ to $M_2$ can be written as

$$
\Delta(BIC_{12}) = BIC_{(M_1)} - BIC_{(M_2)}
$$

$$
= \log(\hat{\theta}_1 | y) - \log(\hat{\theta}_2 | y) - \frac{1}{2}(p_1 - p_2) \log(n)
$$
Rules of thumb have been developed to assess the quality of the evidence favoring one hypothesis over another using Bayes factors and the comparison of BIC values from two competing models. Using $M_1$ as the reference model,
- Rules of thumb have been developed to assess the quality of the evidence favoring one hypothesis over another using Bayes factors and the comparison of BIC values from two competing models. Using $M_1$ as the reference model,

<table>
<thead>
<tr>
<th>BIC Difference</th>
<th>Bayes Factor</th>
<th>Evidence against $M_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 to 2</td>
<td>1 to 3</td>
<td>Weak</td>
</tr>
<tr>
<td>2 to 6</td>
<td>3 to 20</td>
<td>Positive</td>
</tr>
<tr>
<td>6 to 10</td>
<td>20 to 150</td>
<td>Strong</td>
</tr>
<tr>
<td>&gt; 10</td>
<td>&gt; 150</td>
<td>Very strong</td>
</tr>
</tbody>
</table>
The BIC (ironically) is not fundamentally Bayesian

An explicitly Bayesian approach to model comparison has been developed based on the notion of *Bayesian deviance*.

Define *Bayesian deviance* as

\[
D(\theta) = -2 \log[p(y|\theta)] + 2 \log[h(y)],
\]

(22)

To make this Bayesian, we obtain a posterior mean over \( \theta \) by defining

\[
D(\theta) = E_{\theta}[ -2 \log[p(y|\theta)|y] + 2 \log[h(y)] ].
\]

(23)
- The BIC (ironically) is not fundamentally Bayesian
- An explicitly Bayesian approach to model comparison has been developed based on the notion of Bayesian deviance.
- Define Bayesian deviance as

\[
D(\theta) = -2 \log[p(y|\theta)] + 2 \log[h(y)],
\]

(22)
Deviance Information Criterion

- The BIC (ironically) is not fundamentally Bayesian

- An explicitly Bayesian approach to model comparison has been developed based on the notion of *Bayesian deviance*.

- Define *Bayesian deviance* as


\[
D(\theta) = -2 \log[p(y|\theta)] + 2 \log[h(y)],
\]

(22)

- To make this Bayesian this, we obtain a posterior mean over \( \theta \) by defining
Deviance Information Criterion

- The BIC (ironically) is not fundamentally Bayesian
- An explicitly Bayesian approach to model comparison has been developed based on the notion of *Bayesian deviance*.
- Define *Bayesian deviance* as

\[
D(\theta) = -2 \log [p(y|\theta)] + 2 \log [h(y)],
\]

(22)

To make this Bayesian this, we obtain a posterior mean over \( \theta \) by defining

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

(23)
Let \( D(\bar{\theta}) \) be a posterior estimate of \( \theta \). We can define the effective dimension of the model as \( q_D = D(\theta) - D(\bar{\theta}) \).

We then add the model fit term \( D(\theta) \) to obtain the deviance information criterion (DIC) - namely,
Let $D(\bar{\theta})$ be a posterior estimate of $\theta$. We can define the effective dimension of the model as $q_D = D(\theta) - D(\bar{\theta})$. We then add the model fit term $\overline{D(\theta)}$ to obtain the deviance information criterion (DIC) - namely,

$$DIC = \overline{D(\theta)} + q_D = 2\overline{D(\theta)} - D(\bar{\theta}).$$
Let $D(\bar{\theta})$ be a posterior estimate of $\theta$. We can define the effective dimension of the model as $q_D = D(\bar{\theta}) - D(\bar{\theta})$.

We then add the model fit term $D(\bar{\theta})$ to obtain the deviance information criterion (DIC) - namely,

$$DIC = D(\bar{\theta}) + q_D = 2D(\bar{\theta}) - D(\bar{\theta}).$$ (24)

The DIC can be obtained by calculating equation (23) over MCMC samples.

Models with the lowest DIC values are preferred.
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, Raftery & Madigan (1999) who write
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, Raftery & Madigan (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)
Bayesian Model Averaging

- 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, Raftery & Madigan (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 approach to addressing the problem is the method of Bayesian model averaging (BMA). We will show this in the regression example.
Consider a quantity of interest such as a future observation denoted as $\Upsilon$.

Next, consider a set of competing models $M_k$, $k = 1, 2, \ldots, K$ that are not necessarily nested.

The posterior distribution of $\Upsilon$ given data $y$ can be written as

$$p(\Upsilon | y) = \sum_{k=1}^{K} p(\Upsilon | M_k) p(M_k | y).$$

where

$$p(M_k | y) = \frac{p(y | M_k) p(M_k)}{\sum_{l=1}^{K} p(y | M_l) p(M_l), l \neq k}.$$
Bayesian Model Averaging

- Consider a quantity of interest such as a future observation denoted as $\Upsilon$.

- Next, consider a set of competing models $M_k$, $k = 1, 2, \ldots, K$ that are not necessarily nested.

- The posterior distribution of $\Upsilon$ given data $y$ can be written as

$$p(\Upsilon|y) = \sum_{k=1}^{K} p(\Upsilon|M_k)p(M_k|y).$$  \hspace{1cm} (25)

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

Consider a quantity of interest such as a future observation denoted as $\Upsilon$.

Next, consider a set of competing models $M_k$, $k = 1, 2, \ldots, K$ that are not necessarily nested.

The posterior distribution of $\Upsilon$ given data $y$ can be written as

$$p(\Upsilon|y) = \sum_{k=1}^{K} p(\Upsilon|M_k)p(M_k|y).$$

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

$$p(M_k|y) = \frac{p(y|M_k)p(M_k)}{\sum_{l=1}^{K} p(y|M_l)p(M_l)}, \quad l \neq k.$$
The interesting feature of equation (26) is that \( p(M_k|y) \) will likely be different for different models.

The term \( p(y|M_k) \) can be expressed as an integrated likelihood
The interesting feature of equation (26) is that \( p(M_k | y) \) will likely be different for different models.

The term \( p(y | M_k) \) can be expressed as an integrated likelihood

\[
p(y | M_k) = \int p(y | \theta_k, M_k) p(\theta_k | M_k) d\theta_k, \tag{27}
\]

where \( p(\theta_k | M_k) \) is the prior distribution of \( \theta_k \) under model \( M_k \) (Raftery et al., 1997).
The interesting feature of equation (26) is that \( p(M_k|y) \) will likely be different for different models.

The term \( p(y|M_k) \) can be expressed as an integrated likelihood

\[
p(y|M_k) = \int p(y|\theta_k, M_k)p(\theta_k|M_k)d\theta_k,
\]

(27)

where \( p(\theta_k|M_k) \) is the prior distribution of \( \theta_k \) under model \( M_k \) (Raftery et al., 1997).

BMA provides an approach for combining models specified by researchers.

Madigan and Raftery (1994) show that BMA provides better predictive performance than that of a single model based on a log-score rule.
BMA is difficult to implement.
BMA is difficult to implement.

The number of terms in
\[ p(\mathcal{Y} | y) = \sum_{k=1}^{K} p(\mathcal{Y} | M_k)p(M_k | y) \]
can be quite large and the corresponding integrals are hard to compute.
• BMA is difficult to implement.

1. The number of terms in \( p(\gamma | y) = \sum_{k=1}^{K} p(\gamma | M_k) p(M_k | y) \) can be quite large and the corresponding integrals are hard to compute.

2. Eliciting \( p(M_k) \) may not be straightforward. The uniform prior \( \frac{1}{M} \) is often used.
BMA is difficult to implement.

1. The number of terms in $p(\gamma | y) = \sum_{k=1}^{K} p(\gamma | M_k)p(M_k | y)$ can be quite large and the corresponding integrals are hard to compute.

2. Eliciting $p(M_k)$ may not be straightforward. The uniform prior $1/M$ is often used.

3. Choosing the class of models to average over is also challenging.
BMA is difficult to implement.

1. The number of terms in \( p(\mathcal{Y}|y) = \sum_{k=1}^{K} p(\mathcal{Y}|M_k)p(M_k|y) \) can be quite large and the corresponding integrals are hard to compute.

2. Eliciting \( p(M_k) \) may not be straightforward. The uniform prior \( 1/M \) is often used.

3. 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 has led to solutions based on *Occam’s window* and *Markov Chain Monte Carlo Model Composition* – \( M^3 \) (Madigan and Raftery, 1994).
Bayesian Linear Regression

- We will start with the basic multiple regression model.

- Let \( y \) be an \( n \)-dimensional vector \((y_1, y_2, \ldots, y_n)'\) \((i = 1, 2, \ldots, n)\) of reading scores from \( n \) students on the PIRLS reading assessment, and let \( X \) be an \( n \times k \) matrix containing \( k \) background and attitude measures. Then, the normal linear regression model can be written as
We will start with the basic multiple regression model.

Let $\mathbf{y}$ be an $n$-dimensional vector $(y_1, y_2, \ldots, y_n)'$ $(i = 1, 2, \ldots, n)$ of reading scores from $n$ students on the PIRLS reading assessment, and let $\mathbf{X}$ be an $n \times k$ matrix containing $k$ background and attitude measures. Then, the normal linear regression model can be written as

$$ y = \mathbf{X}\beta + \mathbf{u}, $$ (28)
Bayesian Linear Regression

- We will start with the basic multiple regression model.

- Let $y$ be an $n$-dimensional vector $(y_1, y_2, \ldots, y_n)'$ ($i = 1, 2, \ldots, n$) of reading scores from $n$ students on the PIRLS reading assessment, and let $X$ be an $n \times k$ matrix containing $k$ background and attitude measures. Then, the normal linear regression model can be written as

\[
y = X\beta + u,
\]  
(28)

- All the usual regression assumptions apply

- We assume that student level PIRLS reading scores are generated from a normal distribution.
- Recall that the components of Bayes’ theorem require priors on all model parameters.

- We will write the probability distribution for the regression model as
Recall that the components of Bayes’ theorem require priors on all model parameters.

We will write the probability distribution for the regression model as

\[ p(X, y|\beta, \sigma^2) \] (29)
Recall that the components of Bayes’ theorem require priors on all model parameters.

We will write the probability distribution for the regression model as

\[ p(X, y | \beta, \sigma^2) \] (29)

Conventional statistics stops here and estimates the model parameters with either maximum likelihood estimation or ordinary least square.

But for Bayesian regression we need to specify the priors for all model parameters.
First consider non-informative priors.

In the context of the normal linear regression model, the uniform distribution is typically used as a non-informative prior.

That is, we assign an improper uniform prior to the regression coefficient $\beta$ that allows $\beta$ to take on values over the support $[-\infty, \infty]$.

This can be written as $p(\beta) \propto c$, where $c$ is a constant.
Next, we assign a uniform prior to $\log(\sigma^2)$ because this transformation also allows values over the support $[0, \infty]$.

From here, the joint posterior distribution of the model parameters is obtained by multiplying the prior distributions of $\beta$ and $\sigma^2$ by the likelihood given in equation (29).

Assuming that $\beta$ and $\sigma^2$ are independent, we obtain

$$p(\beta, \sigma^2 | y, X) \propto p(X, y | \beta, \sigma^2) p(\beta) p(\sigma^2).$$  \hspace{1cm} (30)

In virtually all packages, non-informative or weakly informative priors are the default.
Next, we assign a uniform prior to $\log(\sigma^2)$ because this transformation also allows values over the support $[0, \infty]$.

From here, the joint posterior distribution of the model parameters is obtained by multiplying the prior distributions of $\beta$ and $\sigma^2$ by the likelihood given in equation (29).

Assuming that $\beta$ and $\sigma^2$ are independent, we obtain

$$p(\beta, \sigma^2 | y, X) \propto p(X, y | \beta, \sigma^2)p(\beta)p(\sigma^2).$$  

(30)
Next, we assign a uniform prior to $\log(\sigma^2)$ because this transformation also allows values over the support $[0, \infty]$.

From here, the joint posterior distribution of the model parameters is obtained by multiplying the prior distributions of $\beta$ and $\sigma^2$ by the likelihood given in equation (29).

Assuming that $\beta$ and $\sigma^2$ are independent, we obtain

$$p(\beta, \sigma^2 | y, X) \propto p(X, y | \beta, \sigma^2)p(\beta)p(\sigma^2).$$  \hspace{1cm} (30)

In virtually all packages, non-informative or weakly informative priors are the default.
What about informative priors?

The most sensible conjugate prior distribution for the vector of regression coefficients $\beta$ of the linear regression model is the multivariate normal prior.

The conjugate prior for the variance of the disturbance term $\sigma^2$ is the inverse-Gamma distribution, with shape and scale hyperparameters $a$ and $b$, respectively.

From here, we can obtain the joint posterior distribution of all model parameters using conjugate priors based on expert opinion or prior research.
Data come from a sub-sample of 5000 Canadian students who took part in the IEA Program for International Reading Literacy Study (PIRLS 2011).

The international population for PIRLS 2011 consisted of students in the grade that represents four years of schooling, provided that the mean age at the time of testing was at least 9.5 years.
Variables in this model are:

1. male (1=male)
2. ASBG04 = # of books in home
3. ASBGSBS = Bullying/teasing at school (higher values mean less bullying/teasing)
4. ASBGSMR = Students motivated to read
5. ASBGSCR = Students confidence in their reading
6. ASBR05E = Teacher is easy to understand
7. ASBR05F = Interested in what teacher says
8. ASBR05G = Teacher gives interesting things to read
Bayesian Multiple Regression: Non-informative Priors

```r
### Multiple Regression Model ###
## Install and load packages ##
install.packages("rjags")
require(rjags)

## Read in data ##
Canadareg <- read.csv(file.choose(),header=TRUE)  # browse to select data "Canada.csv"
male <- ifelse(Canadareg$ITSEX==2,0,1)
Canadareg <- cbind(Canadareg,male)
Canadareg <- subset(Canadareg, select=c(ASRREA01,male,ASBG04,ASBGSBS,ASBGSMR,ASBGSCR,
                                ASBR05E,ASBR05F,ASBR05G))
Canadareg[Canadareg==999999]=NA
Canadareg <- na.omit(Canadareg)
Canadareg <- Canadareg[sample(1:nrow(Canadareg), 5000,replace=F),]
nData=NROW(Canadareg)

## Begin JAGS Code ##
modelstring = "

# Likelihood
model {  
  for (i in 1:nData) {  
    ASRREA01[i] ˜ dnorm(mu[i], tau)  
    mu[i] <- a + b1*male[i] + b2*ASBG04[i] + b3*ASBGSBS[i] + b4*ASBGSMR[i]  
      + b5*ASBGSCR[i] + b6*ASBR05E[i] + b7*ASBR05F[i] + b8*ASBR05G[i]  
  }

```
Bayesian Regression: Non-informative Priors

# Priors

\[
a \sim \text{dnorm}(0, .0001) \\
b_1 \sim \text{dnorm}(0, .0001) \\
b_2 \sim \text{dnorm}(0, .0001) \\
b_3 \sim \text{dnorm}(0, .0001) \\
b_4 \sim \text{dnorm}(0, .0001) \\
b_5 \sim \text{dnorm}(0, .0001) \\
b_6 \sim \text{dnorm}(0, .0001) \\
b_7 \sim \text{dnorm}(0, .0001) \\
b_8 \sim \text{dnorm}(0, .0001) \\
\tau \sim \text{dgamma}(.01, 0.01) \\
\sigma \sim \text{dunif}(0, 100)
\]

```r
}
```

""
```
## End JAGS Code ##
```
Bayesian Regression: Non-informative Priors

attach(Canadareg)
writeLines(modelstring, con="model.bug")
Canadareg <- list(ASRREA01=ASRREA01,male=male,ASBG04=ASBG04,ASBGSBS=ASBGSBS,
                   ASBGSMR=ASBGSMR,ASBGSCR=ASBGSCR,
                   ASBR05E=ASBR05E,ASBR05F=ASBR05F,ASBR05G=ASBR05G,nData=nData)
parameters = c("a","b1","b2","b3","b4","b5","b6","b7","b8","tau")
lapply(Canadareg,summary)
Canadareg1 <-jags.model("model.bug", data=Canadareg,n.chains=2, n.adapt=5000)
cat("Burning in the MCMC chain ...\n")
update(Canadareg1, n.iter=5000)
cat("Sampling from the final MCMC chain ... \n")
codaSamples1 = coda.samples(Canadareg1,variable.names=parameters,
                             n.iter=10000, thin=10,seed=5555)

## Use coda to obtain diagnostic plots ##
par(mar=c(2,2,2,2))
plot(codaSamples1)
acf(codaSamples1[[1]])
geweke.diag(codaSamples1)
gelman.plot(codaSamples1)
gelman.diag(codaSamples1)

## Summarize output ##
options(scipen=999) # A nice trick to remove scientific notation
summary(codaSamples1[[1]])
Figure 7: Trace and density plots for regression example
Figure 8: Trace and density plots for regression example
Figure 9: Selected ACF plots for regression example
Geweke Diagnostic
[[1]]

Fraction in 1st window = 0.1
Fraction in 2nd window = 0.5

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>b1</td>
<td>b2</td>
<td>b3</td>
<td>b4</td>
<td>b5</td>
</tr>
<tr>
<td>0.2561</td>
<td>-0.4423</td>
<td>-0.6049</td>
<td>0.1415</td>
<td>-0.0185</td>
<td>0.1620</td>
</tr>
<tr>
<td>b6</td>
<td>b7</td>
<td>b8</td>
<td>tau</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-0.7107</td>
<td>0.5939</td>
<td>-0.3631</td>
<td>0.8690</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

[[2]]

Fraction in 1st window = 0.1
Fraction in 2nd window = 0.5

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>b1</td>
<td>b2</td>
<td>b3</td>
<td>b4</td>
<td>b5</td>
</tr>
<tr>
<td>-0.08788</td>
<td>1.19659</td>
<td>-0.14634</td>
<td>-0.19179</td>
<td>0.31039</td>
<td>0.31930</td>
</tr>
<tr>
<td>b6</td>
<td>b7</td>
<td>b8</td>
<td>tau</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-0.11029</td>
<td>0.10144</td>
<td>-0.41664</td>
<td>0.06028</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Potential scale reduction factors:

<table>
<thead>
<tr>
<th></th>
<th>Point est.</th>
<th>Upper C.I.</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>b1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>b2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>b3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>b4</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>b5</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>b6</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>b7</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>b8</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>tau</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Multivariate psrf

1
Iterations = 5010:105000
Thinning interval = 10
Number of chains = 1
Sample size per chain = 10000

1. Empirical mean and standard deviation for each variable, plus standard error of the mean:

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>SD</th>
<th>Naive SE</th>
<th>Time-series SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>330.81</td>
<td>7.52</td>
<td>0.0752</td>
<td>0.2285</td>
</tr>
<tr>
<td>b1</td>
<td>2.30</td>
<td>1.76</td>
<td>0.0176</td>
<td>0.0126</td>
</tr>
<tr>
<td>b2</td>
<td>12.06</td>
<td>0.83</td>
<td>0.0083</td>
<td>0.0126</td>
</tr>
<tr>
<td>b3</td>
<td>5.73</td>
<td>0.47</td>
<td>0.0047</td>
<td>0.0103</td>
</tr>
<tr>
<td>b4</td>
<td>-2.66</td>
<td>0.53</td>
<td>0.0053</td>
<td>0.0106</td>
</tr>
<tr>
<td>b5</td>
<td>12.94</td>
<td>0.48</td>
<td>0.0048</td>
<td>0.0103</td>
</tr>
<tr>
<td>b6</td>
<td>12.93</td>
<td>0.47</td>
<td>0.0047</td>
<td>0.0103</td>
</tr>
<tr>
<td>b7</td>
<td>1.24</td>
<td>1.47</td>
<td>0.0147</td>
<td>0.0321</td>
</tr>
<tr>
<td>b8</td>
<td>-4.86</td>
<td>1.42</td>
<td>0.0142</td>
<td>0.0316</td>
</tr>
<tr>
<td>tau</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

2. Quantiles for each variable:

<table>
<thead>
<tr>
<th></th>
<th>2.5%</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
<th>97.5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>315.93</td>
<td>325.75</td>
<td>330.76</td>
<td>335.93</td>
<td>345.39</td>
</tr>
<tr>
<td>b1</td>
<td>-1.14</td>
<td>1.15</td>
<td>2.29</td>
<td>3.47</td>
<td>5.77</td>
</tr>
<tr>
<td>b2</td>
<td>10.45</td>
<td>11.52</td>
<td>12.06</td>
<td>12.62</td>
<td>13.68</td>
</tr>
<tr>
<td>b3</td>
<td>4.79</td>
<td>5.43</td>
<td>5.74</td>
<td>6.46</td>
<td>6.66</td>
</tr>
<tr>
<td>b4</td>
<td>-3.69</td>
<td>-2.99</td>
<td>-2.64</td>
<td>-2.29</td>
<td>-1.64</td>
</tr>
<tr>
<td>b5</td>
<td>11.98</td>
<td>12.61</td>
<td>12.93</td>
<td>13.25</td>
<td>13.85</td>
</tr>
<tr>
<td>b6</td>
<td>2.87</td>
<td>4.77</td>
<td>5.76</td>
<td>6.77</td>
<td>8.69</td>
</tr>
<tr>
<td>b7</td>
<td>-1.67</td>
<td>0.27</td>
<td>1.23</td>
<td>2.21</td>
<td>4.11</td>
</tr>
<tr>
<td>b8</td>
<td>-7.66</td>
<td>-5.81</td>
<td>-4.86</td>
<td>-3.90</td>
<td>-2.05</td>
</tr>
<tr>
<td>tau</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>
### Multiple Regression Model with PPC and DIC ###

#### Install and load packages ####

```r
install.packages("rjags")
require(rjags)
```

#### Read in data ####

```r
Canadareg <- read.csv(file.choose(), header=TRUE) # browse to select data "Canada.csv"
male <- ifelse(Canadareg$ITSEX==2,0,1)
Canadareg <- cbind(Canadareg, male)
Canadareg <- subset(Canadareg, select=c(ASRREA01, male, ASBG04, ASBGSBS, ASBGSLR, ASBGSMR, ASBGSCR, ASBR05E, ASBR05F, ASBR05G))
Canadareg[Canadareg==999999]=NA
Canadareg <- na.omit(Canadareg)
Canadareg <- Canadareg[sample(1:nrow(Canadareg), 5000, replace=F),]
nData=NROW(Canadareg)
```

#### Begin JAGS Code ####

```r
modelstring = "
# Likelihood
model {
  for (i in 1:nData) {
    ASRREA01[i] ~ dnorm(mu[i], tau)
    mu[i] <- a + b1*male[i] + b2*ASBG04[i] + b3*ASBGSBS[i] + b4*ASBGSMR[i]
    + b5*ASBGSCR[i] + b6*ASBR05E[i] + b7*ASBR05F[i] + b8*ASBR05G[i]
  }

  # Obtaining information for posterior pred. checks
  res[i] <- ASRREA01[i] - mu[i]
  ASRREA01.new[i] ~ dnorm(mu[i], tau)
  res.new[i] <- ASRREA01.new[i] - mu[i]
}
""
```

102 / 165
# Priors

\[
\begin{align*}
    a & \sim \text{dnorm}(0, 0.0001) \\
    b_1 & \sim \text{dnorm}(0, 0.0001) \\
    b_2 & \sim \text{dnorm}(0, 0.0001) \\
    b_3 & \sim \text{dnorm}(0, 0.0001) \\
    b_4 & \sim \text{dnorm}(0, 0.0001) \\
    b_5 & \sim \text{dnorm}(0, 0.0001) \\
    b_6 & \sim \text{dnorm}(0, 0.0001) \\
    b_7 & \sim \text{dnorm}(0, 0.0001) \\
    b_8 & \sim \text{dnorm}(0, 0.0001) \\
    \tau & \sim \text{dgamma}(0.01, 0.01) \\
    \sigma & \sim \text{dunif}(0, 100)
\end{align*}
\]

# Derived parameters for post. pred. checks #

```r
fit <- sum(res[ ])
fit.new <- sum(res.new[ ])
```

```
```

## End JAGS Code ##
attach(Canadareg)
writeLines(modelstring, con="model.bug")
Canadareg <- list(ASRREA01=ASRREA01, male=male, ASBG04=ASBG04, ASBGBS=ASBGBS, ASBGSMR=ASBGSMR, ASBGSCR=ASBGSCR, ASBR05E=ASBR05E, ASBR05F=ASBR05F, ASBR05G=ASBR05G, nData=nData)
parameters = c("a","b1","b2","b3","b4","b5","b6","b7","b8","tau","fit","fit.new")
lapply(Canadareg, summary)
Canadareg1 <- jags.model("model.bug", data=Canadareg, n.chains=2, n.adapt=5000)

cat("Burning in the MCMC chain ...
")
update(Canadareg1, n.iter=5000)
cat("Sampling from the final MCMC chain ...
")
codaSamples1 = coda.samples(Canadareg1, variable.names=parameters, n.iter=10000, thin=10, seed=5555)

# Deviance Information Criterion
dic.pD <- dic.samples(Canadareg1, 5000, "pD")

## Use coda to obtain diagnostic plots ##
plot(codaSamples1)
acf(codaSamples1[[1]])
geweke.diag(codaSamples1)
geweke.plot(codaSamples1)
gelman.diag(codaSamples1)
gelman.plot(codaSamples1)

## Summarize output ##
options(scipen=999) # A nice trick to remove scientific notation
summary(codaSamples1[[1]])
# Posterior predictive check and DIC #

```r
fit.all <- unlist(codaSamples1[, 'fit'])  # unlist removes list params and places in vector 
fitnew.all <- unlist(codaSamples1[, 'fit.new'])

bayespval <- mean(fit.all > fitnew.all)
bayespval

plot(fit.all, fitnew.all, xlab = "Actual Dataset", ylab = "Simulated Dataset",
     main = paste("Posterior Predictive Check", "\n", "Bayesian P-value =",
                    round(bayespval, 2)))
abline(0,1)
```

# Deviance Information Criterion

dic.pD
**Figure 10:** Posterior predictive distribution plot and Bayesian p-value – Noninformative priors. Penalized deviance = 55315.
Bayesian Model Averaging

## Bayesian Model Averaging ###

###------ Select DV and IVs and run BMA ----------###

install.packages("BMA")
require(BMA)

#------- Read in data and delete missing data ---------#
Canadareg <- read.csv(file.choose(),header=TRUE) #browse to select data "Canada.csv"
male <- ifelse(Canadareg$ITSEX==2,0,1)
Canadareg <- cbind(Canadareg,male)
Canadareg[Canadareg==999999]=NA
Canadareg <- na.omit(Canadareg)
Canadareg <- Canadareg[sample(1: nrow(Canadareg), 5000,replace=F),]
CanadaBMA <- subset(Canadareg, select=c(ASRREA01,male,ASBG04,ASBGSBS,ASBGSMR,ASBGSCR,
                                               ASBR05E,ASBR05F,ASBR05G))

y <- CanadaBMA[,1]
x <- CanadaBMA[,2:9]

CanadaBMA1 <- bicreg(x,y)
summary(CanadaBMA1)
par(mar=c(1,1,1,1))
plot(CanadaBMA1)
Call:
\[
bicreg(x = x, y = y)
\]
5 models were selected
Best 5 models (cumulative posterior probability = 1):

<table>
<thead>
<tr>
<th></th>
<th>p!=0</th>
<th>EV</th>
<th>SD</th>
<th>model 1</th>
<th>model 2</th>
<th>model 3</th>
<th>model 4</th>
<th>model 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>100.0</td>
<td>356.48</td>
<td>8.676</td>
<td>360.30</td>
<td>351.76</td>
<td>359.58</td>
<td>350.99</td>
<td>353.88</td>
</tr>
<tr>
<td>male</td>
<td>16.0</td>
<td>0.636</td>
<td>1.611</td>
<td>.</td>
<td>.</td>
<td>3.962</td>
<td>3.991</td>
<td>.</td>
</tr>
<tr>
<td>ASBG04</td>
<td>100.0</td>
<td>12.05</td>
<td>0.839</td>
<td>12.05</td>
<td>12.05</td>
<td>12.01</td>
<td>12.02</td>
<td>12.02</td>
</tr>
<tr>
<td>ASBG05S</td>
<td>100.0</td>
<td>4.57</td>
<td>0.472</td>
<td>4.747</td>
<td>4.563</td>
<td>4.687</td>
<td>4.504</td>
<td>4.649</td>
</tr>
<tr>
<td>ASBG05MR</td>
<td>100.0</td>
<td>-3.83</td>
<td>0.507</td>
<td>-3.73</td>
<td>-3.98</td>
<td>-3.77</td>
<td>-4.03</td>
<td>-3.59</td>
</tr>
<tr>
<td>ASBG05CR</td>
<td>100.0</td>
<td>13.69</td>
<td>0.464</td>
<td>13.76</td>
<td>13.63</td>
<td>13.74</td>
<td>13.60</td>
<td>13.64</td>
</tr>
<tr>
<td>ASBR05E</td>
<td>44.8</td>
<td>1.828</td>
<td>2.258</td>
<td>.</td>
<td>3.932</td>
<td>.</td>
<td>3.951</td>
<td>5.043</td>
</tr>
<tr>
<td>ASBR05F</td>
<td>6.1</td>
<td>-0.19</td>
<td>0.839</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>-3.20</td>
</tr>
<tr>
<td>ASBR05G</td>
<td>0.0</td>
<td>0.000</td>
<td>0.000</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>4</th>
<th>5</th>
<th>5</th>
<th>6</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>nVar</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>r2</td>
<td>0.229</td>
<td>0.230</td>
<td>0.230</td>
<td>0.231</td>
<td>0.231</td>
</tr>
<tr>
<td>BIC</td>
<td>-1268.21</td>
<td>-1267.48</td>
<td>-1265.01</td>
<td>-1264.36</td>
<td>-1264.16</td>
</tr>
<tr>
<td>post prob</td>
<td>0.459</td>
<td>0.320</td>
<td>0.093</td>
<td>0.067</td>
<td>0.061</td>
</tr>
</tbody>
</table>
Figure 11: BMA posterior distributions for model parameters. Note that the narrow spike corresponds to \( p(\beta_k = 0 | y) \). Thus, in the case of ASBR05F, note that \( p(\beta > 0 | y) = .06 \) and therefore \( p(\beta = 0 | y) = 0.94 \). corresponding to the spike found in the BMA posterior distribution plot for ASBR05F.
Bayesian Regression: Informative Priors

```r
## Begin JAGS Code ##

modelstring = "

# Likelihood

model {
  for (i in 1:nData) {
    ASRREA01[i] ˜ dnorm(mu[i], tau)
    mu[i] <- a + b1*male[i] + b2*ASBG04[i] + b3*ASBG0S[i] + b4*ASBGMR[i]
    + b5*ASBGSCR[i] + b6*ASBR05E[i] + b7*ASBR05F[i] + b8*ASBR05G[i]
  }

# Informative Priors

  a ˜ dnorm(315, .0001)
  b1 ˜ dnorm(-1.1, 1)
  b2 ˜ dnorm(10, 1)
  b3 ˜ dnorm(5, 1)
  b4 ˜ dnorm(-3.6, 1)
  b5 ˜ dnorm(12, 1)
  b6 ˜ dnorm(3, 1)
  b7 ˜ dnorm(-1.7, 1)
  b8 ˜ dnorm(-7.7, 1)
  tau ˜ dgamma(.01, 0.01)
  sigma ˜ dunif(0, 100)
}
"

## End JAGS Code ##
```
1. Empirical mean and standard deviation for each variable, plus standard error of the mean:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>SD</th>
<th>Naive SE</th>
<th>Time-series SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>348.05</td>
<td>7.27</td>
<td>0.23</td>
<td>0.73</td>
</tr>
<tr>
<td>b1</td>
<td>-1.34</td>
<td>0.87</td>
<td>0.03</td>
<td>0.03</td>
</tr>
<tr>
<td>b2</td>
<td>11.85</td>
<td>0.64</td>
<td>0.02</td>
<td>0.02</td>
</tr>
<tr>
<td>b3</td>
<td>4.83</td>
<td>0.41</td>
<td>0.01</td>
<td>0.02</td>
</tr>
<tr>
<td>b4</td>
<td>-3.02</td>
<td>0.44</td>
<td>0.01</td>
<td>0.03</td>
</tr>
<tr>
<td>b5</td>
<td>13.46</td>
<td>0.39</td>
<td>0.01</td>
<td>0.03</td>
</tr>
<tr>
<td>b6</td>
<td>4.83</td>
<td>0.82</td>
<td>0.03</td>
<td>0.04</td>
</tr>
<tr>
<td>b7</td>
<td>0.80</td>
<td>0.79</td>
<td>0.03</td>
<td>0.04</td>
</tr>
<tr>
<td>b8</td>
<td>-5.12</td>
<td>0.77</td>
<td>0.03</td>
<td>0.04</td>
</tr>
<tr>
<td>tau</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

2. Quantiles for each variable:

<table>
<thead>
<tr>
<th>Variable</th>
<th>2.5%</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
<th>97.5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>334.6</td>
<td>342.8</td>
<td>348.1</td>
<td>353.0</td>
<td>362.2</td>
</tr>
<tr>
<td>b1</td>
<td>-3.04</td>
<td>-1.92</td>
<td>-1.33</td>
<td>-0.77</td>
<td>0.35</td>
</tr>
<tr>
<td>b2</td>
<td>10.62</td>
<td>11.43</td>
<td>11.86</td>
<td>12.26</td>
<td>13.09</td>
</tr>
<tr>
<td>b3</td>
<td>4.00</td>
<td>4.54</td>
<td>4.83</td>
<td>5.12</td>
<td>5.61</td>
</tr>
<tr>
<td>b4</td>
<td>-3.84</td>
<td>-3.32</td>
<td>-3.01</td>
<td>-2.71</td>
<td>-2.12</td>
</tr>
<tr>
<td>b5</td>
<td>12.69</td>
<td>13.19</td>
<td>13.45</td>
<td>13.69</td>
<td>14.23</td>
</tr>
<tr>
<td>b6</td>
<td>3.25</td>
<td>4.28</td>
<td>4.86</td>
<td>5.40</td>
<td>6.37</td>
</tr>
<tr>
<td>b7</td>
<td>-0.87</td>
<td>0.29</td>
<td>0.83</td>
<td>1.32</td>
<td>2.34</td>
</tr>
<tr>
<td>b8</td>
<td>-6.59</td>
<td>-5.63</td>
<td>-5.11</td>
<td>-4.58</td>
<td>-3.71</td>
</tr>
<tr>
<td>tau</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>
Figure 12: Posterior predictive distribution plot and Bayesian p-value: Informative priors case. Penalized Deviance = 55457.
Day 2 PM: Student Analyses: Regression

1. Choose a country.

2. Run the full analysis with non-informative priors - obtain PPCs and DIC values.

3. Run full analysis again with informative priors on regression coefficients only based on Canada results - obtain PPCs and DIC values.

4. Compare DIC values.

5. Make a note of what happens to the PPIs when estimating the model with informative priors.
A common feature of data collection in the social sciences is that units of analysis (e.g. students or employees) are nested in higher organizational units (e.g. schools or companies, respectively).

The PIRLS study deliberately samples schools (within a country) and then takes a sample of 4th grade students within sampled schools.

Such data collection plans are generically referred to as *clustered sampling designs*. 
In addition to being able to incorporate priors directly into a multilevel model, the Bayesian conception of multilevel modeling has another advantage – namely it clears up a great deal of confusion in the presentation of multilevel models.

The literature on multilevel modeling attempts to make a distinction between so-called “fixed-effects” and “random-effects”.

Gelman and Hill have recognized this issue and present five different definitions of fixed and random effects.

The advantage of the Bayesian approach is that all parameters are assumed to be random. When conceived as a Bayesian hierarchical model, much of the confusion around terminology disappears.
Perhaps the most basic multilevel model is the random effects analysis of variance model.

As a simple example consider whether there are differences among $G$ schools ($g = 1, 2, \ldots, G$) on the outcome of student reading performance $y$ obtained from $n$ students ($i = 1, 2, \ldots, n$).

In this example, it is assumed that the $G$ schools are a random sample from a population of schools.
The model can be written as a two level hierarchical linear model as follows: Let
The model can be written as a two level hierarchical linear model as follows: Let

**Level - 1**

\[
y_{ig} = \beta_g + r_{ig},
\tag{31}
\]
The model can be written as a two level hierarchical linear model as follows: Let

**Level - 1**

\[ y_{ig} = \beta_g + r_{ig}, \]  

(31)

The model for the school random effect can be written as
The model can be written as a two level hierarchical linear model as follows: Let

**Level - 1**

\[ y_{ig} = \beta_g + r_{ig}, \quad (31) \]

The model for the school random effect can be written as

**Level - 2**

\[ \beta_g = \mu + u_g, \quad (32) \]
The model can be written as a two level hierarchical linear model as follows: Let

**Level - 1**

\[ y_{ig} = \beta_g + r_{ig}, \quad (31) \]

The model for the school random effect can be written as

**Level - 2**

\[ \beta_g = \mu + u_g, \quad (32) \]

Inserting equation (32) into equation (31) yields
The model can be written as a two level hierarchical linear model as follows: Let

**Level - 1**

\[
y_{ig} = \beta_g + r_{ig}, \tag{31}
\]

The model for the school random effect can be written as

**Level - 2**

\[
\beta_g = \mu + u_g, \tag{32}
\]

Inserting equation (32) into equation (31) yields

\[
y_{ig} = \mu + u_g + r_{ig}. \tag{33}
\]
For equation (33), we first specify the distribution of the reading performance outcome $y_{ig}$ given the school effect $u_g$ and the within school variance $\sigma^2$. Specifically,
For equation (33), we first specify the distribution of the reading performance outcome $y_{ig}$ given the school effect $u_g$ and the within school variance $\sigma^2$. Specifically,

$$y_{ig} \mid u_g, \sigma^2 \sim N(u_g, \sigma^2).$$  \hspace{1cm} (34)
For equation (33), we first specify the distribution of the reading performance outcome $y_{ig}$ given the school effect $u_g$ and the within school variance $\sigma^2$. Specifically,

$$y_{ig} | u_g, \sigma^2 \sim N(u_g, \sigma^2). \quad (34)$$

Next specify the prior distribution on the remaining model parameters. For this model, we specify conjugate priors
For equation (33), we first specify the distribution of the reading performance outcome $y_{ig}$ given the school effect $u_g$ and the within school variance $\sigma^2$. Specifically,

$$y_{ig} \mid u_g, \sigma^2 \sim N(u_g, \sigma^2).$$

Next specify the prior distribution on the remaining model parameters. For this model, we specify conjugate priors

$$u_g \mid \mu, \omega^2 \sim N(0, \omega^2),$$

$$\mu \sim N(b_0, B_0),$$

$$\sigma^2 \sim \text{inverse-Gamma}(a, b),$$

$$\omega^2 \sim \text{inverse-Gamma}(a, b).$$
We can arrange all of the parameters of the random effects ANOVA model into a vector $\theta$ and write the prior distribution as
We can arrange all of the parameters of the random effects ANOVA model into a vector $\theta$ and write the prior distribution as

$$p(\theta) = p(u_1, u_2, \ldots, u_G, \mu, \sigma^2, \omega^2), \quad (39)$$

where under the assumption of exchangeability of the school effects $u_g$ we obtain
We can arrange all of the parameters of the random effects ANOVA model into a vector $\theta$ and write the prior distribution as

$$p(\theta) = p(u_1, u_2, \ldots, u_G, \mu, \sigma^2, \omega^2),$$  \hspace{1cm} (39)

where under the assumption of exchangeability of the school effects $u_g$ we obtain

$$p(\theta) = \prod_{g=1}^{G} p(u_g | \mu, \omega^2)p(\mu)p(\sigma^2)p(\omega^2).$$  \hspace{1cm} (40)
Bayesian Multilevel Regression using rjags

```
## ONE-WAY RANDOM EFFECTS ANOVA: Non-Informative Priors ##
## Install and load packages ##
install.packages("rjags")
require(rjags)
# Read in data and make transformations #

Canada <- read.csv(file.choose(),header=TRUE) # browse to select data "Canada.csv"
Canada[Canada==999999]=NA
Canada1 <- na.omit(Canada)
Canada$id <- as.numeric(as.factor(Canada$IDSCHOOL))  # This creates a new sequential ID.
Canadasub <- subset(Canada1, select=c(ASRREA01,id))
Canada2 <- Canadasub[1:5000,] # Just to reduce the size of the dataset.
nData=NROW(Canada2)
nGroups = length(unique(Canada2$id))

## JAGS Code ##
modelstring = "

# Likelihood

for(i in 1:nData) {
    ASRREA01[i] ~ dnorm(mu.y[i], tau1)
    mu.y[i] <- beta[id[i]]  # Assigns the correct group mean to the specific student
}

for (g in 1:nGroups) {
    beta[g] ~ dnorm(mu, tau2)
}
```

Bayesian Multilevel Regression using rjags

```r
# Priors
mu  ~ dnorm(0.0,0.001)
tau1 ~ dgamma(.001,.001)
tau2 ~ dgamma(.001,.001)
sigma1 <- 1/tau1
sigma2 <- 1/tau2
ICC <- sigma2/(sigma1+sigma2)
```

```
attach(Canada2)
writeLines(modelstring,con="model.bug")
Canada2 <- list(ASRREA01=ASRREA01, id=id,nData=nData,nGroups=nGroups)
parameters = c("mu","tau1", "tau2","sigma1","sigma2","ICC")
lapply(Canada2,summary)
CanadaM1 <-jags.model("model.bug", data=Canada2, n.chains=2, n.adapt=1000)
cat("Burning in the MCMC chain ...
")
update(CanadaM1, n.iter=5000)
cat("Sampling from the final MCMC chain ... 
")
codaSamples1 = coda.samples(CanadaM1,variable.names=parameters,
  n.iter=100000, thin=10,seed=5555)

## Use coda to obtain diagnostic plots ##
plot(codaSamples1)
acf(codaSamples1[[1]])
geweke.diag(codaSamples1)
geweke.diag(codaSamples1)
gelman.diag(codaSamples1)
gelman.plot(codaSamples1)

## Summarize output ##
options(scipen=999) # A nice trick to remove scientific notation
summary(codaSamples1[[1]])
```
Figure 13: Trace and density plot for one-way random-effects ANOVA
Figure 14: ACF plots for one-way random-effects ANOVA
[[1]]

Fraction in 1st window = 0.1
Fraction in 2nd window = 0.5

\[
\begin{array}{cccccc}
\text{ICC} & \text{mu} & \sigma_1 & \sigma_2 & \tau_1 & \tau_2 \\
-0.3539 & 1.7102 & -2.0538 & -0.7617 & 2.0515 & 0.9777 \\
\end{array}
\]

[[2]]

Fraction in 1st window = 0.1
Fraction in 2nd window = 0.5

\[
\begin{array}{cccccc}
\text{ICC} & \text{mu} & \sigma_1 & \sigma_2 & \tau_1 & \tau_2 \\
-0.2456 & -0.6950 & 0.2759 & -0.1807 & -0.2853 & 0.2354 \\
\end{array}
\]

> gelman.diag(codaSamples1)
Potential scale reduction factors:

<table>
<thead>
<tr>
<th>Point est.</th>
<th></th>
<th>Upper C.I.</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>ICC</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>mu</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>\sigma_1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>\sigma_2</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>\tau_1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>\tau_2</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

Multivariate psrf

1
Iterations = 5010:105000
Thinning interval = 10
Number of chains = 1
Sample size per chain = 10000

1. Empirical mean and standard deviation for each variable, plus standard error of the mean:

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>SD</th>
<th>Naive SE</th>
<th>Time-series SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICC</td>
<td>0.1865017</td>
<td>0.015332535</td>
<td>0.00015332535</td>
<td>0.00015388992</td>
</tr>
<tr>
<td>mu</td>
<td>552.4813990</td>
<td>1.667775858</td>
<td>0.01667775858</td>
<td>0.01667775858</td>
</tr>
<tr>
<td>sigma1</td>
<td>3793.7522713</td>
<td>79.370977405</td>
<td>0.79370977405</td>
<td>0.79370977405</td>
</tr>
<tr>
<td>sigma2</td>
<td>870.8763015</td>
<td>84.689699722</td>
<td>0.84689699722</td>
<td>0.86524045885</td>
</tr>
<tr>
<td>tau1</td>
<td>0.0002637</td>
<td>0.000005515</td>
<td>0.00000005515</td>
<td>0.00000005515</td>
</tr>
<tr>
<td>tau2</td>
<td>0.0011592</td>
<td>0.000113268</td>
<td>0.00000113268</td>
<td>0.00000113268</td>
</tr>
</tbody>
</table>

2. Quantiles for each variable:

<table>
<thead>
<tr>
<th></th>
<th>2.5%</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
<th>97.5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICC</td>
<td>0.1576265</td>
<td>0.175838</td>
<td>0.1862509</td>
<td>0.1965435</td>
<td>0.2175658</td>
</tr>
<tr>
<td>mu</td>
<td>549.1980070</td>
<td>551.371135</td>
<td>552.4835792</td>
<td>553.6074288</td>
<td>555.7565580</td>
</tr>
<tr>
<td>sigma1</td>
<td>3641.4297289</td>
<td>3739.188086</td>
<td>3792.7877359</td>
<td>3845.9567853</td>
<td>3955.0834783</td>
</tr>
<tr>
<td>sigma2</td>
<td>714.0057804</td>
<td>812.468735</td>
<td>867.9593784</td>
<td>924.9670708</td>
<td>1047.7049816</td>
</tr>
<tr>
<td>tau1</td>
<td>0.00002528</td>
<td>0.0000260</td>
<td>0.00002637</td>
<td>0.00002674</td>
<td>0.00002746</td>
</tr>
<tr>
<td>tau2</td>
<td>0.00009545</td>
<td>0.0001081</td>
<td>0.00011521</td>
<td>0.00012308</td>
<td>0.00014005</td>
</tr>
</tbody>
</table>
Suppose interest centers on individual and school socio-economic factors in term of their prediction of reading performance in PIRLS.

We can write a substantive model as
Suppose interest centers on individual and school socio-economic factors in term of their prediction of reading performance in PIRLS.

We can write a substantive model as

\[ \text{READING}_{ig} = \beta_{0ig} + \beta_{1ig}(\text{ASBH17A}) + \beta_{2ig}(\text{ASBH17B}) + r_{ig}, \]

where \( \beta_{kig} \ (k = 1, 2, \ldots, K + 1) \) are the intercept and regression coefficients (slopes) that are allowed to vary over the \( G \) schools. The variables ASBG17A = Fathers Education; ASBG17B = Mothers Education
We can model the intercept and slopes as a function of school level predictors.
We can model the intercept and slopes as a function of school level predictors.

**Level - 2**

\begin{align*}
\beta_{0g} &= \gamma_{00} + \gamma_{01}(ACBG03A)_g + u_{0g}, \\
\beta_{1g} &= \gamma_{10} + \gamma_{11}(ACBG03A)_g + u_{1g}, \\
\beta_{2g} &= \gamma_{20} + \gamma_{21}(ACBG03A)_g + u_{2g},
\end{align*}

(41a) \quad (41b) \quad (41c)

where \(\gamma\)'s are the coefficients relating \(\beta_{kg}\) to the school level predictors, and \(ACBG03A = \) percent of students in school \(g\) from economically disadvantaged homes.
The mixed model can be written as
The mixed model can be written as

\[
\text{READING}_{ig} = \gamma_0 + \gamma_1 (ACBG03A) + \gamma_2 (ASBG17A) + \gamma_3 (ASBG17B) + \gamma_4 (ACBG03A) + \gamma_5 (ACBG17B) + \gamma_6 (ACBG17B) + u_0 + u_1 (ASBG17A) + u_2 (ASBG17B) + u_3 + r_{ig}.
\]
The mixed model can be written as

**Mixed Model**

\[
READING_{ig} = \gamma_{00} + \gamma_{01}(ACBG03A) + \gamma_{10}(ASBG17A) + \gamma_{11}(ASBG17B)(ACBG03A) + \gamma_{20}(ASBG17B) + \gamma_{21}(ACBG17B)(ACBG03A) + u_{0g} + u_{1g}(ASBG17A) + u_{2g}(ASBG17B) + u_{3g} + r_{ig}.
\]

In terms of a Bayesian hierarchical model, the priors would have to be chosen for \(\sigma_g^2\) and the hyperparameters \(\gamma_g\) and \(\omega_k^2\).
Bayesian Multilevel Regression
Intercept-Only Model

### Multilevel Regression: Intercept with covariate model, Non-Informative Priors ###

#### Install and load packages ####

```r
install.packages("rjags")
require(rjags)
```

#### Read in data and make transformations ####

```r
Canada <- read.csv(file.choose(), header=TRUE) # browse to select data "Canada.csv"
Canada[Canada==999999]=NA
Canadal <- na.omit(Canada)
Canadal$schid <- as.numeric(as.factor(Canadal$IDSCHOOL)) # This creates a new sequential ID.
Canadasub <- subset(Canadal, select=c(ASRREA01,ASBH17A,ASBH17B,ACBG03A,schid))
Canada2 <- Canadasub[1:5000,] # Just to reduce the size of the dataset.
```

```r
nData=NROW(Canada2)
nGroups = length(unique(Canada2$schid))
```

#### JAGS Code ####

```r
modelstring = 

model {
  for (i in 1:nData) {
    mu[i] <- alpha[schid[i]] + beta1[schid[i]]*ASBH17A[i] + beta2[schid[i]]*ASBH17B[i]
    ASRREA01[i] ~ dnorm(mu[i], tau.c)
  }

  for (j in 1:nGroups) {
    alpha[j] ~ dnorm(gam00 + gam10*ACBG03A[j], alpha.tau)
    beta1[j] ~ dnorm(beta1.mu, beta1.tau)
    beta2[j] ~ dnorm(beta2.mu, beta2.tau)
  }
}
```

```r
129 / 165"
# Priors
gam00  ~  dnorm(0, 0.0001)
gam10  ~  dnorm(0, 0.0001)
beta1.mu  ~  dnorm(0, 1.0E-4)
beta2.mu  ~  dnorm(0, 1.0E-4)
tau.c  ~  dgamma(1.0E-3, 1.0E-3)
alpha.tau  ~  dgamma(1.0E-3, 1.0E-3)
beta1.tau  ~  dgamma(1.0E-3, 1.0E-3)
beta2.tau  ~  dgamma(1.0E-3, 1.0E-3)

# Transformations
alpha.sigma <- 1.0/sqrt(alpha.tau)
beta1.sigma <- 1.0/sqrt(beta1.tau)
beta2.sigma <- 1.0/sqrt(beta2.tau)
sigma.c <- 1.0/sqrt(tau.c)

# Derived parameters for post. pred. checks#
  fit <- sum(res[ ])
  fit.new <- sum(res.new[ ])


attach(Canada2)
writeLines(modelstring, con="model.bug")
Canada2 <- list(ASRREA01=ASRREA01, ASBH17A=ASBH17A, ASBH17B=ASBH17B, ACBG03A=ACBG03A, schid=schid, nData=nData, nGroups=nGroups)
parameters = c("beta1.mu", "beta2.mu", "tau.c", "alpha.tau", "beta1.tau", "beta2.tau",
"alpha.sigma", "beta1.sigma", "beta2.sigma", "sigma.c", "gam00", "gam10", "fit", "fit.new")
lapply(Canada2, summary)
CanadaM1 <- jags.model("model.bug", data=Canada2, n.chains=2, n.adapt=5000)

cat("Burning in the MCMC chain ...
")
update(CanadaM1, n.iter=5000)
cat("Sampling from the final MCMC chain ...
")
codaSamples1 = coda.samples(CanadaM1, variable.names=parameters,
 n.iter=50000, thin=10, seed=5555)

# Deviance Information Criterion
dic1.pD <- dic.samples(CanadaM1, 5000, "pD")

## Use coda to obtain diagnostic plots ##

plot(codaSamples1)
acf(codaSamples1[[1]])
geweke.diag(codaSamples1)
gelman.diag(codaSamples1)
gelman.plot(codaSamples1)

## Summarize output ##

options(scipen=999)  # A nice trick to remove scientific notation
summary(codaSamples1[[1]])
# Obtain posterior predictive check

```r
# Obtain posterior predictive check

fit.all <- unlist(codaSamples1[, 'fit'])
fitnew.all <- unlist(codaSamples1[, 'fit.new'])

bayespval <- mean(fit.all > fitnew.all)
bayespval

pdf("~/desktop/multilevel1ppc.pdf")
plot(fit.all, fitnew.all, xlab = "Actual Dataset", ylab = "Simulated Dataset",
     main = paste("Posterior Predictive Check", "\n", "Bayesian P-value =",
                 round(bayespval, 2)))
abline(0,1)
dev.off()

# Deviance Information Criterion

dic1.pD
```
[[1]]  

Fraction in 1st window = 0.1  
Fraction in 2nd window = 0.5  

<table>
<thead>
<tr>
<th></th>
<th>alpha.sigma</th>
<th>alpha.tau</th>
<th>betal.mu</th>
<th>betal.sigma</th>
<th>betal.tau</th>
<th>beta2.mu</th>
<th>beta2.sigma</th>
</tr>
</thead>
<tbody>
<tr>
<td>Day 1 PM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Day 2 AM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Day 2 PM</td>
<td>1.6123</td>
<td>-6.8907</td>
<td>0.6724</td>
<td>-4.4081</td>
<td>4.3770</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Day 3 AM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Day 3 PM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Fraction in 1st window = 0.1  
Fraction in 2nd window = 0.5  

<table>
<thead>
<tr>
<th></th>
<th>alpha.sigma</th>
<th>alpha.tau</th>
<th>betal.mu</th>
<th>betal.sigma</th>
<th>betal.tau</th>
<th>beta2.mu</th>
<th>beta2.sigma</th>
</tr>
</thead>
<tbody>
<tr>
<td>Day 1 AM</td>
<td>-1.4316</td>
<td>1.3128</td>
<td>2.0892</td>
<td>0.5495</td>
<td>0.1426</td>
<td>-4.0355</td>
<td>0.6756</td>
</tr>
<tr>
<td>Day 1 PM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Day 2 AM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Day 2 PM</td>
<td>-1.3198</td>
<td>-1.7727</td>
<td>0.9650</td>
<td>0.6738</td>
<td>-0.6693</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Day 3 AM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Day 3 PM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Potential scale reduction factors:  
Point est. Upper C.I.

<table>
<thead>
<tr>
<th></th>
<th>Point est.</th>
<th>Upper C.I.</th>
</tr>
</thead>
<tbody>
<tr>
<td>alpha.sigma</td>
<td>1.08</td>
<td>1.30</td>
</tr>
<tr>
<td>alpha.tau</td>
<td>1.09</td>
<td>1.34</td>
</tr>
<tr>
<td>betal.mu</td>
<td>1.24</td>
<td>1.82</td>
</tr>
<tr>
<td>betal.sigma</td>
<td>1.04</td>
<td>1.16</td>
</tr>
<tr>
<td>betal.tau</td>
<td>1.14</td>
<td>1.24</td>
</tr>
<tr>
<td>beta2.mu</td>
<td>1.77</td>
<td>4.40</td>
</tr>
<tr>
<td>beta2.sigma</td>
<td>1.78</td>
<td>4.04</td>
</tr>
<tr>
<td>beta2.tau</td>
<td>1.37</td>
<td>2.91</td>
</tr>
<tr>
<td>gam00</td>
<td>1.03</td>
<td>1.14</td>
</tr>
<tr>
<td>gam10</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>sigma.c</td>
<td>1.00</td>
<td>1.01</td>
</tr>
<tr>
<td>tau.c</td>
<td>1.00</td>
<td>1.01</td>
</tr>
</tbody>
</table>

Multivariate psrf

1.89
1. Empirical mean and standard deviation for each variable, 
   plus standard error of the mean:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Mean</th>
<th>SD</th>
<th>Naive SE</th>
<th>Time-series SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>alpha.sigma</td>
<td>25.8467877</td>
<td>1.61711038</td>
<td>0.02286940362</td>
<td>0.09791926814</td>
</tr>
<tr>
<td>alpha.tau</td>
<td>0.0015152</td>
<td>0.000198336</td>
<td>0.00000280490</td>
<td>0.00001223128</td>
</tr>
<tr>
<td>beta1.mu</td>
<td>7.0357425</td>
<td>0.525386594</td>
<td>0.00743008847</td>
<td>0.11540910189</td>
</tr>
<tr>
<td>beta1.sigma</td>
<td>0.3706912</td>
<td>0.457392750</td>
<td>0.00646851030</td>
<td>0.07845048441</td>
</tr>
<tr>
<td>beta1.tau</td>
<td>234.7152012</td>
<td>480.698653137</td>
<td>6.79810554680</td>
<td>80.05237700237</td>
</tr>
<tr>
<td>beta2.mu</td>
<td>4.9892031</td>
<td>0.880284909</td>
<td>0.01244910857</td>
<td>0.19489310019</td>
</tr>
<tr>
<td>beta2.sigma</td>
<td>0.6932699</td>
<td>0.562461338</td>
<td>0.00795440453</td>
<td>0.09300373934</td>
</tr>
<tr>
<td>beta2.tau</td>
<td>64.3524557</td>
<td>226.300731351</td>
<td>3.20037563451</td>
<td>27.03025807615</td>
</tr>
<tr>
<td>gam00</td>
<td>481.5273343</td>
<td>7.335247174</td>
<td>0.10373606037</td>
<td>0.47988158616</td>
</tr>
<tr>
<td>gam10</td>
<td>1.3938938</td>
<td>1.658904517</td>
<td>0.02346045266</td>
<td>0.05213100139</td>
</tr>
<tr>
<td>sigma.c</td>
<td>60.2621432</td>
<td>0.635819024</td>
<td>0.00899183888</td>
<td>0.00964639716</td>
</tr>
<tr>
<td>tau.c</td>
<td>0.0002755</td>
<td>0.000005814</td>
<td>0.00000008222</td>
<td>0.00000008222</td>
</tr>
</tbody>
</table>

2. Quantiles for each variable:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>2.5%</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
<th>97.5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>alpha.sigma</td>
<td>22.4757574</td>
<td>24.8497642</td>
<td>25.8732026</td>
<td>26.9358635</td>
<td>28.8772000</td>
</tr>
<tr>
<td>alpha.tau</td>
<td>0.0011992</td>
<td>0.0013783</td>
<td>0.0014938</td>
<td>0.0016194</td>
<td>0.0019796</td>
</tr>
<tr>
<td>beta1.sigma</td>
<td>0.0239098</td>
<td>0.0662179</td>
<td>0.1638606</td>
<td>0.4925364</td>
<td>1.6452589</td>
</tr>
<tr>
<td>beta1.tau</td>
<td>0.3694295</td>
<td>4.1221588</td>
<td>37.2435313</td>
<td>228.0597676</td>
<td>1749.227855</td>
</tr>
<tr>
<td>beta2.mu</td>
<td>3.3489585</td>
<td>4.2156964</td>
<td>5.0813464</td>
<td>5.7009877</td>
<td>6.4458430</td>
</tr>
<tr>
<td>beta2.sigma</td>
<td>0.0376242</td>
<td>0.2342708</td>
<td>0.5348191</td>
<td>1.0643539</td>
<td>2.0241580</td>
</tr>
<tr>
<td>beta2.tau</td>
<td>0.2440682</td>
<td>0.8827300</td>
<td>3.4961194</td>
<td>18.2206481</td>
<td>706.4243779</td>
</tr>
<tr>
<td>gam00</td>
<td>466.8756075</td>
<td>476.5797154</td>
<td>481.6285088</td>
<td>486.4800117</td>
<td>495.6771803</td>
</tr>
<tr>
<td>gam10</td>
<td>-1.8043387</td>
<td>0.2673627</td>
<td>1.3644368</td>
<td>2.5045133</td>
<td>4.6860523</td>
</tr>
<tr>
<td>sigma.c</td>
<td>59.02450342</td>
<td>59.8239919</td>
<td>60.2593910</td>
<td>60.6938671</td>
<td>61.5096669</td>
</tr>
<tr>
<td>tau.c</td>
<td>0.0002643</td>
<td>0.0002715</td>
<td>0.0002754</td>
<td>0.0002794</td>
<td>0.0002868</td>
</tr>
</tbody>
</table>

Mean deviance: 55205
penalty 320.3
Penalized deviance: 55525
Figure 15: PPC for intercept-only model
Bayesian Multilevel Regression Slopes and Intercept Model

## Multilevel Regression: Intercept and slope with covariate model, Non-Informative Priors ##

## Install and load packages ##

```r
code
install.packages("rjags")
require(rjags)
```

# Read in data and make transformations #

```r
code
Canada <- read.csv(file.choose(), header=TRUE) # browse to select data "Canada.csv"
Canada[Canada==999999]=NA
Canadal <- na.omit(Canada)
Canadal$schid <- as.numeric(as.factor(Canadal$IDSCHOOL)) # This creates a new sequential ID.
Canadasub <- subset(Canadal, select=c(ASRREA01, ASBH17A, ASBH17B, ACBG03A, schid))
Canada2 <- Canadasub[1:5000,] # Just to reduce the size of the dataset.
```

```r
code
nData=NROW(Canada2)
nGroups = length(unique(Canada2$schid))

#### JAGS Code ####

```r
code
modelstring = "

# Likelihood
model {
  for (i in 1:nData) {
    mu[i] <- alpha[schid[i]] + beta1[schid[i]]*ASBH17A[i] + beta2[schid[i]]*ASBH17B[i]
    ASRREA01[i] ~ dnorm(mu[i], tau.c)
  
  
  # Obtaining information for posterior pred. checks 
  res[i] <- ASRREA01[i] - mu[i]
  ASRREA01.new[i] ~ dnorm(mu[i], tau.c)
  res.new[i] <- ASRREA01.new[i] - mu[i]
  }
" 
```

# Bayesian Multilevel Regression Slopes and Intercept Model
for (j in 1:nGroups) {
    alpha[j] ~ dnorm(gam00 + gam10*ACBG03A[j], alpha.tau)
    beta1[j] ~ dnorm(beta1.mu, beta1.tau)
    beta2[j] ~ dnorm(gam01 + gam11*ACBG03A[j], beta2.tau)
}

# Priors

gam00 ~ dnorm(0,.0001)
gam10 ~ dnorm(0,.0001)
gam01 ~ dnorm(0,.0001)
gam11 ~ dnorm(0,.0001)
beta1.mu ~ dnorm(0, 1.0E-4)
beta2.mu ~ dnorm(0, 1.0E-4)
tau.c ~ dgamma(1.0E-3, 1.0E-3)
alpha.tau ~ dgamma(1.0E-3, 1.0E-3)
beta1.tau ~ dgamma(1.0E-3, 1.0E-3)
beta2.tau ~ dgamma(1.0E-3, 1.0E-3)

# Transforming precision to standard deviation

alpha.sigma <- 1.0/sqrt(alpha.tau)
beta1.sigma <- 1.0/sqrt(beta1.tau)
beta2.sigma <- 1.0/sqrt(beta2.tau)
sigma.c <- 1.0/sqrt(tau.c)

# Derived parameters for post. pred. checks #

fit <- sum(res[, ])
fit.new <- sum(res.new[, ])

} "

### End of JAGS Code ###
attach(Canada2)
writeLines(modelstring, con="model.bug")
Canada2 <- list(ASRREA01=ASRREA01, ASBH17A=ASBH17A, ASBH17B=ASBH17B, ACBG03A=ACBG03A, schid=schid, nData=nData, nGroups=nGroups)
parameters = c("beta1.mu","tau.c","alpha.tau","beta1.tau","beta2.tau","alpha.sigma","beta1.sigma","beta2.sigma","sigma.c","gam00","gam10","gam01","gam11")
lapply(Canada2, summary)
CanadaM2 <- jags.model("model.bug", data=Canada2, n.chains=2, n.adapt=1000)

cat("Burning in the MCMC chain ...
")
update(CanadaM2, n.iter=5000)
dic2.pD <- dic.samples(CanadaM2, 5000, "pD") # Deviance Information Criterion
cat("Sampling from the final MCMC chain ...
")
codaSamples1 = coda.samples(CanadaM2, variable.names=parameters, 
n.iter=50000, thin=10, seed=5555)

## Use coda to obtain diagnostic plots ##
plot(codaSamples1)
acf(codaSamples1[[1]])
geweke.diag(codaSamples1)
gelman.diag(codaSamples1)
gelman.plot(codaSamples1)
## Summarize output ##

```r
options(scipen=999) # A nice trick to remove scientific notation
summary(codaSamples1[[1]])

# Obtain posterior predictive check

fit.all <- unlist(codaSamples1[, 'fit'])
fitnew.all <- unlist(codaSamples1[, 'fit.new'])

bayespval <- mean(fit.all > fitnew.all)
bayespval

plot(fit.all, fitnew.all, xlab = "Actual Dataset", ylab = "Simulated Dataset",
     main = paste("Posterior Predictive Check", "\n", "Bayesian P-value =",
                 round(bayespval, 2)))
abline(0,1)

# Deviance Information Criterion and DIC Difference from Model 1 #

dic2.pD
diffdic(dic1.pD, dic2.pD)
```
1. Empirical mean and standard deviation for each variable, plus standard error of the mean:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>SD</th>
<th>Naive SE</th>
<th>Time-series SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>alpha.sigma</td>
<td>26.1470609</td>
<td>1.552307516</td>
<td>0.02195294342</td>
<td>0.08742986958</td>
</tr>
<tr>
<td>alpha.tau</td>
<td>0.0014787</td>
<td>0.000182438</td>
<td>0.00000258006</td>
<td>0.00001016877</td>
</tr>
<tr>
<td>beta1.mu</td>
<td>7.2461117</td>
<td>0.623289307</td>
<td>0.00881464191</td>
<td>0.27903545459</td>
</tr>
<tr>
<td>beta1.sigma</td>
<td>0.1853041</td>
<td>0.2074473773</td>
<td>0.0293369797</td>
<td>0.02588946802</td>
</tr>
<tr>
<td>beta1.tau</td>
<td>242.5570675</td>
<td>398.661396022</td>
<td>5.63792353049</td>
<td>46.66551288483</td>
</tr>
<tr>
<td>beta2.sigma</td>
<td>0.5058206</td>
<td>0.50755968</td>
<td>0.00717792534</td>
<td>0.08639145534</td>
</tr>
<tr>
<td>beta2.tau</td>
<td>93.2323788</td>
<td>253.880423751</td>
<td>3.59041138490</td>
<td>33.54776990275</td>
</tr>
<tr>
<td>gam00</td>
<td>474.2382962</td>
<td>14.739020474</td>
<td>0.20844122650</td>
<td>5.14104773211</td>
</tr>
<tr>
<td>gam01</td>
<td>6.1213248</td>
<td>2.194435837</td>
<td>0.03103400922</td>
<td>0.98394110125</td>
</tr>
<tr>
<td>gam10</td>
<td>3.0432823</td>
<td>3.849583227</td>
<td>0.0544132809</td>
<td>0.99255029344</td>
</tr>
<tr>
<td>gam11</td>
<td>-0.3046821</td>
<td>0.628570497</td>
<td>0.00888932921</td>
<td>0.2763470958</td>
</tr>
<tr>
<td>sigma.c</td>
<td>60.2879355</td>
<td>0.63628435</td>
<td>0.00899189450</td>
<td>0.00968073858</td>
</tr>
<tr>
<td>tau.c</td>
<td>0.0002752</td>
<td>0.000005811</td>
<td>0.0000008218</td>
<td>0.0000008832</td>
</tr>
</tbody>
</table>

2. Quantiles for each variable:

<table>
<thead>
<tr>
<th>Variable</th>
<th>2.5%</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
<th>97.5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>alpha.sigma</td>
<td>23.002468</td>
<td>25.1842139</td>
<td>26.1627259</td>
<td>27.177529</td>
<td>29.1652713</td>
</tr>
<tr>
<td>alpha.tau</td>
<td>0.001176</td>
<td>0.0013539</td>
<td>0.0014609</td>
<td>0.001577</td>
<td>0.0018900</td>
</tr>
<tr>
<td>beta1.mu</td>
<td>6.201361</td>
<td>6.8354337</td>
<td>7.2414476</td>
<td>7.866749</td>
<td>8.122391</td>
</tr>
<tr>
<td>beta1.sigma</td>
<td>0.026203</td>
<td>0.0616696</td>
<td>0.1112855</td>
<td>0.230629</td>
<td>0.7320091</td>
</tr>
<tr>
<td>beta1.tau</td>
<td>1.866244</td>
<td>18.8006242</td>
<td>80.7462729</td>
<td>262.940314</td>
<td>1456.447438</td>
</tr>
<tr>
<td>beta2.sigma</td>
<td>0.034131</td>
<td>0.1367330</td>
<td>0.3017579</td>
<td>0.733402</td>
<td>1.8960134</td>
</tr>
<tr>
<td>beta2.tau</td>
<td>0.278174</td>
<td>1.8591544</td>
<td>10.9820639</td>
<td>53.487712</td>
<td>858.4246073</td>
</tr>
<tr>
<td>gam00</td>
<td>448.314005</td>
<td>464.2906292</td>
<td>473.9135428</td>
<td>483.445131</td>
<td>507.6447931</td>
</tr>
<tr>
<td>gam01</td>
<td>1.081862</td>
<td>4.7623234</td>
<td>6.2586801</td>
<td>7.182836</td>
<td>9.9804648</td>
</tr>
<tr>
<td>gam10</td>
<td>-5.500616</td>
<td>0.6950690</td>
<td>3.1138073</td>
<td>5.695323</td>
<td>9.9758779</td>
</tr>
<tr>
<td>gam11</td>
<td>-1.302419</td>
<td>-0.6708085</td>
<td>-0.3121674</td>
<td>0.030134</td>
<td>1.2610496</td>
</tr>
<tr>
<td>sigma.c</td>
<td>59.048764</td>
<td>59.8632633</td>
<td>60.2856261</td>
<td>60.704551</td>
<td>61.5407563</td>
</tr>
<tr>
<td>tau.c</td>
<td>0.000264</td>
<td>0.0002714</td>
<td>0.0002752</td>
<td>0.000279</td>
<td>0.0002868</td>
</tr>
</tbody>
</table>

Mean deviance: 55188
Penalty deviance: 269.6
Penalized deviance: 55458
Difference: 67.38174
Bayesian Factor Analysis

- We write the confirmatory factor analysis model as
We write the confirmatory factor analysis model as

\[ y = \alpha + \Lambda \eta + \epsilon, \]  

(43)
We write the confirmatory factor analysis model as

\[ y = \alpha + \Lambda \eta + \epsilon, \]  

(43)

Under conventional assumptions we obtain the model expressed in terms of the population covariance matrix \( \Sigma \) as
We write the confirmatory factor analysis model as

\[ y = \alpha + \Lambda \eta + \epsilon, \quad (43) \]

Under conventional assumptions we obtain the model expressed in terms of the population covariance matrix \( \Sigma \) as

\[ \Sigma = \Lambda \Phi \Lambda' + \Psi, \quad (44) \]
Let $\theta_{\text{norm}} = \{\alpha, \Lambda\}$ be the set of free model parameters that are assumed to follow a normal distribution and let $\theta_{\text{IW}} = \{\Phi, \Psi\}$ be the set of free model parameters that are assumed to follow an inverse-Wishart distribution. Thus,
Conjugate Priors for Factor Analysis

Parameters

Let $\theta_{\text{norm}} = \{\alpha, \Lambda\}$ be the set of free model parameters that are assumed to follow a normal distribution and let $\theta_{\text{IW}} = \{\Phi, \Psi\}$ be the set of free model parameters that are assumed to follow an inverse-Wishart distribution. Thus,

$$\theta_{\text{norm}} \sim N(\mu, \Omega), \quad (45)$$
Conjugate Priors for Factor Analysis

Parameters

- Let $\theta_{\text{norm}} = \{\alpha, \Lambda\}$ be the set of free model parameters that are assumed to follow a normal distribution and let $\theta_{\text{IW}} = \{\Phi, \Psi\}$ be the set of free model parameters that are assumed to follow an inverse-Wishart distribution. Thus,

$$\theta_{\text{norm}} \sim N(\mu, \Omega), \quad (45)$$

- The uniqueness covariance matrix $\Psi$ is assumed to follow an inverse-Wishart distribution. Specifically,
Let $\theta_{\text{norm}} = \{\alpha, \Lambda\}$ be the set of free model parameters that are assumed to follow a normal distribution and let $\theta_{\text{IW}} = \{\Phi, \Psi\}$ be the set of free model parameters that are assumed to follow an inverse-Wishart distribution. Thus,

$$\theta_{\text{norm}} \sim N(\mu, \Omega), \quad (45)$$

The uniqueness covariance matrix $\Psi$ is assumed to follow an inverse-Wishart distribution. Specifically,

$$\theta_{\text{IW}} \sim IW(R, \delta), \quad (46)$$
Conjugate Priors for Factor Analysis

- Let $\theta_{norm} = \{\alpha, \Lambda\}$ be the set of free model parameters that are assumed to follow a normal distribution and let $\theta_{IW} = \{\Phi, \Psi\}$ be the set of free model parameters that are assumed to follow an inverse-Wishart distribution. Thus,

\[ \theta_{norm} \sim N(\mu, \Omega), \quad (45) \]

- The uniqueness covariance matrix $\Psi$ is assumed to follow an inverse-Wishart distribution. Specifically,

\[ \theta_{IW} \sim IW(R, \delta), \quad (46) \]

- Different choices for $R$ and $\delta$ will yield different degrees of “informativeness” for the inverse-Wishart distribution.
install.packages("rjags")
require(rjags)
# READ IN DATA AND PREPARE FOR JAGS

CanadaCFA <- read.csv(file.choose(), header=TRUE) # browse to select data "Canada.csv"
CanadaCFA <- subset(CanadaCFA, select=c(ASBR08A, ASBR08B, ASBR08C, ASBR08D, ASBR08E, ASBR08F, ASBR08G))
CanadaCFA[CanadaCFA==999999]=NA
CanadaCFA <- na.omit(CanadaCFA)
CanadaCFA <- CanadaCFA[sample(1: nrow(CanadaCFA), 5000, replace=F),]

y = as.matrix(CanadaCFA)
nData=NROW(y)

# Specify model in JAGS
modelstring = " # Likelihood
model{
for (i in 1 : nData) {
for (j in 1 : 7) { # NCOL(y) = 7
y[i,j] ~ dnorm(mu[i,j], psi[j])
}
mu[i,1] <- a[1]+1*xi[i,1] # Factor 1, lam[1]=1
xi[i,1:2] ~ dmnorm(u[1:2], phi[1:2,1:2]) # Specify distribution of factors
}"
# Priors on factor loadings (Non-informative)

```
lam[1] <- 1
lam[2] ~ dnorm(0, 10^-6)  # prior mean and prior precision
lam[4] ~ dnorm(0, 10^-6)
lam[6] ~ dnorm(0, 10^-6)
```

```
lam[3] <- 1
lam[5] ~ dnorm(0, 10^-6)
lam[7] ~ dnorm(0, 10^-6)
```

# Priors on intercepts (Non-informative)

```
a[1] ~ dnorm(0, 10^-6)
a[2] ~ dnorm(0, 10^-6)
a[4] ~ dnorm(0, 10^-6)
a[6] ~ dnorm(0, 10^-6)
a[3] ~ dnorm(0, 10^-6)
a[5] ~ dnorm(0, 10^-6)
a[7] ~ dnorm(0, 10^-6)
```

# Priors on Precisions (Non-informative)

```
for(j in 1:7) {
  psi[j] ~ dgamma(1, 0.001)  # Precision of uniqueness
  uniq[j] <- 1/psi[j]  # unique variance
}
```

```
phi[1:2,1:2] ~ dwish(R[1:2,1:2], 3)  # phi is precision matrix
Sigma[1,1] <- phi[2,2]/(phi[1,1]*phi[2,2]-phi[2,1]^2)  # Sigma is covariance matrix
Sigma[2,1] <- -phi[1,2]/(phi[1,1]*phi[2,2]-phi[2,1]^2)
Sigma[2,2] <- phi[1,1]/(phi[1,1]*phi[2,2]-phi[2,1]^2)
Sigma[1,2] <- Sigma[2,1]
```

"
writeLines(modelstring, con="model.bug")
# Specify parameters

u=c(0,0)  # Factor means set to zero
R=matrix(c(1,0,0,1), nrow=2)
CanadaCFA <- list(y=y, nData=nData, u=u, R=R)

# RUN CHAIN
# Initialize Model
parameters = c("a", "lam", "Sigma", "uniq", "phi") # Specify the Parameters to Be Estimated

cfaModel1 = jags.model("model.bug", data=CanadaCFA, n.chains=2, n.adapt=500)

# Obtain the Posterior Sample of Factor Analysis Parameters: #
cat("Burning in the MCMC chain ...
")
update(cfaModel1, n.iter=500)
cat("Sampling from the final MCMC chain ...
")
codaSamples1 = coda.samples(cfaModel1, variable.names=parameters, 
n.iter=1000, thin=1, seed=5555)
## Diagnostics ##

# Selected Trace plots and Density plots for factor loadings #
plot(codaSamples1[[1]][,c(6,8,9,10,11)])

# Selected Trace plots and Density plots for variance terms #
plot(codaSamples1[[1]][,c(16:22)])

# Auto-correlation plots for selected factor loadings (first chain) #
par(mfrow=c(3,2))
acf(codaSamples1[[1]][,6],main="Item 2")
acf(codaSamples1[[1]][,8],main="Item 4")
acf(codaSamples1[[1]][,9],main="Item 5")
acf(codaSamples1[[1]][,10],main="Item 6")
acf(codaSamples1[[1]][,11],main="Item 7")

par(mfrow=c(2,2))
acf(codaSamples1[[1]][,1],main="Variance of Factor TEABEHA")
acf(codaSamples1[[1]][,2],main="Covariance betw TEABEHA and STUDBEHA")
acf(codaSamples1[[1]][,4],main="Variance of Factor STUDBEHA")
acf(codaSamples1[[1]][,22],main="Error Variance of Item 1")

# Geweke and Gelman Diagnostic and Plot
gegeweke.diag(codaSamples1[[1]])
gelman.diag(codaSamples1, multivariate = FALSE)
## Summary

# Posterior Mean, posterior SD and posterior probability interval (PPI)
```
options(scipen=999)
summary(codaSamples1[[1]])
```

# posterior mean of factor correlation #
```
options(scipen=999)
mean(codaSamples1[[1]][,2]/sqrt(codaSamples1[[1]][,1]*codaSamples1[[1]][,4]))
```

# posterior SD of factor correlation #
```
options(scipen=999)
sd(codaSamples1[[1]][,2]/sqrt(codaSamples1[[1]][,1]*codaSamples1[[1]][,4]))
```

# 95% PPI of factor corr.#
```
options(scipen=999)
quantile(codaSamples1[[1]][,2]/sqrt(codaSamples1[[1]][,1]*codaSamples1[[1]][,4]),c(.025,.975))
```
1. Empirical mean and standard deviation for each variable, plus standard error of the mean:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>SD</th>
<th>Naive SE</th>
<th>Time-series SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sigma[1,1]</td>
<td>12.1957649</td>
<td>0.3129577</td>
<td>0.009896591</td>
<td>0.08129586</td>
</tr>
<tr>
<td>Sigma[2,1]</td>
<td>5.4785041</td>
<td>0.1634803</td>
<td>0.005169701</td>
<td>0.05874001</td>
</tr>
<tr>
<td>Sigma[1,2]</td>
<td>5.4785041</td>
<td>0.1634803</td>
<td>0.005169701</td>
<td>0.05874001</td>
</tr>
<tr>
<td>Sigma[2,2]</td>
<td>3.7166390</td>
<td>0.0951442</td>
<td>0.003008724</td>
<td>0.03366946</td>
</tr>
<tr>
<td>a[1]</td>
<td>0.1246684</td>
<td>0.0302611</td>
<td>0.000956940</td>
<td>0.02425277</td>
</tr>
<tr>
<td>a[2]</td>
<td>1.6291422</td>
<td>0.0324200</td>
<td>0.001025213</td>
<td>0.00489975</td>
</tr>
<tr>
<td>a[3]</td>
<td>0.0791766</td>
<td>0.0181704</td>
<td>0.000574601</td>
<td>0.01234002</td>
</tr>
<tr>
<td>a[4]</td>
<td>2.4556130</td>
<td>0.0513908</td>
<td>0.001625120</td>
<td>0.00910447</td>
</tr>
<tr>
<td>a[5]</td>
<td>1.5167426</td>
<td>0.0287328</td>
<td>0.000908612</td>
<td>0.00251472</td>
</tr>
<tr>
<td>a[6]</td>
<td>1.7455273</td>
<td>0.0613882</td>
<td>0.001941264</td>
<td>0.01318495</td>
</tr>
<tr>
<td>a[7]</td>
<td>0.7244733</td>
<td>0.0222405</td>
<td>0.000703309</td>
<td>0.00197731</td>
</tr>
<tr>
<td>lam[1]</td>
<td>1.0000000</td>
<td>0.0000000</td>
<td>0.00000000</td>
<td>0.00000000</td>
</tr>
<tr>
<td>lam[2]</td>
<td>0.5620899</td>
<td>0.0110310</td>
<td>0.000348831</td>
<td>0.00203238</td>
</tr>
<tr>
<td>lam[3]</td>
<td>1.0000000</td>
<td>0.0000000</td>
<td>0.00000000</td>
<td>0.00000000</td>
</tr>
<tr>
<td>lam[4]</td>
<td>0.2698516</td>
<td>0.0139446</td>
<td>0.000409688</td>
<td>0.00238223</td>
</tr>
<tr>
<td>lam[5]</td>
<td>0.4753352</td>
<td>0.0143017</td>
<td>0.000452262</td>
<td>0.00120354</td>
</tr>
<tr>
<td>lam[6]</td>
<td>0.4348302</td>
<td>0.0185991</td>
<td>0.000581565</td>
<td>0.00423937</td>
</tr>
<tr>
<td>lam[7]</td>
<td>0.5134271</td>
<td>0.0165553</td>
<td>0.000330949</td>
<td>0.0076241</td>
</tr>
<tr>
<td>phi[1,1]</td>
<td>0.2428598</td>
<td>0.0048972</td>
<td>0.000154864</td>
<td>0.00026402</td>
</tr>
<tr>
<td>phi[2,1]</td>
<td>-0.3579597</td>
<td>0.0079196</td>
<td>0.000250440</td>
<td>0.00027702</td>
</tr>
<tr>
<td>phi[1,2]</td>
<td>-0.3579597</td>
<td>0.0079196</td>
<td>0.000250440</td>
<td>0.00027702</td>
</tr>
<tr>
<td>phi[2,2]</td>
<td>0.7969053</td>
<td>0.0154632</td>
<td>0.000488990</td>
<td>0.00050366</td>
</tr>
<tr>
<td>uniq[1]</td>
<td>0.0003776</td>
<td>0.0000914</td>
<td>0.000002893</td>
<td>0.00004753</td>
</tr>
<tr>
<td>uniq[2]</td>
<td>0.3347643</td>
<td>0.0070560</td>
<td>0.000223131</td>
<td>0.00023797</td>
</tr>
<tr>
<td>uniq[3]</td>
<td>0.0008867</td>
<td>0.0004810</td>
<td>0.000015214</td>
<td>0.00030639</td>
</tr>
<tr>
<td>uniq[4]</td>
<td>0.7729350</td>
<td>0.0150742</td>
<td>0.000476689</td>
<td>0.00047669</td>
</tr>
<tr>
<td>uniq[5]</td>
<td>0.9446526</td>
<td>0.0194162</td>
<td>0.000613996</td>
<td>0.00061400</td>
</tr>
<tr>
<td>uniq[6]</td>
<td>0.7013185</td>
<td>0.0138156</td>
<td>0.000436888</td>
<td>0.00045147</td>
</tr>
<tr>
<td>uniq[7]</td>
<td>0.5927233</td>
<td>0.0119940</td>
<td>0.000379285</td>
<td>0.00036103</td>
</tr>
</tbody>
</table>
2. Quantiles for each variable:

<table>
<thead>
<tr>
<th></th>
<th>2.5%</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
<th>97.5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sigma[1,1]</td>
<td>11.5758862</td>
<td>11.9746797</td>
<td>12.2030811</td>
<td>12.4045067</td>
<td>12.7691310</td>
</tr>
<tr>
<td>Sigma[2,1]</td>
<td>5.1684885</td>
<td>5.3659367</td>
<td>5.4814118</td>
<td>5.5825546</td>
<td>5.7911302</td>
</tr>
<tr>
<td>Sigma[1,2]</td>
<td>5.1684885</td>
<td>5.3659367</td>
<td>5.4814118</td>
<td>5.5825546</td>
<td>5.7911302</td>
</tr>
<tr>
<td>a[1]</td>
<td>0.0809455</td>
<td>0.0982338</td>
<td>0.1186369</td>
<td>0.1452640</td>
<td>0.1816772</td>
</tr>
<tr>
<td>a[2]</td>
<td>1.5706381</td>
<td>1.6070419</td>
<td>1.6269780</td>
<td>1.6489623</td>
<td>1.7000288</td>
</tr>
<tr>
<td>a[3]</td>
<td>0.0462116</td>
<td>0.0674402</td>
<td>0.0784316</td>
<td>0.0855467</td>
<td>0.1146649</td>
</tr>
<tr>
<td>a[4]</td>
<td>2.3532690</td>
<td>2.4169692</td>
<td>2.4590488</td>
<td>2.4886161</td>
<td>2.5564497</td>
</tr>
<tr>
<td>a[5]</td>
<td>1.4631486</td>
<td>1.4970458</td>
<td>1.5160299</td>
<td>1.5375097</td>
<td>1.5743647</td>
</tr>
<tr>
<td>a[6]</td>
<td>1.6358590</td>
<td>1.7009827</td>
<td>1.7469751</td>
<td>1.7879749</td>
<td>1.8700960</td>
</tr>
<tr>
<td>a[7]</td>
<td>0.6823280</td>
<td>0.7094545</td>
<td>0.7242721</td>
<td>0.7397865</td>
<td>0.7676564</td>
</tr>
<tr>
<td>lam[1]</td>
<td>1.0000000</td>
<td>1.0000000</td>
<td>1.0000000</td>
<td>1.0000000</td>
<td>1.0000000</td>
</tr>
<tr>
<td>lam[2]</td>
<td>0.5401068</td>
<td>0.5542617</td>
<td>0.5625471</td>
<td>0.5698119</td>
<td>0.5832293</td>
</tr>
<tr>
<td>lam[3]</td>
<td>1.0000000</td>
<td>1.0000000</td>
<td>1.0000000</td>
<td>1.0000000</td>
<td>1.0000000</td>
</tr>
<tr>
<td>lam[4]</td>
<td>0.2421761</td>
<td>0.2609658</td>
<td>0.2688939</td>
<td>0.2790611</td>
<td>0.2991232</td>
</tr>
<tr>
<td>lam[5]</td>
<td>0.4469051</td>
<td>0.4657162</td>
<td>0.4755650</td>
<td>0.4851953</td>
<td>0.5027134</td>
</tr>
<tr>
<td>lam[6]</td>
<td>0.3966582</td>
<td>0.4212391</td>
<td>0.4343933</td>
<td>0.4491616</td>
<td>0.4672959</td>
</tr>
<tr>
<td>lam[7]</td>
<td>0.4936497</td>
<td>0.5062420</td>
<td>0.5130664</td>
<td>0.5201479</td>
<td>0.5340757</td>
</tr>
<tr>
<td>phi[1,1]</td>
<td>0.2335481</td>
<td>0.2394177</td>
<td>0.2429008</td>
<td>0.2461530</td>
<td>0.2528012</td>
</tr>
<tr>
<td>phi[2,1]</td>
<td>-0.3729033</td>
<td>-0.3634434</td>
<td>-0.3581288</td>
<td>-0.3526214</td>
<td>-0.3431805</td>
</tr>
<tr>
<td>phi[1,2]</td>
<td>-0.3729033</td>
<td>-0.3634434</td>
<td>-0.3581288</td>
<td>-0.3526214</td>
<td>-0.3431805</td>
</tr>
<tr>
<td>phi[2,2]</td>
<td>0.7655042</td>
<td>0.7867512</td>
<td>0.7966551</td>
<td>0.8074313</td>
<td>0.8273683</td>
</tr>
<tr>
<td>uniq[1]</td>
<td>0.0002435</td>
<td>0.0003018</td>
<td>0.0003713</td>
<td>0.0004437</td>
<td>0.0005796</td>
</tr>
<tr>
<td>uniq[2]</td>
<td>0.3217185</td>
<td>0.3295547</td>
<td>0.3344764</td>
<td>0.3394534</td>
<td>0.3494967</td>
</tr>
<tr>
<td>uniq[3]</td>
<td>0.0004859</td>
<td>0.0005651</td>
<td>0.0006602</td>
<td>0.0010475</td>
<td>0.0022210</td>
</tr>
<tr>
<td>uniq[4]</td>
<td>0.7435506</td>
<td>0.7625202</td>
<td>0.7724621</td>
<td>0.7828628</td>
<td>0.8032921</td>
</tr>
<tr>
<td>uniq[5]</td>
<td>0.9083740</td>
<td>0.9319449</td>
<td>0.9447111</td>
<td>0.9576743</td>
<td>0.9823026</td>
</tr>
<tr>
<td>uniq[6]</td>
<td>0.6754621</td>
<td>0.6920695</td>
<td>0.7014459</td>
<td>0.7104762</td>
<td>0.7290440</td>
</tr>
<tr>
<td>uniq[7]</td>
<td>0.5687867</td>
<td>0.5849021</td>
<td>0.5927457</td>
<td>0.6003985</td>
<td>0.6168780</td>
</tr>
</tbody>
</table>
> # posterior mean of factor correlation #
> options(scipen=999)
> mean(codaSamples1[[1]][,2]/sqrt(codaSamples1[[1]][,1]*codaSamples1[[1]][,4]))
> [1] 0.8136581
>
> # posterior SD of factor correlation #
> options(scipen=999)
> sd(codaSamples1[[1]][,2]/sqrt(codaSamples1[[1]][,1]*codaSamples1[[1]][,4]))
> [1] 0.005619092
>
> # 95% PPI of factor corr. #
> options(scipen=999)
> quantile(codaSamples1[[1]][,2]/sqrt(codaSamples1[[1]][,1]*codaSamples1[[1]][,4]),c(.025,.975))
> 2.5% 97.5%
> 0.8022316 0.8245019
>
> # posterior mean of factor correlation #
> options(scipen=999)
> mean(codaSamples1[[1]][,2]/sqrt(codaSamples1[[1]][,1]*codaSamples1[[1]][,4]))
> [1] 0.8136581
>
> # posterior SD of factor correlation #
> options(scipen=999)
> sd(codaSamples1[[1]][,2]/sqrt(codaSamples1[[1]][,1]*codaSamples1[[1]][,4]))
> [1] 0.005619092
>
> # 95% PPI of factor corr. #
> options(scipen=999)
> quantile(codaSamples1[[1]][,2]/sqrt(codaSamples1[[1]][,1]*codaSamples1[[1]][,4]),c(.025,.975))
> 2.5% 97.5%
> 0.8022316 0.8245019
Bayesian statistics represents a powerful alternative to frequentist (classical) statistics, and is therefore, controversial.

The controversy lies in differing perspectives regarding the nature of probability, and the implications for statistical practice that arise from those perspectives.

The frequentist framework views probability as synonymous with long-run frequency, and that the infinitely repeating coin-toss represents the canonical example of the frequentist view.

In contrast, the Bayesian viewpoint regarding probability was, perhaps, most succinctly expressed by de Finetti.
Probability does not exist.

- Bruno de Finetti
That is, probability does not have an objective status, but rather represents the quantification of our experience of uncertainty.

For de Finetti, probability is only to be considered in relation to our subjective experience of uncertainty, and, for de Finetti, uncertainty is all that matters.
That is, probability does not have an objective status, but rather represents the quantification of our experience of uncertainty.

For de Finetti, probability is only to be considered in relation to our subjective experience of uncertainty, and, for de Finetti, uncertainty is all that matters.

“The only relevant thing is uncertainty – the extent of our known knowledge and ignorance. The actual fact that events considered are, in some sense, determined, or known by other people, and so on, is of no consequence.” (pg. xi).
That is, probability does not have an objective status, but rather represents the quantification of our experience of uncertainty.

For de Finetti, probability is only to be considered in relation to our subjective experience of uncertainty, and, for de Finetti, uncertainty is all that matters.

“The only relevant thing is uncertainty – the extent of our known knowledge and ignorance. The actual fact that events considered are, in some sense, determined, or known by other people, and so on, is of no consequence.” (pg. xi).

The only requirement then is that our beliefs be coherent, consistent, and have a reasonable relationship to any observable data that might be collected.
The major advantages of Bayesian statistical inference over frequentist statistical inference are:

- Coherence
- Handling non-nested models
- Flexibility in handling complex data structures
- Inferences based on data actually observed
- Quantifying evidence
- Incorporating prior knowledge
The major advantages of Bayesian statistical inference over frequentist statistical inference are:

- **Coherence**
The major advantages of Bayesian statistical inference over frequentist statistical inference are:

1. Coherence
2. Handling non-nested models
Summarizing the Bayesian Advantage

The major advantages of Bayesian statistical inference over frequentist statistical inference are:

1. Coherence
2. Handling non-nested models
3. Flexibility in handling complex data structures
Summarizing the Bayesian Advantage

The major advantages of Bayesian statistical inference over frequentist statistical inference are:

1. Coherence
2. Handling non-nested models
3. Flexibility in handling complex data structures
4. Inferences based on data actually observed
Summarizing the Bayesian Advantage

- The major advantages of Bayesian statistical inference over frequentist statistical inference are

  1. Coherence
  2. Handling non-nested models
  3. Flexibility in handling complex data structures
  4. Inferences based on data actually observed
  5. Quantifying evidence
The major advantages of Bayesian statistical inference over frequentist statistical inference are:

1. **Coherence**
2. **Handling non-nested models**
3. **Flexibility in handling complex data structures**
4. **Inferences based on data actually observed**
5. **Quantifying evidence**
6. **Incorporating prior knowledge**
Subjective v. Objective Bayes

- See papers in *Bayesian Analysis, 1 (3), 2006.*
Subjective v. Objective Bayes

- See papers in *Bayesian Analysis, 1 (3), 2006*.

- Subjective Bayesian practice attempts to bring prior knowledge directly into an analysis. This prior knowledge represents the analysts (or others) degree-of-uncertainty.

- An analyst’s degree-of-uncertainty is encoded directly into the prior distribution, and specifically in the degree of precision around the parameter of interest.

- The advantages according to Press (2003) include
Subjective v. Objective Bayes

- See papers in *Bayesian Analysis, 1 (3), 2006*.

- Subjective Bayesian practice attempts to bring prior knowledge directly into an analysis. This prior knowledge represents the analysts (or others) degree-of-uncertainty.

- An analyst’s degree-of-uncertainty is encoded directly into the prior distribution, and specifically in the degree of precision around the parameter of interest.

- The advantages according to Press (2003) include
  - Subjective priors are proper
Subjective v. Objective Bayes

- See papers in *Bayesian Analysis, 1 (3), 2006.*

- Subjective Bayesian practice attempts to bring prior knowledge directly into an analysis. This prior knowledge represents the analysts (or others) degree-of-uncertainty.

- An analyst’s degree-of-uncertainty is encoded directly into the prior distribution, and specifically in the degree of precision around the parameter of interest.

- The advantages according to Press (2003) include
  1. Subjective priors are proper
  2. Priors can be based on factual prior knowledge
Subjective v. Objective Bayes

- See papers in *Bayesian Analysis, 1 (3), 2006*.

- Subjective Bayesian practice attempts to bring prior knowledge directly into an analysis. This prior knowledge represents the analysts (or others) degree-of-uncertainty.

- An analyst’s degree-of-uncertainty is encoded directly into the prior distribution, and specifically in the degree of precision around the parameter of interest.

- The advantages according to Press (2003) include:
  1. Subjective priors are proper
  2. Priors can be based on factual prior knowledge
  3. Small sample sizes can be handled.
The disadvantages to the use of subjective priors according to Press (2003) are
The disadvantages to the use of subjective priors according to Press (2003) are

- It is not always easy to encode prior knowledge into probability distributions.
The disadvantages to the use of subjective priors according to Press (2003) are:

1. It is not always easy to encode prior knowledge into probability distributions.

2. Subjective priors are not always appropriate in public policy or clinical situations.
The disadvantages to the use of subjective priors according to Press (2003) are

1. It is not always easy to encode prior knowledge into probability distributions.

2. Subjective priors are not always appropriate in public policy or clinical situations.

3. Prior distributions may be analytically intractable unless they are conjugate priors.
Within objective Bayesian statistics, there is disagreement about the use of the term “objective”, and the related term “non-informative”.

Specifically, there are a large class of so-called reference priors (Kass and Wasserman, 1996).

An important viewpoint regarding the notion of objectivity in the Bayesian context comes from Jaynes (1968).

For Jaynes, the “personalistic” school of probability is to be reserved for
Within objective Bayesian statistics, there is disagreement about the use of the term “objective”, and the related term “non-informative”.

Specifically, there are a large class of so-called *reference priors* (Kass and Wasserman, 1996).

An important viewpoint regarding the notion of objectivity in the Bayesian context comes from Jaynes (1968).

For Jaynes, the “personalistic” school of probability is to be reserved for

“...the field of psychology and has no place in applied statistics. Or, to state this more constructively, objectivity requires that a statistical analysis should make use, not of anybody’s personal opinions, but rather the specific factual data on which those opinions are based.” (pg. 228)
In terms of advantages Press (2003) notes that...
In terms of advantages Press (2003) notes that

Objective priors can be used as benchmarks against which choices of other priors can be compared
In terms of advantages Press (2003) notes that

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
In terms of advantages Press (2003) notes that

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. An objective prior provides results equivalent to those based on a frequentist analysis
In terms of advantages Press (2003) notes that

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. An objective prior provides results equivalent to those based on a frequentist analysis.

4. Objective priors are sensible public policy priors.
In terms of disadvantages of objective priors, Press (2003) notes that

1. Objective priors can lead to improper results when the domain of the parameters lies 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.
3. Expressing complete ignorance about a parameter via an objective prior leads to incorrect inferences about functions of the parameter.
In terms of disadvantages of objective priors, Press (2003) notes that

- Objective priors can lead to improper results when the domain of the parameters lie on the real number line.
In terms of disadvantages of objective priors, 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.
In terms of disadvantages of objective priors, 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.

3. Expressing complete ignorance about a parameter via an objective prior leads to incorrect inferences about functions of the parameter.
Kadane (2011) states, among other things:
Kadane (2011) states, among other things:

“The purpose of an algorithmic prior is to escape from the responsibility to give an opinion and justify it. At the same time, it cuts off a useful discussion about what is reasonable to believe about the parameters. Without such a discussion, appreciation of the posterior distribution of the parameters is likely to be less full, and important scientific information may be neglected.” (pg. 445)
The subjectivist school allows for personal opinion to be elicited and incorporated into a Bayesian analysis.

The subjectivist school places no restriction on the source, reliability, or validity of the elicited opinion.

The objectivist school views personal opinion as the realm of psychology with no place in a statistical analysis.

The objectivist school would require formal rules for choosing reference priors.
“Subjectivism” within the Bayesian framework runs the gamut from the elicitation of personal beliefs to making use of the best available historical data available to inform priors.

I agree with Jaynes (1968) – namely that the requirements of science demand reference to “specific, factual data on which those opinions are based” (pg. 228).

This view is also consistent with Leamer’s hierarchy of confidence on which priors should be ordered.

We may refer to this view as an evidence-based form of subjective Bayes which acknowledges (1) the subjectivity that lies in the choice of historical data; (2) the encoding of historical data into hyperparameters of the prior distribution; and (3) the choice among competing models to be used to analyze the data.
What if factual historical data are not available?

Berger (2006) states that reference priors should be used “in scenarios in which a subjective analysis is not tenable”, although such scenarios are probably rare.

The goal, nevertheless, is to shift the practice of Bayesian statistics away from the elicitation of personal opinion (expert or otherwise) which could, in principle, bias results toward a specific outcome, and instead move Bayesian practice toward the warranted use prior objective empirical data for the specification of priors.

The specification of any prior should be explicitly warranted against observable, empirical data and available for critique by the relevant scholarly community.
To conclude, the Bayesian school of statistical inference is, arguably, superior to the frequentist school as a means of creating and updating new knowledge in the social sciences.

An evidence-based focus that ties the specification of priors to objective empirical data provides stronger warrants for conclusions drawn from a Bayesian analysis.

In addition predictive criteria should always be used as a means of testing and choosing among Bayesian models.

As always, the full benefit of the Bayesian approach to research in the social sciences will be realized when it is more widely adopted and yields reliable predictions that advance knowledge.
Tusen takk