Bayesian Estimation of Regression Models

In: The SAGE Handbook of Regression Analysis and Causal Inference

By: Susumu Shikano
Edited by: Henning Best & Christof Wolf
Pub. Date: 2013
Access Date: October 29, 2018
Publishing Company: SAGE Publications Ltd
City: London
Print ISBN: 9781446252444
Online ISBN: 9781446288146
DOI: http://dx.doi.org/10.4135/9781446288146
Print pages: 31-54

Bayesian Estimation of Regression Models

SusumuShikano

Introduction to the Method

Bayesian statistics provide an alternative to the maximum likelihood principle for the estimation of regression models. While maximum likelihood assumes a true value for a parameter and describes its estimator, including the standard error, the Bayesian statistics framework assumes a certain distribution as an inherent part of the parameter. This seemingly trivial difference has some important consequences in parameter estimation of statistical models, including regression models. To make this point clear, we explicate the difference between Bayesian statistics and maximum likelihood in what follows.

Basic Idea of Bayesian Estimation

In the maximum likelihood framework, a single data set is interpreted as one realization of potential data sets. This is very simple to understand if one has sampled data from a larger population. From the population, different samples can be drawn and the sample at hand is only one of them. Even if one has non-sampled data one can interpret a data set if one views its variables of interest as random variables. Based on this view of the data set, maximum likelihood focuses the probability of data (x) in hand given certain parameters (θ): $\Pr(x|\theta)$. This conditional probability is called likelihood. The goal of maximum likelihood estimation is to find the best possible parameter values which maximize $\Pr(x|\theta)$. For this purpose, one needs to have a function of the parameter given the data. This is obtained by $L(\theta|x) = \Pr(x|\theta)$ based on Bayes' theorem:

$$\Pr(\theta|x) = \frac{\Pr(\theta)\Pr(x|\theta)}{\Pr(x)}. \qquad (3.1)$$

Inference based on maximum likelihood assumes that $\Pr(\theta)$ and $\Pr(x)$ are constant for all possible estimates of θ. Therefore, maximization of $L(\theta|x) = \Pr(x|\theta)$ and maximization of $\Pr(\theta|x)$ are equivalent to each other. Based on this logic, the equivalent likelihood function is constructed by simplifying $\Pr(x|\theta)$. Thus we search for a value for θ which maximizes $L(\theta|x) = \Pr(x|\theta)$.

In contrast to maximum likelihood, Bayesian statistics relaxes the assumption that $\Pr(\theta)$ is constant for all possible estimates of θ. Since $\Pr(x)$ is constant for all possible estimates of θ, equation (3.1) can be reformulated as

$$\Pr(\theta|x) \propto \Pr(\theta)\Pr(x|\theta). \qquad (3.2)$$

Using this equation, Bayesian statistics calculates Pr(θ|*x*). One might wonder what Pr(θ) means. This is the probability of certain estimates for θ prior to data collection. Therefore, it is called prior probability. In contrast, Pr(θ|*x*) is the probability of certain estimates for θ given the data collected, the so-called posterior probability. Accordingly, Bayes' theorem expresses the process in which a certain prior probability about parameter estimates is updated by the likelihood (Pr(*x*|θ)) to obtain the posterior probability of the parameter. Here, one can assume a constant prior probability for all parameter estimates as in maximum likelihood. However, it is not necessary, as one can also use different prior probabilities for different parameter estimates. The corresponding information can come from the existing literature, past data analysis or the researcher's subjective belief. The most important characteristic of Bayesian inference is the use of prior information. Bayesian inference can be understood as a process of updating the prior information to the posterior information using the data collected.

How to Derive the Posterior Probability

To get an intuition, we begin with a very simple example model with a single parameter. Suppose we are interested in the unemployment rate and would like to obtain an estimate *p*. Furthermore, assume we found that one out of 10 persons investigated is currently unemployed ($x = 1$, $N = 10$). Which parameter is most likely to generate the data? In the maximum likelihood framework $p = 0.1$ is most likely to obtain data with one unemployed person out of 10. Using the binomial distribution, the likelihood of data for different parameter values is quite simple to calculate:

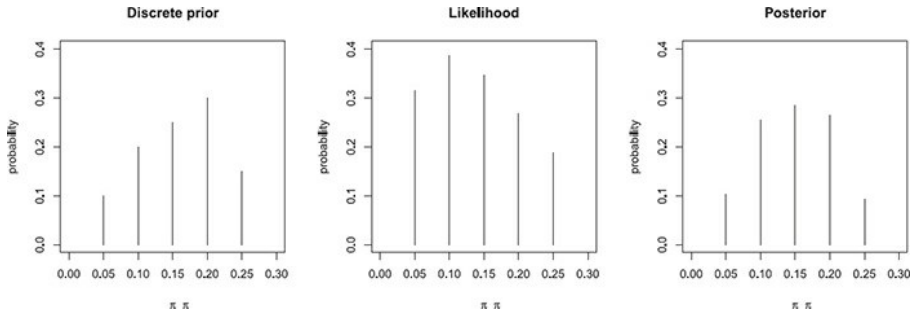
$$\Pr(x = 1|p = 0.1, N = 10) = \binom{10}{1} 0.1^1 (1 - 0.1)^{10-1} \approx 0.387. \qquad (3.3)$$

This means that, given $p = 0.1$, we would obtain the data ($x = 1$, $N = 10$) with probability 38.7%. If we use another value for *p* instead of 0.1 we obtain a smaller value, that is, a lower probability of this figure being the unemployment rate.

For Bayesian inference we first need to specify some prior information which represents our pre-existing knowledge, information or beliefs. Let us assume that we know for certain reasons the unemployment rate is either 0.05, 0.1, 0.15, 0.2 or 0.25. Furthermore, we also know with a probability of 10% that the unemployment rate equals 0.05. Analogously, we also know that the probability of the other values is 20%, 25%, 30% and 15%. This prior knowledge can be graphically displayed as in the left-hand panel of Figure 3.1. In the next step, we calculate the likelihood given our data using the binomial distribution for the possible values for π: {0.05, 0.1, 0.15, 0.2, 0.25}. The calculated likelihood values are presented in the middle panel of Figure 3.1. Note that the likelihood for $p = 0.1$ is identical with the result of equation (3.3). That is, the middle panel of Figure 3.1 verifies that $p = 0.1$ has the maximal value of the likelihood.

Bayesian inference, however, proceeds further to calculate the posterior information using both prior and likelihood. The calculation is simply done based on Bayes' theorem (equation (3.1)). We can multiply the prior probability and likelihood for each value of *p*, which corresponds to to the right-hand side of equation (3.2), $\Pr(\theta)\Pr(x|\theta)$. These probabilities are standardized to obtain the posterior probability whose sum is one. The posterior information thus obtained is shown in the right-hand panel of Figure 3.1.

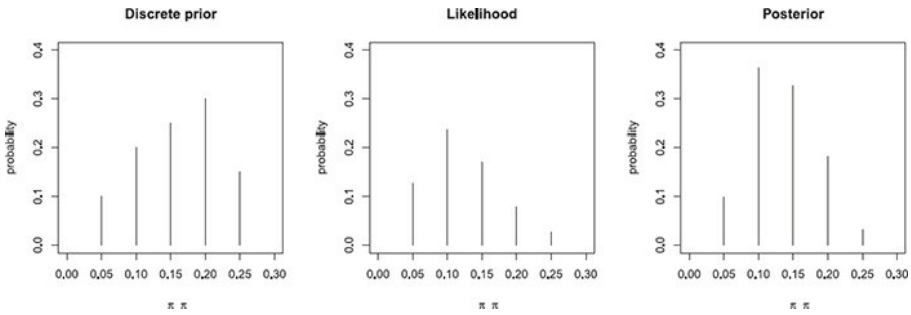
Figure 3.1 Discrete prior, likelihood and posterior (data with one person unemployed out of 10)



We can compare the posterior information with the prior and the likelihood. In the prior information, we thought that the unemployment rate was most probably 0.2. According to the data collected, by contrast, the unemployment rate is most likely to be 0.1. Using this information from the data, we updated our prior information to the posterior according to which the unemployment rate is most probably 0.15. We can observe here that the posterior information is a kind of mixture of the prior and likelihood. The way to systematically mix both pieces of information is provided by the Bayesian framework of inference.

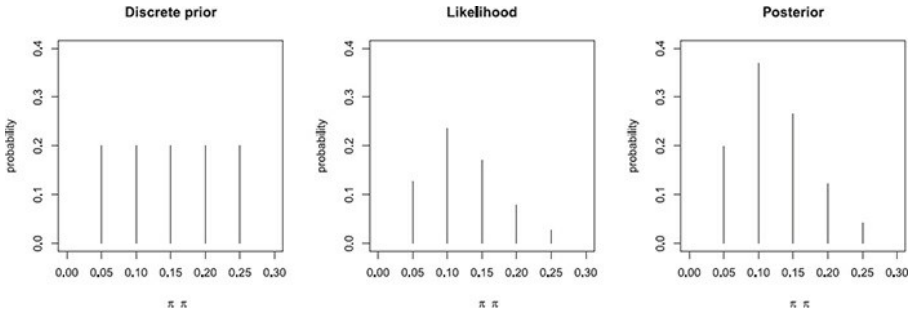
Let us now take another data set with three out of 30 persons unemployed. The unemployment rate in the data is 0.1, which is identical to the previous data (one unemployed out of 10). However, the new data set has more respondents. Using this data set, we can repeat the procedure introduced above. The corresponding prior, likelihood and posterior can be found in Figure 3.2. According to the new posterior, the unemployment rate of 0.1 is most probable. That is, the information provided by the data has a greater impact on the posterior than in the last analysis. This is because the new data has more information (30 observations instead of 10).

Figure 3.2 Discrete prior, likelihood and posterior (data with three persons unemployed out of 30)



Finally, we use yet another kind of prior information. We have no idea which of the possible unemployment rates is more or less probable. Correspondingly, we assign the probability of 20% to each of five possible unemployment rates (left-hand panel of Figure 3.3). The impact of the data information is even stronger than in the previous analysis, so that the unemployment rate of 0.1 is much more probable in the posterior information. This is because the prior has less information than in the previous cases.

Figure 3.3 Discrete prior, likelihood and posterior (uninformative prior, data with three persons unemployed out of 30)

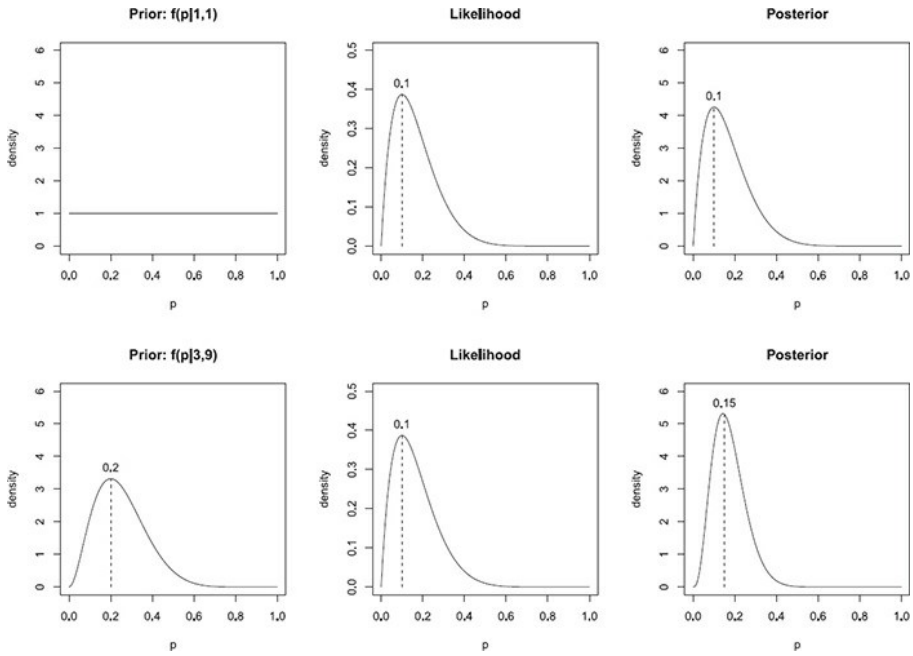


Note that the description of the posterior above was concentrated on the most likely unemployment rate, that is, the mode of the posterior distribution. This is, however, only one possible way to describe a posterior. It is also possible to use further point estimates (mean, median, etc.) as well as interval estimates (credible intervals). This might sound similar to the maximum likelihood estimator; however, one has to be careful about the difference between frequentist and Bayesian inference in interpretation. While maximum likelihood postulates a certain true value for a parameter, Bayesian inference views the parameter as random quantity. This difference can be illustrated by the frequentist confidence interval and the Bayesian credible interval. If one constructs a 95% confidence interval using a data set one postulates that the data is one of a large number of potential samples. Depending on the observed sample, one can construct many further possible 95% confidence intervals, and 95% of them should include the true value of the parameter. In this frequentist view, the randomness comes

from the observation process, including sampling and measurement. In contrast, Bayesian inference views the parameter itself as random entity and does not postulate a single true value behind the parameter. Thus, the Bayesian credible interval expresses the randomness of the parameter *per se*, and a 95% credible interval simply means that the parameter takes the value in the interval with 95% probability.

The posterior thus far has been quite simple to calculate since we have a prior information that π can take only five values: {0.05, 0.1, 0.15, 0.2, 0.25}. This kind of prior is called discrete prior. In contrast, we can also have a prior which can take all values on a certain scale: a continuous prior. In our example, we usually know that the unemployment rate can take all values on the scale between 0 and 1. To express this kind of prior, we can use, for example, a beta distribution. The beta distribution has two parameters, which enables us to express different shapes of a prior distribution. If we have no idea about the unemployment rate, all values between 0 and 1 are equally probable. This can be expressed as $f_{\beta}(p|1, 1)$, as shown in the upper left-hand panel of Figure 3.4. As can clearly be seen from the form of the distribution, we call this kind of prior a flat prior. If we have a continuous prior we also have to consider the likelihood for all possible values of π. This can be realized by using the density function of the binomial distribution. For the data with one person out of 10 unemployed, we have the likelihood function given in the upper middle panel of Figure 3.4. The posterior is calculated analogously to the examples above with the discrete prior. That is, we calculate the density for all possible p by multiplying the corresponding prior and likelihood. For the flat prior, this is quite simple. Since all values of p are equally probable in the prior, the posterior mode is identical to that of the likelihood (see the upper right-hand panel of Figure 3.4). We can now repeat this exercise using another prior. Assume that we have the prior given in the lower left-hand panel of Figure 3.4. That is, we believe that the unemployment rate of 0.2 is more probable than any other values. In contrast, we assume we have the same data as in the last exercise, so that the likelihood function is identical (the lower middle panel of Figure 3.4). In this case, too, the posterior can be calculated by multiplying the prior and likelihood for all π. Fortunately, it is known that the posterior is also a beta distribution and the parameters can simply be calculated using the prior parameter of the prior and some information from the data (lower right-hand panel of Figure 3.4).

Figure 3.4 Continuous prior, likelihood and posterior (data with one person unemployed out of 10)



The property that the prior and posterior have the same probability form is quite important since this offers the analytical way to derive the posterior. Note that this property depends on the form of the distribution used to calculate the likelihood. In the examples above, we used a binomial distribution to obtain the likelihood function. To this kind of distribution, it is known that a beta distribution is always *conjugate*. That is, if the prior has the form of a beta distribution and the likelihood comes from a binomial distribution, the posterior also takes the form of a beta distribution. This conjugacy is important in multiple senses. First, as we have seen in the example above, the calculation of the posterior is simplified through distribution parameters. If we have no conjugate prior we have to rely on numerical solution using Markov chain Monte Carlo (MCMC) techniques which will be introduced later. Second, the posterior obtained in a Bayesian inference can later be used as prior for a further inference using the same framework. Imagine that we obtained the posterior in the lower right-hand panel of Figure 3.4. Thereafter, we collected another data set. We do not use the same prior as in the previous analysis again, but its posterior as prior since we have updated our knowledge using the last data set. In this way, Bayesian inference enables us to combine information from different data sets.

To summarize, Bayesian inference provides a systematic way to integrate the prior information and data collected into the posterior information. In this process, data gains more impact on the posterior if the data contains more information (e.g. more observations) or if the prior has less information (e.g. larger dispersion). This applies not only to the simple examples presented here but also to regression analysis and further kinds of analysis with more complex statistical

models.

Bayesian Estimation of Regression Models

The examples in the previous subsection are quite simple in the sense that we had only one parameter. In estimation of regression models, we have more parameters to estimate. Even if a simple bivariate linear regression model with only one independent variable is estimated, we still have three parameters: two regression coefficients (intercept and slope) and error variance. That is, we need to build the joint posterior distribution using the prior and the data.

As in the previous subsection, we first need to specify the form of prior information for the coefficients and error variance. A conjugate prior is widely used for the reasons discussed in the previous subsection. In particular, the normal inverse gamma distribution is known to be conjugate to the likelihood based on a multivariate normal distribution. The normal inverse gamma distribution constitutes the joint distribution for the regression coefficients and the error variance. The marginal distribution of the coefficients corresponds to a multivariate *t* distribution, that of the error variance to an inverse gamma distribution.

By multiplying the prior with the likelihood which is known from the conventional likelihood-based methods, we can obtain the posterior distribution. Due to the conjugacy we again obtain a normal inverse gamma distribution for the posterior. As shown above, if we use a flat prior for the regression coefficients and variance error, the posterior corresponds to the likelihood so that we obtain the same result for the posterior mode as in conventional maximum likelihood estimation. If we use a specific prior, in contrast, the posterior becomes a mixture of the prior and the likelihood. Here, too, the rule is same as in the previous subsection. If the data has more observations or the prior is more widely dispersed the data has a greater impact on the posterior so that the likelihood and the posterior are similar, and vice versa.

If no conjugate prior is known for the likelihood function or if one wishes to specify a non-conjugate prior, obtaining the posterior is difficult for two reasons. First, to obtain certain posterior information (e.g. expectation) we need not only to multiply the likelihood and prior but also to divide it by Pr(*y*) (see equation (3.1)). Calculation of this denominator requires an integral which often has no analytical solution. Second, in most applications, including regression models, we have a statistical model with multiple parameters. That is, the posterior distribution is a multidimensional joint distribution. However, we are generally interested in the marginal distribution for individual parameters which can be obtained by integrating out the other parameters. Similarly, we often have no analytical solution for integrals of this kind. These are the reasons why Bayesian inference was for a long time limited mainly to conjugacy analysis and consequently had only limited impact on applied social sciences. For several

decades, however, rapid growth of computational capacity has led to an alternative numerical toolkit for obtaining the posterior information: the Markov chain Monte Carlo techniques. These techniques enable us to randomly draw numbers from the target joint posterior distribution (Monte Carlo techniques). The joint posterior distribution is reached by a random process with the Markov property (Markov chain) which is specified by the available information. In particular, two specific classes of MCMC techniques are widely used: Gibbs sampling and the Metropolis–Hastings algorithm. For Gibbs sampling, one needs the conditional posterior distribution for individual parameters. By successively applying the conditional posteriors, one can obtain the joint posterior distribution. In contrast, the Metropolis–Hastings algorithm requires no conditional posteriors. Instead, by comparing the product of the likelihood and the prior among different sets of parameters, the Markov chains go through the parameter space and reach the joint distribution. More detailed descriptions of both methods can be found in the next section.

Independently of the methods to obtain the posterior distribution, its interpretation is quite straightforward. The posterior distribution obtained is simply described using for instance its mean, standard deviation or certain percentiles. If one has a hypothesis about the parameter, the posterior probability that the hypothesis is correct is simply calculated. For example, we have an alternative hypothesis that an independent variable should have a positive impact on the dependent variable. From the corresponding marginal posterior distribution, we can obtain the probability that the regression coefficient of an independent variable has positive sign.

Mathematical Foundations

This section begins with a more formal presentation of Bayesian inference. In this context, it will also be clear why conjugacy is an important concept and when we need MCMC methods. The section then goes on to introduce the regression model in a formal way.

Bayesian Inference

In terms of densities, Bayes' theorem states that

$$f(\theta|x) = \frac{f(x|\theta)f(\theta)}{f(x)} = \frac{f(x|\theta)f(\theta)}{\int f(x|\theta)f(\theta)\,\mathrm{d}\,\theta}, \qquad (3.4)$$

where the denominator $f(x) = \int f(x|\theta)f(\theta)d\theta$ is a normalizing constant which ensures that $\int f(\theta|x)d\theta$.

Based on $f(\theta|x)$, we can derive different kinds of point estimators. The posterior expectation is the expected value of the posterior distribution, which is given by

$$E(\theta|x) = \int \theta f(\theta|x)\,\mathrm{d}\,\theta. \qquad (3.5)$$

The posterior median is given by

$$\hat{\theta} = \text{Med}(\theta|x) \equiv \int_{-\infty}^{\hat{\theta}} f(\theta|x)\,\mathrm{d}\,\theta = 0.5. \qquad (3.6)$$

The posterior mode is given by

$$\text{Mod}(\theta|x) = \arg\max_{\theta} f(\theta|x)\,. \qquad (3.7)$$

Note that this posterior mode is in a close relationship to the point estimate based on maximum likelihood. Maximum likelihood finds an estimator which maximizes $L(\theta|x) = f(x|\theta)$. We would obtain the same estimator if $f(\theta)$ were constant, since $f(\theta|x) \propto f(x|\theta)f(\theta)$.

In addition to the point estimators, we can also obtain interval estimators. A credible interval at level 1 − α is defined by

$$\int_{t_a}^{t_b} f(\theta|x)\,\mathrm{d}\,\theta = 1 - \alpha, \qquad (3.8)$$

where t_a, $t_b \in \Theta$. According to this definition, θ lies in the interval between t_a and t_b with posterior probability 1 − α. Note the difference between this and the frequentist confidence interval, which would include the true value of θ with probability 1 − α if one has a large number of repeated samples. In contrast, the Bayesian credible interval views θ as a random quantity.

Note, further, that equation (3.8) has no unique solution for t_a and t_b. Usually one uses the α/2 and 1-α/2 quantile of the posterior distribution for t_a and t_b, respectively. Alternatively, one can calculate the highest posterior density (HPD) interval. Accordingly, an interval $C = [t_a, t_b] \subset \Theta$ is the 1 − α HPD interval for θ if equation (3.8) holds and

$$f(\theta|x) \geq f(\tilde{\theta}|x), \quad \forall\theta \in C \text{ and } \tilde{\theta} \notin C \qquad (3.9)$$

That is, the HPD interval includes the parameter values with the highest posterior density.

Analytical Solution with Conjugate Priors

In the previous subsection, we introduced different Bayesian estimators. Among them, the posterior mode is relatively easy to calculate since we can ignore the denominator of equation (3.4) and simply find the parameter values which maximize the numerator. In contrast, the posterior expectation requires information $f(x) = \int f(x|\theta)f(\theta)\,\mathrm{d}\,\theta$ as well as integration over θ as given by equation (3.5). In most cases the integrals in $f(x)$ and/or calculation of the posterior expectation have no solution in closed form. The exception is those instances where the prior distribution is conjugate to the distribution defining the likelihood.

A prior distribution is conjugate to the distribution defining the likelihood if the derived posterior

density has the same functional form as the prior density with different parameter values. For example, a beta distribution is the conjugate to a binomial distribution. Using this property, we derived the posterior distribution in Figure 3.4. In general, the binomial distribution models the process in which N experiments yielding success with probability p are independently repeated. The probability that we obtain x successes can be calculated:

$$f(x|p,N) = \binom{N}{x} p^x (1-p)^{N-x}. \tag{3.10}$$

Note that here we have two kinds of parameter, p and N, and we are generally interested in p. For this p we can assume a beta distribution as the prior distribution, whose density is given by

$$f(p|\alpha,\beta) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)}, \tag{3.11}$$

$$B(\alpha,\beta) = \int_0^1 p^{\alpha-1}(1-p)^{\beta-1}\,\mathrm{d}p = \frac{\Gamma(\alpha)\,\Gamma(\beta)}{\Gamma(\alpha+\beta)}. \tag{3.12}$$

$B(\alpha, \beta)$ is a normalizing constant which ensures that $\int f(p|\alpha,\beta)\,\mathrm{d}p = 1$. Further, the expected value of the beta distribution is also known:

$$E(p|\alpha,\beta) = \int_0^1 p f(p|\alpha,\beta)\,\mathrm{d}p = \frac{\alpha}{\alpha+\beta}. \tag{3.13}$$

The mode of the beta distribution is

$$\mathrm{Mod}(p|\alpha,\beta) = \arg\max_p f(p|\alpha,\beta) = \frac{\alpha-1}{\alpha+\beta-2}. \tag{3.14}$$

Both parameters of the beta distribution, α and β, are called hyperparameters since they are the parameters of a distribution which describes another distribution's parameter (here p).

By substituting both likelihood and prior into Bayes' theorem, we obtain:

$$\begin{aligned} f(p|x) &= \frac{f(x|p,N)f(p|\alpha,\beta)}{f(x)} \\ &= \frac{\binom{N}{x} p^x (1-p)^{N-x} \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)}}{f(x)} \\ &= \frac{\binom{N}{x} B^{-1}(\alpha,\beta)\, p^{\alpha+x-1}(1-p)^{\beta+N-x-1}}{f(x)} \\ &= \frac{p^{\alpha+x-1}(1-p)^{\beta+N-x-1}}{B(x+\alpha, N-x+\beta)}. \end{aligned} \tag{3.15}$$

This means that the posterior distribution is also a beta distribution with parameters α + x and β

+ $N - x$. Therefore the expectation and mode can be easily obtained via equations (3.13) and (3.14):

$$E(p|x) = \frac{\alpha + x}{\alpha + \beta + N}, \quad (3.16)$$

$$\text{Mod}(p|x) = \frac{\alpha + x - 1}{\alpha + \beta + N - 2}. \quad (3.17)$$

We can now return to the example in Figure 3.4. In the example we have data $x = 1$ and $N = 10$. In the upper panels, we had a beta prior with $\alpha = \beta = 1$. As is clear in the figure, the prior is completely flat between 0 and 1. The posterior expectation and mode can be calculated as follows:

$$E(p|x) = \frac{\alpha + x}{\alpha + \beta + N} = \frac{x + 1}{N + 2}, \quad (3.18)$$

$$\text{Mod}(p|x) = \frac{\alpha + x - 1}{\alpha + \beta + N - 2} = \frac{1 + x - 1}{1 + 1 + N - 2} = \frac{x}{N}. \quad (3.19)$$

While the posterior mode coincides with the maximum likelihood estimates (x/N) the posterior expectation differs slightly and shrinks towards the prior expectation ($\frac{1}{2}$).

Conjugate priors exist for other likelihood functions. In particular, it is known that a conjugate prior exists for likelihood functions which belong to the exponential family. Table 3.1 lists conjugate priors for likelihood functions frequently used in social science research. Among these conjugate distributions, the normal inverse gamma distribution is relevant for the linear regression model since the model generally has a likelihood function using the normal distribution with unknown mean and variance.

Table 3.1 Likelihood functions with conjugate priors

Likelihood	Model parameter	Conjugate prior
Bernoulli	p: probability	Beta
Binomial	p: probability	Beta
Negative binomial	p: probability	Beta
Poisson	γ: rate	Gamma
Multinomial	**p**: probability vector	Dirichlet
Normal	μ: mean	Normal
Normal	μ^2: variance	Inverse gamma
Normal	σ^2: variance	Scaled inverse chi-squared
Normal	μ and σ^2	Normal inverse gamma

Linear Regression Model with Conjugate Priors

A linear regression model consists of a dependent variable **y** and some independent variables **X**. Through a linear combination of **X** using β we can predict the dependent variable. The residual ε is assumed to be distributed normal with mean zero and variance $\sigma^2\mathbf{I}$:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{3.20}$$

$$\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}). \tag{3.21}$$

This can also be expressed as

$$(\mathbf{y}|\boldsymbol{\beta}, \sigma^2, \mathbf{X}) \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}), \tag{3.22}$$

with **I** as the identity matrix. That is, our dependent variable is assumed to be a random variable from normal distribution with unknown mean and variance. The corresponding likelihood function is

$$f(\mathbf{y}|\boldsymbol{\beta}, \sigma^2) = \prod_i^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y_i - \mathbf{X}\boldsymbol{\beta})'(y_i - \mathbf{X}\boldsymbol{\beta})}{2\sigma^2}\right\}. \tag{3.23}$$

It is known that this likelihood function has several conjugate prior distributions. Among others, the normal inverse gamma distribution is typically used to derive the posterior. The normal inverse gamma distribution is a product of a normal distribution for β and an inverse gamma distribution for σ^2. An inverse gamma distribution has two parameters, the scale parameter *a* and the shape parameter *d*, and its probability density function is given by

$$f_{\Gamma^{-1}}(\sigma^2|a, d) = \frac{a^d}{\Gamma(d)}\sigma^{2(-d-1)} \exp\left(-\frac{a}{\sigma^2}\right). \tag{3.24}$$

Multiplication with a univariate normal distribution yields a normal inverse gamma distribution:

$$\begin{aligned} f_{N-\Gamma^{-1}}(\beta, \sigma^2|\mu, \lambda, a, d) &= f_N(\beta|\mu, \sigma^2/\lambda) f_{\Gamma^{-1}}(\sigma^2|a, d) \\ &= \frac{\sqrt{\lambda}}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{\lambda(\beta - \mu)^2}{2\sigma^2}\right\} \frac{a^d}{\Gamma(d)}\sigma^{2(-d-1)} \exp\left(-\frac{a}{\sigma^2}\right) \\ &= \frac{\sqrt{\lambda}}{\sqrt{2\pi\sigma^2}} \frac{a^d}{\Gamma(d)} \left(\frac{1}{\sigma^2}\right)^{d+1} \exp\left\{-\frac{\lambda(\beta - \mu)^2 + 2a}{2\sigma_0^2}\right\}. \end{aligned} \tag{3.25}$$

In the linear regression with *k* independent variables (including a constant), β is a *k* × 1 parameter vector and $\sigma^2\sigma$ is its variance–covariance matrix. For this reason, the prior is more precisely a *multivariate* normal inverse gamma distribution:

$$\begin{aligned} f_{N-\Gamma^{-1}}(\boldsymbol{\beta}, \sigma^2|\boldsymbol{\mu}, \boldsymbol{\Sigma}, a, d) &= f_N(\boldsymbol{\beta}|\boldsymbol{\mu}, \sigma^2\boldsymbol{\Sigma}) f_{\Gamma^{-1}}(\sigma^2|a, d) \\ &= \frac{\sqrt{\lambda}}{\sqrt{2\pi\sigma^k|\boldsymbol{\Sigma}|}} \frac{a^d}{\Gamma(d)} \left(\frac{1}{\sigma^2}\right)^{d+1} \\ &\quad \times \exp\left\{-\frac{(\boldsymbol{\beta} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta} - \boldsymbol{\mu}) + 2a}{2\sigma^2}\right\}. \end{aligned} \tag{3.26}$$

Now we specify the prior as follows:

$$f(\boldsymbol{\beta}, \sigma^2) = f_{N-\Gamma^{-1}}(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0, a_0, d_0). \quad (3.27)$$

It is known that multiplication of prior and likelihood yields another multivariate normal inverse gamma distribution:

$$f(\boldsymbol{\beta}, \sigma^2|\mathbf{y}) = f_{N-\Gamma^{-1}}(\boldsymbol{\beta}^*, \boldsymbol{\Sigma}^*, a^*, b^*), \quad (3.28)$$

with

$$\boldsymbol{\beta}^* = (\boldsymbol{\Sigma}_0^{-1} + \mathbf{X}'\mathbf{X})^{-1}(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\beta}_0 + \mathbf{X}'\mathbf{y}), \quad (3.29)$$

$$\boldsymbol{\Sigma}^* = (\boldsymbol{\Sigma}_0^{-1} + \mathbf{X}'\mathbf{X})^{-1}, \quad (3.30)$$

$$a^* = a + \frac{1}{2}(\boldsymbol{\beta}_0'\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\beta}_0 + \mathbf{y}'\mathbf{y} - \boldsymbol{\beta}^{*\prime}\boldsymbol{\Sigma}^{*-1}\boldsymbol{\beta}^*), \quad (3.31)$$

$$d^* = d_0 + \frac{n}{2}. \quad (3.32)$$

Note that this is a joint posterior distribution of β and σ^2. In general, σ^2 is a nuisance parameter and we are only interested in the marginal posterior distribution of β, which can be obtained by integrating out σ^2. It is known that this marginal posterior follows a multivariate Student *t* distribution with v = *n* − *k* degrees of freedom:

$$f(\boldsymbol{\beta}|\mathbf{y}) = \int f(\boldsymbol{\beta}, \sigma^2|\mathbf{y})\, \mathrm{d}\sigma^2 = f_t(\nu, \boldsymbol{\beta}^*, \boldsymbol{\Sigma}^*). \quad (3.33)$$

Here, we can compare the posterior of β with the estimates via ordinary least squares and maximum likelihood. As can be seen in Chapter 2 of this volume, ordinary least squares and maximum likelihood give the same point estimate of β:

$$\boldsymbol{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \quad (3.34)$$

The posterior approximately yields this result if *n* approaches infinity:

$$\lim_{n\to\infty} \boldsymbol{\beta}^* = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \quad (3.35)$$

This is because $\boldsymbol{\Sigma}_0^{-1}$ has less weight as *n* increases. For the same reason, increasing $\boldsymbol{\Sigma}_0$ also leads to the same estimate as ordinary least squares and maximum likelihood.

Another important aspect is multicollinearity. Differently from ordinary least squares and maximum likelihood, it is also possible to obtain the posterior even in situations of perfect multicollinearity. In the case of perfect multicollinearity, that is, if there is an exact linear relationship among independent variables, **X'X** does not have full rank and invertibility. Therefore, we can calculate neither $(\mathbf{X'X})^{-1}$ nor β. In calculating β^*, by contrast, we do not have to invert **X'X**. Instead, if $\boldsymbol{\Sigma}_0^{-1} + \mathbf{X}'\mathbf{X}$ has full rank we can obtain β^*.

Another Approach: A Numerical Solution Via Gibbs Sampling

In the previous subsection, we derived the posterior distribution of β and σ^2 jointly using the conjugate normal inverse gamma prior. Alternatively, we can derive the posterior of β and σ_2 separately. That is, we specify a (multivariate) normal distribution for β and an inverse gamma for σ^2:

$$f(\boldsymbol{\beta}) = f_{MN}(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0),$$
$$f(\sigma^2) = f_{\Gamma^{-1}}(a_0, d_0).$$

Both priors are conjugate to the likelihood function from a normal distribution, and the posteriors are given by

$$f(\boldsymbol{\beta}|\sigma^2, \mathbf{y}) = f_{MN}(\boldsymbol{\beta}^*, \boldsymbol{\Sigma}^*), \quad (3.36)$$
$$f(\sigma^2|\boldsymbol{\beta}, \mathbf{y}) = f_{\Gamma^{-1}}(a^*, d^*), \quad (3.37)$$

with

$$\boldsymbol{\beta}^* = (\sigma^{-2}\mathbf{X}'\mathbf{X} + \boldsymbol{\Sigma}_0^{-1})^{-1}(\sigma^{-2}\mathbf{X}'\mathbf{y} + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\beta}_0), \quad (3.38)$$
$$\boldsymbol{\Sigma}^* = (\sigma^{-2}\mathbf{X}'\mathbf{X} + \boldsymbol{\Sigma}_0^{-1})^{-1}, \quad (3.39)$$
$$a^* = a_0 + \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \quad (3.40)$$
$$d^* = d_0 + \frac{n}{2}. \quad (3.41)$$

Note that the derived posteriors of individual parameters are conditioned by the other parameter values. In contrast, we are generally interested in the marginal probability which gives an average picture of the posterior distribution over all possible parameter combinations. For this purpose, we utilize a numerical method called Gibbs sampling, one of the MCMC methods, instead of integrating out the other parameters.

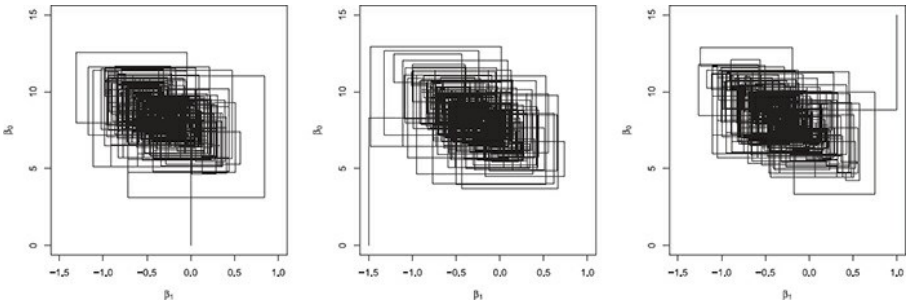
Here, we wish to generate random draws from a joint posterior f (β, $\sigma^2|\mathbf{y}$). Gibbs sampling can be described by the following iterative steps for t = 1, …, T, with $\beta^{(t)}$ and $\sigma 2(t)$ being the values generated in the tth iteration and $\beta^{(0)}$ and $\sigma 2(0)$ arbitrary selected initial values:

1. Draw a random number from f ($\beta|\sigma^{2(t-1)}$, **y**) and save as $\beta^{(t)}$.
2. Draw a random number from f ($\sigma^2|\beta^{(t)}$, **y**) and save as $\sigma_{2(t)}$.
3. Go to step 1 for the (t + 1)th iteration.

By repeating these steps we can build a Markov chain. This kind of Markov chain (Gibbs chain) is known to have a unique invariant distribution. That is, after a large enough number of

iterations chains converges to a stable state independently of their initial states. The invariant distribution obtained corresponds to the desired joint distribution. This is best illustrated by Figure 3.5 which shows three Markov chains for β with different initial values. These chains are set up by a simple bivariate regression model which will be described later in the example section. For simple statistical models of this kind, the Markov chain converges very quickly. In Figure 3.5 we can also see that all three chains reached a common region after just a few steps, that is, the posterior distribution. By summarizing the information for the individual parameters after convergence, we can also describe the individual marginal distribution (see Figure 3.7).

Figure 3.5 Three Markov chains with different initial values



An important issue in practical Bayesian analysis is to evaluate whether the Markov chain reached convergence. While there are several methods and criteria suggested, it is most important to run not just one, but multiple chains with different initial values and to carry out a visual inspection. If we can confirm that multiple chains are converged in a certain region and well mixed we discard the part of Markov chain before convergence. This discarded phase of the Markov chain is called *burn-in*. After burn-in we should run the Markov chain further to obtain samples from the joint posterior. If the chain after the burn-in is not run long enough the joint posterior cannot be captured well enough. One simple way to evaluate whether one has run the Markov chain long enough is to observe whether the Markov chain changes the form of the captured posterior substantially. If it does not, we can stop running the Markov chain and begin summarizing the posterior.

One of the important advantages of Gibbs sampling and other MCMC techniques is their simplicity in describing posterior distributions. To calculate the posterior expectation one does not need to integrate, but simply take the average value of the random draws:

$$E(\theta|x) = \frac{1}{T}\sum_{t}\theta^{(t)}. \qquad (3.42)$$

The posterior median is given by the median value. The credible interval can be constructed using quantiles. Only the posterior mode cannot be calculated in a straightforward way. For this

purpose, one first needs a kernel density estimate based on which one can estimate the mode.

A More General Numerical Solution: Metropolis–Hastings Algorithm

Gibbs sampling is a powerful technique which enables us to obtain the joint and marginal posterior in a very simple way. The algorithm, however, requires the full set of the conditional posterior for all parameters. This is not always the case. For example, a binary logistic regression with data of the form $y_i = 0$ or 1 has likelihood function

$$f(\mathbf{y}|\boldsymbol{\beta},\mathbf{X}) = \prod_{i=1}^{n}(\text{logit}^{-1}(\mathbf{x}_i'\boldsymbol{\beta}))^{y_i}(1-\text{logit}^{-1}(\mathbf{x}_i'\boldsymbol{\beta}))^{1-y_i}, \qquad (3.43)$$

For this likelihood function, no conjugate prior distribution is known. If we take a multivariate normal prior for β,

$$f(\boldsymbol{\beta}) = f(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0), \qquad (3.44)$$

we can write down the posterior:

$$f(\boldsymbol{\beta}|\mathbf{y}) \propto \prod_{i=1}^{n}(\text{logit}^{-1}(\mathbf{x}_i'\boldsymbol{\beta}))^{y_i}(1-\text{logit}^{-1}(\mathbf{x}_i'\boldsymbol{\beta}))^{1-y_i} \times \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\right\}. \qquad (3.45)$$

However, a closed form of the solution is known for neither the joint nor the conditional posterior. Therefore, neither conjugacy analysis nor Gibbs sampling proves useful for this statistical model. Fortunately, we can still use another class of MCMC algorithm and a general form of Gibbs sampling: the Metropolis–Hastings algorithm.

Let us assume that we wish to generate random draws from a posterior $f(\beta|\mathbf{y})$. The Metropolis–Hastings algorithm can be described by the following iterative steps for $t = 1, \ldots, T$, with $\beta_{(t)}$ being the vector of generated values in the tth iteration and $\beta^{(0)}$ arbitrary selected initial values:

1. Set $\beta = \beta_{(t-1)}$.
2. Generate new candidate values β' from a proposal distribution $q(\beta_0|\beta)$.
3. Calculate $\alpha = \min\left(1, \frac{f(\beta'|y)q(\beta|\beta')}{f(\beta|y)q(\beta'|\beta)}\right) = \min\left(1, \frac{f(y|\beta')f(\beta')q(\beta|\beta')}{f(y|\beta)f(\beta)q(\beta'|\beta)}\right)$.
4. Update $\beta^{(t)} = \beta_0$ with probability α. Otherwise set $\beta^{(t)} = \beta$.

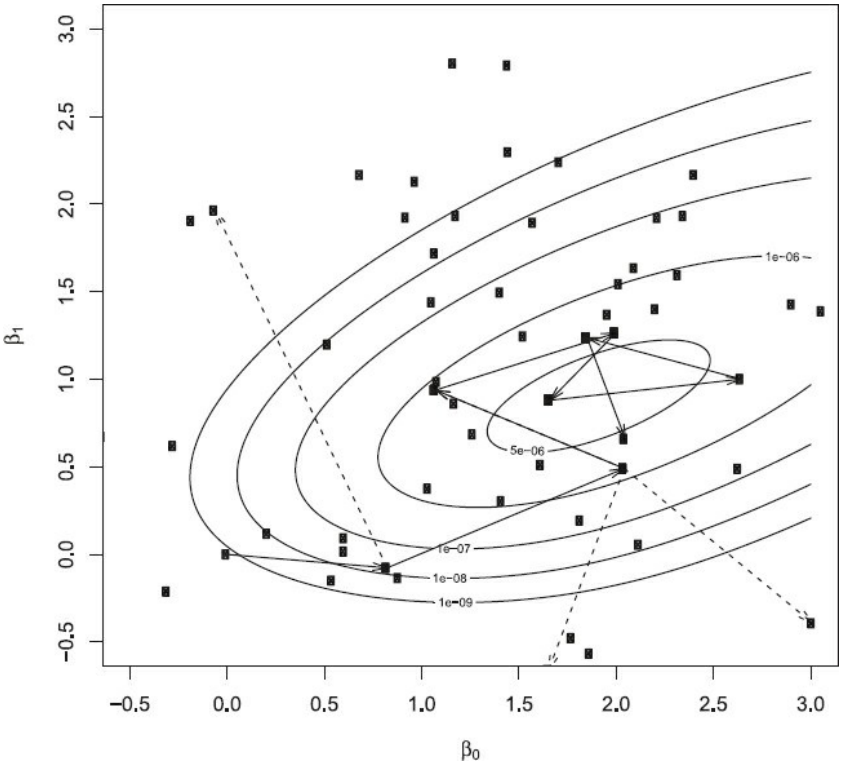
The underlying idea is quite simple. If the candidate values have larger density in $f(\beta|\mathbf{y})$ they are selected; if not, depending on the relationship of the density of the current and candidate values, one set of both is drawn. This process can be illustrated by Figure 3.6, which presents an example case of a binary logit model with one covariate. Therefore, we only need to estimate

two parameters, β_0 and β_1. We specify $\{\beta_0, \beta_1\} = \{0, 0\}$ as initial values of the Markov chain. Furthermore, we use a normal distribution with variance 1 as the proposal distribution independently for β_0 and β_1:

$$\beta_0' \sim N(\beta_0, 1), \quad (3.46)$$

$$\beta_1' \sim N(\beta_1, 1). \quad (3.47)$$

Figure 3.6 Graphical illustration of the Metropolis–Hastings algorithm. The solid arrows are the accepted moves of the Markov chain. The dashed arrows are (a part of) the rejected moves. The circles are rejected candidate proposals. The contour lines gives the product of the likelihood and the prior (equation (3.45))



The normal distribution has an advantage due to its symmetrical form which results in $q(\beta_0|\beta) = q(\beta|\beta_0)$. Therefore, the calculation of α can be simplified:

$$\alpha = \min\left(1, \frac{f(\mathbf{y}|\boldsymbol{\beta}')f(\boldsymbol{\beta}')}{f(\mathbf{y}|\boldsymbol{\beta})f(\boldsymbol{\beta})}\right). \quad (3.48)$$

The proposal distribution randomly generated a candidate vector β_0 =(0.823, −0.077). α can now be calculated based on equation (3.45). This candidate vector has a higher density, resulting in α = 1. In this case, the Markov chain moves to this candidate vector with probability 1. In the next iteration, the normal distributions centred on the current status of the Markov chain, (0.823, −0.077), generated a new candidate vector, β, =(-0.062, 1.963). As clearly seen in Figure 3.6, whose contour lines give the density level, the new candidate vector's density is

much lower than that of the current status. Correspondingly, α is very low (2.662×10^{-6}). This does not exclude that the Markov chain moves to this candidate vector. However, in this example case, the Markov chain stayed in the current status. That is, $\beta_0 = (0.823, -0.077)$ was drawn for the second time. In the next iteration, the proposal distributions generated a new candidate vector whose density is higher than that of the current status. Therefore, the Markov chain moved to the new vector. After the process described above is iterated a large number of times we obtain the samples from the joint posterior distribution. Figure 3.6 gives an example Markov chain with only 60 iterations. The solid dots are the selected β and the circles are the rejected candidate vectors.

The beauty of this algorithm is at least twofold. First, it is known that the Metropolis–Hastings algorithm will converge to its equilibrium distribution independently of the proposal distribution *q* being used. That is, we do not necessarily use the normal distribution as in the example above. We could also use a normal distribution with a different variance. However, a very small variance of the proposal distribution can slow down convergence. Second, one does not need to have the full density function with the normalizing constant since $\frac{f(\boldsymbol{\beta}'|\mathbf{y})}{f(\boldsymbol{\beta}|\mathbf{y})}$ cancels out the normalizing constant. Indeed, we had no full description of the posterior distribution in equation (3.45) which ignores the denominator of the posterior ($f(\mathbf{y})$) and also some irrelevant component of the prior $f(\beta)$. Independently of these components, α can be calculated and the Markov chain can proceed.

Figure 3.6 may remind some readers of the iterative algorithm (see Chapter 2 in this volume) in the maximum likelihood estimation framework. Here, we have to note that the iterative algorithm in maximum likelihood and the MCMC techniques in the Bayesian inference have two different goals. The goal of maximum likelihood estimation is to find the set of parameter values which maximizes the likelihood. Therefore, the search of the iterative algorithm ends in the maximum of the likelihood surface. In contrast, MCMC techniques aim to draw samples from the target posterior distribution. Thus, there is no clear end-point of the Markov chain. Furthermore, a chain can also sometimes move away from the maximum of the posterior density surface. This can be also seen in Figure 3.6. This is because parameter values with lower density can be drawn so long as they have a certain density.

Example Analysis

In this section we apply the method sketched above to a bivariate regression model. As data we use the European Social Survey (ESS) which consists of multiple rounds of cross-sectional data. By using this data it is also demonstrated how results from the previous round can be

integrated to the analysis of the current round.

In particular, we regress the opinion concerning European integration as a specific issue attitude on the left–right scale as a more general ideological orientation. The corresponding variables are measured by using the following questionnaire items:

- European integration (euftf) – 'Now thinking about the European Union, some say European unification should go further. Others say it has already gone too far. Using this card, what number on the scale best describes your position?' [0: Unification has already gone too far, …, 10: Unification should go further].
- Ideological orientation (lrscale) – 'In politics people sometimes talk of "left" and "right". Using this card, where would you place yourself on this scale, where 0 means the left and 10 means the right?'

The model parameters are estimated by using German respondents who neither are EU citizens nor were born in Germany. The reason for using this special group is to see whether the ideological orientation can have an impact on the specific issue attitude when the issue has less relevance. Thus, the analysis can provide an interesting test concerning the impact of ideology. However, the problem is that we do not have many respondents in the individual surveys: the second round has 57 respondents, third 43 respondents, and the fourth 40 respondents.[1] These figures reduced further to 38, 33 and 28 after listwise deletion due to missing values.

Before we proceed to the Bayesian inference, we first check the estimation results using conventional ordinary least squares in Table 3.2. Accordingly, all three estimates of the impact of ideological orientation on attitudes to the EU have negative sign, which means that more left-oriented respondents support further European integration, and vice versa. These effects have, however, relatively large standard errors, so that none of them is significant at the 5% level. Therefore, we cannot make any statements concerning the impact of ideology based on these data. If we look at the number of observations, however, we also realize that individual rounds each provide quite small pieces of information. Indeed, if we pooled all three rounds into one regression model the effect is significant. Another alternative is Bayesian inference which is useful for combining information from individual rounds.

Table 3.2 Estimates via OLS

	2nd round		3rd round		4th round		Pooled	
	Est.	SE	Est.	SE	Est.	SE	Est.	SE
β_0	8.216	1.402	8.754	2.097	7.594	2.372	8.633	1.025
β_1	−0.288	0.320	−0.618	0.390	−0.383	0.509	−0.528	0.213
R^2	0.022		0.075		0.021		0.059	
n	38		33		28		99	

What can we do if we apply Bayesian inference to the same data? One possible approach is that we begin with a flat prior belief and sequentially update our belief using the data from individual rounds. In other words, with no idea at first about whether the ideological orientation has an impact on attitudes to the EU, we can collect data and update our belief about the existence of the effect.

Conjugacy Analysis of the Second Round Data with Flat Prior

At the beginning of the analysis, we assume we have no information about the impact of ideological orientation. Therefore, we specify a conjugate prior using a normal inverse gamma distribution $f_{N-\Gamma^{-1}}(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0, a, d)$ with the following parameter values:

$$\boldsymbol{\beta}_0 = (0,0), \tag{3.49}$$

$$\boldsymbol{\Sigma}_0 = \begin{pmatrix} 10000 & 0 \\ 0 & 10000 \end{pmatrix}, \tag{3.50}$$

$$a = 0.0001, \tag{3.51}$$

$$d = 0.0001. \tag{3.52}$$

Substituting these hyperparameter values and information from data into equations (3.29)–(3.32), we obtain the following parameters of the posterior:

$$\boldsymbol{\beta}^* = (8.216, -0.288), \tag{3.53}$$

$$\boldsymbol{\Sigma}^* = \begin{pmatrix} 0.239 & -0.051 \\ -0.051 & 0.012 \end{pmatrix}, \tag{3.54}$$

$$a^* = 148.160, \tag{3.55}$$

$$d^* = 19.000. \tag{3.56}$$

Note that β^* is very similar to the ordinary least squares point estimates. This is because we specified a very flat prior distribution.

An inverse gamma distribution with parameter values a and d has expectation $a/(d-1)$. Therefore:

$$E(\sigma^2|\mathbf{y}) = \frac{148.160}{19.000 - 1} = 8.231. \tag{3.57}$$

Thus, the posterior variance–covariance matrix of β has expectation

$$E(\sigma^2|\mathbf{y}) \cdot \boldsymbol{\Sigma}_0 = \begin{pmatrix} 1.965 & -0.423 \\ -0.423 & 0.102 \end{pmatrix}. \quad (3.58)$$

The square root of the diagonal elements is (1.402, 0.320). This corresponds to the standard error of the ordinary least squares estimates.

Drawing the Posterior Using Gibbs Sampling

In this subsection, we construct the same posterior distribution using a numerical method, namely Gibbs sampling. As in the conjugacy analysis above, we again assume we have no information about the impact of ideological orientation. Here, however, we specify our prior using a multivariate normal and an inverse gamma distribution instead of the normal inverse gamma distribution. The parameters of both distributions are the same as in equations (3.49) to (3.52).

The conditioned posteriors for β and σ^2 correspond to equations (3.36) and (3.37), respectively. By using these posteriors we set up the Gibbs sampling. As initial values we use the following three sets of values:

$$\{\beta_0, \beta_1, \sigma^2\} = \{0, 0, 1\},\ \{0, -1.5, 2\},\ \{15, 1, 3\}.$$

That is, we run three Markov chains. Figure 3.5 traces the Markov chains of β (β_0 and β_1). The three chains started from different initial values; however, they converged after a few steps to a certain common region. This common region corresponds to the joint posterior distribution of β_0 and β_1. The individual points in the Markov chains correspond to the samples from the joint posterior distribution. Therefore, we can simply summarize the information of the sampled data from the posterior. Before summarizing the data, we discarded the first 1000 iterations as burn-in. Thereafter, we ran a further 2000 iterations to sample from the posterior distribution.

Figure 3.7 shows the density distributions for the individual parameters, which visualize the corresponding marginal posterior distribution. We can also use the mean, standard deviation or quantiles to obtain further information about the posterior which is presented in Table 3.3. The right-hand half of the same table presents the corresponding summary statistics of the posterior distributions which were derived in the conjugacy analysis in the previous subsection. The results are almost identical, which means that the Gibbs sampling worked very well. Slight differences are due to the random draws in the Gibbs sampling. In the previous subsection we have already found that the conjugate posterior is almost identical to the ordinary least squares estimates (Table 3.2), which is of course also the case for the posterior via Gibbs sampling. This does not mean, however, that we have to make exactly same statements using ordinary least squares and Bayesian estimation. To see this, we again look at Figure 3.7 which also presents

the interval between the 5th and 95th percentiles. This is not called a 90% confidence interval, but a 90% credible interval, and its interpretation is not the same. The confidence interval gives information about how frequently the true parameter value is inside the interval in repeated experiments or equivalent (frequentist probability). In contrast, the posterior distribution expresses our belief about the parameter with specific uncertainty. Therefore, we can simply state that the parameter value is in the credible interval with probability 90%. Consequently, we can also calculate the probability that a certain parameter lies in a certain interval. The middle panel of Figure 3.7 also gives the probability of β_1 having a negative value. By using this information we can also state that right-wing ideology leads to an anti-European attitude with 81% probability. This stands in clear contrast to the conventional inference where we can only either accept or reject certain hypotheses.

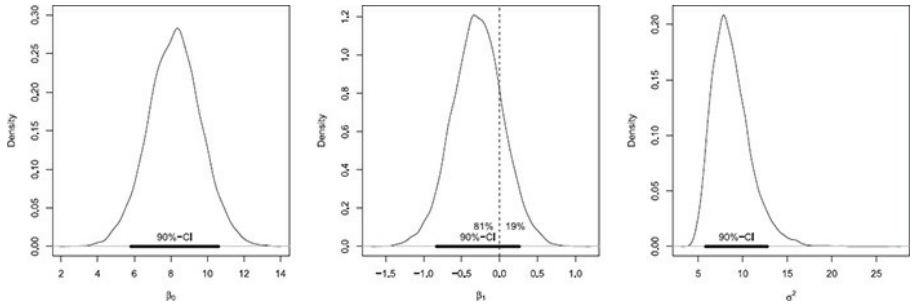
Table 3.3 Summary statistics of the posterior distributions (second round data)

	Gibbs sampling				Conjugacy analysis			
	Est.	SD	5%	95%	Est.	SD	5%	95%
β_0	8.223	1.456	5.819	10.568	8.216	1.402	5.849	10.583
β_1	−0.290	0.331	−0.834	0.264	−0.288	0.320	−0.828	0.253
σ^2	8.732	2.205	5.817	12.844	8.231	1.996	5.551	11.908

Updating Belief Using Data from Further Rounds

In the analysis of the second round data in the previous two subsections we derived the posterior using certain flat priors. That is, we assumed we had no clear belief about the parameter values of the statistical model of interest. Having done the first analysis, however, we formed certain beliefs which we can now use as priors in further analyses. In particular, the ESS data allows us to estimate the same statistical model using the third and fourth round data.

Figure 3.7 (Marginal) posterior distribution



Here, we can exploit the advantage of conjugacy analysis. Due to the conjugacy of the normal inverse gamma distribution to the normal likelihood function, we can use the parameters of the posterior distribution as those of the prior in further analysis. In the analysis of the third round data, we can substitute the posterior's parameter values in equations (3.53)–(3.56) as prior

parameter into equations (3.29)–(3.32). This results in the following parameter values of the new posterior distribution which again has the form of the normal inverse gamma distribution:

$$\boldsymbol{\beta}^* = (8.900, -0.553), \tag{3.59}$$

$$\boldsymbol{\Sigma}^* = \begin{pmatrix} 0.122 & -0.024 \\ -0.024 & 0.005 \end{pmatrix}, \tag{3.60}$$

$$a^* = 386.625, \tag{3.61}$$

$$d^* = 35.500. \tag{3.62}$$

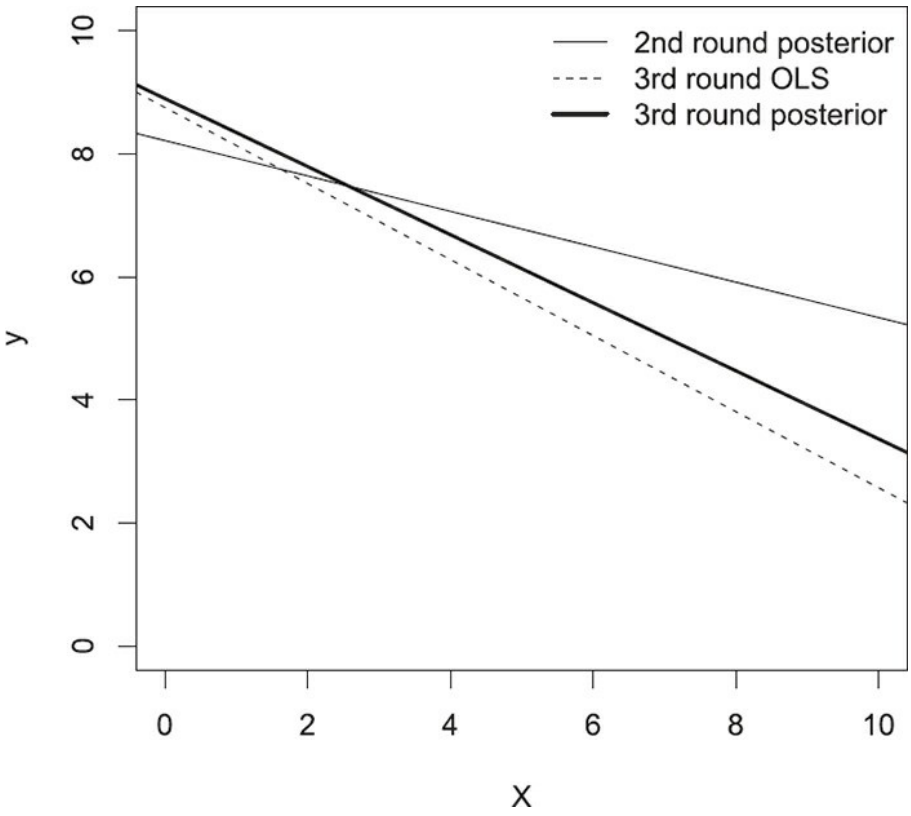
Using these parameter values, we can calculate the summary statistics of the (marginal) posterior distribution (the left-hand half of Table 3.4). If we look at β_1, which is substantively of most interest to us, we find a smaller impact in the posterior mean (-0.553) than in the ordinary least squares estimate (-0.618). This is because we used a prior distribution with a mean of −0.288. While we find a smaller magnitude of the effect in the mean, its uncertainty is relatively small. While the standard error in the corresponding ordinary least squares estimation was 0.390, the standard deviation of the posterior is only 0.241. This is because we have the prior as an information source in addition to the third round data. Therefore, we can be more certain regarding our results. Consequently, the probability that β_1 has negative sign increases from the previous analysis to 99%.

Table 3.4 Summary statistics of the posterior distributions (third and fourth round data)

	3rd round				4th round			
	Est.	SD	5%	95%	Est.	SD	5%	95%
β_0	8.900	1.172	6.914	10.886	8.632	1.025	6.885	10.381
β_1	−0.554	0.241	−0.962	−0.145	−0.527	0.213	−0.892	−0.164
σ^2	11.206	1.936	8.435	14.700	10.192	1.479	8.023	12.831

Here one might wonder that the posterior mean of β_0 (8.900) is higher than that both of the prior (8.223) and the ordinary least squares estimate (8.754). Indeed, we learned at the beginning of this chapter that the posterior is a mixture of the prior and the likelihood. However, we are now not working on the posterior with a single parameter, but on the joint posterior with multiple parameters. That is, the posterior of a parameter depends on the prior and the likelihood as well as the other parameters. In this particular case, the posterior after the second round (i.e. the prior for the third round analysis) has a relatively flat regression line with a higher value for the constant (the thin solid line in Figure 3.8). That is, the dependent variable is in general at a higher level. The ordinary least squares result of the third round analysis shows, by contrast, a steeper regression line (the dashed line in Figure 3.8). Now the posterior based on the third round data is a mixture of this steeper regression line and a generally high level of the dependent variable (the thick solid line in Figure 3.8). And both factors raise the value of constant.

Figure 3.8 Estimated regression lines from second and third round data

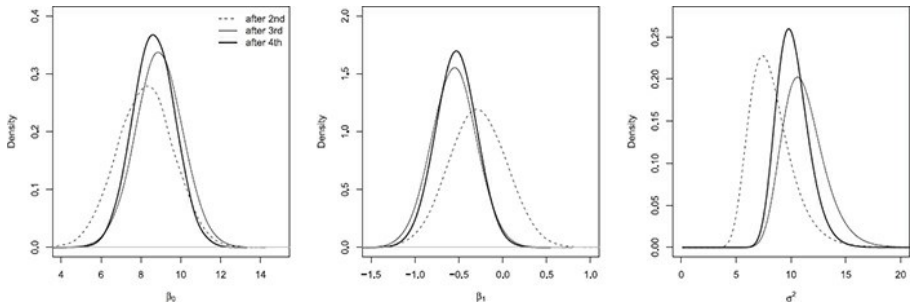


After updating our belief using the third round data, we can now further update our belief using the fourth round data. The procedure is analogous to the analysis of the third round data. The only difference is that we use the posterior from the third round data as prior. The results appear in the right-hand half of Table 3.4. If we compare the posterior of β_1 with the corresponding ordinary least squares estimate, it is clear that the prior has a stronger impact on the posterior than the data. This is reasonable if we consider the following points. First, the fourth round data provides only 28 observations. Second, the standard error of β_1 in the ordinary least squares estimation is larger than those in the earlier rounds. Furthermore, the prior (posterior from the third round) is smaller than the standard errors from ordinary least squares estimation as discussed above. For these reasons, the prior information is much more weighted than the information from the fourth round data.

Figure 3.9 gives a graphical presentation of the development of the posteriors. In the posteriors of β_0 and β_1 we can clearly observe that the uncertainty about the parameter was reduced in the course of analysis. Concerning β_1 we can be quite sure after the analysis of fourth round data that the ideological orientation has an impact on the EU attitude. In contrast, σ^2 has a different development. After building the first posterior using the second round, the next posterior has a larger mean and also a larger variance. This suggests that the prior from the

second round and the data from the third round data provided conflicting pieces of information concerning σ^2. Therefore, after the analysis we could not be as sure as after the first analysis. However, the further data from the fourth round provided more information supporting a smaller value of σ^2, so that our posterior became similar again to that after the first analysis.

Figure 3.9 Sequential updating of the (marginal) posterior distributions



Careful readers will surely have realized that the posterior from the fourth round data coincides with the pooled ordinary least squares result (Table 3.2). This is always the case if we conduct this kind of sequential updating of the posterior. To illustrate this relationship, we can return to the normal inverse gamma posterior (equations (3.29)–(3.32)). Here, we focus on β and σ, but the same holds also for the other parameters (*a* and *d*). Denote the data from the second, third and fourth round by $\mathbf{y}_2$, $\mathbf{y}_3$, $\mathbf{y}_4$, $\mathbf{X}_2$, $\mathbf{X}_3$ and $\mathbf{X}_4$. We first derive the posterior based on a flat prior and second round data, which corresponds to the ordinary least squares estimate:

$$\boldsymbol{\beta}_2^* = (\mathbf{X}_2'\mathbf{X}_2)^{-1}(\mathbf{X}_2'\mathbf{y}_2), \tag{3.63}$$

$$\boldsymbol{\Sigma}_2^* = (\mathbf{X}_2'\mathbf{X}_2)^{-1}. \tag{3.64}$$

We substitute these parameters into β_0 and σ_0 in equations (3.29) and (3.30) to obtain the posterior after the third round data analysis:

$$\boldsymbol{\beta}_3^* = (\underbrace{\mathbf{X}_2'\mathbf{X}_2}_{\boldsymbol{\Sigma}_0^{-1}} + \mathbf{X}_3'\mathbf{X}_3)^{-1}(\underbrace{(\mathbf{X}_2'\mathbf{X}_2)}_{\boldsymbol{\Sigma}_0^{-1}}\underbrace{(\mathbf{X}_2'\mathbf{X}_2)^{-1}(\mathbf{X}_2'\mathbf{y}_2)}_{\boldsymbol{\beta}_0} + \mathbf{X}_3'\mathbf{y}_3) \tag{3.65}$$

$$= (\mathbf{X}_2'\mathbf{X}_2 + \mathbf{X}_3'\mathbf{X}_3)^{-1}(\mathbf{X}_2'\mathbf{y}_2 + \mathbf{X}_3'\mathbf{y}_3), \tag{3.66}$$

$$\boldsymbol{\Sigma}_3^* = (\underbrace{\mathbf{X}_2'\mathbf{X}_2}_{\boldsymbol{\Sigma}_0^{-1}} + \mathbf{X}_3'\mathbf{X}_3)^{-1}. \tag{3.67}$$

Analogously, we can obtain the posterior after the fourth round data analysis:

$$\boldsymbol{\beta}_4^* = (\mathbf{X}_2'\mathbf{X}_2 + \mathbf{X}_3'\mathbf{X}_3 + \mathbf{X}_4'\mathbf{X}_4)^{-1}(\mathbf{X}_2'\mathbf{y}_2 + \mathbf{X}_3'\mathbf{y}_3 + \mathbf{X}_4'\mathbf{y}_4), \tag{3.68}$$

$$\boldsymbol{\Sigma}_4^* = (\mathbf{X}_2'\mathbf{X}_2 + \mathbf{X}_3'\mathbf{X}_3 + \mathbf{X}_4'\mathbf{X}_4)^{-1}. \tag{3.69}$$

Here, it is clear that the sequential posterior updating corresponds exactly to the pooled analysis using the OLS or maximum likelihood framework. Of course, this is not the case if we use certain informative priors for the analysis of the second round data.

From this observation another advantage of the Bayesian inference is clear in that we can also estimate statistical models with a smaller number of observations. This is because the prior belief serves as a further observation in estimation. Furthermore, Bayesian inference calculates parameters' posterior distribution directly instead of constructing an estimator. Thus it is free from the conventional asymptotic theory. For these reasons, the Bayesian inference is also possible for smaller numbers of observations than would be sufficient for ordinary least squares and/or maximum likelihood.

Caveats and Frequent Errors

In using a Bayesian inferential framework one always has to be aware of the underlying conception about prior beliefs, data, probability and the manner of inference. In particular, one should be careful in interpreting the estimation results. Bayesian inference, for example, does not know the null hypothesis significance test. Therefore, it is completely wrong to discuss the significance level, rejection or acceptance of the null hypothesis, etc. Further, we have already discussed the difference between the confidence interval in conventional inference and the Bayesian credible interval. The interpretation of the credible interval *per se* is quite straightforward, and one should be rather careful in the interpretation of the conventional confidence interval. However, one still has to be careful not to make significance-test-style statements using a credible interval.

The use of prior information and its systematic integration into the posterior is one of the most important features of Bayesian inference. Again, one should be aware of the conception of priors and the idea of Bayesian updating. The basic idea of Bayesian inference is to update our prior beliefs using new information from the data. A belief is prior here in the sense that we have the information prior to the data collection. Therefore, the choice of prior distribution cannot be reasoned by information from the data which is used to update the prior information. A prior can be also specified after data collection. This is the case even in many concrete contexts. However, the legitimation of the choice of certain priors can be never made using information from the data. If one is sufficiently conscious of this point the criticism against the Bayesian manner of inference for its use of prior information does not have to be taken seriously. It is obvious that the choice of priors is not grounded by the data. This kind of a priori decision is, however, also made by conventional inferential statistics. To conduct statistical analyses, we have to make a series of assumptions, such as the choice of independent variables and the distribution form of the stochastic term, in order to identify the statistical model. If these assumptions can be made we can also make assumptions about parameter values – of course, if we have good reasons for that.

Even if one is convinced of the use of prior information, it is difficult to bring the information into a certain probability form. As one possibility this chapter has presented an example which uses past estimates as prior. If one has no information from other comparable analyses one needs to build one's own prior distribution in a discrete or continuous form. In this case, one can have a certain choice between different probability forms. If this is the case and one cannot decide on a specific probability form with a strong argument, one should use multiple priors to check whether the results based on the posterior are strongly affected by the choice of prior. This kind of check is called a sensitivity test.

In our favorable case with prior information from the other analysis, one still has to be careful about the previous data used to obtain the prior. If a strong measurement and/or sampling error exists which systematically affects both past data and current data, the posterior result can contain strongly boosted errors. The example above might also suffer from this problem. We used only those respondents born outside Germany who do not possess EU citizenship. If there are any systematic sampling errors which are relevant to the relationship of interest to us, our posterior provides information biased towards the error. In contrast, if we can expect the errors to be balanced out in the course of the updating process we do not have to worry about this problem.

While one can never overemphasize the role of prior information, it is not the only advantage of the Bayesian inference. In particular, Bayesian inference can work with a wide range of statistical models, thanks in particular to the MCMC techniques. For example, some statistical models may cause difficulties in finding the maximum likelihood due to some local maxima or a very flat likelihood surface. This difficulty is less relevant for Bayesian inference due to the random walk property of MCMC. Another advantage concerns missing-data problems. In the framework of MCMC one can treat missing data as a random entity and a parameter to be estimated. Furthermore, hierarchical models which have been increasingly utilized in social sciences fit the basic idea of Bayesian inference. In hierarchical models parameters at one level are conceptualized as random variables and modelled by higher-level parameters, which corresponds to the idea of hyperparameters. And this idea is quite simple to implement in the MCMC framework.

In using MCMC, we can never be careful enough in evaluating the convergence of Markov chains. This topic is crucial since random draws in the burn-in phase can offer completely different pictures of the target posterior. For this reason running multiple chains with significantly different initial values is inevitable. At the same time, the phase after the burn-in is also important. In this phase, we can collect the information from the posterior distribution. If

one runs the Markov chain not long enough the collected information may be biased due to random draws. In particular, such information as quantiles requires a sufficiently large number of iterations. In principle, the more iterations after burn-in the more accurate the information obtained about the posterior. To evaluate whether one has reached a certain precision through sampling after the burn-in phase one can utilize the *Monte Carlo error*, which is based on the idea of the variance in autocorrelated samples.

Further Reading

The aim of this chapter is to provide a first look at Bayesian inference for the readers who are unfamiliar with it and to make the relevant literature accessible. For this reason, the topics dealt with in this chapter are deliberately limited. In particular, this chapter only presents the basic idea of Bayesian inference and its application to simple linear regression using conjugacy analysis and Gibbs sampling. Beyond these topics, there are a large number of statistical models and their accompanying topics. For those who have just started with this chapter, some journal articles are recommended as further reading.

Gill (1999) elegantly summarizes the differences between conventional and Bayesian inference. Simon Jackman, a political scientist who helped disseminate Bayesian inference in the social sciences, has written several articles which provide a short introduction. Among them, Jackman (2000) delves deeply into the relationship between maximum likelihood and Bayesian inference and gives a good introduction to the Markov chain Monte Carlo method. Western and Jackman (1994) is also a well-written introductory article whose focus is more on its application and advantages in the social sciences. Due to their limited length, journal articles always have deficits in terms of broad coverage of topics. Having gained a feel for Bayesian statistics, those who wish to extend their knowledge of Bayesian analysis would be well advised to read Gelman and Hill (2007). This book is about the regression model and multilevel modelling, but also covers further important and relevant topics such as missing values, and offers many practical tips on programming and computation.

For those who would like to extend their knowledge further, Gill (2002) and Gelman et al. (2003) deal extensively with diverse topics in Bayesian inference. These books are very technical and more appropriate for those who are familiar with Bayesian inference and have specific problems in mind.

Note

1 The first round is not considered here since this round does not include the questionnaire item concerning European integration.

References

Gelman, A., **Carlin, J. B.**, **Stern, H. S.**, and **Rubin, D. B.** (**2003**). *Bayesian Data Analysis*, 2nd edn. Boca Raton, FL: Chapman & Hall/CRC.

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

Gill, J. (**1999**). The insignificance of null hypothesis significance testing. *Political Research Quarterly*, 52(3), pp. 647–674. http://dx.doi.org/10.1177/106591299905200309

Gill, J. (**2002**). *Bayesian Methods: A Social and Behavioral Sciences Approach*. Boca Raton, FL: Chapman & Hall/CRC.

Jackman, S. (**2000**). Estimation and inference via Bayesian simulation: An introduction to Markov chain Monte Carlo. *American Journal of Political Science*, 44(2), pp. 375–404. http://dx.doi.org/10.2307/2669318

Western, B. and **Jackman, S.** (**1994**). Bayesian inference for comparative research. *American Political Science Review*, 88(2), pp. 412–423. http://dx.doi.org/10.2307/2944713

http://dx.doi.org/10.4135/9781446288146.n3