Bayesian Econometrics using BayES™

Grigorios Emvalomatis

February 21, 2020

The most up-to-date version of this document can be found at:
http://www.bayeconsoft.com/pdf/BayesianEconometricsUsingBayES.pdf

Contents

Preface

This document is intended to serve as an introductory textbook for a postgraduate or advanced undergraduate course on Bayesian econometrics or as a reference for the applied econometrician who never got exposed to the Bayesian approach. Although Bayesian econometrics is increasingly being used in applied research, programs of study in economics usually include courses only in frequentist econometrics, even if time allows for more than a single course. Apart from tradition, this bias towards the frequentist approach to statistical inference can be attributed to the lack of specialized software for Bayesian econometrics. This is changing rapidly, with mainstream econometric software packages incorporating Bayesian techniques and the emergence of new software packages that make application of Bayesian methods considerably easier.

This textbook aims at covering the basics of Bayesian econometrics, focusing on the application of the methods, rather than the techniques themselves (deriving full conditionals and coding). It does so by relying heavily on BayES for estimating the models presented in it and, as such, it can also be used as a gentle introduction to the software. BayES was chosen, apart from the obvious reason that the author is also the developer of the software, because of the nature of the software. BayES is designed from the beginning exclusively for Bayesian econometrics and it provides an intuitive graphical interface that allows first-time users to run models without having to spend hours reading the documentation. Additionally, it features a compact matrix language, which can be used by advanced users to code samplers for their own models, if these are not yet available in BayES. Equally importantly, it provides interfaces to other statistical software packages, both Bayesian and frequentist, which allow estimation of specialized models available in them.

Chapter 1 starts by defining the modern meaning of the term econometrics and proceeds to present the fundamentals of Bayesian inference and techniques. This chapter is, by far, the most challenging, as it deals with the meaning and interpretation of probability, a concept that appears straightforward until one really starts thinking about it, as well as with the process of using data to update prior beliefs. The chapter has very little to do with economics and can be viewed as a crash course in Bayesian inference for readers who have never seen the concepts and methods before. Simulation methods are also covered in this chapter, as it is hard to separate Bayesian estimation theory from modern estimation techniques. The extend of coverage of inference methods may seem unconventional to readers who have been exposed to frequentist econometrics, but one has to keep in mind that most readers of frequentist econometrics textbooks usually have already had a course in frequentist statistics and hypothesis testing, while this is rarely the case for Bayesian methods.

The following two chapters cover the basic models used in econometrics and which can be estimated with the methods presented in the first chapter. These include the linear model and systems of equations and the user will be referred back to them on multiple occasions. Chapter 4 discusses data augmentation, the method that enables Bayesian inference to deal with complex models on which frequentist methods usually "choke". Although initially discussed at a high level of abstraction, two applications of data augmentation are also presented in this chapter, as extensions to the linear model. This chapter is definitely recommended to all readers, as it forms the basis for much of the material covered in the remainder of the textbook. From this point onwards the reader could concentrate on the models of interest without any break in continuity.

The textbook follows the development of BayES and, as such, it can be considered incomplete. As BayES' coverage extends to include more models, this document will evolve as well. Nevertheless, the material included in the current version covers the basics of Bayesian inference and the most popular econometric models for cross-sectional and panel data. Therefore, it can already be used for a semester-long course on Bayesian econometrics or to provide the fundamentals for an applied econometrician.

This textbook has already started being used by academics and applied researchers. Among them, Prof John Burkett deserves a grateful acknowledgement for providing comments and

suggestions, which have greatly improved exposition in many places.

Chapter 1

Econometrics and Bayesian Inference

1.1 Overview

This chapter starts by defining the modern use of the term econometrics and by briefly comparing the frequentist and Bayesian approaches to statistical inference. Its primary purpose, however, is to introduce the reader to Bayesian concepts and methods, as well as to the algorithms that revolutionized the way applied Bayesian research is conducted. Therefore, the presentation abstracts from economic theory as much as possible and concentrates only on statistical concepts. At places exposition may appear repetitive and this is because the fundamental concepts are initially presented in a way that allows the reader to form a complete picture of the approach, before delving into the details. Ideas are fixed using very simple examples that have close to nothing to do with economics. Finally, Markov chain Monte Carlo (MCMC) methods are presented in an algorithmic fashion and only some intuition is provided, while the reader is directed to other textbooks and book chapters for formal definitions and proofs.

1.2 Econometrics: Frequentist and Bayesian Approaches

The modern meaning of the term *econometrics* was coined by Frisch (1933) as the unification of three aspects of quantitative economic analysis: (i) statistical, (ii) quantitative theoretical, and (iii) mathematical. Although this definition appears to enjoy general acceptance, the statistical aspect is undoubtedly stressed in the way econometrics is taught at the undergraduate level and presented in modern econometric textbooks and this approach is followed in this textbook as well. According to such an approach, mathematics is assumed to have been used at a preliminary stage of the process of econometric analysis to express a theoretical model in a form that is amenable to statistical analysis, while economic theory is used to derive refutable hypotheses, which can then be confronted with the data. These steps involve elements which are viewed as being too problem-specific to be covered in the main part of the text. Therefore, with some notable exceptions, the presentation of statistical methods abstracts from specific economic models, but economic theory is reintroduced in particular applications. On the other hand, mathematics is integrated seamlessly into the statistical derivations.

In applied work econometrics uses data to accomplish three primary tasks: (i) to estimate the parameters of statistical models suggested by economic theory, (ii) to evaluate the plausibility of statements or compare alternative models/theories when these are confronted with data, and (iii) to predict or forecast the values of quantities of interest. Almost always these

tasks are pursued in the order presented above, with the analysis in later steps being informed by the results obtained in preceding steps. Notice that, again, the statistical aspect of econometrics is emphasized, while the development of the theories to be tested or compared is not explicitly stated as an additional task.

The two preceding paragraphs defined econometrics from a compositional and a functional perspective, respectively. These definitions are generic enough to encompass both major branches of modern econometrics, namely *classical* or *frequentist* and *Bayesian* econometrics. However, the similarities between these two branches end as soon as the approach to statistical inference is concerned. The differences stem from the way randomness in the data is transformed into uncertainty with respect to the values of the parameters of an econometric model (or other quantities of interest) and this simple discrepancy is enough to make the two approaches largely incompatible.

Although both frequentist and Bayesian statistical methods are based on the axioms and laws of probability, they take a different view on the fundamental concept of probability itself. In the frequentist approach the probability of an event occurring is quantified by repeating a random experiment multiple times and calculating the proportion of times the event actually occurred. In the Bayesian approach, probability is used to express a state of knowledge or belief about the likelihood of the event. On one hand, the Bayesian view on probability is much more practical because it can be used in problems where no random experiment can be conceived, which can be repeated multiple times. On the other hand, quantifying beliefs introduces subjectivity to the analysis[1] and this has spurred a great deal of criticism of the Bayesian approach and equally many attempts from the Bayesian side to defend its methods. This debate extends far beyond econometrics or statistics and well into the realm of philosophy and, in particular, probabilistic logic.[2] The main arguments used in defense of the Bayesian approach and the possible ways of reducing the influence of subjective beliefs on the final results are briefly reviewed at the end of the following section. The interested reader is directed to Howie (2004) for a more in-depth discussion on the topic.

The different views on probability taken by the frequentist and the Bayesian approach lead to slightly different meanings for the term "estimation". In the frequentist approach observed data are used to construct a *confidence interval* for a parameter or any other quantity of interest and for a predetermined *confidence level*, say 95%. The confidence interval is such that, if the entire sampling and estimation process were to be repeated multiple times, then in 95% of the repetitions the constructed interval would contain the true value of the quantity of interest. Notice that, even with a single dataset, the frequentist approach has to rely on a conceptual repetition of the sampling and estimation process. On the other hand, the Bayesian approach uses the data to update prior beliefs about the value of the quantity of interest and the end result is usually a probability density function, which quantifies the uncertainty with respect to the true parameter value, after having seen the data. Because the Bayesian approach takes the data as given, interpretation of the results is much more intuitive. Furthermore, the entire process of estimation, model comparison and prediction becomes straightforward and can be very concisely presented because it relies only on the basic laws of probability.

Although conceptually straightforward, until the end of the previous century Bayesian methods were only marginally used in applied econometric work. This is because the mathematics involved in an application of Bayesian methods to most modern econometric models would make the approach either impractical or too restrictive. This has changed over the last two decades for three primary reasons: (i) the development of efficient sampling methods that allow very complex models to be considered, (ii) the increase in computing power of personal computers or clusters of computers, which facilitates the application of these sampling algorithms, and (iii) the incorporation of Bayesian techniques to standard econometrics/statistics software or the emergence of new software designed to automate calculations, thus relieving the researcher from tedious algebraic manipulations and the burden of coding the procedures

[1] This is not so much the case when the Bayesian view on probability is used to express the state of knowledge. The subtle difference between expressing a "state of knowledge" and a "degree of belief" has led to a further subdivision of the Bayesian view on probability to *objective* and *subjective* Bayesian probability.

[2] See Alder (2005a,b) for a very entertaining discussion on this issue.

necessary to perform Bayesian inference.

1.3 Bayes' Theorem and Bayesian Inference

Bayesian inference gets its name from Bayes' theorem, a result in probability theory that is used to update beliefs regarding the value of parameters or other random quantities using evidence from the data. Bayes' theorem follows from the formula of conditional probability and holds irrespective of whether one takes the frequentist or the Bayesian view on probability.

To derive Bayes' theorem let A and B be two events defined in relation to a random experiment, with probabilities Prob(A) and Prob(B), respectively. The probability of both A and B occurring in a repetition of the experiment is denoted by Prob(A, B). Assuming that Prob(B) $\neq 0$, the probability of A occurring, given that B has occurred is:

$$\text{Prob}(\mathtt{A}|\mathtt{B}) = \frac{\text{Prob}(\mathtt{A},\mathtt{B})}{\text{Prob}(\mathtt{B})} \tag{1.1}$$

This *conditional probability* formula is easier to interpret after some rearrangement:

$$\text{Prob}(\mathtt{A},\mathtt{B}) = \text{Prob}(\mathtt{A}|\mathtt{B}) \cdot \text{Prob}(\mathtt{B}) \tag{1.2}$$

which, in words, says:

> *the probability of* A *and* B *occurring is equal to the probability of* B *occurring times the probability of* A *occurring given that* B *has occurred*

We could think of this formula as a way of calculating the joint probability of A and B by first calculating the probability of B and then examining the probability of A, while treating B as having already occurred. However, the time dimension introduced here, where we think of B as occurring before A, is used only to give some intuitive interpretation of the formula. We could reverse the roles of A and B on the right-hand side and write:

$$\text{Prob}(\mathtt{A},\mathtt{B}) = \text{Prob}(\mathtt{B}|\mathtt{A}) \cdot \text{Prob}(\mathtt{A}) \tag{1.3}$$

Bayes' theorem follows by equating the right-hand sides of (1.2) and (1.3) and rearranging:

$$\text{Prob}(\mathtt{A}|\mathtt{B}) = \frac{\text{Prob}(\mathtt{B}|\mathtt{A}) \cdot \text{Prob}(\mathtt{A})}{\text{Prob}(\mathtt{B})} \tag{1.4}$$

Bayes' theorem reverses the roles of the two events in conditioning and, by doing so, allows calculation of Prob(A|B) using knowledge of Prob(B|A). More precisely and in the context of Bayesian inference, knowledge of Prob(B|A) and Prob(B) allows updating the belief of A occurring after obtaining evidence of B having occurred. Example 1.1 provides a simple application of Bayes' theorem, which illustrates the use of the theorem to update beliefs or knowledge about the likelihood of an event, when new information is obtained.

♦ **Example 1.1 Bayes' Theorem**

Increased air traffic near a small regional airport during high season causes delays in landing. In response to passengers' and airlines' concerns and complaints, the airport's management released the following information:

- the probability of an airplane landing at the airport with more than 10 minutes delay is 30%
- 60% of delayed landings are due to delayed departure from the airport of origin
- of the airplanes that land at the small airport, 20% leave their airport of origin with a delay

Suppose that you are picking a friend from the airport who just called you to tell you that her airplane will depart with a delay. What is the probability that the airplane your friend is in will land with more than 10 minutes delay?

To transform the information provided by the airport's management into probability statements define the events:

DL: *an airplane lands at the airport with a delay of more than 10 minutes*
DD: *an airplane leaves the airport of origin with a delay*

Using these definitions, the three bits of information above become:

- $\text{Prob}(\texttt{DL}) = 0.3$
- $\text{Prob}(\texttt{DD}|\texttt{DL}) = 0.6$
- $\text{Prob}(\texttt{DD}) = 0.2$

Prior to receiving the information that your friend's airplane departed with a delay, the probability that it would land with a delay of more than 10 minutes is simply $\text{Prob}(\texttt{DL}) = 0.3$. Given the additional information of delayed departure, the probability of delayed landing becomes:

$$\text{Prob}(\texttt{DL}|\texttt{DD}) = \frac{\text{Prob}(\texttt{DD}|\texttt{DL}) \cdot \text{Prob}(\texttt{DL})}{\text{Prob}(\texttt{DD})} = \frac{0.6 \cdot 0.3}{0.2} = 0.9$$

Bayes' theorem was presented here using events, but it can be shown that it also holds when considering random variables. Let X and Y be two random variables with probability density functions $\text{p}(x)$ and $\text{p}(y)$, respectively. Using these probability density functions Bayes' theorem can be expressed as:

$$\text{p}(x|y) = \frac{\text{p}(y|x) \cdot \text{p}(x)}{\text{p}(y)} \tag{1.5}$$

In Bayesian inference x plays the role of the parameters of a stochastic model and y the role of the data. Using notation that will persist throughout this textbook, by collecting all parameters in a vector $\boldsymbol{\theta}$ and the data in a vector $\mathbf{y}$ and by renaming some of the densities, the theorem becomes:

$$\pi\left(\boldsymbol{\theta}|\mathbf{y}\right) = \frac{\text{p}(\mathbf{y}|\boldsymbol{\theta}) \cdot \text{p}(\boldsymbol{\theta})}{m\left(\mathbf{y}\right)} \tag{1.6}$$

Of course, if the model involves more than a single parameter and more than a single data point is used then all densities in the last expression will be multivariate. The last expression involves four densities:

1. $\pi\left(\boldsymbol{\theta}|\mathbf{y}\right)$ is the *posterior density* and it is the primary quantity of interest in Bayesian inference. It expresses our knowledge about the values of the model's parameters *after* we see the data.
2. $\text{p}(\mathbf{y}|\boldsymbol{\theta})$ is the *likelihood function* and it is the main part of the model specification. The likelihood function is the density of the data given the values of the model's parameters and it depends on the assumptions imposed by the researcher on the *data-generating process*.
3. $\text{p}(\boldsymbol{\theta})$ is the *prior density* of the model's parameters and it is an additional element of the model specification. The prior density expresses knowledge or beliefs about the values of the parameters *before* we look at the data.
4. $m\left(\mathbf{y}\right)$ is the *marginal likelihood* and, as its name suggests, it is the density of the data marginally with respect to the parameters. Its form depends on the specification of model (likelihood function and prior density) and can be obtained by integrating $\boldsymbol{\theta}$ from the numerator of (1.6). In most applications it will be difficult to perform this integration analytically, but, given that the primary quantity of interest is the posterior density of the parameters and that $m\left(\mathbf{y}\right)$ does not involve $\boldsymbol{\theta}$, the denominator in the last expression can be viewed as a constant of proportionality for $\pi\left(\boldsymbol{\theta}|\mathbf{y}\right)$. This constant is irrelevant for purposes of estimation and can be ignored at this stage. In practice, therefore, it is most often ommited from (1.6) and Bayes' theorem is expressed as:

$$\pi\left(\boldsymbol{\theta}|\mathbf{y}\right) \propto \text{p}(\mathbf{y}|\boldsymbol{\theta}) \cdot \text{p}(\boldsymbol{\theta}) \tag{1.7}$$

with the symbol "$\propto$" taken to mean "proportional to".

The last expression in words says:

the posterior density of the model's parameters is proportional to the likelihood times the prior density

The right-hand side contains the complete specification of the model. It is stressed that a model specification in Bayesian inference consists of both the likelihood function and the prior density of the parameters. Because of the fundamental role that the three densities that appear in (1.7) play in Bayesian inference, each one is examined in detail in the following three subsections. To fix ideas, the discussion is augmented with a simple but very extensive example that runs throughout this section.

1.3.1 The Likelihood Function

The *likelihood function* constitutes part of the specification of a stochastic model and it conveys the assumptions on the process that generates the data. It is expressed as a density of the form $\mathrm{p}(\mathbf{y}|\boldsymbol{\theta})$, where $\mathbf{y}$ are the data and $\boldsymbol{\theta}$ the model's parameters. The meaning of $\boldsymbol{\theta}$ is straightforward, but the meaning of $\mathbf{y}$ deserves some discussion. With a given dataset at hand, $\mathbf{y}$ will be populated by numerical values. In stochastic models these values are viewed as realizations of random variables; the realizations are what we observe (fixed) but the underlying *data-generating process* is what we are interested in. This is because statistical inference is not concerned with simply describing the dataset at hand, but its primary objective is to make statements about the values of $\boldsymbol{\theta}$ in the population. And the only way this can be achieved is by considering the process that generates the data in the population. Of course, the data are used in the process of statistical inference to provide information about the values of the parameters.

In words, and given the foregoing discussion, the likelihood function is the probability density function of a potential dataset, evaluated at the observed data points, given the values of the parameters. The values of the parameters are not known yet and conditioning on them may appear bizarre. However, Bayes' theorem can be used to reverse the roles of $\mathbf{y}$ and $\boldsymbol{\theta}$ in conditioning, such that we get the density of the parameters given the observed data.

Because the likelihood function expresses the assumptions on the process that generates data, different models will have different likelihood functions, simply because they concern different phenomena. Therefore, specification of the likelihood is not possible at the level of generality considered here. Nevertheless, this specification will be the major part of the chapters that follow, which deal with specific statistical and economic models. A simple example is provided here only to fix ideas.

♦ **Example 1.2 Customer Arrival Rate**

A sandwich store which is located at the commercial district of a large city becomes very busy during rush hour, when employees of nearby businesses have their lunch break. The manager knows that the store can serve 4 customers per minute but she needs an estimate of the rate at which customers are added to the queue (arrival rate). For that purpose, she stood at the door of the sandwich store and recorded the time that elapsed between each successive customer arrival, until 100 customers entered the store.

Let y_i be the time that elapses between the i^{th} and the following arrival and let $\mathbf{y}$ be a vector that contains the observed data. Therefore, $\mathbf{y}$ is an $N \times 1$ vector, where $N = 100$ (the number of observations).

We now need to specify a model for the process that generated these data. By far the most popular distribution used in this type of queueing problems is the Exponential. The Exponential distribution's probability density function is $p(x) = \lambda e^{-\lambda x}$, where λ is the rate parameter. In our application λ is the primary quantity of interest, as it measures the expected number of customers that enter the store, per minute. One additional assumption that we will make here is that the time elapsing between two successive arrivals is independent of the time elapsing between preceding or following arrivals. These assumptions lead to a model where each observed y_i is a draw from an exponential distribution with rate λ: $y_i \sim \mathrm{Exp}(\lambda)$. The likelihood function is the density of all data points and, because of the independence assumption, this density can be expressed as:

$$\mathrm{p}(\mathbf{y}|\lambda) = \prod_{i=1}^{N} \lambda e^{-\lambda y_i} = \lambda^N e^{-\lambda \sum_{i=1}^{N} y_i}$$

1.3.2 The Prior Density

The *prior density* constitutes the second part of the specification of a model in Bayesian inference and it conveys prior beliefs or knowledge about the values of the parameters of a model. These beliefs are prior in the sense that they are formed without using information contained in the dataset at hand. Like the likelihood function, the prior density takes the form of a probability density function, expressed in general terms as $\mathrm{p}(\boldsymbol{\theta})$. In practice this density will belong to a conveniently chosen parametric family and it will be augmented by this family's own parameters, called *hyperparameters*.[3]

Both the family of the prior density and the values of the hyperparameters are chosen by the researcher and, as any form of model specification, they may have a considerable impact on the conclusions drawn from the analysis. Most importantly, because these choices are not updated in the process of estimation or inference, they have the potential of introducing a degree of subjectivity into the analysis. Therefore, a lot of effort has gone into deriving priors that do not impose too harsh restrictions on the data or that have minimal impact on the results.

Priors can be classified into the following three broad categories:

- *subjective priors*: When using subjective priors, the researcher can be thought of as expressing his/her own beliefs about the values of the parameters.

  In an extreme scenario, a researcher with very strong beliefs would pick a prior density that would have a large spike around a certain parameter value and be zero elsewhere. Such a prior would dominate the information contained in the data and very little can be learned from them. The conclusions obtained using such a prior will still be in line with the researcher's strong views, but probably of limited usefulness to anyone else, at least as far as the statistical analysis is concerned.

  On the other hand, a prior density that is chosen such that it is positive over a wide range of possible values of the parameters and contains no spikes, can be used to expresses rather "vague" beliefs. Given enough information in the data, such a prior is likely to have minimal impact on the results and conclusions. In this, case the prior is dominated by the likelihood, especially if the dataset at hand contains many observations. Furthermore, the impact of alternative specifications of the prior density on the results can be examined by repeating the analysis multiple times and with different priors, something that is known as *sensitivity analysis*.

  There is a continuum of approaches that can be taken and lie between the extremes of refusing to learn from the data and taking great care in devising "*vague*" priors. Berger (1985) presents some practical ways of devising subjective priors that can be used to express beliefs.

- *objective priors*: In this case the priors are formed using knowledge available either from theory or from previous statistical analyses (of other datasets than the one at hand). Objective priors contain information about the density of the parameters, but this information can be justified.

  As an example of using economic theory to form priors, consider an aggregate production function with capital and labor as the two inputs and value added as the measure of output. Most economic models suggest that, at the aggregate level, the production function exhibits constant returns to scale and the specification of the prior could take this into account. The prior could be imposing the constant-returns-to-scale assumption very strongly, by allowing the values of the parameters to deviate only marginally from it, or rather "vaguely".

  When results from previous analyses are available, these can also be taken into account when forming priors. Continuing with the previous example, if the production function

[3] Throughout this textbook Greek letters are used to denote parameters and Latin for hyperparameters.

takes the form a Cobb-Douglas then previous analyses have shown that the output elasticity with respect to labor is close to $\frac{2}{3}$ and that of capital close to $\frac{1}{3}$. On top of that, from the results of previous analyses the researcher could obtain approximations to the entire distribution of the production function's parameters and incorporate them in the prior density. The task of Bayesian inference in this context could be, not to disregard previous knowledge, but to update it using new data.

- *noninformative priors*: These priors are designed with the objective of having minimal impact on the results in general situations. Other terms used for them are *reference*, *vague*, *flat* or *diffuse* priors.

  There have been multiple attempts in the literature to derive a general way of constructing noninformative priors and the most well known is Jeffreys' approach. Jeffreys' priors satisfy the *invariance principal*, according to which a prior density for a parameter θ should convey the same amount of information as for a monotonic transformation of θ, when the model is re-parameterized. Although invariance could be a desirable property, it is not clear in what sense such a prior is noninformative. Furthermore, generalization to multiple parameters of Jeffreys' priors is rather controversial. Most importantly, Jeffreys' priors are almost always improper (they do not integrate to unity as a density function should) and this creates problems in some complex models and hamper model comparison via *Bayes factors*.

  Another approach for constructing noninformative priors starts by formally defining what is meant by saying that the prior should have minimal impact on the results. The approach uses concepts from information theory to find the prior that can be "maximally dominated by the data" and the term *reference prior* is almost exclusively associated with it. Like Jeffreys' priors, reference priors are almost always improper and in problems with a single parameter they take the same form as Jeffreys' priors. A review of the approach by the authors that contributed the most in its development can be found in Berger et al. (2009).

As it can be seen from the discussion above, the three categories of approaches used for deriving prior densities do not have clear boundaries. For one thing, the terms "vague" and "diffuse" can be used in the contexts of subjective, objective and noninformative priors. The current tendency is to move away from the use of the term "noninformative prior", as it is largely recognized that it is nearly impossible to construct priors entirely free of prior beliefs. Taking the argument to the extreme, expressing complete ignorance about the value of a parameter still conveys some amount of information. The interested reader is directed to chapter 3 from Berger (1985), chapter 1 from Lancaster (2004), chapter 2 from Gelman et al. (2013), and chapter 4 from Greenberg (2013) for discussions that take different viewpoints on the subject.

The subjective approach to forming priors will be used in this textbook, but always with special care such that no priors are chosen which would restrict what we can learn from the data. When information from economic theory or previous studies is available, this will be incorporated in the priors, thus moving towards the objective approach. Avoiding Jeffreys' priors is done both for generality and for practical purposes. In most cases Jeffreys' priors can be obtained by letting the parameters of a prior density (the *hyperparameters*) to go towards specific values, such as 0 or ∞. In practice, relatively flat priors (coverage of the range of possible values of the parameters and without spikes) are sufficient to guarantee minimal impact on the results. Finally, because model comparison will be performed using *Bayes factors* we will need proper priors for the models' parameters.

Returning to the practical problem of defining a prior, when using Jeffreys' or reference priors the respective procedures will suggest a single formula and nothing else needs to be done at this stage. In the subjective and objective approaches the researcher has or gets to pick: (i) a family of density functions for the prior, and (ii) values for the hyperparameters of this density. Picking a parametric family is usually based on grounds of straightforward reasoning and convenience. For parameters which must be restricted to be positive, such as variance

or scale parameters, a density function should be used that is defined only for positive values of its argument. On the other hand, densities with support on the real line should be used for location parameters, which could be either positive or negative. In terms of convenience, analysis is greatly simplified if the prior is chosen such that, when combined with the likelihood function, it results in a posterior that belongs to a known parametric family of distributions. There is no guaranty that such a prior will exist for every parameter in a model, but when it does the posterior usually belongs to the same parametric family as the prior. A prior density for a parameter vector, $\boldsymbol{\theta}$, in a model is called *conjugate* if it leads to a posterior for $\boldsymbol{\theta}$ that belongs to the same parametric family as the prior.

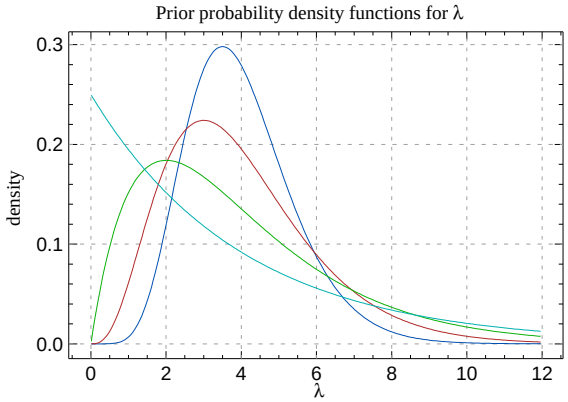
♦ **Example 1.2 Customer Arrival Rate (Continued)**
The sandwich-store example presents a model for the process that generates inter-arrival times, where each data point is assumed to be a draw from an Exponential distribution with rate λ. The rate parameter of an Exponential distribution is always positive and we can consider a Gamma distribution with shape parameter a and rate parameter b as a prior for it:

$$\mathrm{p}(\lambda) = \frac{b^a}{\Gamma(a)}\lambda^{a-1}e^{-b\lambda}$$

a and b are the hyperparameters and we need to pick values for them to complete the specification of the prior.

The expected value of a Gamma-distributed random variable is $\frac{a}{b}$ and the variance $\frac{a}{b^2}$. Suppose that, before we see the data, we expect that, on average, about four customers enter the store per minute. This implies that λ should be close to four and we can express this in the prior by picking values for a and b such that $a = 4b$. Of course, if we want to learn anything from the data we should allow λ to deviate from this expectation, if the data suggest so. We can increase the variance of λ in the prior by picking smaller values for both a and b.

The following table presents possible values for the hyperparameters and their implications for the expected value and variance of λ in the prior. The resulting prior densities are plotted in the following figure. Notice that as the values of a and b become smaller the prior becomes more vague.



Hyperparameters	$\mathrm{E}(\lambda)$	$\mathrm{V}(\lambda)$
$a = 8,\ b = 2$	4	2
$a = 4,\ b = 1$	4	4
$a = 2,\ b = 0.5$	4	8
$a = 1,\ b = 0.25$	4	16

The figure above can be recreated in BayES by placing the code contained in the following box in the Script Editor window and and hitting `Ctrl+R`.

```
// create a range of values from 0.01 to 12.0, on which the pdf of the
// Gamma distribution will be evaluated
x = range (0.01 , 12 , 0.05) ;

// calculate the pdf of the Gamma distribution at each point x and for
// varying values of the hyperparameters
y1 = gampdf(x, 8, 2);
y2 = gampdf(x, 4, 1);
y3 = gampdf(x, 2, 0.5);
y4 = gampdf(x, 1, 0.25);

// plot the four different pdfs against x
plot ([ y1 , y2, y3, y4 ], x,
    "title" = "Prior probability density functions for \lambda",
    "xlabel" = "\lambda", "ylabel" = "density", "grid" = "on");
```

1.3.3 The Posterior Density

The *posterior density* is the end product of a Bayesian inference exercise, at least as far as parameter estimation is concerned. This density takes the general form $\pi(\boldsymbol{\theta}|\mathbf{y})$ and it expresses our knowledge about $\boldsymbol{\theta}$ after having seen the data. The posterior density is obtained from an application of Bayes' theorem:

$$\pi(\boldsymbol{\theta}|\mathbf{y}) = \frac{\mathrm{p}(\mathbf{y}|\boldsymbol{\theta}) \cdot \mathrm{p}(\boldsymbol{\theta})}{m(\mathbf{y})} \propto \mathrm{p}(\mathbf{y}|\boldsymbol{\theta}) \cdot \mathrm{p}(\boldsymbol{\theta}) \tag{1.8}$$

which makes apparent that it depends on the modeling assumptions that are incorporated in both the likelihood and the prior.

When the model involves a single parameter, the posterior density is this parameter's probability density function. When there are more parameters in the model, the posterior is the joint density of all parameters. Mathematically, the posterior takes the form of a formula which, in most cases, provides little intuition about the values of the parameter(s). The task now becomes one of extracting the information contained in $\pi(\boldsymbol{\theta}|\mathbf{y})$ and presenting it in a way that is easy to comprehend. An obvious way to proceed is to plot a graph of each parameter's posterior density, marginally with respect to the rest. However, this approach becomes impractical if the model has more than a few parameters. Therefore, it is customary in applied work to present in a table the first two moments of the marginal posterior density of each parameter in $\boldsymbol{\theta}$. For example, if $\boldsymbol{\theta}$ consists of two parameters, θ_1 and θ_2, the results are presented in a table of the following form:

Parameter	**Mean**	**St.dev.**
θ_1	$\mathrm{E}(\theta_1\|\mathbf{y})$	$\sqrt{\mathrm{V}(\theta_1\|\mathbf{y})}$
θ_2	$\mathrm{E}(\theta_2\|\mathbf{y})$	$\sqrt{\mathrm{V}(\theta_2\|\mathbf{y})}$

where:

$$\begin{aligned} \mathrm{E}(\theta_1|\mathbf{y}) &= \int_{\Theta_1} \theta_1 \cdot \pi(\theta_1|\mathbf{y})\ \mathrm{d}\theta_1 \\ \mathrm{V}(\theta_1|\mathbf{y}) &= \int_{\Theta_1} \Big(\theta_1 - \mathrm{E}(\theta_1|\mathbf{y})\Big)^2 \cdot \pi(\theta_1|\mathbf{y})\ \mathrm{d}\theta_1 \end{aligned} \tag{1.9}$$

and $\pi(\theta_1|\mathbf{y})$ is the marginal posterior density of θ_1:

$$\pi(\theta_1|\mathbf{y}) = \int_{\Theta_2} \pi(\theta_1, \theta_2|\mathbf{y})\ \mathrm{d}\theta_2 \tag{1.10}$$

Similar calculations should be performed for θ_2.

If the model contains more than two parameters, calculating the moments involves multidimensional integration. However, these integrals are rarely evaluated analytically in practice. If $\pi(\boldsymbol{\theta}|\mathbf{y})$ belongs to a known family of densities, most frequently the marginal moments will be available in closed form and the only thing one has to do is to evaluate the formulas using the dataset at hand. If $\pi(\boldsymbol{\theta}|\mathbf{y})$ does not belong to a known family or the marginal densities are not available analytically, the moments presented above can be approximated using simulation.[4]

The posterior density function can also be used to make probability statements about the values of the parameters. For example, the probability of a parameter θ_1 being within an interval $[c_1, c_2]$ can be expressed as:

$$\mathrm{Prob}(c_1 \leq \theta_1 \leq c_2) = \int_{c_1}^{c_2} \pi(\theta_1|\mathbf{y})\ \mathrm{d}\theta_1 \tag{1.11}$$

[4]Simulation methods are covered later in this chapter, but it is worth mentioning at this point that the way Bayesian inference is conducted was revolutionized by simulation methods because they provide a way of approximating these integrals.

The integral above can be evaluated analytically if the marginal cumulative density function of θ_1 is known in closed form or, as with the previous integrals, approximated using simulation methods.

Finally, it has become common practice in applied research to present, along with the moments of a parameter, its 90% or 95% *credible interval*. The credible interval is constructed using a simplified version of (1.11) by picking the numbers c_1 and c_2 such that the left-hand side probability is equal to 0.9 or 0.95, respectively. For example, a 90% credible interval for θ_1 can be obtained by setting c_1 equal to the value that satisfies $\text{Prob}(\theta_1 \leq c_1) = 0.05$ and c_2 to the value that satisfies $\text{Prob}(\theta_1 \leq c_2) = 0.95$. The credible interval $[c_1, c_2]$ constructed in this way presents another way of quantifying the uncertainty regarding the value of θ_1, as it states that the probability of θ_1 being between c_1 and c_2 is 90%. In this example the credible interval was constructed such that equal probability mass was discarded from the lower and upper tails of the posterior distribution. There exist other ways to construct credible intervals and a popular alternative is to construct the *shortest possible credible interval*. However, when the posterior density cannot be obtained in closed form, construction of the shortest possible interval may become very challenging.

Because credible intervals provide a way of performing interval estimation, they can be viewed as the Bayesian counterpart to frequentist confidence intervals. The differences between the two concepts, however, become apparent once the meaning of a confidence interval is examined in detail. A 90% confidence interval for θ_1 is to be interpreted in the following way: if one were able to obtain many datasets generated from the assumed process and repeat the process of estimation and construction of the confidence interval on all of these datasets, then in 90% of the repetitions the constructed confidence interval will contain the true parameter value. On the other hand, the credible interval has a much more intuitive interpretation: θ_1 lies within the 90% credible interval with probability 90%.

♦ **Example 1.2 Customer Arrival Rate (Continued)**
The two previous parts of the sandwich store example present a model specification with likelihood function and prior density:

$$\text{p}(\mathbf{y}|\lambda) = \lambda^N e^{-\lambda \sum_{i=1}^{N} y_i} \qquad \text{and} \qquad \text{p}(\lambda) = \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b\lambda}$$

respectively. Using Bayes' theorem, the posterior density of λ from this model is:

$$\begin{aligned} \pi(\lambda|\mathbf{y}) &\propto \text{p}(\mathbf{y}|\lambda) \cdot \text{p}(\lambda) \\ &= \lambda^N e^{-\lambda \sum_{i=1}^{N} y_i} \times \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b\lambda} \\ &\propto \lambda^{(N+a)-1} e^{-\lambda\left(\sum_{i=1}^{N} y_i + b\right)} \end{aligned}$$

where $\frac{b^a}{\Gamma(a)}$ is dropped from the final expression and, because it does not involve λ, becomes part of the constant of proportionality.

The resulting posterior looks like a Gamma probability density function with shape parameter $\tilde{a} = N + a$ and rate parameter $\tilde{b} = \sum_{i=1}^{N} y_i + b$. All that is missing is a constant of proportionality. This constant can be obtained using the fact that proper density functions integrate to unity:

$$\int_0^\infty \pi(\lambda|\mathbf{y})\, \text{d}\lambda = 1 \Rightarrow \int_0^\infty c \cdot \lambda^{\tilde{a}-1} e^{-\lambda\tilde{b}}\, \text{d}\lambda = 1$$

The constant, c, that satisfies the last equation is precisely the one that would make the posterior density equal (not proportional) to a Gamma density: $c = \frac{\tilde{b}^{\tilde{a}}}{\Gamma(\tilde{a})}$. Therefore:

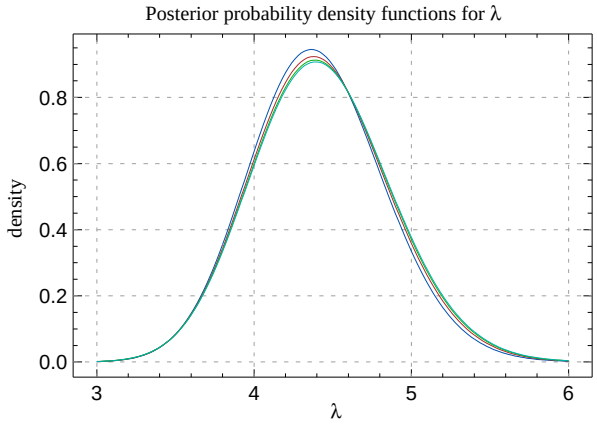
$$\lambda|\mathbf{y} \sim \text{Gamma}\left(\tilde{a}, \tilde{b}\right)$$

From the properties of the Gamma distribution we get $\text{E}(\lambda|\mathbf{y}) = \frac{\tilde{a}}{\tilde{b}}$ and $\text{V}(\lambda|\mathbf{y}) = \frac{\tilde{a}}{\tilde{b}^2}$. Furthermore, because we have a single parameter in the model, we can plot the entire posterior probability density function.

The only thing left to do is to feed the formulas with the data. The file `WaitingTimes.csv` contains values for 100 draws from an Exponential distribution. Using the formulas derived here and the values of the hyperparameters defined in the second part of the example, we obtain the results in the following table.

Hyperparameters	$\mathrm{E}(\lambda\|\mathbf{y})$	$\mathrm{V}(\lambda\|\mathbf{y})$
$a = 8,\ b = 2$	4.4048	0.1797
$a = 4,\ b = 1$	4.4220	0.1880
$a = 2,\ b = 0.5$	4.4312	0.1925
$a = 1,\ b = 0.25$	4.4359	0.1948

The posterior densities of λ using each of the four pairs of values for the hyperparameters are presented in the following figure.



Notice that as the prior becomes more vague (a and b go towards zero), the posterior expectation moves away from the prior expected value of $\frac{a}{b} = 4$, but also that the posterior variance increases. This is an indication that the data tend to support a value for $\mathrm{E}(\lambda|\mathbf{y})$ greater than 4 and, as the prior becomes more vague, they are allowed to express this more freely. Nevertheless, differences in the posterior are small and quite different values for the hyperparameters produce similar posterior densities. Even with 100 observations, information from the data can dominate information from the prior.

We note in passing that, in a frequentist setting, both the maximum-likelihood and the method of moments techniques would produce a point estimate of λ as:

$$\hat{\lambda}_{\mathsf{MLE}} = \frac{N}{\sum_{i=1}^{N} y_i}$$

The only difference between the formulas for $\hat{\lambda}_{\mathsf{MLE}}$ and $\mathrm{E}(\lambda|\mathbf{y})$ from the Bayesian approach is that in the latter, the values of the hyperparameters are added to the numerator and denominator, respectively. As the prior becomes more vague, $\mathrm{E}(\lambda|\mathbf{y})$ converges to the frequentist point estimate. Using the same dataset, the maximum-likelihood estimate of λ is 4.4407.

The results and figure presented above can be generated in BayES using the code contained in the following box.

```
// import the data into a dataset called Data
Data = webimport("www.bayeconsoft.com/datasets/WaitingTimes.csv");

// get the number of observations in the dataset
N = rows(Data);

// calculate the sum of the values in y
sumy = sum(Data.y);

// define values for the hyperparameters
a = [ 8; 4;   2;    1];     // 4x1 vector
b = [ 2; 1; 0.5; 0.25];     // 4x1 vector
```

```
// calculate the posterior parameters for each pair of hyperparameters
a_tilde = N + a;
b_tilde = sumy + b;

// calculate the posterior moments for each set of hyperparameters
E_lambda = a_tilde./b_tilde;
V_lambda = a_tilde./(b_tilde.^2);

print( [E_lambda, V_lambda]);

// calculate the maximum-likelihood estimate of lambda
lambda_MLE = N/sumy;
print(lambda_MLE);

// plot the posterior densities
x = range(3, 6, 0.02) ;
y1 = gampdf(x, a_tilde(1), b_tilde(1));
y2 = gampdf(x, a_tilde(2), b_tilde(2));
y3 = gampdf(x, a_tilde(3), b_tilde(3));
y4 = gampdf(x, a_tilde(4), b_tilde(4));

plot([ y1 , y2, y3, y4 ], x,
    "title" = "Posterior probability density functions for \lambda",
    "xlabel" = "\lambda", "ylabel" = "density", "grid" = "on");
```

1.3.4 Model Comparison and the Marginal Likelihood

Estimating the parameters of a stochastic model, or to put it better, updating our knowledge about the values of the parameters using evidence from the data, accomplishes the first task of econometrics. After this step is completed, the analysis can move to comparing alternative models in terms of their ability to accommodate the data. The Bayesian approach provides a very intuitive device for model comparison. Although multiple models can be compared, to keep exposition simple, the presentation is restricted here to two models, with generalizations provided at the end of this subsection.

Suppose that the researcher has two competing theories, which suggest alternative models:

$$\begin{array}{lll} \text{Model 0:} & \mathrm{p}_0\left(y_i|\boldsymbol{\theta}_0,\bullet\right), & \mathrm{p}_0\left(\boldsymbol{\theta}_0\right) \\ \text{Model 1:} & \mathrm{p}_1\left(y_i|\boldsymbol{\theta}_1,\bullet\right), & \mathrm{p}_1\left(\boldsymbol{\theta}_1\right) \end{array} \tag{1.12}$$

In the context of statistical inference, the two models can be expressed as different data-generating processes, captured by the density of a potential observation, augmented with the prior density for the parameters. Labels 0 and 1 are used here to distinguish the elements of these processes. The general formulation above allows for differences between the data-generating processes in the form of the likelihood function, the number of parameters or the set of conditioning variables, denoted by "$\bullet$", as well as differences in the prior densities. Notice, however, that both processes describe the generation of the same variable, y_i.

In accordance with the labels used above, define M as a discrete random variable which can assume two values, 0 or 1, and whose value indicates which of the two models better describes the phenomenon under study.[5] Next, associate prior probabilities with the events $M=0$ and $M=1$. These probabilities express prior beliefs or knowledge about the relative plausibility of the two models. As with parameter estimation, the *prior model probabilities* do not contain information from the dataset at hand and the data are only used to update these priors. By

[5] A convenient mechanism to conceptualize the problem is to think of it as if one of the two models is the "true model" that generates the data, but we are uncertain as to whether this is Model 0 or Model 1. The use of the term "true model", however, is controversial and it has to be recognized that this mechanism involves a great abstraction from reality. According to George Box *"all models are wrong, but some are useful"*. Even if one disagrees with this statement, it can be argued that, by entertaining only two models, it is highly unlikely that we have included the "true model" in the analysis.

applying Bayes' theorem to Model 0 we obtain:

$$\text{Prob}(M=0|\mathbf{y}) = \frac{m(\mathbf{y}|M=0)\cdot\text{Prob}(M=0)}{m(\mathbf{y})} \tag{1.13}$$

where $\text{Prob}(M=0)$ and $\text{Prob}(M=0|\mathbf{y})$ are, respectively, the *prior* and *posterior model probabilities* for Model 0. $m(\mathbf{y}|M=0)$ is the density of the data given the assumptions incorporated in Model 0, but marginally with respect to the parameters that appear in Model 0. The assumptions made by Model 0 involve both the likelihood function and the prior and this density can be expressed as:

$$m(\mathbf{y}|M=0) = \int_{\Theta_0} \text{p}_0(\mathbf{y}|\theta_0, \bullet)\cdot\text{p}_0(\theta_0)\,\text{d}\theta_0 \tag{1.14}$$

It becomes apparent from this discussion that $m(\mathbf{y}|M=0)$ is precisely the *marginal likelihood* function that appears in the denominator of Bayes' theorem and, when estimating the model's parameters, it was treated as a normalizing constant and ignored. The only difference is that now we explicitly recognize that there are alternative models that could have generated the data and these models result in different marginal likelihood functions.

Finally, $m(\mathbf{y})$ in (1.13) is the density of the data marginally with respect to both parameters and modeling assumptions. Because there is no conditioning information at all associated with $m(\mathbf{y})$, very little can be said about about its form. A convenient way to avoid having to calculate this normalizing constant is to apply Bayes' theorem to Model 1 to obtain:

$$\text{Prob}(M=1|\mathbf{y}) = \frac{m(\mathbf{y}|M=1)\cdot\text{Prob}(M=1)}{m(\mathbf{y})} \tag{1.15}$$

and then divide (1.13) by (1.15). This process gives the *posterior odds ratio* between Model 0 and Model 1:

$$\frac{\text{Prob}(M=0|\mathbf{y})}{\text{Prob}(M=1|\mathbf{y})} = \frac{m(\mathbf{y}|M=0)}{m(\mathbf{y}|M=1)}\cdot\frac{\text{Prob}(M=0)}{\text{Prob}(M=1)} \tag{1.16}$$

The posterior odds ratio is equal to the *prior odds ratio* times the ratio of marginal likelihoods from the two models. The latter is known as the *Bayes factor*. The posterior odds ratio indicates the relative plausibility of the two models *after* we see the data. The information contained in this ratio is usually presented by normalizing the posterior model probabilities such that they sum to unity. That is, once a value, say c, is obtained for the posterior odds ratio, one may solve the system of equations:

$$\left.\begin{array}{l}\text{Prob}(M=0|\mathbf{y}) = c\cdot\text{Prob}(M=1|\mathbf{y}) \\ \text{Prob}(M=0|\mathbf{y}) + \text{Prob}(M=1|\mathbf{y}) = 1\end{array}\right\} \tag{1.17}$$

for the posterior model probabilities. Keep in mind, however, that this is just a normalization used to facilitate interpretation and claims involving only one of the two models being the "true model" should, optimally, be avoided (see also footnote 5).

Generalization of the procedure described above for model comparison to the case of $J > 2$ models is straightforward. The procedure consists of the following steps:

1. estimating all J models and calculating the values of the respective marginal likelihoods
2. assigning prior model probabilities to the J models
3. obtaining $J-1$ posterior odds ratios using (1.16)
4. solving the system of equations consisting of the $J-1$ posterior odds ratios and the equation $\sum_{j=1}^{J}\text{Prob}(M=j|\mathbf{y}) = 1$

The most challenging step in this procedure is estimating each model and, especially, calculating the value of the marginal likelihood. As it can be seen from (1.14), calculating $m(\mathbf{y}|M=j)$ involves an integral, which will rarely have an analytical solution. Approximations to this integral can be obtained by various methods, but we will not go further into the details here. The interested reader is directed to Gelfand & Dey (1994), Chib (1995), Chib & Jeliazkov (2001) and Lewis & Raftery (1997) for some popular approaches.

1.3.5 Prediction and Forecasting

The third task of econometrics is to make predictions about the values of the variables being modeled. These predictions make use of information contained in the observed data and depend on the model specification. To fix ideas, suppose that the phenomenon being studied involves modeling the data-generating process of a single variable, y. As before, let $\mathbf{y}$ be a vector containing the observed values of this variable (realizations from the data-generating process). Now, define $\mathbf{y}_*$ as a vector of random variables, each one of them associated with the value of y, either in different repetitions of the data-generating process or at different points in time. As with parameter estimation, prediction involves expressing the uncertainty about $\mathbf{y}_*$, using information from the data. This uncertainty is quantified using the *posterior predictive density*, which can be expressed as:

$$\mathrm{p}(\mathbf{y}_*|\mathbf{y}) = \int_\Theta \mathrm{p}(\mathbf{y}_*, \boldsymbol{\theta}|\mathbf{y})\ \mathrm{d}\theta = \int_\Theta \mathrm{p}(\mathbf{y}_*|\boldsymbol{\theta}, \mathbf{y}) \cdot \pi(\boldsymbol{\theta}|\mathbf{y})\ \mathrm{d}\theta \tag{1.18}$$

Notice that the posterior predictive density is the probability density function of $\mathbf{y}_*$ conditional only on the observed data. This density is obtained above, in two steps: first by marginalizing $\mathrm{p}(\mathbf{y}_*, \boldsymbol{\theta}|\mathbf{y})$ with respect to $\boldsymbol{\theta}$ and then by conditioning on $\boldsymbol{\theta}$ inside the integral. $\pi(\boldsymbol{\theta}|\mathbf{y})$ in the second step is the posterior density of the parameters, obtained by the application of Bayes' theorem on the original problem of parameter estimation. In all but the simplest models, the last integral will be impossible to evaluate analytically. However, once again, simulation methods can be used for approximating it.

In many cases independence assumptions made on the data-generating process will allow simplifying $\mathrm{p}(\mathbf{y}_*|\boldsymbol{\theta}, \mathbf{y})$ to $\mathrm{p}(\mathbf{y}_*|\boldsymbol{\theta})$. In such cases, all information contained in the observed data is transmitted to the parameters and $\mathbf{y}$ contains no additional information on $\mathbf{y}_*$. Especially in time-series contexts, however, this simplification will not be possible due to the dependence of current values of y on its past values and one needs to work with the general version of the posterior predictive density.

As it was the case with the parameters' posterior density, the posterior predictive density may not be the best device to communicate uncertainty with respect to the values of $\mathbf{y}_*$. Therefore, one additional step is taken in the context of prediction/forecasting, that of summarizing the information contained in the posterior predictive density. This is usually done by presenting the moments of $\mathbf{y}_*$ (expected value, variance, etc.) along with the corresponding credible intervals. Calculation of the moments or credible intervals involves additional integration, which is most often performed by simulation.

♦ **Example 1.2 Customer Arrival Rate (Continued)**

In the sandwich-store example we assumed that each inter-arrival time, y_i is a draw from an Exponential distribution with rate λ. Using a Gamma prior we expressed the posterior density of λ as a Gamma density with shape parameter $\tilde{a} = N + a$ and rate parameter $\tilde{b} = \sum_{i=1}^{N} y_i + b$, where a and b are the hyperparameters.

Suppose now that a customer just entered the sandwich store and we want to predict how much time will elapse until the next customer enters. Let y_* be the next inter-arrival time. The posterior predictive density is:

$$\mathrm{p}(y_*|\mathbf{y}) = \int_0^\infty \mathrm{p}(y_*|\lambda, \mathbf{y}) \cdot \pi(\lambda|\mathbf{y})\ \mathrm{d}\lambda = \int_0^\infty \mathrm{p}(y_*|\lambda) \cdot \pi(\lambda|\mathbf{y})\ \mathrm{d}\lambda$$

where $\mathrm{p}(y_*|\lambda, \mathbf{y})$ simplifies to $\mathrm{p}(y_*|\lambda)$, because, due to having assumed that each $y_i \sim \mathrm{Exp}(\lambda)$, the value of y_* does not depend on previous inter-arrival times once we condition on the value of λ. Plugging the formulas for $\mathrm{p}(y_*|\lambda)$ and $\pi(\lambda|\mathbf{y})$ in the last expression leads to:

$$\mathrm{p}(y_*|\mathbf{y}) = \int_0^\infty \lambda e^{-\lambda y_*} \cdot \frac{\tilde{b}^{\tilde{a}}}{\Gamma(\tilde{a})} \lambda^{\tilde{a}-1} e^{-\lambda \tilde{b}}\ \mathrm{d}\lambda = \frac{\tilde{b}^{\tilde{a}}}{\Gamma(\tilde{a})} \int_0^\infty \lambda^{\tilde{a}} e^{-\lambda\left(\tilde{b}+y_*\right)}\ \mathrm{d}\lambda$$

Finally, by evaluating the last integral, we obtain an expression which can be used to derive, for example, the expected value of y_* or the probability of y_* being within a certain interval:

$$\mathrm{p}(y_*|\mathbf{y}) = \tilde{a} \cdot \tilde{b}^{\tilde{a}} \cdot \left(\tilde{b} + y_*\right)^{-\tilde{a}-1}$$

1.3.6 Discussion

Both the Bayesian and the frequentist approaches to statistical inference can accomplish the three primary tasks of econometrics: parameter estimation, model comparison and prediction or forecasting. In the Bayesian approach uncertainty with respect to the values of the quantities of interest (parameters, forecasts, etc.) is considered from the outset by expressing prior beliefs or knowledge about the values of the parameters. These beliefs are then updated using evidence from the data and this process provides a natural way of reducing prior uncertainty. On the contrary, uncertainty with respect to the value of the quantities of interest in the frequentist approach is introduced using the conceptual device of repeated sampling. This device avoids introducing subjectivity into the analysis, at least at this stage, but makes interpretation of final results considerably more cumbersome.

Naturally, the Bayesian approach has been criticized for its need to express prior beliefs, because this introduces subjectivity into the analysis. This critique, however, loses much of its credibility once we consider what constitutes a model in both approaches. A model in Bayesian inference consists of the specification of a data-generating process, which leads to the likelihood function, and the specification of the prior density. In the frequentist approach only specification of the likelihood is needed. However, assumptions incorporated in the likelihood function can have a tremendous impact on the conclusions drawn from the analysis. That is, the frequentist approach is, obviously, not immune to misspecification. Of course, removing a component of a model that can cause problems reduces the risk of misspecification, but one should avoid taking an extreme stance on the issue: subjectivity can enter the analysis in many more ways than through the prior density.

The response from proponents of the Bayesian approach to this critique has been to construct generic ways of obtaining priors which have as small an impact on the final results as possible. Whether one chooses to use the proposed priors remains a matter of preference or convenience. In most models, when prior beliefs are adequately vague, they will be quickly dominated by the data in the sense that their impact on the final results will diminish as the number of observations increases. Therefore, given the same specification of a data-generating process and a large dataset, the frequentist and Bayesian approaches will produce very similar results.

A question that rises naturally is whether one should use the Bayesian or the frequentist approach when analyzing a phenomenon. Many textbooks on statistics have attempted in the past to answer this question in a definite way, most frequently using more arguments against the opposing approach rather than in favor of the approach they are advocating. The polemic shows strong tendencies to decline in recent years and a general consensus tends to emerge, according to which each approach has its benefits and shortcomings and one should exercise discretion when picking between them. Simple models are equally well tackled by either approach and, because frequentist statistical and econometric software packages are more readily available, researchers tend to use frequentist methods to estimate these models' parameters. The Bayesian approach appears to be preferred when considering more complicated models in which maximum likelihood "chokes". Therefore, the division between frequentist and Bayesian statisticians/econometricians tends to fade, with most researchers nowadays picking different methods to tackle different problems.

1.4 Estimation by Simulation

Bayes' theorem provides a convenient and intuitive device for performing the three primary tasks of econometrics. However, except in the simplest models, obtaining the posterior densities in closed form or summarizing the information conveyed by them in a way that is easy to comprehend involves multidimensional integration. Reliance on analytical solutions seriously limited the applicability of Bayesian methods in the past. The development or extension of techniques for obtaining random draws from multivariate distributions of non-standard form and the increasing speed of computers over the past decades provided an alternative to having to evaluate these integrals. Instead, simulation is used extensively nowadays to approximate

the integrals and much more complicated models can be considered. Due to the use of simulation methods, Bayesian estimation techniques are considerably more computationally intensive than using a frequentist approach, at least when considering simple models. As model complexity increases, however, the computational requirements of integration by simulation usually increase at a slower rate when compared to the optimization methods used in frequentist inference.

The following section reviews some fundamental results that justify the use of simulation methods to approximate complicated integrals, which can be expressed as expectations. This review is followed by a brief discussion of *Markov-chain Monte Carlo* (MCMC) techniques, where the methods are presented in an algorithmic fashion. The reader is directed to Chib (2001) or chapters 6 and 7 from Greenberg (2013) for a formal treatment of the matter.

1.4.1 The Strong Law of Large Numbers and a Central Limit Theorem

Summarizing the information contained in a posterior density, either of the parameters or of another quantity of interest, involves calculating the moments of the associated random variable. These moments can always be expressed as expectations and a law of large numbers can be invoked to justify approximation of the integrals by simulation.[6] There are a few versions of laws of large numbers, but to avoid going into the details of each one of them, we will be using the *Strong Law of Large Numbers* (SLLN).[7]

THEOREM 1.1: Strong Law of Large Numbers
Let $X_1, X_2, \ldots$ be a sequence of G independent and identically distributed random variables with $\mathrm{E}(|X|) < \infty$. Then:

$$\hat{\mu}_G \equiv \frac{\sum_{g=1}^{G} X_g}{G} \overset{\text{a.s.}}{\to} \mathrm{E}(X)$$

where "$\overset{\text{a.s.}}{\to}$" denotes almost sure convergence.

Formally defining *almost sure convergence* requires delving deep into the fundamentals of probability theory. Instead of going through a series of technical definitions, the usefulness of the SLLN in the context of estimation by simulation will be illustrated here. For this purpose, consider a model with a single parameter, θ, whose posterior density function is $\pi(\theta|\mathbf{y})$. To summarize the information contained in this posterior density we would like, first of all, to evaluate the expectation:

$$\mathrm{E}(\theta|\mathbf{y}) = \int_{\Theta} \theta \cdot \pi(\theta|\mathbf{y}) \, \mathrm{d}\theta \tag{1.19}$$

Suppose that we can obtain G random draws from the distribution whose probability density function is $\pi(\theta|\mathbf{y})$ and denote these draws by $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(G)}$. The SLLN states that, as G becomes larger, the sample mean of these random draws:

$$\bar{x}_\theta = \frac{\sum_{g=1}^{G} \theta^{(g)}}{G} \tag{1.20}$$

converges to $\mathrm{E}(\theta|\mathbf{y})$. Notice that G is controlled by the researcher and the larger this number is, the closer the sample mean of the draws is likely to be to the true expectation. Furthermore, the variance of θ can also be expressed as an expectation:

$$\mathrm{V}(\theta|\mathbf{y}) = \int_{\Theta} \left(\theta - \mathrm{E}(\theta|\mathbf{y})\right)^2 \cdot \pi(\theta|\mathbf{y}) \, \mathrm{d}\theta \tag{1.21}$$

[6] An alternative term used frequently in the context of Bayesian inference instead of simulation methods is is *Monte Carlo* methods.

[7] See Billingsley (1995) for a discussion of the differences between the versions of the laws of large numbers and, in particular, page 282 for a proof of the SLLN.

Given the same G draws from the posterior of θ, we can invoke the SLLN a second time to approximate this variance by the sample variance:

$$s_\theta^2 = \frac{\sum_{g=1}^{G} \left(\theta^{(g)} - \bar{x}_\theta\right)^2}{G-1} \tag{1.22}$$

Even probability statements about the value of θ can expressed as expectations. For example, the probability that θ is between two fixed numbers, c_1 and c_2, can be written as:

$$\text{Prob}(c_1 \leq \theta \leq c_2|\mathbf{y}) = \int_{c_1}^{c_2} \theta \cdot \pi\left(\theta|\mathbf{y}\right) \, \mathrm{d}\theta = \int_{\Theta} \mathbb{1}(c_1 \leq \theta \leq c_2) \cdot \pi\left(\theta|\mathbf{y}\right) \, \mathrm{d}\theta \tag{1.23}$$

where $\mathbb{1}(\cdot)$ is the *indicator function*: its value is equal to one if the statement it takes as an argument is true and zero otherwise. The last expression can be interpreted as an expectation of a function of θ and the SLLN suggests that the formula:

$$\frac{\sum_{g=1}^{G} \mathbb{1}\left(c_1 \leq \theta^{(g)} \leq c_2\right)}{G} \tag{1.24}$$

can be used to approximate this probability. In practical terms, this formula reduces to calculating the proportion of random draws from the posterior that fall into the $[c_1, c_2]$ interval, relative to the total number of draws.

The SLLN states that as G goes to infinity, the simulation-based approximation to the theoretical expectation becomes better and better, although in a stochastic sense: as we get more draws from the posterior the approximation could temporarily move away from the theoretical expectation, but we should be "almost certain" that, with a large enough G, these deviations will become so small that we can ignore them. In practice, however, we will always have to work with a finite G and the approximations will always remain imperfect. The *Central Limit Theorem* (CLT) presents a way of quantifying the probability of the approximation being a certain degree off the quantity it is meant to approximate.[8]

THEOREM 1.2: Central Limit Theorem
Let $X_1, X_2, \ldots$ be a sequence of G independent and identically distributed random variables with $\mathrm{E}(X) = \mu$ and $\mathrm{V}(X) = \sigma^2$. Then the sample mean, $\hat{\mu}_G \equiv \frac{1}{G}\sum_{g=1}^{G} X_g$, converges in distribution to a Normal:

$$\sqrt{G}\left(\hat{\mu}_G - \mu\right) \xrightarrow{\mathrm{d}} \mathrm{N}\left(0, \sigma^2\right)$$

Loosely speaking and in the context of the single-parameter model used above, the CLT suggests that, as G increases, the distribution of the discrepancy between the theoretical expectations and their simulation-based approximations[9] becomes indistinguishable from a Normal distribution with mean zero and variance $\frac{\sigma^2}{G}$. Based on this result, one can use the Normal distribution to evaluate the probability of $|\hat{\mu}_G - \mu|$ being above any given threshold. Therefore, it has become common practice to report, along with the Monte Carlo approximations to the expectations, the corresponding *Monte Carlo standard error*, defined as $\sqrt{s_\theta^2/G}$, and let the reader decide whether the approximation is precise enough.[10] Keep in mind that, in a Bayesian inference setting, G represents the number of draws from the posterior distribution and, as such, it is controlled by the researcher. By increasing the number of draws, the Monte Carlo standard error can be reduced, although, in practice, not indefinitely (Flegal et al., 2008).

[8] See Billingsley (1995, p.356-359) for a proof of the CLT.

[9] $\hat{\mu}_G - \mu$ is a random variable because $\hat{\mu}_G$ is a random variable. In turn, $\hat{\mu}_G$ is a random variable because it is a function of G random variables.

[10] Theoretically, the variance of $\hat{\mu}_G - \mu$ should be $\frac{\sigma^2}{G}$. However, because σ^2 is unknown, it is replaced by the sample variance, which is a consistent estimator of the population variance: as G goes to infinity, s_θ^2 converges in probability to σ^2.

♦ **Example 1.2 Customer Arrival Rate (Continued)**

Using a Gamma prior for the arrival-rate parameter, λ, in the sandwich-store example we expressed the posterior density of λ as a Gamma density with shape parameter $\tilde{a} = N + a$ and rate parameter $\tilde{b} = \sum_{i=1}^{N} y_i + b$, where a and b are the hyperparameters. We then used the properties of the Gamma distribution to calculate $\mathrm{E}(\lambda|\mathbf{y}) = \frac{\tilde{a}}{\tilde{b}}$ and $\mathrm{V}(\lambda|\mathbf{y}) = \frac{\tilde{a}}{\tilde{b}^2}$. We will now approximate these two posterior moments by drawing random numbers from $\pi(\lambda|\mathbf{y})$. Of course, in this simple problem there is no real need for simulation and the example is provided only for illustrating the use of simulation methods.

Towards this end, let's fix the values of the hyperparameters as $a = 1$ and $b = 0.25$. The following table presents the simulation-based estimates of the moments, along with the associated Monte Carlo standard errors, and for increasing numbers of draws from the posterior (G). Compare these results to the ones obtained using the analytical formulas: $\mathrm{E}(\lambda|\mathbf{y}) = 4.4359$ and $\mathrm{V}(\lambda|\mathbf{y}) = 0.1948$.

# of draws (G)	$\mathrm{E}(\lambda\|\mathbf{y})$	$\mathrm{V}(\lambda\|\mathbf{y})$	Monte Carlo standard error
100	4.4933	0.2486	0.0499
1,000	4.4407	0.2036	0.0143
10,000	4.4432	0.1934	0.0044

The results presented above can be obtained in BayES by changing the value of G in the code contained in the following box.

```
// import the data and get N and the sum of the values in y
Data = webimport("www.bayeconsoft.com/datasets/WaitingTimes.csv");
N = rows(Data);
sumy = sum(Data.y);

// calculate the posterior parameters
a_tilde = N + 1;
b_tilde = sumy + 0.25;

// draw samples from the posterior
G = 100;
x = gamrnd(a_tilde, b_tilde, G, 1);

// calculate the moments and the Monte Carlo standard error
E_lambda = mean(x);
V_lambda = var(x);
MCse = sqrt(V_lambda/G);

// print the results
print( [E_lambda, V_lambda, MCse] );
```

In closing this section, we note that the usefulness of the SLLN and the CLT extends beyond the case of single-parameter models. In multiple-parameter models, however, the expectations considered above need to be taken with respect to the marginal density of each parameter. Let's consider, for example, a model which involves two parameters, θ_1 and θ_2 and suppose that we can obtain G random draws from this joint density. Denote these draws by $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots, \boldsymbol{\theta}^{(G)}$, where each $\boldsymbol{\theta}^{(g)}$ is a two-dimensional vector. Equations (1.9) and (1.10) suggest that, to approximate $\mathrm{E}(\theta_1|\mathbf{y})$, one needs to first integrate-out θ_2 from $\pi(\theta_1, \theta_2|\mathbf{y})$:

$$\mathrm{E}(\theta_1|\mathbf{y}) = \int_{\Theta_1} \theta_1 \cdot \pi(\theta_1|\mathbf{y})\, \mathrm{d}\theta_1 = \int_{\Theta_1} \theta_1 \cdot \left[\int_{\Theta_2} \pi(\theta_1, \theta_2|\mathbf{y})\, \mathrm{d}\theta_2 \right] \mathrm{d}\theta_1 \tag{1.25}$$

Given that the draws are from the joint posterior density of θ_1 and θ_2, simply summarizing the values of θ_1 contained in $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots, \boldsymbol{\theta}^{(G)}$ would take care of both integrals:

$$\mathrm{E}(\theta_1|\mathbf{y}) \approx \frac{\sum_{g=1}^{G} \theta_1^{(g)}}{G} \tag{1.26}$$

Therefore, the posterior moments of a parameter, as obtained by simulation, are always marginal with respect to the values of the remaining parameters in the model.

1.4.2 Markov-Chain Monte Carlo (MCMC)

The results presented in the preceding section can be used to summarize the properties of the posterior distribution of a model's parameters, $\pi(\boldsymbol{\theta}|\mathbf{y})$. However, their application requires a procedure for obtaining independent random draws from this posterior distribution; something that is not straightforward except in very simple cases. The term *Markov-chain Monte Carlo* (MCMC) is used to denote a set of closely related methods that are designed to generate random draws from complex distributions. The associated algorithms work by constructing and drawing random numbers from Markov chains, whose stationary distributions are the same as *target distribution*, which in a Bayesian estimation problem, is simply the posterior distribution of the parameters.

The discussion on why and under what conditions these algorithms work in practice becomes very technical, very quickly and, as such, goes beyond the purposes of this book. The interested reader is directed to Chib (2001) for a complete presentation. The algorithms themselves, however, and the intuition behind them will be provided here, because understanding how MCMC methods work is essential for interpreting the results of the algorithms, as well as avoiding some pitfalls in applying them.

Before we proceed we note that, when the draws from the posterior are generated using Markov chains, they are no longer independent. Therefore, the SLLN and CLT that we encountered before no longer apply. Nevertheless, versions of the two theorems exist when the random draws are generated from weakly-dependent processes.[11] The important difference in the case of correlated draws is that, given a sequence of draws from the Markov chain, $X_1, X_2, \ldots$, the Monte Carlo standard error, $\sqrt{\sigma^2/G}$, now involves the variance:

$$\sigma^2 = \mathrm{V}(X_1) + 2\sum_{j=2}^{\infty} \mathrm{Cov}(X_1, X_j) \tag{1.27}$$

where $\mathrm{V}(X_1)$ is the variance of a random variable that follows the same distribution as the stationary distribution of the Markov chain, and $\mathrm{Cov}(X_1, X_j)$ is the covariance of this random variable with another variable, j steps ahead in the Markov chain. If the process that generates the draws is weakly dependent, then this covariance will go to zero as j tends to infinity. The practical implication of this result is that, when the draws are autocorrelated, the Monte Carlo standard error is larger than what could be achieved with independent draws. The *inefficiency factor* in this context is defined as:

$$\hat{\kappa} = \frac{\hat{\sigma}^2}{s^2/G} \tag{1.28}$$

where $\hat{\sigma}^2$ is an estimate of the quantity in (1.27) and s^2 an estimate of the variance of the draws, were they independent. Obtaining $\hat{\sigma}^2$ and s^2 presents challenges and, apart from the theoretical justification of this quantity, Chib (2001) provides a range of approaches for estimating $\hat{\sigma}^2$. The inefficiency factor is also known as the *autocorrelation time* and the inverse of it was first defined in Geweke (1992) as the *relative numerical efficiency*.

In expectation, the inefficiency factor will be greater than one and it can be interpreted as the factor by which one needs to divide the number of autocorrelated draws obtained from an MCMC sampler, G, to get the number of independent draws, $\tilde{G}$, that would lead to the same Monte Carlo standard error. $\tilde{G}$ is appropriately called the *effective sample size*. When designing an MCMC sampling scheme, the algorithm should be tuned such that the inefficiency factor is as close to unity as possible, or, to put it differently, to reduce the autocorrelation of draws from the posterior. If no such effort is made or if the problem is ill-conditioned given the data, then the information content of the draws may be very limited and a vast amount or draws may be needed until the Monte Carlo standard error is reduced to reasonable levels.

[11] See for example Theorem 27.4 in Billingsley (1995, p.364) or Chan & Geyer (1994).

The Metropolis-Hastings algorithm

The most general MCMC method was first proposed by Metropolis et al. (1953) and extended by Hastings (1970), leading to an algorithm known as *Metropolis-Hastings*. Almost all other algorithms used in Bayesian inference can be viewed as special cases of the Metropolis-Hastings algorithm. This algorithm comes in many flavors and can be adjusted to take advantage of the specificities of a particular problem. To fix ideas, we will consider a model with K parameters and a posterior density $\pi(\boldsymbol{\theta}|\mathbf{y})$. The algorithm starts by fixing an initial value for $\boldsymbol{\theta}$. Based on this initial value it proposes a move to a new value, $\boldsymbol{\theta}^\star$, using a *proposal density*, $q(\boldsymbol{\theta}, \boldsymbol{\theta}^\star|\mathbf{y})$. The proposal density is chosen by the researcher and, at least in theory, could be any proper probability density function. Lastly, the proposed value, $\boldsymbol{\theta}^\star$, is accepted with probability:

$$\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^\star|\mathbf{y}) = \min\left\{\frac{\pi(\boldsymbol{\theta}^\star|\mathbf{y})}{\pi(\boldsymbol{\theta}|\mathbf{y})} \cdot \frac{q(\boldsymbol{\theta}^\star, \boldsymbol{\theta}|\mathbf{y})}{q(\boldsymbol{\theta}, \boldsymbol{\theta}^\star|\mathbf{y})}, 1\right\} \tag{1.29}$$

This means that the new value of $\boldsymbol{\theta}$ becomes $\boldsymbol{\theta}^\star$ and the process is repeated multiple times. If the move to $\boldsymbol{\theta}^\star$ is rejected, the state of the Markov chain remains unchanged in the current iteration. The product of the two fractions inside the minimization operator is known as the *Metropolis-Hastings ratio*.

Very frequently in practice the proposal density is chosen to be the multivariate Normal, centered at the current value of $\boldsymbol{\theta}$ and with covariance matrix $\mathbf{C}$:

$$q(\boldsymbol{\theta}, \boldsymbol{\theta}^\star|\mathbf{y}) = \frac{|\mathbf{C}|^{-1/2}}{(2\pi)^{K/2}} \exp\left\{-\tfrac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^\star)' \mathbf{C}^{-1}(\boldsymbol{\theta} - \boldsymbol{\theta}^\star)\right\} \tag{1.30}$$

leading to the *random-walk Metropolis-Hastings* algorithm. This choice is convenient because $q(\boldsymbol{\theta}^\star, \boldsymbol{\theta}|\mathbf{y}) = q(\boldsymbol{\theta}, \boldsymbol{\theta}^\star|\mathbf{y})$ for all $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^\star$ and, thus, only the ratio of posterior densities needs to be evaluated when calculating the acceptance probability.

Notice that the acceptance probability involves the ratio of the posterior evaluated at $\boldsymbol{\theta}^\star$ and $\boldsymbol{\theta}$. Thus, for the Metropolis-Hastings algorithm to be applied one needs to know the posterior density only up to a constant of proportionality. This fact makes the algorithm very well-suited for Bayesian parameter estimation, given that constants of proportionality are frequently unknown (see, for example, equation (1.7)).

The only thing left to do before applying the random-walk version of algorithm is to chose the value of the covariance matrix, $\mathbf{C}$, in the proposal. In theory, any positive-definite matrix would do. However, different choices would lead to different degrees of autocorrelation of the draws, which in turn, may have severe consequences for the computational efficiency of the algorithm. One simple choice is to set $\mathbf{C}$ equal to $\mathcal{T} \cdot \mathbf{I}_K$, where $\mathbf{I}_K$ is the $K \times K$ identity matrix and $\mathcal{T}$ is a tuning parameter, with its value chosen such that approximately 30%-45% of the proposed moves are accepted. This acceptance rate is an approximation to the optimal acceptance rate when the target distribution is multivariate Normal, and the precise value depends on the value of K (Roberts et al., 1997). Intuitively, when $\mathcal{T}$ is set to a large value, the proposed $\boldsymbol{\theta}^\star$ may be too erratic and, therefore, rarely accepted. When $\mathcal{T}$ is set to a small value, the proposed $\boldsymbol{\theta}^\star$ will be frequently accepted, but this will be because $\boldsymbol{\theta}^\star$ is very close to the current state of the Markov chain, $\boldsymbol{\theta}$. In both cases the draws will be highly autocorrelated.

Finally, to avoid dependence of the results on the initially chosen state of the chain, it is common practice to let the Markov chain run for a few iterations before start storing the draws. This is called the *burn-in* phase of the algorithm. It is important to run a burn-in because the Metropolis-Hastings algorithm produces draws from a Markov chain whose stationary distribution is the same as the target distribution, $\pi(\boldsymbol{\theta}|\mathbf{y})$. If the chain starts at an initial value far away from its stationary distribution, the initial draws will not be from the target distribution because the chain is still moving towards its stationary distribution. Additional adjustments can be made during this burn-in phase, for example, getting a better value for the tuning parameter, $\mathcal{T}$, or a rough estimate of the covariance matrix of $\boldsymbol{\theta}$ such that the proposal density is further tailored to the specific model and dataset.

The simplest form of the Metropolis-Hastings algorithm, as described above, is given in Algorithm 1.1 and an application in the context of the sandwich-store example follows, where the algorithm is implemented in BayES' language.

Algorithm 1.1 Simple Metropolis-Hastings

set the number of burn-in iterations, D
set the number of draws to be retained, G
set $\boldsymbol{\theta}$ to a reasonable starting value
for $g = 1:(D+G)$ **do**
  draw $\boldsymbol{\theta}^\star$ from the proposal, $q(\boldsymbol{\theta}, \boldsymbol{\theta}^\star|\mathbf{y})$
  accept the move (set current $\boldsymbol{\theta}$ equal to $\boldsymbol{\theta}^\star$) with probability:

$$\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^\star|\mathbf{y}) = \min\left\{\frac{\pi(\boldsymbol{\theta}^\star|\mathbf{y})}{\pi(\boldsymbol{\theta}|\mathbf{y})}\frac{q(\boldsymbol{\theta}^\star, \boldsymbol{\theta}|\mathbf{y})}{q(\boldsymbol{\theta}, \boldsymbol{\theta}^\star|\mathbf{y})}, 1\right\}$$

  if $g > D$ **then**
    store the current value of $\boldsymbol{\theta}$
  end if
end for

♦ **Example 1.2 Customer Arrival Rate (Continued)**
Consider for a last time the sandwich-store example in which we will continue to assume that each inter-arrival time, y_i, is a draw from an Exponential distribution with rate λ, but we will now use a log-Normal prior, with hyperparameters m and s^2:

$$p(\lambda) = \frac{1}{\lambda\sqrt{2\pi s^2}}\exp\left\{-\frac{(\log\lambda - m)^2}{2s^2}\right\}$$

With this choice of prior, the posterior density of λ becomes:

$$\pi(\lambda|\mathbf{y}) \propto \lambda^{N-1} \times \exp\left\{-\lambda\sum_{i=1}^{N} y_i - \frac{(\log\lambda - m)^2}{2s^2}\right\}$$

The posterior density does not belong to any known parametric family and we have to use simulation to summarize the information contained in it. Because λ must be positive, we will use a log-Normal proposal, with location parameter equal to the logarithm of λ in the current iteration:

$$q(\lambda, \lambda^\star|\mathbf{y}) = \frac{\mathcal{T}^{-1/2}}{\lambda^\star(2\pi)^{1/2}}\exp\left\{-\frac{(\log\lambda^\star - \log\lambda)^2}{2\mathcal{T}}\right\}$$

where $\mathcal{T}$ will be used as the tuning parameter. The logarithm of the Metropolis-Hastings ratio becomes:

$$\begin{aligned}\log \mathsf{MH}(\lambda, \lambda^\star) &= \log\pi(\lambda^\star|\mathbf{y}) - \log\pi(\lambda|\mathbf{y}) + \log q(\lambda^\star, \lambda|\mathbf{y}) - \log q(\lambda, \lambda^\star|\mathbf{y}) \\ &= N(\log\lambda^\star - \log\lambda) - (\lambda^\star - \lambda)\sum_{i=1}^{N} y_i - \frac{(\log\lambda^\star - m)^2}{2s^2} + \frac{(\log\lambda - m)^2}{2s^2}\end{aligned}$$

After setting the values of the hyperparameters as $m = 1.4$ and $s^2 = 0.9$, as well as the value of the tuning parameter as $\mathcal{T} = 0.3$, we are ready to implement the algorithm. Note that these values of the hyperparameters lead to a prior density for λ similar to a Gamma density with shape $a = 2$ and rate $b = 0.5$, so the results obtained here should be comparable to the ones we got with the Gamma prior. Additionally, the value of $\mathcal{T}$ is chosen such that approximatively 38% of the proposed moves are accepted.

An implementation of the algorithm in BayES' language is given in the following box. The first two posterior moments of λ obtained after running this code are $\mathrm{E}(\lambda|\mathbf{y}) = 4.4445$ and $\mathrm{V}(\lambda|\mathbf{y}) = 0.1907$, respectively.

```
// import the data and get N and the sum of the values in y
Data = webimport("www.bayeconsoft.com/datasets/WaitingTimes.csv");
N = rows(Data);
sumy = sum(Data.y);

// set the values of the hyperparameters and the tuning parameter
m = 1.4;    s2 = 0.9;   tuning = 0.3;

// set the number of iterations
D = 3000;        // # of burn-in iterations
G = 10000;       // # of retained draws

// set the starting value for lambda and calculate its logarithm
lambda = 2;
loglambda = log(lambda);

// initialize a vector to store the draws
draws = zeros(G,1);

// start the algorithm
for (g=1:D+G)
    // draw log-lambda star from the proposal and calculate its exponential
    loglambda_star = normrnd(loglambda,tuning);
    lambda_star = exp(loglambda_star);

    // calculate the logarithm of the Metropolis-Hastings ratio
    logMH = N*(loglambda_star-loglambda)
        - (lambda_star-lambda)*sumy
        - ((loglambda_star-m)^2 - (loglambda-m)^2)/(2*s2);

    // accept/reject the proposed move
    if ( log(unifrnd()) < logMH )
        lambda = lambda_star;
        loglambda = loglambda_star;
    end

    // store the results from the current iteration ================
    if (g>D)
        draws(g-D) = lambda;
    end
end

// summarize the draws from the posterior
print( [mean(draws); var(draws)] );
```

The multiple-block Metropolis-Hastings algorithm

In complex models that contain multiple parameters to be estimated, the simple version of the Metropolis-Hastings algorithm may become very inefficient, in the sense that it produces very highly autocorrelated draws from the posterior. This is because, when $\boldsymbol{\theta}$ has high dimensions or it contains parameters with different roles in the model, such as location and scale parameters, it becomes harder to tailor the proposal to the specific model and dataset. The *multiple-block Metropolis-Hastings* algorithm is an extension to the algorithm described above, which is designed to work in such complex circumstances.

The multiple-block version of the Metropolis-Hastings algorithm works by first partitioning the parameter vector, $\boldsymbol{\theta}$, into $B \geq 2$ blocks, $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \ldots, \boldsymbol{\theta}_B$. Next, the density of each block, $\boldsymbol{\theta}_b$, called this block's *full conditional*, is obtained from the posterior density of $\boldsymbol{\theta}$ by treating the parameters contained in all other blocks except $\boldsymbol{\theta}_b$ as fixed. The full conditional for block $\boldsymbol{\theta}_b$ is denoted by $\pi\left(\boldsymbol{\theta}_b|\mathbf{y}, \boldsymbol{\theta}_1 \ldots, \boldsymbol{\theta}_{b-1}, \boldsymbol{\theta}_{b+1} \ldots \boldsymbol{\theta}_B\right)$, or, more compactly, by $\pi\left(\boldsymbol{\theta}_b|\bullet\right)$. In this notation the origin of the term "full conditional" becomes apparent: $\pi\left(\boldsymbol{\theta}_b|\bullet\right)$ is the density of $\boldsymbol{\theta}_b$ conditional on everything else in the model, both data and parameters. Quantities that involve the parameters in other blocks and which enter the posterior density multiplicatively, become

part of the constant of proportionality of $\pi(\boldsymbol{\theta}_b|\bullet)$. Quantities that enter the posterior density in forms other than multiplicative remain in the full conditional and, during the implementation of the algorithm, are evaluated using these parameters' current values.

Once the full conditionals of all blocks have been derived and simplified, the values of parameters in each block are updated in succession using the simple form of the Metropolis-Hastings algorithm. A compete iteration of the multiple-block Metropolis-Hastings involves B steps: values $\boldsymbol{\theta}_b^\star$ are proposed from block b's proposal density and they are accepted or rejected using the acceptance probability in (1.29) and while using $\pi(\boldsymbol{\theta}_b|\bullet)$ in place of $\pi(\boldsymbol{\theta}|\mathbf{y})$. This process is then repeated for the next block and while conditioning on the current values of all other parameters, irrespective of whether previous moves have been accepted or rejected.

Without further discussion, the multiple-block Metropolis-Hastings algorithm is given in Algorithm 1.2. An application of the algorithm in a problem of estimating the parameters of a Normal distribution follows.

Algorithm 1.2 Multiple-Block Metropolis-Hastings

set the number of burn-in iterations, D
set the number of draws to be retained, G
set $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_B$ to reasonable starting values
for $g = 1:(D+G)$ **do**
  for $b = 1:B$ **do**
    draw $\boldsymbol{\theta}_b^\star$ from its proposal, $q(\boldsymbol{\theta}_b, \boldsymbol{\theta}_b^\star|\mathbf{y})$
    accept the move (set current $\boldsymbol{\theta}_b$ equal to $\boldsymbol{\theta}_b^\star$) with probability:

$$\alpha_b(\boldsymbol{\theta}_b, \boldsymbol{\theta}_b^\star|\mathbf{y}) = \min\left\{\frac{\pi(\boldsymbol{\theta}_b^\star|\bullet)}{\pi(\boldsymbol{\theta}_b|\bullet)}\frac{q(\boldsymbol{\theta}_b^\star, \boldsymbol{\theta}_b|\mathbf{y})}{q(\boldsymbol{\theta}_b, \boldsymbol{\theta}_b^\star|\mathbf{y})}, 1\right\}$$

  end for

  if $g > D$ **then**
    store the current value of $\boldsymbol{\theta}$
  end if
end for

♦ **Example 1.3 Crab Size**

In this example we will consider part of the dataset used by Brockmann (1996) to examine the mating patterns of horseshoe crabs. The dataset consists of 173 observations on multiple characteristics of female crabs, but we will use only the variable which measures, in centimeters, the crab's carapace width. We will assume that the natural logarithm of carapace width, y_i, for each potential observation, i, is a draw from a Normal distribution with mean μ and variance σ^2. This assumption precludes negative values for the carapace width, which are physically impossible, as it implies that width itself follows a log-Normal distribution. When working with scale parameters in Bayesian inference, notation becomes considerably simpler if we re-parameterize the problem in terms of the *precision parameter*, $\tau \equiv \frac{1}{\sigma^2}$. Therefore, the model suggests that $y_i \sim \mathrm{N}\left(\mu, \frac{1}{\tau}\right)$, leading to the likelihood:

$$\mathrm{p}(\mathbf{y}|\mu,\tau) = \prod_{i=1}^{N}\frac{\tau^{1/2}}{(2\pi)^{1/2}}\exp\left\{-\frac{\tau(y_i-\mu)^2}{2}\right\} = \frac{\tau^{N/2}}{(2\pi)^{N/2}}\exp\left\{-\frac{\tau}{2}\sum_{i=1}^{N}(y_i-\mu)^2\right\}$$

We will place a Normal prior on μ, with mean m and precision t, and a Gamma prior on τ with shape and rate parameters a and b, respectively. Application of Bayes' rule leads to:

$$\pi(\mu,\tau|\mathbf{y}) \propto \tau^{N/2}\exp\left\{-\frac{\tau}{2}\sum_{i=1}^{N}(y_i-\mu)^2\right\} \times \exp\left\{-\frac{t}{2}(\mu-m)^2\right\} \times \tau^{a-1}e^{-b\tau}$$

We will treat μ and τ as the two blocks of the Metropolis-Hastings algorithm, with full conditionals:

$$\pi(\mu|\bullet) \propto \exp\left\{-\frac{\tau}{2}\sum_{i=1}^{N}(y_i-\mu)^2 - \frac{t}{2}(\mu-m)^2\right\}$$

and:

$$\pi\left(\tau|\bullet\right) \propto \tau^{N/2+a-1} \exp\left\{-\tau\left[\frac{1}{2}\sum_{i=1}^{N}\left(y_i-\mu\right)^2+b\right]\right\}$$

respectively. Finally, we will use a Normal distribution as a proposal for μ and a log-Normal distribution for τ. The mean of the proposal for μ will be the current value of parameter and its precision parameter, $\mathcal{T}_\mu$, will be used as the tuning parameter. The location parameter for τ's proposal is set equal to the logarithm of the current value of τ and its scale parameter, $\mathcal{T}_\tau$, is used for tuning (see the previous example for more details).

An implementation of the multiple-block Metropolis-Hastings algorithm in BayES' language is given in the following box. Running this code in BayES produces the results in following table.

Parameter	Mean	Variance
μ	3.26548	3.104e-05
τ	157.192	278.411
σ	0.08010	1.834e-05

```
// import the data into a dataset called Data
Data = webimport("www.bayeconsoft.com/datasets/CrabSize.csv");

// set the values of the hyperparameters
m = 0;      t = 0.001;  // prior mean and precision for mu
a = 0.001;  b = 0.001;  // prior shape and rate for tau

// set the number of iterations
D = 3000;   // # of burn-in iterations
G = 10000;  // # of retained draws

// set starting values for mu and tau and take the logarithm of tau
mu = 0;     tau = 1;    logtau = log(tau);

// set the values of the tuning parameters
tuning_mu = 5.0;    tuning_tau = 0.3;

// initialize a matrix to store the draws
draws = zeros(G,2);

// calculate some quantities used multiple times
y = log(Data.width);
a_tilde = 0.5*rows(y) + a;

// start the algorithm
for (g=1:D+G)
    // sample for mu ================================================
    mu_star = normrnd(mu,1/sqrt(tuning_mu));
    logMH = -0.5*tau*(sum((y-mu_star).^2) - sum((y-mu).^2))
        -0.5*t*((mu_star-m)^2 - (mu-m)^2);
    if ( log(unifrnd()) < logMH )
        mu = mu_star;
    end

    // sample for tau ===============================================
    logtau_star = normrnd(logtau,tuning_tau);
    tau_star = exp(logtau_star);
    logMH = a_tilde*(logtau_star - logtau)
        -(tau_star-tau)*(0.5*sum((y-mu).^2) + b);
    if ( log(unifrnd()) < logMH )
        tau = tau_star;
        logtau = logtau_star;
    end

    // store the results from the current iteration ================
    if (g>D)
        draws(g-D,1) = mu;
        draws(g-D,2) = tau;
    end
end
```

```
// add sigma = 1/sqrt(tau) to the draws matrix and summarize
draws_sigma = ones(G,1) ./ sqrt(draws(:,2));
draws = [draws, draws_sigma];
print( [mean(draws); var(draws)] );
```

The Gibbs Algorithm

The Gibbs algorithm has its roots in statistical physics and the work of Josiah Willard Gibbs. It was first described by Geman & Geman (1984), who named the algorithm the *Gibbs sampler*, and became popular among statisticians after Gelfand & Smith (1990) demonstrated its general applicability to Bayesian inference problems. The Gibbs sampler can be viewed as a special case of the multiple-block Metropolis-Hastings algorithm in the case where the full conditionals of all blocks of parameters belong to known parametric families.

To fix ideas, consider a problem that involves K parameters and suppose that the parameter vector, $\boldsymbol{\theta}$, has been partitioned into B blocks, $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_B$. Suppose also that the full conditional of block $\boldsymbol{\theta}_b$, $\pi(\boldsymbol{\theta}_b|\bullet)$, takes a form that can be recognized as the probability density function of a distribution for which there exist fast algorithms to generate random draws from.[12] Then, the proposal density for $\boldsymbol{\theta}_b$, $q(\boldsymbol{\theta}_b, \boldsymbol{\theta}_b^\star|\mathbf{y})$ can be set to be independent of the current values of $\boldsymbol{\theta}$ and equal to $\pi(\boldsymbol{\theta}_b^\star|\bullet)$. In this case the Metropolis-Hastings ratio simplifies to one for all values of $\boldsymbol{\theta}_b^\star$ and the proposed move is accepted with probability one. This fact simplifies the procedure of updating the values of $\boldsymbol{\theta}_b$ considerably and, if these simplifications can be performed for all blocks of $\boldsymbol{\theta}$, one can implement a pure Gibbs sampler. If the full conditionals of only some of the blocks can be derived without missing a constant of proportionality, then Gibbs updates can be used for these blocks and complete Metropolis-Hastings updates for the remaining blocks. A term that is used for such a hybrid algorithm is *Metropolis-Hastings within Gibbs*.

A pure Gibbs sampler is much simpler and succinct than the multiple-block Metropolis-Hastings algorithm and it is given in Algorithm 1.3. Following this, the algorithm is implemented in BayES' language in the context of the problem of estimating the mean and variance of horseshoe crab carapace width, assuming that the logarithm of this width follows a Normal distribution.

Algorithm 1.3 Gibbs Sampler

set the number of burn-in iterations, D
set the number of draws to be retained, G
set $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_B$ to reasonable starting values
for $g = 1:(D+G)$ **do**
  for $b = 1:B$ **do**
    draw $\boldsymbol{\theta}_b^\star$ from $\pi(\boldsymbol{\theta}_b^\star|\bullet)$ and set the current value of $\boldsymbol{\theta}_b$ to $\boldsymbol{\theta}_b^\star$
  end for

  if $g > D$ **then**
    store the current value of $\boldsymbol{\theta}$
  end if
end for

♦ **Example 1.3 Crab Size (Continued)**
Consider again the problem of estimating the parameters of a Normal distribution for the random variable that measures the natural logarithm of a crab's carapace width. Using a Normal prior for the

[12] In most cases where the full conditional is a member of a known parametric family, $\pi(\boldsymbol{\theta}_b|\bullet)$ can be derived without missing a constant of proportionality and the term *complete conditional* may be used to refer to $\pi(\boldsymbol{\theta}_b|\bullet)$.

mean parameter, μ, of this distribution and a Gamma prior for the precision parameter, τ, we obtained the following full conditionals:

$$\pi\left(\mu|\bullet\right) \propto \exp\left\{-\frac{\tau}{2}\sum_{i=1}^{N}\left(y_i-\mu\right)^2-\frac{t}{2}\left(\mu-m\right)^2\right\}$$

and:

$$\pi\left(\tau|\bullet\right) \propto \tau^{N/2+a-1}\exp\left\{-\tau\left[\frac{1}{2}\sum_{i=1}^{N}\left(y_i-\mu\right)^2+b\right]\right\}$$

respectively. Let's start from the full conditional of τ. This density resembles the probability density function of Gamma-distributed random variable with shape and rate parameters $\tilde{a}=\frac{N}{2}+a$ and $\tilde{b}=\frac{1}{2}\sum_{i=1}^{N}\left(y_i-\mu\right)^2+b$, respectively. The only thing missing is the constant of proportionality, which can be derived from the fact that the probability density function of a random variable must integrate to unity. Therefore:

$$\tau|\mathbf{y},\mu\sim\text{Gamma}\left(\tilde{a},\tilde{b}\right)$$

Additional algebra is required to to show that the full conditional of μ is proportional to $\exp\left\{-\frac{\tilde{t}}{2}\left(\mu-\tilde{m}\right)^2\right\}$, with $\tilde{t}=\tau N+t$ and $\tilde{m}=\frac{1}{\tilde{t}}\left(\tau\sum_{i=1}^{N}y_i+t\cdot m\right)$. The steps involved in this derivation are expanding the squares in the original expression for the full conditional of μ, simplifying and, finally, completing the square by adding and subtracting $\tilde{m}^2$ inside the $\exp$ operator. We can now recognize that the full conditional of μ resembles a Normal probability density function with mean $\tilde{m}$ and precision parameter $\tilde{t}$. With $\frac{\tilde{t}^{1/2}}{(2\pi)^{1/2}}$ as the constant of proportionality, we get:

$$\mu|\mathbf{y},\tau\sim\text{N}\left(\tilde{m},\tfrac{1}{\tilde{t}}\right)$$

We are now ready to implement a Gibbs sampler for this problem. The code contained in the following box provides an implementation in BayES' language. Running this code in BayES produces the results in following table.

Parameter	Mean	Variance
μ	3.26638	3.713e-05
τ	156.934	289.268
σ	0.08018	1.928e-05

which are almost the same as the ones obtained using the multiple-block Metropolis-Hastings algorithm. Any discrepancies in the results between the two algorithms are solely due to approximation error of the integrals, inherent in Monte Carlo methods. Using longer chains should further reduce these discrepancies.

```
// import the data into a dataset called Data
Data = webimport("www.bayeconsoft.com/datasets/CrabSize.csv");

// set the values of the hyperparameters
m = 0;      t = 0.001;  // prior mean and precision for mu
a = 0.001;  b = 0.001;  // prior shape and rate for tau

// set the number of iterations
D = 3000;   // # of burn-in iterations
G = 10000;  // # of retained draws

// set the starting values for mu and tau
mu = 0;     tau = 1;

// initialize a matrix to store the draws
draws = zeros(G,2);

// calculate some quantities used multiple times
y = log(Data.width);
N = rows(y);
sumy = sum(y);
tm = t*m;
a_tilde = 0.5*N + a;
```

```
// start the algorithm
for (g=1:D+G)
    // sample for mu ============================================
    t_tilde = tau*N + t;
    m_tilde = (tau*sumy + tm)/t_tilde;
    mu = normrnd(m_tilde, 1/sqrt(t_tilde));

    // sample for tau ===========================================
    b_tilde = 0.5*sum((y-mu).^2) + b;
    tau = gamrnd(a_tilde, b_tilde);

    // store the results from the current iteration ==============
    if (g>D)
        draws(g-D,1) = mu;
        draws(g-D,2) = tau;
    end
end

// add sigma = 1/sqrt(tau) to the draws and summarize
draws_sigma = ones(G,1) ./ sqrt(draws(:,2));
draws = [draws, draws_sigma];
print( [mean(draws); var(draws)] );
```

General Comments on MCMC Methods and Extensions

Before closing this section we make some brief remarks on the application and use of MCMC methods:

1. The Gibbs sampler was presented here as a special case of the Metropolis-Hastings algorithm. However, other methods can be used within a Gibbs sampler to generate random draws for a block, even if the full conditional of this block is known only up to a constant of proportionality. Choices include, methods based on *rejection sampling* and its extensions or *composition sampling*. The interested reader is referred to Chib (2001) or chapter 10 from Gelman et al. (2013) for details.
2. The Gibbs sampler and the multiple-block Metropolis-Hastings algorithm can be shown to work when, in every iteration, a single block, $\boldsymbol{\theta}_b$, is chosen at random from the B blocks and updated, either by sampling directly from its full conditional or using its proposal and accept/reject steps. In practice however, almost invariably, the algorithms are implemented in a way such that all blocks are updated sequentially, not randomly, in an iteration. This is done for two reasons: it is slightly easier to implement an algorithm that works sequentially and, most importantly, sequential updating tends to produce less autocorrelated draws from the posterior.
3. When the full conditional of a parameter block belongs to a known family of distributions, most statistical packages will provide built-in procedures for sampling from this distribution directly. Even if this is not the case, the *inverse probability transform* method can be used, especially if the full conditional of the block under consideration contains only a single parameter. Sampling directly from the full conditional should be preferred over using a Metropolis-Hastings step for two reasons: (i) producing a draw directly from a distribution most often involves many fewer computations than evaluating the Metropolis-Hastings ratio and then deciding whether to accept or reject the proposed move, and (ii) direct sampling usually results in much lower inefficiency factors than Metropolis-Hastings updates, thus requiring fewer samples from the posterior to approximate the integrals to a given degree of accuracy.
4. An extension to the simple Gibbs sampler, called *collapsed Gibbs sampler* can take advantage of analytical integration results, when these are available, to reduce the degree of autocorrelation in the draws generated by the sampler. A collapsed Gibbs sampler works by analytically marginalizing a block from one or more full conditionals. For example, consider a model that contains three blocks of parameters, $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$ and $\boldsymbol{\theta}_3$. Suppose also

that $\boldsymbol{\theta}_2$ can be integrated out analytically from the joint density of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ given $\boldsymbol{\theta}_3$, resulting in an expression of the form:

$$\pi\left(\boldsymbol{\theta}_1|\mathbf{y},\boldsymbol{\theta}_3\right) = \int_{\Theta_2} \pi\left(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2|\mathbf{y},\boldsymbol{\theta}_3\right) \mathrm{d}\boldsymbol{\theta}_2 \tag{1.31}$$

If sampling from $\pi\left(\boldsymbol{\theta}_1|\mathbf{y},\boldsymbol{\theta}_3\right)$ can be accomplished easily, a collapsed Gibbs sampler would take draws from $\pi\left(\boldsymbol{\theta}_1|\mathbf{y},\boldsymbol{\theta}_3\right)$ instead of $\pi\left(\boldsymbol{\theta}_1|\mathbf{y},\boldsymbol{\theta}_2,\boldsymbol{\theta}_3\right)$ and from the regular full conditionals of $\boldsymbol{\theta}_2$ and $\boldsymbol{\theta}_3$. In such a simple setting, the collapsed Gibbs sampler can be justified by blocking $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ together: instead of sampling iteratively from $\pi\left(\boldsymbol{\theta}_1|\mathbf{y},\boldsymbol{\theta}_2,\boldsymbol{\theta}_3\right)$, $\pi\left(\boldsymbol{\theta}_2|\mathbf{y},\boldsymbol{\theta}_1,\boldsymbol{\theta}_3\right)$ and $\pi\left(\boldsymbol{\theta}_3|\mathbf{y},\boldsymbol{\theta}_1,\boldsymbol{\theta}_2\right)$, one could think of $(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)$ as constituting a single block and sample iteratively from $\pi\left(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2|\mathbf{y},\boldsymbol{\theta}_3\right)$ and $\pi\left(\boldsymbol{\theta}_3|\mathbf{y},\boldsymbol{\theta}_1,\boldsymbol{\theta}_2\right)$. The former density can be expressed as:

$$\pi\left(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2|\mathbf{y},\boldsymbol{\theta}_3\right) = \pi\left(\boldsymbol{\theta}_2|\mathbf{y},\boldsymbol{\theta}_1,\boldsymbol{\theta}_3\right) \times \pi\left(\boldsymbol{\theta}_1|\mathbf{y},\boldsymbol{\theta}_3\right) \tag{1.32}$$

thus justifying sampling from $\pi\left(\boldsymbol{\theta}_1|\mathbf{y},\boldsymbol{\theta}_3\right)$ instead of sampling from $\pi\left(\boldsymbol{\theta}_1|\mathbf{y},\boldsymbol{\theta}_1,\boldsymbol{\theta}_3\right)$. An issue to recognize here is that draws from the posterior distribution in a collapsed Gibbs sampler must be generated in a particular order: as the last expression makes clear, in the example above it is important to first draw $\boldsymbol{\theta}_1$ from $\pi\left(\boldsymbol{\theta}_1|\mathbf{y},\boldsymbol{\theta}_3\right)$ and condition on this value when drawing from $\pi\left(\boldsymbol{\theta}_2|\mathbf{y},\boldsymbol{\theta}_1,\boldsymbol{\theta}_3\right)$, while the drawing from $\pi\left(\boldsymbol{\theta}_3|\mathbf{y},\boldsymbol{\theta}_1,\boldsymbol{\theta}_2\right)$ must not intervene between these two steps. Collapsed Gibbs samplers can be implemented in more complex cases and the interested reader is directed to J. S. Liu (1994) and van Dyk & Park (2008) for further details and some caveats.

5. In some complex models, inefficiency factors for all or some blocks may be extremely large, no matter how well the Metropolis-Hastings algorithm is tuned or tailored to the problem. In such cases, to achieve a certain degree of accuracy in the approximation of the integrals, one may need to take an immense number of draws from the posterior. However, if these draws need to be stored in memory so that they are available for processing after the algorithm completes, machine memory limitations may become an issue, especially in models with many parameters. It has become a common practice in these cases to use a *thinning parameter* to reduce the autocorrelation in the draws from the posterior. The thinning parameter is an integer greater than one and indicates how many draws from the posterior are drawn consecutively before one is stored in memory. For example, if the thinning parameter is set equal to 10, then the algorithm may still need to be run for a large number of iterations, but only one in ten draws from the posterior is stored. Thinning almost always leads to a reduction in the accuracy with which the integrals of interest are approximated because it throws away autocorrelated, yet relevant, information and should be used only in cases of limited machine memory (Link & Eaton, 2012).

6. Due to the nature of MCMC methods and their reliance on Markov chains, the draws from the posterior are generated in a sequential fashion. However, if multiple computing nodes are available on a machine, one may take advantage of the resources by running multiple chains in parallel. This approach still requires a burn-in phase, either common to all chains or chain-specific, but after this phase completes, each chain can contribute draws from the posterior, effectively reducing the amount of time it takes to produce a certain number of draws. BayES provides built-in facilities for running multiple chains in parallel for the models it supports.

7. All MCMC algorithms and their variants require a starting point, chosen by the researcher, which can be far away from the stationary distribution of the underlying Markov chain. When the algorithm is left to run for some iterations, it will, under general conditions, converge to its stationary distribution and this is precisely the role of the burn-in phase. Beyond that point every draw generated from the algorithm will be a draw from the stationary distribution and, thus, from the target distribution. However, it is rarely

possible to theoretically derive the rate of convergence of a chain to its stationary distribution. Therefore, using a short burn-in phase may result in draws that are not from the target distribution and can lead to invalid inferences. Although there have been multiple attempts to produce formal convergence diagnostics, the easiest way to check whether a chain has converged is to plot the draws for some or all parameters in the order they have been produced by the algorithm and examine the plot for any tendency of the values to move mainly in one direction. If this is the case then most likely the chain is still moving towards its stationary distribution and the number of burn-in iterations needs to be increased. When multiple chains are run in parallel, the usual practice is to plot the draws for a parameter produced by each chain on the same plot and examine visually if the the draws from different chains tend to overlap or, using MCMC jargon, examine if the chains *mix* well. Such plots can also reveal the degree of autocorrelation in the draws from the posterior, although the inefficiency factor, defined in equation (1.28), can provide a numerical measure of this autocorrelation.

1.5 Synopsis

After defining the modern meaning of the term econometrics, this chapter presented the Bayesian approach to statistical inference as an alternative to frequentist statistics. The three fundamental quantities in Bayesian inference, the likelihood function, the prior and the posterior densities, were defined and discussed. The theory behind model comparison and prediction was also presented, at a high level of abstraction. The basic simulation methods used in Bayesian inference were then described, in an algorithmic fashion. This was done, primarily, to introduce the reader to the terminology of Markov chain Monte Carlo, which will be used throughout this book, as well as to point out some common pitfalls when applying these simulation methods. It should be stressed that statistical software like BayES make the application of MCMC methods easy in the sense that the algorithms used for sampling from the posterior distribution are already coded for many popular models. However, one should still be very careful when applying these algorithms and extensive analysis of the results, visual or otherwise, should follow every application.

Chapter 2

The Linear Model

2.1 Overview

This chapter presents an extensive discussion of the multiple linear regression model with Normally distributed errors. Regardless of its simplicity, or maybe because of it, the linear regression model is one of the most widely used models in applied econometrics and, very frequently in practice, comprises the first attempt to confront economic theory with data. Indeed many of the more elaborate econometric models can be viewed as direct extensions to this model. This is more so the case in a Bayesian setting for an additional reason: the Bayesian response to increasing model complexity is, usually, to introduce latent variables in such a way that the complex model can be represented as a linear regression. Therefore, the techniques discussed here will be useful for the most part of the material that follows and the reader will be frequently referred back to this chapter.

The chapter starts with the setup of the linear regression model and its interpretation as a conditional mean model. Presentation of the likelihood function, conjugate priors and the derivation of the full conditionals for the model's parameters follows. Specification issues, model comparison and prediction are also discussed in the context of the model.

2.2 Model Setup and Interpretation

By being largely quantitative, modern economic theory posits relationships among economic variables in the general form $y = f(x_1, x_2, \ldots, x_K)$. In this expression the xs are variables that can be thought of as driving, causing or determining the value of the response variable, y. In a consumer's problem, for example, y would be the quantity of a good demanded by a consumer and the x variables would include this good's price, the prices of complementary and substitute goods, consumer income and any other consumer characteristics that may affect preferences. Most often economic theory also provides predictions on whether the effect of a variable, x_k, on y is positive or negative, or otherwise bounds the magnitude of this effect. For example an increase in the good's own price would lead, except in the case of Giffen goods, to a reduction in quantity demanded. On the other hand, theory is usually silent about the form that $f(\cdot)$ takes. The *linear regression model* is a stochastic model that quantifies causal relationships of this general form by expressing them as functions which are linear in unknown parameters:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_K x_{iK} + \varepsilon_i \tag{2.1}$$

where $\beta_1, \beta_2, \ldots, \beta_K$ are parameters to be estimated using data, $x_{i1}, x_{i2}, \ldots, x_{iK}$ are the values of the causal variables for a potential observation, i, from the population under study and y_i is

the value of the response variable for the same i. ε_i is the *disturbance* or *error term* associated with observation i and it captures any remaining variability in y that cannot be explained by the xs in this model. The error term may be non-zero for multiple reasons. Firstly, even the original functional relationship between y and the xs is a simplification of reality and cannot be expected to hold exactly for all subjects in a population. Secondly, the linear model uses an approximation to the unknown function $f(\cdot)$ and, by definition, this approximation will result in a some error. However, we would like this error to be, on average, equal to zero or, in other words, the model to be, on average, able to determine the value of y, given values for the xs.

It becomes apparent from the preceding discussion, where we already used terms related to statistics such as population and potential observation, that to proceed with the analysis of the model we need to formalize the discussion in a statistical context. Before doing so we define some terminology: y in the linear regression model is called the *dependent variable* and the xs are the *independent variables*. Other terms can be used for these two sets of variables, but we will use the ones presented here, as these are, by far, the most popular among economists. Finally, we can use the following succinct notation to represent the linear regression model:

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i \tag{2.2}$$

where $\mathbf{x}_i$ is a $K \times 1$ vector that stores the values of the independent variables for a potential observation i and $\boldsymbol{\beta}$ is a $K \times 1$ vector that contains the β parameters.

The linear regression model, as written in either equation (2.1) or (2.2), is a model for the determination of y in the population. In this context, y_i and ε_i are random variables and $\mathbf{x}_i$ is a random vector. We will assume in this chapter that the error term is independent of $\mathbf{x}_i$ and that it follows a Normal distribution with mean zero and variance σ^2. The variance of the error term is another parameter to be estimated along with the βs and, to simplify the algebra necessary to proceed with estimation, we will express the distribution of ε_i in terms of the *precision parameter*: $\varepsilon_i \sim \mathrm{N}\left(0, \frac{1}{\tau}\right)$, where $\tau \equiv \frac{1}{\sigma^2}$. From the properties of the Normal distribution and given the independence assumption, we get that $y_i|\mathbf{x}_i \sim \mathrm{N}\left(\mathbf{x}_i'\boldsymbol{\beta}, \frac{1}{\tau}\right)$. This, in turn, implies that:

$$\mathrm{E}(y_i|\mathbf{x}_i) = \mathbf{x}_i'\boldsymbol{\beta} \tag{2.3}$$

In words, the expected value of y, conditional on the independent variables, is a linear function of the independent variables and the parameters. If we knew the values of the parameters and somebody gave us values to plug in for $\mathbf{x}_i$, then we would be able to state what the expected value of y_i is.

The discussion in the preceding paragraph may be dense, but using a concrete example will help illuminate any points that remain unclear. Suppose that our objective is to quantify the role of prices and household characteristics in determining the monthly expenditure on food products by each household, i, within a certain geographical region. Our population is, therefore, all households located within this region and, at this stage, i is used to index households in this population. In the context of this example, y_i is the amount of money that household i spends on food products and $\mathbf{x}_i$ contains the prices that this same household faces, as well as other household characteristics. Notice that we have not yet said anything about the availability of data. Nevertheless, we can write down a model that describes how y is determined in the population, even before we obtain the data. While still being in the design phase of the research, suppose that we plan to administer a questionnaire to a sample of N randomly selected households in the region of interest, so that we collect data on the dependent and independent variables. Before we actually record the responses of household i, we cannot know the values of these variables. Of course, this is to be expected: y_i and $\mathbf{x}_i$ are random variables at this stage. However, our model provides a rule that we expect these variables to follow: y_i follows a Normal distribution, conditional on $\mathbf{x}_i$, with expected value $\mathbf{x}_i'\boldsymbol{\beta}$ and variance $\frac{1}{\tau}$. That is, the model can be thought of as an assumption on the process that will generate our data, the *data-generating process*. The data that we may collect on the dependent and independent variables are only useful for estimating the parameters, not for defining the model.

The interpretation of the linear regression model as a specification of a conditional expectation gives a direct meaning to the values of the β parameters. For example, the parameter associated with the k-th independent variable measures the effect of a change in the value of x_k on the expected value of y, given that no other values change in the model:

$$\beta_k = \frac{\partial \mathrm{E}\left(y_i|\mathbf{x}_i\right)}{\partial x_{ik}} \tag{2.4}$$

Using econometric jargon, β_k is the *marginal effect* of x_k on y.

Before we close this section, we note that the linear regression model can be expressed even more compactly using notation which will probably be familiar to readers with prior exposure to frequentist econometrics, but which also has the potential to create a lot of confusion. With N potential observations on the dependent and independent variables and by stacking these potential observations one under the other, the linear model can be written as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{2.5}$$

where:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{bmatrix} \quad \text{and} \quad \mathbf{X} = \begin{bmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_N' \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1K} \\ x_{21} & x_{22} & \cdots & x_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{NK} \end{bmatrix}$$

In this notation, $\boldsymbol{\varepsilon}$ follows a multivariate Normal distribution with expected value equal to an $N \times 1$ vector of zeros. Assuming that the error terms across observations are, conditional on $\mathbf{X}$, independent from one another, the covariance of this Normal distribution is $\frac{1}{\tau}\mathbf{I}_N$. From the properties of the Normal distribution, this representation implies that $\mathbf{y}|\mathbf{X} \sim \mathrm{N}\left(\mathbf{X}\boldsymbol{\beta}, \frac{1}{\tau}\mathbf{I}_N\right)$ and $\mathrm{E}\left(\mathbf{y}|\mathbf{X}\right) = \mathbf{X}\boldsymbol{\beta}$. This representation still describes the data-generating process in the population and the term "potential observation" was used multiple times to stress that $\mathbf{y}$ and $\mathbf{X}$ are random. One can think of this as the analysis being still at the design phase and the plan is to get N observations from the population. Before the data are actually observed, all quantities in (2.5) are random.

2.3 Likelihood, Priors and Posterior

By definition, the *likelihood function* is the probability density function of a potential dataset, given the values of the parameters and evaluated at the observed data points. Intuitively, the likelihood function gives the likelihood of observing the data we do actually observe, if the model is correctly specified and if we knew the values of the parameters. From the assumptions made by the linear regression model we know that each observed y_i is a draw from a Normal distribution. Given the conditional independence assumption made on the error terms, this density can be expressed as:

$$\mathrm{p}(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \tau) = \prod_{i=1}^{N} \mathrm{p}(y_i|\mathbf{x}_i, \boldsymbol{\beta}, \tau) = \prod_{i=1}^{N} \frac{\tau^{1/2}}{(2\pi)^{1/2}} \exp\left\{-\frac{\tau}{2}\left(y_i - \mathbf{x}_i'\boldsymbol{\beta}\right)^2\right\} \tag{2.6}$$

Collecting terms and using the matrix representation of the model in equation (2.5) leads to:

$$\mathrm{p}(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \tau) = \frac{\tau^{N/2}}{(2\pi)^{N/2}} \exp\left\{-\frac{\tau}{2}\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)'\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)\right\} \tag{2.7}$$

Some clarification is in place here because the notation used, although standard in both frequentist and Bayesian treatments of the model, may be misleading. $\mathbf{y}$, $\mathbf{X}$ and $\boldsymbol{\varepsilon}$ in equation (2.5) are random variables and, as such, have a probability density function. By definition, (2.7) expresses the probability density function of $\mathbf{y}|\mathbf{X}$. And here comes the tricky part: in

the context of estimation, where we need to evaluate this probability density function at the observed data points, $\mathbf{y}$ and $\mathbf{X}$ are also used to represent the vector and matrix, respectively, that store the observed data. To put it differently, while in basic statistics one typically uses upper-case letters to represent random variables and the corresponding lower-case letters to represent possible values or realizations of these random variables, in the linear regression model the same symbols are used to represent both. This is a fine point that may catch even seasoned statisticians off-guard when asked to state the assumptions behind a stochastic model without reference to the data.

We can now move to the specification of the priors for the model's parameters. The linear regression model provides a natural blocking of its parameters: $\boldsymbol{\beta}$ and τ. This is because the slope coefficients in $\boldsymbol{\beta}$ define the conditional mean of $\mathbf{y}$ and enter the likelihood function as a group, while τ defines the variance of $\mathbf{y}$ and appears in the likelihood function in entirely different places. Additionally, we do not need to impose any restrictions on the values of $\boldsymbol{\beta}$, except if economic theory requires so, while we need to restrict τ to be positive.

The functional form of the priors makes a big difference for the estimation of the model's parameters. If the priors are *conjugate* or, otherwise, lead to full conditionals that belong to parametric families from which it is easy to draw random numbers directly, then one can use a Gibbs sampler instead of full Metropolis-Hastings updates. For the linear regression model Zellner (1971, chapter 3) shows that the Normal-Gamma prior is conjugate. The Normal-Gamma prior is rather peculiar: the prior for τ is Gamma with hyperparameters a and b and the prior for $\boldsymbol{\beta}$ conditional on τ is multivariate Normal with mean $\mathbf{m}$ and variance matrix $\frac{1}{\tau}\mathbf{V}$, where $\mathbf{m}$ and $\mathbf{V}$ are $\boldsymbol{\beta}$'s hyperparameters. Because this prior is conjugate, the joint posterior density of $\boldsymbol{\beta}$ and τ is also Normal-Gamma and there is no need to use simulation to approximate its moments, since they can be obtained from the marginal posterior densities of the two blocks (Koop, 2003). It is stressed, however, that this prior and the results associated with it were proposed in a period before application of MCMC methods became widespread, and when forcing the posterior density to be a member of a known parametric family was, almost, a necessity.

In the Normal-Gamma prior, the prior variance of $\boldsymbol{\beta}$ depends on τ's hyperparameters and this may pose some problems when eliciting the values of a, b, $\mathbf{m}$ and $\mathbf{V}$ such that they conform to prior beliefs. Throughout this textbook we will, instead, use independent priors for the two blocks: we will keep assuming that τ follows a Gamma distribution, but we will assume that $\boldsymbol{\beta}$ follows a multivariate Normal distribution, marginally with respect to τ:

$$\mathrm{p}(\boldsymbol{\beta}) = \frac{|\mathbf{P}|^{1/2}}{(2\pi)^{K/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}-\mathbf{m})' \mathbf{P} (\boldsymbol{\beta}-\mathbf{m})\right\} \qquad \text{and} \qquad \mathrm{p}(\tau) = \frac{b^a}{\Gamma(a)}\tau^{a-1}e^{-b\tau} \tag{2.8}$$

It is stressed that, although we assume that $\boldsymbol{\beta}$ and τ are independent in the prior, this will not be the case in the posterior. Finally, notice that, as it was the case until now with scale parameters, we express the density of $\boldsymbol{\beta}$ in terms of the inverse of the variance matrix: $\mathbf{P}$ is the prior *precision matrix*. This re-parameterization simplifies algebraic manipulations considerably.

We are now ready to derive the posterior density of the parameters. By a standard application of Bayes' theorem we get:

$$\begin{aligned} \pi(\boldsymbol{\beta}, \tau | \mathbf{y}, \mathbf{X}) &\propto \mathrm{p}(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \tau) \times \mathrm{p}(\boldsymbol{\beta}) \times \mathrm{p}(\tau) \\ &= \frac{\tau^{N/2}}{(2\pi)^{N/2}} \exp\left\{-\frac{\tau}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right\} \\ &\quad \times \frac{|\mathbf{P}|^{1/2}}{(2\pi)^{K/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}-\mathbf{m})' \mathbf{P} (\boldsymbol{\beta}-\mathbf{m})\right\} \\ &\quad \times \frac{b^a}{\Gamma(a)}\tau^{a-1}e^{-b\tau} \end{aligned} \tag{2.9}$$

The posterior density can be simplified in quite a few ways, but we should keep in mind what is our objective: we want to derive the full conditionals of $\boldsymbol{\beta}$ and τ so that we are able

to implement a Gibbs sampler and estimate their posterior moments. This is done in the following section.

2.4 Full Conditionals and Parameter Estimation

The task of deriving the full conditionals of $\boldsymbol{\beta}$ and τ from the posterior density in (2.9) may appear daunting. The algebra required is indeed tedious, but also quite simple. We will derive these full conditionals here step-by-step because the derivation can be used as a detailed example on how the algebra works. However, we will refrain from presenting similar derivations elsewhere, since the entire purpose of using a software like BayES is to avoid having to go through this exercise.

Let's start by deriving the full conditional for τ. By dropping terms from the posterior that enter multiplicatively and do not involve τ we get:

$$\pi\left(\tau|\bullet\right) \propto \tau^{N/2} \exp\left\{-\frac{\tau}{2}\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)'\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)\right\} \times \tau^{a-1}e^{-b\tau} \tag{2.10}$$

Next, collecting terms leads to:

$$\pi\left(\tau|\bullet\right) \propto \tau^{N/2+a-1} \exp\left\{-\tau\left[\frac{1}{2}\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)'\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)+b\right]\right\} \tag{2.11}$$

which looks like the probability density function of a Gamma-distributed random variable with shape and rate parameters, $\tilde{a}=\frac{N}{2}+a$ and $\tilde{b}=\frac{1}{2}\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)'\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)+b$, respectively. The only thing missing is a constant of proportionality, which must be equal to $\frac{\tilde{b}^{\tilde{a}}}{\Gamma(\tilde{a})}$ so that $\pi\left(\tau|\bullet\right)$ integrates to unity. Therefore:

$$\pi\left(\tau|\bullet\right) = \frac{\tilde{b}^{\tilde{a}}}{\Gamma\left(\tilde{a}\right)}\tau^{\tilde{a}-1}e^{-\tilde{b}\tau} \tag{2.12}$$

Deriving the full conditional of $\boldsymbol{\beta}$ is slightly more challenging. Let's start by dropping terms from the posterior that enter multiplicatively and which do not involve $\boldsymbol{\beta}$:

$$\pi\left(\boldsymbol{\beta}|\bullet\right) \propto \exp\left\{-\frac{\tau}{2}\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)'\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)\right\} \times \exp\left\{-\frac{1}{2}\left(\boldsymbol{\beta}-\mathbf{m}\right)'\mathbf{P}\left(\boldsymbol{\beta}-\mathbf{m}\right)\right\} \tag{2.13}$$

Carrying out the multiplications, dropping, for a second time, terms that do not involve $\boldsymbol{\beta}$ and collecting terms leads to:

$$\pi\left(\boldsymbol{\beta}|\bullet\right) \propto \exp\left\{-\frac{1}{2}\left[\boldsymbol{\beta}'\left(\tau\mathbf{X}'\mathbf{X}+\mathbf{P}\right)\boldsymbol{\beta}-\boldsymbol{\beta}'\left(\tau\mathbf{X}'\mathbf{y}+\mathbf{P}\mathbf{m}\right)-\left(\tau\mathbf{y}'\mathbf{X}+\mathbf{m}'\mathbf{P}\right)\boldsymbol{\beta}\right]\right\} \tag{2.14}$$

The next step, which is not intuitive, is to simplify the expression by defining $\tilde{\mathbf{P}}=\tau\mathbf{X}'\mathbf{X}+\mathbf{P}$ and $\tilde{\mathbf{m}}=\tilde{\mathbf{P}}^{-1}\left(\tau\mathbf{X}'\mathbf{y}+\mathbf{P}\mathbf{m}\right)$. With these definitions the expression inside the square brackets becomes $\boldsymbol{\beta}'\tilde{\mathbf{P}}\boldsymbol{\beta}-\boldsymbol{\beta}'\tilde{\mathbf{P}}\tilde{\mathbf{m}}-\tilde{\mathbf{m}}'\tilde{\mathbf{P}}\boldsymbol{\beta}$. The final step requires "completing the square" in this expression. By adding $\tilde{\mathbf{m}}'\tilde{\mathbf{P}}\tilde{\mathbf{m}}$ inside the square brackets[1] and collecting terms we obtain:

$$\pi\left(\boldsymbol{\beta}|\bullet\right) \propto \exp\left\{-\frac{1}{2}\left(\boldsymbol{\beta}-\tilde{\mathbf{m}}\right)'\tilde{\mathbf{P}}\left(\boldsymbol{\beta}-\tilde{\mathbf{m}}\right)\right\} \tag{2.15}$$

From this expression it is easy to see that $\pi\left(\boldsymbol{\beta}|\bullet\right)$ is proportional to a multivariate Normal density. Again we are missing a constant of proportionality, which has to be equal to $\frac{|\tilde{\mathbf{P}}|^{1/2}}{(2\pi)^{K/2}}$ for the density to integrate to unity. Therefore:

$$\pi\left(\boldsymbol{\beta}|\bullet\right) = \frac{|\tilde{\mathbf{P}}|^{1/2}}{(2\pi)^{K/2}} \exp\left\{-\frac{1}{2}\left(\boldsymbol{\beta}-\tilde{\mathbf{m}}\right)'\tilde{\mathbf{P}}\left(\boldsymbol{\beta}-\tilde{\mathbf{m}}\right)\right\} \tag{2.16}$$

These results are presented here in the form of a theorem, so that we can refer back to them whenever need arises.

[1] Adding this quantity inside the square brackets is equivalent to multiplying the entire expression of the full conditional by $\exp\left\{-\frac{1}{2}\tilde{\mathbf{m}}'\tilde{\mathbf{P}}\tilde{\mathbf{m}}\right\}$ and, thus, affects only the constant of proportionality.

THEOREM 2.1: Full Conditionals for the Linear Model
In the linear regression model with Normally distributed error and K independent variables:

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i, \qquad \varepsilon_i \sim \mathrm{N}\left(0, \tfrac{1}{\tau}\right)$$

and with a Normal prior for $\boldsymbol{\beta}$ and a Gamma prior for τ:

$$\mathrm{p}(\boldsymbol{\beta}) = \frac{|\mathbf{P}|^{1/2}}{(2\pi)^{K/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}-\mathbf{m})'\mathbf{P}(\boldsymbol{\beta}-\mathbf{m})\right\} \quad \text{and} \quad \mathrm{p}(\tau) = \frac{b^a}{\Gamma(a)}\tau^{a-1}e^{-b\tau}$$

the full conditionals of $\boldsymbol{\beta}$ and τ are, respectively, Normal and Gamma:

$$\pi(\boldsymbol{\beta}|\bullet) = \frac{|\tilde{\mathbf{P}}|^{1/2}}{(2\pi)^{K/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}-\tilde{\mathbf{m}})'\tilde{\mathbf{P}}(\boldsymbol{\beta}-\tilde{\mathbf{m}})\right\} \quad \text{and} \quad \pi(\tau|\bullet) = \frac{\tilde{b}^{\tilde{a}}}{\Gamma(\tilde{a})}\tau^{\tilde{a}-1}e^{-\tilde{b}\tau}$$

where:

- $\tilde{\mathbf{P}} = \tau\mathbf{X}'\mathbf{X} + \mathbf{P}$
- $\tilde{\mathbf{m}} = (\tau\mathbf{X}'\mathbf{X} + \mathbf{P})^{-1}(\tau\mathbf{X}'\mathbf{y} + \mathbf{P}\mathbf{m})$
- $\tilde{a} = \frac{N}{2} + a$
- $\tilde{b} = \frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) + b$

With the full conditionals at hand we are ready to implement the Gibbs sampler. This sampler would involve sampling from the full conditional of each of the two blocks, $\boldsymbol{\beta}$ and τ, within each iteration. We will not present an implementation here, but the interested reader is directed to the `lm.bsf` script file, which can be found in the directory `"Samples/4-Functions"` (created during BayES' installation), which contains such an implementation. Instead, the following example contains an application of the linear regression model, using BayES' built-in sampler.

♦ **Example 2.1 Expenditure on Food Products by Households in the Netherlands**
In this example we will consider part of the dataset used by Adang & Melenberg (1995). The dataset contains information on 90 households located in the Netherlands, each one of them observed for 42 consecutive months. The variables in the dataset are:

`expFood`	:	monthly expenditure on food products, in 100's of Guilders
`expOther`	:	monthly expenditure on other products, in 100's of Guilders
`pFood`	:	a price index for food products, April 1984=100
`pOther`	:	a price index for other products, April 1984=100
`Hsize`	:	number of household members
`Children`	:	number of household members younger than 11 years

Our objective is to estimate the parameters of a model that determines monthly household expenditure as a function of prices and other household characteristics. The model for the population (all households located in the Netherlands) is:

$$\mathsf{expFood}_i = \beta_1 + \beta_2\mathsf{pFood}_i + \beta_3\mathsf{pOther}_i + \beta_4\mathsf{expOther}_i + \beta_5\mathsf{Hsize}_i + \beta_6\mathsf{Children}_i + \varepsilon_i$$

Using BayES' `lm()` function and given the data at hand, we obtain the results in the following table.

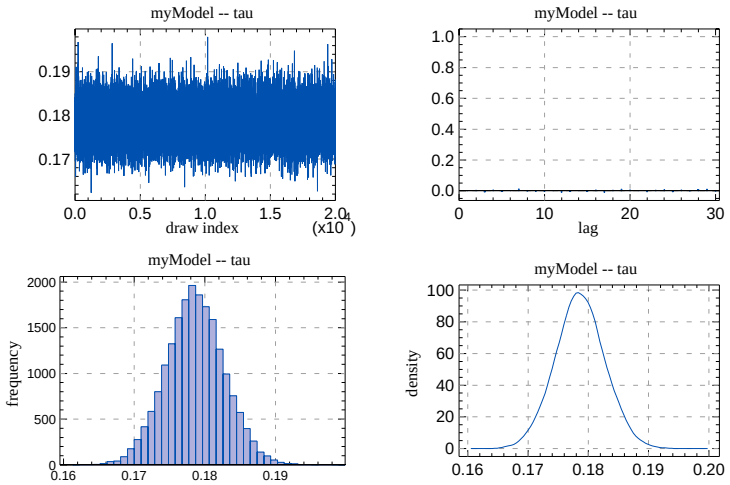
	Mean	Median	Sd.dev.	5%	95%
constant	-4.72133	-4.73423	5.43244	-13.6764	4.21482
pFood	0.0725127	0.0730401	0.0445887	-0.00084805	0.145471
pOther	0.0140819	0.0140129	0.0311635	-0.0374249	0.0655787
expOther	0.0188497	0.0188519	0.00155341	0.016298	0.0214029
Hsize	0.429952	0.429473	0.0499011	0.348149	0.513194
Children	-0.207939	-0.207817	0.0731264	-0.328807	-0.0886875
tau	0.178622	0.178583	0.00411218	0.171855	0.185456
sigma_e	2.36657	2.36636	0.0272516	2.3221	2.41223

The first column of the table contains the names of the variables with which each parameter is associated. For example, β_1 is associated with a constant variable (β_1 is always multiplied by

one in the model), β_2 is associated with pFood, etc. The second column contains the posterior mean of each parameter, while the third and fourth columns the corresponding posterior medians and standard deviations. The last two columns give the endpoints of the 90% credible intervals. These are constructed by dropping 5% of the draws from each tail of the marginal distribution of each parameter. The bottom block of the table presents similar results for the precision parameter of the error and for its standard deviation, using the relation $\sigma = \frac{1}{\sqrt{\tau}}$.

Let's interpret the posterior means one by one. If all independent variables were equal to zero except for the constant term, then we would expect household expenditure to be about -472 Guilders. Of course this number does not make sense, but restricting prices or household size to zero does not make much sense either. All other things equal, if the price index for food products increases by one unit, expenditure for food products is expected to increase by about 7 Guilders, while if the price index of other products increases by one unit then expected expenditure increases by about 1.4 Guilders. Notice, however, that the 90% credible intervals for each of the last two variables contain zero, suggesting that a strictly positive effect is not very likely. When expenditure on other products increases by 100 Guilders we should expect expenditure on food products to increase by about 1.9 Guilders. Finally, an additional member to the household increases expected expenditure on food products by about 430 Guilders, while if this additional member is younger than 11 years of age, the increase in expected expenditure is lower: 430-208 = 222 Guilders.

To assess the performance of the MCMC sampler, one can use BayES' `diagnostics()` function. This function will produce a table that contains estimates of the MCMC standard error, as well as of the inefficiency factor, per parameter. Visually assessing the performance of the MCMC sampler can be achieved by drawing diagnostics plots for each parameter, using BayES' `plotdraws()` function. The figure below contains such a plot for the τ parameter of the model. The two subplots at the top present a history and a correlogram of the draws for τ and indicate that these are not autocorrelated. The two remaining subplots present a histogram and an estimate of the kernel density of the draws. Both of them are smooth, suggesting that the sampler did not get trapped for any considerable amount of draws in specific regions of the sample space.



Obtaining the table and plot presented above using BayES can be achieved using the code in the following box. Note that we have used the default values for the hyperparameters, which may not be appropriate for all applications. The reader is directed to BayES' documentation for details on what these default values are and how to alter them.

```
// import the data into a dataset called Data
Data = webimport("www.bayeconsoft.com/datasets/FoodExp.csv");

// construct the constant term (a variable always equal to one)
Data.constant = 1;

// run the model
myModel = lm(expFood ~ constant pFood pOther expOther Hsize Children);

// plot the draws for tau
plotdraws(tau, "model"=myModel);
```

2.5 Other Functional Forms and Marginal Effects

The linear regression model expresses a relationship of the general form $y = f(x_1, x_2, \ldots, x_K)$ as a linear function of the parameters and assumes that noise enters the resulting expression additively. This two qualifications do no require the relationship between y and the xs to be linear. To take this issue further, consider a simplified version, where $K = 3$ and that x_1 is always equal to one. The following model is, of course, linear in the parameters:

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i \tag{2.17}$$

Furthermore, the marginal effects of x_2 and x_3 are β_2 and β_3, respectively. However, the model:

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i2}^2 + \beta_5 x_{i2} x_{i3} + \varepsilon_i \tag{2.18}$$

is linear in the parameters as well, although it is non-linear in the xs: x_2 enters this expression once linearly and associated with β_2, once as a *quadratic term* and associated with β_4 and once as an *interaction term* with x_3 and associated with β_5. That is, the xs may affect y non-linearly, while the results described above can still be applied to this model. To proceed with estimation we simply need to treat the constant term, x_{i2}, x_{i3}, as well as x_{i2}^2 and $x_{i2}x_{i3}$ as independent variables in the linear regression model and define $\mathbf{x}_i$, accordingly, as a 5×1 vector.

Due to the non-linear fashion in which the xs enter the model in (2.18), their marginal effects are more complex. The marginal effect of x_{i2} and x_{i3} on y_i are:

$$\frac{\partial \mathrm{E}(y_i|\mathbf{x}_i)}{\partial x_{i2}} = \beta_2 + 2\beta_4 x_{i2} + \beta_5 x_{i3} \qquad \text{and} \qquad \frac{\partial \mathrm{E}(y_i|\mathbf{x}_i)}{\partial x_{i3}} = \beta_3 + \beta_5 x_{i2} \tag{2.19}$$

respectively. The marginal effects now depend on the values of the xs themselves, but this is to be expected: if x_2 affects y in a non-linear fashion, then its effect on y should vary by the value of x_2.

Because the marginal effects vary by the values of the xs, a question that arises is at what point should these effects be calculated and reported. Sometimes, interest may center on the marginal effect of a variable at specific values of the xs and, depending on the research question, economic theory may provide a natural point at which the effects should be evaluated. Quite often, however, no such natural point exists and an approach used frequently in practice is to evaluate the marginal effects at the sample means of the observed data. This would amount, for example, to reporting the marginal effect of x_3 on y as the number that results from plugging into the relevant expression from (2.19) the values for β_3, β_5 and the sample mean of x_{i2}, $\bar{x}_2$. Of course, the uncertainty around the values of β_3 and β_5 is transmitted to the marginal effect and one could get the posterior moments of this marginal effect by evaluating it at all draws of β_3 and β_5 from the posterior and summarizing.

Let's return to the general problem of representing the possibly non-linear relationship among K xs and y. If no argument can be made in favor of a specific functional form of this relationship, one could use a *flexible functional form*, obtained by approximating $f(\cdot)$ by a second-order *Taylor-series expansion* around a $K \times 1$ vector of zeros:

$$f(x_1, x_2, \ldots, x_K) \approx f(0, 0, \ldots, 0) + \sum_{k=1}^{K} \frac{\partial f}{\partial x_k} \cdot x_k + \frac{1}{2} \sum_{k=1}^{K} \sum_{\ell=1}^{K} \frac{\partial^2 f}{\partial x_k \partial x_\ell} \cdot x_k \cdot x_\ell \tag{2.20}$$

The first term in this expression is the value of $f(\cdot)$ evaluated at the vector of zeros and becomes the constant term in a linear regression model. Similarly, the first- and second-order derivatives, all evaluated at the vector of zeros, become parameters associated with the original xs and their interactions, respectively. Notice that from *Young's theorem* we get:

$$\frac{\partial^2 f}{\partial x_k \partial x_\ell} = \frac{\partial^2 f}{\partial x_\ell \partial x_k} \tag{2.21}$$

and the parameter associated with $x_k \cdot x_\ell$ should be equal to the one associated with $x_\ell \cdot x_k$. For example, when $y = f(x_2, x_3)$[2], the Taylor-series expansion leads to the linear regression model:

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i2}^2 + \beta_5 x_{i2} x_{i3} + \beta_6 x_{i3}^2 + \varepsilon_i \tag{2.22}$$

where β_2 and β_3 are the partial derivatives of $f(\cdot)$ with respect to x_2 and x_3, β_5 is the second cross-derivative and β_4 and β_6 are equal to $\frac{1}{2}$ times the second-order derivatives with respect to x_2 and x_3, respectively, and all these derivatives are evaluated at the point $(x_2, x_3) = (0, 0)$. It becomes apparent that, although flexible, this approach leads to a proliferation of parameters to be estimated and it is, therefore, mostly useful in situations with large numbers of observations.

Sometimes economic theory may suggest functional forms for the relationship between the causal variables and the response variable. For example, a popular choice for representing the aggregate production function in growth economics is the *Cobb-Douglas* form: $Y = A(t) K^\alpha L^\beta$, where Y is the amount of produced output and K and L are the amounts of capital and labor input employed during production. $A(t)$ is a productivity index and α and β are parameters to be estimated, which, however, enter the model non-linearly. Nevertheless, taking the natural logarithm of both sides of the last expression leads to $\log Y = \log A(t) + \alpha \log K + \beta \log L$. After appending an error term and assuming that $\log A(t)$ can be adequately represented as $A_0 + \gamma t$, the parameters of the Cobb-Douglas function can be estimated using a linear regression model of the form:

$$\log Y_i = \beta_1 + \beta_2 \log K_i + \beta_3 \log L_i + \beta_4 t + \varepsilon_i \tag{2.23}$$

where $\log Y_i$, $\log K_i$ and $\log L_i$ are the dependent and independent variables, t is a time trend and $\beta_1 \equiv A_0$, $\beta_2 \equiv \alpha$, $\beta_3 \equiv \beta$ and $\beta_4 \equiv \gamma$ are the parameters to be estimated. That is, by applying monotonic transformations to both sides of a relationship that is implied by economic theory, we may be able to turn a model which is not linear in the parameters into one that is.

Although the model in (2.23) is linear in the parameters, the relationship between the inputs and output is not. Using the conditional expectation interpretation of the linear regression model, equation (2.23) implies that $\mathrm{E}(\log Y_i | K_i, L_i, t) = \beta_1 + \beta_2 \log K_i + \beta_3 \log L_i + \beta_4 t$ and the marginal effect of $\log K_i$ on $\log Y_i$ is:

$$\beta_2 = \frac{\partial \mathrm{E}(\log Y_i | K_i, L_i, t)}{\partial \log K_i} = \frac{\partial \log \mathrm{E}(Y_i | K_i, L_i, t)}{\partial \log K_i} = \frac{\partial \mathrm{E}(Y_i | K_i, L_i, t)}{\partial K_i} \cdot \frac{K_i}{\mathrm{E}(Y_i | K_i, L_i, t)} \tag{2.24}$$

which has the form of an *elasticity*. Therefore, β_2 in (2.23) gives the percentage change in output caused by a 1% increase in the amount of capital. This argument carries over to the interpretation of β_3 as an elasticity and extends to all models in double-log form, $\log y_i = (\log \mathbf{x}_i)' \boldsymbol{\beta} + \varepsilon_i$, or in forms where only some of the independent variables are in logarithms, $\log y_i = (\log \mathbf{x}_i)' \boldsymbol{\beta} + \mathbf{z}_i' \boldsymbol{\gamma} + \varepsilon_i$.

It is stressed that, in general:

$$\frac{\partial \log \mathrm{E}(y_i | \mathbf{x}_i)}{\partial \log x_{ik}} \neq \frac{\partial \mathrm{E}(\log y_i | \mathbf{x}_i)}{\partial \log x_{ik}} \tag{2.25}$$

and either definition of elasticity can be used in stochastic models, with the choice usually being based on convenience. Wooldridge (2002, p.17) shows that the two definitions are equivalent if the error term is independent of $\mathbf{x}_i$; an assumption that we are maintaining throughout this chapter. In particular, if $\varepsilon_i | \mathbf{x}_i \sim \mathrm{N}\left(0, \frac{1}{\tau}\right)$ in the double-log model $\log y_i = (\log \mathbf{x}_i)' \boldsymbol{\beta} + \varepsilon_i$, then $\log y_i | \mathbf{x}_i$ follows a Normal distribution as well. By definition, $y_i | \mathbf{x}_i$ follows a log-Normal distribution and from the properties of this distribution we get:

$$\log \mathrm{E}(y_i | \mathbf{x}_i) = (\log \mathbf{x}_i)' \boldsymbol{\beta} + \tfrac{1}{2\tau} \tag{2.26}$$

From this expression we can see that $\frac{\partial \log \mathrm{E}(y_i | \mathbf{x}_i)}{\partial \log x_{ik}} = \beta_k = \frac{\partial \mathrm{E}(\log y_i | \mathbf{x}_i)}{\partial \log x_{ik}}$.

[2]Notice that we dropped x_1 from the set of causal variables to keep notation consistent with previous expressions.

There are a few practical reasons why one may choose to model the logarithm of y_i instead of its level. By taking the logarithm of the dependent variable the model restricts the value of y_i to be strictly positive and this restriction is very frequently a reasonable one, given the nature of economic data. Additionally, the distribution of $\log y_i$ may be much more concentrated and symmetric around its mean than that of y_i. Although the linear regression model requires the dependent variable to follow a Normal distribution only conditionally on the independent variables, it is still easier to model a variable that follows a more-or-less symmetric distribution, even marginally with respect to $\mathbf{x}_i$. This is because, when the density of the dependent variable has a long tail to the right, observations with extremely large values on this variable have a disproportionate impact on the parameter estimates. The role of the logarithmic transformation in such cases is to remove the long tail from the density of the dependent variable.

Even when the the relationship between y and the xs is modeled in logarithms, there may still be scope for using as independent variables squared and interactions terms of the logarithms of the xs. The *translog* is another flexible functional form, obtained by taking a second-order Taylor-series expansion of $\log y = \log f(\log x_1, \log x_2, \ldots, \log x_K)$ around a vector of ones (for the xs):

$$\begin{aligned}\log f(\log x_1, \log x_2, \ldots, \log x_K) \approx \log f(0, 0, \ldots, 0) + \sum_{k=1}^{K} \frac{\partial \log f}{\partial \log x_k} \cdot \log x_k \\ + \frac{1}{2} \sum_{k=1}^{K} \sum_{\ell=1}^{K} \frac{\partial^2 \log f}{\partial \log x_k \partial \log x_\ell} \cdot \log x_k \cdot \log x_\ell\end{aligned} \tag{2.27}$$

In the translog specification the marginal effects are still interpreted as elasticities, which now vary by the levels of the independent variables.

♦ **Example 2.2 Aggregate Production Function**

In this example we will use aggregate data, taken from the Penn World Table, Version 9.0 (Feenstra et al., 2015). The dataset contains annual information on a series of aggregate variables for the EU-15 Member States from 1970 to 2014. The ones we will use here are:

`Y`	:	real GDP at constant national prices (in mil. \$2011)
`K`	:	capital stock at constant national prices (in mil. \$2011)
`L`	:	number of persons engaged (in millions), adjusted for human capital
`trend`	:	a trend variable running from -22 to $+22$

We will start by estimating an aggregate production function, assuming that it can be adequately described by the Cobb-Douglas form:

$$\log \mathsf{Y}_i = \beta_1 + \beta_2 \log \mathsf{K}_i + \beta_3 \log \mathsf{L}_i + \beta_4 \mathsf{trend}_i + \varepsilon_i$$

The results obtained using BayES' `lm()` function are presented in the following table. For this model we use the default priors for the βs and the precision parameters of the error term: the 4×1 vector $\boldsymbol{\beta}$ is assumed to follow a Normal distribution with mean equal to a vector of zeros and a diagonal precision matrix, with its diagonal elements set equal to 10^{-4}, while τ is assumed to follow a Gamma distribution in the prior, with both shape and rate parameters set equal to 10^{-4}. The priors for both blocks are very vague in the context of the application.

	Mean	Median	Sd.dev.	5%	95%
constant	3.24514	3.24896	0.279077	2.784	3.69835
logK	0.583076	0.582767	0.0238982	0.544447	0.622634
logL	0.441425	0.441675	0.0225753	0.403951	0.477781
trend	0.00119668	0.00120461	0.000508724	0.000353659	0.0020245
tau	72.1455	72.0801	3.9129	65.778	78.7437
sigma_e	0.117862	0.117787	0.00320426	0.112692	0.123302

From this table we see that the posterior expected value of the output elasticity with respect to capital is about 0.583 and that it is within the interval $[0.544, 0.623]$ with probability 90%. The

corresponding value and interval for the elasticity with respect to labor are 0.441 and $[0.404, 0.478]$. Finally, output tends to increase by 0.12% per year due to autonomous technological progress.

We will now extend the Cobb-Douglas model and estimate a translog production function:

$$\begin{aligned}\log \mathsf{Y}_i = \beta_1 &+ \beta_2 \log \mathsf{K}_i + \beta_3 \log \mathsf{L}_i + \beta_4 \mathsf{trend}_i \\ &+ \beta_5 \log \mathsf{K}_i \log \mathsf{K}_i + \beta_6 \log \mathsf{K}_i \log \mathsf{L}_i + \beta_7 \log \mathsf{L}_i \log \mathsf{L}_i \\ &+ \beta_8 \mathsf{trend}_i \log \mathsf{K}_i + \beta_9 \mathsf{trend}_i \log \mathsf{L}_i + \varepsilon_i\end{aligned}$$

If we proceed to create the interaction terms and estimate the model, we will get parameter estimates for the βs, which will depend on the units of measurement of the independent variables. Of course, marginal effects will still be in the form of elasticities, but obtaining them requires additional calculations. However, if we transform the independent variables by subtracting their sample means from the observed values:

$$\widetilde{\log \mathsf{K}}_i = \log \mathsf{K}_i - \overline{\log \mathsf{K}} \qquad \text{and} \qquad \widetilde{\log \mathsf{L}}_i = \log \mathsf{L}_i - \overline{\log \mathsf{L}}$$

before creating interaction terms and estimating the model, this will make the parameters associated with the first-order terms directly interpretable as elasticities evaluated at the point defined by the geometric means of K and L (arithmetic means of $\log K$ and $\log L$) across observations. This is because, the marginal effect, for example, of capital in the transformed variables will be:

$$\frac{\partial \, \mathrm{E}\left(\log \mathsf{Y}_i \middle| \widetilde{\log \mathsf{K}}_i, \widetilde{\log \mathsf{L}}_i\right)}{\partial \widetilde{\log \mathsf{K}}_i} = \beta_2 + 2\beta_5 \widetilde{\log \mathsf{K}}_i + \beta_6 \widetilde{\log \mathsf{L}}_i + \beta_8 \mathsf{trend}_i$$

and if we evaluate this expression at the arithmetic means of the right-hand side variables we will get β_2, as the means of the transformed variables are zero (notice that the sample mean of the trend variable is already zero).

After performing these transformations and running the model using BayES' `lm()` function, we obtain the results presented in the following table. Again, we use BayES' default priors for the `lm()` function. The results for the parameters associated with the first-order terms (β_2, β_3 and β_4) are very similar to the ones obtained from the Cobb-Douglas model, although the credible intervals for some of the parameters on the second-order terms do not include zero. For example the parameter associated with variable tlogK is negative with probability greater than 95% and the parameter associated with variable tlogL is positive with probability greater than 95%, both of these findings suggesting that technological progress during the period covered by the data was not neutral.

	Mean	**Median**	**Sd.dev.**	**5%**	**95%**
constant	12.7573	12.7573	0.0064693	12.7467	12.7678
logK	0.570639	0.570463	0.0260929	0.527734	0.613743
logL	0.44325	0.443427	0.0247698	0.402362	0.484039
trend	0.00140262	0.00140398	0.000524615	0.000547946	0.00226372
logKlogK	0.196592	0.196859	0.10778	0.0194922	0.374709
logKlogL	-0.29448	-0.294986	0.210733	-0.642433	0.0504442
logLlogL	0.0963105	0.0962573	0.102469	-0.0717326	0.265435
tlogK	-0.00700459	-0.00700388	0.00300928	-0.0119623	-0.00205764
tlogL	0.00491162	0.00492376	0.00288508	0.000153932	0.00963616
tau	81.4347	81.3652	4.47955	74.1769	88.9722
sigma_e	0.11094	0.110862	0.00305806	0.106017	0.116109

The results presented above can be obtained in BayES using the code in the following box.

```
// import the data into a dataset called Data
Data = webimport("www.bayeconsoft.com/datasets/PWT.csv");

// construct the constant term and take logs of inputs and output
Data.constant = 1;
Data.logY = log(Data.Y);
Data.logK = log(Data.K);
Data.logL = log(Data.L);

// run the Cobb-Douglas model
CobbDouglas = lm(logY ~ constant logK logL trend);
```

```
// normalize inputs and create interaction terms
Data.logK = Data.logK - mean(Data.logK);
Data.logL = Data.logL - mean(Data.logL);
Data.logKlogK = Data.logK.*Data.logK;
Data.logKlogL = Data.logK.*Data.logL;
Data.logLlogL = Data.logL.*Data.logL;
Data.tlogK = Data.trend.*Data.logK;
Data.tlogL = Data.trend.*Data.logL;

// run the translog model
Translog = lm(logY ~ constant logK logL trend
    logKlogK logKlogL logLlogL tlogK tlogL);
```

2.6 Post-Estimation Inference

Estimating the parameters of a linear regression model accomplishes the first task of econometrics. The remaining two tasks are evaluating the plausibility of statements that involve the parameters or otherwise comparing alternative models/theories and, in the context of the linear regression model, producing predictions for the values of the dependent variable. We will use the term *post-estimation inference* to describe these tasks, although we will treat them separately in the following subsections.

2.6.1 Imposing Parametric Restrictions and Evaluating their Plausibility

Although economic theory is rarely informative about the functional form of the relationship between the causal variables and the response variable, it frequently provides restrictions on the signs of the marginal effects or the interdependence of these effects. Because marginal effects are always functions of the parameters in the linear regression model, the plausibility of restrictions prescribed by theory can be evaluated using evidence from the data or the restrictions can even imposed on a model.

Let's consider the example of the aggregate production function, assuming that this takes the Cobb-Douglas form. The linear regression model resulting from this assumption is:

$$\log Y_i = \beta_1 + \beta_2 \log K_i + \beta_3 \log L_i + \beta_4 t + \varepsilon_i \tag{2.28}$$

In this model β_4 measures the rate of increase in output only due to the passage of time and the role of the time trend in this model is precisely to capture technological progress.[3] Arguably, technological progress moves in one direction, with any innovations that result from research and development efforts being adopted by firms only if these innovations increase productivity. This argument suggests that β_4 should be positive and any evidence from the data against this would raise questions about the theory itself or the adequacy of the Cobb-Douglas form to represent the actual production function. Given the uncertainty associated with the value of a parameter, what would constitute evidence against the hypothesis would be showing that the probability of $\beta_4 < 0$ is non-negligible. In a Bayesian setting, where the draws from the posterior for β_4 are draws from its marginal (with respect to the other parameters) distribution given the data, this probability can be approximated simply by calculating the proportion of draws that satisfy $\beta_4 < 0$.

This approach of evaluating the plausibility of statements that come from economic theory can be extended when the statement involves functions of a single or multiple parameters. Continuing with the example of the Cobb-Douglas aggregate production function, the long-run properties of economic growth models depend crucially on whether this function exhibits decreasing, constant or increasing returns to scale. In the Cobb-Douglas function returns to scale are constant if the sum of β_2 and β_3 is equal to one, decreasing if the sum is smaller than one and increasing otherwise. The probability of the production function exhibiting, for example,

[3] Recall that β_4 came from the assumption that the logarithm of the productivity index in a Cobb-Douglas production function can be expressed as a linear function of time.

increasing returns to scale can be approximated by calculating the proportion of draws from the posterior that satisfy $\beta_2 + \beta_3 > 1$. This is because the draws from the posterior for β_2 and β_3 are draws from the joint distribution of these two parameters, conditional on the data and marginally with respect to to the remaining parameters. Furthermore, the plausibility of the joined statement "*the production function exhibits increasing returns to scale and technological progress*" can be evaluated by the proportion of draws from the posterior distribution that satisfy both $\beta_2 + \beta_3 > 1$ and $\beta_4 > 0$, at the same time.

We can now generalize the preceding discussion. Consider a generic linear regression model:

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i \tag{2.29}$$

and q statements or hypotheses about the values of functions of the model's parameters. These q statements can be expressed in the form $\mathbf{g}(\boldsymbol{\beta}) > \mathbf{r}$, where $\mathbf{g}$ is a vector-valued function and $\mathbf{r}$ a $q \times 1$ vector.[4] Then the plausibility of all q statements taken together can be evaluated by calculating the proportion of draws from the posterior that satisfy all q of them at the same time. We note in passing that this simple approach does not extend well to cases where the restrictions are expressed in the form of equalities. This is because $\boldsymbol{\beta}$ is a continuous random variable and the probability of a function of it being exactly equal to a given vector, $\mathbf{r}$, is zero. An obvious workaround is to calculate the probability of $\mathbf{g}(\boldsymbol{\beta})$ being within a certain small interval around $\mathbf{r}$, but we will describe below a procedure for evaluating the plausibility of statements that involve equalities, based on the general Bayesian model-comparison approach.

Sometimes, the researcher may want to impose parametric restrictions on a model, rather than simply evaluating their plausibility. This may be done so that the parameter estimates are consistent with theory or to introduce prior knowledge about the values of the parameters, given that the validity of the theory is not to be questioned. Imposing restrictions of a general form may become quite demanding, but the process simplifies considerably if these restrictions are linear in the parameters:

$$\mathbf{R}\boldsymbol{\beta} = \mathbf{r} \tag{2.30}$$

where $\mathbf{R}$ is a $q \times K$ matrix and $\mathbf{R}\boldsymbol{\beta}$ assumes the role of the possibly non-linear function $\mathbf{g}(\boldsymbol{\beta})$, defined above. If this system of equations has multiple solutions, then the dimensionality of $\boldsymbol{\beta}$ can be reduced by expressing one or more of the βs as functions of the remaining and of the coefficients in $\mathbf{R}$ and $\mathbf{r}$. This leads to a smaller number of parameters to be estimated and the restrictions can be imposed on the model by appropriately transforming the data. This approach, however, is case specific and the discussion cannot proceed any further without a concrete case.

The Bayesian approach provides a more direct and natural device for imposing this type of restrictions, through the prior density of the parameters. Taking this route starts from treating the restrictions as stochastic, rather than deterministic. An intuitive way to think about this is to consider the prior density of $\boldsymbol{\beta}$. If the uncertainty about the value of $\boldsymbol{\beta}$ is reflected in this prior density, then $\boldsymbol{\beta}$ is a random vector and, therefore, $\mathbf{r}$ should also be treated as random. However, economic theory provides values for $\mathbf{r}$ and to impose the restrictions through the prior we should get the prior density of $\boldsymbol{\beta}$ given $\mathbf{r}$. Thus, apart from accommodating the restrictions, $\mathrm{p}(\boldsymbol{\beta}|\mathbf{r})$ also expresses how forcefully we want to impose them or, in other words how far away are we willing to allow $\boldsymbol{\beta}$ to be from satisfying the restrictions. Defining the priors such that they reflect prior beliefs on the validity of the restrictions may become complex if many interrelated restrictions are to be imposed, but a sequential definition of priors and an application of Bayes' rule can achieve this in two simple steps. The first one is to define a prior for $\boldsymbol{\beta}$ that does not impose the restrictions or, to put it differently, a prior that disregards economic theory:

$$\mathrm{p}(\boldsymbol{\beta}) = \frac{|\mathbf{P}|^{1/2}}{(2\pi)^{K/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{\beta} - \mathbf{m})'\mathbf{P}(\boldsymbol{\beta} - \mathbf{m})\right\} \tag{2.31}$$

[4] In the previous example $\mathbf{g}(\boldsymbol{\beta})$ and $\mathbf{r}$ are:

$$\mathbf{g}(\boldsymbol{\beta}) = \begin{bmatrix}\beta_1 + \beta_2 \\ \beta_4\end{bmatrix} \qquad \text{and} \qquad \mathbf{r} = \begin{bmatrix}1 \\ 0\end{bmatrix}$$

respectively.

The second step imposes a distribution on $\mathbf{r}|\boldsymbol{\beta}$:

$$\mathrm{p}(\mathbf{r}|\boldsymbol{\beta}) = \frac{|\boldsymbol{\Xi}|^{1/2}}{(2\pi)^{q/2}} \exp\left\{-\frac{1}{2}(\mathbf{r}-\mathbf{R}\boldsymbol{\beta})' \boldsymbol{\Xi} (\mathbf{r}-\mathbf{R}\boldsymbol{\beta})\right\} \tag{2.32}$$

This density suggests, first of all, that the value of $\mathbf{r}$ is random and there is no guarantee that the restrictions will be satisfied exactly. However, they should be satisfied in expectation (the expected value of $\mathbf{r}$ is $\mathbf{R}\boldsymbol{\beta}$) and the variability of $\mathbf{r}$ around its mean can be controlled by the value of the precision matrix, $\boldsymbol{\Xi}$. For example, expressing strong prior beliefs that $\mathbf{R}\boldsymbol{\beta}$ should be close to $\mathbf{r}$ can be achieved by setting $\boldsymbol{\Xi}$ equal to a diagonal matrix with large values on the diagonal. Alternatively, a diagonal $\boldsymbol{\Xi}$ with small values on the diagonal would allow $\mathbf{r}$ to be far from what economic theory prescribes. Thus, $\boldsymbol{\Xi}$ can be set such that the model moves continuously from not imposing the restrictions at all to imposing them with greater conviction.

Applying Bayes' theorem on the last two densities leads to:

$$\mathrm{p}(\boldsymbol{\beta}|\mathbf{r}) = \frac{\mathrm{p}(\mathbf{r}|\boldsymbol{\beta}) \times \mathrm{p}(\boldsymbol{\beta})}{\mathrm{p}(\mathbf{r})} = \frac{|\check{\mathbf{P}}|^{1/2}}{(2\pi)^{K/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}-\check{\mathbf{m}})' \check{\mathbf{P}} (\boldsymbol{\beta}-\check{\mathbf{m}})\right\} \tag{2.33}$$

where $\check{\mathbf{P}} = (\mathbf{R}'\boldsymbol{\Xi}\mathbf{R}+\mathbf{P})$ and $\check{\mathbf{m}} = (\mathbf{R}'\boldsymbol{\Xi}\mathbf{R}+\mathbf{P})^{-1}(\mathbf{R}'\boldsymbol{\Xi}\mathbf{r}+\mathbf{P}\mathbf{m})$, while the procedure for obtaining this result is almost identical to the the one used for getting the full conditional for $\boldsymbol{\beta}$ in Section 2.4. The resulting density incorporates the restrictions and can be used as the "adjusted" or "updated" prior for $\boldsymbol{\beta}$ in the linear regression model.

♦ **Example 2.2 Aggregate Production Function (Continued)**
Consider again the data from the Penn World Table and assume that the aggregate production function takes the Cobb-Douglas form:

$$\log \mathsf{Y}_i = \beta_1 + \beta_2 \log \mathsf{K}_i + \beta_3 \log \mathsf{L}_i + \beta_4 \mathsf{trend}_i + \varepsilon_i$$

We estimated this model in the previous part of this example. Using the results from this model we approximate the probability of β_4 being positive (technological progress) as:

$$\mathrm{Prob}(\beta_4 > 0|\mathbf{y}) \approx 0.9908$$

A Cobb-Douglas production function exhibits increasing returns to scale if $\beta_2 + \beta_3 > 1$. Using the results from the estimated model, the probability of increasing returns to scale is approximated as:

$$\mathrm{Prob}(\beta_2 + \beta_3 > 1|\mathbf{y}) \approx 1$$

That is, given the data, we are almost certain that the production technology is characterized by increasing returns to scale. Finally, we can evaluate the plausibility of both statements at the same time by approximating $\mathrm{Prob}(\beta_4 > 0, \beta_2 + \beta_3 > 1|\mathbf{y})$. The following table presents the results for this compound statement as presented by BayES. From this table we see that BayES approximates the probability of each statement being true separately, before calculating the probability of the compound statement to 0.9908. In this example the probability of the compound statement being true is equal to the probability of the first statement only because the second statement is almost always true.

Condition	Cases	Successes	Probability
$\beta_4 > 0$	20000	19816	0.9908
$\beta_2 + \beta_3 > 1$	20000	20000	1
Overall	20000	19816	0.9908

Although the results above indicate that the technology is characterized by increasing returns to scale, we can still impose the restriction of constant returns to scale ($\beta_1 + \beta_2 = 1$); we are simply restricting the data to conform to a model/data-generating process that they seem not to support. The following table presents the results obtained by using BayES' `lm()` function while imposing this restriction with great conviction: we set the $\boldsymbol{\Xi}$ matrix that defines the precision of the right-hand side of the constrain $\beta_1 + \beta_2 = r$ equal to 10^8. The output elasticities with respect to the two inputs change slightly when compared to the unrestricted model we estimated in the previous part of this example and now sum up to one. Notice also that the standard deviation of the error term has increased because we are now forcing the data to conform to a specific theory/data-generating process.

	Mean	Median	Sd.dev.	5%	95%
constant	4.08074	4.08425	0.256233	3.65764	4.49684
logK	0.516158	0.515854	0.0222997	0.480052	0.553002
logL	0.483844	0.484147	0.0222999	0.447011	0.519957
trend	0.00251447	0.00251857	0.000482089	0.00171773	0.0032989
tau	67.927	67.8723	3.68118	61.934	74.1212
sigma_e	0.121467	0.121383	0.00329981	0.116153	0.127071

The following box contains code that can be used to reproduce the results presented in this example.

```
// import the data and transform the variables
Data = webimport("www.bayeconsoft.com/datasets/PWT.csv");
Data.constant = 1;          Data.logY = log(Data.Y);
Data.logK = log(Data.K);    Data.logL = log(Data.L);

// run the unconstrained Cobb-Douglas model
unconCD = lm(logY ~ constant logK logL trend);

// calculate the probability of beta4>0
test( unconCD.trend > 0 );

// calculate the probability of beta2+beta3>1
test( unconCD.logK + unconCD.logL > 1 );

// calculate the probability of (beta4>0) AND (beta2+beta3>1)
test( unconCD.trend > 0,
    unconCD.logK + unconCD.logL > 1.0 );

// estimate the constrained model (beta1+beta2=1)
conCD = lm(logY ~ constant logK logL trend,
    "constraints" = {logK+logL=1}, "Xi" = 1e9);
```

2.6.2 Model Comparison in the Linear Regression Model

Model comparison in the context of the linear regression model is a direct application of the general principles presented in subsection 1.3.4. To simplify the discussion we will consider the comparison of only two models, but extension to multiple models is straightforward. Suppose that we have the following two candidate models, which are derived from alternative economic theories:

$$\begin{aligned} \text{Model 0:} \quad & y_i = \mathbf{x}_i'\boldsymbol{\beta}_0 + \varepsilon_i, \qquad \varepsilon_i \sim \mathrm{N}\left(0, \tfrac{1}{\tau_0}\right) \\ \text{Model 1:} \quad & y_i = \mathbf{z}_i'\boldsymbol{\beta}_1 + \xi_i, \qquad \xi_i \sim \mathrm{N}\left(0, \tfrac{1}{\tau_1}\right) \end{aligned} \tag{2.34}$$

along with their priors, $\mathrm{p}_0\left(\boldsymbol{\beta}_0, \tau_0\right)$ and $\mathrm{p}_1\left(\boldsymbol{\beta}_1, \tau_1\right)$. We will keep using Normal and Gamma priors for the two $\boldsymbol{\beta}$s and two τs, respectively, but the prior hyperparameters may differ between the two models. The dependent variable is the same in both models, but there is no restriction on the independent variables: $\mathbf{x}$ could contain a subset of the variables in $\mathbf{z}$ or the other way around, the sets of independent variables in the two vectors could overlap only partially, with some of the variables in $\mathbf{x}$ not appearing in $\mathbf{z}$ and some of the variables in $\mathbf{z}$ not appearing in $\mathbf{x}$, or the sets of independent variables could be disjoint or overlap completely. In the latter case of the two models having exactly the same independent variables, the only difference between the models will be in the prior hyperparameters. For example, one model may use hyperparameters that stochastically restrict the values of some of its βs, as we saw in the preceding subsection, while the other model does not impose these restrictions. Model comparison here provides an indirect way of evaluating the plausibility of the restrictions.

As we did in subsection 1.3.4, we define M as a discrete random variable which can assume two values, 0 or 1, and which indicates which of the two models better describes how the data on the dependent variable are generated in the population. Setting $\mathrm{Prob}(M=0)$ and

$\text{Prob}(M=1)$ to values that express prior beliefs on the relative suitability of each model to describe the data-generating process, the *posterior odds ratio* is:

$$\frac{\text{Prob}(M=0|\mathbf{y})}{\text{Prob}(M=1|\mathbf{y})} = \frac{m(\mathbf{y}|M=0)}{m(\mathbf{y}|M=1)} \cdot \frac{\text{Prob}(M=0)}{\text{Prob}(M=1)} \tag{2.35}$$

The only thing left to do is to calculate the ratio of the two marginal likelihoods that define the *Bayes factor*. In the linear regression model these are given by:

$$m(\mathbf{y}|M=j) = \int_0^{\infty} \int_{-\infty}^{\infty} \text{p}_j\left(\mathbf{y}|\boldsymbol{\beta}_j, \tau_j, \bullet\right) \cdot \text{p}_j\left(\boldsymbol{\beta}_j, \tau_j\right) \, \text{d}\boldsymbol{\beta}_j \text{d}\tau_j \tag{2.36}$$

where $\text{p}_j\left(\mathbf{y}|\boldsymbol{\beta}_j, \tau_j, \bullet\right)$ and $\text{p}_j\left(\boldsymbol{\beta}_j, \tau_j\right)$ are the likelihood function and the prior, respectively, for model j and for $j=0,1$. These integrals cannot be calculated analytically even for such simple models as the ones we are considering here. There exist, however, a few approaches to approximate them:

- Gelfand & Dey (1994) use an identity that expresses the reciprocal of the marginal likelihood as an expectation of a ratio that involves known quantities and a probability density function, appropriately chosen by the researcher. The expectation is then approximated by simulation and using the draws from the posterior which were generated during estimation, in a way that resembles importance sampling. This approach is very general and can be directly applied to the linear regression model. Numerical instability issues may arise, however, in models with missing data[5] or if the free density is not chosen carefully.
- Lewis & Raftery (1997) propose using the *Laplace approximation* to the integrals that appear in the definition of the marginal likelihood. This method requires derivation of the mode of the function inside the integral. But if the posterior joint density of $\boldsymbol{\beta}$ and τ is approximately Normal, its mode will coincide with the mean and the posterior expected values of the parameters, along with their posterior covariance matrix, can be used for the approximation. The precision of this approximation, however, reduces if the posterior density is far from the Normal.
- Chib (1995) develops a technique that can be used to approximate the integral using additional simulations in a reduced Gibbs sampler. This approach requires a point for $\boldsymbol{\beta}$ and τ at which the posterior density is non-zero, but it does not have to be the mode of the posterior. Chib's approach works only when the full conditionals are known exactly (they have no missing constants of proportionality), but Chib & Jeliazkov (2001) extend the method to Metropolis-Hastings within Gibbs samplers.

♦ **Example 2.2 Aggregate Production Function (Continued)**
Using once again the data from the Penn World Table we will now compare the following two models for the aggregate production function:

$$\log \mathsf{Y}_i = \beta_1 + \beta_2 \log \mathsf{K}_i + \beta_3 \log \mathsf{L}_i + \beta_4 \mathsf{trend}_i + \varepsilon_i$$

and:

$$\begin{aligned} \log \mathsf{Y}_i = {} & \beta_1 + \beta_2 \log \mathsf{K}_i + \beta_3 \log \mathsf{L}_i + \beta_4 \mathsf{trend}_i \\ & + \beta_5 \log \mathsf{K}_i \log \mathsf{K}_i + \beta_6 \log \mathsf{K}_i \log \mathsf{L}_i + \beta_7 \log \mathsf{L}_i \log \mathsf{L}_i \\ & + \beta_8 \mathsf{trend}_i \log \mathsf{K}_i + \beta_9 \mathsf{trend}_i \log \mathsf{L}_i + \varepsilon_i \end{aligned}$$

The Cobb-Douglas model can be obtained from the translog model by restricting β_5 to β_9 in the latter to zero. Therefore, comparing the two models is equivalent to testing a set of five linear restrictions. Although The Cobb-Douglas model is nested withing the translog, this is not required for model comparison, in general.

The following two tables present the results obtained after estimating the two models, using the Lewis and Raftery and the Chib and Jeliazkov approximations to the logarithm of the marginal likelihood, respectively. Both the Cobb-Douglas and translog models were estimated using the vague default priors defined in BayES.

[5] We will encounter models with missing or *latent data* in following chapters.

Model	Log-Marginal Likelihood	Type of log-ML Approximation	Prior Model Probability	Posterior Model Probability
CobbDouglas	442.663	Lewis & Raftery	0.5	0.834778
Translog	441.043	Lewis & Raftery	0.5	0.165222

Model	Log-Marginal Likelihood	Type of log-ML Approximation	Prior Model Probability	Posterior Model Probability
CobbDouglas	442.671	Chib & Jeliazkov	0.5	0.833955
Translog	441.057	Chib & Jeliazkov	0.5	0.166045

The results do not change substantially when using the two alternative approximations: with equal prior model probabilities, the posterior model probability is considerably higher for the Cobb-Douglas model, although if we restrict attention only to these two models, the probability that the translog model expresses the "true" data-generating process is non-negligible ($\approx 16.5\%$).

The finding of the Cobb-Douglas model being preferred by the data in this application is driven mostly by the relatively little non-data information provided during estimation through the priors. Because the translog model nests the Cobb-Douglas, it should be able to accurately mimic the data-generating process, even if this process were the one described by the Cobb-Douglas specification. However, the translog model contains many more parameters that need to be estimated than the Cobb-Douglas and the lack of prior information on the values of these extra parameters penalize this large model heavily. With more informative priors for the parameters associated with the interaction terms of the translog specification, this model could turn out as being preferred by the data over the Cobb-Douglas.

The results presented above can be obtained in BayES using the code in the following box.

```
// import the data and transform the variables
Data = webimport("www.bayeconsoft.com/datasets/PWT.csv");
Data.constant = 1;          Data.logY = log(Data.Y);
Data.logK = log(Data.K);    Data.logL = log(Data.L);

// normalize inputs and create interaction terms
Data.logK = Data.logK - mean(Data.logK);
Data.logL = Data.logL - mean(Data.logL);
Data.logKlogK = Data.logK.*Data.logK;
Data.logKlogL = Data.logK.*Data.logL;
Data.logLlogL = Data.logL.*Data.logL;
Data.tlogK = Data.trend.*Data.logK;
Data.tlogL = Data.trend.*Data.logL;

// run the Cobb-Douglas model and request the calculation of the Chib and
// Jeliazkov approximation to the logarithm of the marginal likelihood
CobbDouglas = lm(logY ~ constant logK logL trend,
    "logML_CJ" = true);

// run the translog model and request the calculation of the Chib and
// Jeliazkov approximation to the logarithm of the marginal likelihood
Translog = lm(logY ~ constant logK logL trend
    logKlogK logKlogL logLlogL tlogK tlogL,
    "logML_CJ" = true);

// compare the two models using the Lewis-Raftery approximation
pmp( { CobbDouglas, Translog } );

// compare the two models using the Chib-Jeliazkov approximation
pmp( { CobbDouglas, Translog }, "logML_CJ"=true );
```

2.6.3 Predicting the Values of the Dependent Variable

A linear regression model of the general form $y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i$, along with the distributional assumption $\varepsilon_i \sim \mathrm{N}\left(0, \frac{1}{\tau}\right)$, expresses the assumptions on the data-generating process for the dependent variable in the population. This is done without reference to the data and the role of the data in applied research is to estimate the parameters of the model and evaluate the plausibility of statements that involve the values of these parameters. However, because the

model applies to the population, it is possible to use it to make stochastic statements about the values of the dependent variable, conditional on values for the independent variables which are not observed in the data.

To formalize this point, let $\mathbf{x}_*$ be a $K \times 1$ vector of values for the independent variables, chosen by the researcher, and let y_* be a random variable that is generated by the assumed data-generating process. From the properties of the model we get $y_*|\boldsymbol{\beta}, \tau \sim \mathrm{N}\left(\mathbf{x}_*'\boldsymbol{\beta}, \frac{1}{\tau}\right)$ and if the values of the parameters were known, we could use the Normal distribution to obtain the most likely value of y_* or to calculate the probability of it being within a certain interval. For doing so, one could use the most likely values of $\boldsymbol{\beta}$ and τ or their posterior expected values and plug them into the formulas for the mean and variance of y_*. Such an approach, however, would ignore the uncertainty associated with the values of the parameters. Explicitly accounting for this uncertainty requires integrating out $\boldsymbol{\beta}$ and τ from the joint density of y_* and the parameters, conditional on the observed data:

$$\mathrm{p}(y_*|\mathbf{x}_*, \mathbf{y}) = \int_0^{\infty} \int_{-\infty}^{\infty} \mathrm{p}(y_*, \boldsymbol{\beta}, \tau|\mathbf{x}_*, \mathbf{y}) \, \mathrm{d}\boldsymbol{\beta}\mathrm{d}\tau = \int_0^{\infty} \int_{-\infty}^{\infty} \mathrm{p}(y_*|\boldsymbol{\beta}, \tau, \mathbf{x}_*, \mathbf{y}) \, \pi(\boldsymbol{\beta}, \tau|\mathbf{y}) \, \mathrm{d}\boldsymbol{\beta}\mathrm{d}\tau \quad (2.37)$$

where $\pi(\boldsymbol{\beta}, \tau|\mathbf{y})$ is the posterior density of the parameters and $\mathrm{p}(y_*|\boldsymbol{\beta}, \tau, \mathbf{x}_*, \mathbf{y})$ is the probability density function of y_*, conditional on the values of the parameters:

$$\mathrm{p}(y_*|\boldsymbol{\beta}, \tau, \mathbf{x}_*, \mathbf{y}) = \frac{\tau^{1/2}}{(2\pi)^{1/2}} \exp\left\{-\frac{\tau}{2}\left(y_* - \mathbf{x}_*'\boldsymbol{\beta}\right)^2\right\} \quad (2.38)$$

Notice that, once we condition on $\boldsymbol{\beta}$ and τ, the density of y_* does not depend on the observed data, $\mathbf{y}$, because we have assumed that the value of the dependent variable for each potential observation is independent of the values of y_i for other potential observations. This is another way of seeing that the role of the observed data in the model is to provide information on the values of the parameters. Once this information has been extracted from the data and cast into information about $\boldsymbol{\beta}$ and τ, the observed data have nothing more to to say in relation to the value of y_*.

Equation (2.37) is the *posterior predictive density* in the context of the linear regression model. The integral cannot be evaluated analytically when independent Normal and Gamma priors are used for $\boldsymbol{\beta}$ and τ, but the moments of y_* or the probability of it being within a certain interval can be expressed as expectations and approximated by simulation. For example, the expected value of y_* is:

$$\mathrm{E}(y_*|\mathbf{x}_*, \mathbf{y}) = \int_{-\infty}^{\infty} y_* \cdot \mathrm{p}(y_*|\mathbf{x}_*, \mathbf{y}) \, \mathrm{d}y_* \quad (2.39)$$

Because $y_*|\boldsymbol{\beta}, \tau, \mathbf{x}_* \sim \mathrm{N}\left(\mathbf{x}_*'\boldsymbol{\beta}, \frac{1}{\tau}\right)$, samples from $\mathrm{p}(y_*|\mathbf{x}_*, \mathbf{y})$ can be obtained by using the G retained draws from the posterior for $\boldsymbol{\beta}$ and τ and then generating R draws for y_* from a Normal distribution, given the values from the g-th iteration of the Gibbs sampler. Given $Q = G \cdot R$ such draws, $y_*^{(1)}, y_*^{(2)}, \ldots y_*^{(Q)}$, the expectation can be approximated as:

$$\mathrm{E}(y_*|\mathbf{x}_*, \mathbf{y}) \approx \frac{1}{Q} \sum_{q=1}^{Q} y_*^{(q)} \quad (2.40)$$

Approximating other moments or functions of y_* involves simply changing the way y_* enters the summation in the expression above. For example, the variance of y_* can be approximated using the same Q draws from the posterior predictive density and the formula:

$$\mathrm{V}(y_*|\mathbf{x}_*, \mathbf{y}) \approx \frac{1}{Q} \sum_{q=1}^{Q} \left(y_*^{(q)} - \bar{y}_*\right)^2 \quad (2.41)$$

while the probability of y_* being within the interval $[c_1, c_2]$ can be approximated as:

$$\text{Prob}(c_1 \leq y_* \leq c_2 | \mathbf{x}_*, \mathbf{y}) \approx \frac{1}{Q} \sum_{q=1}^{Q} \mathbb{1}\Big(c_1 \leq y_*^{(q)} \leq c_2\Big) \tag{2.42}$$

where $\mathbb{1}(\bullet)$ is the *indicator function*.

2.7 Synopsis

This chapter introduced and covered the linear regression model in great detail. The error term in the model was assumed, throughout, to follow a Normal distribution with mean zero and precision parameter, τ, common to all potential observations. This assumption can be relaxed in various ways and this will be done in following chapters. We used a Normal prior for the slope parameters of the model and an independent Gamma prior for the precision parameter and showed that these priors are conjugate. This chapter was slightly more extensive than the ones that will follow for two reasons: (i) some concepts, like marginal effects and the distinction between the population and the sample, were encountered and discussed for the first time, and (ii) the Bayesian way of comparing models, evaluating the plausibility of statements that involve the model's parameters, as well as predicting the values of the dependent variable, were also discussed in the context of the linear model, so as to provide a concrete econometric example.

Chapter 3

Seemingly Unrelated Regressions

3.1 Overview

This chapter covers the *Seemingly Unrelated Regressions* (SUR) model, first introduced by Zellner (1962), who also coined the term for it. The SUR model is a direct extension of the linear regression model to the case where multiple dependent variables are modeled simultaneously. It is useful in its own right, as it provides additional information in relation to running multiple linear regressions separately, as well as a means of testing statements that involve restrictions of parameters which appear in different equations. In a Bayesian setting the SUR model also emerges as an intermediate step in estimating more complex models, such as multiple discrete response models.

The chapter starts with the setup of the SUR model, discussing its assumptions and the interpretation of its parameters. After presentation of the likelihood function, priors and full conditionals, a separate section is dedicated to imposing linear restrictions on the parameters of the model.

3.2 The System Approach to Linear Regression

In analogy to the linear regression model, economic theory may posit causal relationships among multiple independent and multiple response variables. The general form of such relationships can be expressed mathematically as $\mathbf{y} = \mathbf{f}(\mathbf{x})$, where $\mathbf{y}$ is a vector of dependent variables, $\mathbf{x}$ is a vector of independent variables and $\mathbf{f}(\cdot)$ is now a vector-valued function. The SUR model assumes that this function is linear in unknown parameters, as well as stochastic. In a model with M dependent variables, the SUR model expresses these relationships for a potential observation, i, as:

$$\begin{array}{ccccc} y_{1i} & = & \mathbf{x}_{1i}'\boldsymbol{\beta}_1 & + & \varepsilon_{1i} \\ y_{2i} & = & \mathbf{x}_{2i}'\boldsymbol{\beta}_2 & + & \varepsilon_{2i} \\ \vdots & & \vdots & & \vdots \\ y_{Mi} & = & \mathbf{x}_{Mi}'\boldsymbol{\beta}_M & + & \varepsilon_{Mi} \end{array} \tag{3.1}$$

where y_{mi} is the m-th dependent variable and it is assumed to be determined as the inner product of a $K_m \times 1$ vector of independent variables, $\mathbf{x}_{mi}$, and a $K_m \times 1$ vector of parameters, $\boldsymbol{\beta}_m$, plus an error term, ε_{mi}, and this interpretation holds for $m = 1, 2, \ldots, M$. Some or all of the independent variables may appear in multiple equations or each equation could have entirely different independent variables.

The unit of analysis in such a model could be a household, a firm, a country, etc., and this is what i is used as an index for. All M equations apply to the same unit and the only thing that changes from one equation to the other is the dependent variable being modeled and, possibly, the independent variables. A few typical examples where the SUR model can be applied are the following:

- consumer theory: the unit of analysis is a household or individual consumer and the dependent variables are the shares of expenditures on M categories of goods in total household expenditure. The independent variables in this context would be the prices or price indexes of the M categories of goods, as well as household income and socio-demographic characteristics.
- production theory: the unit of analysis is a firm and the dependent variables are the shares of costs from employing, renting or using different factors of production in total cost. The independent variables in this context would be the prices of the M factors of production, the amount of produced output and any other relevant firm characteristics.
- student performance: the unit of analysis is a student and the dependent variables are the scores obtained in different forms of assessment or different courses. The independent variables in this context would be a measure of the effort that the student put into studying for the assessment/course, the student's prior knowledge of the subject and any other relevant student characteristics.

One important thing to keep in mind is that in all applications of the SUR model a variable that appears as a dependent variable in one equation cannot be included in the set of independent variables in any other equation. That is, the independent variables are assumed to be determining the values of dependent variables simultaneously and that there is no effect from one dependent to another, once we condition on the xs.

The expression in (3.1) looks like a set of M linear regressions, stacked one under the other and, when viewing the model as such, a valid approach for estimating the M vectors of parameters would be to run M separate linear regression models. However, these *seemingly unrelated regressions* may be connected with each other, depending on the assumptions we impose on the error terms. If we assume that the ε_is are independent from each other across equations, then the regressions are indeed unrelated. If, on the other hand, we allow the error terms to be dependent, then the regressions are related and joint estimation can take advantage of the information contained in these error terms.

In general, we will assume that the $M \times 1$ vector of εs follows a multivariate Normal distribution, with expected value equal to a vector of zeros and precision matrix (inverse covariance matrix) $\mathbf{\Omega}$. The model in which the error terms are independent across equations can be obtained by restricting the precision matrix in the general model to be diagonal. By the properties of the multivariate Normal distribution one can deduce that, even when $\mathbf{\Omega}$ is not diagonal, the marginal distribution of each ε_{mi} is still Normal. Therefore, running M separate linear regressions does not contradict any of the assumptions of the SUR model. Why then should we consider joint estimation of all M equations? The answer is because, by imposing additional structure on the model, we can extract more information from the data. Intuitively, if the error terms associated with two equations in the system are correlated, then knowing the value of one can help us better predict the value of the other. Of course, we will never know the values of the error terms in practice, but modeling them using their joint density rather than their marginal densities can provide additional information. This information is then translated into higher precision in the estimation of the βs.

It may appear as if one has nothing to lose by estimating the M equations in a system. After all, if the error terms are independent, then the off-diagonal elements of $\mathbf{\Omega}$ should turn out to be zero or close to zeros and we are back to the case of running M linear regressions separately. There are two ways in which this argument may fail. First, if the error terms are indeed independent, by allowing them to be dependent when estimating a system we are over-parameterizing the model by having to estimate the off-diagonal elements of the precision matrix. Imposing a correct restriction on the data would lead to more precise estimates of

the βs than allowing $\mathbf{\Omega}$ to be non-diagonal. Second, misspecification of a single equation in the system, for example by missing a relevant variable, will be transmitted to all equations through the error terms, even if all other equations are correctly specified. As it is most often the case in applied econometrics, by imposing additional structure on the data, one runs the risk of imposing invalid assumptions on the data-generating process.

Just like the single-equation linear regression model, the SUR model can be given a conditional expectation interpretation: $\mathrm{E}\left(y_{mi}|\mathbf{x}_{mi}\right) = \mathbf{x}_{mi}'\boldsymbol{\beta}_m$ for $m = 1, 2, \ldots, M$. Similarly, nothing changes with respect to meaning of the βs in the SUR model: if β_{mk} is the coefficient associated with the k-th independent variable in the m-th equation of the system, then this is to be interpreted as the change in the expected value of y_m caused by a small change in the associated independent variable. It is possible for some of the independent variables to enter the model squared or in interactions with other independent variables. It is also possible for the statistical model to be specified in monotonic transformations, such as the logarithmic function, of the original variables that are implied by economic theory. In these cases the discussion in Section 2.5 around marginal effects extends to the SUR model, with the only difference being that now one has to keep track of the dependent variable that the marginal effect applies to.

The interpretation of $\mathbf{\Omega}$, on the other hand, requires some attention. It may be more natural to start by interpreting the variance matrix of the error terms, $\mathbf{\Omega}^{-1}$, rather than $\mathbf{\Omega}$ itself. The diagonal of this variance matrix stores the variances of each of the error terms, marginally with respect to the remaining error terms. The off-diagonal elements of $\mathbf{\Omega}^{-1}$ are the pairwise covariances of the error terms, again marginally with respect to the remaining εs. Mathematically:

$$\mathbf{\Omega}^{-1} = \begin{bmatrix} \mathrm{V}\left(\varepsilon_{1i}\right) & \mathrm{Cov}\left(\varepsilon_{1i}, \varepsilon_{2i}\right) & \cdots & \mathrm{Cov}\left(\varepsilon_{1i}, \varepsilon_{Mi}\right) \\ \mathrm{Cov}\left(\varepsilon_{2i}, \varepsilon_{1i}\right) & \mathrm{V}\left(\varepsilon_{2i}\right) & \ldots & \mathrm{Cov}\left(\varepsilon_{2i}, \varepsilon_{Mi}\right) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}\left(\varepsilon_{Mi}, \varepsilon_{1i}\right) & \mathrm{Cov}\left(\varepsilon_{Mi}, \varepsilon_{2i}\right) & \cdots & \mathrm{V}\left(\varepsilon_{Mi}\right) \end{bmatrix} \tag{3.2}$$

where all variances and covariances are taken marginally with respect to the error terms that do not appear as their arguments. If $\mathbf{\Omega}^{-1}$ is diagonal, all covariances are zero and the error terms are uncorrelated. Given the assumption of joint Normality of the error terms, if the error terms are uncorrelated, then they are also independent. On the other hand, if any of the off-diagonal elements are non-zero, then the associated error terms are not independent.

However, the model, as it is presented here, is parameterized in terms of $\mathbf{\Omega}$, not its inverse. It turns out that the precision matrix also has an intuitive interpretation, albeit not as direct. To start with, $\mathbf{\Omega}^{-1}$ will be diagonal if and only if $\mathbf{\Omega}$ is diagonal. Therefore, and given that the εs jointly follow a multivariate Normal distribution, whether the error terms across equations are independent or not can be inferred directly from $\mathbf{\Omega}$. The off-diagonal elements of the precision matrix can be used to obtain the partial correlations of the error terms, conditional on the remaining εs. In particular, the partial correlation coefficient of ε_{mi} and $\varepsilon_{\ell i}$ can be obtained as $-\frac{\omega_{m\ell}}{\sqrt{\omega_{mm}\cdot\omega_{\ell\ell}}}$, where $\omega_{m\ell}$ is the element of $\mathbf{\Omega}$ in row m, column ℓ.

Before closing this section we introduce an equivalent and more compact representation of the model in (3.1), which will be very useful in expressing the likelihood function, priors and full conditionals. The SUR model can be written as:

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i, \qquad \boldsymbol{\varepsilon}_i \sim \mathrm{N}\left(\mathbf{0}, \mathbf{\Omega}^{-1}\right) \tag{3.3}$$

where:

$$\underset{M\times 1}{\mathbf{y}_i} = \begin{bmatrix} y_{1i} \\ y_{2i} \\ \vdots \\ y_{Mi} \end{bmatrix}, \qquad \underset{M\times K}{\mathbf{X}_i} = \begin{bmatrix} \mathbf{x}_{1i}' & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{x}_{2i}' & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \mathbf{x}_{Mi}' \end{bmatrix}, \qquad \underset{K\times 1}{\boldsymbol{\beta}} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_M \end{bmatrix} \quad \text{and} \quad \underset{M\times 1}{\boldsymbol{\varepsilon}_i} = \begin{bmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \\ \vdots \\ \varepsilon_{Mi} \end{bmatrix}$$

and K is the total number of β parameters that appear across all equations: $K = \sum_{m=1}^{M} K_m$. From this representation it is easy to see that the SUR model expresses the relationship among

M dependent variables and K independent variables of the general form $\mathbf{y} = \mathbf{f}(\mathbf{x})$ as a linear function of the parameters and appends to it a Normally-distributed vector of error terms.

3.3 Likelihood, Priors and Full Conditionals

Given the assumption that the error term follows a multivariate Normal distribution, the density of $\mathbf{y}_i$ is also multivariate Normal, with mean $\mathbf{X}_i\boldsymbol{\beta}$ and precision matrix $\boldsymbol{\Omega}$. Adding to this an independence assumption on the error terms across observations, the likelihood of the SUR model with N observations is:

$$\begin{aligned} \mathrm{p}(\{\mathbf{y}_i\}|\boldsymbol{\beta},\boldsymbol{\Omega}) &= \prod_{i=1}^{N} \frac{|\boldsymbol{\Omega}|^{1/2}}{(2\pi)^{M/2}} \exp\left\{-\frac{1}{2}(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta})'\boldsymbol{\Omega}(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta})\right\} \\ &= \frac{|\boldsymbol{\Omega}|^{N/2}}{(2\pi)^{MN/2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{N}(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta})'\boldsymbol{\Omega}(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta})\right\} \end{aligned} \tag{3.4}$$

The parameters to be estimated in the model are the $K\times 1$ vector $\boldsymbol{\beta}$ and the $M\times M$ matrix $\boldsymbol{\Omega}$. A multivariate Normal prior for $\boldsymbol{\beta}$ appears as the natural choice and it turns out to be conjugate. $\boldsymbol{\Omega}$ is a precision matrix and it needs to be restricted by the prior to be symmetric and positive definite. Recall that in single-equation models we used a Gamma prior for the precision parameter, which restricted the value of the parameter to be positive and it also turned out to be conjugate. A natural choice for the prior of the precision matrix is the generalization of the Gamma distribution to multiple dimensions. The *Wishart distribution* constitutes such a generalization and it represents a distribution over symmetric and non-negative-definite matrices. Its probability density function is:

$$\mathrm{p}(\boldsymbol{\Omega}) = \frac{|\boldsymbol{\Omega}|^{\frac{n-M-1}{2}}|\mathbf{V}^{-1}|^{n/2}}{2^{nM/2}\Gamma_M\left(\frac{n}{2}\right)} \exp\left\{-\tfrac{1}{2}\operatorname{tr}\left(\mathbf{V}^{-1}\boldsymbol{\Omega}\right)\right\} \tag{3.5}$$

where n is the degrees-of-freedom parameter, $\mathbf{V}$ is the scale matrix and $\Gamma_M(\cdot)$ is the M-dimensional Gamma function. In the context of Bayesian inference, n and $\mathbf{V}$ will be the prior hyperparameters. An important point to keep in mind is that, for the density of the Wishart distribution to integrate to unity, n needs to be greater than or equal to M. This becomes particularly relevant if model comparison is to be performed after the estimation of the model, because improper priors invalidate any procedures used for approximating the marginal likelihood.

The expected value of a Wishart-distributed random matrix, $\boldsymbol{\Omega}$, is $n\mathbf{V}$ and the variance of the element in row m, column ℓ of $\boldsymbol{\Omega}$, marginally with respect to the remaining elements, is $n\left(v_{m\ell}^2 + v_{mm}\cdot v_{\ell\ell}\right)$, where $v_{m\ell}$ is the element in row m, column ℓ of $\mathbf{V}$. For a given value of n and using the expected-value formula only, a reasonable choice for $\mathbf{V}$ would be $\frac{1}{n}\mathbf{Q}$, where $\mathbf{Q}$ is a prior guess on the value of $\boldsymbol{\Omega}$. The choice of n and $\mathbf{V}$, however, also affect the variance of $\boldsymbol{\Omega}$. If we restrict attention to values of the hyperparameters that lead to proper priors and given that $\mathbf{V}$ is set to $\frac{1}{n}\mathbf{Q}$, the least informative prior for $\boldsymbol{\Omega}$ is obtained by setting n equal to the dimension of the problem. This is because, if $\mathbf{V} = \frac{1}{n}\mathbf{Q}$, the marginal variances are maximized for large values of $\mathbf{V}$ (therefore, small values of n), since the values of $\mathbf{V}$ enter the formula for the marginal variances squared, but n does not.

The posterior is, as always, proportional to the likelihood times the prior. The full conditionals for $\boldsymbol{\beta}$ and $\boldsymbol{\Omega}$ are obtained by simplifying the posterior and transforming the resulting expressions so that they resemble the probability density function of known distributions. The procedure for deriving the full conditional of $\boldsymbol{\beta}$ is similar to what was done for the linear regression model. Derivation of the full conditional of $\boldsymbol{\Omega}$ requires some transformations that involve the properties of the trace operator, but the process is rather straightforward. The interested reader is directed to Greenberg (2013, pp.136-137) for a step-by-step presentation of the process. The important thing for our purposes is that the Normal and Wishart priors are conjugate in the context of the SUR model and this simplifies sampling from the posterior considerably. These results are presented below in the form of a theorem.

THEOREM 3.1: Full Conditionals for the SUR Model
In the SUR model with Normally distributed error and M equations:

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i, \qquad \boldsymbol{\varepsilon}_i \sim \mathrm{N}\left(\mathbf{0}, \boldsymbol{\Omega}^{-1}\right)$$

and with a Normal prior for $\boldsymbol{\beta}$ and a Wishart prior for $\boldsymbol{\Omega}$:

$$\mathrm{p}(\boldsymbol{\beta}) = \frac{|\mathbf{P}|^{1/2}}{(2\pi)^{K/2}} \exp\left\{-\frac{1}{2}\left(\boldsymbol{\beta}-\mathbf{m}\right)'\mathbf{P}\left(\boldsymbol{\beta}-\mathbf{m}\right)\right\}$$

and:

$$\mathrm{p}(\boldsymbol{\Omega}) = \frac{|\boldsymbol{\Omega}|^{\frac{n-M-1}{2}}|\mathbf{V}^{-1}|^{n/2}}{2^{nM/2}\Gamma_M\left(\frac{n}{2}\right)} \exp\left\{-\tfrac{1}{2}\operatorname{tr}\left(\mathbf{V}^{-1}\boldsymbol{\Omega}\right)\right\}$$

the full conditionals of $\boldsymbol{\beta}$ and $\boldsymbol{\Omega}$ are, respectively, Normal and Wishart:

$$\pi\left(\boldsymbol{\beta}|\bullet\right) = \frac{|\tilde{\mathbf{P}}|^{1/2}}{(2\pi)^{K/2}} \exp\left\{-\frac{1}{2}\left(\boldsymbol{\beta}-\tilde{\mathbf{m}}\right)'\tilde{\mathbf{P}}\left(\boldsymbol{\beta}-\tilde{\mathbf{m}}\right)\right\}$$

and:

$$\pi\left(\boldsymbol{\Omega}|\bullet\right) = \frac{|\boldsymbol{\Omega}|^{\frac{\tilde{n}-M-1}{2}}|\tilde{\mathbf{V}}|^{-\frac{\tilde{n}}{2}}}{2^{\frac{\tilde{n}M}{2}}\Gamma_M\left(\frac{\tilde{n}}{2}\right)} \exp\left\{-\frac{1}{2}\operatorname{tr}\left(\tilde{\mathbf{V}}^{-1}\boldsymbol{\Omega}\right)\right\}$$

where:

- $\tilde{\mathbf{P}} = \sum_{i=1}^{N}\mathbf{X}_i'\boldsymbol{\Omega}\mathbf{X}_i + \mathbf{P}$
- $\tilde{\mathbf{m}} = \left(\sum_{i=1}^{N}\mathbf{X}_i'\boldsymbol{\Omega}\mathbf{X}_i + \mathbf{P}\right)^{-1}\left(\sum_{i=1}^{N}\mathbf{X}_i'\boldsymbol{\Omega}\mathbf{y}_i + \mathbf{P}\mathbf{m}\right)$
- $\tilde{n} = N + n$
- $\tilde{\mathbf{V}}^{-1} = \sum_{i=1}^{N}\left(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}\right)\left(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}\right)' + \mathbf{V}^{-1}$

Using these formulas we can discuss a well-known result in the frequentist treatment of the SUR model: if all equations in the system have the same independent variables, then running a SUR model results in the same point estimate for $\boldsymbol{\beta}$ as running the M linear regressions, equation by equation. This result does not hold exactly when the model is estimated using Bayesian methods because the M equations are now connected through the priors, as well as through the likelihood. However, as the role of the prior precision matrix in the formulas diminishes, either because it is set close to a matrix of zeros or because there are many observations available, the posterior mean of $\boldsymbol{\beta}$ from the SUR model converges to the posterior mean obtained by the separate regressions. To see how this works, it is convenient to set $\mathbf{P}$ exactly equal to a matrix of zeros, before considering what happens when the prior for $\boldsymbol{\beta}$ is proper.

In the case where all equations contain the same independent variables in a $k \times 1$ vector $\mathbf{x}_i$, $\mathbf{X}_i$ can be written as $\mathbf{I}_M \otimes \mathbf{x}_i'$, where $\mathbf{I}_M$ is the $M \times M$ identity matrix and $\otimes$ denotes the Kronecker product operator. With this expression and using the properties of the Kronecker product, $\mathbf{X}_i'\boldsymbol{\Omega}\mathbf{X}_i$ becomes $\boldsymbol{\Omega} \otimes \mathbf{x}_i\mathbf{x}_i'$ and $\mathbf{X}_i'\boldsymbol{\Omega}\mathbf{y}_i$ becomes $\left(\boldsymbol{\Omega} \otimes \mathbf{I}_k\right)\left(\mathbf{I}_M \otimes \mathbf{x}_i\right)\mathbf{y}_i$. With $\mathbf{P} = \mathbf{0}$, the expression for $\tilde{\mathbf{m}}$ can be written as:

$$\begin{aligned}\tilde{\mathbf{m}} &= \left(\boldsymbol{\Omega} \otimes \sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\left(\left(\boldsymbol{\Omega} \otimes \mathbf{I}_k\right)\sum_{i=1}^{N}\left(\mathbf{I}_M \otimes \mathbf{x}_i\right)\mathbf{y}_i\right) \\ &= \left(\mathbf{I}_M \otimes \left(\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\right)\sum_{i=1}^{N}\left(\mathbf{I}_M \otimes \mathbf{x}_i\right)\mathbf{y}_i\end{aligned} \tag{3.6}$$

Carrying-out the multiplications in this expression leads to:

$$\tilde{\mathbf{m}} = \begin{bmatrix} \left(\sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i'\right)^{-1} \sum_{i=1}^{N} \mathbf{x}_i y_{1i} \\ \left(\sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i'\right)^{-1} \sum_{i=1}^{N} \mathbf{x}_i y_{2i} \\ \vdots \\ \left(\sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i'\right)^{-1} \sum_{i=1}^{N} \mathbf{x}_i y_{Mi} \end{bmatrix} \tag{3.7}$$

which would be the posterior mean of $\boldsymbol{\beta}$ if one had used zero precision matrices to run the M separate linear regressions and stacked these means one under the other. However, when $\mathbf{P}$ is different from zero the second equality in (3.6) does not hold. Nevertheless, if $\mathbf{P}$ is not too restrictive, the two sums over i that appear in the expression for $\tilde{\mathbf{m}}$ will become larger as the sample size increases and the relationship will hold approximately.

♦ **Example 3.1 Secondary School Student Performance**

In this example we will consider part of the dataset used by Cortez & Silva (2008). The dataset contains information on 382 secondary school students in Portugal regarding their performance in two subjects, mathematics and Portuguese, as well as student characteristics and student effort per subject. The variables in the dataset are:

`mGrade, pGrade`	:	final grade obtained by the student in mathematics and Portuguese, respectively, on a 0-20 scale
`age`	:	age of the student in years
`female`	:	dummy variable, 1 if female
`mStudy, pStudy`	:	amount of time spent studying for mathematics and Portuguese, respectively, coded such that higher values correspond to greater effort
`mFails, pFails`	:	number of past class failures for mathematics and Portuguese, respectively
`mPaid, pPaid`	:	dummy variable, 1 if the student took extra paid classes in mathematics and Portuguese, respectively
`mAbsent, pAbsent`	:	number of absences from mathematics and Portuguese, respectively

The unit of analysis is the student and our objective is to estimate the effects of student behavior in relation to each subject and overall student characteristics on the grades obtained. We will consider the following two-equation model:

$$\begin{aligned} \text{mGrade}_i &= \beta_{11} + \beta_{12}\text{age}_i + \beta_{13}\text{female}_i \\ &\quad + \beta_{14}\text{mStudy}_i + \beta_{15}\text{mFails}_i + \beta_{16}\text{mPaid}_i + \beta_{17}\text{mAbsent}_i + \varepsilon_{1i} \\ \text{pGrade}_i &= \beta_{21} + \beta_{22}\text{age}_i + \beta_{23}\text{female}_i \\ &\quad + \beta_{24}\text{pStudy}_i + \beta_{25}\text{pFails}_i + \beta_{26}\text{pPaid}_i + \beta_{27}\text{pAbsent}_i + \varepsilon_{2i} \end{aligned}$$

The justification for using a SUR model in this example is that unobserved student characteristics may affect performance in both subjects. But because these characteristics are not observed, their effect will be absorbed by the error terms, making them correlated.

We will first run linear regression models for each equation separately before we compare the results to a SUR model. The following two tables present the results for the final grade in mathematics and Portuguese, respectively, obtained using BayES' `lm()` function.

	Mean	Median	Sd.dev.	5%	95%
constant	20.0977	20.1022	3.14519	14.917	25.242
age	-0.567237	-0.566791	0.188453	-0.878842	-0.255136
female	-1.56698	-1.56346	0.460623	-2.32726	-0.808936
mStudy	0.381196	0.382687	0.278552	-0.0795295	0.841523
mFails	-2.23463	-2.23236	0.310932	-2.75392	-1.71992
mPaid	0.346688	0.346526	0.449231	-0.395886	1.09017
mAbsent	0.0417686	0.0415376	0.0292148	-0.0057057	0.0903097
tau	0.0555064	0.0554304	0.00405269	0.0489308	0.0623213
sigma_e	4.25305	4.24743	0.156127	4.00583	4.5208

	Mean	Median	Sd.dev.	5%	95%
constant	11.0178	11.0221	2.02516	7.68222	14.3327
age	0.0239318	0.0241175	0.121614	-0.177171	0.225257
female	0.612066	0.613343	0.290043	0.133815	1.0894
pStudy	0.629213	0.629955	0.175254	0.338025	0.91804
pFails	-1.59277	-1.59117	0.284222	-2.06223	-1.12235
pPaid	-0.923214	-0.925777	0.551325	-1.829	-0.0156636
pAbsent	-0.0577804	-0.0580629	0.0286406	-0.104306	-0.0106136
tau	0.140246	0.140058	0.0102397	0.123635	0.157442
sigma_e	2.67563	2.67206	0.0982199	2.52029	2.84401

We now run a SUR model on the two equations using BayES' `sur()` function. The results are presented in the following table. This table contains information on the posterior moments and the associated 90% credible intervals for all 14 parameters that appear in he two equations. One thing to notice is that the posterior means are not radically different in the SUR and separate linear regression models. However, standard deviations are mostly smaller for the SUR model and the associated credible intervals shorter. This is to be expected, given that the SUR model uses additional information on the correlation of the error terms in the two equations.

	Mean	Median	Sd.dev.	5%	95%
mGrade					
constant	20.659	20.6385	3.14089	15.4885	25.8599
age	-0.63891	-0.638812	0.188163	-0.950864	-0.328369
female	-1.62363	-1.62252	0.46032	-2.38742	-0.872054
mStudy	0.539939	0.539651	0.275466	0.0857904	0.994168
mFails	-1.61059	-1.61056	0.29144	-2.09169	-1.13296
mPaid	0.439109	0.43924	0.397885	-0.214804	1.09452
mAbsent	0.0623385	0.0621403	0.026261	0.0194588	0.105657
pGrade					
constant	10.7444	10.7409	2.02199	7.43297	14.0455
age	0.0259941	0.0259337	0.121296	-0.171782	0.225373
female	0.582283	0.581926	0.28726	0.111596	1.05765
pStudy	0.68499	0.68342	0.17308	0.400274	0.968099
pFails	-1.68026	-1.68083	0.257794	-2.1041	-1.25697
pPaid	-0.709772	-0.710286	0.488809	-1.51059	0.108863
pAbsent	-0.0199308	-0.0200388	0.0263046	-0.0636699	0.0235892

BayES does not present summaries of the draws from the posterior distribution of $\boldsymbol{\Omega}$, but these draws are stored in memory and become available for post-estimation analysis if a left-hand side value is provided when running the SUR model. The posterior mean of $\boldsymbol{\Omega}$ is:

$$\mathrm{E}\left(\boldsymbol{\Omega}|\bullet\right) = \begin{bmatrix} 0.07169 & -0.05440 \\ -0.05440 & 0.18254 \end{bmatrix}$$

and using the draws stored in memory we can calculate the partial correlation coefficient between the two errors:

$$\mathrm{E}\left(\rho_{\varepsilon_1,\varepsilon_2}|\bullet\right) \approx \frac{1}{G}\sum_{g=1}^{G} -\frac{\omega_{12}^{(g)}}{\sqrt{\omega_{11}^{(g)}\omega_{22}^{(g)}}} = 0.47498$$

Creating a credible interval for $\rho_{\varepsilon_1,\varepsilon_2}$ or calculating the probability of it being above or below a certain threshold is straightforward. For example, $\mathrm{Prob}(\rho_{\varepsilon_1,\varepsilon_2}|\bullet) > 0.5$ can be approximated using BayES' `test()` function; for the current example this probability is 0.27525.

The results presented above can be obtained in BayES using the code in the following box.

```
// import the data and create a constant term
Data = webimport("www.bayeconsoft.com/datasets/Students.csv");
Data.constant = 1;

// run the two regressions separately
mathM = lm(mGrade ~ constant age female mStudy mFails mPaid mAbsent);
portM = lm(pGrade ~ constant age female pStudy pFails pPaid pAbsent);
```

```
// run the two regressions in a SUR system
jointM = sur( {
    mGrade ~ constant age female mStudy mFails mPaid mAbsent,
    pGrade ~ constant age female pStudy pFails pPaid pAbsent
    });

// print the posterior expected value of the precision matrix
print(jointM.Omega);

// calculate the partial correlation coefficient for the error terms
draws_rho = -jointM.Omega_2_1 ./ sqrt(jointM.Omega_1_1 .* jointM.Omega_2_2);
print(mean(draws_rho));

// test whether the partial correlation coefficient is greater than 0.5
test(draws_rho>0.50);
```

3.4 Cross-Equation Restrictions and the SUR Model

Apart from providing general associations among causal and response variables, economic theory often suggests restrictions on the ways these variables may interact with each other. If these theoretical restrictions can be expressed as constraints on the values of a model's parameters, then they can be imposed on the model and their validity can be examined using the procedures discussed in Section 2.6. In the case of models with multiple response variables the constraints may involve parameters that appear in different equations and the SUR model can be used as a means of imposing the constraints during estimation. It is difficult to proceed with the discussion at such an abstract level and without reference to particular constrains. Therefore, we will discuss in detail two examples, one from consumer theory and one from production theory, before returning to the statistical approach of imposing restrictions suggested by economic theory.

3.4.1 Demand Systems

Deaton & Muellbauer (1980) propose a system of demand equations for M goods or categories of goods, which satisfies many of the properties implied by consumer theory. Due to this feature, the system is called the *Almost Ideal Demand System*. The specification starts from an unobserved expenditure function of the form:

$$\begin{aligned} e\left(u, p_1, p_2, \ldots, p_M\right) = \alpha_0 + \sum_{m=1}^{M} \alpha_m \log p_m + \frac{1}{2} \sum_{m=1}^{M} \sum_{\ell=1}^{M} \alpha_{m\ell} \log p_m \log p_\ell \\ + u \cdot \left(\beta_0 \cdot \prod_{m=1}^{M} \beta_m p_m \right) \end{aligned} \tag{3.8}$$

where u is the utility level and $p_1, p_2, \ldots, p_M$ are the prices of the M goods. The αs and βs are the parameters of the model, but as it will become apparent in a while, not all of them can be estimated.

If consumers make their choices based on real rather than nominal prices, then the expenditure function should be homogeneous of degree one in prices: if all prices increase by a certain proportion, the expenditure required to achieve the same utility level will increase by that same proportion. This homogeneity restriction implies that the parameters should satisfy the following restrictions:

- $\sum_{m=1}^{M} \alpha_m = 1$
- $\sum_{m=1}^{M} \alpha_{m\ell} = 0$
- $\sum_{\ell=1}^{M} \alpha_{m\ell} = 0$
- $\sum_{m=1}^{M} \beta_m = 0$

Finally, due to Young's theorem, it should hold $\alpha_{m\ell} = \alpha_{\ell m}$ for all m and ℓ.

Because the expenditure function contains the unobserved utility level, u, in the right-hand side, it is impossible to estimate the parameters that appear in (3.8) directly from it. However,

Shephard's lemma can be used to obtain the demand functions implied by any expenditure function:

$$h_m(u, p_1, p_2, \ldots, p_M) = \frac{\partial e(u, p_1, p_2, \ldots, p_M)}{\partial p_m}, \qquad m = 1, 2, \ldots, M \tag{3.9}$$

where $h_m(u, p_1, p_2, \ldots, p_M)$ is the *Hicksian demand function* for good m. These demand functions are rather complex, but can be used to derive equations that express the expenditure on each good as a share in total expenditure:

$$s_m = \alpha_m + \sum_{\ell=1}^{M} \alpha_{m\ell} \log p_\ell + \beta_m \log \frac{E}{P}, \qquad m = 1, 2, \ldots, M \tag{3.10}$$

where E is total expenditure in all M goods and P is a price index constructed using the original prices and the α parameters. Due to this index, the αs enter the model in a non-linear fashion and, in practice, most applications replace P with an approximation calculated beforehand and without reference to these parameters. The linearized version of the demand system becomes:

$$\begin{array}{rcl} s_{1i} &=& \alpha_1 + \sum\limits_{\ell=1}^{M} \alpha_{1\ell} \log p_{\ell i} + \beta_1 \log \frac{E_i}{P_i} + \varepsilon_{1i} \\ s_{2i} &=& \alpha_2 + \sum\limits_{\ell=1}^{M} \alpha_{2\ell} \log p_{\ell i} + \beta_2 \log \frac{E_i}{P_i} + \varepsilon_{2i} \\ \vdots & & \vdots \qquad\qquad\qquad \vdots \\ s_{Mi} &=& \alpha_M + \sum\limits_{\ell=1}^{M} \alpha_{M\ell} \log p_{\ell i} + \beta_M \log \frac{E_i}{P_i} + \varepsilon_{Mi} \end{array} \tag{3.11}$$

where i indexes potential observations (consumers, households, different time periods, etc.). This system of equations resembles a SUR model and the restrictions mentioned above need to be imposed such that the linear-homogeneity property of the expenditure function is satisfied. Furthermore, the M shares appearing as the dependent variables in the system need to sum to one, by construction. Ensuring that the shares always sum to one requires, on top of the linear homogeneity constraints, also that the error terms across all equations and for each potential observation sum to zero. This requirement, however, renters the covariance matrix of $\boldsymbol{\varepsilon}$ singular. To take this issue into account, one out of the M equations is dropped from the system and the parameters associated with this equation are obtained from the parameters in the remaining equations and the parametric restrictions. Nevertheless, α_0 and β_0 that appear in the original expenditure function cannot be estimated, but this is something to be expected, given that utility is measured only on an ordinal scale.

3.4.2 Cost Functions and Cost Share Equations

Berndt & Wood (1975) appear to be the first to estimate a cost function along with the cost share equations implied by Shephard's lemma. The specification of the model starts from a translog cost function:

$$\begin{aligned} \log C_i = \beta_0 + \sum_{m=1}^{M} \beta_m \log w_{mi} + \frac{1}{2} \sum_{m=1}^{M} \sum_{\ell=1}^{M} \beta_{m\ell} \log w_{mi} \log w_{\ell i} \\ + \delta_1 \log y_i + \tfrac{1}{2} \delta_2 \log^2 y_i + \sum_{m=1}^{M} \gamma_m \log y_i \log w_{mi} + \varepsilon_i \end{aligned} \tag{3.12}$$

where C_i is the cost of production for a potential observation i, $w_{1i}, w_{2i}, \ldots, w_{Mi}$ are the prices of the M factors of production faced by i and y_i the amount of output produced by i. Similarly to the expenditure function, the cost function should be homogeneous of degree one in the input prices: if all input prices change proportionally, then the cost of producing the same amount of output should also change by the same proportion. Linear homogeneity in the ws implies that the parameters of the cost function should satisfy the following constraints:

- $\sum_{m=1}^{M} \beta_m = 1$
- $\sum_{m=1}^{M} \beta_{m\ell} = 0$
- $\sum_{\ell=1}^{M} \beta_{m\ell} = 0$
- $\sum_{m=1}^{M} \gamma_m = 0$

while from Young's theorem we get $\beta_{m\ell} = \beta_{\ell m}$ for all m and ℓ.

By applying Shephard's lemma on this cost function, one can obtain the input demand equations as functions of the parameters, input prices and output. Using these input demand functions, the cost of purchasing or renting each input m as a share of total cost is a linear function in the parameters of the cost function:

$$\begin{array}{rcl} s_{1i} &=& \beta_1 + \sum_{\ell=1}^{M} \beta_{1\ell} \log w_{\ell i} + \gamma_1 \log y_i + \varepsilon_{1i} \\ s_{2i} &=& \beta_2 + \sum_{\ell=1}^{M} \beta_{2\ell} \log w_{\ell i} + \gamma_2 \log y_i + \varepsilon_{2i} \\ \vdots & & \vdots \qquad\qquad \vdots \\ s_{Mi} &=& \beta_M + \sum_{\ell=1}^{M} \beta_{M\ell} \log w_{\ell i} + \gamma_M \log y_i + \varepsilon_{Mi} \end{array} \tag{3.13}$$

Again, we end up with a system of M equations with restrictions on the parameters that span multiple equations. Because the dependent variables should sum to one by construction, once we impose these restrictions, the covariance matrix of the error terms must become singular. Therefore, one of the share equations is typically dropped from the system before estimation.

The process of deriving a system of cost share equations in the production context is almost identical to what was done in the demand system example. However, there is a big difference between the two models. The original equation in the demand system (expenditure function) contains unobserved quantities and, therefore, cannot be used to estimate the parameters. Therefore, using the resulting system of expenditure shares was a necessity. In the case of the cost function, all quantities that appear in (3.12) are observable and one can get estimates of all parameters by running a linear regression model on it. Why then, should we bother to estimate the system of cost shares? One reason may be that the objective of the research is to confront production theory with the data or to infer whether producers are cost minimizers, in the context of the application. This can be done by evaluating the plausibility of the parametric constraints across equations. A second and probably more important reason is that the system approach allows reusing the observed data points twice to estimate a fixed set of parameters. This can be done by estimating a system of M equations, consisting of the original cost function and $M-1$ cost share equations. Imposing restrictions that stem from economic theory, not only imposes additional structure on the data-generating process, but also extracts more information from the data.

3.4.3 Imposing Linear Restrictions

We now turn to the task of imposing restrictions on the values of the parameters in a general SUR model. For this purpose it is convenient to use the representation:

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta} + \boldsymbol{\varepsilon}_i, \qquad \boldsymbol{\varepsilon}_i \sim \mathrm{N}\left(\mathbf{0}, \boldsymbol{\Omega}^{-1}\right) \tag{3.14}$$

where $\mathbf{y}_i$ and $\boldsymbol{\varepsilon}_i$ are $M \times 1$ vectors, $\boldsymbol{\beta}$ is $K \times 1$ and $\mathbf{X}_i$ is an $M \times K$ matrix. This is a convenient representation because all slope parameters are contained in a single vector and constraints on the values of $\boldsymbol{\beta}$ can be incorporated in the prior. If these constraints are linear in the slope parameters, that is if they take the form:

$$\mathbf{R}\boldsymbol{\beta} = \mathbf{r} \tag{3.15}$$

then the procedure discussed in Section 2.6 can be used for doing so. In the expression above $\mathbf{R}$ is a $q \times K$ matrix and $\mathbf{r}$ a $q \times 1$ vector, where q is the number of linear constraints to be imposed. In short, a multivariate Normal prior is placed on $\boldsymbol{\beta}$, with mean $\mathbf{m}$ and precision matrix $\mathbf{P}$, and which ignores the constraints altogether. Next, a multivariate Normal prior for

$\mathbf{r}|\boldsymbol{\beta}$, with mean $\mathbf{R}\boldsymbol{\beta}$ and precision matrix $\boldsymbol{\Xi}$, is used to define how far away $\boldsymbol{\beta}$ is allowed to be from satisfying the restrictions. Finally, an application of Bayes' theorem updates the prior for $\boldsymbol{\beta}$ with the information contained in the constraints.

♦ **Example 3.2 Aggregate Cost Function**

In this example we will consider again the data from the Penn World Table (Feenstra et al., 2015), which we first used in Example 2.2. This dataset contains annual information on a series of aggregate variables for the EU-15 Member States from 1970 to 2014. Apart from the variables that we used before (real GDP, capital stock and labor input), we will now also use the following variables from the dataset:

`w`	:	annual compensation of a unit of labor, adjusted for human capital ($2011)
`r`	:	rental price of capital (proportion of the value of capital stock)
`C`	:	cost of production

We specify a model for the aggregate cost function and the resulting cost share equations, along the lines presented in subsection 3.4.2:

$$\begin{aligned}\log \mathsf{C}_i = \beta_0 &+ \beta_K \log \mathsf{r}_i + \beta_L \log \mathsf{w}_i + \tfrac{1}{2}\beta_{KK} \log^2 \mathsf{r}_i + \beta_{KL} \log \mathsf{r}_i \log \mathsf{w}_i + \tfrac{1}{2}\beta_{LL} \log^2 \mathsf{w}_i \\ &+ \delta_1 \log \mathsf{Y}_i + + \delta_2 \log^2 \mathsf{Y}_i + \gamma_K \log \mathsf{Y}_i \log \mathsf{r}_i + \gamma_L \log \mathsf{Y}_i \log \mathsf{w}_i + \varepsilon_{1i} \\ \mathsf{sK}_i &= \beta_K + \beta_{KK} \log \mathsf{r}_i + \beta_{KL} \log \mathsf{w}_i + \gamma_K \log \mathsf{Y}_i + \varepsilon_{2i} \\ \mathsf{sL}_i &= \beta_L + \beta_{KL} \log \mathsf{r}_i + \beta_{LL} \log \mathsf{w}_i + \gamma_L \log \mathsf{Y}_i + \varepsilon_{3i}\end{aligned}$$

where sK and sL are the shares of capital and labor inputs, respectively, in the cost of production. The restrictions that we need to impose on the βs and γs for the cost function to be homogeneous of degree one in r and w are:

- $\beta_K + \beta_L = 1$
- $\beta_{KK} + \beta_{KL} = 0$
- $\beta_{KL} + \beta_{LL} = 0$
- $\gamma_K + \gamma_L = 0$

During estimation we also need to make sure that the values of the parameters that appear in more that a single equation are restricted to be the same. Due to singularity of the error covariance matrix when all parametric constraints are imposed, we need to drop one of the cost share equations from the system and, for this example, we will drop the labor share equation.

The following table presents the results obtained by estimating the system consisting of the cost function and the capital share equation using BayES' `sur()` function. From this table we can see that the constraints hold approximately at the posterior means and medians of the parameters. A parameter of particular interest when estimating a cost function is the one associated with the logarithm of output. In this example, $\mathrm{E}(\delta_1|\bullet) \approx 0.976$ and, because it is smaller than one, it suggests that the underlying production technology is characterized by increasing returns to scale at the geometric mean of the data; a result that we also obtained by estimating the parameters of a production function in Example 2.2.

	Mean	Median	Sd.dev.	5%	95%
logC					
constant	12.7149	12.7148	0.00582343	12.7052	12.7244
logr	0.397411	0.397414	0.00234082	0.393545	0.401282
logw	0.602527	0.602525	0.00234095	0.598654	0.606402
logrlogr	0.0930304	0.0930325	0.00352821	0.0872569	0.0988219
logrlogw	-0.18606	-0.18606	0.00705639	-0.197617	-0.174525
logwlogw	0.0930302	0.0930295	0.00352831	0.0872615	0.0988132
logY	0.97556	0.975545	0.00377648	0.969294	0.981731
logYlogY	0.0127822	0.0127701	0.00169557	0.00999777	0.0155466
logYlogr	-0.00747926	-0.00751856	0.00176828	-0.0103636	-0.00453448
logYlogw	0.00747897	0.00751737	0.00176843	0.00453253	0.0103668
sK					
constant	0.397411	0.397414	0.00234071	0.393538	0.401288
logr	0.186062	0.186061	0.00705629	0.17452	0.197649
logw	-0.186059	-0.18606	0.00705653	-0.197616	-0.174531
logY	-0.00747984	-0.0075215	0.00176798	-0.0103651	-0.00453182

The results presented above can be obtained in BayES using the code in the following box.

```
// import the data into a dataset called Data
Data = webimport("www.bayeconsoft.com/datasets/PWT.csv");

// construct the constant term  and the cost share of capital
Data.constant = 1;
Data.sK = Data.r .* Data.K ./ Data.C;

// take logs or relevant variables
Data.logC = log(Data.C);
Data.logY = log(Data.Y);
Data.logr = log(Data.r);
Data.logw = log(Data.w);

// normalize variables and create interaction terms
Data.logr = Data.logr - mean(Data.logr);
Data.logw = Data.logw - mean(Data.logw);
Data.logY = Data.logY - mean(Data.logY);
Data.logrlogr = Data.logr.*Data.logr;
Data.logrlogw = Data.logr.*Data.logw;
Data.logwlogw = Data.logw.*Data.logw;
Data.logYlogY = Data.logY.*Data.logY;
Data.logYlogr = Data.logY.*Data.logr;
Data.logYlogw = Data.logY.*Data.logw;

// run the SUR model while imposing all eight constraints
SURmodel = sur( {
    logC ~ constant logr logw logrlogr logrlogw logwlogw
        logY logYlogY logYlogr logYlogw,
    sK ~ constant logr logw logY
    }, "constraints" = {
    logC$logr + logC$logw = 1,
    2*logC$logrlogr + logC$logrlogw = 0,
    2*logC$logwlogw + logC$logrlogw = 0,
    logC$logYlogr + logC$logYlogw = 0,
    sK$constant - logC$logr = 0,
    sK$logr - 2*logC$logrlogr = 0,
    sK$logw - logC$logrlogw = 0,
    sK$logY - logC$logYlogr = 0
    }, "Xi" = 1e9*eye(8,8));
```

3.5 Synopsis

This chapter covered in detail the Seemingly Unrelated Regressions (SUR) model. The model was introduced as a direct extension to the single-equation linear model and its parameters interpreted by viewing it as another conditional expectation specification. We used Normal and Wishart priors for the slope parameters and the precision matrix, respectively, which are both conjugate. Two extensive examples, one from consumer theory and one from production theory, were presented to motivate the need for a model that can impose parametric constrains that span multiple equations.

Chapter 4

Data Augmentation

4.1 Overview

This chapter introduces and discusses the *data-augmentation* technique. Data augmentation was formalized by Tanner & Wong (1987), but has its roots in the work of Rubin (1978, 1980) and Li (1988), who were dealing with problems of imputing missing values in datasets. Tanner & Wong were the first to connect the method to the *Expectation-Maximization* (EM) *algorithm* (Dempster et al., 1977), which works in models estimated by maximum-likelihood, and to introduce missing or *latent data* artificially into the analysis for the purpose of facilitating computations. Data augmentation can be extremely useful in complex models, where sampling directly from the posterior distribution of the parameters may be challenging, but once the problem is cast into a latent-data model, the likelihood simplifies considerably.

The following section provides the mathematical/probabilistic justification of data augmentation, in a general setting. A version of the Gibbs sampler in latent-data problems is presented next, along with a brief discussion around the potential usefulness of the byproducts of the sampler. The last section of this chapter presents two interesting applications of data augmentation: the linear regression model with heteroskedastic error and the stochastic-frontier model.

4.2 Data Augmentation in Latent-Data Problems

Consider a general econometric model with a $K \times 1$ vector of parameters to be estimated, $\boldsymbol{\theta}$, and likelihood function $\mathrm{p}(\mathbf{y}|\boldsymbol{\theta})$, where $\mathbf{y}$ is a vector that will store the *observed data*. Suppose also that $\mathrm{p}(\boldsymbol{\theta})$ is the prior density of $\boldsymbol{\theta}$. Bayesian inference in such a model would proceed by sampling from the posterior distribution of the parameters:

$$\pi\left(\boldsymbol{\theta}|\mathbf{y}\right) \propto \mathrm{p}(\mathbf{y}|\boldsymbol{\theta}) \cdot \mathrm{p}(\boldsymbol{\theta}) \tag{4.1}$$

and summarizing the draws. This is feasible, given the generality of the Gibbs sampler and the Metropolis-Hastings algorithms, even for the most complex models. However, if the likelihood function is unconventional and no conjugate priors exist, tailoring the samplers to the problem may become very challenging and the algorithms may be plagued by very large autocorrelation times. Suppose, that there exists a random variable, conditionally upon which the likelihood simplifies considerably. Let $\mathbf{z}$ be a vector that would store the values of this random variable. Because $\mathbf{z}$ is, by assumption, not observable it represents the *latent data*.

Since $\mathbf{z}$ is not observable, inferences about $\boldsymbol{\theta}$ cannot be made conditionally on the latent data. The obvious approach would then be to integrate the latent data from the joint density of $\boldsymbol{\theta}$ and $\mathbf{z}$, given the observed data:

$$\pi(\boldsymbol{\theta}|\mathbf{y}) = \int_Z \pi(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y})\, \mathrm{d}\mathbf{z} = \int_Z \pi(\boldsymbol{\theta}|\mathbf{y}, \mathbf{z})\, \mathrm{p}(\mathbf{z}|\mathbf{y})\, \mathrm{d}\mathbf{z} \tag{4.2}$$

The last integral contains two densities that are not known, but depend on the specification of the model:

- $\pi(\boldsymbol{\theta}|\mathbf{z}, \mathbf{y})$ is the posterior density of the parameters, conditional on both observed and latent data. It can be obtained from Bayes' theorem:

$$\pi(\boldsymbol{\theta}|\mathbf{y}, \mathbf{z}) \propto \mathrm{p}(\mathbf{y}, \mathbf{z}|\boldsymbol{\theta})\, \mathrm{p}(\boldsymbol{\theta}) = \mathrm{p}(\mathbf{y}|\boldsymbol{\theta}, \mathbf{z})\, \mathrm{p}(\mathbf{z}|\boldsymbol{\theta})\, \mathrm{p}(\boldsymbol{\theta}) \tag{4.3}$$

  where $\mathrm{p}(\mathbf{y}|\boldsymbol{\theta}, \mathbf{z})$ is easy to handle by construction: this is the reason the latent data are introduced to the problem in the first place. $\mathrm{p}(\mathbf{y}, \mathbf{z}|\boldsymbol{\theta})$ is the density of both observed and latent data and is appropriately called the *complete-data likelihood*, so that it can be distinguished from $\mathrm{p}(\mathbf{y}|\boldsymbol{\theta})$, which is called the *incomplete-* or *observed-data likelihood*. $\mathrm{p}(\mathbf{z}|\boldsymbol{\theta})$ is the density of latent data, conditionally on the parameters, but marginally with respect to the observed data.

- $\mathrm{p}(\mathbf{z}|\mathbf{y})$ is the predictive density of the latent data, given the observed. It can be obtained by integrating $\boldsymbol{\theta}$ from the joint density of $\mathbf{z}$ and $\boldsymbol{\theta}$, given $\mathbf{y}$:

$$\mathrm{p}(\mathbf{z}|\mathbf{y}) = \int_\Theta \pi(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y})\, \mathrm{d}\boldsymbol{\theta} = \int_\Theta \mathrm{p}(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta})\, \pi(\boldsymbol{\theta}|\mathbf{y})\, \mathrm{d}\boldsymbol{\theta} \tag{4.4}$$

Expressing the predictive density of the latent data as an integral that involves $\pi(\boldsymbol{\theta}|\mathbf{y})$ gives the impression that we are going around in circles: to get $\pi(\boldsymbol{\theta}|\mathbf{y})$ from (4.2) we need $\mathrm{p}(\mathbf{z}|\mathbf{y})$ and to get $\mathrm{p}(\mathbf{z}|\mathbf{y})$ from (4.4) we need $\pi(\boldsymbol{\theta}|\mathbf{y})$. But that is the purpose of the exercise: Tanner & Wong substitute (4.4) into (4.2), change the order of integration and view the resulting expression as an operator fixed-point equation.[1] They then motivate the following iterative algorithm as a successive-substitution method for solving fixed-point problems:

(a) sample $\mathbf{z}^{(1)}, \mathbf{z}^{(2)}, \ldots, \mathbf{z}^{(Q)}$, $Q \geq 1$, from $\mathrm{p}(\mathbf{z}|\mathbf{y})$, given the current approximation to $\pi(\boldsymbol{\theta}|\mathbf{y})$. This step is further broken into the following items:

  (a1) sample $\boldsymbol{\theta}$ from the current approximation of $\pi(\boldsymbol{\theta}|\mathbf{y})$

  (a2) sample $\mathbf{z}^{(1)}, \mathbf{z}^{(2)}, \ldots, \mathbf{z}^{(Q)}$ from $\mathrm{p}(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta})$ and using the value for $\boldsymbol{\theta}$ that was generated in **(a1)**

(b) update the current approximation to $\pi(\boldsymbol{\theta}|\mathbf{y})$ using $\frac{1}{Q}\sum_{q=1}^{Q} \pi\left(\boldsymbol{\theta}\middle|\mathbf{y}, \mathbf{z}^{(q)}\right)$ and repeat

One has to keep in mind that at the time the paper by Tanner & Wong was published the Gibbs sampler had not yet gained prominence among statisticians and this fixed-point view on the problem circumvents the issue of sampling from complex distributions. In practice, data-augmentation algorithms are usually implemented by setting Q in step **(a1)** above, equal to one, while treating the latent data as additional quantities to be estimated, along with the parameters. In this context, the joint posterior density of $\boldsymbol{\theta}$ and $\mathbf{z}$ is given by:

$$\pi(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y}) = \frac{\mathrm{p}(\mathbf{y}, \mathbf{z}|\boldsymbol{\theta})\, \mathrm{p}(\boldsymbol{\theta})}{\mathrm{p}(\mathbf{y})} \propto \mathrm{p}(\mathbf{y}|\boldsymbol{\theta}, \mathbf{z})\, \mathrm{p}(\mathbf{z}|\boldsymbol{\theta})\, \mathrm{p}(\boldsymbol{\theta}) \tag{4.5}$$

[1] One subtle point in the derivation is that we need to distinguish between $\boldsymbol{\theta}$ as an argument of $\pi(\bullet|\mathbf{y})$ and $\pi(\bullet|\mathbf{y}, \mathbf{z})$ and the dummy of integration in (4.4). Tanner & Wong use $\boldsymbol{\phi}$ as the dummy of integration to derive the fixed-point equation.

A Gibbs sampler can now be implemented, which iterates between sampling from the full conditionals of $\boldsymbol{\theta}$ and $\mathbf{z}$, either in one or multiple blocks for each one of them. It is stressed that these full conditionals are based on the complete-data likelihood and the prior density of $\boldsymbol{\theta}$, as these appear on the numerator of the fraction in the last expression. Once the Gibbs sampler completes, ignoring the draws on the latent data and considering only the draws on the parameters amounts to integrating-out $\mathbf{z}$ from $\pi(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y})$, as expressed in the first equality of (4.2).

A generic version of the Gibbs sampler with data-augmentation is presented in Algorithm 4.1. Sampling for $\boldsymbol{\theta}$ corresponds to step **(a1)** in the Tanner & Wong formulation, but with $Q = 1$. In this case, $\mathbf{z}$ is integrated-out from $\pi(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y})$ using only a single draw and step **(b)** becomes degenerate. Sampling for $\mathbf{z}$ in the Gibbs sampler corresponds to step **(a2)**.

Algorithm 4.1 Gibbs Sampler with Data Augmentation

```
set the number of burn-in iterations, D
set the number of draws to be retained, G
set θ to a reasonable starting value
set z to a reasonable starting value
for g = 1:(D+G) do
  draw θ from π(θ|y, z), either in one or multiple blocks
  draw z from p(z|y, θ), either in one or multiple blocks

  if g > D then
    store the current value of θ
    possibly store the current value of z
  end if
end for
```

One interesting feature of the Gibbs sampler in Algorithm 4.1 is that it allows for storing the draws on $\mathbf{z}$. This is because, in the context of an application, the latent data can have a meaningful interpretation and drawing inferences on them may be part of the objectives of the analysis. Because the Gibbs sampler produces draws from the joint posterior density of $\boldsymbol{\theta}$ and $\mathbf{z}$, the draws on $\mathbf{z}$ alone are from the posterior density of the latent data, marginally with respect to $\boldsymbol{\theta}$ and conditional only on the observed data. Therefore, these draws can be used to make probabilistic statements regarding the values of $\mathbf{z}$.

4.3 Applications of Data Augmentation

This section considers two applications of data augmentation. Both of them are interesting in their own right and give rise to classes of models which can be viewed as direct extensions to the linear regression model. Although the parameters of many of the models in the two classes can be estimated without making use of data augmentation, application of the technique simplifies the analysis considerably.

4.3.1 The Linear Model with Heteroskedastic Error

In the treatment of the linear regression model in Chapter 2 we maintained the assumption that the error term for each potential observation, i, follows a Normal distribution with mean zero and precision τ. In the linear regression model with *heteroskedastic error* we will relax the assumption that the ε_is have the same precision parameter for all potential observations, while we will keep assuming that they follow a Normal distribution and are independent of each other. Mathematically, the model becomes:

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i, \qquad \varepsilon_i \sim \mathrm{N}\left(0, \tfrac{1}{\tau_i}\right) \tag{4.6}$$

Of course, in an application with N observations we should not expect to be able to estimate all N precision parameters, along with the K βs, unless very informative priors are imposed on them; there is simply not enough information in the data. An alternative course of action is to impose some additional structure, in a *hierarchical* fashion, on the way the τ_is are determined in the population. There are a few approaches for doing so. Koop (2003, pp.124-130) describes a procedure where each τ_i is defined as the product of a common precision parameter and an observation-specific random variable, which is assumed to follow an Exponential distribution. A model with similar assumptions is presented in Greenberg (2013, pp.51-52), where it is shown that the model is equivalent to assuming that the error term follows a Student-t distribution (see also Geweke, 1993 on this). We will take here a different approach, which is general enough to account for heteroskedasticity of unknown form, as well as allow for estimating the effect of particular variables on the τ_is.

Because precision parameters need to be positive, we will assume that the logarithm of each τ_i follows a Normal distribution:

$$\log \tau_i \sim \mathrm{N}\left(\mathbf{w}_i'\boldsymbol{\delta}, \tfrac{1}{\phi}\right) \tag{4.7}$$

where $\mathbf{w}_i$ is an $L \times 1$ vector of observable variables which affect the precision of ε_i, and $\boldsymbol{\delta}$ is an $L \times 1$ vector of parameters. ϕ is another precision parameter to be estimated. The model allows for $\mathbf{w}_i$ to consist of only a constant term, in which case each $\log \tau_i$ follows a Normal distribution with common mean. Keep in mind that the expression above is not a prior density in the sense we have been using priors until now. Rather, $\boldsymbol{\delta}$ and ϕ are additional parameters to be estimated, while viewing the model from an incomplete-data perspective leads to a natural interpretation of the τ_is as latent data: if the τ_is were observable then we would be able to estimate the model's parameters using very similar full conditionals to the ones presented in Theorem 2.1. Data augmentation provides a way of integrating-out the uncertainty associated with the unobserved τ_is when drawing inferences on the model's parameters: $\boldsymbol{\beta}$, $\boldsymbol{\delta}$ and ϕ.

Given that each τ_i follows a log-Normal distribution, the complete-data likelihood for this model is:

$$\begin{aligned}
\mathrm{p}\left(\mathbf{y}, \{\tau_i\} \middle| \mathbf{X}, \mathbf{W}, \boldsymbol{\beta}, \boldsymbol{\delta}, \phi\right) &= \prod_{i=1}^{N} \mathrm{p}\left(y_i \middle| \mathbf{x}_i, \boldsymbol{\beta}, \tau_i\right) \mathrm{p}\left(\tau_i \middle| \mathbf{w}_i, \boldsymbol{\delta}, \phi\right) \\
&= \prod_{i=1}^{N} \frac{\tau_i^{1/2}}{(2\pi)^{1/2}} \exp\left\{-\frac{\tau_i \left(y_i - \mathbf{x}_i'\boldsymbol{\beta}\right)^2}{2}\right\} \\
&\quad \times \prod_{i=1}^{N} \frac{\phi^{1/2}}{\tau_i \, (2\pi)^{1/2}} \exp\left\{-\frac{\phi \left(\log \tau_i - \mathbf{w}_i\boldsymbol{\delta}\right)^2}{2}\right\}
\end{aligned} \tag{4.8}$$

where $\mathbf{W}$ is an $N \times L$ matrix that stores the values of the variables that affect the precision of the error term, for all N observations. The first equality above is obtained by expressing the joint density of observed and latent data as the product of a conditional density and a marginal density. Once, however, we condition on the τ_is, $\mathbf{y}$ no longer depends on $\mathbf{W}$, $\boldsymbol{\delta}$ and ϕ. Likewise, the τ_is depend on no other parameters or data, once we condition on $\mathbf{W}$, $\boldsymbol{\delta}$ and ϕ. The second equality results from the Normality of the error terms and log-Normality of the τ_is and the conditional independence of the y_is, as well as of the τ_is.

As in the linear regression model with *homoskedastic* error, we will use a Normal prior for $\boldsymbol{\beta}$, with mean $\mathbf{m}_\beta$ and precision matrix $\mathbf{P}_\beta$. For $\boldsymbol{\delta}$ we will, again, use a Normal prior, with mean $\mathbf{m}_\delta$ and precision matrix $\mathbf{P}_\delta$, while for ϕ we will use a Gamma prior, with shape and rate parameters a and b, respectively. Specification of the priors completes the specification of the model and by applying Bayes' theorem we get:

$$\pi\left(\boldsymbol{\beta}, \boldsymbol{\delta}, \phi, \{\tau_i\} \middle| \mathbf{y}, \mathbf{X}, \mathbf{W}\right) \propto \mathrm{p}\left(\mathbf{y}, \{\tau_i\} \middle| \mathbf{X}, \mathbf{W}, \boldsymbol{\beta}, \boldsymbol{\delta}, \phi\right) \mathrm{p}(\boldsymbol{\beta}) \, \mathrm{p}(\boldsymbol{\delta}) \, \mathrm{p}(\phi) \tag{4.9}$$

The densities in the right-hand side of last expression are all known: the the first density is given in (4.8), $\mathrm{p}(\boldsymbol{\beta})$ and $\mathrm{p}(\boldsymbol{\delta})$ are multivariate-Normal densities and $\mathrm{p}(\phi)$ is a Gamma density.

Before we can implement a Gibbs sampler we need to derive the full conditionals of all unobserved quantities: $\boldsymbol{\beta}$, $\boldsymbol{\delta}$, ϕ and the τ_is (all N of them). It is stressed that the τ_is are not parameters in the model and, therefore, have no priors associated with them. They are, however, unobserved random variables and data augmentation requires sampling from their conditional (on everything else in the model) density to integrate-out the uncertainty associated with them. Although the algebraic transformations required to get the full conditionals of the parameters and latent-data are tedious, we have seen versions of many of them before:

- deriving the full conditional of $\boldsymbol{\beta}$ in the linear model with heteroskedastic error is very similar to the case of homoskedastic error
- deriving the full conditionals of $\boldsymbol{\delta}$ and ϕ requires exactly the same transformations presented above Theorem 2.1, with $\log \tau_i$ assuming the role of y_i and $\mathbf{w}_i$ that of $\mathbf{x}_i$
- deriving the full conditional of each τ_i is different from what we encountered until now, but the algebraic transformations are straightforward

The important thing is that, again, all three priors used here are conjugate for their respective parameters and this simplifies sampling from their full conditionals. On the other hand, the full conditional of the τ_is does not belong any known parametric family and a different approach, other than direct sampling, must be used to obtain samples from it. Metropolis-Hastings updates for each τ_i are certainly feasible, albeit not necessarily the most efficient choice, at least as far as computational burden is concerned. The results are presented below in the form of a theorem, followed by an application to an aggregate production function.

THEOREM 4.1: Full Conditionals for the Heteroskedastic Linear Model
In the linear regression model with Normally distributed, heteroskedastic error and K independent variables in the observed equation:

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i, \qquad \varepsilon_i \sim \mathrm{N}\left(0, \tfrac{1}{\tau_i}\right)$$

and L independent variables in the precision equation:

$$\log \tau_i = \mathbf{w}_i'\boldsymbol{\delta} + v_i, \qquad v_i \sim \mathrm{N}\left(0, \tfrac{1}{\phi}\right)$$

and with Normal priors for $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ and a Gamma prior for ϕ:

$$\mathrm{p}(\boldsymbol{\beta}) = \frac{|\mathbf{P}_\beta|^{1/2}}{(2\pi)^{K/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}-\mathbf{m}_\beta)'\mathbf{P}_\beta(\boldsymbol{\beta}-\mathbf{m}_\beta)\right\},$$

$$\mathrm{p}(\boldsymbol{\delta}) = \frac{|\mathbf{P}_\delta|^{1/2}}{(2\pi)^{L/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{\delta}-\mathbf{m}_\delta)'\mathbf{P}_\delta(\boldsymbol{\delta}-\mathbf{m}_\delta)\right\} \qquad \text{and} \qquad \mathrm{p}(\phi) = \frac{b^a}{\Gamma(a)}\phi^{a-1}e^{-b\phi}$$

the full conditionals of $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ are Normal:

$$\pi(\boldsymbol{\beta}|\bullet) = \frac{|\tilde{\mathbf{P}}_\beta|^{1/2}}{(2\pi)^{K/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}-\tilde{\mathbf{m}}_\beta)'\tilde{\mathbf{P}}_\beta(\boldsymbol{\beta}-\tilde{\mathbf{m}}_\beta)\right\}$$

$$\pi(\boldsymbol{\delta}|\bullet) = \frac{|\tilde{\mathbf{P}}_\delta|^{1/2}}{(2\pi)^{L/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{\delta}-\tilde{\mathbf{m}}_\delta)'\tilde{\mathbf{P}}_\delta(\boldsymbol{\delta}-\tilde{\mathbf{m}}_\delta)\right\}$$

and the full conditional of ϕ is Gamma:

$$\pi(\phi|\bullet) = \frac{\tilde{b}^{\tilde{a}}}{\Gamma(\tilde{a})}\phi^{\tilde{a}-1}e^{-\tilde{b}\phi}$$

where:

- $\tilde{\mathbf{P}}_\beta = \sum_{i=1}^{N} \tau_i \mathbf{x}_i \mathbf{x}_i' + \mathbf{P}_\beta$ and $\tilde{\mathbf{m}}_\beta = \left(\sum_{i=1}^{N} \tau_i \mathbf{x}_i \mathbf{x}_i' + \mathbf{P}_\beta \right)^{-1} \left(\sum_{i=1}^{N} \tau_i \mathbf{x}_i y_i + \mathbf{P}_\beta \mathbf{m}_\beta \right)$
- $\tilde{\mathbf{P}}_\delta = \phi \mathbf{W}'\mathbf{W} + \mathbf{P}_\delta$ and $\tilde{\mathbf{m}}_\delta = (\phi \mathbf{W}'\mathbf{W} + \mathbf{P}_\delta)^{-1} (\phi \mathbf{W}'\mathbf{z}^* + \mathbf{P}_\delta \mathbf{m}_\delta)$
- $\tilde{a} = \frac{N}{2} + a$ and $\tilde{b} = \frac{1}{2} (\mathbf{z}^* - \mathbf{W}\boldsymbol{\delta})' (\mathbf{z}^* - \mathbf{Z}\boldsymbol{\delta}) + b$
- $\mathbf{z}^*$ is the $N \times 1$ vector obtained by stacking the $\log \tau_i$s

The full conditional of τ_i, $i = 1, 2, \ldots, N$, is:

$$\pi(\tau_i | \bullet) \propto \tau_i^{-1/2} \exp\left\{ -\frac{\tau_i (y_i - \mathbf{x}_i'\boldsymbol{\beta})^2}{2} - \frac{\phi (\log \tau_i - \mathbf{w}_i'\boldsymbol{\delta})^2}{2} \right\}$$

♦ **Example 4.1 Aggregate Production with Heteroskedasticity**

In this example we will use again the data from the Penn World Table (Feenstra et al., 2015), which we first used in Example 2.2 to estimate an aggregate production function. We will assume here that the production function is Cobb-Douglas and we will add a time trend to capture technological progress:

$$\log \mathsf{Y}_i = \beta_1 + \beta_2 \log \mathsf{K}_i + \beta_3 \log \mathsf{L}_i + \beta_4 \mathsf{trend}_i + \varepsilon_i$$

Apart from the homoskedastic model we estimated in Example 2.2, we will consider two models with heteroskedastic error, one in which the precision of the error term follows a log-Normal distribution with common location parameter, γ:

$$\log \tau_i = \gamma + v_i, \qquad v_i \sim \mathsf{N}\left(0, \tfrac{1}{\phi}\right)$$

and an extension in which the location parameter is a function of the logarithms of the two inputs and time:

$$\log \tau_i = \gamma_1 + \gamma_2 \log \mathsf{K}_i + \gamma_3 \log \mathsf{L}_i + \gamma_4 \mathsf{trend}_i + v_i, \qquad v_i \sim \mathsf{N}\left(0, \tfrac{1}{\phi}\right)$$

For ease of comparison, the results from the model with homoskedastic error are reproduced in the next table. The following two tables contain the results from the models with heteroskedastic error.

	Mean	Median	Sd.dev.	5%	95%
constant	3.24514	3.24896	0.279077	2.784	3.69835
logK	0.583076	0.582767	0.0238982	0.544447	0.622634
logL	0.441425	0.441675	0.0225753	0.403951	0.477781
trend	0.00119668	0.00120461	0.000508724	0.000353659	0.0020245
tau	72.1455	72.0801	3.9129	65.778	78.7437
sigma_e	0.117862	0.117787	0.00320426	0.112692	0.123302

	Mean	Median	Sd.dev.	5%	95%
logY					
constant	2.60779	2.60556	0.236055	2.22318	3.00122
logK	0.638856	0.639117	0.0203548	0.604912	0.672072
logL	0.390394	0.390249	0.0197409	0.358244	0.423338
trend	-0.000753112	-0.000761663	0.000426506	-0.00142877	-3.1367e-05
logtau					
constant	4.70099	4.7012	0.099249	4.53613	4.86476
phi	1.00228	0.943447	0.308244	0.625406	1.615
sigma_v	1.02992	1.02956	0.1422	0.786939	1.26465

The first thing to notice from these results is that the parameters of the production function change slightly when moving from the homoskedastic to the heteroskedastic models, as well as from the first heteroskedastic model to the second. An interesting pattern appears in the results of the heteroskedastic model with observation-specific location parameter for τ_i: as the amount of capital employed in the production process increases, the precision of the error term increases as well, while the opposite tendency appears for the amount of labor (although the 90% credible interval for the associated parameter contains zero). This could be due to, for example, the standardization of production in capital-intensive processes, leading to smaller margins of error. On the other, the precision of the error term decreases over time.

	Mean	Median	Sd.dev.	5%	95%
logY					
constant	2.09591	2.09506	0.193072	1.78358	2.41586
logK	0.683124	0.683164	0.0165634	0.655633	0.709897
logL	0.34707	0.346883	0.0162974	0.320587	0.374088
trend	-0.00147632	-0.00148201	0.000370967	-0.00208157	-0.000866163
logtau					
constant	-6.82863	-6.83205	4.93942	-14.8837	1.29765
logK	0.901962	0.903622	0.422066	0.209257	1.59166
logL	-0.454889	-0.456639	0.387832	-1.08508	0.184916
trend	-0.0176653	-0.0176734	0.00897148	-0.0324105	-0.00287653
phi	1.04366	0.992658	0.277183	0.71938	1.49932
sigma_v	1.00025	1.00372	0.114119	0.816719	1.17906

After estimating the three models, we can compare them using Bayes factors. The following two tables present model-comparison results based on the Lewis and Raftery and the Chib and Jeliazkov approximations to the logarithm of the marginal likelihood, respectively. With equal prior model probabilities, the data clearly favor the second heteroskedastic model.

Model	Log-Marginal Likelihood	Type of log-ML Approximation	Prior Model Probability	Posterior Model Probability
homosked	442.739	Lewis & Raftery	0.333333	2.25774e-12
heterosked1	462.762	Lewis & Raftery	0.333333	0.00112074
heterosked2	469.554	Lewis & Raftery	0.333333	0.998879

Model	Log-Marginal Likelihood	Type of log-ML Approximation	Prior Model Probability	Posterior Model Probability
homosked	442.746	Chib & Jeliazkov	0.333333	2.55552e-12
heterosked1	462.417	Chib & Jeliazkov	0.333333	0.000892387
heterosked2	469.438	Chib & Jeliazkov	0.333333	0.999108

Obtaining the results presented above using BayES can be achieved using the code in the following box.

```
// import the data and transform the variables
Data = webimport("www.bayeconsoft.com/datasets/PWT.csv");
Data.constant = 1;          Data.logY = log(Data.Y);
Data.logK = log(Data.K);    Data.logL = log(Data.L);

// run a homoskedastic Cobb-Douglas model
homosked = lm(logY ~ constant logK logL trend,
    "logML_CJ"=true);

// run a simple heteroskedastic Cobb-Douglas model
heterosked1 = lm(logY ~ constant logK logL trend | constant,
    "logML_CJ"=true);

// run a heteroskedastic Cobb-Douglas model with determinants in log-tau
heterosked2 = lm(logY ~ constant logK logL trend | constant logK logL trend,
    "logML_CJ"=true);

// compare the three models
pmp( { homosked, heterosked1, heterosked2 } );
pmp( { homosked, heterosked1, heterosked2 }, "logML_CJ"=true );
```

4.3.2 The Stochastic Frontier Model

The *stochastic-frontier* model was introduced independently by Meeusen & van den Broeck (1977) and Aigner et al. (1977) as a way of estimating the parameters of a production function, while recognizing that producers may not be exploiting the full potential of the production

technology, in the sense that they are not producing the maximum possible output, given the amount of inputs they are using. Jondrow et al. (1982) proposed an approach to estimating producer-specific *technical-efficiency* scores after the estimation of the model and this development lead to a surge of applications of the model, whose primary objective was not the estimation of production function, but the benchmarking of producers based on their technical efficiency. All three papers use frequentist methods and the first Bayesian treatment of the model appeared almost two decades after its introduction (van den Broeck et al., 1994).

The specification of the model starts by representing the production technology as a function of production factors, $f(\boldsymbol{x})$, while explicitly recognizing that this production function returns the maximum possible output, y, that can be produced given $\boldsymbol{x}$ and that this maximum output may not always be attained by producers. The next step is to define the technical efficiency of a potential observation, i:

$$\mathrm{TE}_i = \frac{y_i}{f(\boldsymbol{x}_i)} \tag{4.10}$$

Technical efficiency is the ratio of observed output, y_i, over maximum possible output, $f(\boldsymbol{x}_i)$, and, as such, it assumes values on the unit interval. By taking the logarithm of both sides of this expression and rearranging, the estimable form of a production frontier becomes:

$$\log y_i = \log f(\boldsymbol{x}_i) - u_i + v_i \tag{4.11}$$

where $u_i \equiv -\log \mathrm{TE}_i$. Because $\mathrm{TE}_i \in (0,1]$, u_i is non-negative and during estimation it is treated as an one-sided error term. v_i on the other hand, is the typical error term in stochastic models, which captures statistical noise.

To proceed with estimation one needs to specify the functional form of $f(\boldsymbol{x})$, as well as a distribution for u_i. Typically, a Cobb-Douglas or translog form is assumed for the production function, leading to a model where the logarithm of output is a linear function of parameters and the logarithms of inputs and possibly their interactions. The distributional assumption imposed on u_i is not as straightforward. Meeusen & van den Broeck (1977) assumed that u_i follows an Exponential distribution, while Aigner et al. (1977) used a half-Normal distribution. Many more distributions with support on the interval $[0,+\infty)$ have since been proposed, giving rise to alternative stochastic-frontier models. We will provide here an extensive treatment of the Exponential model and compare it only to the half-Normal model, before demonstrating the use of these two models in an application, at the end of this subsection.

To simplify notation, let y_i denote the logarithm of output for a potential observation, i, $\log y_i$, and let $\mathbf{x}_i$ denote the $K \times 1$ vector of values of the independent variables that enter the specification of the production function, for the same i. If the production function is Cobb-Douglas then $\mathbf{x}_i$ is simply equal to $\log \boldsymbol{x}_i$. If, on the other hand, the production function is translog then $\mathbf{x}_i$ will contain the logarithms of inputs, as well as their squared terms and their interactions. With these definitions, the statistical model becomes:

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} - u_i + v_i, \qquad v_i \sim \mathrm{N}\left(0, \tfrac{1}{\tau}\right), \qquad u_i \sim \mathrm{Exp}(\lambda) \tag{4.12}$$

The parameters of the model are $\boldsymbol{\beta}$, τ and λ. The u_is are unknown and, in a data-augmentation setting, will be treated as latent data. Notice, however, that the u_is now have an interesting interpretation: because we defined u_i as $-\log \mathrm{TE}_i$, the technical efficiency score of a potential observation i can be estimated by inverting this relationship. In every iteration of the Gibbs sampler, random draws from the posterior distribution of each u_i will be produced and, if these draws are stored in memory, then a point estimate of TE_i can be obtained by calculating the sample mean of e^{-u_i} across these draws.

The complete-data likelihood for the model is:

$$\begin{aligned} \mathrm{p}(\mathbf{y}, \mathbf{u} | \mathbf{X}, \boldsymbol{\beta}, \tau, \lambda) &= \prod_{i=1}^{N} \mathrm{p}(y_i | \mathbf{x}_i, u_i, \boldsymbol{\beta}, \tau)\, \mathrm{p}(u_i | \lambda) \\ &= \prod_{i=1}^{N} \frac{\tau^{1/2}}{(2\pi)^{1/2}} \exp\left\{ -\frac{\tau (y_i - \mathbf{x}_i'\boldsymbol{\beta} + u_i)^2}{2} \right\} \times \prod_{i=1}^{N} \lambda e^{-\lambda u_i} \end{aligned} \tag{4.13}$$

where, as before, $\mathbf{y}$ and $\mathbf{X}$ are the vector and matrix of the stacked values for the dependent and independent variables, respectively, and $\mathbf{u}$ is the $N \times 1$ vector of stacked values for the u_is. The first equality above expresses the joint density of observed and latent data as the product of a conditional and a marginal density, while the second equality comes from the distributional assumptions of v_i and u_i and the conditional independence of these two error components.

We will keep using a multivariate-Normal prior for $\boldsymbol{\beta}$, with mean $\mathbf{m}$ and precision matrix $\mathbf{P}$ and a Gamma prior for τ, with shape and rate parameters a_τ and b_τ, respectively. The values of the hyperparameters can be set such that these priors are very vague. For λ, however, we need to use a more informative prior and we follow van den Broeck et al. (1994) in using a Gamma prior for it, with shape parameter, a_λ, equal to one and rate parameter, b_λ, equal to $-\log r^\star$, where $r^\star$ is the prior median efficiency. With the likelihood function and the prior densities at hand, we can express the posterior density of the parameters and the latent data as:

$$\pi\left(\boldsymbol{\beta}, \tau, \lambda, \mathbf{u} \middle| \mathbf{y}, \mathbf{X}\right) \propto \mathrm{p}\left(\mathbf{y}, \mathbf{u} \middle| \mathbf{X}, \boldsymbol{\beta}, \tau, \lambda\right) \mathrm{p}(\boldsymbol{\beta})\, \mathrm{p}(\tau)\, \mathrm{p}(\lambda) \tag{4.14}$$

where the first density in the right-hand side of the last expression is given in (4.13) and the the last three densities are the prior densities for the three blocks of parameters. By defining $\mathbf{y}^*$ as $\mathbf{y} - \mathbf{u}$, it is easy to see that the full conditionals of $\boldsymbol{\beta}$ and τ are the same as the ones presented in Theorem 2.1, with $\mathbf{y}^*$ assuming the role of $\mathbf{y}$. The full conditional of λ is Gamma, with shape parameter equal $N + a_\lambda$ and rate parameter $\sum_i u_i + b_\lambda$, while the transformations required to obtain this full conditional are exactly the same as the ones performed in Example 1.2, with u_i assuming the role of the data. It requires some additional work to show that the full conditional of every u_i is Normal, truncated from below at zero. Once again, we present the full conditionals of the Exponential stochastic-frontier model in the form of a theorem, for ease of reference.

THEOREM 4.2: Full Conditionals for the Exponential Stochastic-Frontier Model

In the stochastic-frontier model with Normally-distributed noise and Exponentially-distributed inefficiency component of the error term and with K independent variables:

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} - u_i + v_i, \qquad v_i \sim \mathrm{N}\left(0, \tfrac{1}{\tau}\right), \qquad u_i \sim \mathrm{Exp}\left(\lambda\right)$$

and with Normal prior for $\boldsymbol{\beta}$ and Gamma priors for τ and λ:

$$\mathrm{p}(\boldsymbol{\beta}) = \frac{|\mathbf{P}|^{1/2}}{(2\pi)^{K/2}} \exp\left\{-\frac{1}{2}\left(\boldsymbol{\beta} - \mathbf{m}\right)' \mathbf{P}\left(\boldsymbol{\beta} - \mathbf{m}\right)\right\},$$

$$\mathrm{p}(\tau) = \frac{b_\tau^{a_\tau}}{\Gamma\left(a_\tau\right)} \tau^{a_\tau - 1} e^{-b_\tau \tau} \qquad \text{and} \qquad \mathrm{p}(\lambda) = \frac{b_\lambda^{a_\lambda}}{\Gamma\left(a_\lambda\right)} \lambda^{a_\lambda - 1} e^{-b_\lambda \lambda}$$

the full conditional of $\boldsymbol{\beta}$ is Normal:

$$\pi\left(\boldsymbol{\beta} | \bullet\right) = \frac{|\tilde{\mathbf{P}}|^{1/2}}{(2\pi)^{K/2}} \exp\left\{-\frac{1}{2}\left(\boldsymbol{\beta} - \tilde{\mathbf{m}}\right)' \tilde{\mathbf{P}}\left(\boldsymbol{\beta} - \tilde{\mathbf{m}}\right)\right\}$$

and the full conditionals of τ and λ are Gamma:

$$\pi\left(\tau | \bullet\right) = \frac{\tilde{b}_\tau^{\tilde{a}_\tau}}{\Gamma\left(\tilde{a}_\tau\right)} \tau^{\tilde{a}_\tau - 1} e^{-\tilde{b}_\tau \tau} \qquad \text{and} \qquad \pi\left(\lambda | \bullet\right) = \frac{\tilde{b}_\lambda^{\tilde{a}_\lambda}}{\Gamma\left(\tilde{a}_\lambda\right)} \lambda^{\tilde{a}_\lambda - 1} e^{-\tilde{b}_\lambda \lambda}$$

where:

- $\tilde{\mathbf{P}} = \mathbf{X}'\mathbf{X} + \mathbf{P}$ and $\tilde{\mathbf{m}} = \left(\mathbf{X}'\mathbf{X} + \mathbf{P}\right)^{-1}\left(\mathbf{X}'\mathbf{y}^* + \mathbf{P}\mathbf{m}\right)$
- $\tilde{a}_\tau = \frac{N}{2} + a_\tau$ and $\tilde{b}_\tau = \frac{1}{2}\left(\mathbf{y}^* - \mathbf{X}\boldsymbol{\beta}\right)'\left(\mathbf{y}^* - \mathbf{X}\boldsymbol{\beta}\right) + b_\tau$

- $\tilde{a}_\lambda = N + a_\lambda$ and $\tilde{b}_\lambda = \sum_i u_i + b_\lambda$
- $\mathbf{y}^* = \mathbf{y} + \mathbf{u}$

The full conditional of u_i, $i = 1, 2, \ldots, N$, is Normal, truncated from below at zero:

$$\pi\left(u_i|\bullet\right) = \frac{\tau^{1/2}}{(2\pi)^{1/2}\Phi\left(\tau^{1/2}\mu_i\right)} \exp\left\{-\frac{\tau}{2}\left(u_i - \mu_i\right)^2\right\} \mathbb{1}(u_i \geq 0)$$

where $\mu_i = -\left(y_i - \mathbf{x}_i'\boldsymbol{\beta}\right) - \frac{\lambda}{\tau}$.

Before we proceed to an application, we note that the derivation of the full conditionals of the parameter blocks in stochastic-frontier models with different specifications of the distribution of u_i are exactly the same for all blocks except, obviously, for u_i itself and any parameters that enter the assumed density of u_i. For example, in the half-Normal stochastic-frontier model it is assumed that $u_i \sim \mathrm{N}^+\left(0, \frac{1}{\phi}\right)$. The full conditionals of $\boldsymbol{\beta}$ and τ are exactly the same as in the Exponential stochastic-frontier model. A Gamma prior for ϕ is conjugate and the full conditional of ϕ is Gamma with shape and rate parameters, $\frac{N}{2} + a_\phi$ and $\frac{\mathbf{u}'\mathbf{u}}{2} + b_\phi$, respectively. Finally, the full conditional of u_i is Normal, truncated from below at zero, but with different location and scale parameters:

$$\pi\left(u_i|\bullet\right) = \frac{(\tau+\phi)^{1/2}}{(2\pi)^{1/2}\Phi\left((\tau+\phi)^{1/2}\mu_i\right)} \exp\left\{-\frac{\tau+\phi}{2}\left(u_i - \mu_i\right)^2\right\} \mathbb{1}(u_i \geq 0)$$

where $\mu_i = -\frac{\tau}{\tau+\phi}\left(y_i - \mathbf{x}_i'\boldsymbol{\beta}\right)$.

♦ **Example 4.2 US Electric Utilities**

In this example we will use part of the dataset constructed and first used by Rungsuriyawiboon & Stefanou (2007). This version of the dataset contains information on 81 US investor-owned electric utilities, each one of them observed annually from 1986 to 1997, on the following variables:

`q`	:	megawatt hours of electric power generated
`K`	:	real capital stock at replacement cost
`L`	:	deflated value of the cost of labor and maintenance
`F`	:	deflated value of the cost of fuel used for power generation
`trend`	:	a trend variable running from -6 to 5

To concentrate on the inefficiency part of the model, we will assume that the production function is Cobb-Douglas in the three inputs and we will add a time trend to capture technological progress:

$$\log \mathsf{q}_i = \beta_1 + \beta_2 \log \mathsf{K}_i + \beta_3 \log \mathsf{L}_i + \beta_4 \log \mathsf{F}_i + \beta_5 \mathsf{trend}_i - u_i + v_i$$

We will first run a simple linear model, which disregards any inefficiency in production (u_i=0), to serve as a benchmark for comparisons. The results from this model appear in the following table.

	Mean	Median	Sd.dev.	5%	95%
constant	5.11508	5.11499	0.130217	4.90232	5.32902
logK	0.179785	0.179915	0.0234004	0.141448	0.218299
logL	0.196873	0.197032	0.0191965	0.165351	0.228522
logF	0.620872	0.620971	0.023638	0.58204	0.659569
trend	0.0151138	0.0151111	0.00251212	0.0109742	0.0192037
tau	14.584	14.5772	0.660596	13.5131	15.6864
sigma_e	0.262057	0.261917	0.00595144	0.252489	0.272037

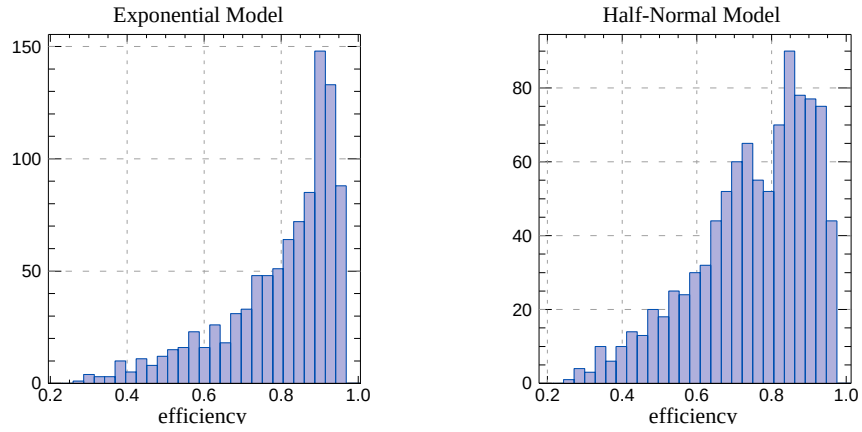
We next consider two stochastic-frontier models, one where u_i is assumed to follow an Exponential distribution and one where it follows a half-Normal distribution. The results from these two models are given in the following tables.

	Mean	Median	Sd.dev.	5%	95%
constant	5.64397	5.64403	0.094784	5.48794	5.79964
logK	0.184463	0.184559	0.0204281	0.150623	0.217666
logL	0.235713	0.235865	0.0140363	0.212521	0.258588
logF	0.56514	0.565127	0.0217979	0.529248	0.600886
trend	0.0215448	0.0215612	0.00185282	0.018475	0.02456
tau	84.3254	83.181	12.3741	65.9133	106.366
lambda	4.01486	4.00855	0.197964	3.7016	4.35148
sigma_v	0.109762	0.109645	0.00794179	0.096969	0.123176
sigma_u	0.249678	0.249468	0.0122746	0.229811	0.270172

	Mean	Median	Sd.dev.	5%	95%
constant	5.80095	5.80058	0.102632	5.6305	5.96848
logK	0.169237	0.169645	0.0230587	0.130457	0.206874
logL	0.236839	0.236838	0.0128786	0.215705	0.258086
logF	0.574338	0.573809	0.0231652	0.537095	0.613029
trend	0.0212594	0.0212574	0.0018476	0.0182299	0.0243108
tau	160.573	156.645	32.2774	114.452	219.885
phi	6.17676	6.16406	0.387879	5.56205	6.8369
sigma_v	0.0800751	0.0798994	0.00785511	0.0674414	0.0934788
sigma_u	0.402958	0.402779	0.0126301	0.382448	0.424018

The estimates of the parameters of the production function are very similar in the two stochastic-frontier models, but are slightly different from the ones obtained from the simple linear model. Furthermore, the inclusion of the inefficiency term in the stochastic-frontier models leads to a dramatic increase in the precision of the noise component of the error term. This is to be expected: by allowing for inefficiency in production, what was treated by the linear model as noise is separated by the stochastic-frontier models into noise and inefficiency.

After the estimation of a stochastic-frontier model we can get observation-specific estimates of the efficiency scores. These are obtained by summarizing the exponential of minus the draws from the full conditionals of the u_is, which are generated when running the Gibbs sampler. The histograms of the efficiency-score estimates from the two stochastic-frontier models appear in the following figure. The distributions of efficiency scores from both models have the typical long tail to the left (fewer firms are more inefficient), but are quite different from each other.



The relative plausibility of each model given the data can be assessed using Bayes factors. Based on the Lewis and Raftery approximation of the log-marginal likelihood, we conclude that the data strongly favor the Exponential stochastic-frontier model over the simple linear model, as well as the half-Normal model (see the following table).

Model	Log-Marginal Likelihood	Type of log-ML Approximation	Prior Model Probability	Posterior Model Probability
LM	-124.249	Lewis & Raftery	0.333333	3.23169e-44
SF_Exp	-24.1139	Lewis & Raftery	0.333333	0.994596
SF_hNorm	-29.3292	Lewis & Raftery	0.333333	0.00540351

The results presented above can be obtained in BayES using the code in the following box.

```
// import the data and transform the variables
Data = webimport("www.bayeconsoft.com/datasets/USUtilities.csv");
Data.constant = 1;
Data.logq = log(Data.q);    Data.logK = log(Data.K);
Data.logL = log(Data.L);    Data.logF = log(Data.F);

// run a linear model
LM = lm(logq ~ constant logK logL logF trend);

// run an Exponential stochastic-frontier model
SF_Exp = sf(logq ~ constant logK logL logF trend, "udist"="exp");

// run a half-Normal stochastic-frontier model
SF_hNorm = sf(logq ~ constant logK logL logF trend, "udist"="hnorm");

// store the efficiency scores as variables in the dataset
store( eff_i, eff_Exp, "model"=SF_Exp );
store( eff_i, eff_hNorm, "model"=SF_hNorm );

// plot histograms of the efficiency scores from the two models
myFigure = multiplot(1,2);
hist( subplot(myFigure, 1, 1), Data.eff_Exp,
    "title" = "Exponential Model", "xlabel" = "efficiency", "grid"="on" );
hist( subplot(myFigure, 1, 2), Data.eff_hNorm,
    "title" = "Half-Normal Model", "xlabel" = "efficiency", "grid"="on" );

// compare the three models
pmp( { LM, SF_Exp, SF_hNorm } );
```

4.4 Marginal Data Augmentation

By artificially introducing latent data into a complex model, data augmentation can vastly simplify the implementation of a sampling algorithm. However, the resulting simplifications often induce high autocorrelation in the draws obtained from such an algorithm. Meng & van Dyk (1999) and J. Liu & Wu (1999) independently proposed a technique designed to speed-up the convergence rate and mixing properties of an MCMC algortithm. Both proposals represent direct extensions of previous work conducted in the context of the EM algorithm (Meng & van Dyk, 1997; C. Liu et al., 1998), but use sightly different terminology: in Meng & van Dyk's terminology the technique is called *marginal data augmentation*, while J. Liu & Wu use the term *parameter-expanded data augmentation*.

The technique works by introducing yet another artificial quantity, α, into the problem, called the *working* (Meng & van Dyk's terminology) or *expansion parameter* (J. Liu & Wu's terminology). This parameter is then integrated-out from the complete-data likelihood via simulation. Contrary to the model's parameters, however, α has the unique property of being identified only from the complete data, leaving the observed-data likelihood unaffected by conditioning. Using the notation introduced in section 4.2, this property is expressed mathematically as:

$$\int_{Z} \mathrm{p}(\mathbf{y}, \mathbf{z}|\boldsymbol{\theta}, \alpha)\ \mathrm{d}\mathbf{z} = \mathrm{p}(\mathbf{y}|\boldsymbol{\theta}) \tag{4.15}$$

This expression makes clear that, by ignoring the presence of a working parameter in standard data augmentation, we are effectively conditioning on a particular value of α. Instead of

constructing a sampler conditional on α, we could marginalize the working parameter by multiplying both sides by a prior density, $\mathrm{p}(\alpha|\boldsymbol{\theta})$, and then integrating over α. This procedure leads, after a change in the order of integration on the left-had side, to:

$$\int_Z \left[\int_A \mathrm{p}(\mathbf{y}, \mathbf{z}|\boldsymbol{\theta}, \alpha)\, \mathrm{p}(\alpha|\boldsymbol{\theta})\, \mathrm{d}\alpha \right] \mathrm{d}\mathbf{z} = \mathrm{p}(\mathbf{y}|\boldsymbol{\theta}) \tag{4.16}$$

If we carry out the integration inside the square brackets analytically, we obtain $\mathrm{p}(\mathbf{y}, \mathbf{z}|\boldsymbol{\theta})$ and, thus, revert back to standard data augmentation, where we iteratively sample from $\mathrm{p}(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta})$ and $\mathrm{p}(\boldsymbol{\theta}|\mathbf{y}, \mathbf{z})$. As Meng & van Dyk (1999) explain, the key to the computational advantage of marginalizing the working parameter over conditioning upon it is that the model based on $\int_A \mathrm{p}(\mathbf{y}, \mathbf{z}|\boldsymbol{\theta}, \alpha)\, \mathrm{d}\alpha$ is likely more diffuse than the one based on $\mathrm{p}(\mathbf{y}, \mathbf{z}|\boldsymbol{\theta}, \alpha)$. This is desirable because, in a standard data augmentation setting, we could achieve zero autocorrelation in the draws if we could sample iteratively from $\mathrm{p}(\mathbf{z}|\mathbf{y})$ and $\mathrm{p}(\boldsymbol{\theta}|\mathbf{y}, \mathbf{z})$. Thus, we should aim at having as diffuse a $\mathrm{p}(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta})$ as possible, up to the limit of $\mathrm{p}(\mathbf{z}|\mathbf{y})$. But, because:

$$\mathrm{p}(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}) = \frac{\mathrm{p}(\mathbf{y}, \mathbf{z}|\boldsymbol{\theta})}{\mathrm{p}(\mathbf{y}|\boldsymbol{\theta})} \tag{4.17}$$

and the denominator is unaffected by the introduction of α, inducing a more diffuse numerator would make $\mathrm{p}(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta})$ more diffuse as well. This line of argument implies that the prior on the working parameter should also be diffuse, although making it improper could alter the properties of the sampler.[2]

Although marginal data augmentation may work in the general setting presented above, Meng & van Dyk (1999) showed formally that if the prior imposed on the working parameter is proper and independent of the identifiable parameters, $\boldsymbol{\theta}$, then marginal data augmentation can only improve the geometric rate of convergence of the sampler. Using slightly different notation than Meng & van Dyk, the procedure starts by defining a one-to-one and differentiable mapping in the space of the latent data, $\mathcal{D}_a(\mathbf{w})$. This mapping is indexed by the working parameter and, according to Meng & van Dyk (1997), prominent choices are:

- rescaling: $\mathcal{D}_a(\mathbf{z}) = \alpha\mathbf{z}$
- recentering: $\mathcal{D}_a(\mathbf{z}) = \alpha + \mathbf{z}$
- affine transformations: $\mathcal{D}_a(\mathbf{z}) = \alpha_1 + \alpha_2\mathbf{z}$

Let $\mathbf{w}$ denote the transformed latent data that result from applying $\mathcal{D}_a(\bullet)$ on $\mathbf{z}$. Finally, let $\mathrm{p}(\alpha)$ denote the prior for the working parameter. The Gibbs sampler then iterates between the steps:

(a) draw $\mathbf{w}$ from $\mathrm{p}(\mathbf{w}|\mathbf{y}, \boldsymbol{\theta})$ by sampling for $(\mathbf{w}, \alpha)$ and discarding the draw on α; in most cases this step is further broken into the steps:

  (a1) draw α from $\mathrm{p}(\alpha)$
  (a2) draw $\mathbf{w}$ from $\mathrm{p}(\mathbf{w}|\mathbf{y}, \boldsymbol{\theta}, \alpha)$

(b) draw $\boldsymbol{\theta}$ from $\mathrm{p}(\boldsymbol{\theta}|\mathbf{w}, \mathbf{y})$ by sampling for $(\boldsymbol{\theta}, \alpha)$ and discarding the draw on α; in most cases this step is further broken into the steps:

  (b1) draw α from $\mathrm{p}(\alpha|\mathbf{y}, \mathbf{w})$
  (b2) draw $\boldsymbol{\theta}$ from $\mathrm{p}(\boldsymbol{\theta}|\mathbf{y}, \mathbf{w}, \alpha)$

The working parameter in this procedure is marginalized in both main steps, but it is possible to implement a sampler where α is updated in step **(b)** and this value used in step **(a2)**, thus skipping step **(a1)**. More possibilities regarding marginalization arise when $\boldsymbol{\theta}$ is

[2] When this is not the case, improper priors on the working parameter are to be preferred. See, for example, J. Liu & Wu (1999).

broken into multiple blocks, but the procedures are case specific and we will not go into the details of any particular problem here.

Before closing this section we mention in passing that there are two additional potential uses of marginal data augmentation:

1. Because the working parameter can be identified only from the complete data, marginal data augmentation can be used to impose restrictions in models which require such restrictions for identification, without the use of "exotic" priors. We will see such an application of the technique in Section 6.6, as well as in Chapter 7.
2. Although marginal data augmentation was developed in the context of latent-data models, the technique can also be used to improve mixing in situations without any missing data. van Dyk (2010) presents such an example, where the working parameter effectively induces a re-parameterization of the original problem.

4.5 Synopsis

This chapter introduced and covered in detail the technique of data augmentation. Data augmentation artificially introduces latent data into complex models such that they become amenable to statistical analysis. It is a very powerful technique and gives Bayesian methods a clear advantage over frequentist analysis of very complex models: as model complexity increases, the Bayesian approach introduces latent data and, in most cases, casts the model into a linear regression one. The latent data are subsequently integrated-out from the posterior density, such that inferences about the model's parameters can be drawn. This chapter also covered two simple applications of data augmentation: the linear regression model with heteroskedastic error and the stochastic-frontier model. Data augmentation will be used extensively in the following chapters, even when discussing models whose parameters can be estimated without the introduction of latent data. This is because the technique simplifies the analysis considerably, both from an analytical and a computational perspective, and we can, thus, build on the results obtained in the context of the linear regression model.

Chapter 5

The Linear Model with Panel Data

5.1 Overview

This chapter extends the linear regression model such that it can accommodate panel data. The availability of panel data opens up an array of possibilities for flexible modeling of the phenomenon of interest, as it allows for controlling for any group-invariant factors that may affect the dependent variable(s). Although we will use the frequentist terms "fixed effects", "random effects" and "random coefficients" to describe the alternative panel-data models, we do so while recognizing that the terms themselves may appear as conveying information that they should not. In fact, the use of these terms is rather controversial in a Bayesian setting because parameters or "effects" are always random in this context. As McCulloch & Rossi (1994) put it, "in the Bayesian point of view, there is no distinction between fixed and random effects, only between hierarchical and non-hierarchical models".

The following section defines what a panel dataset is and discusses the assumptions behind the alternative panel-data models. Although we will almost exclusively use the group-time definition of panel data, it is mentioned in passing that the models can be applied also in contexts where there is a natural classification of observations in groups and where no time dimension appears in the data. Estimation procedures for the alternative models are described using data augmentation, while multiple examples are used to illustrate their application.

The techniques presented here can be applied to models other than the linear regression one. However, the chapter focuses on the linear regression model, as this is arguably the simplest practical model that is extensively used in the econometrics literature. Subsequent chapters will frequently contain subsections that discuss extensions of the models presented therein to cases where panel data are available.

5.2 Panel Data and Alternative Panel-Data Models

A *panel dataset* is a collection of observations on a set of random variables for a number of *groups*, each one of which is observed over multiple time periods. Some examples of panel datasets are the following:

- expenditure on different categories of goods for N households is observed monthly, over a period of T months
- input and output quantities are observed annually for N firms and over a period of T years

- Gross Domestic Product (GDP) per capita, the aggregate savings rate, the population growth rate and the rate of technological progress are observed annually for N countries and over a period of T years

The unit of analysis in these three examples is, respectively, the household, the firm and the country, and the term 'group' will be used to refer to this unit. Time represents the second dimension of the panel. Typically, panel datasets consist of many groups (large N) and few time observations per group (small T), although this is not always the case. Panels for which all groups are observed for the same number of time periods are called *balanced*, while when the number of time observations varies by group the panel is said to be *unbalanced*.

Because a panel dataset has two dimensions, we will use a double subscript to refer to a potential observation. For example, y_{it} will be the value of a random variable for a potential observation for group i and in period t. With this notation, the linear regression model with panel data and independent Normally-distributed errors becomes:

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}, \qquad \varepsilon_{it} \sim \mathrm{N}\left(0, \tfrac{1}{\tau}\right) \tag{5.1}$$

Of course, this slight change in notation does not invalidate the procedure of estimating the parameters of the model which was covered in Chapter 2. That is, one can disregard the panel nature of the dataset and estimate the parameters by pooling together the time observations across multiple groups. Such an approach amounts to estimating what is called the *pooled model*. The real usefulness of panel data, however, comes from the possibilities it presents for controlling for group-specific *unobserved heterogeneity*. To proceed with this point, suppose that the phenomenon under question involves a stochastic model where the dependent variable is determined by a set of time varying independent variables, $\mathbf{x}$, as well as a set of time-invariant independent variables, $\mathbf{w}$:

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\gamma} + \varepsilon_{it}, \qquad \varepsilon_{it} \sim \mathrm{N}\left(0, \tfrac{1}{\tau}\right) \tag{5.2}$$

where $\boldsymbol{\beta}$ is a $K \times 1$ of parameters associated with the time-varying variables and $\boldsymbol{\gamma}$ is a vector of parameters associated with the time-invariant variables.[1] If both $\mathbf{x}$ and $\mathbf{w}$ are observed, then the pooled model can produce estimates of both $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. With the availability of panel data, estimates of $\boldsymbol{\beta}$ can be obtained even if the time invariant variables are unobserved. By defining $\alpha_i \equiv \mathbf{w}_i'\boldsymbol{\gamma}$ as the unobserved *group effect*, the model above becomes:

$$y_{it} = \alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}, \qquad \varepsilon_{it} \sim \mathrm{N}\left(0, \tfrac{1}{\tau}\right) \tag{5.3}$$

The group effects in this formulation become additional parameters, which can be estimated given that there are multiple time observations for each unit, i. Two points deserve some attention. First, by introducing the group effects into the model we are effectively controlling for any group-specific/time-invariant characteristic that may affect the dependent variable, whether these characteristics are observable or not. This is a very powerful result as it eliminates the potential of omitting relevant time-invariant variables from the specification of the model when drawing inferences about $\boldsymbol{\beta}$. Second, because typical panel datasets consist of many groups but have a short time dimension, the number of α_is that need to be estimated may become very large, while very little information is available (only T observations) to estimate each one of them. Two approaches to deal with this potential problem are usually employed, one that imposes a hierarchical structure on the group effects and one that avoids estimating them altogether.

The first approach assumes that the α_is are independent from the variables in $\mathbf{x}$ and that each one of them follows a Normal distribution with a common mean and precision ω. If the set of independent variables includes a constant term, this model can be expressed as:

$$y_{it} = \alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}, \qquad \varepsilon_{it} \sim \mathrm{N}\left(0, \tfrac{1}{\tau}\right), \qquad \alpha_i \sim \mathrm{N}\left(0, \tfrac{1}{\omega}\right) \tag{5.4}$$

where the group effect assumes the role of a unit-specific error term. Its mean can be restricted to zero because any non-zero mean will be absorbed by the parameter associated with the

[1] Notice that $\mathbf{w}$ has only an i subscript because it does not vary over t.

constant term in $\mathbf{x}$. This *hierarchical model* is known in the frequentist econometrics literature as the *random-effects* model, presumably because the group effects are treated as random variables, which are drawn from a common distribution.

The second type of model treats the group effects as additional parameters to be estimated and, for this reason the term used in the frequentist econometrics literature to describe it is the *fixed-effects* model. Although rarely estimated in this form, this model would require placing priors on $\boldsymbol{\beta}$, τ and all α_is. In a panel dataset with a short time dimension, the priors placed on each α_i can have a large impact on the results, because there is not enough information in the data to dominate these priors. Furthermore, it takes a lot of effort to elicit appropriate priors for all group effects, especially if the number groups is large. It is worth mentioning here, however, that the advantage of of the fixed-effects over the random-effects model is that it does not require the assumption that the group effects are independent of the variables in $\mathbf{x}$.

Estimation of the α_is can be avoided in the fixed-effects model by using a simple transformation of the data. Towards this end, let $\bar{y}_{i.}$ be the sample mean over time of the dependent variable for group i: $\bar{y}_{i.} = \frac{1}{T}\sum_{t=1}^{T} y_{it}$, and define $\bar{\mathbf{x}}_{i.}$ and $\bar{\varepsilon}_{i.}$ accordingly. If (5.3) holds in the population, then it should also hold:

$$\bar{y}_{i.} = \alpha_i + \bar{\mathbf{x}}_{i.}'\boldsymbol{\beta} + \bar{\varepsilon}_{i.} \tag{5.5}$$

This result is obtained simply by adding the T equations for group i by parts and dividing by T. Finally, subtracting by parts (5.5) from (5.3) removes the group effects:

$$(y_{it} - \bar{y}_{i.}) = (\mathbf{x}_{it} - \bar{\mathbf{x}}_{i.})'\boldsymbol{\beta} + (\varepsilon_{it} - \bar{\varepsilon}_{i.}) \tag{5.6}$$

By the properties of the Normal distribution, $\varepsilon_{it} - \bar{\varepsilon}_{i.}$ also follows a Normal distribution, albeit with a precision parameter which is a complex function of the original τ that appears in the distribution of each ε_{it}. Therefore, $\boldsymbol{\beta}$ can be estimated using the same procedure as the one used in the typical linear regression model, but with dependent and independent variables constructed as deviations from the group means. Since there is nothing new in terms of statistical procedures in this model, we will not cover it further in this chapter, except only in the examples. We note, however, that this approach of removing the group effects relies heavily on the assumption that each ε_{it} follows a Normal distribution and the results do not extend to models where the error term has more elaborate structure (for example, in a stochastic-frontier model).

Both fixed- and random-effects models are designed to deal with group-specific unobserved heterogeneity that is due to time-invariant variables and which enters the model additively. With multiple time observations per group, however, unobserved heterogeneity can be modeled as entering in the form of group-specific slope parameters. In this context, the model becomes:

$$y_{it} = \mathbf{z}_{it}'\boldsymbol{\gamma}_i + \varepsilon_{it}, \qquad \varepsilon_{it} \sim \mathrm{N}\left(0, \tfrac{1}{\tau}\right) \tag{5.7}$$

If the number of time observations per group is greater than the number of independent variables in $\mathbf{z}$, then the individual $\boldsymbol{\gamma}_i$s can be estimated by applying the procedure for estimating the parameters of a linear regression model, in a group-by-group basis. Using such an approach, however, is likely to magnify the problems that appear in the fixed-effects model: eliciting priors for each $\boldsymbol{\gamma}_i$ and having limited information to estimate each $\boldsymbol{\gamma}_i$ in isolation. On the other hand, the *random-coefficients* model imposes a hierarchical structure on the $\boldsymbol{\gamma}_i$s in a similar fashion the random-effects model does on the α_is. Compared to running N regressions separately, this hierarchical structure allows information about $\boldsymbol{\gamma}_i$ to be transmitted from one group to the other. A random-coefficients model can be estimated even when there are more independent variables in the model than time observations per group because all groups contribute jointly to the estimation of the parameters. The term used to describe this effect is *borrowing of strength*.

However, we need to be careful regarding what constitutes a parameter in this model. Because each $\boldsymbol{\gamma}_i$ is now a $K \times 1$ random vector, the typical assumption made is that $\boldsymbol{\gamma}_i$s are draws from a multivariate-Normal distribution with common mean and precision matrix:

$$\boldsymbol{\gamma}_i \sim \mathrm{N}\left(\bar{\boldsymbol{\gamma}}, \boldsymbol{\Omega}^{-1}\right) \tag{5.8}$$

where $\bar{\boldsymbol{\gamma}}$ is a $K\times 1$ vector of parameters to be estimated and $\boldsymbol{\Omega}$ is a $K\times K$ precision matrix, which contains $\frac{(K-1)K}{2}+K$ unique parameters to be estimated. Estimation of the random-coefficients model is deferred to the following section.

Before closing this section we note that the panel-data models presented above can be applied to datasets with no time dimension, as long as there is another natural way of grouping the observations. For example, when modeling the profitability of a particular type of businesses, if data are available on individual businesses located in different countries, then the country can assume the role of the group and the individual businesses the role of the time dimension. A random-effects model in this context would control for unobserved heterogeneity at the country level that is due to, for example, the entrepreneurial environment, exchange rates, or any other variable that is the same for all businesses within a country.

5.3 Estimation of the Hierarchical Panel-Data Models

This section describes the estimation process of the two hierarchical panel-data models: random effects and random coefficients. Estimation of the fixed-effects model is not covered here because the procedure for estimating $\boldsymbol{\beta}$ from (5.6) is exactly the same as the one used for the linear regression model. Furthermore, estimating the α_is in the fixed-effects model by brute force and not by transforming the data in deviations from the group means is something that is rarely done in Bayesian econometrics. In the linear model with Normally-distributed error both the random-effects and the random-coefficients models can be estimated by analytically integrating the unobserved effects (α_is or $\boldsymbol{\gamma}_i$s) from the likelihood. However, estimation by data augmentation is much simpler and can be extended to models where the error term follows more elaborate distributions and for this reason we will follow this approach here.

5.3.1 Estimation of the Random-Effects Model

The random-effects model takes the form:

$$y_{it}=\alpha_i+\mathbf{x}_{it}'\boldsymbol{\beta}+\varepsilon_{it},\qquad \varepsilon_{it}\sim \mathrm{N}\left(0,\tfrac{1}{\tau}\right),\qquad \alpha_i\sim \mathrm{N}\left(0,\tfrac{1}{\omega}\right) \tag{5.9}$$

where $\mathbf{x}_{it}$ is a $K\times 1$ vector of independent variables. We will assume that we have N groups in the dataset and that each one of them is observed for T time periods. Extension to unbalanced panels is straightforward, but notation can become cumbersome. The parameters to be estimated are $\boldsymbol{\beta}$, τ and ω, while in a data-augmentation setting, the α_is are the latent data.

The complete-data likelihood for the random-effects model is:

$$\begin{aligned}\mathrm{p}\left(\mathbf{y},\{\alpha_i\}\,|\mathbf{X},\boldsymbol{\beta},\tau,\omega\right)&=\prod_{i=1}^{N}\left[\prod_{t=1}^{T}\mathrm{p}\left(y_{it}|\mathbf{x}_{it},\boldsymbol{\beta},\tau,\alpha_i\right)\right]\times\mathrm{p}\left(\alpha_i|\omega\right)\\&=\frac{\tau^{NT/2}}{(2\pi)^{NT/2}}\exp\left\{-\frac{\tau}{2}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(y_{it}-\alpha_i-\mathbf{x}_{it}'\boldsymbol{\beta}\right)^2\right\}\\&\quad\times\frac{\omega^{N/2}}{(2\pi)^{N/2}}\exp\left\{-\frac{\omega}{2}\sum_{i=1}^{N}\alpha_i^2\right\}\end{aligned} \tag{5.10}$$

where $\mathbf{y}$ and $\mathbf{X}$ are the vector and matrix of the dependent and independent variables, respectively, stacked over both time observations and groups. The first factor in the complete-data likelihood comes from the fact that each ε_{it} follows a Normal distribution and the second factor is due to each α_i following a Normal distribution with zero mean.

As in the linear regression model, we will use a multivariate-Normal prior for $\boldsymbol{\beta}$ and a Gamma prior for τ. Furthermore, since ω is another precision parameter, we will use a Gamma prior for it as well. By letting $\mathbf{y}^*$ be the $NT\times 1$ vector of stacked values of $y_{it}-\alpha_i$, it becomes apparent that the full conditionals of $\boldsymbol{\beta}$ and τ are exactly the same as the ones presented in Theorem 2.1, with $\mathbf{y}^*$ taking the place of $\mathbf{y}$. Similar transformations as the ones presented

above the same Theorem can be used to show that the full conditional of ω is Gamma and that the full conditional of each α_i is Normal. These results are presented here in the form of a theorem, before we move to an application of the fixed- and random-effects models to the estimation of the aggregate production function.

THEOREM 5.1: Full Conditionals for the Random-Effects Linear Model
In the random-effects linear model with Normally-distributed error and group effects and K independent variables:

$$y_{it} = \alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}, \qquad \varepsilon_{it} \sim \mathrm{N}\left(0, \tfrac{1}{\tau}\right), \qquad \alpha_i \sim \mathrm{N}\left(0, \tfrac{1}{\omega}\right)$$

and with a Normal prior for $\boldsymbol{\beta}$ and Gamma priors for τ and ω:

$$\mathrm{p}(\boldsymbol{\beta}) = \frac{|\mathbf{P}|^{1/2}}{(2\pi)^{K/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}-\mathbf{m})'\mathbf{P}(\boldsymbol{\beta}-\mathbf{m})\right\},$$

$$\mathrm{p}(\tau) = \frac{b_\tau^{a_\tau}}{\Gamma(a_\tau)}\tau^{a_\tau-1}e^{-b_\tau\tau} \qquad \text{and} \qquad \mathrm{p}(\omega) = \frac{b_\omega^{a_\omega}}{\Gamma(a_\omega)}\omega^{a_\omega-1}e^{-b_\omega\omega}$$

the full conditional of $\boldsymbol{\beta}$ is Normal:

$$\pi(\boldsymbol{\beta}|\bullet) = \frac{|\tilde{\mathbf{P}}|^{1/2}}{(2\pi)^{K/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}-\tilde{\mathbf{m}})'\tilde{\mathbf{P}}(\boldsymbol{\beta}-\tilde{\mathbf{m}})\right\}$$

and the full conditionals of τ and ω are Gamma:

$$\pi(\tau|\bullet) = \frac{\tilde{b}_\tau^{\tilde{a}_\tau}}{\Gamma(\tilde{a}_\tau)}\tau^{\tilde{a}_\tau-1}e^{-\tilde{b}_\tau\tau} \qquad \text{and} \qquad \pi(\omega|\bullet) = \frac{\tilde{b}_\omega^{\tilde{a}_\omega}}{\Gamma(\tilde{a}_\omega)}\omega^{\tilde{a}_\omega-1}e^{-\tilde{b}_\omega\omega}$$

where:

- $\tilde{\mathbf{P}} = \mathbf{X}'\mathbf{X} + \mathbf{P}$ and $\tilde{\mathbf{m}}_\beta = (\mathbf{X}'\mathbf{X} + \mathbf{P})^{-1}(\mathbf{X}'\mathbf{y}^* + \mathbf{P}\mathbf{m})$
- $\tilde{a}_\tau = \frac{NT}{2} + a_\tau$ and $\tilde{b}_\tau = \frac{1}{2}(\mathbf{y}^* - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y}^* - \mathbf{X}\boldsymbol{\beta}) + b_\tau$
- $\tilde{a}_\omega = \frac{N}{2} + a_\omega$ and $\tilde{b}_\omega = \frac{1}{2}\sum_{i=1}^{N}\alpha_i^2 + b_\omega$
- $\mathbf{y}^*$ is the $NT \times 1$ vector obtained by stacking $y_{it}^* = y_{it} - \alpha_i$ over T and N

The full conditional of α_i, $i = 1, 2, \ldots, N$, is Normal:

$$\pi(\alpha_i|\bullet) = \frac{(\tau T + \omega)^{1/2}}{(2\pi)^{1/2}} \exp\left\{-\frac{(\tau T + \omega)}{2}(\alpha_i - \tilde{m}_i)^2\right\}$$

where $\tilde{m}_i = \frac{\tau}{\tau T + \omega}\sum_{t=1}^{T}(y_{it} - \mathbf{x}_{it}\boldsymbol{\beta})$

♦ **Example 5.1 Fixed and Random Effects in the Aggregate Production Function**
In this example we will use the data from the Penn World Table (Feenstra et al., 2015) to estimate the aggregate production function with fixed and random effects. The dataset contains annual information on value added, capital and labor use and a time trend for the EU-15 Member States from 1970 to 2014. The unit of analysis here is the Member State and the panel is balanced because each group is observed for 45 years.

We will assume that the aggregate production function is Cobb-Douglas and, for comparison purposes, we will first consider the pooled model:

$$\log \mathsf{Y}_{it} = \beta_1 + \beta_2 \log \mathsf{K}_{it} + \beta_3 \log \mathsf{L}_{it} + \beta_4 \mathsf{trend}_{it} + \varepsilon_{it}, \qquad \varepsilon_{it} \sim \mathsf{N}\left(0, \tfrac{1}{\tau}\right)$$

The results from the pooled model are presented in the following table.

	Mean	Median	Sd.dev.	5%	95%
constant	3.24514	3.24896	0.279077	2.784	3.69835
logK	0.583076	0.582767	0.0238982	0.544447	0.622634
logL	0.441425	0.441675	0.0225753	0.403951	0.477781
trend	0.00119668	0.00120461	0.000508724	0.000353659	0.0020245
tau	72.1455	72.0801	3.9129	65.778	78.7437
sigma_e	0.117862	0.117787	0.00320426	0.112692	0.123302

The the slope coefficients in the fixed-effects model:

$$\log \mathsf{Y}_{it} = \alpha_i + \beta_2 \log \mathsf{K}_{it} + \beta_3 \log \mathsf{L}_{it} + \beta_4 \mathsf{trend}_{it} + \varepsilon_{it}, \qquad \varepsilon_{it} \sim \mathsf{N}\left(0, \tfrac{1}{\tau}\right)$$

can be estimated using BayES' `lm()` function, after transforming the dependent and independent variables by taking their deviations from their respective means. The results from this model are given in the following table.

	Mean	Median	Sd.dev.	5%	95%
dlogK	0.393942	0.393992	0.0333159	0.338807	0.448848
dlogL	0.563322	0.563276	0.0327412	0.509985	0.617021
dtrend	0.00489146	0.00489257	0.000855796	0.00348505	0.00630568
tau	176.757	176.619	9.69914	161.188	192.928
sigma_e	0.0753014	0.0752457	0.00207175	0.0719966	0.0787679

Finally, BayES' `lm_re()` function can be used to estimate the random-effects model:

$$\log \mathsf{Y}_{it} = \alpha_i + \beta_2 \log \mathsf{K}_{it} + \beta_3 \log \mathsf{L}_{it} + \beta_4 \mathsf{trend}_{it} + \varepsilon_{it}, \qquad \varepsilon_{it} \sim \mathsf{N}\left(0, \tfrac{1}{\tau}\right), \qquad \alpha_i \sim \mathsf{N}\left(0, \tfrac{1}{\omega}\right)$$

These results appear in the following table.

	Mean	Median	Sd.dev.	5%	95%
constant	5.21101	5.20963	0.383709	4.58129	5.84385
logK	0.419055	0.419135	0.0312374	0.36744	0.470369
logL	0.576326	0.576208	0.0293342	0.528282	0.624983
trend	0.00396158	0.0039528	0.000652133	0.00290387	0.00504271
tau	172.605	172.478	9.56803	156.993	188.641
omega	93.0941	87.6235	37.8641	41.6574	162.728
sigma_e	0.0762035	0.076144	0.00212013	0.0728087	0.0798106
sigma_alpha	0.110479	0.106834	0.0240874	0.0783917	0.154945

We can see in these results that the posterior means of the slope coefficients are quite similar in the fixed- and random-effects models, but there are substantial differences from the pooled model. Furthermore, the output elasticities with respect to capital and labor in the models that account for group unobserved heterogeneity are closer to the $\frac{1}{3}/\frac{2}{3}$ split, suggested by economic theory, if factors of production are compensated by their marginal products. As expected, the precision of the error term increases considerably when moving from the pooled model to the models that account for unobserved heterogeneity.

Bayes factors can be used to examine how well the data conform to the assumptions made by the three models. It should be noted that the dependent variable in these models is not the same: the fixed-effects used the deviations from the group means of log-output as the dependent variable, while the other two models use log-output itself. However, we can view the fixed-effects model as having $\log \mathsf{Y}_{it}$ as the dependent variable, while the group means of log-output are treated as forming an additional independent variable, associated with a coefficient equal to one. Running the fixed-effects model in this format would generate the same value for the log-marginal likelihood. The results in the following table indicate that data clearly favor the fixed-effects model.

Model	Log-Marginal Likelihood	Type of log-ML Approximation	Prior Model Probability	Posterior Model Probability
LM	442.739	Lewis & Raftery	0.333333	1.44989e-136
FE	755.519	Lewis & Raftery	0.333333	1
RE	702.433	Lewis & Raftery	0.333333	8.81473e-24

The results presented above can be obtained in BayES using the code in the following box.

```
// import the data and transform the variables
Data = webimport("www.bayeconsoft.com/datasets/PWT.csv");

// construct the constant term and take logs of inputs and output
Data.constant = 1;          Data.logY = log(Data.Y);
Data.logK = log(Data.K);    Data.logL = log(Data.L);

// declare the dataset as panel
set_pd(Year,CountryID);

// run a simple linear model
LM = lm(logY ~ constant logK logL trend);

// run a fixed-effects model
Data.dlogY = Data.logY - groupmeans(logY);
Data.dlogK = Data.logK - groupmeans(logK);
Data.dlogL = Data.logL - groupmeans(logL);
Data.dtrend = Data.trend - groupmeans(trend);
FE = lm(dlogY ~ dlogK dlogL dtrend);

// run a random-effects model
RE = lm_re(logY ~ constant logK logL trend);

// compare the three models
pmp( { LM, FE, RE } );
```

5.3.2 Estimation of the Random-Coefficients Model

The random-coefficients model takes the form:

$$y_{it} = \mathbf{z}_{it}'\boldsymbol{\gamma}_i + \varepsilon_{it}, \qquad \varepsilon_{it} \sim \mathrm{N}\left(0, \tfrac{1}{\tau}\right), \qquad \boldsymbol{\gamma}_i \sim \mathrm{N}\left(\bar{\boldsymbol{\gamma}}, \boldsymbol{\Omega}^{-1}\right) \tag{5.11}$$

where $\mathbf{z}_{it}$ is a $K\times 1$ vector of independent variables and each $\boldsymbol{\gamma}_i$ is a $K\times 1$ vector of associated group effects. As before, we will assume that we have N groups in the dataset and that each one of them is observed for T time periods. While extension to unbalanced panels is again straightforward, this comes at the cost much more complex notation. The parameters of the model to be estimated are $\bar{\boldsymbol{\gamma}}$, $\boldsymbol{\Omega}$ and τ. The $\boldsymbol{\gamma}_i$s represent the latent data, although they may be of interest in some applications and estimates of them can be obtained as byproducts of the Gibbs sampler.

The complete-data likelihood for the random-coefficients model is:

$$\begin{aligned}
\mathrm{p}\left(\mathbf{y}, \{\boldsymbol{\gamma}_i\} \,\middle|\, \mathbf{Z}, \bar{\boldsymbol{\gamma}}, \boldsymbol{\Omega}, \tau, \right) &= \prod_{i=1}^{N}\left[\prod_{t=1}^{T}\mathrm{p}\left(y_{it}\middle|\mathbf{z}_{it}, \tau, \boldsymbol{\gamma}_i\right)\right] \times \mathrm{p}\left(\boldsymbol{\gamma}_i\middle|\bar{\boldsymbol{\gamma}}, \boldsymbol{\Omega}\right) \\
&= \frac{\tau^{NT/2}}{(2\pi)^{NT/2}} \exp\left\{-\frac{\tau}{2}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(y_{it} - \mathbf{z}_{it}'\boldsymbol{\gamma}_i\right)^2\right\} \\
&\quad \times \frac{|\boldsymbol{\Omega}|^{N/2}}{(2\pi)^{NK/2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{N}\left(\boldsymbol{\gamma}_i - \bar{\boldsymbol{\gamma}}\right)'\boldsymbol{\Omega}\left(\boldsymbol{\gamma}_i - \bar{\boldsymbol{\gamma}}\right)\right\}
\end{aligned}$$

where $\mathbf{y}$ and $\mathbf{Z}$ are the vector and matrix of the dependent and independent variables, respectively, stacked over both time observations and groups. The first factor in the complete-data likelihood comes from the fact that each ε_{it} follows a Normal distribution and the second factor is due to each $\boldsymbol{\gamma}_i$ following a multivariate-Normal distribution.

As in the random-effects model, we will use a multivariate-Normal prior for $\bar{\boldsymbol{\gamma}}$ and a Gamma prior for τ. Because $\boldsymbol{\Omega}$ is a precision matrix, we will place a Wishart prior on it, with degrees-of-freedom parameter n and scale matrix $\mathbf{V}$. All three priors are conjugate, while the derivations of the full conditionals for $\bar{\boldsymbol{\gamma}}$ and τ follow similar steps as the ones used for the linear regression

model. Deriving the full conditional of $\boldsymbol{\Omega}$ requires transformations similar to the ones used in SUR model. Finally, the full conditional of each $\boldsymbol{\gamma}_i$ is multivariate Normal and its derivation is similar to the way the full conditional of $\boldsymbol{\beta}$ was derived in the linear regression model. We present all these results in the form of a theorem, before moving on to apply the random-coefficients model in estimating the aggregate production function.

THEOREM 5.2: Full Conditionals for the Random-Coefficients Linear Model
In the random-coefficients linear model with Normally-distributed error and group effects and K independent variables:

$$y_{it} = \mathbf{z}_{it}'\boldsymbol{\gamma}_i + \varepsilon_{it}, \qquad \varepsilon_{it} \sim \mathrm{N}\left(0, \tfrac{1}{\tau}\right), \qquad \boldsymbol{\gamma}_i \sim \mathrm{N}\left(\bar{\boldsymbol{\gamma}}, \boldsymbol{\Omega}^{-1}\right) \tag{5.12}$$

and with a Normal prior for $\bar{\boldsymbol{\gamma}}$, a Wishart prior for $\boldsymbol{\Omega}$ and a Gamma prior for τ:

$$\mathrm{p}(\bar{\boldsymbol{\gamma}}) = \frac{|\mathbf{P}|^{1/2}}{(2\pi)^{K/2}} \exp\left\{-\frac{1}{2}\left(\bar{\boldsymbol{\gamma}} - \mathbf{m}\right)' \mathbf{P}\left(\bar{\boldsymbol{\gamma}} - \mathbf{m}\right)\right\},$$

$$\mathrm{p}(\boldsymbol{\Omega}) = \frac{|\boldsymbol{\Omega}|^{\frac{n-K-1}{2}} |\mathbf{V}^{-1}|^{n/2}}{2^{nK/2}\Gamma_K\left(\frac{n}{2}\right)} \exp\left\{-\tfrac{1}{2}\operatorname{tr}\left(\mathbf{V}^{-1}\boldsymbol{\Omega}\right)\right\} \qquad \text{and} \qquad \mathrm{p}(\tau) = \frac{b^a}{\Gamma(a)}\tau^{a-1}e^{-b\tau}$$

the full conditional of $\bar{\boldsymbol{\gamma}}$ is Normal:

$$\pi\left(\bar{\boldsymbol{\gamma}}|\bullet\right) = \frac{|\tilde{\mathbf{P}}|^{1/2}}{(2\pi)^{K/2}} \exp\left\{-\frac{1}{2}\left(\bar{\boldsymbol{\gamma}} - \tilde{\mathbf{m}}\right)' \tilde{\mathbf{P}}\left(\bar{\boldsymbol{\gamma}} - \tilde{\mathbf{m}}\right)\right\}$$

the full conditionals of $\boldsymbol{\Omega}$ is Wishart and the full conditional of τ is Gamma:

$$\pi\left(\boldsymbol{\Omega}|\bullet\right) = \frac{|\boldsymbol{\Omega}|^{\frac{\tilde{n}-K-1}{2}} |\tilde{\mathbf{V}}^{-1}|^{\tilde{n}/2}}{2^{\tilde{n}K/2}\Gamma_K\left(\frac{\tilde{n}}{2}\right)} \exp\left\{-\tfrac{1}{2}\operatorname{tr}\left(\tilde{\mathbf{V}}^{-1}\boldsymbol{\Omega}\right)\right\} \qquad \text{and} \qquad \pi\left(\tau|\bullet\right) = \frac{\tilde{b}^{\tilde{a}}}{\Gamma(\tilde{a})}\tau^{\tilde{a}-1}e^{-\tilde{b}\tau}$$

where:

- $\tilde{\mathbf{P}} = N\boldsymbol{\Omega} + \mathbf{P}$ and $\tilde{\mathbf{m}} = \left(N\boldsymbol{\Omega} + \mathbf{P}\right)^{-1}\left(\boldsymbol{\Omega}\sum_{i=1}^{N}\boldsymbol{\gamma}_i + \mathbf{P}\mathbf{m}\right)$
- $\tilde{n} = N + n$, $\tilde{\mathbf{V}}^{-1} = \mathbf{C}\mathbf{C}' + \mathbf{V}^{-1}$ and $\mathbf{C} = \begin{bmatrix}\boldsymbol{\gamma}_1 - \bar{\boldsymbol{\gamma}} & \boldsymbol{\gamma}_2 - \bar{\boldsymbol{\gamma}} & \dots & \boldsymbol{\gamma}_I - \bar{\boldsymbol{\gamma}}\end{bmatrix}$
- $\tilde{a}_\tau = \frac{NT}{2} + a_\tau$ and $\tilde{b}_\tau = \frac{1}{2}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(y_{it} - \mathbf{z}_{it}'\boldsymbol{\gamma}_i\right)^2 + b_\tau$

The full conditional of $\boldsymbol{\gamma}_i$, $i = 1, 2, \dots, N$, is multivariate Normal:

$$\pi\left(\boldsymbol{\gamma}_i|\bullet\right) = \frac{|\tilde{\boldsymbol{\Omega}}_i|^{1/2}}{(2\pi)^{K/2}} \exp\left\{-\tfrac{1}{2}\left(\boldsymbol{\gamma}_i - \tilde{\boldsymbol{\gamma}}_i\right)' \tilde{\boldsymbol{\Omega}}_i\left(\boldsymbol{\gamma}_i - \tilde{\boldsymbol{\gamma}}_i\right)\right\}$$

where:

- $\tilde{\boldsymbol{\Omega}} = \tau\mathbf{z}_i'\mathbf{z}_i + \boldsymbol{\Omega}$
- $\tilde{\boldsymbol{\gamma}}_i = \left(\tau\mathbf{z}_i'\mathbf{z}_i + \boldsymbol{\Omega}\right)^{-1}\left(\tau\mathbf{z}_i'\mathbf{y}_i + \boldsymbol{\Omega}\bar{\boldsymbol{\gamma}}\right)$
- $\mathbf{y}_i$ and $\mathbf{z}_i$ are the vector and matrix of the dependent and the independent variables, respectively, for group i and stacked over the time dimension

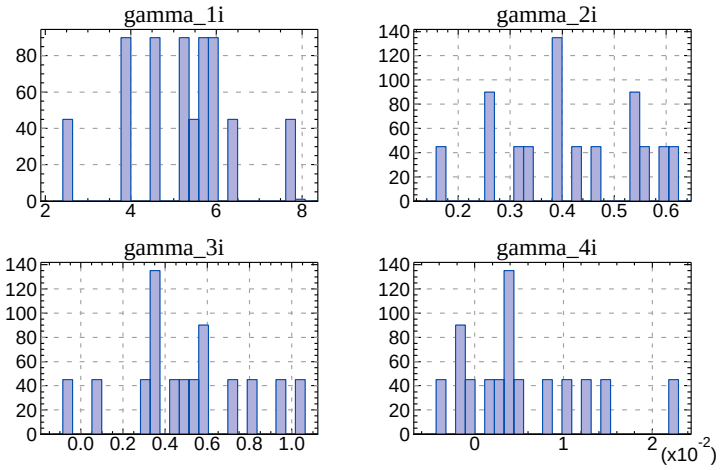
♦ **Example 5.2 Random Coefficients in the Aggregate Production Function**
In this example we will keep using the data from the Penn World Table to estimate the aggregate production function. As in the previous example, we will assume that the aggregate production function is Cobb-Douglas, but that each country has each own vector of coefficients:

$$\log \mathsf{Y}_{it} = \gamma_{1i} + \gamma_{2i}\log \mathsf{K}_{it} + \gamma_{3i}\log \mathsf{L}_{it} + \gamma_{4i}\mathsf{trend}_{it} + \varepsilon_{it}, \qquad \varepsilon_{it} \sim \mathsf{N}\left(0, \tfrac{1}{\tau}\right), \qquad \boldsymbol{\gamma}_i \sim \mathsf{N}\left(\bar{\boldsymbol{\gamma}}, \boldsymbol{\Omega}^{-1}\right)$$

This model can be estimated in BayES using the `lm_rc()` function. The results are presented in the following table.

	Mean	Median	Sd.dev.	5%	95%
constant	5.36874	5.64416	0.85051	3.57618	6.20717
logK	0.418731	0.401558	0.0658398	0.344499	0.555542
logL	0.49582	0.495558	0.0774272	0.369143	0.623302
trend	0.00537389	0.00532298	0.0105028	-0.0118546	0.0226852
tau	630.637	619.458	58.1705	550.6	740.025
sigma_e	0.0399441	0.0401789	0.00179197	0.0367607	0.0426187

These results are slightly different from the ones produced by the other panel-data models. One has to keep in mind, however, that the parameters reported in this table are the means of the output elasticities, across all countries. Country-specific estimates of these elasticities can be obtained as a byproduct of the Gibbs sampler. The following figure presents histograms of the posterior means of the country-specific output elasticities. From this figure it is obvious that there is quite large variability in these coefficients across countries.



Apart from γ, the random-coefficients model contains additional parameters. The posterior mean of $\mathbf{\Omega}$ for this model is:

$$\mathrm{E}\left(\mathbf{\Omega}|\bullet\right) = \begin{bmatrix} 252.71951 & 28.404319 & -40.135201 & -0.50773612 \\ 28.404319 & 554.45376 & 65.313201 & 20.203856 \\ -40.135201 & 65.313201 & 111.26204 & 14.386714 \\ -0.50773612 & 20.203856 & 14.386714 & 738.18898 \end{bmatrix}$$

Finally, the Lewis and Raftery approximation of the log-marginal likelihood is 1004.268, which is much larger than what the pooled, fixed-effects and random-effects models produced (see the results in Example 5.1). Therefore, the data clearly favor the random-coefficients model.

Obtaining the results presented above using BayES can be achieved using the code in the following box.

```
// import the data and transform the variables
Data = webimport("www.bayeconsoft.com/datasets/PWT.csv");

// construct the constant term and take logs of inputs and output
Data.constant = 1;          Data.logY = log(Data.Y);
Data.logK = log(Data.K);    Data.logL = log(Data.L);

// declare the dataset as panel
set_pd(Year,CountryID);

// run a random-coefficients model
RC = lm_rc(logY ~ constant logK logL trend);

// store the estimates of the country-specific coefficients
store( gamma_i, gamma_i_, "model" = RC );
```

```
// plot histograms of the country-specific coefficients
myFigure = multiplot(2,2);
hist( subplot(myFigure, 1, 1), Data.gamma_i_constant,
    "title" = "gamma_1i", "grid"="on" );
hist( subplot(myFigure, 1, 2), Data.gamma_i_logK,
    "title" = "gamma_2i", "grid"="on" );
hist( subplot(myFigure, 2, 1), Data.gamma_i_logL,
    "title" = "gamma_3i", "grid"="on" );
hist( subplot(myFigure, 2, 2), Data.gamma_i_trend,
    "title" = "gamma_4i", "grid"="on" );

// print the posterior mean of Omega
print(RC.Omega);

// print the approximation to the log-marginal likelihood
print(RC.logML);
```

5.4 Extensions to Other Panel-Data Models

Before closing this chapter we briefly discuss a few straightforward extensions to the hierarchical models for panel data. We first consider a model which can be viewed as a hybrid between the fixed- and the random-effects models and discuss along the way two approaches that account for possible correlation of the group effects with the time-varying variables by including additional independent variables. We next discuss extensions to the random-coefficients model, where some parameters are common to all groups or where a more elaborate hierarchical structure is imposed on the random coefficients.

5.4.1 Correlated Random Effects

Consider the original formulation of panel-data models:

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{w}_i'\boldsymbol{\gamma} + \varepsilon_{it}, \qquad \varepsilon_{it} \sim \mathrm{N}\left(0, \tfrac{1}{\tau}\right) \tag{5.13}$$

where $\mathbf{x}_{it}$ is a $K\times 1$ vector of time-varying and $\mathbf{w}_i$ a vector of time-invariant independent variables. Both the fixed- and random-effects models provide ways for controlling for group-specific unobserved heterogeneity, in case not all relevant time-invariant variables are observable. In specific applications, however, some of these time-invariant variables may be observed and estimating their associated coefficients may be a major objective of the analysis. Achieving this in a random-effects setting is as simple as giving an alternative interpretation to the group effect: let $\mathbf{z}_i$ be an $L\times 1$ vector of observed time-invariant variables and $\boldsymbol{\delta}$ an $L\times 1$ vector of parameters associated with these variables. $\mathbf{w}_i$ now represents a vector of unobserved time-invariant variables, which can be linearly combined to form the group effect, α_i. Thus, by estimating the model:

$$y_{it} = \alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\delta} + \varepsilon_{it}, \qquad \varepsilon_{it} \sim \mathrm{N}\left(0, \tfrac{1}{\tau}\right), \qquad \alpha_i \sim \mathrm{N}\left(0, \tfrac{1}{\omega}\right) \tag{5.14}$$

one can produce estimates for the parameters associated with the time-varying variables, $\boldsymbol{\beta}$, as well as for the parameters associated with the observed time-invariant variables, $\boldsymbol{\delta}$, while at the same time controlling for time-invariant unobserved heterogeneity. An additional assumption is made in the background: the α_is are independent of both the time-varying variables in $\mathbf{x}$ and the observed time-invariant in $\mathbf{z}$.

Although simple, this approach does not work in the fixed-effects model. This is because any time-invariant variables, whether observed or unobserved, will drop from a model estimated in deviations from the group means of the dependent and independent variables. An alternative approach, first proposed by Mundlak (1978), uses random effects instead of transforming the data, but expresses the group effects as a linear function of the the group means of the time-varying variables, in an attempt to capture possible correlations between these effects and the variables in $\mathbf{x}$:

$$\alpha_i = \bar{\mathbf{x}}_{i\cdot}'\boldsymbol{\lambda} + v_i \tag{5.15}$$

where v_i is a Normally-distributed error term. Substituting the last expression in (5.14) leads to:

$$y_{it} = v_i + \bar{\mathbf{x}}_{i.}'\boldsymbol{\lambda} + \mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\delta} + \varepsilon_{it}, \qquad \varepsilon_{it} \sim \mathrm{N}\left(0, \tfrac{1}{\tau}\right), \qquad v_i \sim \mathrm{N}\left(0, \tfrac{1}{\omega}\right) \tag{5.16}$$

From this expression it becomes apparent that v_i takes the place of α_i in the typical random-effects model, while we still need to assume that α_i is uncorrelated with $\mathbf{z}_i$. Estimation of $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ is now feasible by random effects, at the cost of including K additional independent variables in the model and having to estimate the associated parameters, $\boldsymbol{\lambda}$. At the same time, any correlation of the form expressed in (5.15) between the time-varying independent variables and the α_is is taken into account. This approach of casting a fixed-effects model into a random-effects by including the group means as additional independent variables is known as *Mundlak's approach*. Mundlak (1978) shows that in a frequentist setting and for the linear regression model only, running random effects on the augmented model produces exactly the same point estimates for $\boldsymbol{\beta}$. This result, however, does not hold exactly in a Bayesian setting, because the posterior mean of $\boldsymbol{\beta}$ depends also on the priors placed on the additional independent variables.

Chamberlain (1982, 1984) goes a step further and suggests expressing the original group-specific effects as linear functions of the the time-varying variables for each group and in every period:

$$\alpha_i = \mathbf{x}_{i1}'\boldsymbol{\lambda}_1 + \mathbf{x}_{i2}'\boldsymbol{\lambda}_2 + \ldots + \mathbf{x}_{iT}'\boldsymbol{\lambda}_T + v_i \tag{5.17}$$

where $\mathbf{x}_{it}$ is a $K\times 1$ vector of the values of $\mathbf{x}$ for group i and in period t and $\boldsymbol{\lambda}_t$ is the associated $K\times 1$ vector of coefficients. Substituting this expression in (5.14) and collecting terms leads to:

$$y_{it} = v_i + \sum_{s\neq t} \mathbf{x}_{is}'\boldsymbol{\lambda}_s + \mathbf{x}_{it}'(\boldsymbol{\beta} + \boldsymbol{\lambda}_t) + \mathbf{z}_i'\boldsymbol{\delta} + \varepsilon_{it}, \qquad \varepsilon_{it} \sim \mathrm{N}\left(0, \tfrac{1}{\tau}\right), \qquad v_i \sim \mathrm{N}\left(0, \tfrac{1}{\omega}\right) \tag{5.18}$$

The model is then estimated by random effects and the approach is known as *Chamberlain's approach* or *correlated random effects*, although the later term is frequently used also to describe Mundlak's approach.

Both specifications proposed by Mundlak and Chamberlain split the original group effects into a part that is correlated with the time-varying variables and a part that is not, while the uncorrelated part assumes the role of the unobserved-heterogeneity component in a random-effects model. Both approaches include additional observed independent variables in the model and, especially Chamberlain's approach, can lead to a proliferation of parameters and, if there is limited variability over time in the variables in $\mathbf{x}$, to severe multicollinearity problems. Additionally, they require that any observed time-invariant variables included in the model are independent of the α_is; otherwise $\boldsymbol{\delta}$ will capture, apart from the effect of the variables in $\mathbf{w}$ on y, also part of the effect of the unobserved time-invariant variables. The two approaches are particularly useful, however, outside the linear regression model, where a transformation of the data in deviations from group means, as the fixed-effects model requires, makes the distribution of the resulting error term intractable. In such a setting, usually the issue is not one of estimating the parameters associated with time-invariant variables, but allowing the group effects to be correlated with the time-invariant variables.

5.4.2 Models with Group-Specific and Common Coefficients

The random-coefficients model:

$$y_{it} = \mathbf{z}_{it}'\boldsymbol{\gamma}_i + \varepsilon_{it}, \qquad \varepsilon_{it} \sim \mathrm{N}\left(0, \tfrac{1}{\tau}\right), \qquad \boldsymbol{\gamma}_i \sim \mathrm{N}\left(\bar{\boldsymbol{\gamma}}, \boldsymbol{\Omega}^{-1}\right) \tag{5.19}$$

captures unobserved heterogeneity by allowing the coefficients associated with the time-varying independent variables to vary by group. In a model with many independent variables, however, the number of parameters to be estimated may become very large. This is because $\boldsymbol{\Omega}$ needs to be estimated along with $\bar{\boldsymbol{\gamma}}$ and the number of unique elements in $\boldsymbol{\Omega}$ is a quadratic function of the number of independent variables. Furthermore, the resulting group-specific $\boldsymbol{\gamma}_i$s may be unrealistic if the number of time observations per group is small relative to the number of

independent variables. To put it differently, a random-coefficients model may be too flexible to allow reasonable inferences to be drawn about $\bar{\boldsymbol{\gamma}}$ if all coefficients are group specific.

An obvious remedy to this issue is to restrict some of the independent variables in large models to be common to all groups. The model then becomes:

$$y_{it} = \mathbf{z}_{it}'\boldsymbol{\gamma}_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}, \qquad \varepsilon_{it} \sim \mathrm{N}\left(0, \tfrac{1}{\tau}\right), \qquad \boldsymbol{\gamma}_i \sim \mathrm{N}\left(\bar{\boldsymbol{\gamma}}, \boldsymbol{\Omega}^{-1}\right) \tag{5.20}$$

where $\mathbf{x}_{it}$ is an $L \times 1$ vector of independent variables which are associated with a parameter vector, $\boldsymbol{\beta}$, which is common to all is. The way of splitting the set of independent variables into those associated with group-specific coefficients and those associated with common parameters should be motivated by economic theory and may not always be obvious. In general, group-specific coefficients are reserved for variables whose impact on the dependent variable is expected to differ substantially among groups, given their unobserved characteristics.

The parameters to be estimated in this model are, $\bar{\boldsymbol{\gamma}}$, $\boldsymbol{\Omega}$ and τ, as in the basic random-coefficients model, plus $\boldsymbol{\beta}$, and this presents a natural blocking for the Gibbs sampler. Deriving the full conditionals of the four blocks is straightforward. The only thing that is required is to replace y_{it} by $y_{it} - \mathbf{x}_{it}'\boldsymbol{\beta}$ in the full conditional for $\boldsymbol{\gamma}_i$, by $y_{it} - \mathbf{z}_{it}'\boldsymbol{\gamma}_i$ in the full conditional of $\boldsymbol{\beta}$ in the typical linear regression model, and use the complete residual, $y_{it} - \mathbf{z}_{it}'\boldsymbol{\gamma}_i - \mathbf{x}_{it}'\boldsymbol{\beta}$, in the full conditional of τ.

5.4.3 Random-Coefficients Models with Determinants of the Means

The random-coefficients model introduces flexibility into the model but also allows for "borrowing of strength" by imposing a hierarchical structure on the $\boldsymbol{\gamma}_i$s. The assumption that we maintained until now in this model is that each $\boldsymbol{\gamma}_i$ is a draw from a multivariate-Normal distribution with common mean, $\bar{\boldsymbol{\gamma}}$. It is feasible, however, to allow this mean to be different for different groups by specifying it as a linear function of time-invariant variables and additional parameters to be estimated. More formally, the structure imposed on $\boldsymbol{\gamma}_i$ becomes:

$$\boldsymbol{\gamma}_i \sim \mathrm{N}\left(\mathbf{W}_i\boldsymbol{\xi}, \boldsymbol{\Omega}^{-1}\right) \tag{5.21}$$

where $\mathbf{W}_i$ is a matrix constructed by the variables which affect the mean of the $\boldsymbol{\gamma}_i$s and $\boldsymbol{\xi}$ contains the associated parameters, both of them taking the form described below equation (3.3), in the SUR model. This specification can, alternatively, be expressed as:

$$\boldsymbol{\gamma}_i = \mathbf{W}_i\boldsymbol{\xi} + \mathbf{u}_i, \qquad \mathbf{u}_i \sim \mathrm{N}\left(\mathbf{0}, \boldsymbol{\Omega}^{-1}\right) \tag{5.22}$$

where the similarities to the SUR model become apparent. As such, no additional work is required to derive the full conditionals of the parameters for this hierarchical model: the full conditional of τ is the same as in the typical random-coefficients model and the full conditionals of $\boldsymbol{\xi}$ and $\boldsymbol{\Omega}$ are the ones presented in Theorem 3.1, with $\boldsymbol{\gamma}_i$ assuming the role of $\mathbf{y}_i$, $\mathbf{W}_i$ that of $\mathbf{X}_i$ and $\boldsymbol{\xi}$ that of $\boldsymbol{\beta}$. Finally, the full conditional of each $\boldsymbol{\gamma}_i$ is the same as in the typical random-coefficients model, with $\mathbf{W}_i\boldsymbol{\xi}$ replacing $\bar{\boldsymbol{\gamma}}$.

5.5 Synopsis

This chapter covered the basic panel-data models in the context of the linear regression model. After motivating the use of panel-data models as a way of accounting for group-specific unobserved heterogeneity, we made the distinction between fixed-effects, random-effects and random-coefficients models. The parameters of the fixed-effects model which are common to all groups can be estimated simply by transforming the dependent and independent variables to deviations from their respective group means and using the procedure discussed in Chapter 2. Data augmentation was used for the estimation of the two hierarchical models, where the full conditionals of the models' parameters were found to be very similar to the ones coming from the linear regression and SUR models. Finally, three straightforward extensions to the basic panel-data models were discussed: correlated random effects using Mundlak's and Chamberlain's approach, models with both group-specific coefficients and parameters common to all

groups and random-coefficients models where the mean vector of the group-specific coefficients was itself expressed as a function of independent variables and parameters.

Apart from the fixed-effects model which, in deviations from group means, works only in the linear model with Normally-distributed error, the use of the other panel-data approaches presented in this chapter extends to models with more elaborate structures in the error term. As it will become apparent in the following chapters, extending non-linear models such that they accommodate group effects is relatively easy. One of the major advantages of the Bayesian approach to statistical inference is that, when coupled with data augmentation, increasing model complexity can be handled by artificially including latent data (group effects in the context of panel-data models) and then integrating them out by simulation.

Chapter 6

Models for Binary Response

6.1 Overview

This chapter introduces the simplest models that can be used to draw inferences in problems where the response variable is qualitative. In particular, it deals with models which can be used to determine the probability of the response variable being true or false and which are, therefore, appropriately called models for binary response. Because the response variable in such models is qualitative, any numerical values used to code the two states it can be in are arbitrary. This creates some complications for statistical analysis, which always has to rely on numerical data. Instead of modeling the response variable directly, binary-response models specify the probability of this variable being true. Since this probability is unobserved, a new conceptual device is required to estimate the models' parameters, as well as to interpret their results.

There are a few alternative, yet equivalent, representations of binary-respose models, some of which are useful for estimating the models' parameters, while others facilitate interpretation of the results. After defining formally what a binary-response model is, the statistical formulation is presented in the following section. Binary-response models can be given an economic interpretation within the random-utility framework. This task is taken up in Subsection 6.2.1. Estimation of the most popular models for binary choice, Probit and Logit models, is covered in Section 6.3 and the section that follows deals with the calculation and interpretation of marginal effects from such models. Section 6.5 extends the binary-response models to accommodate panel data, while Section 6.6 provides an extension to multivariate binary-response models.

The models covered in this chapter form the basis for more complex statistical models in which the response variable is qualitative and it can be in one out of multiple possible states. These models are covered in the following chapter, as they present additional complications, both in terms of interpretation of the results and in terms of estimation.

6.2 The Nature of Binary-Response Models

Binary-response or *binary-choice* models are used to draw inferences in problems where the response variable is qualitative and it can be in one of two states: true or false. In this type of problems interest revolves around the probability of occurrence of a specific economic phenomenon, corresponding to the response variable being true, as well as on the effects that any relevant factors may have on this probability. Typically, and for mathematical convenience,

the response variable is coded such that occurrence of the phenomenon under investigation is indicated by a value of one, while non-occurrence by zero. Oftentimes, the term "success" is associated with the occurrence of the phenomenon and "failure" with non-occurrence.

A few examples on which binary-response models can be applied to quantify the probability of success are the following:

- a customer buys a specific brand of milk during a visit to the grocery store (success) or not (failure)
- a household owns the house it resides in (success) or not (failure)
- an individual is employed (success) or unemployed (failure)
- an economy experiences unemployment rate greater than 10% in a given year (success) or not (failure)

As it is the case in almost all econometric models, the researcher is rarely interested only in estimating the probability of occurrence of a phenomenon. Rather, quantifying the magnitude of the causal effect of relevant economic variables on this probability is of primary importance. Such a causal relationship can be expressed, in general terms, as $y = f(x_1, x_2 \ldots, x_K)$, where y is the response variable and which can be equal to either zero or one and $x_1, x_2 \ldots, x_K$ are K independent variables that drive y. This causal relationship can be expressed mathematically once numbers are used to code the values of y. However, the mapping of the two states of the response variable to numerical values is largely arbitrary: we could equally well choose values other than one and zero to code the true/false states of the response variable and still communicate the same information about its state.

To circumvent these issues, instead of attempting to determine the value of the dependent variable directly, binary-response models specify the probability of success, conditional on the values of the independent variables. Formally, the quantity being modeled is $\text{Prob}(y = 1|\mathbf{x})$, where $\mathbf{x}$ is a $K \times 1$ random vector constructed by the K independent variables. To use a formulation similar to the linear regression model, define p_i as $\text{Prob}(y_i = 1|\mathbf{x}_i)$ for a potential observation, i. If p_i were observable then we would be able to specify and estimate the parameters of a model in the form:

$$p_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i \tag{6.1}$$

In practice, however, only y_i can be observed, not p_i. Nevertheless, the observed y_i depends heavily on the unobserved p_i and this dependence is precisely what discrete-response models exploit. Assuming that y is coded such that $y_i = 1$ whenever the economic phenomenon under investigation occurs and $y_i = 0$ whenever it does not, y_i follows a Bernoulli distribution with probability mass function:

$$\text{p}(y_i) = p_i^{y_i} \cdot (1 - p_i)^{1-y_i} \tag{6.2}$$

This expression implies that y_i is equal to one with probability p_i and equal to zero with probability $1 - p_i$.

With the connection between y_i and p_i revealed, we could proceed by inserting the specification of p_i from (6.1) into (6.2) and deriving the likelihood function. We should keep in mind, however, that p_i is a probability and it needs to be restricted on the unit interval for any possible value of $\mathbf{x}_i'\boldsymbol{\beta}$. A convenient way to achieve this is to use a monotonically-increasing function, $F(\cdot)$, which is defined on the real line and its range is the unit interval. We can then specify:

$$p_i = F(\mathbf{x}_i'\boldsymbol{\beta}) \tag{6.3}$$

$F(\cdot)$ is known as the *index function*. Inserting p_i from the last expression into (6.2) and assuming that potential observations are independent from each other leads to the likelihood function:

$$\text{p}(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}) = \prod_{i=1}^{N} \left[F(\mathbf{x}_i'\boldsymbol{\beta})\right]^{y_i} \left[1 - F(\mathbf{x}_i'\boldsymbol{\beta})\right]^{1-y_i} \tag{6.4}$$

where $\mathbf{y}$ is the $N \times 1$ vector that stores the values (0 or 1) of the response variable for N potential observations and $\mathbf{X}$ is the $N \times K$ matrix that stores the corresponding values of the independent variables.

Specification of the likelihood function is still incomplete because we have not yet chosen the form of the index function, which is used to map $\mathbf{x}_i'\boldsymbol{\beta}$ onto the unit interval. The only requirements for this function are that: (i) its domain is $(-\infty, +\infty)$, (ii) its range is $[0, 1]$ and (iii) it is monotonically increasing. Although one could think of many possible functions that satisfy these three requirements, the two functions that are used almost exclusively in applied research are the cumulative density function of a standard-Normally distributed random variable:

$$\Phi\left(\mathbf{x}_i'\boldsymbol{\beta}\right) = \int_{-\infty}^{\mathbf{x}_i'\boldsymbol{\beta}} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{t^2}{2}\right\} \mathrm{d}t \tag{6.5}$$

and the cumulative density function of a standard-Logistically distributed random variable:

$$\Lambda\left(\mathbf{x}_i'\boldsymbol{\beta}\right) = \frac{1}{1 + e^{-\mathbf{x}_i'\boldsymbol{\beta}}} = \frac{e^{\mathbf{x}_i'\boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i'\boldsymbol{\beta}}} \tag{6.6}$$

These two choices give rise, respectively, to the binary *Probit* and *Logit* models. Because the cumulative probability density function of a standard-Logistically distributed random variable is available in closed form, the likelihood in (6.4) is simple to evaluate for any given $\boldsymbol{\beta}$. This made Logit the model of choice when computers were too slow to accurately approximate the value of $\Phi\left(\mathbf{x}_i'\boldsymbol{\beta}\right)$, which is not available in closed form. With the increase of computing power in the last few decades, however, this issue became immaterial and the Probit model became the most prominent device for modeling binary response variables.

Notice that when moving from (6.1) to (6.3) we dropped the error term, ε_i. This is because, by modeling the probability of the response variable being true (rather than modeling y_i directly), the model already accounts for noise: the value of y_i is random even if p_i is deterministic and there is no need for another error term in the model. This becomes apparent when using the latent-variable representation of a binary-choice model. This formulation presents an alternative way of dealing with the complications arising from having to model a qualitative variable. It works by introducing a continuous unobserved variable, y_i^*, whose value determines whether the observed response variable, y_i, is equal to zero or one in the following way:

$$\begin{aligned} y_i^* &= \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i \\ y_i &= \begin{cases} 1 & \text{if} \quad y_i^* > 0 \\ 0 & \text{if} \quad y_i^* \leq 0 \end{cases} \end{aligned} \tag{6.7}$$

In this representation y_i^* can assume any real value and is determined by the values of the independent variables and the associated parameters, while ε_i is there to capture statistical noise. The first expression in (6.7), therefore, resembles a typical linear regression model. Because only y_i is observed, we cannot estimate the model's parameter without relying on the relationship between y_i^* and y_i. Given the specification of this relationship and because y_i^* is a random variable, we can only make probabilistic statements about the value of y_i. In particular:

$$\begin{aligned} \mathrm{Prob}\left(y_i = 1 \middle| \mathbf{x}_i\right) &= \mathrm{Prob}\left(y_i^* > 0 \middle| \mathbf{x}_i\right) \\ &= \mathrm{Prob}\left(\varepsilon_i > -\mathbf{x}_i'\boldsymbol{\beta} \middle| \mathbf{x}_i\right) \\ &= 1 - \mathrm{Prob}\left(\varepsilon_i \leq -\mathbf{x}_i'\boldsymbol{\beta} \middle| \mathbf{x}_i\right) \end{aligned} \tag{6.8}$$

To proceed, let $F\left(\cdot\right)$ be the cumulative density function of ε_i. If this function is such that the probability density function of ε_i, $F'(\cdot)$, is symmetric around zero, then $1 - F\left(z\right) = F\left(-z\right)$ for all real z. Thus:

$$\mathrm{Prob}\left(y_i = 1 \middle| \mathbf{x}_i\right) = F\left(\mathbf{x}_i'\boldsymbol{\beta}\right) \tag{6.9}$$

which is exactly what was specified in (6.3). Therefore, if we assume that ε_i follows a standard-Normal distribution, then the latent-variable representation leads to the Probit model, while if ε_i follows a standard Logistic distribution we obtain the Logit model.[1]

A slight difference between the two formulations of the statistical model remains: the first formulation uses either the standard-Normal or standard-Logistic distribution function, while in the second formulation only the mean of these distributions needs to be restricted to zero. Is the second formulation more flexible, given that we do not have to restrict the scale parameter of ε_i to one? Suppose that we allow this scale parameter, σ^2, to be different from one. Then:

$$\text{Prob}\big(y_i = 1\big|\mathbf{x}_i\big) = \text{Prob}\big(\varepsilon_i \leq \mathbf{x}_i'\boldsymbol{\beta}\big|\mathbf{x}_i\big) = \text{Prob}\left(\frac{\varepsilon_i}{\sigma} \leq \frac{\mathbf{x}_i'\boldsymbol{\beta}}{\sigma}\middle|\mathbf{x}_i\right) \tag{6.10}$$

Because ε_i/σ has a standardized distribution, the conditional probability of success is equal to the standard-Normal or standard-Logistic cumulative density function, evaluated at $\mathbf{x}_i'\boldsymbol{\beta}/\sigma$. This makes apparent that multiplying $\boldsymbol{\beta}$ and σ by any real number, z, the cumulative density function of the standardized distribution will return exactly the same value for the combination $(z\boldsymbol{\beta}, z\sigma)$ as for the combination $(\boldsymbol{\beta}, \sigma)$. In other words, there is an infinite number of parameters that produce exactly the same probability of success and, therefore, $\boldsymbol{\beta}$ and σ cannot be identified separately. There are a few different approaches for solving this non-identification problem, but the one employed most frequently in practice is to restrict σ to unity. In this way, ε_i is restricted to follow either a standard-Normal or standard-Logistic distribution and the two formulations of binary-response models are equivalent.

6.2.1 Random Utility: An Underlying Framework for Binary Choice

The two alternative formulations of binary-response models presented above deal with the problem of having to specify how a qualitative variable is determined in the population, primarily, from a statistical/technical perspective. Although departing from different starting points, both of them provide the same solution. Because they concentrate only on the technical aspects, however, these approaches are devoid of any economic meaning. For example, in a problem where the objective is to uncover the factors that determine whether a consumer buys a specific product or not, binary-choice models prescribe that the relevant quantity to be modeled is the probability of buying the product, but provide no guidance on which factors may be relevant. The *random-utility* framework is a conceptual device that connects the statistical formulations of the problem with economic theory. This framework is particularly useful when analyzing individual decisions, such as consumer choices.

To make things concrete, suppose that consumer i faces the decision of whether to buy a product or not. The utility that this consumer derives from purchasing the product or not is assumed to be a function of her characteristics, as well as of the characteristics of the product:

$$u_{1i}^* = \mathbf{x}_i'\boldsymbol{\gamma}_1 + \epsilon_{1i} \tag{6.11}$$

if she buys the product, and:

$$u_{0i}^* = \mathbf{x}_i'\boldsymbol{\gamma}_0 + \epsilon_{0i} \tag{6.12}$$

if she does not.[2] A rational consumer would be maximizing utility and, thus, purchase the product if $u_{1i}^* > u_{0i}^*$. We may assume that the consumer knows with certainty which option leads to higher utility or even know exactly how much utility she derives from each option. However, the mechanism that determines the values of u_{0i}^* and u_{1i}^* is not precisely known to the researcher. The two error terms are added to the assumed process that determines the utility levels to capture statistical noise and, due to this noise, we can only make probabilistic

[1] Notice that when the mean of the Normal or Logistic distribution is zero, the corresponding probability density function is symmetric around zero, as required by the argument made here.

[2] We use a star as a superscript on the utility levels because these are not unobservable. We will continue using such a superscript to denote unobserved quantities in this and the following chapters.

statements about whether the consumer will purchase the product or not. The probability of the consumer buying the product is:

$$\begin{aligned}\text{Prob}\left(u_{1i}^* > u_{0i}^* \middle| \mathbf{x}_i\right) &= \text{Prob}\left(\mathbf{x}_i'\left(\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_0\right) + \left(\epsilon_{1i} - \epsilon_{0i}\right) > 0 \middle| \mathbf{x}_i\right) \\ &= \text{Prob}\left(\mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i > 0 \middle| \mathbf{x}_i\right) \\ &= \text{Prob}\left(\varepsilon_i > -\mathbf{x}_i'\boldsymbol{\beta} \middle| \mathbf{x}_i\right)\end{aligned} \tag{6.13}$$

where $\boldsymbol{\beta} \equiv \boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_0$ and $\varepsilon_i \equiv \epsilon_{1i} - \epsilon_{0i}$. With these definitions we have reached the same result presented in (6.8). Thus, the random-utility framework leads to the latent-variable representation of the model, with the latent variable, y_i^*, being the difference in the utility the consumer derives from buying the product and from not buying: $y_i^* \equiv u_{1i}^* - u_{0i}^*$. Furthermore, if we assume that the two error terms are Normally distributed, then so is ε_i and we obtain the Probit model. To get the Logit model we need to assume that ϵ_{1i} and ϵ_{0i} are independent of each other and each one follows a *type I extreme-value distribution*.[3]

The random-utility framework presents a way of incorporating the assumption of rational consumers and utility-maximizing behavior into binary-choice models. Once this is done, the researcher has some guidance on which consumer characteristics or product attributes may be important in determining the decision to buy or not. This framework also suggests why we need to restrict the scale parameter of the error term to unity: utility is measured on an ordinal scale and the only thing that matters for the decision to purchase the product or not is whether $u_{1i}^* > u_{0i}^*$, not the absolute levels of u_{1i}^* and u_{0i}^*. In particular, multiplying both utility values by the same positive constant will preserve this inequality. This constant will also rescale the error term and, thus, we can implicitly pick the scaling constant's value such that the scale parameter of ε_i is equal to one.

6.3 Estimation of Binary-Response Models

There are two alternative approaches for estimating the parameters of binary-response models, one that is based on the likelihood function presented in (6.4) and one that is based on the latent-variable formulation presented in (6.7). The first of these approaches views the model as a *generalized linear model*:

$$g\left(\mu_i\right) = \mathbf{x}_i'\boldsymbol{\beta} \tag{6.14}$$

where μ_i is the expected value of the dependent variable and $g\left(\cdot\right)$ is the associated *link function*. In binary-response models and when the response variable is coded as 0/1, y_i follows a Bernoulli distribution and its expected value is simply the probability of success, p_i. Going back to (6.3), we see that p_i is specified as a monotonically increasing function of $\mathbf{x}_i'\boldsymbol{\beta}$ and, therefore, the inverse of the link function is the cumulative density function of a standard-Normally or standard-Logistically distributed random variable, respectively, for the Probit and Logit models. Viewing the models through the generalized-linear-model lens allows using a general approach for drawing inferences in these models.[4] Although useful for binary-response models, this approach cannot be easily extended to accommodate more complex discrete-response models.

The latent-variable representation, on the other hand, makes the models amenable to estimation by data augmentation and no new concepts or results need to be developed. More importantly, estimation via data augmentation scales well when more complex discrete-response

[3] The probability density and cumulative density functions of a random variable that follows a type I extreme-value distribution are, respectively, $\text{p}(z) = e^{-z}e^{-e^{-z}}$ and $\text{P}(z) = e^{-e^{-z}}$. The random variable that is obtained as the difference of two independent type-I extreme value distributed random variables follows a standard-Logistic distribution. To show this, suppose that Z and W are two random variables that follow the type I extreme-value distribution and let $U = Z - W$. Then, the cumulative density function of U is:

$$\text{Prob}(U<u) = \int_{-\infty}^{\infty} \text{Prob}(U<u|w)\,\text{p}(w)\,\text{d}w = \int_{-\infty}^{\infty} \text{P}(u+w)\,\text{p}(w)\,\text{d}w = \int_{-\infty}^{\infty} e^{-w-e^{-w}(1+e^{-u})}\,\text{d}w = \frac{e^u}{1+e^u}$$

which is the cumulative density function of a standard-Logistically distributed random variable.

[4] See chapter 16 in Gelman et al. (2013) for a concise overview of generalized linear models.

models are considered in the following chapter. For these reasons, only the latter approach will be covered in this section. Estimation of the Probit and Logit models is covered separately in the following two subsections.

6.3.1 Estimation of the Binary Probit Model

Consider the latent-variable representation of the Probit model:

$$\begin{aligned} y_i^* &= \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i, \qquad \varepsilon_i \sim \mathrm{N}(0,1) \\ y_i &= \begin{cases} 1 & \text{if} \quad y_i^* > 0 \\ 0 & \text{if} \quad y_i^* \leq 0 \end{cases} \end{aligned} \tag{6.15}$$

Naturally, in a data-augmentation setting y_i assumes the role of the observed data for observation i and y_i^* the latent data for the same observation. The contribution of a potential observation i to the complete-data likelihood is $\mathrm{p}(y_i, y_i^*|\mathbf{x}_i, \boldsymbol{\beta}) = \mathrm{p}\left(y_i\middle|y_i^*\right) \times \mathrm{p}(y_i^*|\mathbf{x}_i, \boldsymbol{\beta})$. Given that ε_i follows a standard-Normal distribution, the second factor in this expression is:

$$\mathrm{p}(y_i^*|\mathbf{x}_i, \boldsymbol{\beta}) = (2\pi)^{-1/2} \exp\left\{-\tfrac{1}{2}\left(y_i^* - \mathbf{x}_i'\boldsymbol{\beta}\right)^2\right\} \tag{6.16}$$

Although the latent-variable representation of the binary Probit model explicitly specifies the relationship between y_i and y_i^*, expressing $\mathrm{p}\left(y_i\middle|y_i^*\right)$ in a single-line formula is necessary for deriving an expression for the complete-data likelihood function. There are a few alternative ways of doing this, but a convenient one is the following:

$$\mathrm{p}\left(y_i\middle|y_i^*\right) = \mathbb{1}(y_i^* > 0)^{y_i} \cdot \mathbb{1}(y_i^* \leq 0)^{1-y_i} \tag{6.17}$$

where $\mathbb{1}(\cdot)$ is the indicator function. Let's spend a few moments to see how this expression works. First of all, this expression looks like the probability mass function of a Bernoulli-distributed random variable (y_i in our case), with probability of success equal to $\mathbb{1}(y_i^* > 0)$. But when $y_i^* > 0$ this probability becomes equal to one and the value of y_i is guaranteed to be equal to one. On the other hand, when $y_i^* \leq 0$ the probability of success is equal to zero and the value of y_i is guaranteed to be equal to zero as well.

Finally, with N independent observations, the complete-data likelihood function becomes:

$$\mathrm{p}\left(\mathbf{y}, \mathbf{y}^*\middle|\mathbf{X}, \boldsymbol{\beta}\right) = \prod_{i=1}^{N} \left[\mathbb{1}(y_i^* > 0)^{y_i} \cdot \mathbb{1}(y_i^* \leq 0)^{1-y_i} \times (2\pi)^{-1/2} \exp\left\{-\tfrac{1}{2}\left(y_i^* - \mathbf{x}_i'\boldsymbol{\beta}\right)^2\right\}\right] \tag{6.18}$$

where $\mathbf{y}$ and $\mathbf{y}^*$ are $N \times 1$ vectors that store the values of the observed and latent data, respectively, and $\mathbf{X}$ is the $N \times K$ matrix that stores the values of the independent variables for these observations.

All parameters of the binary Probit model are contained in $\boldsymbol{\beta}$. As we have done until now for slope coefficients, we will place a multivariate-Normal prior on $\boldsymbol{\beta}$, with mean vector $\mathbf{m}$ and precision matrix $\mathbf{P}$. A standard application of Bayes' theorem leads to the following posterior density for $\boldsymbol{\beta}$ and the latent data:

$$\begin{aligned} \pi\left(\boldsymbol{\beta}, \mathbf{y}^*\middle|\mathbf{y}, \mathbf{X}\right) &\propto \mathrm{p}\left(\mathbf{y}, \mathbf{y}^*\middle|\mathbf{X}, \boldsymbol{\beta}\right) \mathrm{p}(\boldsymbol{\beta}) \\ &= \prod_{i=1}^{N} \left[\mathbb{1}(y_i^* > 0)^{y_i} \cdot \mathbb{1}(y_i^* \leq 0)^{1-y_i} \times (2\pi)^{-1/2} \exp\left\{-\tfrac{1}{2}\left(y_i^* - \mathbf{x}_i'\boldsymbol{\beta}\right)^2\right\}\right] \\ &\quad \times \frac{|\mathbf{P}|^{1/2}}{(2\pi)^{K/2}} \exp\left\{-\tfrac{1}{2}\left(\boldsymbol{\beta} - \mathbf{m}\right)' \mathbf{P}\left(\boldsymbol{\beta} - \mathbf{m}\right)\right\} \end{aligned} \tag{6.19}$$

In deriving the full conditional of $\boldsymbol{\beta}$ we first need to drop all terms from the posterior density that do not involve $\boldsymbol{\beta}$ and which enter the function multiplicatively. Doing so results in a full conditional that has the same form as the one we encountered in the linear regression model:

$$\pi\left(\boldsymbol{\beta}|\bullet\right) \propto \exp\left\{-\tfrac{1}{2}\left(\mathbf{y}^* - \mathbf{X}\boldsymbol{\beta}\right)'\left(\mathbf{y}^* - \mathbf{X}\boldsymbol{\beta}\right)\right\} \times \exp\left\{-\tfrac{1}{2}\left(\boldsymbol{\beta} - \mathbf{m}\right)' \mathbf{P}\left(\boldsymbol{\beta} - \mathbf{m}\right)\right\} \tag{6.20}$$

The only difference from the expression in (2.13) for the linear regression model is that τ, the precision parameter of the error term, is now restricted to be equal to one. Following exactly the same steps as the ones bellow equation (2.13) leads to the full conditional of $\boldsymbol{\beta}$ being a multivariate-Normal density with mean $(\mathbf{X}'\mathbf{X}+\mathbf{P})^{-1}(\mathbf{X}'\mathbf{y}^*+\mathbf{Pm})$ and precision matrix $\mathbf{X}'\mathbf{X}+\mathbf{P}$.

The task of deriving the full conditional of the latent data may appear daunting at first, mainly due to the peculiar-looking first factor in the expression:

$$\pi(y_i^*|\bullet) \propto \mathbb{1}(y_i^*>0)^{y_i}\cdot\mathbb{1}(y_i^*\leq 0)^{1-y_i}\times(2\pi)^{-1/2}\exp\left\{-\tfrac{1}{2}(y_i^*-\mathbf{x}_i'\boldsymbol{\beta})^2\right\} \tag{6.21}$$

However, one has to keep in mind that the full conditional of y_i^* is also conditional on the observed y_i and information on its value leads to great simplifications. In particular, if $y_i=1$ then the first factor becomes $\mathbb{1}(y_i^*>0)$ and the full conditional of y_i^* is a Normal density with mean $\mathbf{x}_i'\boldsymbol{\beta}$ and precision one, truncated from below at zero. A similar argument in the case where $y_i=0$ shows that the only difference in the full conditional of y_i^* is that the Normal density is now truncated at zero from above.

These results are presented below in the form of a theorem. An application of the binary Probit model to determining the probability of an individual being a member of a trade union follows in Example 6.1.

THEOREM 6.1: Full Conditionals for the Binary Probit Model
In the binary-Probit model with K independent variables:

$$y_i^* = \mathbf{x}_i'\boldsymbol{\beta}+\varepsilon_i, \qquad \varepsilon_i\sim \mathrm{N}(0,1)$$
$$y_i = \begin{cases} 1 & \text{if} \quad y_i^*>0 \\ 0 & \text{if} \quad y_i^*\leq 0 \end{cases}$$

and with a Normal prior for $\boldsymbol{\beta}$:

$$\mathrm{p}(\boldsymbol{\beta}) = \frac{|\mathbf{P}|^{1/2}}{(2\pi)^{K/2}}\exp\left\{-\frac{1}{2}(\boldsymbol{\beta}-\mathbf{m})'\mathbf{P}(\boldsymbol{\beta}-\mathbf{m})\right\}$$

the full conditional of $\boldsymbol{\beta}$ is Normal:

$$\pi(\boldsymbol{\beta}|\bullet) = \frac{|\tilde{\mathbf{P}}|^{1/2}}{(2\pi)^{K/2}}\exp\left\{-\frac{1}{2}(\boldsymbol{\beta}-\tilde{\mathbf{m}})'\tilde{\mathbf{P}}(\boldsymbol{\beta}-\tilde{\mathbf{m}})\right\}$$

where:

- $\tilde{\mathbf{P}} = \mathbf{X}'\mathbf{X}+\mathbf{P}$
- $\tilde{\mathbf{m}} = (\mathbf{X}'\mathbf{X}+\mathbf{P})^{-1}(\mathbf{X}'\mathbf{y}^*+\mathbf{Pm})$

The full conditional of y_i^*, $i=1,2,\ldots,N$, is Normal, truncated from below or above at zero, depending on the value of y_i:

$$\mathrm{p}(y_i^*|\bullet) = \begin{cases} \frac{1}{(2\pi)^{1/2}}\exp\left\{-\frac{1}{2}(y_i^*-\mathbf{x}_i'\boldsymbol{\beta})^2\right\}\mathbb{1}(y_i^*>0) & \text{if} \quad y_i=1 \\ \frac{1}{(2\pi)^{1/2}}\exp\left\{-\frac{1}{2}(y_i^*-\mathbf{x}_i'\boldsymbol{\beta})^2\right\}\mathbb{1}(y_i^*\leq 0) & \text{if} \quad y_i=0 \end{cases}$$

♦ **Example 6.1 Union Membership**
In this example we will use a panel dataset of young males who lived in the Netherlands for the years covered by the data (1980-87). The data were originally collected in the context of the Dutch National Longitudinal Survey and were compiled and first used in this form by Vella & Verbeek (1998). The part of the dataset that we will use here contains annual information on 545 individuals, each one observed for 8 years, on the following variables:

`union`	:	indicator variable: 1 if the individual reported that his wage was set in a collective bargaining agreement in the year under question
`hours`	:	number of hours worked during the year (in thousands)
`married`	:	dummy variable: 1 if the individual is married
`black`	:	dummy variable: 1 if the individual is black
`hisp`	:	dummy variable: 1 if the individual is Hispanic
`health`	:	dummy variable: 1 if the individual has a health disability

Our objective is to model the probability of an individual's wage being set in a collective bargaining agreement. We will, for now, ignore the panel nature of the data and assume that this probability is equal to the standard-Normal cumulative density function, evaluated at a linear combination of the individual's characteristics and the associated coefficients:

$$\text{Prob}(\text{union}_i = 1) = \Phi\left(\beta_1 + \beta_2 \text{hours}_i + \beta_3 \text{married}_i + \beta_4 \text{black}_i + \beta_5 \text{hisp}_i + \beta_6 \text{health}_i\right)$$

where i is used to index observations across both individuals and time. These assumptions lead to a Probit model and by using BayES' `probit()` function, we obtain the results in the following table.

	Mean	Median	Sd.dev.	5%	95%
constant	-0.506081	-0.506252	0.0854881	-0.64798	-0.365264
hours	-0.167226	-0.167187	0.038776	-0.231133	-0.10355
married	0.196219	0.196309	0.0436326	0.124272	0.26797
black	0.490516	0.490638	0.0629497	0.38709	0.594179
hisp	0.189867	0.190251	0.0575819	0.0946097	0.28406
health	-0.459651	-0.457347	0.19373	-0.779257	-0.146054

From these results we can conclude that the number of hours worked by an individual, as well as the individual having a health disability, reduce the probability of his wage being set in a collective bargaining agreement. On the other hand, being married, black or Hispanic has a positive effect on this probability. Because the probability is modeled as a non-linear function of these characteristics, we can only interpret the signs, but not the magnitudes of the estimates.

The results presented above can be obtained in BayES using the code in the following box.

```
// import the data into a dataset called Data
Data = webimport("www.bayeconsoft.com/datasets/UnionMembership.csv");

// generate a constant term
Data.constant = 1;

// run the Probit model
Probit = probit( union ~ constant hours married black hisp health );
```

6.3.2 Estimation of the Binary Logit Model

The latent-variable representation of the binary Logit model:

$$\begin{aligned} y_i^* &= \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i, \qquad \varepsilon_i \sim \text{Logistic}\,(0,1) \\ y_i &= \begin{cases} 1 & \text{if} \quad y_i^* > 0 \\ 0 & \text{if} \quad y_i^* \leq 0 \end{cases} \end{aligned} \tag{6.22}$$

differs from the binary Probit model only in the distributional assumption imposed on the error term. Therefore, derivation of the complete-data likelihood follows along the same lines. With N independent observations, the complete-data likelihood function is:

$$\text{p}\left(\mathbf{y}, \mathbf{y}^* \middle| \mathbf{X}, \boldsymbol{\beta}\right) = \prod_{i=1}^{N} \left[\mathbb{1}(y_i^* > 0)^{y_i} \cdot \mathbb{1}(y_i^* \leq 0)^{1-y_i} \times \frac{e^{y_i^* - \mathbf{x}_i'\boldsymbol{\beta}}}{\left(1 + e^{y_i^* - \mathbf{x}_i'\boldsymbol{\beta}}\right)^2} \right] \tag{6.23}$$

where the second factor inside the square brackets is the probability density function of a Logistically-distributed random variable, y_i^*, with mean $\mathbf{x}_i'\boldsymbol{\beta}$ and scale parameter equal to one.

Placing a multivariate-Normal prior on $\boldsymbol{\beta}$, with mean vector $\mathbf{m}$ and precision matrix $\mathbf{P}$ and applying Bayes' theorem leads to the following posterior density:

$$\begin{aligned}\pi\left(\boldsymbol{\beta}, \mathbf{y}^* \middle| \mathbf{y}, \mathbf{X}\right) &\propto \mathrm{p}\left(\mathbf{y}, \mathbf{y}^* \middle| \mathbf{X}, \boldsymbol{\beta}\right) \mathrm{p}(\boldsymbol{\beta}) \\ &= \prod_{i=1}^{N} \left[\mathbb{1}(y_i^* > 0)^{y_i} \cdot \mathbb{1}(y_i^* \leq 0)^{1-y_i} \times \frac{e^{y_i^* - \mathbf{x}_i'\boldsymbol{\beta}}}{\left(1 + e^{y_i^* - \mathbf{x}_i'\boldsymbol{\beta}}\right)^2} \right] \\ &\quad \times \frac{|\mathbf{P}|^{1/2}}{(2\pi)^{K/2}} \exp\left\{ -\frac{1}{2} (\boldsymbol{\beta} - \mathbf{m})' \mathbf{P} (\boldsymbol{\beta} - \mathbf{m}) \right\}\end{aligned} \tag{6.24}$$

The assumption of a Logistically-distributed error term has severe implications for the full conditional of $\boldsymbol{\beta}$. In particular, after dropping terms that enter the complete-data likelihood multiplicatively and which do not involve $\boldsymbol{\beta}$, this full conditional becomes:

$$\pi(\boldsymbol{\beta}|\bullet) \propto \prod_{i=1}^{N} \left[\frac{e^{y_i^* - \mathbf{x}_i'\boldsymbol{\beta}}}{\left(1 + e^{y_i^* - \mathbf{x}_i'\boldsymbol{\beta}}\right)^2} \right] \times \exp\left\{ -\frac{1}{2} (\boldsymbol{\beta} - \mathbf{m})' \mathbf{P} (\boldsymbol{\beta} - \mathbf{m}) \right\} \tag{6.25}$$

and can hardly be simplified any further. The posterior density of $\boldsymbol{\beta}$ does not belong to any known parametric family and we cannot sample directly from it. A random-walk Metropolis-Hastings approach is, nevertheless, still feasible.

These results are presented below in the form of a theorem. The binary Logit model is then applied to the problem of determining the probability of union membership, examined in Example 6.1.

THEOREM 6.2: Full Conditionals for the Binary Logit Model
In the binary-Logit model with K independent variables:

$$\begin{aligned} y_i^* &= \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i, \qquad \varepsilon_i \sim \text{Logistic}\,(0,1) \\ y_i &= \begin{cases} 1 & \text{if} \quad y_i^* > 0 \\ 0 & \text{if} \quad y_i^* \leq 0 \end{cases}\end{aligned}$$

and with a Normal prior for $\boldsymbol{\beta}$:

$$\mathrm{p}(\boldsymbol{\beta}) = \frac{|\mathbf{P}|^{1/2}}{(2\pi)^{K/2}} \exp\left\{ -\frac{1}{2} (\boldsymbol{\beta} - \mathbf{m})' \mathbf{P} (\boldsymbol{\beta} - \mathbf{m}) \right\}$$

the full conditional of $\boldsymbol{\beta}$ is:

$$\pi(\boldsymbol{\beta}|\bullet) \propto \prod_{i=1}^{N} \left[\frac{e^{y_i^* - \mathbf{x}_i'\boldsymbol{\beta}}}{\left(1 + e^{y_i^* - \mathbf{x}_i'\boldsymbol{\beta}}\right)^2} \right] \times \exp\left\{ -\frac{1}{2} (\boldsymbol{\beta} - \mathbf{m})' \mathbf{P} (\boldsymbol{\beta} - \mathbf{m}) \right\}$$

The full conditional of y_i^*, $i = 1, 2, \ldots, N$, is Logistic, truncated from below or above at zero, depending on the value of y_i:

$$\mathrm{p}\left(y_i^* \middle| \bullet\right) = \begin{cases} \frac{e^{y_i^* - \mathbf{x}_i'\boldsymbol{\beta}}}{\left(1 + e^{y_i^* - \mathbf{x}_i'\boldsymbol{\beta}}\right)^2} \mathbb{1}(y_i^* > 0) & \text{if} \quad y_i = 1 \\ \frac{e^{y_i^* - \mathbf{x}_i'\boldsymbol{\beta}}}{\left(1 + e^{y_i^* - \mathbf{x}_i'\boldsymbol{\beta}}\right)^2} \mathbb{1}(y_i^* \leq 0) & \text{if} \quad y_i = 0 \end{cases}$$

♦ **Example 6.1 Union Membership (Continued)**
We will use here again the data from Vella & Verbeek (1998) to model the probability of an individual's wage being set in a collective bargaining agreement. This time, however, we will assume that this probability is equal to the standard-Logistic cumulative density function, evaluated at a linear combination of the individual's characteristics and the associated coefficients:

$$\text{Prob}(\mathsf{union}_i = 1) = \Lambda\left(\beta_1 + \beta_2 \mathsf{hours}_i + \beta_3 \mathsf{married}_i + \beta_4 \mathsf{black}_i + \beta_5 \mathsf{hisp}_i + \beta_6 \mathsf{health}_i\right)$$

which leads to the Logit model. The results obtained using BayES' `logit()` function to estimate this model are presented in the following table.

	Mean	Median	Sd.dev.	5%	95%
constant	-0.837459	-0.840063	0.145692	-1.07483	-0.602593
hours	-0.276154	-0.276099	0.0639649	-0.380458	-0.168628
married	0.339109	0.339522	0.0729668	0.221854	0.461872
black	0.825599	0.823883	0.104185	0.65558	0.998639
hisp	0.323381	0.322948	0.0973918	0.166631	0.485795
health	-0.833753	-0.82473	0.362296	-1.42905	-0.266384

From these results we can see again that the number of hours worked by an individual and the individual having a health disability, reduce the probability being modeled, while being married, black or Hispanic increase this probability. The signs of the parameter estimates are the same as in the binary Probit model, but their magnitudes are very different. This, however, should be expected: the Probit and Logit models make different assumptions on the functional form of the relationship between the probability and the independent variables and, for this reason, the parameters are not comparable in terms of magnitude.

Obtaining the results presented above using BayES can be achieved using the code in the following box.

```
// import the data into a dataset called Data
Data = webimport("www.bayeconsoft.com/datasets/UnionMembership.csv");

// generate a constant term
Data.constant = 1;

// run the Logit model
Logit = logit( union ~ constant hours married black hisp health );
```

6.4 Interpretation of Parameters and Marginal Effects

The response variable in a binary-response model is qualitative and, for mathematical convenience, is coded as 0 or 1. However, the quantity actually being modeled is the probability of the response variable being equal to one. The relationship between this probability and the model's independent variables is non-linear and, for this reason, the magnitude of the parameters cannot be interpreted directly. Nevertheless, their signs can and this is because function $F(\cdot)$ that projects $\mathbf{x}_i'\boldsymbol{\beta}$ from the real line onto the unit interval is monotonically increasing.[5] This implies that, if $\mathbf{x}_i'\boldsymbol{\beta}$ changes due to a small change in the k-th independent variable, the probability of success will always change in the same direction.

If we want to obtain a quantitative measure of how large is the effect of a change in the k-th independent variable on the probability of success, we have to calculate the marginal effect for this variable. Applying the chain rule to a generic index function leads to the following expression for this effect:

$$\frac{\partial \operatorname{Prob}(y_i = 1|\mathbf{x}_i)}{\partial x_{ik}} = f\left(\mathbf{x}_i'\boldsymbol{\beta}\right) \cdot \beta_k \tag{6.26}$$

where $f(\cdot)$ is the derivative function of $F(\cdot)$. When $F(\cdot)$ is a cumulative density function, $f(\cdot)$ is the associated probability density function. Thus, this expression becomes:

$$\frac{\partial \operatorname{Prob}(y_i = 1|\mathbf{x}_i)}{\partial x_{ik}} = \left[(2\pi)^{-1/2} \exp\left\{-\tfrac{1}{2}\left(\mathbf{x}_i'\boldsymbol{\beta}\right)^2\right\}\right] \cdot \beta_k \tag{6.27}$$

for the Probit model and:

$$\frac{\partial \operatorname{Prob}(y_i = 1|\mathbf{x}_i)}{\partial x_{ik}} = \left[\frac{e^{\mathbf{x}_i'\boldsymbol{\beta}}}{\left(1 + e^{\mathbf{x}_i'\boldsymbol{\beta}}\right)^2}\right] \cdot \beta_k \tag{6.28}$$

[5] $F(\cdot)$ is the standard-Normal or standard-Logistic cumulative density function for the Probit and Logit models, respectively.

for the Logit model. Due to the first factor in each expression being non-linear in $\mathbf{x}_i'\boldsymbol{\beta}$, the marginal effects depend on the point, $\mathbf{x}_i$, at which they are evaluated. Although this can be any point of particular interest, typical choices are the mean or median of the independent variables, as they are observed in the dataset used for estimation. An alternative approach is to calculate the marginal effect for each data point and then report the sample average of these observation-specific effects.

Contrary to the model's parameters, marginal effects can be given an interpretation in terms of the units of measurement of the independent variables. For example, if the marginal effect of the k-th independent variable, evaluated at a particular point, is equal to z, then a unit increase in this independent variable leads to a change in the probability of success by z. Interpretation of the marginal effect of a dummy independent variable requires some attention. Because dummy variables can only assume two values, 0 or 1, it does not make sense to ask questions that involve a "small change" in their value. Typically, if the k-th independent variable is a dummy variable, its "marginal effect" is calculated as the difference in the probability of success when the value of the dummy variable changes from 0 to 1:

$$\text{Prob}(y_i=1|\mathbf{x}_{i1}) - \text{Prob}(y_i=1|\mathbf{x}_{i0}) = F\left(\mathbf{x}_{i1}'\boldsymbol{\beta}\right) - F\left(\mathbf{x}_{i0}'\boldsymbol{\beta}\right) \tag{6.29}$$

where $\mathbf{x}_{i0}$ is a vector that consists of the values of the independent variables at the point at which the marginal effect is evaluated, but with a zero in the k-th place. $\mathbf{x}_{i1}$ is a similar vector, but with a one in the k-th place.

An important thing to recognize about the marginal effects is that they are also random variables. Even if the point at which a marginal effect is evaluated is treated as fixed (for example, chosen by the researcher), the value of the marginal effect depends on the values of the βs. Nevertheless, uncertainty with respect to the values of the parameters can be taken into account by evaluating the marginal effect at each draw obtained from the full conditional of $\boldsymbol{\beta}$, generated by the Gibbs sampler. This approach amounts to simulation-based approximation of the moments of marginal effects and the researcher can choose which of these moments to report.

♦ **Example 6.1 Union Membership (Continued)**

We will keep using here the data from Vella & Verbeek (1998) that were used in the two previous parts of this example to model the probability of an individual's wage being set in a collective bargaining agreement, using a Probit and Logit model, respectively. If the models estimated using the BayES' `probit()` and a `logit()` functions are stored in memory (they are given a left-hand-side value), then the `mfx()` function can be used to calculate the marginal effects for the models' independent variables. The marginal effects for the Probit model, evaluated at the sample means of the independent variables are given in the following table. This table is followed by a table that contains the corresponding marginal effects obtained from the Logit model.

dProb(y=1)/dx	Mean	Median	Sd.dev.	5%	95%
hours	-0.0519124	-0.0518528	0.0120186	-0.071646	-0.032153
*married	0.0613966	0.0613737	0.0137269	0.0388949	0.0840416
*black	0.169879	0.169763	0.0234874	0.131463	0.2089
*hisp	0.0617634	0.0617471	0.0193936	0.0300097	0.093782
*health	-0.114603	-0.118348	0.040164	-0.173875	-0.0429363

*Marginal effect is calculated for discrete change from 0 to 1.

	Mean	Median	Sd.dev.	5%	95%
hours	-0.050128	-0.0500867	0.0116633	-0.0691871	-0.0305501
*married	0.0621888	0.0622301	0.0134484	0.0404735	0.0850314
*black	0.172435	0.172149	0.023942	0.133665	0.213077
*hisp	0.062288	0.0618445	0.019634	0.0314598	0.0955617
*health	-0.114687	-0.118931	0.0394839	-0.171541	-0.0453364

*Marginal effect is calculated for discrete change from 0 to 1.

In the last part of this example we recognized that the parameter estimates from the Probit and Logit models differ substantially in magnitude. This is to be expected as the parameters themselves play a different role in each model. The marginal effects, however, measure the same underlying

quantity (change in probability caused by a "small change" in the independent variable) and their magnitude is very similar in the two models. For example, the marginal effect of the hours variable from the Probit model suggests that increasing the number of hours worked during the year by 1,000 reduces the probability of success, in expectation, by 0.052. The corresponding reduction from the Logit model is 0.050 units. Similarly, a married person has higher probability of having his wage set by a collective agreement relative to a single person by 0.061/0.062 units, according to each model.

The results presented above can be obtained in BayES using the code in the following box.

```
// import the data into a dataset and generate a constant term
Data = webimport("www.bayeconsoft.com/datasets/UnionMembership.csv");
Data.constant = 1;

// run the Probit and Logit models
Probit = probit( union ~ constant hours married black hisp health );
Logit = logit( union ~ constant hours married black hisp health );

// calculate marginal effects from the two models at the means of the data
mfx( "model"=Probit, "point"="mean" );
mfx( "model"=Logit, "point"="mean" );
```

6.5 Binary-Response Models for Panel Data

Extending binary-response models so that they can accommodate panel data is straightforward when data augmentation is used. In close analogy to a linear regression model with individual effects, a random-effects binary-response model assumes that the probability of success is determined by a monotonic transformation of the group effects and the sum of interactions between independent variables and parameters to be estimated:

$$\text{Prob}(y_{it}=1|\mathbf{x}_{it},\alpha_i)=F\left(\alpha_i+\mathbf{x}_{it}'\boldsymbol{\beta}\right) \tag{6.30}$$

$F(\cdot)$ in this expression is a generic index function that maps its argument onto the unit interval. As with the simple binary-response models, estimation of random-effects models is easier to handle using the latent-variable representation. In this formulation the Probit model becomes:

$$y_{it}^*=\alpha_i+\mathbf{x}_{it}'\boldsymbol{\beta}+\varepsilon_{it},\qquad \varepsilon_{it}\sim \text{N}(0,1),\qquad \alpha_i\sim \text{N}\left(0,\tfrac{1}{\omega}\right)$$
$$y_{it}=\begin{cases}1 & \text{if} \quad y_{it}^*>0\\ 0 & \text{if} \quad y_{it}^*\leq 0\end{cases}$$

where the assumption that each group effect follows a Normal distribution is added. The only difference in a random-effects Logit model is that ε_{it} follows a standard-Logistic distribution.

No new results are required for estimating the binary-Probit model with random effects and only minor changes are necessary for the random-effects binary-Logit model. The full conditionals of $\boldsymbol{\beta}$ and y_{it}^* in these two models are similar to those that appear in Theorems 6.1 and 6.2 and need only to be amended such that the means of the corresponding densities are $\alpha_i+\mathbf{x}_{it}'\boldsymbol{\beta}$ instead of $\mathbf{x}_i'\boldsymbol{\beta}$. For the Probit model, the values of the y_{it}^*s generated by the Gibbs sampler replace y_{it} in the full conditional of the α_is, as they are presented in Theorem 5.1, while no changes are required for the full conditional of ω from the linear model. For the Logit model, however, the full conditionals of the α_is no longer belong to a parametric family from which random numbers can be drawn directly:

$$\pi(\alpha_i|\bullet)\propto\prod_{t=1}^{T}\left[\frac{e^{y_{it}^*-\alpha_i-\mathbf{x}_{it}'\boldsymbol{\beta}}}{\left(1+e^{y_{it}^*-\alpha_i-\mathbf{x}_{it}'\boldsymbol{\beta}}\right)^2}\right]\times\exp\left\{-\tfrac{\omega\alpha_i^2}{2}\right\} \tag{6.31}$$

As always, Metropolis-Hastings updates can be used to sample from the full conditional of each α_i.

Although no real conceptual differences appear when moving from simple binary-response models to models with random effects, calculation and interpretation of marginal effects present

some new challenges. The marginal effect of the k-th independent variable with a generic index function, $F(\cdot)$, is:

$$\frac{\partial \operatorname{Prob}(y_{it}=1|\mathbf{x}_{it},\alpha_i)}{\partial x_{it,k}} = f(\alpha_i + \mathbf{x}'_{it}\boldsymbol{\beta}) \cdot \beta_k \tag{6.32}$$

where $f(\cdot)$ is the derivative function of $F(\cdot)$. However, the group effects are not observed and, therefore, the point at which $f(\cdot)$ has to be evaluated is unknown. There are two approaches for dealing with this issue:

1. restrict each α_i to zero, as this is both its most likely and its expected value according to the assumption $\alpha_i \sim \mathrm{N}\left(0, \frac{1}{\omega}\right)$. This leads to *conditional marginal effects*.
2. integrate uncertainty with respect to the values of α_i from the marginal effects. This leads to *averaged marginal effects*.

The method of calculating averaged marginal effects is easier to describe using the latent-variable representation of binary-response models and much easier to perform in the Probit model. The latent-variable representation of the Probit model implies:

$$y_{it} = \begin{cases} 1 & \text{if} \quad \alpha_i + \varepsilon_{it} > -\mathbf{x}'_{it}\boldsymbol{\beta} \\ 0 & \text{if} \quad \alpha_i + \varepsilon_{it} \leq -\mathbf{x}'_{it}\boldsymbol{\beta} \end{cases} \tag{6.33}$$

Notice that because α_i is unobserved, it is treated here as part of the error term. Let $w_{it} \equiv \alpha_i + \varepsilon_{it}$. Since w_{it} is the convolution of two Normally-distributed random variables, it also follows a Normal distribution with mean zero and precision $\xi = \frac{\omega}{1+\omega}$. Therefore:

$$\begin{aligned} \operatorname{Prob}\left(y_{it}=1\middle|\mathbf{x}_{it}\right) &= \operatorname{Prob}\left(w_{it} > -\mathbf{x}'_{it}\boldsymbol{\beta}\middle|\mathbf{x}_{it}\right) \\ &= \operatorname{Prob}\left(z_{it} > -\sqrt{\xi}\cdot\mathbf{x}'_{it}\boldsymbol{\beta}\middle|\mathbf{x}_{it}\right) \\ &= \Phi\left(\sqrt{\xi}\cdot\mathbf{x}'_{it}\boldsymbol{\beta}\right) \end{aligned} \tag{6.34}$$

where $z_{it} \sim \mathrm{N}(0,1)$. Given this result, the averaged marginal effect for the k-th independent variable, evaluated at point $\mathbf{x}_{it}$ is:

$$\frac{\partial \operatorname{Prob}(y_{it}=1)}{\partial x_{it,k}} = (2\pi)^{-1/2} \exp\left\{-\frac{\left(\sqrt{\xi}\cdot\mathbf{x}'_{it}\boldsymbol{\beta}\right)^2}{2}\right\} \cdot \sqrt{\xi}\cdot\beta_k \tag{6.35}$$

This procedure takes care of uncertainty with respect to the value of α_i. As before, uncertainty with respect to the values of the parameters can be integrated out by evaluating the averaged marginal effects at each pair of draws, $\left(\boldsymbol{\beta}^{(g)}, \omega^{(g)}\right)$, produced by the Gibbs sampler.

It may appear at first that the same procedure, with minor modifications, can be followed when calculating the averaged marginal effects for a Logit model. This procedure, however, hits onto a wall when having to define the cumulative density function of $w_{it} \equiv \alpha_i + \varepsilon_{it}$. Because in the Logit model ε_{it} follows a standard-Logistic distribution, w_{it} is the convolution of a Normally and a Logistically distributed random variable and its cumulative density function has a very complex form (Gupta & Nadarajah, 2008). Instead of approximating this complex function, α_i can be treated as an unknown and integrated from the marginal effect using a Guass-Hermite quadrature. Formally, the averaged marginal effect for the k-th variable is:

$$\frac{\partial \operatorname{Prob}(y_{it}=1|\mathbf{x}_{it})}{\partial x_{it,k}} = \beta_k \int_{-\infty}^{\infty} \frac{\partial \operatorname{Prob}(y_{it}=1|\mathbf{x}_{it},\alpha)}{\partial x_{it,k}} \,\mathrm{p}(\alpha)\,\mathrm{d}\alpha \tag{6.36}$$

were the fact that α_i is independent of the x variables is used to simplify the expression.

♦ **Example 6.2 Union Membership with Random Effects**

In this example we will use again the data from Vella & Verbeek (1998). We will employ both Probit and Logit models with random effects of the following form:

$$\text{Prob}(\text{union}_{it} = 1) = F\left(\alpha_i + \beta_1 + \beta_2 \text{hours}_{it} + \beta_3 \text{married}_{it} + \beta_4 \text{black}_{it} + \beta_5 \text{hisp}_{it} + \beta_6 \text{health}_{it}\right)$$

to model the probability of an individual's wage being set in a collective bargaining agreement. Using BayES' `probit_re()` and `logit_re()` functions, we obtain the results in the following two tables.

We notice again that the posterior means of the parameters differ substantially between the two models. Furthermore, these estimates differ from the ones obtained using the Probit and Logit models without random effects, which, along with the posterior mean of ω being relatively small, indicates that individual-specific unobserved heterogeneity has a large impact on the results. As in the simple binary-response models, we can only interpret the signs of the parameter estimates, but not their magnitudes.

	Mean	Median	Sd.dev.	5%	95%
constant	-1.26135	-1.25964	0.18163	-1.56251	-0.965651
hours	-0.173859	-0.173641	0.0658939	-0.281828	-0.0659159
married	0.148267	0.148004	0.0827272	0.0126782	0.284378
black	0.980143	0.97843	0.26235	0.551895	1.41573
hisp	0.497062	0.496222	0.233149	0.115129	0.882007
health	-0.426416	-0.423766	0.273909	-0.88632	0.0184441
omega	0.344765	0.342964	0.0394689	0.283148	0.41243
sigma_alpha	1.71148	1.70756	0.0982653	1.55714	1.87934

	Mean	Median	Sd.dev.	5%	95%
constant	-2.23299	-2.2333	0.314293	-2.75218	-1.71597
hours	-0.32676	-0.326845	0.117889	-0.520654	-0.133128
married	0.285391	0.285187	0.146128	0.0426119	0.52498
black	1.78431	1.79204	0.437257	1.05695	2.49308
hisp	0.872266	0.875463	0.420684	0.173618	1.56974
health	-0.78715	-0.763513	0.495039	-1.6381	-0.0183024
omega	0.108519	0.10798	0.0130879	0.0879224	0.130807
sigma_alpha	3.05228	3.04319	0.185364	2.76494	3.37253

The magnitude of the corresponding marginal effects, on the other hand, can be interpreted in terms of the units of measurement of the independent variables. The following two tables present the averaged marginal effects for the Probit and Logit models, respectively, both evaluated at the sample means of the independent variables. As in the case of binary-response models without random effects, the posterior moments of the marginal effects from the random-effects Probit and Logit models are very similar.

dProb(y=1)/dx	Mean	Median	Sd.dev.	5%	95%
hours	-0.0273372	-0.0273097	0.0104264	-0.044536	-0.0103453
*married	0.0234191	0.0233288	0.0131467	0.0019967	0.0451452
*black	0.172155	0.171653	0.0488668	0.0931708	0.253171
*hisp	0.0832189	0.0824989	0.0401423	0.0183486	0.150572
*health	-0.0597496	-0.0613837	0.0363692	-0.117014	0.00294351

*Marginal effect is calculated for discrete change from 0 to 1.

dProb(y=1)/dx	Mean	Median	Sd.dev.	5%	95%
hours	-0.0290922	-0.0290373	0.0105196	-0.0463726	-0.0119123
*married	0.0254943	0.0254845	0.0131122	0.00380234	0.0471321
*black	0.174426	0.174704	0.0457013	0.0992204	0.249219
*hisp	0.0814554	0.0812937	0.0400345	0.0156925	0.147785
*health	-0.0623797	-0.0631842	0.0363061	-0.121534	-0.00161118

*Marginal effect is calculated for discrete change from 0 to 1.

Finally the conditional marginal effects for the Probit and Logit models, again evaluated at the sample means of the independent variables, are given in following two tables. From these results we notice that the averaged and conditional marginal effects differ only slightly.

dProb(y=1|a=0)

dx	Mean	Median	Sd.dev.	5%	95%
hours	-0.0263526	-0.025962	0.0105964	-0.0443266	-0.00965044
*married	0.0229053	0.0223969	0.0133352	0.00187171	0.0456316
*black	0.237393	0.232529	0.0836294	0.10916	0.383237
*hisp	0.0985098	0.0937853	0.0540631	0.0183407	0.195316
*health	-0.0431709	-0.0462635	0.0249444	-0.0782462	0.00281976

*Marginal effect is calculated for discrete change from 0 to 1.

dProb(y=1|a=0)

dx	Mean	Median	Sd.dev.	5%	95%
hours	-0.023151	-0.0227135	0.00899897	-0.0386072	-0.00919084
*married	0.0206655	0.0202458	0.0111638	0.00298201	0.0398194
*black	0.22872	0.223768	0.0815691	0.103695	0.371767
*hisp	0.0851502	0.0800911	0.0500504	0.0129658	0.176661
*health	-0.0369545	-0.0388756	0.0201243	-0.0666451	-0.00121787

*Marginal effect is calculated for discrete change from 0 to 1.

Given that the probability of occurrence of the same event is modeled by both Probit and Logit models, we can perform formal model comparison. The results in the following table indicate that the data clearly favor the Logit model.

Model	Log-Marginal Likelihood	Type of log-ML Approximation	Prior Model Probability	Posterior Model Probability
REProbit	-1698.31	Lewis & Raftery	0.5	0.00300232
RELogit	-1692.5	Lewis & Raftery	0.5	0.996998

The results presented above can be obtained in BayES using the code in the following box.

```
// import the data into a dataset and generate a constant term
Data = webimport("www.bayeconsoft.com/datasets/UnionMembership.csv");
Data.constant = 1;

// declare the dataset as a panel
set_pd(year,id);

// run a random-effects Probit model
REProbit = probit_re( union ~ constant hours married black hisp health,
    "chains"=2, "thin"=10 );

// run a random-effects Logit model
RELogit = logit_re( union ~ constant hours married black hisp health,
    "chains"=2, "thin"=10 );

// calculate averaged marginal effects for the two models
mfx( "model"=REProbit, "point"="mean", "type"=1 );
mfx( "model"=RELogit, "point"="mean", "type"=1 );

// calculate conditional marginal effects for the two models
mfx( "model"=REProbit, "point"="mean", "type"=2 );
mfx( "model"=RELogit, "point"="mean", "type"=2 );

// compare the two models
pmp( { REProbit, RELogit } );
```

6.6 Multivariate Binary-Response Models

Until now we worked with models that have a single response variable, y_i, which can be in one out of two possible states. In specific settings, however, the phenomenon under investigation may involve binary-response variables, the states of which may not be independent. For

example, we may be interested on the magnitudes of the effects of various socioeconomic and demographic characteristics of an individual (income, occupation type, age, marital status, etc.) on the probability of the individual purchasing a life-insurance policy and making voluntary contributions to a pension scheme. In principle, we could model the two binary decisions using two separate binary-response models. If y_{i1} and y_{i2} are two binary random variables indicating the individual's choices regarding the first and second decisions, respectively, the two models in the latent-variable representation would be:

$$\begin{array}{rcl} y_{1i}^* & = & \mathbf{x}_{1i}'\boldsymbol{\beta}_1 + \varepsilon_{1i} \\ y_{2i}^* & = & \mathbf{x}_{2i}'\boldsymbol{\beta}_2 + \varepsilon_{2i} \end{array} \tag{6.37}$$

where, for generality, we also allow for the possibility of the vectors of independent variables to differ between equations. In practice, however, we would expect the two decisions to be interrelated. For example, the degree of risk aversion of the decision maker, which is rarely observed and, thus, not included in either $\mathbf{x}_{1i}$ or $\mathbf{x}_{2i}$, would affect both y_{1i}^* and y_{2i}^* in the same direction. This effect would be captured by the error terms, making them positively correlated and inducing a positive correlation between the observed y_{1i} and y_{2i}. By modelling the two decisions separately, we ignore this correlation, as we work with the marginal distribution of each error term relative to the other. The main implication of this is that we are not exploiting all the information available in the data, something that usually results in wider credible intervals for the parameters in $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ compared to the case where this information is fully exploited. Multivariate binary-response models take into account the possible interrelationship between the random variables using a setup similar to the Seemingly Unrelated Regressions (SUR) models (see Chapter 3). Another reason for using a multivariate binary-response model versus multiple univariate ones is that the multivariate model can make predictions for the probabilities of combinations of values for the response variables. In the example above, a bivariate binary-response model could predict the probability of and individual both purchasing a life insurance and making voluntary contributions, $\text{Prob}(y_{1i}=1, y_{2i}=1|\mathbf{x}_{1i}, \mathbf{x}_{2i})$, doing only one of the two, $\text{Prob}(y_{1i}=1, y_{2i}=0|\mathbf{x}_{1i}, \mathbf{x}_{2i})$ or $\text{Prob}(y_{1i}=0, y_{2i}=1|\mathbf{x}_{1i}, \mathbf{x}_{2i})$, or neither, $\text{Prob}(y_{1i}=0, y_{2i}=0|\mathbf{x}_{1i}, \mathbf{x}_{2i})$.

In the most general setting, we consider an M-dimensional random vector $\mathbf{y}_i$, each element of which is a random variable that can assume two possible values: 0 and 1. Similarly to univariate binary-response models, we will model the probability of $\mathbf{y}_i$ being equal to an M-dimensional vector, $\mathbf{z}$, of 0s and 1s.[6] Contrary to univariate models, however, it is much harder to work with this probability directly, as this now involves a multidimensional integral. Instead, we will start from the latent-variable representation of the model. Toward this end, for each element y_{mi} of $\mathbf{y}_i$ we define a latent variable, y_{mi}^*, that satisfies:

$$y_{mi} = \begin{cases} 1 & \text{if} \quad y_{mi}^* > 0 \\ 0 & \text{if} \quad y_{mi}^* \leq 0 \end{cases} \qquad \forall m = 1, 2. \ldots, M \tag{6.38}$$

We then assume that each y_{mi}^* can be expressed as a linear function of independent variables and parameters to be estimated plus statistical noise:

$$\begin{array}{rcl} y_{1i}^* & = & \mathbf{x}_{1i}'\boldsymbol{\beta}_1 + \varepsilon_{1i} \\ y_{2i}^* & = & \mathbf{x}_{2i}'\boldsymbol{\beta}_2 + \varepsilon_{2i} \\ \vdots & & \vdots \\ y_{Mi}^* & = & \mathbf{x}_{Mi}'\boldsymbol{\beta}_M + \varepsilon_{Mi} \end{array} \tag{6.39}$$

This system of equations is almost identical to the one in (3.1) that we encountered in the SUR model and the only difference is that the dependent variables are now unobserved. A distributional assumption on the error terms is enough to complete the specification of the model and, throughout this section, we will assume that the ε_{mi}s jointly follow a multivariate-Normal distribution with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$. This gives rise to the *multivariate*

[6] Vector $\mathbf{z}$ is just an element of the sample space of $\mathbf{y}_i$. Notice that in univariate binary-response models $\mathbf{z}$ was typically set equal to 1, while we could easily get $\text{Prob}(y_i=0|\mathbf{x}_i) = 1 - \text{Prob}(y_i=1|\mathbf{x}_i)$.

Probit model, which in compact form can be written as:

$$\begin{aligned} \mathbf{y}_i^* &= \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i, \qquad \boldsymbol{\varepsilon}_i \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}) \\ y_{mi} &= \begin{cases} 1 & \text{if} \quad y_{mi}^* > 0 \\ 0 & \text{if} \quad y_{mi}^* \leq 0 \end{cases} \qquad \forall m = 1, 2 \ldots, M \end{aligned} \tag{6.40}$$

where:

$$\underset{M\times 1}{\mathbf{y}_i^*} = \begin{bmatrix} y_{1i}^* \\ y_{2i}^* \\ \vdots \\ y_{Mi}^* \end{bmatrix}, \qquad \underset{M\times K}{\mathbf{X}_i} = \begin{bmatrix} \mathbf{x}_{1i}' & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{x}_{2i}' & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \mathbf{x}_{Mi}' \end{bmatrix}, \qquad \underset{K\times 1}{\boldsymbol{\beta}} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_M \end{bmatrix} \quad \text{and} \quad \underset{M\times 1}{\boldsymbol{\varepsilon}_i} = \begin{bmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \\ \vdots \\ \varepsilon_{Mi} \end{bmatrix}$$

There are two important notes to make about the specification of the model:

1. We specify the model in terms of the covariance matrix, $\boldsymbol{\Sigma}$, of $\boldsymbol{\varepsilon}_i$, not in terms of the precision matrix, $\boldsymbol{\Omega} \equiv \boldsymbol{\Sigma}^{-1}$. This is done because it simplifies the estimation process and the interpretation of the results.
2. There is no direct counterpart to a multivariate Logit model. Although the logistic distribution can be generalized in multiple dimensions (Malik & Abraham, 1973), this generalization restricts the covariance matrix of $\boldsymbol{\varepsilon}_i$ to a constant, instead of allowing estimation of its elements. Nevertheless, some workarounds have been proposed in the literature (Carey et al., 1993; Glonek & McCullagh, 1995)

The multivariate Probit model was introduced by Ashford & Sowden (1970), who proposed estimation by maximum likelihood and Amemiya (1974) considered alternative methods of estimation. Both papers considered only frequentist methods and, due to the computational complexity of the model, restricted attention to bivariate Probit models ($M = 2$). Chib & Greenberg (1998) proposed Bayesian estimation of the model, which, through data augmentation, can accommodate any number of response variables.

An equivalent representation of the multivariate Probit model is:

$$\begin{aligned} \mathrm{Prob}(\mathbf{y}_i = \mathbf{z}|\mathbf{X}_i) &= \int_{\mathcal{A}_{Mi}} \cdots \int_{\mathcal{A}_{1i}} \frac{|\boldsymbol{\Sigma}|^{-1/2}}{(2\pi)^{M/2}} \exp\left\{-\frac{1}{2}(\mathbf{y}_i^* - \mathbf{X}_i\boldsymbol{\beta})' \boldsymbol{\Sigma}^{-1} (\mathbf{y}_i^* - \mathbf{X}_i\boldsymbol{\beta})\right\} \mathrm{d}\mathbf{y}_i^* \\ \mathcal{A}_{mi} &= \begin{cases} (0, +\infty) & \text{if} \quad z_m = 1 \\ (-\infty, 0] & \text{if} \quad z_m = 0 \end{cases} \qquad \forall m = 1, 2 \ldots, M \end{aligned} \tag{6.41}$$

As complex as this expression may appear, it is nothing more than the multivariate generalization of (6.5). Due to (6.40), $\boldsymbol{\varepsilon}_i = \mathbf{y}_i^* - \mathbf{X}_i\boldsymbol{\beta}$ follows a multivariate-Normal distribution, the density of which appears inside the M-dimensional integral. Because each ε_{mi} is positive whenever $z_m = 1$ and negative otherwise, the range of integration in each dimension m is restricted accordingly. Despite the intuition behind (6.41), the multidimensional integral does not have a closed form and calculating its value is challenging, but can be accomplished by the Geweke-Hajivassiliou-Keane (*GHK*) *algorithm* (Geweke, 1991; Hajivassiliou et al., 1996; Keane, 1994).[7] Nevertheless, when data augmentation is used, estimation can be based on the latent-variable representation of the model and there is no need to evaluate this integral. Evaluation, however, is necessary for making predictions for the probabilities of joint events.

It may appear that we now have all necessary ingredients for constructing a Gibbs sampler. The first step in an iteration of the sampler would be to draw $\mathbf{y}_i^*$ from a multivariate-Normal distribution, truncated appropriately given the values of $\mathbf{z}$, as we did in the univariate case. In the second step we would sample for $\boldsymbol{\beta}$ and $\boldsymbol{\Omega} \equiv \boldsymbol{\Sigma}^{-1}$ using the complete conditionals

[7] The GHK algorithm uses simulation to approximate the probability of a multivariate-Normally distributed random vector assuming a value in a linearly restricted space. It does this by expressing the joint probability $\mathrm{Prob}(\mathbf{y}_i = \mathbf{z}|\mathbf{X}_i)$ as a series of conditional probabilities, thus requiring drawing from truncated-Normal distributions in one dimension at a time. The algorithm itself is rather complex and we will not cover it here.

that we developed for the SUR model. However, as the model was presented until now, we cannot uniquely identify the magnitudes of the parameters. In particular, if instead of $(\boldsymbol{\beta}, \boldsymbol{\Sigma})$ we use the parameterization $\left(\breve{\boldsymbol{\beta}}, \breve{\boldsymbol{\Sigma}}\right)$, then $\text{Prob}(\mathbf{y}_i|\mathbf{X}_i, \boldsymbol{\beta}, \boldsymbol{\Sigma}) = \text{Prob}\left(\mathbf{y}_i\middle|\mathbf{X}_i, \breve{\boldsymbol{\beta}}, \breve{\boldsymbol{\Sigma}}\right)$ whenever $\boldsymbol{\beta}_m = \breve{\sigma}_{mm}^{-1/2}\breve{\boldsymbol{\beta}}_m$ and $\boldsymbol{\Sigma} = \mathbf{D}\breve{\boldsymbol{\Sigma}}\mathbf{D}$, where $\mathbf{D}$ is a diagonal matrix with $\breve{\sigma}_{mm}^{-1/2}$ on the m^{th} position on the diagonal. We encountered the same issue in the univariate Probit model and we resolved it by restricting the variance of ε_i to unity. In the multivariate case we need to impose a similar normalization for each of the response variables. To start with, note that due to $\mathbf{y}_i^*$ following a multivariate-Normal distribution, each y_{mi}^* follows, marginally with respect to the remaining $y_{\ell i}^*$s, a univariate-Normal distribution. If we want the multivariate Probit model to be a true generalization of the univariate Probit, then we need to set all the diagonal elements of $\boldsymbol{\Sigma}$ to one. Thus, the obvious way to resolve the non-identification issue is to scale $\boldsymbol{\Sigma}$ in such a way that it becomes a proper correlation matrix. The problem with this normalization is that there is no conjugate prior for a correlation matrix and implementing a Metropolis-Hastings step for $\boldsymbol{\Sigma}$ inside the Gibbs sampler is complicated by the fact that $\boldsymbol{\Sigma}$ must remain positive semi-definite in every iteration of the sampler.

Three main types of approaches can be used to construct a Gibbs sampler for the multivariate Probit model. The first one is to ignore the non-identification issue and treat $\boldsymbol{\Sigma}$ as a general covariance matrix when sampling from the complete conditionals of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$, but normalize the draws before reporting the moments of the posterior distribution. Chib & Greenberg (1998) argue against this approach, as it can quickly lead to numerical instability and convergence problems, unless the prior on $\boldsymbol{\Sigma}$ is rather informative. They instead advocate using a Metropolis-Hastings step for updating $\boldsymbol{\Sigma}$, with the draw rejected whenever the proposed move results in a non-positive-definite matrix. X. Liu & Daniels (2006) review the relevant literature and develop an algorithm for sampling from the full conditional of a correlation matrix using parameter expansion (see Section 4.4). We will present here a similar approach, although using a setup and terminology that is closer to marginal data augmentation, rather than parameter expansion.

We start by defining $\breve{\boldsymbol{\Sigma}}$ as an $M \times M$ positive-definite, but otherwise unrestricted matrix. We then place an inverse-Wishart prior on it, with degrees-of-freedom parameter n and scale matrix $\mathbf{S}$:

$$\text{p}\left(\breve{\boldsymbol{\Sigma}}\right) = \frac{|\breve{\boldsymbol{\Sigma}}|^{-\frac{n+M+1}{2}}|\mathbf{S}|^{n/2}}{2^{nM/2}\Gamma_M\left(\frac{n}{2}\right)} \exp\left\{-\frac{\text{tr}\left(\mathbf{S}\breve{\boldsymbol{\Sigma}}^{-1}\right)}{2}\right\} \tag{6.42}$$

This prior is equivalent to imposing a Wishart prior on $\boldsymbol{\Omega} \equiv \boldsymbol{\Sigma}^{-1}$ with the same degrees-of-freedom parameter and scale matrix $\mathbf{S}^{-1}$. We then transform from $\breve{\boldsymbol{\Sigma}}$ to $(\boldsymbol{\alpha}, \boldsymbol{\Sigma})$ by defining $\boldsymbol{\Sigma} = \mathbf{D}_{\boldsymbol{\alpha}}\breve{\boldsymbol{\Sigma}}\mathbf{D}_{\boldsymbol{\alpha}}$, where $\boldsymbol{\alpha} = \begin{bmatrix}\breve{\sigma}_{11}^{-1/2} & \breve{\sigma}_{22}^{-1/2} & \cdots & \breve{\sigma}_{MM}^{-1/2}\end{bmatrix}'$ and $\mathbf{D}_{\boldsymbol{\alpha}}$ is a diagonal matrix constructed by putting the elements of $\boldsymbol{\alpha}$ on the diagonal. In a marginal data-augmentation setup $\boldsymbol{\alpha}$ plays the role of the working parameter and the $\boldsymbol{\Sigma}$ matrix constructed in this way is always a correlation matrix. It can be shown that the determinant of the Jacobian of the transformation from $\breve{\boldsymbol{\Sigma}}$ to $(\boldsymbol{\alpha}, \boldsymbol{\Sigma})$ is equal to $\frac{1}{2^M}\prod_{m=1}^M \alpha_m^{M+2}$ and an application of the multivariate change-of-variables theorem leads to the following probability density function for $(\boldsymbol{\alpha}, \boldsymbol{\Sigma})$:

$$\text{p}(\boldsymbol{\alpha}, \boldsymbol{\Sigma}) = \frac{|\boldsymbol{\Sigma}|^{-\frac{n+M+1}{2}}|\mathbf{S}|^{n/2}\prod_{m=1}^M \alpha_m^{n-1}}{2^{\frac{(n-2)M}{2}}\Gamma_M\left(\frac{n}{2}\right)} \exp\left\{-\frac{\text{tr}\left(\mathbf{S}\mathbf{D}_{\boldsymbol{\alpha}}\boldsymbol{\Sigma}^{-1}\mathbf{D}_{\boldsymbol{\alpha}}\right)}{2}\right\} \tag{6.43}$$

In the specific case where $\mathbf{S}$ is diagonal, $\boldsymbol{\alpha}$ can be integrated-out from this density analytically, leading to a prior density for $\boldsymbol{\Sigma}$ of the form:

$$\text{p}(\boldsymbol{\Sigma}) = |\boldsymbol{\Sigma}|^{-\frac{n+M+1}{2}} \times \prod_m \left(\boldsymbol{\Sigma}_{[mm]}^{-1}\right)^{-\frac{n}{2}} \times \frac{\Gamma^M\left(\frac{n}{2}\right)}{\Gamma_M\left(\frac{n}{2}\right)} \times \mathbb{1}\big(\text{diag}(\boldsymbol{\Sigma}) = \mathbf{1}\big) \tag{6.44}$$

where $\boldsymbol{\Sigma}_{[mm]}^{-1}$ is the m^{th} diagonal element of $\boldsymbol{\Sigma}^{-1}$. In short, the prior imposed on $\breve{\boldsymbol{\Sigma}}$ in (6.42) implies the density in (6.44) when $\mathbf{S}$ is diagonal. Notice that in this case, the magnitudes of

the elements of $\mathbf{S}$ have no impact on the prior for $\boldsymbol{\Sigma}$. On the contrary, very few simplifications can be made to (6.43) when $\mathbf{S}$ is not diagonal.

The marginal data augmentation setup of the model starts from the original modeling assumption: $\mathbf{y}_i^* = \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i$, with $\boldsymbol{\varepsilon}_i \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma})$. Pre-multiplying both sides of the system of equations by $\mathbf{D}_{\boldsymbol{\alpha}}^{-1}$ leads to $\breve{\mathbf{y}}_i^* \equiv \mathbf{D}_{\boldsymbol{\alpha}}^{-1}\mathbf{y}_i^* = \mathbf{D}_{\boldsymbol{\alpha}}^{-1}\mathbf{X}_i\boldsymbol{\beta} + \mathbf{D}_{\boldsymbol{\alpha}}^{-1}\boldsymbol{\varepsilon}_i$ with $\boldsymbol{\varepsilon}_i \sim \mathrm{N}\left(\mathbf{0}, \mathbf{D}_{\boldsymbol{\alpha}}^{-1}\boldsymbol{\Sigma}\mathbf{D}_{\boldsymbol{\alpha}}^{-1}\right)$, where $\mathbf{D}_{\boldsymbol{\alpha}}^{-1}\boldsymbol{\Sigma}\mathbf{D}_{\boldsymbol{\alpha}}^{-1} = \breve{\boldsymbol{\Sigma}}$ is a positive-definite matrix on which we impose the inverse-Wishart prior in (6.42). Finally, we define $\mathbf{D}_K$ as a $K \times K$ diagonal matrix, constructed by replicating each α_m as many times as the number of independnet variables in equation m:[8]

$$\mathbf{D}_K = \begin{bmatrix} \alpha_1\mathbf{I}_{K_1} & \mathbf{0}_{K_1\times K_2} & \cdots & \mathbf{0}_{K_1\times K_M} \\ \mathbf{0}_{K_1\times K_1} & \alpha_2\mathbf{I}_{K_2} & \cdots & \mathbf{0}_{K_2\times K_M} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_{K_M\times K_1} & \mathbf{0}_{K_M\times K_2} & \cdots & \alpha_M\mathbf{I}_{K_M} \end{bmatrix} \tag{6.45}$$

Then $\mathbf{D}_{\boldsymbol{\alpha}}^{-1}\mathbf{X}_i\boldsymbol{\beta} = \mathbf{X}_i\mathbf{D}_K^{-1}\boldsymbol{\beta}$ and the system of equations can be written as:

$$\breve{\mathbf{y}}_i^* = \mathbf{X}_i\breve{\boldsymbol{\beta}} + \breve{\boldsymbol{\varepsilon}}_i, \qquad \breve{\boldsymbol{\varepsilon}}_i \sim \mathrm{N}\left(\mathbf{0}, \breve{\boldsymbol{\Sigma}}\right) \tag{6.46}$$

where $\breve{\boldsymbol{\beta}} \equiv \mathbf{D}_K^{-1}\boldsymbol{\beta}$.

The main steps of the Gibbs sampler with marginal data augmentation, at a very abstract level, are:

(a) draw $(\{\mathbf{y}_i^*\}, \boldsymbol{\alpha})$ from the transition kernel, $\mathscr{K}$, of a Markov chain with stationary distribution $\mathrm{p}\left(\{\mathbf{y}_i^*\}, \boldsymbol{\alpha} \middle| \{\mathbf{y}_i\}, \boldsymbol{\beta}, \boldsymbol{\Sigma}\right)$ and transform to $\breve{\mathbf{y}}_i^*$; this is further broken into the steps:

- **(a1)** draw $\{\mathbf{y}_i^*\}$ from $\mathrm{p}\left(\{\mathbf{y}_i^*\} \middle| \{\mathbf{y}_i\}, \boldsymbol{\beta}, \boldsymbol{\Sigma}\right)$
- **(a2)** draw $\boldsymbol{\alpha}$ from $\mathrm{p}\left(\boldsymbol{\alpha} \middle| \boldsymbol{\Sigma}\right)$
- **(a3)** transform each $\mathbf{y}_i^*$ to $\breve{\mathbf{y}}_i^* = \mathbf{D}_{\boldsymbol{\alpha}}^{-1}\mathbf{y}_i^*$

(b) draw $\breve{\boldsymbol{\beta}}$ from $\mathrm{p}\left(\breve{\boldsymbol{\beta}} \middle| \{\breve{\mathbf{y}}_i^*\}, \boldsymbol{\Sigma}, \boldsymbol{\alpha}\right)$ and transform to $\boldsymbol{\beta}$; this is further broken into the following steps:

- **(b1)** draw $\breve{\boldsymbol{\beta}}$ from $\mathrm{p}\left(\breve{\boldsymbol{\beta}} \middle| \{\breve{\mathbf{y}}_i^*\}, \boldsymbol{\Sigma}, \boldsymbol{\alpha}\right)$
- **(b2)** construct $\mathbf{D}_K$, transform $\breve{\boldsymbol{\beta}}$ to $\boldsymbol{\beta} = \mathbf{D}_K\breve{\boldsymbol{\beta}}$ and store this $\boldsymbol{\beta}$

(c) draw $\breve{\boldsymbol{\Sigma}}$ from $\mathrm{p}\left(\breve{\boldsymbol{\Sigma}} \middle| \{\breve{\mathbf{y}}_i^*\}, \breve{\boldsymbol{\beta}}\right)$ and transform; this is further broken into the following steps:

- **(c1)** draw $\breve{\boldsymbol{\Sigma}}$ from $\mathrm{p}\left(\breve{\boldsymbol{\Sigma}} \middle| \{\breve{\mathbf{y}}_i^*\}, \breve{\boldsymbol{\beta}}\right)$ and construct $\mathbf{D}_{\boldsymbol{\alpha}}$ from the elements of $\breve{\boldsymbol{\Sigma}}$
- **(c2)** set $\boldsymbol{\Sigma} = \mathbf{D}_{\boldsymbol{\alpha}}\breve{\boldsymbol{\Sigma}}\mathbf{D}_{\boldsymbol{\alpha}}$ and store this $\boldsymbol{\Sigma}$
- **(c3)** construct $\mathbf{D}_K$ and transform each $\breve{\mathbf{y}}_i^*$ to $\mathbf{y}_i^* = \mathbf{D}_{\boldsymbol{\alpha}}\breve{\mathbf{y}}_i^*$ and $\breve{\boldsymbol{\beta}}$ to $\boldsymbol{\beta} = \mathbf{D}_K\breve{\boldsymbol{\beta}}$

If we use a multivariate-Normal prior for $\boldsymbol{\beta}$ with mean $\mathbf{m}$ and precision $\mathbf{P}$, almost all densities from which we need to sample are standard:

1. The complete conditional of each $\mathbf{y}_i^*$ in step **(a1)** is a multivariate-Normal distribution with mean $\mathbf{X}_i\boldsymbol{\beta}$ and variance $\boldsymbol{\Sigma}$, truncated from below at zero in the dimensions for which $y_{mi} = 1$ and from above at zero in the dimensions for which $y_{mi} = 0$. Although challenging, sampling from the complete conditional of each $\mathbf{y}_i^*$ can be done using either rejection or Gibbs sampling.[9]

[8] If all M equations have the same number of independent variables, K_m, then $\mathbf{D}_K$ can be expressed as $\mathbf{D}_{\boldsymbol{\alpha}} \otimes \mathbf{I}_{K_m}$.

[9] Rejection sampling consists of nothing more than repeatedly sampling from the respective unrestricted multivariate-Normal distribution until the resulting draw for $\mathbf{y}_i^*$ is in the required range. The probability of generating a draw in the required range depends on M, the current value of $\mathbf{X}_i\boldsymbol{\beta}$ in the Gibbs sampler and $\mathbf{y}_i$, and could be very low. Thus, a pure rejection sampling procedure could become very inefficient, with a vast number of draws generated before one is accepted. Gibbs sampling generates a draw from the complete conditional of $\mathbf{y}_i^*$ by sampling from a univariate truncated-Normal distribution, given the values of $\mathbf{y}_i^*$ in all remaining dimensions.

2. The complete conditional of $\breve{\boldsymbol{\beta}}$ in step **(b1)** is a multivariate Normal with precision $\tilde{\mathbf{P}} = \sum_{i=1}^{N} \mathbf{X}_i' \mathbf{D}_{\boldsymbol{\alpha}} \boldsymbol{\Sigma}^{-1} \mathbf{D}_{\boldsymbol{\alpha}} \mathbf{X}_i + \mathbf{D}_K \mathbf{P} \mathbf{D}_K$ and mean $\tilde{\boldsymbol{\beta}} = \tilde{\mathbf{P}}^{-1} \left(\sum_{i=1}^{N} \mathbf{X}_i' \mathbf{D}_{\boldsymbol{\alpha}} \boldsymbol{\Sigma}^{-1} \mathbf{D}_{\boldsymbol{\alpha}} \breve{\mathbf{y}}_i^* + \mathbf{D}_K \mathbf{P} \mathbf{m} \right)$.
3. The complete conditional of $\breve{\boldsymbol{\Sigma}}$ in step **(c1)** is an inverse-Wishart with degrees-of-freedom parameter $\tilde{n} = N + n$ and scale matrix $\tilde{\mathbf{S}} = \sum_{i=1}^{N} \left(\breve{\mathbf{y}}_i^* - \mathbf{X}_i \breve{\boldsymbol{\beta}} \right) \left(\breve{\mathbf{y}}_i^* - \mathbf{X}_i \breve{\boldsymbol{\beta}} \right)' + \mathbf{S}$.

The only challenging part is sampling from $\mathrm{p}\left(\boldsymbol{\alpha} \middle| \boldsymbol{\Sigma}\right)$ in step **(a2)**. However, if the scale matrix in the prior for $\breve{\boldsymbol{\Sigma}}$ is diagonal, then each element of $\boldsymbol{\alpha}$ follows a *Nakagami-m* distribution, which is easy to sample from. If this scale matrix is not diagonal, we need to resort to a Metropolis-Hastings step.

The following example uses the multivariate Probit model to examine the effect of individual-level socioeconomic and demographic characteristics on the probability of three outcomes: investing in the stock market, having other types of savings to use in retirement, and being a homeowner.

♦ **Example 6.3 Saving and Investing**

In this example we will use part of the dataset constructed through the 2017 Cooperative Congressional Election Study (Schaffner & Ansolabhere, 2019). The original dataset contains 18,200 observations (US-based individuals above the voting age) and information on 267 variables. We will use here only 5,000 randomly drawn observations from this dataset and the following variables:

`investor`	:	dummy variable, 1 if the respondent has money invested in the stock market
`savings`	:	dummy variable, 1 if the respondent has savings accounts, pensions, or other investments that they could use in retirement
`homeowner`	:	dummy variable: 1 if the respondent owns their own home
`age`	:	age of the respondent in years
`educ`	:	educational level of the respondent, coded from 1 to 6 with larger values corresponding to higher educational level
`male`	:	dummy variable: 1 if the respondent is male
`faminc`	:	annual family income bracket, coded from 1 (less that \$10,000) to 16 (more than \$500,000) and increasing by \$10,000 when income is bellow \$100,000 and by larger amounts after that

We will model the state of the first three variables above (0 or 1) by using the model:

$$\begin{aligned} \text{investor}_i^* &= \beta_{11} + \beta_{12}\text{age}_i + \beta_{13}\text{educ}_i + \beta_{14}\text{male}_i + \beta_{15}\text{faminc}_i + \varepsilon_{1i} \\ \text{savings}_i^* &= \beta_{21} + \beta_{22}\text{age}_i + \beta_{23}\text{educ}_i + \beta_{24}\text{male}_i + \beta_{25}\text{faminc}_i + \varepsilon_{2i} \\ \text{homeowner}_i^* &= \beta_{31} + \beta_{32}\text{age}_i + \beta_{33}\text{educ}_i + \beta_{34}\text{male}_i + \beta_{35}\text{faminc}_i + \varepsilon_{3i} \end{aligned}$$

where the three variables in the left-hand side of the equations can be thought of as the differences in utilities that individual i derives from using the respective method of saving or not. We will also assume that the ε_{mi}s follow a multivariate-Normal distribution and allow them to be correlated: $\boldsymbol{\varepsilon}_i \sim \mathrm{N}\left(\mathbf{0}, \boldsymbol{\Sigma}\right)$, where $\boldsymbol{\varepsilon}_i \equiv \begin{bmatrix} \varepsilon_{1i} & \varepsilon_{2i} & \varepsilon_{3i} \end{bmatrix}'$ and $\boldsymbol{\Sigma}$ is scaled such that its diagonal elements are equal to one. Taken together, these assumptions lead to a multivariate Probit model with $M = 3$. Note that we use here the same independent variables in all three equations, This is not necessary, however, and we could use different sets of variables per equation, disjoint or not. Also, because $\boldsymbol{\varepsilon}_i$ follows a multivariate-Normal distribution, each of the error terms follows, marginally with respect to the remaining $\varepsilon_{\ell i}$s, a univariate-Normal distribution with mean zero an variance one. Thus, we could instead estimate the βs in the model above using three separate univariate Probit models. This approach, however, cannot provide estimates of the off-diagonal elements of $\boldsymbol{\Sigma}$.

The results obtained using BayES' `mvprobit()` function to estimate the multivariate-Probit model are presented in the following table. Except for the parameters associated with the individual's educational level and gender in the equation where the response variable captures whether the respondent is a homeowner or not, all other slope parameters are positive and their 90% credible intervals do not contain zero. Because each latent dependent variable follows, marginally with respect to the other, a univariate-Normal distribution we can interpret the signs of these parameters. For example, older individuals are more likely to invest in the stock market, as well as to have other forms of savings

and to own a house. However, as the probability of success in each of the three response variables is a non-linear function of the parameters, we should refrain from interpreting the magnitudes of these parameters.

	Mean	Median	Sd.dev.	5%	95%
investor					
constant	-2.49983	-2.49978	0.0875499	-2.64399	-2.35776
age	0.0109609	0.0109589	0.00115749	0.00906729	0.012877
educ	0.149862	0.149812	0.0142048	0.126563	0.17347
male	0.273818	0.273965	0.0392544	0.209114	0.338338
faminc	0.158985	0.158981	0.00668294	0.148021	0.170084
savings					
constant	-2.21019	-2.21015	0.0877345	-2.35536	-2.06527
age	0.0162371	0.0162379	0.00118512	0.0142868	0.0181883
educ	0.148181	0.148172	0.0149679	0.123253	0.172899
male	0.297574	0.297704	0.0402341	0.231492	0.363458
faminc	0.184913	0.184829	0.00715524	0.173186	0.196749
homeowner					
constant	-1.77696	-1.7766	0.0819438	-1.91279	-1.64407
age	0.0262803	0.0262838	0.00118565	0.0243446	0.0282255
educ	0.0127839	0.0126896	0.0143497	-0.0105772	0.0368126
male	-0.00951553	-0.00984216	0.0398539	-0.0747032	0.0557755
faminc	0.140924	0.140939	0.00679899	0.129771	0.152214

We note in passing that the posterior expectation of $\boldsymbol{\Sigma}$ is:

$$\mathrm{E}\left(\boldsymbol{\Sigma}\middle|\left\{\mathbf{y}_i\right\}\right) = \begin{bmatrix} 1 & 0.68381436 & 0.25318785 \\ 0.68381436 & 1 & 0.24996288 \\ 0.25318785 & 0.24996288 & 1 \end{bmatrix}$$

which clearly shows that the error terms in the three equations are positively correlated. This implies that unobserved factors at the individual level tend to affect the probability of success in the three response variables in the same direction.

Obtaining the results presented above using BayES can be achieved using the code in the following box.

```
// import the data into a dataset called Data
Data = webimport("www.bayeconsoft.com/datasets/CCES2017.csv");

// generate a constant term
Data.constant = 1;

// run the model
myMVProbit = mvprobit( {
    investor ~ constant age educ male faminc,
    savings ~ constant age educ male faminc,
    homeowner ~ constant age educ male faminc
    } );

// print the posterior mean of Sigma
print(myMVProbit.Sigma);
```

As in the univariate Probit and Logit models, due to the probability of success being a non-linear function of the independent variables and the parameters, in multivariate Probit models we can interpret the signs of these parameters, but not their magnitudes. We can interpret, however, the magnitudes of marginal effects of the form:

$$\frac{\partial \operatorname{Prob}(y_{mi}=1|\mathbf{X}_i)}{\partial x_{\ell ik}} \tag{6.47}$$

where $x_{\ell ik}$ is the k^{th} independent variable in equation ℓ. ℓ could be the same as m, in which case

we evaluate the marginal effect of a variable that appears in the m^{th} equation on the probability of success for the m^{th} response variable, or it could be different from m. There are two types of effects that we could calculate with respect to an independent variable, $x_{\ell ik}$, both of which could be of interest in the context of the application: the effect on the probability of $y_{mi}=1$ marginally with respect to the remaining response variables or conditional on them being equal to 0 or 1. Because the distribution of each y^*_{mi} marginally with respect the remaining $y^*_{\ell i}$s is univariate Normal, the first type of marginal effects is straightforward to calculate:

$$\frac{\partial \operatorname{Prob}(y_{mi}=1|\mathbf{X}_i)}{\partial x_{\ell ik}} = \begin{cases} \left[(2\pi)^{-1/2}\exp\left\{-\frac{\left(\mathbf{x}'_{mi}\boldsymbol{\beta}_m\right)^2}{2}\right\}\right]\cdot\beta_{mk} & \text{if} \quad \ell = m \\ 0 & \text{if} \quad \ell \neq m \end{cases} \tag{6.48}$$

Notice that the upper branch of this expression is the same as the marginal effect of a variable in the univariate Probit model, as presented in equation (6.27). In this type of marginal effects only the independent variables in equation m affect the probability of y_{mi} being equal to one.

Calculating marginal effects on $\operatorname{Prob}(y_{mi}=1|\mathbf{X}_i)$ conditional on the values of the remaining response variables requires some additional transformations. To start with, let $\mathbf{z}_{/m}$ be an $(M-1)\times 1$ vector with entries equal to zero for the y_ℓs that are to be restricted to zero when conditioning and entries equal to one for the y_ℓs that are to be restricted to one. Notice that $\mathbf{z}_{/m}$ does not contain a value for y_{mi} itself because this is allowed, at least conceptually, to change from zero to one when calculating the marginal effect. For example, when $M=4$ and $m=2$, $\mathbf{z}_{/2}=\begin{bmatrix}1 & 0 & 0\end{bmatrix}'$ is taken to mean that we are conditioning on $y_{1i}=1$ and $y_{3i}=y_{4i}=0$ while calculating the marginal effect on $\operatorname{Prob}(y_{2i}=1|\mathbf{X}_i)$. With this notation, the probability on which we want to calculate marginal effects can be expressed as:

$$\operatorname{Prob}\left(y_{mi}=1|\mathbf{y}_{i/m}=\mathbf{z}_{/m}\right) = \frac{\operatorname{Prob}\left(y_{mi}=1,\mathbf{y}_{i/m}=\mathbf{z}_{/m}\right)}{\operatorname{Prob}\left(\mathbf{y}_{i/m}=\mathbf{z}_{/m}\right)} \tag{6.49}$$

where $\mathbf{y}_{i/m}$ is an $(M-1)\times 1$ random vector that contains the random variables in $\mathbf{y}_i$ except y_{mi}. This expression for the example given above becomes:

$$\begin{aligned} \operatorname{Prob}\left(y_{2i}=1\middle|y_{1i}=1,y_{3i}=0,y_{4i}=0\right) &= \frac{\operatorname{Prob}(y_{2i}=1,y_{1i}=1,y_{3i}=0,y_{4i}=0)}{\operatorname{Prob}(y_{1i}=1,y_{3i}=0,y_{4i}=0)} \\ &= \frac{\operatorname{Prob}(y^*_{2i}>0,y^*_{1i}>0,y^*_{3i}\leq 0,y^*_{4i}\leq 0)}{\operatorname{Prob}(y^*_{1i}>0,y^*_{3i}\leq 0,y^*_{4i}\leq 0)} \end{aligned} \tag{6.50}$$

The probabilities in both the numerator and denominator of the last fraction are in the form of (6.41). Conditional probabilities of the form presented here, however, are easier to approximate directly by the *GHK algorithm*, rather that approximating the ratio of joint probabilities. This is because the algorithm itself works by successive conditioning of joint probabilities.

The expression in (6.49) can be used to make predictions about the value of the probability on the left-hand side. The marginal effect itself can be approximated using finite differences. One thing to notice here is that, if an independent variable appears in the equation for y^*_{mi} as well as in the equation for at least another $y^*_{\ell i}$, then a change in this variable will affect the probabilities in both the numerator and denominator. Furthermore, even if an independent variable does not appear in the equation for y^*_{mi}, it can still affect $\operatorname{Prob}\left(y_{mi}=1|\mathbf{y}_{i/m}=\mathbf{z}_{/m}\right)$ through its effect on $\operatorname{Prob}\left(\mathbf{y}_{i/m}=\mathbf{z}_{/m}\right)$ in the denominator of (6.49).

We now turn back to Example 6.3, where we used a multivariate Probit model to examine the effect of socioeconomic and demographic characteristics on the probability of investing in the stock market, having other types of savings, and being a homeowner. In the following part of this example we calculate both types of marginal effects presented above.

♦ **Example 6.3 Saving and Investing (Continued)**
We will use here again part of the dataset constructed through the 2017 Cooperative Congressional Election Study (Schaffner & Ansolabhere, 2019) and the multivariate-Probit model:

$$\begin{aligned} \mathsf{investor}^*_i &= \beta_{11}+\beta_{12}\mathsf{age}_i+\beta_{13}\mathsf{educ}_i+\beta_{14}\mathsf{male}_i+\beta_{15}\mathsf{faminc}_i+\varepsilon_{1i} \\ \mathsf{savings}^*_i &= \beta_{21}+\beta_{22}\mathsf{age}_i+\beta_{23}\mathsf{educ}_i+\beta_{24}\mathsf{male}_i+\beta_{25}\mathsf{faminc}_i+\varepsilon_{2i} \\ \mathsf{homeowner}^*_i &= \beta_{31}+\beta_{32}\mathsf{age}_i+\beta_{33}\mathsf{educ}_i+\beta_{34}\mathsf{male}_i+\beta_{35}\mathsf{faminc}_i+\varepsilon_{3i} \end{aligned}$$

The marginal effects of each independent variable on the probability of success for each response variable, marginally with respect to the values of the remaining response variables are presented in the following table. These marginal effects are calculated at the sample means of the independent variables. We can see from this table, for example, that an increase of the age of an individual by a year increases the probability of this individual being an investor by 0.41%, having other forms of savings by 0.60%, and being a homeowner by 0.96%. Similarly, being male increases the probability of investing in the stock market by 10.3% and the probability of having other forms of savings by 11.0%, while it reduces the probability of home ownership by 0.35%, although the 90% credible interval for the last marginal effect contains zero. Given that each latent dependent variable in the multivariate Probit model follows, marginally with respect to the remaining latent dependent variables a univariate Normal distribution, the values of these marginal effects should be close to those obtained by running three independent univariate Probit models.

	Mean	Median	Sd.dev.	5%	95%
dProb(y1=1)/dx					
age	0.00412872	0.00412792	0.000435779	0.00341427	0.00484702
educ	0.0564497	0.0564375	0.00534948	0.0476976	0.0653133
*male	0.103289	0.103315	0.0147816	0.0789707	0.127619
faminc	0.0598866	0.0598741	0.00251633	0.0557611	0.0640417
dProb(y2=1)/dx					
age	0.00604344	0.00604502	0.000440553	0.00532363	0.00677011
educ	0.0551519	0.0551641	0.00555627	0.0459007	0.0642902
*male	0.109791	0.109865	0.0146639	0.0855691	0.133709
faminc	0.0688234	0.0688014	0.00262152	0.0645015	0.073151
dProb(y3=1)/dx					
age	0.00964734	0.00964778	0.000432773	0.00893693	0.01036
educ	0.00469271	0.00465784	0.00526681	-0.0038771	0.0135131
*male	-0.00350548	-0.00361579	0.0146345	-0.0274549	0.020458
faminc	0.051732	0.051731	0.00247665	0.0476768	0.0558198

*Marginal effect is calculated for discrete change from 0 to 1.

Using a multivariate Probit model instead of multiple univariate Probit ones can reveal more information on the interrelationship of the response variables and the marginal effects. The following table presents marginal effects of the independent variables in the model on the probability of success of each dependent variable, conditional on the remaining response variables being equal to one. For example, the first block of the table contains marginal effects of the form:

$$\frac{\partial \operatorname{Prob}\left(y_{1i}=1 \middle| y_{2i}=1, y_{3i}=1, \mathbf{X}_i\right)}{\partial x_{\ell ik}}$$

for all independent variables, $x_{\ell ik}$, in the model.

	Mean	Median	Sd.dev.	5%	95%
dProb(y1=1)/dx					
age	0.00136771	0.0013865	0.00161165	-0.00133307	0.0039956
educ	0.0449023	0.0449377	0.0216666	0.00961775	0.0806608
*male	0.0795244	0.0796626	0.0222757	0.0428292	0.115982
faminc	0.037942	0.0379786	0.0125245	0.017199	0.0582635
dProb(y2=1)/dx					
age	0.00184107	0.00183701	0.000482143	0.00105529	0.00264311
educ	0.0132836	0.0132483	0.00626243	0.00309367	0.0236947
*male	0.0297062	0.0295972	0.00861738	0.0157202	0.0439751
faminc	0.0186869	0.0186443	0.00384257	0.012451	0.0250804
dProb(y3=1)/dx					
age	0.00763243	0.00763314	0.000543023	0.00673871	0.00852946
educ	-0.00552058	-0.00555597	0.00670174	-0.0165713	0.00553625
*male	-0.0208181	-0.0208736	0.0132155	-0.0424709	0.00096068
faminc	0.0345723	0.0346319	0.00383884	0.028212	0.040712

*Marginal effect is calculated for discrete change from 0 to 1.

The results presented above can be obtained in BayES using the code in the following box.

```
// import the data into a dataset called Data
Data = webimport("www.bayeconsoft.com/datasets/CCES2017.csv");

// generate a constant term
Data.constant = 1;

// run the model
myMVProbit = mvprobit( {
    investor ~ constant age educ male faminc,
    savings ~ constant age educ male faminc,
    homeowner ~ constant age educ male faminc
    } );

// print the posterior mean of Sigma
print(myMVProbit.Sigma);
```

6.7 Synopsis

This chapter introduced and covered in detail the models which are most widely-used to uncover the causal effects of a set of independent variables on a qualitative response variable. Attention was restricted to response variables which can be in only one of two states: true/-false, success/failure, 0/1, etc. The statistical setup of these models started with the index-function representation, before showing that the latent-variable representation is an equivalent way of looking at the problem. Binary-choice models were also motivated using the random-utility framework. The approach of estimating binary Probit and Logit models using data augmentation and the latent-variable representation was described next. Interpretation of the models' parameters and marginal effects followed, before extending the models to account for group-specific unobserved heterogeneity when panel data are available. Finally, a multivariate generalization of the binary-Probit model was presented as a simple combination of concepts from univariate-Probit and SUR models.

The models covered in this chapter are extended in the following chapter to accommodate qualitative variables which can be in one of more than two states. The discussion there is heavily based on the ideas and techniques presented in this chapter and the reader will be referred back to the relevant concepts and results on multiple occasions.

Chapter 7

Models for Multiple Discrete Response

7.1 Overview

This chapter extends the binary-response models, introduced in Chapter 6, to the case of response variables that can be in one out of a finite set of states. In these models the response variable is still qualitative and the quantity being modeled is the probability of one outcome occurring or alternative chosen, out of a fixed set of mutually exclusive and collectively exhaustive outcomes/alternatives. As in binary-response models, there are two equivalent representations of the models: one based on directly specifying the probability of an outcome occurring as a function of independent variables and another based on the random-utility framework.

Although discrete-response/*discrete-choice models* with multiple outcomes/alternatives can be viewed as direct extensions to binary-response models, complications arise both in terms of interpretation of the results and in estimation. In this chapter we cover models in which it is either undesirable or impossible to find a criterion according to which the available outcomes/alternatives can be ordered objectively, given the context of the problem. The response variable will be coded using integer values for ease of reference, but this does not mean that, for example, the outcome being assigned code 2 is, in any objective way, preferable to the outcomes assigned codes 0 or 1 or, more generally, that outcome 2 is further away from outcome 0 than outcome 1 is, in any meaningful dimension.[1] Regarding the nature of the independent variables, there are two broad classes of models: those in which the independent variables vary by the unit of analysis, called *multinomial models* and those in which the independent variables vary by outcome and possibly by unit of analysis, called *conditional models*. Conditional models are more general than multinomial, but notation for the former is more cumbersome and interpretation of the results considerably more challenging. Therefore, multinomial models will be used when introducing the mathematical setup of discrete-choice models, as well as when presenting the estimation techniques. Conditional models are presented in Section 7.4, where it is also shown how multinomial models can be viewed as restricted versions of conditional models. Before doing so, however, the following section discusses in more detail the conceptual background behind discrete-response models with more than two outcomes.

7.2 The Nature of Discrete-Response Models

Discrete-response/discrete-choice models are used to draw inferences in problems where the response variable is qualitative and it can be in one out of a fixed number of states. As

[1] Models for which the the response variable has an ordinal interpretation are covered in Chapter **??**.

in binary-response models, which form a subset of discrete-response models, interest revolves around the probability of the response variable being in a specific state and how this probability is affected by the independent variables. A few examples on which multiple discrete-response models can be applied are:

- a customer buys a specific brand of milk during a visit to the grocery store out of three brands being offered at the store, plus the option of not purchasing milk at all (four alternatives in total)
- an employee of a large company commutes to work by car, public transportation, on foot, by bike or works from home (five alternatives in total)
- an individual chooses whether to pursue a college degree in formal sciences, natural sciences, social sciences, or not to go to college at all (four alternatives in total)
- a voter choses to vote for one out of four candidates who run for office or not to vote (five alternatives in total)

Notice that in all these examples the decision maker has to choose one and only one alternative. By saying that the available alternatives are *collectively exhaustive* we mean that the decision maker has to select one of them. This is why the outside option (not buying any brand of milk/not commuting/not going to college/not voting) was included in the choice set. Another way of saying that an individual cannot choose more than one alternatives is to say that the alternatives are *mutually exclusive*. Notice also that in none of the examples is it reasonable to assume that one of the available alternatives is, in any objective way, better than another for all decisions makers. One alternative will be chosen by each decision maker and this, as will assume throughout, is the one that maximizes his/her utility. But which alternative leads to maximum utility depends on the *characteristics* of the individual or the *attributes* of the alternative.

Although the context of individual choice is by far the most widely used in economics and marketing, this is not the only context in which multinomial and conditional models can be used. Indeed, statistical procedures designed to deal with problems where the response variable is categorical can be viewed as general classification methods, having applications in fields like medicine, natural language processing or pattern recognition. For example, a multinomial model can be used to model the probability of an individual having a specific blood type, given the values of relevant characteristics. Even though there is no choice to be made by the individual here, the statistical methods described in this chapter can be applied to this case without any modification. Nevertheless, we will present multinomial and conditional models in the context of individual choice and keep using the associated terms (for example we use the term discrete-choice model rather than discrete response), as this approach helps with the understanding of the underlying concepts.

To provide a general treatment of discrete-choice models that does not depend on the "labels" of the alternatives in the choice set (which are specific to each problem), we will map the alternatives to consecutive integer values. In this setup, each individual has to choose one out of $M+1$ alternatives and we will start numbering alternatives from zero. Using this numbering convention is convenient because it makes apparent that by setting M to one we can revert to binary-choice models. Coded in this way, the response variable for individual i in the population, y_i, can assume any integer value from zero to M. As in binary-response models, because the mapping from alternatives to integers is *ad hoc* (any other mapping is equally valid), it does not make sense to use as the dependent variable in a model y_i directly. Instead, discrete-choice models specify the probability of y_i assuming a particular value in the set $\{0, 1, \ldots, M\}$ and model this probability.

7.3 Multinomial Models

The objective of multinomial models is to express the probability of individual i in the population choosing alternative $m \in \{0, 1, \ldots, M\}$ as a function of individual characteristics and

parameters, and then estimate the associated parameters and marginal effects. Mathematically, the probability of interest is $\text{Prob}(y_i = m|\mathbf{x}_i'\boldsymbol{\beta}_m)$, which we will denote by p_{mi}. This is a conditional probability, with $\mathbf{x}_i$ being a $K\times 1$ vector of independent variables (*individual characteristics*) and $\boldsymbol{\beta}_m$ the corresponding $K\times 1$ vector of parameters. Notice that the vector of parameters is specific to each each alternative, m, but the independent variables vary only by individual (are common to all alternatives). This is the feature that distinguishes multinomial from conditional models, with the independent variables in the latter varying also by alternative. These probabilities enter the probability mass function of y_i, which naturally follows a *categorical distribution*:

$$\text{p}(y_i|\mathbf{x}_i) = \prod_{m=0}^{M} p_{mi}^{\mathbb{1}(y_i=m)} \tag{7.1}$$

where $\mathbb{1}(\cdot)$ is the indicator function. The categorical distribution is a direct extension of the *Bernoulli distribution* to the case of $M > 1$ and, although simple, it is often confused with the *multinomial distribution*. This confusion is largely responsible for calling this class of discrete-choice models *multinomial*, although a more appropriate name would be categorical models.[2]

The specification of a multinomial model is complete once the form of the probabilities, p_{mi}s, is specified as a function of the individual characteristics. For each individual i, these probabilities must be non-negative and sum across alternatives to unity. Similarly to binary-choice models, two specifications are used frequently in practice. The first one is:

$$p_{mi} = \frac{e^{\mathbf{x}_i'\boldsymbol{\beta}_m}}{\sum_{\ell=0}^{M} e^{\mathbf{x}_i'\boldsymbol{\beta}_\ell}}, \qquad m = 0, 1, \ldots, M \tag{7.2}$$

which gives rise to the *multinomial Logit* model. Not all $\boldsymbol{\beta}_m$s can be identified here, because multiplying all $\boldsymbol{\beta}_m$s by the same constant, z, will leave the probabilities unaffected. This issue can be easily resolved by normalizing $\boldsymbol{\beta}_0$ to a vector of zeros, as we implicitly did in the binary Logit model.

In the second specification, which leads to the *multinomial Probit* model, the p_{mi}s are specified as multidimensional integrals on a continuous random vector, $\mathbf{u}$, which is assumed to follow an M-dimensional Normal distribution with mean that depends on $\mathbf{x}_i$ and the $\boldsymbol{\beta}_m$s, and precision matrix $\boldsymbol{\Omega}$:

$$p_{mi} = \begin{cases} \int_{-\infty}^{0}\cdots\int_{-\infty}^{0} \dfrac{|\boldsymbol{\Omega}|^{1/2}}{(2\pi)^{M/2}} \exp\left\{-\dfrac{(\mathbf{u}-\mathbf{X}_i\boldsymbol{\beta})'\,\boldsymbol{\Omega}\,(\mathbf{u}-\mathbf{X}_i\boldsymbol{\beta})}{2}\right\} \text{d}\mathbf{u} & \text{if} \quad m = 0 \\ \int_{\mathscr{A}_M}\cdots\int_{\mathscr{A}_1} \dfrac{|\boldsymbol{\Omega}|^{1/2}}{(2\pi)^{M/2}} \exp\left\{-\dfrac{(\mathbf{u}-\mathbf{X}_i\boldsymbol{\beta})'\,\boldsymbol{\Omega}\,(\mathbf{u}-\mathbf{X}_i\boldsymbol{\beta})}{2}\right\} \text{d}\mathbf{u} & \text{if} \quad m > 0 \end{cases} \tag{7.3}$$

and the range of integration in the second branch is $\mathscr{A}_m = (0, +\infty)$, and $\mathscr{A}_\ell = (-\infty, u_m]$ for $\ell \neq m$. The mean vector of the multivariate-Normal distribution, is the product of:

$$\underset{M\times(M\cdot K)}{\mathbf{X}_i} = \mathbf{I}_M \otimes \mathbf{x}_i' = \begin{bmatrix} \mathbf{x}_i' & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{x}_i' & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{x}_i' \end{bmatrix} \qquad \text{and} \qquad \underset{K\times 1}{\boldsymbol{\beta}} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_M \end{bmatrix}$$

[2] To be fair in this criticism, if the probability of y_i being equal to m is not conditioned on any independent variables, then the p_ms themselves become the parameters of the model and they can be estimated using a multinomial, rather than a categorical distribution. This is because, when the probability of each outcome occurring is the same for all individuals, then the information contained in the data can be summarized completely by the number of observed values in each category. This number is what the multinomial distribution models directly.

Notice that in the way $\boldsymbol{\beta}$ is defined here, it does not contain $\boldsymbol{\beta}_0$ and this is because we normalize it to a vector of zeros for identification purposes. $\boldsymbol{\Omega}$ is an $M \times M$ precision matrix to be estimated. However, as it is the case in the binary Probit model, the parameters of a multinomial Probit model cannot be identified unless some restrictions are imposed on them. This is because, if $\boldsymbol{\beta}$ is multiplied by any positive number, z, and $\boldsymbol{\Omega}$ is multiplied by $z^{-\frac{1}{2}}$, then the probabilities in (7.3) will remain unaffected. This issue is usually resolved by restricting $\boldsymbol{\Omega}$ and we will discuss some approaches for doing so when we present the procedure for estimating the model's parameters.

7.3.1 The Random-Utility Setup and the Latent-Variable Representation

The random-utility setup can be used as an underlying framework for multinomial models in the same way as for binary-choice models. We consider an individual i, who maximizes utility by choosing one out of $M+1$ mutually exclusive and collectively exhaustive alternatives. For each alternative, m, we assume that the unobserved utility that the individual derives from it, u^*_{mi}, is a linear function of this individual's characteristics, $\mathbf{x}_i$, plus statistical noise:

$$\begin{array}{rcl} u^*_{0i} & = & \mathbf{x}_i'\boldsymbol{\gamma}_0 + \epsilon_{0i} \\ u^*_{1i} & = & \mathbf{x}_i'\boldsymbol{\gamma}_1 + \epsilon_{1i} \\ \vdots & & \vdots \qquad\quad \vdots \\ u^*_{Mi} & = & \mathbf{x}_i'\boldsymbol{\gamma}_M + \epsilon_{Mi} \end{array} \tag{7.4}$$

Because utility has only ordinal meaning, the ordering of preferences remains the same when we add or subtract a number from the right-hand side of all these equations. Thus an equivalent way of representing the individual's preferences is obtained by subtracting $\mathbf{x}_i'\boldsymbol{\gamma}_0$ from all equations:

$$\begin{array}{rcl} u^*_{0i} & = & 0 + \epsilon_{0i} \\ u^*_{1i} & = & \mathbf{x}_i'\boldsymbol{\beta}_1 + \epsilon_{1i} \\ \vdots & & \vdots \qquad\quad \vdots \\ u^*_{Mi} & = & \mathbf{x}_i'\boldsymbol{\beta}_M + \epsilon_{Mi} \end{array} \tag{7.5}$$

where $\boldsymbol{\beta}_m \equiv \boldsymbol{\gamma}_m - \boldsymbol{\gamma}_0$ for all m. Notice that, by construction, $\boldsymbol{\beta}_0$ is normalized to a vector of zeros, as required for identification also in the direct setup of the models.

As in binary-choice models, the role of statistical noise is to capture any uncertainty with respect to the mechanism that determines the utility levels from the part of the researcher. This uncertainty could be due to, for example, unobserved characteristics of the decision maker that affect their tastes, deviations of the true mechanism from linearity and so on. This uncertainty has nothing to do with the decision makers themselves, who we will assume know exactly how much utility they derive from each alternative.

Utility maximization implies that individual i will choose alternative m if $u^*_{mi} > u^*_{\ell i}$ for all $\ell \neq m$. However, because of the noise terms in the specification of the utility levels, we can only make probabilistic statements about the alternative being chosen. Let y_i be an integer-valued random variable, the value of which indicates the alternative that the individual chooses; thus y_i can assume values in the set $\{0, 1, \ldots, M\}$.[3] Then, the probability of alternative m being chosen is:

$$\begin{aligned} p_{mi} &\equiv \text{Prob}(y_i = m|\mathbf{x}_i) \\ &= \text{Prob}\Big(u^*_{mi} > u^*_{\ell i} \quad \forall \ell \neq m \Big| \mathbf{x}_i\Big) \\ &= \text{Prob}\Big(\mathbf{x}_i'\boldsymbol{\beta}_m + \epsilon_{mi} > \mathbf{x}_i'\boldsymbol{\beta}_\ell + \epsilon_{\ell i} \quad \forall \ell \neq m \Big| \mathbf{x}_i\Big) \\ &= \text{Prob}\Big(\epsilon_{\ell i} < \epsilon_{mi} + \mathbf{x}_i'\left(\boldsymbol{\beta}_m - \boldsymbol{\beta}_\ell\right) \quad \forall \ell \neq m \Big| \mathbf{x}_i\Big) \end{aligned} \tag{7.6}$$

[3] Notice that the u^*_{mi}s are not observable because they represent utility levels. However, the y_is are observable as they are simply codes for the choices that each individual i makes.

If, for the moment, we treat ϵ_{mi} as given, this probability is equal to the joint cumulative density function of all $\epsilon_{\ell i}$s, $\ell \neq m$, evaluated at an M-dimensional vector with all elements equal to $\epsilon_{mi}+\mathbf{x}_i'(\boldsymbol{\beta}_m - \boldsymbol{\beta}_\ell)$. To continue with the analysis we need to impose distributional assumptions on the $\epsilon_{\ell i}$s. Following McFadden (1974),[4] we will first assume that $\epsilon_{\ell i}$, $\ell = 0, 1, \ldots, M$, is independent from all other error terms and follows a *type I extreme-value distribution*, with probability density and cumulative density functions:

$$\mathrm{p}(u) = e^{-u}\exp\left\{-e^{-u}\right\} \qquad \text{and} \qquad \mathrm{P}(u) = \exp\left\{-e^{-u}\right\} \tag{7.7}$$

respectively. Because the error terms are assumed to be independent, the cumulative density function implied by (7.6), but conditional on ϵ_{mi}, can be expressed as the product of the individual cumulative density functions:

$$\mathrm{Prob}\Big(\epsilon_{\ell i} < \epsilon_{mi} + \mathbf{x}_i'(\boldsymbol{\beta}_m - \boldsymbol{\beta}_\ell) \quad \forall \ell \neq m \Big| \mathbf{x}_i, \epsilon_{mi}\Big) = \prod_{\ell \neq m} \exp\left\{-e^{-\epsilon_{mi}-\mathbf{x}_i'(\boldsymbol{\beta}_m-\boldsymbol{\beta}_\ell)}\right\} \tag{7.8}$$

Of course, the researcher does not observe the value of ϵ_{mi} and this can be integrated out from the probability:

$$p_{mi} = \int_{-\infty}^{\infty} \prod_{\ell \neq m} \exp\left\{-e^{-u-\mathbf{x}_i'(\boldsymbol{\beta}_m-\boldsymbol{\beta}_\ell)}\right\} \times e^{-u}\exp\left\{-e^{-u}\right\}\, \mathrm{d}u \tag{7.9}$$

Carrying-out the integration leads to the choice probabilities presented in equation (7.2).[5] McFadden (1974) also shows that the relationship holds in the other direction: the only distribution for the ϵ_{mi}s consistent with the choice probabilities in (7.2) is the type-I extreme-value distribution.

Although leading to choice probabilities that have a closed form and are relatively easy to work with, the distributional assumption imposed on the ϵ_{mi}s imposes a restrictive structure on the relationships among these probabilities. In particular, consider the *odds ratio* between alternatives m and ℓ:

$$\frac{p_{mi}}{p_{\ell i}} = \frac{e^{\mathbf{x}_i'\boldsymbol{\beta}_m}}{e^{\mathbf{x}_i'\boldsymbol{\beta}_\ell}} = e^{\mathbf{x}_i'(\boldsymbol{\beta}_m-\boldsymbol{\beta}_\ell)} \tag{7.10}$$

This odds ratio does not depend on any parameters other than the ones specific to the two alternatives being considered and suggests that the inclusion or exclusion of any alternative other than m and ℓ in the choice set is irrelevant for the choice between m and ℓ. This property of the multinomial Logit model is accordingly termed *Independence of Irrelevant Alternatives*(IIA) and can lead to predictions about choices that are rather unreasonable. The blue bus/red bus example, which is due to McFadden (1974), illustrates this point.[6] Consider an individual who has two options for her daily commute to work: taking a red bus or driving. Given her characteristics and the model's parameters, suppose that the commuter is equally likely to choose any of the two alternatives. Thus, the odds ratio between taking the red bus and driving is one. Suppose now that a third option for commuting to work is added: a blue bus, which is similar to the red bus in every respect except color. Assuming that the commuter does not care about the color of the bus, we could expect her to be equally likely to take the red or blue bus and twice as likely to drive to work. This is because the original probability of taking a bus is now split equally between the red and blue bus, while the probability of driving remains unchanged. Given this intuitive argument, we would expect the odds ratio between taking the red bus and driving to drop to $1/2$. However, the multinomial Logit model cannot

[4] McFadden derived the choice probabilities for the conditional Logit model, rather than the multinomial Logit, and did not assume that the parameters enter the specification of the utility levels linearly. However, the mathematical procedures are not affected by these differences.

[5] This is done by collecting terms that involve e^{-u} and then integrating by changing the integration variable to $s = e^{-u}$. See Train (2009, pp.74-75) for a detailed derivation.

[6] The example was presented in McFadden (1974) in the context of the conditional Logit model, where also the attributes of the alternatives affect individual choice.

take into account the fact that the two buses are perfect substitutes and treats the blue bus option as irrelevant when predicting the odds ratio between the red bus and car. Thus, the multinomial Logit model will keep predicting equal probabilities between taking the red bus and driving to work, even after the blue bus is added to the choice set.

The source of IIA in the multinomial Logit model is the assumption that the error terms in (7.4) are independent. There are two alternative ways to impose a less restrictive structure on these error terms that lead to mathematically tractable models: (i) assume that the ϵ_{mi}s follow a generalization of the extreme-value distribution and (ii) assume that the ϵ_{mi}s follow a multivariate-Normal distribution. The first approach can lead to a variety of models, with the most widely used one being the *nested Logit*. The second approach leads to the *multinomial Probit* model. We will cover only the multinomial Probit model here and direct the reader to Chapter 4 in Train (2009) for a discussion of discrete-choice models that are based on *generalized exterme-value* distributions.

To derive the choice probabilities for the multinomial Probit model we could collect the $M+1$ ϵ_{mi}s into an $(M+1)\times 1$ vector and assume that it follows a multivariate-Normal distribution. Then we could continue in the same way as in the multinomial Logit model. Although valid, taking this approach will lead to integrals that will be very difficult to work with. Instead, we will transform the utility equations in (7.4) yet once more by expressing them in terms of differences in utility. This will also serve as a way of introducing the latent-variable representation of multinomial models. Let $y^*_{mi}=u^*_{mi}-u^*_{0i}$ be the difference between the utility that individual i derives from alternatives m and 0. By subtracting u^*_{0i} from both sides of all equations in (7.4) we obtain:

$$\begin{array}{rcl} y^*_{0i} & = & 0 \\ y^*_{1i} & = & \mathbf{x}_i'\boldsymbol{\beta}_1 + \varepsilon_{1i} \\ \vdots & & \vdots \quad\quad \vdots \\ y^*_{Mi} & = & \mathbf{x}_i'\boldsymbol{\beta}_M + \varepsilon_{Mi} \end{array} \tag{7.11}$$

where, $\boldsymbol{\beta}_m=\boldsymbol{\gamma}_m-\boldsymbol{\gamma}_0$ and $\varepsilon_{mi}=\epsilon_{mi}-\epsilon_{0i}$. In terms of these differences in utility levels, the individual's decision process is the following:

- individual i will choose alternative 0 if $y^*_{\ell i}<0$ for all $\ell=1,2,\ldots,M$. This is because u^*_{0i} is the maximum possible utility across all $M+1$ alternatives if and only if all $y^*_{\ell i}$s are negative.
- individual i will choose alternative $m\neq 0$ if $y^*_{mi}>0$ and $y^*_{mi}>y^*_{\ell i}$ for all $\ell\neq m$. This is because u^*_{mi} is greater than all other $u^*_{\ell i}$s if and only if y^*_{mi} is greater than all other $y^*_{\ell i}$s and positive.

Putting these results into a single expression leads to the latent-variable representation of multinomial models:

$$\begin{array}{rcl} y^*_{1i} & = & \mathbf{x}_i'\boldsymbol{\beta}_1 + \varepsilon_{1i} \\ y^*_{2i} & = & \mathbf{x}_i'\boldsymbol{\beta}_2 + \varepsilon_{2i} \\ \vdots & & \vdots \quad\quad \vdots \\ y^*_{Mi} & = & \mathbf{x}_i'\boldsymbol{\beta}_M + \varepsilon_{Mi} \end{array} \qquad y_i = \left\{ \begin{array}{lllll} 0 & \text{if} & \max_j\left\{y^*_{ji}\right\}\leq 0 & & \\ 1 & \text{if} & \max_j\left\{y^*_{ji}\right\}=y^*_{1i} & \text{and} & y^*_{1i}>0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ M & \text{if} & \max_j\left\{y^*_{ji}\right\}=y^*_{Mi} & \text{and} & y^*_{Mi}>0 \end{array} \right. \tag{7.12}$$

Notice that, by specifying the problem in terms of the differences in utility level between each alternative and alternative 0, we have reduced the dimension of the problem from $M+1$ to M.

It is stressed that this representation encompasses both multinomial Logit and Probit models. The difference between the two models stems from the distributional assumption on the error terms in the last expression. In the multinomial Logit model each $\epsilon_{\ell i}$ is assumed to follow a type I extreme-value distribution and, as we showed in footnote 3 on page 95, each difference of the from $\varepsilon_{\ell i}\equiv\epsilon_{\ell i}-\epsilon_{0i}$ follows a standard-Logistic distribution. Even in the multinomial-Logit case, however, the $\varepsilon_{\ell i}$ are not independent from each other, by construction.[7] For this reason,

[7]This is because each $\varepsilon_{\ell i}$ is constructed using the same ϵ_{0i}, which is a random variable. In fact, the M $\varepsilon_{\ell i}$s jointly follow a multivariate-Logistic distribution, as defined in Malik & Abraham (1973).

it is easier for this model to work with the original $\epsilon_{\ell i}$s, as we did above. In the multinomial Probit model we assume that the $M+1$ $\epsilon_{\ell i}$s follow a multivariate-Normal distribution with mean equal to a vector of zeros. This implies that the $M\times 1$ vector $\boldsymbol{\varepsilon}_i = \begin{bmatrix}\varepsilon_{1i} & \varepsilon_{2i} & \cdots & \varepsilon_{Mi}\end{bmatrix}'$ also follows a multivariate-Normal distribution with mean $\mathbf{0}$. The $\varepsilon_{\ell i}$s are, again, not independent form each other and this is now due, not only to the way they are constructed, but also due to the possibility of the underlying $\epsilon_{\ell i}$s being dependent. Since in the multinomial Probit model we have to work with dependent error terms, whether these are the original $\epsilon_{\ell i}$s or the transformed $\varepsilon_{\ell i}$s, it is slightly easier to use the transformed error terms, as this reduces the dimensionality of the problem from $M+1$ to M.

Let $\boldsymbol{\Omega}$ denote the $M\times M$ precision matrix of $\boldsymbol{\varepsilon}_i$ in the multinomial Probit model. Then, $\mathbf{y}_i^* = \begin{bmatrix}y_{1i}^* & y_{2i}^* & \cdots & y_{Mi}^*\end{bmatrix}'$ follows a multivariate-Normal distribution with mean $\mathbf{X}_i\boldsymbol{\beta}$, where $\mathbf{X}_i$ and $\boldsymbol{\beta}$ are as defined below equation (7.3). We can now express the probability of individual i choosing an alternative in terms of the differences in utility levels. This is easy to do for $m=0$ because the probability we need to calculate can be directly expressed as the cumulative density function of a multivariate-Normal distribution:

$$\begin{aligned} p_{0i} &= \text{Prob}\left(y_{\ell i}^* < 0 \quad \forall \ell \middle| \mathbf{x}_i\right) \\ &= \int_{-\infty}^{0}\cdots\int_{-\infty}^{0} \frac{|\boldsymbol{\Omega}|^{1/2}}{(2\pi)^{M/2}} \exp\left\{-\frac{(\mathbf{y}_i^* - \mathbf{X}_i\boldsymbol{\beta})'\,\boldsymbol{\Omega}\,(\mathbf{y}_i^* - \mathbf{X}_i\boldsymbol{\beta})}{2}\right\} \mathrm{d}\mathbf{y}_i^* \end{aligned} \tag{7.13}$$

This expression is the same as the upper branch of equation (7.3), with the only difference appearing in the integration variable (which will of course disappear if we carry-out the integration). Similarly, for any alternative $m\neq 0$ we get:

$$\begin{aligned} p_{mi} &= \text{Prob}\left(y_{mi}^* > 0,\ y_{mi}^* > y_{\ell i}^* \quad \forall \ell \neq m \middle| \mathbf{x}_i\right) \\ &= \int_{\mathcal{A}_M}\cdots\int_{\mathcal{A}_1} \frac{|\boldsymbol{\Omega}|^{1/2}}{(2\pi)^{M/2}} \exp\left\{-\frac{(\mathbf{y}_i^* - \mathbf{X}_i\boldsymbol{\beta})'\,\boldsymbol{\Omega}\,(\mathbf{y}_i^* - \mathbf{X}_i\boldsymbol{\beta})}{2}\right\} \mathrm{d}\mathbf{y}_i^* \end{aligned} \tag{7.14}$$

The range of integration in this M-dimensional integral is $\mathcal{A}_m = (0, +\infty)$, and $\mathcal{A}_\ell = (-\infty, y_{mi}^*]$ for $\ell\neq m$: we allow y_{mi}^* to assume any positive value and we restrict every $y_{\ell i}^*$ with $\ell\neq m$ to be smaller than y_{mi}^*. Save for the variable of integration, this expression is the same as the lower branch of equation (7.3).

Compared to the rather simple formulas for the choice probabilities implied by the multinomial Logit model, the multidimensional integrals that appear in the choice probabilities of the multinomial Probit model cannot be expressed in closed form or simplified in any way that will facilitate parameter estimation. This is the main reason behind the multinomial Probit model not enjoying the same degree of popularity among economists, at least in frequentist analyses, as the multinomial Logit model. As we will see in the following subsection, however, not being able to get closed-form solutions for the choice probabilities does not present major challenges when a Bayesian approach is used.

Before we close this section we note that the parameters of the multinomial Probit model, as presented here, are not identified. The issue is similar to the one that we encountered in the binary Probit model and has to do with the scale of the error terms. In particular, multiplying $\boldsymbol{\beta}$ by a positive constant, z, and $\boldsymbol{\Omega}$ by $z^{-\frac{1}{2}}$ leaves the choice probabilities unaffected. Using the random-utility framework as the underlying theoretical model, we can give economic content to this issue: returning to the assumed mechanism that determines the utility levels, if we multiply all equations in (7.4) by $z>0$ then the ordering of utility levels will not change and only lead to a rescaling of the $y_{\ell i}^*$s. This should be expected given that utility values have only ordinal meaning. This multiplication, however, will alter the variance of the $\epsilon_{\ell i}$s and, eventually, of the $\varepsilon_{\ell i}$s. The non-identification issue is resolved in the binary Probit model by restricting the variance of the error term to unity. Now that the model involves M $\varepsilon_{\ell i}$s more possibilities are available for normalizing the scale of $\boldsymbol{\varepsilon}_i$ and we will review some of them when discussing estimation of the multinomial Probit model.

7.3.2 Estimation of the Multinomial Logit Model

The observed-data likelihood for the multinomial Logit model is obtained by plugging the choice probabilities in (7.2) into the probability mass function of the categorical distribution in (7.1):

$$\mathrm{p}(\mathbf{y}|\mathbf{X},\boldsymbol{\beta}) = \prod_{i=1}^{N}\left[\prod_{m=0}^{M}\left(\frac{e^{\mathbf{x}_i'\boldsymbol{\beta}_m}}{\sum_{\ell=0}^{M} e^{\mathbf{x}_i'\boldsymbol{\beta}_\ell}}\right)^{\mathbb{1}(y_i=m)}\right] \tag{7.15}$$

where $\mathbf{y}$ is an $N\times 1$ vector that stores the values on the response variable (codes of the alternative chosen) for all N potential observations and $\mathbf{X}$ is an $N\times K$ matrix that stores the corresponding values of the K independent variables (individual characteristics). We should also keep in mind that $\boldsymbol{\beta}_0$ is normalized to a vector of zeros for identification purposes. Although appearing complex, the amount of calculations required to evaluate the complete-data likelihood at specific parameter values and for a given dataset is rather small. This is because each potential observation i will contribute to the likelihood function only one of the $M+1$ terms that appear inside the second product and this term will be the one corresponding to the alternative chosen. This computational simplification makes the multinomial Logit model attractive when a frequentist approach is to be used for estimation. In a Bayesian setting, however, and with the absence of conjugate priors for the $\boldsymbol{\beta}_\ell$s, designing an efficient sampler is challenging. Nevertheless, data augmentation can be used to simplify the computations. Using the latent-variable representation of multinomial models in (7.12), the complete-data likelihood can be expressed as:

$$\mathrm{p}\left(\mathbf{y},\{\mathbf{y}_i^*\}\,\middle|\mathbf{X},\boldsymbol{\beta}\right) = \prod_{i=1}^{N}\mathrm{p}\left(y_i\middle|\mathbf{y}_i^*\right)\times\mathrm{p}(\mathbf{y}_i^*|\mathbf{X},\boldsymbol{\beta}) \tag{7.16}$$

where the second factor is the probability density function of a multivariate Logistic distribution:

$$\mathrm{p}(\mathbf{y}_i^*|\mathbf{x}_i,\boldsymbol{\beta}) = \frac{M!\exp\left\{-\sum_{\ell=1}^{M}\left(y_{\ell i}^*-\mathbf{x}_i'\boldsymbol{\beta}_\ell\right)\right\}}{\left(1+\sum_{\ell=1}^{M}e^{-\left(y_{\ell i}^*-\mathbf{x}_i'\boldsymbol{\beta}_\ell\right)}\right)^{M+1}} \tag{7.17}$$

The multivariate-Logistic distribution, is defined in Malik & Abraham (1973).

There are a couple of alternative ways of expressing the first factor in the complete-data likelihood in a single mathematical expression. A convenient one, because it takes a form similar to the probability mass function of a categorically distributed random variable, is:

$$\mathrm{p}\left(y_i\middle|\mathbf{y}_i^*\right) = \mathbb{1}\left(\max_j\left\{y_{ji}^*\right\}\le 0\right)^{\mathbb{1}(y_i=0)}\times\prod_{\ell=1}^{M}\mathbb{1}\left(\max_j\left\{y_{ji}^*\right\}=y_{\ell i}^*>0\right)^{\mathbb{1}(y_i=\ell)} \tag{7.18}$$

All parameters in the multinomial Logit model are contained in the $(M\cdot K)\times 1$ vector $\boldsymbol{\beta}$. Using a Normal prior for $\boldsymbol{\beta}$, with mean vector $\mathbf{m}$ and precision matrix $\mathbf{P}$, a standard application of Bayes' theorem results in:

$$\begin{aligned}\pi\left(\boldsymbol{\beta},\{\mathbf{y}_i^*\}\,\middle|\mathbf{y},\mathbf{X}\right) &\propto \mathrm{p}\left(\mathbf{y},\{\mathbf{y}_i^*\}\,\middle|\mathbf{X},\boldsymbol{\beta}\right)\mathrm{p}(\boldsymbol{\beta})\\ &= \prod_{i=1}^{N}\left[\mathbb{1}\left(\max_j\left\{y_{ji}^*\right\}\le 0\right)^{\mathbb{1}(y_i=0)}\times\prod_{\ell=1}^{M}\mathbb{1}\left(\max_j\left\{y_{ji}^*\right\}=y_{\ell i}^*>0\right)^{\mathbb{1}(y_i=\ell)}\right.\\ &\qquad\left.\times M!\exp\left\{-\sum_{\ell=1}^{M}\left(y_{\ell i}^*-\mathbf{x}_i'\boldsymbol{\beta}_\ell\right)\right\}\left(1+\sum_{\ell=1}^{M}e^{-\left(y_{\ell i}^*-\mathbf{x}_i'\boldsymbol{\beta}_\ell\right)}\right)^{-(M+1)}\right]\\ &\qquad\times\frac{|\mathbf{P}|^{1/2}}{(2\pi)^{K/2}}\exp\left\{-\frac{1}{2}\left(\boldsymbol{\beta}-\mathbf{m}\right)'\mathbf{P}\left(\boldsymbol{\beta}-\mathbf{m}\right)\right\}\end{aligned} \tag{7.19}$$

Although this expression looks intimidating, many of the terms will drop from it when deriving the full conditionals for $\boldsymbol{\beta}$ and the $\mathbf{y}_i^*$s, as it was the case also in binary-choice models. Dropping the multiplicative terms that do not involve $\boldsymbol{\beta}$ leads to the full conditional:

$$\pi\left(\boldsymbol{\beta}|\bullet\right) \propto \prod_{i=1}^{N}\left[\frac{\exp\left\{-\sum_{\ell=1}^{M}\left(y_{\ell i}^{*}-\mathbf{x}_i'\boldsymbol{\beta}_\ell\right)\right\}}{\left(1+\sum_{\ell=1}^{M}e^{-\left(y_{\ell i}^{*}-\mathbf{x}_i'\boldsymbol{\beta}_\ell\right)}\right)^{M+1}}\right]\times\exp\left\{-\frac{\left(\boldsymbol{\beta}-\mathbf{m}\right)'\mathbf{P}\left(\boldsymbol{\beta}-\mathbf{m}\right)}{2}\right\} \tag{7.20}$$

This density does not belong to any known parametric family and direct sampling from it is not possible. However, a Metropolis-Hastings step is, as always, feasible.

Given $\boldsymbol{\beta}$, $\mathbf{y}_i^*$ follows is a multivariate Logistic distribution, truncated to an appropriate range. For example, if $y_i = 0$ only the term $\mathbb{1}\left(\max_j\left\{y_{ji}^*\right\}\leq 0\right)$ in (7.18) depends on $\mathbf{y}_i^*$ and the resulting full conditional is:

$$\mathrm{p}(\mathbf{y}_i^*|\mathbf{x}_i,\boldsymbol{\beta},y_i=0) \propto \frac{\exp\left\{-\sum_{\ell=1}^{M}\left(y_{\ell i}^{*}-\mathbf{x}_i'\boldsymbol{\beta}_\ell\right)\right\}}{\left(1+\sum_{\ell=1}^{M}e^{-\left(y_{\ell i}^{*}-\mathbf{x}_i'\boldsymbol{\beta}_\ell\right)}\right)^{M+1}}\mathbb{1}\left(\max_j\left\{y_{ji}^*\right\}\leq 0\right) \tag{7.21}$$

Similar simplifications take place for any value of y_i other than zero. Frühwirth-Schnatter & Frülhwirth (2012) provide an algorithm for sampling from these complete conditionals, which is based on first sampling for the underlying $\epsilon_{\ell i}$s.

These results are summarized in Theorem 7.1. An application of the multinomial Logit model is presented in Example 7.1.

THEOREM 7.1: Full Conditionals for the Multinomial Logit Model
In the multinomial Logit model with $M+1$ alternatives and K independent variables:

$$\mathbf{y}_i^* = \mathbf{X}_i\boldsymbol{\beta}+\boldsymbol{\varepsilon}_i, \qquad \boldsymbol{\varepsilon}_i \sim \text{Logistic}\left(\mathbf{0},\mathbf{1}\right)$$

$$y_i = \begin{cases} 0 & \text{if} & \max_j\left\{y_{ji}^*\right\}\leq 0 & & \\ 1 & \text{if} & \max_j\left\{y_{ji}^*\right\}=y_{1i}^* & \text{and} & y_{1i}^*>0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ M & \text{if} & \max_j\left\{y_{ji}^*\right\}=y_{Mi}^* & \text{and} & y_{Mi}^*>0 \end{cases}$$

and with a Normal prior for $\boldsymbol{\beta}$:

$$\mathrm{p}(\boldsymbol{\beta}) = \frac{|\mathbf{P}|^{1/2}}{(2\pi)^{MK/2}}\exp\left\{-\frac{1}{2}\left(\boldsymbol{\beta}-\mathbf{m}\right)'\mathbf{P}\left(\boldsymbol{\beta}-\mathbf{m}\right)\right\}$$

the full conditional of $\boldsymbol{\beta}$ is:

$$\pi\left(\boldsymbol{\beta}|\bullet\right) \propto \prod_{i=1}^{N}\left[\frac{\exp\left\{-\sum_{\ell=1}^{M}\left(y_{\ell i}^{*}-\mathbf{x}_i'\boldsymbol{\beta}_\ell\right)\right\}}{\left(1+\sum_{\ell=1}^{M}e^{-\left(y_{\ell i}^{*}-\mathbf{x}_i'\boldsymbol{\beta}_\ell\right)}\right)^{M+1}}\right]\times\exp\left\{-\frac{1}{2}\left(\boldsymbol{\beta}-\mathbf{m}\right)'\mathbf{P}\left(\boldsymbol{\beta}-\mathbf{m}\right)\right\}$$

The full conditional of $\mathbf{y}_i^*$, $i=1,2,\ldots,N$, is proportional to an M-dimensional Logistic

probability density function, restricted to specific range, depending on the value of y_i:

$$\mathrm{p}\left(\mathbf{y}_i^*|\bullet\right) \propto \begin{cases} \frac{\exp\left\{-\sum_{\ell=1}^{M}\left(y_{\ell i}^*-\mathbf{x}_i'\boldsymbol{\beta}_\ell\right)\right\}}{\left(1+\sum_{\ell=1}^{M} e^{-\left(y_{\ell i}^*-\mathbf{x}_i'\boldsymbol{\beta}_\ell\right)}\right)^{M+1}} \mathbb{1}\left(\max_j\left\{y_{ji}^*\right\}\leq 0\right) & \text{if } y_i=0 \\ \frac{\exp\left\{-\sum_{\ell=1}^{M}\left(y_{\ell i}^*-\mathbf{x}_i'\boldsymbol{\beta}_\ell\right)\right\}}{\left(1+\sum_{\ell=1}^{M} e^{-\left(y_{\ell i}^*-\mathbf{x}_i'\boldsymbol{\beta}_\ell\right)}\right)^{M+1}} \mathbb{1}\left[\max_j\left\{y_{ji}^*\right\}=y_{mi}^*>0\right] & \text{otherwise} \end{cases}$$

♦ **Example 7.1 Preferred Method of Balancing the Budget**

In this example, we will use again part of the dataset constructed through the 2017 Cooperative Congressional Election Study (Schaffner & Ansolabhere, 2019), which we first used in example 6.3. We will use here 5,000 randomly drawn observations from the original 18,200 observations contained in the dataset and the following variables:

`action`	:	stated preferred action for Congress to balance the budget; the available choices are coded as: 0 - "Cut Defense Spending" 1 - "Cut Domestic Spending" 2 - "Raise Taxes"
`age`	:	age of the respondent in years
`educ`	:	educational level of the respondent, coded from 1 to 6 with larger values corresponding to higher educational level
`male`	:	dummy variable: 1 if the respondent is male
`homeowner`	:	dummy variable: 1 if the respondent owns their own home
`ideology`	:	self-reported political viewpoint, coded from 1 (very liberal) to 5 (very conservative)
`faminc`	:	annual family income bracket, coded from 1 (less that \$10,000) to 16 (more than \$500,000) and increasing by \$10,000 when income is bellow \$100,000 and by larger amounts after that

We will assume that the utility each individual derives from each possible action taken by Congress is a linear function of the individual's characteristics:

$$\text{action}_{mi}^* = \gamma_{1m}+\gamma_{2m}\text{age}_i+\gamma_{3m}\text{educ}_i+\gamma_{4m}\text{male}_i+\gamma_{5m}\text{homeowner}_i+\gamma_{6m}\text{ideology}_i+\gamma_{7m}\text{faminc}_i+\epsilon_{mi}$$

for $m=0,1,2$ and where each ϵ_{mi} follows a type I extreme-value distribution. This specification leads to a multinomial Logit model with three alternatives. The results obtained using BayES' `mnlogit()` function to estimate this model are presented in the following table.

	Mean	Median	Sd.dev.	5%	95%
action = 1					
constant	-3.39487	-3.39701	0.219936	-3.77262	-3.02478
age	0.00746804	0.00750233	0.00246472	0.00338422	0.0115466
educ	-0.0949643	-0.0952437	0.0284275	-0.141041	-0.0484418
male	-0.00753653	-0.00931138	0.07004	-0.120805	0.111782
homeowner	0.0983611	0.0983883	0.0855046	-0.0435998	0.238013
ideology	0.981247	0.9804	0.0376342	0.920353	1.04508
faminc	0.0101913	0.0101737	0.0132354	-0.0116894	0.0321513
action = 2					
constant	-2.01301	-2.02137	0.182642	-2.30012	-1.70194
age	0.0259051	0.0258871	0.00235707	0.0220072	0.0298638
educ	-0.0652364	-0.0645008	0.0280207	-0.112605	-0.020312
male	-0.157204	-0.156578	0.0756974	-0.282729	-0.0335256
homeowner	-0.134969	-0.135067	0.0892559	-0.286894	0.00819727
ideology	0.224584	0.224245	0.0347081	0.167559	0.28273
faminc	-0.0261602	-0.0261855	0.0134347	-0.0480857	-0.00412205

There are a couple of things to note here. First, there are no parameter estimates for the base category ($m=0$; "Cut Defense Spending"). The two blocks of the table correspond to parameters $\boldsymbol{\beta}_m \equiv \boldsymbol{\gamma}_m - \boldsymbol{\gamma}_0$ for $m=1$ ("Cut Domestic Spending") and $m=2$ ("Raise Taxes"). Second, the 90% credible intervals for most βs does not contain zero but, because the choice probabilities in the

multinomial Logit model are not linear in the parameters, we cannot interpret the magnitude of the βs; we should instead calculate marginal effects on the choice probabilities. However, these marginal effects are highly non-linear functions of the βs and not guaranteed to have the same sign as the corresponding parameters. Therefore, contrary to the binary Logit model, we should refrain from interpreting even the sign of the βs in a multinomial Logit model.

Obtaining the results presented above using BayES can be achieved using the code in the following box. We note that due to the Metropolis-Hastings step used for sampling from the full conditional of $\boldsymbol{\beta}$, as well as to the many latent variables ($\mathbf{y}_i^*$s), the multinomial Logit model is plagued by very large inefficiency factors. For this reason, a large burn-in and a large value for the thinning parameter was chosen in the code below. Further post-estimation analysis of the results, such as plotting the draws from the posterior distribution of the parameters, is highly recommended.

```
// import the data into a dataset called Data
Data = webimport("www.bayeconsoft.com/datasets/CCES2017.csv");

// generate a constant term
Data.constant = 1;

// run the multinomial-Logit model
myMNLogit = mnlogit( action ~ constant age educ male homeowner ideology faminc,
    "draws"=20000, "burnin"=20000, "thin"=5, "chains"=2 );
```

7.3.3 Estimation of the Multinomial Probit Model

Using the latent-variable representation of multinomial models in (7.12), the complete-data likelihood for the multinomial model can be expressed as:

$$\mathrm{p}\left(\mathbf{y}, \{\mathbf{y}_i^*\} \middle| \mathbf{X}, \boldsymbol{\beta}\right) = \prod_{i=1}^{N} \mathrm{p}\left(y_i \middle| \mathbf{y}_i^*\right) \times \mathrm{p}(\mathbf{y}_i^* | \mathbf{X}, \boldsymbol{\beta}) \tag{7.22}$$

where the second factor is the probability density function of a multivariate-Normal distribution with mean $\mathbf{X}_i\boldsymbol{\beta}$ and precision matrix $\boldsymbol{\Omega}$. The first factor in this likelihood function can be expressed in the same way as in the multinomial Logit model (see equation (7.18)). We will again use a Normal prior for $\boldsymbol{\beta}$, with mean vector $\mathbf{m}$ and precision matrix $\mathbf{P}$. If we ignore for the moment the non-identification issue in the multinomial Probit model, we could use a Wishart prior for $\boldsymbol{\Omega}$, with degrees-of-freedom parameter n and scale matrix $\mathbf{V}$. In this setting the complete conditionals of $\boldsymbol{\beta}$ and $\boldsymbol{\Omega}$ would be the same as in the SUR model (see Theorem 3.1), with $\mathbf{y}_i^*$ taking the place of $\mathbf{y}_i$. Similarly to the multinomial Logit model, the complete conditionals of the $\mathbf{y}_i^*$ take the form of the multivariate-Normal density, restricted to a specific range, depending on the value of y_i:

$$\mathrm{p}\left(\mathbf{y}_i^* \middle| \bullet\right) = \begin{cases} \frac{|\boldsymbol{\Omega}|^{\frac{1}{2}}}{(2\pi)^{\frac{M}{2}}} \exp\left\{-\frac{(\mathbf{y}_i^* - \mathbf{X}_i\boldsymbol{\beta})'\boldsymbol{\Omega}(\mathbf{y}_i^* - \mathbf{X}_i\boldsymbol{\beta})}{2}\right\} \mathbb{1}\left(\max_j \left\{y_{ji}^*\right\} \leq 0\right) & \text{if } y_i = 0 \\ \frac{|\boldsymbol{\Omega}|^{\frac{1}{2}}}{(2\pi)^{\frac{M}{2}}} \exp\left\{-\frac{(\mathbf{y}_i^* - \mathbf{X}_i\boldsymbol{\beta})'\boldsymbol{\Omega}(\mathbf{y}_i^* - \mathbf{X}_i\boldsymbol{\beta})}{2}\right\} \mathbb{1}\left[\max_j \left\{y_{ji}^*\right\} = y_{mi}^* > 0\right] & \text{otherwise} \end{cases} \tag{7.23}$$

Sampling from the complete conditional of each $\mathbf{y}_i^*$ can be done using rejection or Gibbs sampling.[8]

The main issue with estimating the parameters of a multinomial Probit model is related to identification and, as in the binary Probit model, this can be resolved by restricting the scale of the precision matrix or its inverse. A common way of doing this is by restricting the first element of $\boldsymbol{\Omega}^{-1}$ to unity. However, if this approach is taken then $\boldsymbol{\Omega}$ no longer follows a Wishart distribution and this complicates sampling from its full conditional. A simple solution to the arising computational complications is to completely ignore the identification issue during sampling and transform the draws after sampling is complete. For example, McCulloch & Rossi (1994) follow this approach and report only $\sigma_{11}^{-1/2}\boldsymbol{\beta}$, where σ_{11} is the first element

[8] See footnote 9 on page 109.

of the unrestricted $\mathbf{\Omega}^{-1}$. Although simple, the approach can lead to numerical instability problems, especially when coupled with vague priors. This is because there is nothing in the sampling algorithm other than the prior to prevent $\boldsymbol{\beta}$ and $\sigma_{11}^{1/2}$ from growing at the same rate, but possibly without bound. McCulloch et al. (2000) provide an alternative approach that partitions $\mathbf{\Omega}^{-1}$ into three blocks of parameters with full conditionals that are easy to sample from, while Imai & van Dyk (2005) develop a Gibbs sampler that relies on marginal data augmentation for restricting σ_{11} to unity. Restricting the first element of $\mathbf{\Omega}^{-1}$ to unity, however, treats the alternatives asymmetrically and for this reason, Burgette & Nordheim (2012) modify Imai & van Dyk's algorithm such that the scale of $\mathbf{\Omega}^{-1}$ is fixed by restricting its trace to be equal to M. Because we have been using mostly precision (as opposed to covariance) matrices in this textbook, we will present here Burgette & Nordheim's approach, while restricting the trace of $\mathbf{\Omega}$ to M.

We start by defining $\breve{\mathbf{\Omega}}$ as an $M \times M$ positive-definite matrix and imposing a Wishart prior on it, with hyperparameters n and $\mathbf{V}$. We then define $\alpha^2 \equiv {}^{\text{tr}(\breve{\mathbf{\Omega}})}/_{M}$ and $\mathbf{\Omega} \equiv ({}^{1}/_{\alpha^2})\,\breve{\mathbf{\Omega}}$. The last two expressions define a transformation from the unrestricted $\breve{\mathbf{\Omega}}$ to $\left(\alpha^2, \mathbf{\Omega}\right)$, with the trace of $\mathbf{\Omega}$ restricted to M. α^2 in this transformation will be the working parameter in a marginal data augmentation context. The determinant of the Jacobian of the transformation can be shown to be a function of M and an application of the multivariate version of the change-of-variables theorem leads to the probability density function of $\left(\alpha^2, \mathbf{\Omega}\right)$. The density of $\mathbf{\Omega}$ is obtained by integrating-out α^2 from the joint density of $\left(\alpha^2, \mathbf{\Omega}\right)$ and it can be shown to be:

$$\text{p}(\mathbf{\Omega}) = \frac{M \cdot |\mathbf{\Omega}|^{\frac{n-M-1}{2}} |\mathbf{V}^{-1}|^{n/2}}{\text{tr}\left(\mathbf{V}^{-1}\mathbf{\Omega}\right)^{nM/2}} \frac{\Gamma\left(\frac{nM}{2}\right)}{\Gamma_M\left(\frac{n}{2}\right)} \mathbb{1}(\text{tr}\left(\mathbf{\Omega}\right) = M) \tag{7.24}$$

Furthermore, conditional on $\mathbf{\Omega}$, α^2 follows a Gamma distribution with shape ${}^{nM}/_{2}$ and rate ${}^{\text{tr}(\mathbf{V}^{-1}\mathbf{\Omega})}/_{2}$. We can now define the marginal data augmentation mapping, $\mathcal{D}_{\alpha}\left(\cdot\right)$, in a way that facilitates sampling for $\breve{\mathbf{\Omega}}$:

$$\breve{\mathbf{y}}_i^* = \tfrac{1}{\alpha}\left(\mathbf{y}_i^* - \mathbf{X}_i\boldsymbol{\beta}\right) \equiv \breve{\boldsymbol{\varepsilon}}_i, \qquad \breve{\boldsymbol{\varepsilon}}_i \sim \text{N}\left(\mathbf{0}, \breve{\mathbf{\Omega}}\right) \tag{7.25}$$

The resulting Gibbs sampler iterates between the steps:

- **(a)** draw $\left(\{\mathbf{y}_i^*\}, \mathbf{\Omega}\right)$ from the transition kernel, $\mathscr{K}$, of a Markov chain with stationary distribution $\text{p}\left(\{\mathbf{y}_i^*\}, \mathbf{\Omega} \middle| \{\mathbf{y}_i\}; \boldsymbol{\beta}\right)$. This step requires multiple iterations between the following:
  - **(a1)** draw $\{\mathbf{y}_i^*\}$ from $\text{p}\left(\{\mathbf{y}_i^*\} \middle| \{\mathbf{y}_i\}, \mathbf{\Omega}; \boldsymbol{\beta}\right)$; this is the multivariate truncated-Normal distribution given in (7.23)
  - **(a2)** draw α^2 from $\text{p}\left(\alpha^2 \middle| \mathbf{\Omega}\right)$; this is a Gamma distribution with parameters $a = {}^{nM}/_{2}$ and $b = {}^{\text{tr}(\mathbf{V}^{-1}\mathbf{\Omega})}/_{2}$
  - **(a3)** transform each $\mathbf{y}_i^*$ to $\breve{\mathbf{y}}_i^* = \frac{1}{\alpha}\left(\mathbf{y}_i^* - \mathbf{X}_i\boldsymbol{\beta}\right)$
  - **(a4)** draw $\breve{\mathbf{\Omega}}$ from $\text{p}\left(\breve{\mathbf{\Omega}} \middle| \{\breve{\mathbf{y}}_i^*\}\right)$; this is a Wishart distribution as in the SUR model
  - **(a5)** set $\alpha^2 = \frac{1}{M}\,\text{tr}\left(\breve{\mathbf{\Omega}}\right)$ and $\mathbf{\Omega} = \frac{1}{\alpha^2}\breve{\mathbf{\Omega}}$
  - **(a6)** transform each $\breve{\mathbf{y}}_i^*$ to $\mathbf{y}_i^* = \alpha\breve{\mathbf{y}}_i^* + \mathbf{X}_i\boldsymbol{\beta}$
- **(b)** draw $\boldsymbol{\beta}$ from $\text{p}\left(\boldsymbol{\beta} \middle| \{\mathbf{y}_i\}, \{\mathbf{y}_i^*\}, \mathbf{\Omega}\right) = \text{p}\left(\boldsymbol{\beta} \middle| \{\mathbf{y}_i^*\}, \mathbf{\Omega}\right)$; this is a multivariate-Normal distribution as in the SUR model

In this algorithm all draws are obtained from standard distributions. Marginal data augmentation is used only in step **(a)**, while step **(b)** is a typical conditional draw of the Gibbs sampler. In steps **(a1)**-**(a3)** we sample each $\breve{\mathbf{y}}_i^*$ given $\mathbf{\Omega}$, but we do so while using α^2 as a working parameter. In steps **(a4)**-**(a5)** we effectively sample for $\left(\mathbf{\Omega}, \alpha^2\right)$ given $\breve{\mathbf{y}}_i^*$.

The algorithm requires iterating over steps **(a1)**-**(a6)** multiple times within each iteration of the Gibbs sampler. Burgette & Nordheim (2012) use a vector of zeros as the prior mean for

$\boldsymbol{\beta}$, which leads to an algorithm that does not require an inner loop and, thus, is much more efficient from a computational standpoint. Such an algorithm is used by BayES when $\mathbf{m} = \mathbf{0}$.

The following example applies the multinomial Probit model the Cooperative Congressional Election Study dataset and the specification used in Example 7.1.

♦ **Example 7.1 Preferred Method of Balancing the Budget (Continued)**
In this example, we will continue working with the 2017 Cooperative Congressional Election Study dataset (Schaffner & Ansolabhere, 2019) and assuming that the utility an individual derives from each possible action taken by Congress to balance the budget can be expressed as a linear function of the individual's characteristics:

$$\text{action}^*_{mi} = \gamma_{1m} + \gamma_{2m}\text{age}_i + \gamma_{3m}\text{educ}_i + \gamma_{4m}\text{male}_i + \gamma_{5m}\text{homeowner}_i + \gamma_{6m}\text{ideology}_i + \gamma_{7m}\text{faminc}_i + \epsilon_{mi}$$

for $m = 0, 1, 2$. However, we will now assume that the ϵ_{mi}s follow a multivariate-Normal distribution and allow them to be correlated. This assumption implies that $\boldsymbol{\varepsilon}_i \equiv \begin{bmatrix}\epsilon_{1i} - \epsilon_{0i} & \epsilon_{2i} - \epsilon_{0i}\end{bmatrix}'$ follows a bivariate-Normal distribution, the precision matrix of which we denote by $\boldsymbol{\Omega}$. Taken together, the imposed assumptions lead to a multinomial Probit model with three alternatives. The results obtained using BayES' `mnprobit()` function to estimate this model are presented in the following table.

	Mean	Median	Sd.dev.	5%	95%
action = 1					
constant	-5.43837	-4.31663	3.61359	-12.754	-2.51101
age	0.0267974	0.0211903	0.0207602	0.00802298	0.0656532
educ	-0.166731	-0.136008	0.116675	-0.38381	-0.059719
male	-0.0738342	-0.0606747	0.181822	-0.361785	0.165814
homeowner	0.0796718	0.065249	0.198332	-0.182619	0.390793
ideology	1.62346	1.30366	1.04844	0.757805	3.72458
faminc	0.00241351	0.00254456	0.0293319	-0.0396159	0.0441816
action = 2					
constant	-4.16746	-3.49587	2.57237	-8.94124	-1.78317
age	0.040039	0.0350294	0.0206812	0.0194973	0.0780333
educ	-0.13806	-0.113227	0.101639	-0.324423	-0.0368948
male	-0.206685	-0.182003	0.17261	-0.507393	0.0109183
homeowner	-0.0926144	-0.0905784	0.171672	-0.357769	0.163172
ideology	1.02036	0.797808	0.805694	0.303477	2.55343
faminc	-0.0192827	-0.017832	0.0265328	-0.061064	0.0177392

The posterior means of the βs have the same order of magnitude as those from the multinomial Logit model, but are not particularly close to them. This situation is similar to what we encountered in binary-respose models: the βs are scaled differently (through the variances of the error terms) in the two models and, because of this, what we should be comparing is the marginal effects on the choice probabilities. In the case of multinomial models, however, there is an additional reason for the observed differences in parameter estimates. This has to do with the flexibility in the dependence of the ϵ_{mi}s introduced by the multinomial Probit model. If the Independence of Irrelevant Alternatives does not hold in a particular application, then we should expect to see these differences propagating to the marginal effects as, in such a case, the multinomial Logit model would be imposing too restrictive of a structure on the data.

Obtaining the results presented above using BayES can be achieved using the code in the following box. Again, due to to the many latent variables ($\mathbf{y}^*_i$s) in the model, very large inefficiency factors are expected. For this reason, a large burn-in and a large value for the thinning parameter was chosen in the code below. Post-estimation inspection of the draws from the posterior distribution, such as by plotting them per chain, is highly recommended.

```
// import the data into a dataset called Data
Data = webimport("www.bayeconsoft.com/datasets/CCES2017.csv");

// generate a constant term
Data.constant = 1;

// run the multinomial-Probit model
myMNProbit = mnprobit( action ~ constant age educ male homeowner ideology faminc,
    "draws"=20000, "burnin"=20000, "thin"=5, "chains"=2 );
```

7.3.4 Marginal Effects in Multinomial Models

The quantities being modelled in multinomial models are the probabilities of occurrence of each of the $M+1$ mutually exclusive and collectively exhaustive outcomes. These probabilities are highly non-linear functions of the parameters and the independent variables: for the multinomial Logit model these probabilities are given in (7.2), while for the multinomial Probit in (7.3). For the multinomial Probit model the probabilities cannot even be expressed in closed form. Because of this non-linear relationship between the quantities of interest and the independent variables, the effect of a change in, say, the k^{th} independent variable, x_{ik}, on the probability of occurrence of each outcome should be calculated as a partial derivative.

We will first calculate the marginal effects for the Logit model because, at least we have the probabilities of occurrence in closed form. The change in the probability of outcome m occurring in this model, caused by a small change in x_{ik} is:

$$\frac{\partial \operatorname{Prob}(y_i=m|\mathbf{x}_i)}{\partial x_{ik}} = \frac{e^{\mathbf{x}_i'\boldsymbol{\beta}_m}}{\sum_{\ell=0}^{M} e^{\mathbf{x}_i'\boldsymbol{\beta}_\ell}} \left[\beta_{mk} - \frac{\sum_{\ell=0}^{M} \beta_{\ell k} e^{\mathbf{x}_i'\boldsymbol{\beta}_\ell}}{\sum_{\ell=0}^{M} e^{\mathbf{x}_i'\boldsymbol{\beta}_\ell}} \right] = p_{mi} \left[\beta_{mk} - \sum_{\ell=0}^{M} \beta_{\ell k} p_{\ell i} \right] \tag{7.26}$$

where β_{mk} is the k^{th} parameter (coefficient associated with the k^{th} independent variable) in $\boldsymbol{\beta}_m$. We should keep in mind that all parameters in $\boldsymbol{\beta}_0$ are normalized to zero for identification purposes. $p_{mi} \equiv \operatorname{Prob}(y_i=m|\mathbf{x}_i)$ is a non-negative number and, therefore, when multiplied by the first summand inside the square brackets, the sign of the product will be the same as the sign of β_{mk}. However, the second summand inside the square brackets is a non-linear function of the data and all model parameters and, as such, its sign cannot and magnitude cannot be determined theoretically. The practical implication of this is that there is no guarantee that the sign of β_{mk} will be the same as the sign of the associated marginal effect. Thus, contrary to the binary Logit model, in the multinomial Logit model we should not be tempted to interpret even the signs of the parameters; the only interpretable quantity of interest is the marginal effect.

When $\mathbf{x}_i$ contains binary variables, calculating a derivative with respect to the value of the dummy variable does not make sense. The corresponding concept of a marginal effect of a binary independent variable on the probability of outcome m occurring is the difference:

$$\operatorname{Prob}(y_i=m|\mathbf{x}_{i1}) - \operatorname{Prob}(y_i=m|\mathbf{x}_{i0}) = \frac{e^{\mathbf{x}_{1i}'\boldsymbol{\beta}_m}}{\sum_{\ell=0}^{M} e^{\mathbf{x}_{1i}'\boldsymbol{\beta}_\ell}} - \frac{e^{\mathbf{x}_{0i}'\boldsymbol{\beta}_m}}{\sum_{\ell=0}^{M} e^{\mathbf{x}_{0i}'\boldsymbol{\beta}_\ell}} \tag{7.27}$$

where $\mathbf{x}_{i0}$ is a vector that consists of the values of the independent variables at the point at which the marginal effect is evaluated, but with a zero in the k-th place and $\mathbf{x}_{i1}$ is a similar vector, but with a one in the k-th place.

The marginal effects in a multinomial Probit model have a similar interpretation as in the multinomial Logit model and, in theory, should be calculated by differentiating the probabilities in (7.3). However, these probabilities do not have a closed form and neither do their derivatives. Bolduc (1999) provides an algorithm for approximating these derivatives, which is based on the Geweke-Hajivassiliou-Keane (*GHK*) *algorithm*.[9] Although measuring the same underlying concept, marginal effects obtained from a multinomial Logit and a multinomial Probit model can be substantially different from each other when the IIA assumption imposed by the former model is violated.

One feature of the marginal effects in multinomial models that is worth noting is that the sum over all possible alternatives of the marginal effect with respect to an independent variable is always zero:

$$\sum_{m=0}^{M} \frac{\partial \operatorname{Prob}(y_i=m|\mathbf{x}_i)}{\partial x_{ik}} = 0, \quad k = 1, 2, \ldots, K \tag{7.28}$$

[9] See also footnote 7 on page 107.

This is straightforward to show for the multinomial Logit model, using the two formulas above for continuous and discrete x_ks, but quite more challenging for the multinomial Probit model. The result itself is not surprising at all: a small change in the k^{th} independent variable will affect all choice probabilities while enforcing that these probabilities sum to unity, even after the change.

The following example presents and interprets the marginal effects from the multinomial Logit and Probit models used in the two preceding parts of Example 7.1 to model the probability of choosing a proposed action for balancing the budget.

♦ **Example 7.1 Preferred Method of Balancing the Budget (Continued)**
We will now calculate and interpret the marginal effects generated by the multinomial Logit and Probit models, which were estimated in the last two parts of the example using the 2017 Cooperative Congressional Election Study dataset (Schaffner & Ansolabhere, 2019). The following table presents these marginal effects from the multinomial Logit model.

	Mean	Median	Sd.dev.	5%	95%
dProb(y=0)/dx					
age	-0.00379443	-0.00380745	0.000505469	-0.00461301	-0.00294011
educ	0.0202953	0.0202074	0.00589497	0.0106949	0.0299378
*male	0.017625	0.017819	0.0153062	-0.00827157	0.0423618
*homeowner	0.000923984	0.00076394	0.0171921	-0.0271954	0.0294507
ideology	-0.162154	-0.162177	0.00726767	-0.174181	-0.150151
faminc	0.0013299	0.00132785	0.00275236	-0.00321471	0.00588073
dProb(y=1)/dx					
age	-0.000337324	-0.00034084	0.000484574	-0.00113408	0.000469027
educ	-0.0156426	-0.0157603	0.00562302	-0.0247207	-0.0062633
*male	0.0101726	0.00984587	0.0135418	-0.0116254	0.0330215
*homeowner	0.0313129	0.0317581	0.0176621	0.00113834	0.0593831
ideology	0.19544	0.195329	0.00692193	0.18433	0.207033
faminc	0.00417935	0.00416939	0.00263937	-0.000145476	0.00850979
dProb(y=2)/dx					
age	0.00413176	0.00412524	0.000385711	0.00350517	0.00477069
educ	-0.00465276	-0.0046136	0.00461197	-0.0124201	0.00281134
*male	-0.0277976	-0.0276025	0.0121285	-0.0481447	-0.00824178
*homeowner	-0.0322369	-0.0325666	0.016071	-0.058721	-0.00623916
ideology	-0.0332862	-0.0333736	0.0055326	-0.0423157	-0.0240223
faminc	-0.00550925	-0.00549959	0.00224092	-0.00920047	-0.0018341

*Marginal effect is calculated for discrete change from 0 to 1.

Contrary to the parameter estimates, where we had two blocks of parameters, we now have three blocks of marginal effects, with each block containing the marginal effects on one of the three alternatives:

- $m=0$: "Cut Defense Spending"
- $m=1$: "Cut Domestic Spending"
- $m=2$: "Raise Taxes"

From this table we see, for example, that a person who is one year older than the average person in the sample has 0.379% lower probability of indicating that they prefer alternative 0, 0.034% lower probability of preferring alternative 1 and 0.413% higher probability of preferring alternative 2. That these three changes in probability sum to zero is not by coincidence: if older people are more likely to prefer one alternative then, necessarily, they should be less likely to prefer another alternative, because the alternatives considered are collectively exhaustive. Additionally, an increase in the educational level increases the probability of preferring alternative zero, as does being a male, although the 95% credible interval for the last marginal effect contains zero. The largest impact on the probabilities comes from the political viewpoint, with people self-reporting themselves as more conservative being more likely to prefer alternative 1 and less likely to prefer alternatives 0 or 2. An interesting attribute with respect to the ideology variable is that the coefficient associated with it in alternative 2 is positive (this can be seen in the parameter estimates from the multinomial Logit model), while the marginal effect is

negative. As stressed before, in multinomial models there is no guarantee that the sign of the marginal effect will be the same as the sign of the corresponding parameter.

The following table presents the marginal effects from the multinomial Probit model. As a general observation, the marginal effects from the this model are quite close to those from the multinomial Logit model. This is not unexpected, as they both measure the same underlying concept. These differences are parlty due to the the greater flexibility allowed by the multinomial Probit model in the relationship between the error terms in the three alternatives.

	Mean	Median	Sd.dev.	5%	95%
dProb(y=0)/dx					
age	-0.00390703	-0.0039057	0.000480295	-0.00469911	-0.00311865
educ	0.0181802	0.0181492	0.0056827	0.0089015	0.0275763
*male	0.017166	0.0173196	0.0169089	-0.0107641	0.0446347
*homeowner	0.00111666	0.00123993	0.0183443	-0.0290281	0.0313161
ideology	-0.15461	-0.154584	0.00733989	-0.166676	-0.14257
faminc	0.00100175	0.000995773	0.00258082	-0.00321711	0.00522696
dProb(y=1)/dx					
age	-0.000708374	-0.000706684	0.000470647	-0.00148982	6.2426e-05
educ	-0.0149808	-0.014955	0.00550868	-0.0241439	-0.00592573
*male	0.0168277	0.0167579	0.015635	-0.00888957	0.0425712
*homeowner	0.0335841	0.0335181	0.0171731	0.00550791	0.0617718
ideology	0.195973	0.195901	0.00739073	0.183937	0.208261
faminc	0.00381987	0.00381856	0.00250139	-0.000307203	0.00793254
dProb(y=2)/dx					
age	0.0046154	0.00461087	0.000451577	0.00388337	0.00537312
educ	-0.00319934	-0.00323039	0.00488207	-0.0112184	0.00483264
*male	-0.0339937	-0.0340452	0.0142383	-0.0570253	-0.0106831
*homeowner	-0.0347008	-0.0345522	0.0161238	-0.0612176	-0.0085267
ideology	-0.0413633	-0.0413306	0.00641197	-0.0520315	-0.0308671
faminc	-0.00482162	-0.00483443	0.00228921	-0.00859132	-0.00105586

*Marginal effect is calculated for discrete change from 0 to 1.

Finally, we can compare the multinomial Logit and Probit models with respect to their ability to accomodate the features of the data. As the following table shows, with equal prior model probabilities, the data tend to favor the multinomial Probit model.

Model	Log-Marginal Likelihood	Type of log-ML Approximation	Prior Model Probability	Posterior Model Probability
myMNLogit	-4759.96	Lewis & Raftery	0.5	0.149501
myMNProbit	-4758.23	Lewis & Raftery	0.5	0.850499

Obtaining the results presented above using BayES can be achieved using the code in the following box.

```
// import the data into a dataset called Data and generate a constant term
Data = webimport("www.bayeconsoft.com/datasets/CCES2017.csv");
Data.constant = 1;

// run the multinomial Logit and Probit models
myMNLogit = mnlogit( action ~ constant age educ male homeowner ideology,
    "draws"=20000, "burnin"=20000, "thin"=5, "chains"=2 );

myMNProbit = mnprobit( action ~ constant age educ male homeowner ideology,
    "draws"=20000, "burnin"=20000, "thin"=5, "chains"=2 );

// calculate marginal effects for the two models at the means of the
// independent variables
mfx( "model"=myMNLogit, "point"="mean" );
mfx( "model"=myMNProbit, "point"="mean" );

// compare the two models
pmp( { myMNLogit, myMNProbit } );
```

7.4 Conditional Models

Conditional models for discrete outcomes are very similar to multinomial models, as far as the response variable is concerned: y_i is still allowed to be in one out of $M+1$ states and there is no objective way of ordering the alternatives. The feature that distinguishes conditional from multinomial models is the nature of the independent variables: while in multinomial models the independent variables vary only by observation (by individual in a discrete-choice context), in conditional models the independent variables vary by alternative and possibly by individual. In mathematical terms, multinomial models express the right hand side of the equation for the m^{th} latent variable as $\mathbf{x}_i'\boldsymbol{\beta}_m + \epsilon_{mi}$, but conditional models express the right-hand side of the same equation as $\mathbf{z}_{mi}'\boldsymbol{\delta} + \epsilon_{mi}$. Notice that the vector of parameters, $\boldsymbol{\delta}$, in the conditional model no longer has an m subscript, as this subscript has now moved to the vector of independent variables, $\mathbf{z}_{mi}$.

To make maters concrete we will use an example where individuals living in a given city have four alternatives for commuting to work: walk, ride a bike, drive or use public transportation. The idependent variables in a multinomial model would be characteristics of the individual decision maker that may affect the choice. These characteristics could be income, age, distance from workplace and so on, and, for any individual i, these characteristics do not vary by the alternative chosen. In a conditional model the independent variables would be attributes of the alternative, such as an index of how safe each mode of transportation is in that particular city, the cost of each alternative or the time it takes for commuting to work by each alternative. Of course, some attributes could be common to all individuals and some could vary both by alternative and individual. For example, the time it takes to commute to work by car could very well be different for two different individuals, but the safety index for each mode of transportation could be the same for all individuals.

Conditional models for discrete outcomes can be thought of as extensions to multinomial models because, apart from variables that vary by alternative, they can aslo accommodate independent variables that vary only by individual, simply by interacting these variables with alternative-specific dummy variables. Consider the following latent-variable representation of the model:

$$\begin{array}{rcl} y_{1i}^* & = & \mathbf{z}_{1i}'\boldsymbol{\delta} + \mathbf{x}_i'\boldsymbol{\beta}_1 + \varepsilon_{1i} \\ y_{2i}^* & = & \mathbf{z}_{2i}'\boldsymbol{\delta} + \mathbf{x}_i'\boldsymbol{\beta}_2 + \varepsilon_{2i} \\ \vdots & & \vdots \qquad\qquad \vdots \\ y_{Mi}^* & = & \mathbf{z}_{Mi}'\boldsymbol{\delta} + \mathbf{x}_i'\boldsymbol{\beta}_M + \varepsilon_{Mi} \end{array} \tag{7.29}$$

where the variables in the $\mathbf{z}_{mi}$s vary by alternative and possibly by individual ,while the variables in $\mathbf{x}_i$ vary only by individual. We can express this system of equations in matrix form as:

$$\mathbf{y}_i^* = \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i \tag{7.30}$$

where:

$$\underset{M\times 1}{\mathbf{y}_i^*} = \begin{bmatrix} y_{1i}^* \\ y_{2i}^* \\ \vdots \\ y_{Mi}^* \end{bmatrix}, \quad \underset{M\times J}{\mathbf{X}_i} = \begin{bmatrix} \mathbf{z}_{1i}' & \mathbf{x}_i' & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{z}_{2i}' & \mathbf{0} & \mathbf{x}_i' & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{z}_{Mi}' & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{x}_i' \end{bmatrix}, \quad \underset{J\times 1}{\boldsymbol{\beta}} = \begin{bmatrix} \boldsymbol{\delta} \\ \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_M \end{bmatrix}, \quad \text{and} \quad \underset{M\times 1}{\boldsymbol{\varepsilon}_i} = \begin{bmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \\ \vdots \\ \varepsilon_{Mi} \end{bmatrix}$$

and where J, is equal to the number of variables in $\mathbf{z}$ plus M times the number of variables in $\mathbf{x}$. It becomes apparent from $\mathbf{X}_i$ that the way $\mathbf{x}_i$ enters the equation for the m^{th} alternative is equivalent to constructing a group of independent variables that vary by both individual and alternative, but the variability in the alternatives dimension comes from interacting the variables in $\mathbf{x}_i$ with a dummy variable that is equal to one only for the m^{th} alternative. This representation also makes clear that, no new estimation procedures need to be developed for conditional models: once we define $\mathbf{X}_i$ appropriately, we can use the same samplers developed for the corresponding multinomial models.

Before turning to estimation procedures we need to make an important remark about the nature of the variables in $\mathbf{z}$. The random-utility framework is frequently used to justify the latent-variable representation of the model. In this framework we normalize the utility obtained from the first alternative (denoted by 0) to zero. By doing this in the multinomial model we normalized the coefficients associated with the first alternative to zero. In a conditional model, however, not all parameters are specific to an alternative and the previous approach will not work. Let's start from specifying the process that determines the utility levels from each alternative:

$$\begin{array}{rcrcl} u^*_{0i} & = & \mathbf{w}'_{0i}\boldsymbol{\delta} + \mathbf{x}'_i\boldsymbol{\gamma}_0 & + & \epsilon_{0i} \\ u^*_{1i} & = & \mathbf{w}'_{1i}\boldsymbol{\delta} + \mathbf{x}'_i\boldsymbol{\gamma}_1 & + & \epsilon_{1i} \\ \vdots & & \vdots & & \vdots \\ u^*_{Mi} & = & \mathbf{w}'_{Mi}\boldsymbol{\delta} + \mathbf{x}'_i\boldsymbol{\gamma}_M & + & \epsilon_{Mi} \end{array} \tag{7.31}$$

where $\mathbf{w}_{0i}$ contains the attributes of alternative 0, possibly specific to individual i, $\mathbf{w}_{1i}$ the attributes of alternative 1 and so on. To express the model in terms of differences in utility we subtract the first equation from all other equations:

$$\begin{array}{rcrcl} y^*_{0i} & = & 0 & & \\ y^*_{1i} & = & (\mathbf{w}_{1i}-\mathbf{w}_{0i})'\boldsymbol{\delta} + \mathbf{x}'_i(\boldsymbol{\gamma}_1-\boldsymbol{\gamma}_0) & + & (\epsilon_{1i}-\epsilon_{0i}) \\ \vdots & & \vdots & & \vdots \\ y^*_{Mi} & = & (\mathbf{w}_{Mi}-\mathbf{w}_{0i})'\boldsymbol{\delta} + \mathbf{x}'_i(\boldsymbol{\gamma}_M-\boldsymbol{\gamma}_0) & + & (\epsilon_{Mi}-\epsilon_{0i}) \end{array} \tag{7.32}$$

where $y^*_{mi} \equiv u^*_{mi} - u^*_{0i}$, $m = 0, 1, \ldots, M$. To get from this system of equations to the latent-variable representation in (7.29) we define $\mathbf{z}_{mi} \equiv \mathbf{w}_{mi} - \mathbf{w}_{0i}$, $\boldsymbol{\beta}_m \equiv \boldsymbol{\gamma}_m - \boldsymbol{\gamma}_0$ and $\varepsilon_{mi} \equiv \epsilon_{mi} - \epsilon_{mi}$, for all $m = 1, 2, \ldots, M$. The practical implication of this normalization is that the variables in $\mathbf{z}$ in the latent-variable representation of the model are the differences of the values of the attributes of the first alternative from the attributes the remaining M alternatives. In an empirical application the $\mathbf{w}_{mi}$ variables need to be transformed to the $\mathbf{z}_{mi}$s before the conditional model is estimated. This may appear as extra work for the researcher but most software packages, including BayES, automate the task of creating these differences in the values of the attributes. Notice that by taking these differences, a conditional model collapses to a binary-choice model when there are only two alternatives. In such a model the differences in the attributes between outcome 1 and outcome 0 are inserted in the only place a binary-choice model has for specifying independent variables: the equation that specifies the probability of success (outcome 1).

Now that we have seen the main difference between multinomial and conditional models for discrete choice we can formally define the difference between conditional Logit and Probit models: in a *conditional Logit* model the error terms in (7.31) are assumed to be independent from each other and follow a type I extreme-value distribution, while in a *conditional Probit* model[10] the error terms for a particular observation i are assumed to follow a multivariate-Normal distribution. The functional forms of the choice probabilities in the two models are the same as in the multinomial counterparts, with the only thing changing being the specification of the right-hand sides of the equations for the latent variables. The conditional Logit model still imposes the IIA assumption:

$$\frac{p_{mi}}{p_{\ell i}} = \frac{e^{\mathbf{z}'_{mi}\boldsymbol{\delta}+\mathbf{x}'_i\boldsymbol{\beta}_m}}{e^{\mathbf{z}'_{\ell i}\boldsymbol{\delta}+\mathbf{x}'_i\boldsymbol{\beta}_\ell}} = e^{(\mathbf{z}_{mi}-\mathbf{z}_{\ell i})'\boldsymbol{\delta}+\mathbf{x}'_i(\boldsymbol{\beta}_m-\boldsymbol{\beta}_\ell)} \tag{7.33}$$

although when already controlling for differences in the attributes of each alternative, this assumption may be much more unrealistic. This is because the odds ratio between alternatives m and ℓ, as given in the last equation, should naturally depend on whether the attributes of a third alternative are closer to those of m than ℓ. As it is the case with multinomial models, the conditional Probit model relaxes the IIA assumption by allowing the error terms in (7.31) to be correlated across equations.

[10] Although a Logit model with independent variables that vary by alternative is typically called a conditional Logit model, this is not the case for Probit models: most authors prefer to still call what we define here as a conditional Probit, a multinomial Probit model.

7.4.1 Estimation of Conditional Models for Discrete Choice

The latent-variable representation of a conditional model for discrete choice differs from the corresponding representation of a multinomial model only in terms of the variables that appear in the right-hand side of the equations for the y^*_{mi}s. In particular, in a conditional model we get for alternative m:

$$y^*_{mi} = \mathbf{z}'_{mi}\boldsymbol{\delta} + \mathbf{x}'_i\boldsymbol{\beta} + \varepsilon_{mi} \tag{7.34}$$

while the term $\mathbf{z}'_{mi}\boldsymbol{\delta}$ is dropped in a multinomial model. Similarly, the parameters in a conditional model are $\boldsymbol{\beta} = \begin{bmatrix}\boldsymbol{\delta} & \boldsymbol{\beta}_1 & \cdots & \boldsymbol{\beta}_M\end{bmatrix}'$, while the fist block of parameters is missing from a multinomial model. Given these similarities, only minor modifications need to be made in the formulas for the full and complete conditionals derived for the multinomial counterparts. If a multivariate-Normal prior is used for $\boldsymbol{\beta}$ in a conditional Logit model then the full conditional for $\boldsymbol{\beta}$ becomes:

$$\pi\left(\boldsymbol{\beta}|\bullet\right) \propto \prod_{i=1}^{N}\left[\frac{\exp\left\{-\sum_{\ell=1}^{M}\left(y^*_{\ell i} - \mathbf{z}'_{\ell i}\boldsymbol{\delta} - \mathbf{x}'_i\boldsymbol{\beta}_\ell\right)\right\}}{\left(1+\sum_{\ell=1}^{M} e^{-\left(y^*_{\ell i} - \mathbf{z}'_{\ell i}\boldsymbol{\delta} - \mathbf{x}'_i\boldsymbol{\beta}_\ell\right)}\right)^{M+1}}\right] \times \exp\left\{-\frac{\left(\boldsymbol{\beta}-\mathbf{m}\right)'\mathbf{P}\left(\boldsymbol{\beta}-\mathbf{m}\right)}{2}\right\} \tag{7.35}$$

For a conditional Probit model the full and complete conditionals are those reported in Section 7.3.3 and the only difference between the multinomial and conditional Probit models is in the way $\mathbf{X}_i$ is defined:

$$\mathbf{X}_i = \begin{bmatrix} \mathbf{z}'_{1i} & \mathbf{x}'_i & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{z}'_{2i} & \mathbf{0} & \mathbf{x}'_i & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{z}'_{Mi} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{x}'_i \end{bmatrix} \tag{7.36}$$

in the conditional model, while a multinomial model contains only the block-diagonal part of the matrix above.

The following example presents an application of conditional models to a problem of modelling the choice of fishing mode by recreational anglers. It uses as independent variables the attributes of each fishing mode, as well as the characteristics of the individuals.

♦ **Example 7.2 Recreational Fishing Mode**

In this example, we will use the dataset constructed by Thomson & Crooke (1991), as used by Herriges & Kling (1999) to model the mode of fishing chosen by 1182 recreational anglers. The dataset contains information on 1182 individuals on the following variables:

`mode`	:	recreation fishing mode choice: 0 - beach 2 - charter 1 - boat 3 - pier
`price_0`	:	price for beach mode (hundreds of US dollars per trip)
`price_1`	:	price for private boat mode (hundreds of US dollars per trip)
`price_2`	:	price for charter boat mode (hundreds of US dollars per trip)
`price_3`	:	price for pier mode (hundreds of US dollars per trip)
`rate_0`	:	catch rate for beach mode
`rate_1`	:	catch rate for private boat mode
`rate_2`	:	catch rate for charter boat mode
`rate_3`	:	catch rate for pier mode
`income`	:	the individual's monthly income (thousands of US dollars)

In this dataset the price and catch-rate variables (attributes) vary by both alternative and individual, but income varies only by individual (individual characteristic). The price variable includes, apart from any boat fees, the cost of fuel and opportunity costs of round-trip travel. The catch rates are defined in a per-hour basis and depend on the respondents' targeted species. We note that the dataset is provided in 'wide format', where the variables that represent attributes are stored in different columns

and there is a single row per observation. BayES requires the data for conditional models to be in 'wide format', but some software packages require the data in 'long format'. In the 'long format' each attribute variable is stored in a single column for all alternatives, but the data for each individual are spread over multiple rows (four rows in an application with four alternatives) to accommodate the different values of the attributes for each alternative.

We will assume that the utility that individual i derives from each fishing mode is a linear function of the model's attributes and the individual's income:

$$\text{mode}^*_{mi} = \delta_1 \text{price}_{mi} + \delta_2 \text{rate}_{mi} + \gamma_{1m} + \gamma_{2m} \text{income}_i + \epsilon_{mi}$$

for $m = 0, 1, 2, 3$. Notice that each mode of fishing (equation) has its own constant term, γ_{1m}, and its own coefficient associated with the income variable, γ_{2m}. On the contrary, the two variables that vary by alternative are associated with coefficients that are common to all modes. If we assume that each ϵ_{mi} follows a type I extreme value distribution and it is independent of the $\epsilon_{\ell i}$s that appear in other equations, then we end up with a conditional Logit model with four alternatives. A multivariate-Normal assumption on the $M+1$ ϵ_{mi}s leads to a conditional Probit model. The following two tables present results obtained using, respectively, BayES' `clogit()` and `cprobit()` functions to estimate the two models.

	Mean	Median	Sd.dev.	5%	95%
mode					
price	-2.52981	-2.52485	0.166764	-2.81297	-2.2634
rate	0.366761	0.365327	0.109923	0.188559	0.550965
mode = 1					
constant	0.548667	0.552333	0.219605	0.182512	0.906326
income	0.0859984	0.085725	0.0494161	0.00574861	0.168668
mode = 2					
constant	1.7191	1.71642	0.22586	1.35122	2.09803
income	-0.0366796	-0.0367008	0.0507107	-0.121232	0.0469404
mode = 3					
constant	0.799792	0.803484	0.219524	0.435421	1.14951
income	-0.132464	-0.132594	0.0503367	-0.213896	-0.0475816

	Mean	Median	Sd.dev.	5%	95%
mode					
price	-1.58003	-1.55837	0.19682	-1.93473	-1.29969
rate	0.649115	0.636077	0.163492	0.404603	0.937011
mode = 1					
constant	-0.505295	-0.492719	0.338955	-1.07672	0.0328209
income	0.0941459	0.091574	0.0452355	0.0243894	0.171175
mode = 2					
constant	0.635322	0.602065	0.342907	0.146063	1.23368
income	-0.114876	-0.107703	0.0624231	-0.224988	-0.0296475
mode = 3					
constant	0.61948	0.571198	0.31534	0.21298	1.18844
income	-0.0938168	-0.0853114	0.0581785	-0.19812	-0.0187632

Both tables of results are split into four blocks. The upper block contains summary statistics of the draws obtained from the posterior distributions of the parameters which are common to all alternatives (δ_1 and δ_2) and which are associated with the attributes of each fishing mode: price and catch rate. The following three blocks contain similar statistics for the parameters that are specific to each alternative: the constant term and the individual's income. Although there are four alternatives from which an individual can choose, the first alternative is used as the base for identification purposes and the reported statistics correspond to parameters $\boldsymbol{\beta}_m \equiv \boldsymbol{\gamma}_m - \boldsymbol{\gamma}_0$ for $m = 1$ ("boat"), $m = 2$ ("charter") and $m = 3$ ("pier"). BayES always uses as the base category the alternative whose attribute names end with "_0", which in this example corresponds to the "beach" alternative.

Because the choice probabilities are expressed as non-linear functions of the parameters, we should, once again, refrain from interpreting the magnitudes of the parameters and we should, instead, calculate marginal effects on the choice probabilities. As it is the case in multinomial models, the signs of the marginal effects of the individual characteristics could be different from the signs of the corresponding parameters. As will see bellow, however, at least for the conditional Logit model, the signs of the parameters associated with the attributes of the alternatives are always the same as the effect that a marginal increase in the value of an alternative's attribute has on the probability of the individual choosing this alternative. In the context of our application we can thus conclude that an increase in the price of an alternative leads to a reduction in the probability of an individual choosing this alternative (because the posterior mean of δ_1 is negative) and that an increase in the catch rate of an alternative increases the probability of choosing this alternative (because the posterior mean of δ_2 is positive).

Obtaining the results presented above using BayES can be achieved using the code in the following box. We note that due to the Metropolis-Hastings step used in the conditional Logit model, the thinning parameter was to 5 to reduce autocorrelation in the retained draws.

```
// import the data into a dataset called Data
Data = webimport("www.bayeconsoft.com/datasets/FishingMode.csv");

// generate a constant term
Data.constant = 1;

// run the conditional-Logit model
myCLogit = clogit( mode ~ price rate | constant income, "thin"=5, "chains"=2 );

// run the conditional-Probit model
myCProbit = cprobit( mode ~ price rate | constant income, "thin"=5, "chains"=2 );
```

7.4.2 Marginal Effects in Conditional Models for Discrete Choice

As it is the case for any non-linear model, the magnitudes of the parameter estimates cannot be easily interpreted in conditional models for discrete choice. Instead, we should calculate and interpret the signs and magnitudes of the marginal effects of the independent variables on the probability of each alternative being selected by the individual. If a conditional model contains some variables that vary only by individual then the marginal effects on these variables are the same as those for multinomial models (see Section 7.3.4). The marginal effects with respect to the variables that vary by alternative are, again, calculated as marginal changes in each probability of the form $\text{Prob}(y_i = m | \{\mathbf{z}_{\ell i}\}, \mathbf{x}_i)$ caused by a change in the variables in $\mathbf{z}$. However, there are two types of vectors that store such variables: $\mathbf{z}_{mi}$ that contains the differences in attributes between alternative m and the base alternative, and $M-1$ additional $\mathbf{z}_{\ell i}$s that contain the differences in attributes between alternative $\ell \neq m$ and the base alternative. For example, if we are modelling the choice of commuting mode to work, an increase in commuting time for alternative m is expected to lead to a decrease in the probability of alternative m being chosen. However, an increase in commuting time for any other alternative, ℓ, is expected, *a priori*, to lead to an increase in the probability of selecting alternative m. For the conditional Logit model these two types of marginal effects can be obtained in closed form:

$$\begin{aligned}\frac{\partial \,\text{Prob}(y_i = m | \{\mathbf{z}_{\ell i}\}, \mathbf{x}_i)}{\partial w_{mik}} &= \delta_k \left[\frac{e^{\mathbf{w}'_{mi}\boldsymbol{\delta} + \mathbf{x}'_i \boldsymbol{\beta}_m}}{\sum_{\ell=0}^{M} e^{\mathbf{w}'_{\ell i}\boldsymbol{\delta} + \mathbf{x}'_i \boldsymbol{\beta}_\ell}} - \frac{\left(e^{\mathbf{w}'_{mi}\boldsymbol{\delta} + \mathbf{x}'_i \boldsymbol{\beta}_m}\right)^2}{\left(\sum_{\ell=0}^{M} e^{\mathbf{w}'_{\ell i}\boldsymbol{\delta} + \mathbf{x}'_i \boldsymbol{\beta}_\ell}\right)^2} \right] \\ &= \delta_k p_{mi} \left(1 - p_{mi}\right)\end{aligned} \tag{7.37}$$

and:

$$\begin{aligned}\frac{\partial \,\text{Prob}(y_i = m | \{\mathbf{z}_{\ell i}\}, \mathbf{x}_i)}{\partial w_{jik}} &= -\delta_k \frac{e^{\mathbf{w}'_{ji}\boldsymbol{\delta} + \mathbf{x}'_i \boldsymbol{\beta}_\ell}}{\sum_{\ell=0}^{M} e^{\mathbf{w}'_{\ell i}\boldsymbol{\delta} + \mathbf{x}'_i \boldsymbol{\beta}_\ell}} \frac{e^{\mathbf{w}'_{mi}\boldsymbol{\delta} + \mathbf{x}'_i \boldsymbol{\beta}_m}}{\sum_{\ell=0}^{M} e^{\mathbf{w}'_{\ell i}\boldsymbol{\delta} + \mathbf{x}'_i \boldsymbol{\beta}_\ell}} \\ &= -\delta_k p_{mi} p_{ji}\end{aligned} \tag{7.38}$$

where w_{mk} is the k^{th} attribute of alternative m and w_{jk} is the k^{th} attribute of alternative j, $j \neq m$.[11] As in the multinomial Logit model, the marginal effects with respect to the k^{th} attribute sum across alternatives to zero:

$$\sum_{m=0}^{M} \frac{\partial \operatorname{Prob}(y_i = m | \{\mathbf{z}_{\ell i}\}, \mathbf{x}_i)}{\partial w_{jik}} = 0, \qquad k = 1, 2, \ldots, K \tag{7.39}$$

This result is quite intuitive: if the k^{th} attribute of alternative j changes by a small amount, this change will have an impact on the choice probabilities for all $M+1$ alternatives. However, the choice probabilities should sum to unity before and after the change in w_{jik} and, thus, the marginal effects should sum to zero. Another identity that can be easily verified using the formulas for the marginal effects is:

$$\sum_{j=0}^{M} \frac{\partial \operatorname{Prob}(y_i = m | \{\mathbf{z}_{\ell i}\}, \mathbf{x}_i)}{\partial w_{jik}} = 0, \quad m = 0, 1, \ldots, M \tag{7.40}$$

It is easier to interpret this result in the context of the example of selecting a commuting mode to work. A small increase in commuting time for alternative m is expected to lead to a reduction in the probability of a commuter selecting mode m. However, if the commuting time of all $M+1$ modes increases by the same amount, then this does not affect the probability of choosing alternative m, or any other alternative. This makes perfect sense if we consider the random-utility framework that underlies the model: a change in the value of the k^{th} variable in every $\mathbf{w}_{\ell i}$ in (7.31) by Δw_k will lead to a change in all utility levels by $\delta_k \Delta w_k$. But because the choice probabilities depend only on utility differences, not the absolute levels of utility, these choice probabilities will remain unaffected.

The last two results are based on the fundamental properties of conditional models for discrete choice and, as such, hold also for conditional Probit models. In conditional Probit models however, there are no closed-form expressions for the marginal effects and the results are much harder to prove mathematically. Even though no closed-form expressions are available for conditional Probit models, the effects can again be approximated using the GHK algorithm.

We now turn to two additional results that are specific to the marginal effects of conditional Logit models. First, because all p_{mi}s, $(1-p_{mi})$s and p_{ji}s are non-negative, (7.37) and (7.38) suggest that there is a correspondence between the signs of the marginal effects of the attributes of the alternatives and the signs of the respective coefficients: the effect of a change in $\mathbf{w}_{mi}$ on p_{mi} has the same direction as the sign of the corresponding δ_k, but the effect of a change in a variable in $\mathbf{w}_{ji}$, $j \neq m$, on p_{mi} has the opposite direction. This means that we can interpret the signs of the parameters in $\boldsymbol{\delta}$, although not their magnitudes. Second, the marginal effects of conditional Logit models are symmetric with respect to the attributes of the alternatives:

$$\frac{\partial \operatorname{Prob}(y_i = m | \{\mathbf{z}_{\ell i}\}, \mathbf{x}_i)}{\partial w_{jik}} = \frac{\partial \operatorname{Prob}(y_i = j | \{\mathbf{z}_{\ell i}\}, \mathbf{x}_i)}{\partial w_{mik}} \tag{7.41}$$

This symmetry is not as intuitive as the previous identities and comes as a result of the way the choice probabilities are specified in conditional Logit models. Although neither of the last two results holds exactly for conditional Probit models, both of them appear to be approximately satisfied in empirical applications.

The following example is a continuation of Example 7.2, where we used conditional Logit and Probit specifications to model the choice of fishing mode. In this part of the example we calculate and interpret the marginal effects produced by the two models.

♦ **Example 7.2 Recreational Fishing Mode (Continued)**
In this example, we will continue working with the dataset used by Herriges & Kling (1999) to model the choice of fishing mode. The following two tables present statistics from the the posterior distributions of

[11] Notice that we express marginal effects here in terms of the original attributes, $\mathbf{w}_{\ell i}$, not in terms of the differences in attributes relative to the base category, $\mathbf{z}_{\ell i} \equiv \mathbf{w}_{\ell i} - \mathbf{w}_{0i}$.

the marginal effects under a conditional Logit and a conditional Probit model, respectively. Each table contains four relatively long blocks, but we only present here the first two blocks, which correspond to alternatives 0 ("beach") and 1 ("boat").

	Mean	Median	Sd.dev.	5%	95%
dProb(y=0)/dx					
price_0	-0.12415	-0.12359	0.0120506	-0.144926	-0.105301
price_1	0.0549252	0.0546284	0.00553064	0.0462726	0.0643764
price_2	0.0606322	0.0602833	0.00611838	0.0511386	0.0712863
price_3	0.00859255	0.00849544	0.00154957	0.006208	0.0113355
rate_0	0.0181096	0.0177575	0.00597811	0.00888643	0.0284348
rate_1	-0.00800764	-0.00785416	0.00263756	-0.0125446	-0.00393137
rate_2	-0.00883996	-0.00866363	0.00291664	-0.0138842	-0.00434996
rate_3	-0.00126197	-0.00121042	0.000487182	-0.00212885	-0.000563979
income	-0.000496659	-0.00050641	0.00230648	-0.00425132	0.00336569
dProb(y=1)/dx					
price_0	0.0549252	0.0546284	0.00553064	0.0462726	0.0643764
price_1	-0.615507	-0.614619	0.0426639	-0.687958	-0.546662
price_2	0.491633	0.490037	0.0432637	0.422656	0.566211
price_3	0.0689487	0.0687155	0.00656619	0.0585808	0.0800685
rate_0	-0.00800764	-0.00785416	0.00263756	-0.0125446	-0.00393137
rate_1	0.0892116	0.0888839	0.0267315	0.045927	0.13395
rate_2	-0.0711503	-0.0709699	0.0212949	-0.106687	-0.0366008
rate_3	-0.0100537	-0.00985093	0.00329948	-0.0158002	-0.00494428
income	0.0315915	0.0316969	0.00662684	0.020615	0.0424148
dProb(y=2)/dx					
⋮	⋮	⋮	⋮	⋮	⋮
dProb(y=3)/dx					
⋮	⋮	⋮	⋮	⋮	⋮

We start by interpreting the posterior means in the first block of parameters. The number associated with `price_0` suggests that if the price of fishing on the beach increases by one unit ($100), this will, on average, lead to a reduction in the probability of an individual choosing this mode of fishing by 12.4% according to the conditional Logit model and by 16.1% according to the conditional Probit. Similarly, a unit increase in the catch rate of fishing on the beach is expected to increase the probability of choosing this alternative by 1.8% and 6.6%, respectively. The number associated with `price_1` in the first block of the tables suggests that if the price of fishing on a private boat increases by one unit, this will lead to an average increase in the probability of an individual choosing fishing on the beach by 5.5% according to the conditional Logit model and by 7.8% according to the conditional Probit. The corresponding posterior mean of the marginal effect associated with `rate_1` implies that a unit increase in the catch rate of fishing by boat leads to an average reduction in the probability of choosing to fish on the beach by 0.8% and 3.2%, respectively. Finally, an increase in the monthly income of an individual by $1000 is expected to have a miniscule effect on the probability of fishing on the beach (reduction by 0.05% and 0.08% according to the conditional Logit and Probit models, respectively).

The interpretation of the remaining blocks of the marginal-effects tables follows along the same lines. The only difference is that the marginal effects in the second block, for example, are on the probability of an individual choosing the second alternative (fishing by private boat). In the second block we see that an increase in the alternative's own price leads to a reduction in the probability of choosing that alternative, while an increase in the catch rate of the alternative leads to an increase of this probability. Finally, an increase in the individual's income by $1000 leads, on average, to an increase in the probability of fishing on a private boat by 3.2% and 2.7%, respectively from the two models.

	Mean	Median	Sd.dev.	5%	95%
dProb(y=0)/dx					
price_0	-0.161285	-0.159096	0.0249924	-0.205604	-0.124101
price_1	0.0782552	0.0779895	0.0186328	0.0478232	0.10902
price_2	0.0616513	0.0615342	0.0162211	0.0350641	0.0884303
price_3	0.0213781	0.0186491	0.0165523	-0.000315232	0.0520839
rate_0	0.0663813	0.0646423	0.0183071	0.0395106	0.0990871
rate_1	-0.0322954	-0.0310713	0.0110687	-0.0523178	-0.0163819
rate_2	-0.025261	-0.0245107	0.00859937	-0.0406282	-0.0126247
rate_3	-0.00882493	-0.00739523	0.00731041	-0.0224607	0.000134144
income	0.0007846	0.000720556	0.00314279	-0.00425339	0.00604273
dProb(y=1)/dx					
price_0	0.0782379	0.0782139	0.0192069	0.0467666	0.109578
price_1	-0.265342	-0.265326	0.0279359	-0.310818	-0.219266
price_2	0.104429	0.102792	0.0251637	0.0661748	0.148261
price_3	0.0826751	0.0831587	0.0216344	0.0457653	0.117666
rate_0	-0.0322855	-0.0311056	0.0112234	-0.0524222	-0.016052
rate_1	0.108625	0.107493	0.024723	0.0699003	0.15109
rate_2	-0.0422906	-0.0414753	0.0117643	-0.0630251	-0.0243376
rate_3	-0.0340493	-0.0327333	0.0122079	-0.0560727	-0.0164999
income	0.0272776	0.0272413	0.00646498	0.0165572	0.0378988
dProb(y=2)/dx					
⋮	⋮	⋮	⋮	⋮	⋮
dProb(y=3)/dx					
⋮	⋮	⋮	⋮	⋮	⋮

The results presented above can be obtained in BayES using the code in the following box.

```
// import the data into a dataset called Data and generate a constat term
Data = webimport("www.bayeconsoft.com/datasets/FishingMode.csv");
Data.constant = 1;

// run the conditional Logit and Probit models
myCLogit = clogit( mode ~ price rate | constant income, "thin"=5, "chains"=2 );
myCProbit = cprobit( mode ~ price rate | constant income, "thin"=5, "chains"=2 );

// calculate marginal effects for the two models
mfx( "model"=myCLogit );
mfx( "model"=myCProbit );
```

7.5 Synopsis

This chapter introduced and covered in detail models designed to work with response variables that can be in one out of a finite number of states. These models are generalizations of the binary Probit and Logit models, examined in Chapter 6, in which the response variable can be in one out of two states. Such models are used in economics and most social sciences, in general, in the context of individual choice between different alternatives. Thus, they are also known as discrete-choice models and they can be justified using the random-utility framework. The independent variables that enter the specification of discrete-choice models can be the relevant characteristics of the individual making the choice or the attributes of the alternatives available to the decision maker. The type of independent variables is the feature that distinguishes the two main categories of models: multinomial models contain variables that vary only by

individual, while conditional models contain attributes as independent variables, although they can also accommodate individual characteristics.

As with models for binary response, the distributional assumptions imposed on the error terms give rise to Logit and Probit models. The choice probabilities in multinomial and conditional Logit models can be expressed in closed form and this is something that simplifies the analysis considerably. However, Logit models also assume independence of the odds ratio between choosing two alternatives from whether other alternatives are available to the decision maker or not. This assumption, known as Independence of Irrelevant Alternatives (IIA), is unrealistic in most applications. The multinomial and conditional Probit models relax this assumption by allowing the error terms to be correlated. However, this flexibility comes at the cost of having to employ very computationally intensive procedures for parameter estimation and calculation of marginal effects. This is because the choice probabilities in the Probit models cannot be expressed in closed form and they involve integrals, the dimensions of which increase with the number of alternatives. This is not a problem *per se* in a Bayesian context. However, integration by simulation is challenging in these models because it involves sampling from multidimensional truncated-Normal densities.

References

Adang, P., & Melenberg, B. (1995). Nonnegativity constraints and intratemporal uncertainty in a multi-good life-cycle model. *Journal of Applied Econometrics*, *10*(1), 1-15.

Aigner, D., Lovell, C., & Schmidt, P. (1977). Formulation and estimation of stochastic frontier production function models. *Journal of Econometrics*, *6*(1), 21-37.

Alder, M. (2005a). How to be much cleverer than all your friends (so they really hate you), PART I. *Philosophy Now*, *51*, 18-21.

Alder, M. (2005b). How to be much cleverer than all your friends (so they really hate you), PART II. *Philosophy Now*, *52*, 18-21.

Amemiya, T. (1974). Bivariate probit analysis: Minimum chi-square methods. *Journal of the American Statistical Association*, *69*(348), 940-944.

Ashford, J. R., & Sowden, R. R. (1970). Multi-variate probit analysis. *Biometrics*, *26*(3), 535-546.

Berger, J. O. (1985). *Statistical decision theory and bayesian analysis* (2nd ed.). New York, NY: Springer-Verlag.

Berger, J. O., Bernardo, J. M., & Sun, D. (2009). The formal definition of reference priors. *The Annals of Statistics*, *37*(2), 905-938.

Berndt, E. R., & Wood, D. O. (1975). Technology, prices, and the derived demand for energy. *The Review of Economics and Statistics*, *57*(3), 259-268.

Billingsley, P. (1995). *Probability and measure* (3rd ed.). New York, NY: John Wiley & Sons.

Bolduc, D. (1999). A practical technique to estimate multinomial probit models in transportation. *Transportation Research Part B: Methodological*, *33*(1), 63-79.

Brockmann, H. J. (1996). Satellite male groups in horseshoe crabs, Limulus polyphemus. *Ethology*, *102*(1), 1-21.

Burgette, L. F., & Nordheim, E. V. (2012). The trace restriction: An alternative identification strategy for the Bayesian multinomial probit model. *Journal of Business & Economic Statistics*, *30*(3), 404-410.

Carey, V., Zeger, S. L., & Diggle, P. (1993). Modelling multivariate binary data with alternating logistic regressions. *Biometrika*, *80*(3), 517-526.

Chamberlain, G. (1982). Multivariate regression models for panel data. *Journal of Econometrics*, *18*(1), 5-46.

Chamberlain, G. (1984). Panel data. In Z. Griliches & M. D. Intriligator (Eds.), *Handbook of econometrics* (Vol. 2, p. 1247-1318). Elsevier.

Chan, K. S., & Geyer, C. J. (1994). Discussion: Markov chains for exploring posterior distributions. *The Annals of Statistics*, *22*(4), 1747-1758.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, *90*(432), 1313-1321.

Chib, S. (2001). Markov chain Monte Carlo methods: Computation and inference. In J. J. Heckman & E. E. Leamer (Eds.), *Handbook of econometrics* (Vol. 5, p. 3569-3649). Elsevier.

Chib, S., & Greenberg, E. (1998). Analysis of multivariate probit models. *Biometrika*, *85*(2), 347-361.

Chib, S., & Jeliazkov, I. (2001). Marginal likelihood from the Metropolis-Hastings output. *Journal of the American Statistical Association*, *96*(453), 270-281.

Cortez, P., & Silva, A. (2008). Using data mining to predict secondary school student performance. In A. Brito & J. Teixeira (Eds.), *Proceedings of 5th future business technology conference (fubutec 2008)* (p. 5-12). Porto: EUROSIS.

Deaton, A., & Muellbauer, J. (1980). An almost ideal demand system. *American Economic Review*, *70*(3), 312-326.

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

Feenstra, R. C., Inklaar, R., & Timmer, M. P. (2015). The next generation of the Penn World Table. *American Economic Review*, *105*(10), 3150-3182.

Flegal, J. M., Haran, M., & Jones, G. L. (2008). Markov chain Monte Carlo methods: can we trust the third significant figure? *Statistical Science*, *23*(2), 250-260.

Frisch, R. (1933). Editor's note. *Econometrica*, *1*(1), 1-4.

Frühwirth-Schnatter, S., & Frülhwirth, R. (2012). Bayesian inference in the multinomial logit model. *Austrian Journal of Statistics*, *41*(1), 27-43.

Gelfand, A. E., & Dey, D. K. (1994). Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society, Series B: Methodological*, *56*(3), 501-514.

Gelfand, A. E., & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, *85*(410), 398-409.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). Boca Raton, FL: CRC Press.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*(6), 721-741.

Geweke, J. (1991). Efficient simulation from the multivariate normal and student-t distributions subject to linear constraints and the evaluation of constraint probabilities. In E. M. Keramidas & S. M. Kaufman (Eds.), *Computing science and statistics: Proceedings of the 23rd symposium on the interface* (p. 571-578). Interface Foundation of North America, Fairfax Station, VA.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In J. M. Bernardo, J. Berger, A. P. Dawid, & J. F. M. Smith (Eds.), *Bayesian statistics 4* (p. 169-193). Oxford: Oxford University Press.

Geweke, J. (1993). Bayesian treatment of the independent Student-t linear model. *Journal of Applied Econometrics*, *8*, S19-S40.

Glonek, G. F. V., & McCullagh, P. (1995). Multivariate logistic models. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(3), 533-546.

Greenberg, E. (2013). *Introduction to bayesian econometrics* (2nd ed.). New York, NY: Cambridge University Press.

Gupta, A. K., & Nadarajah, S. (2008). Normal and logistic random variables: distribution of the linear combination. *Statistical Papers*, *49*(2), 201-209.

Hajivassiliou, V., McFadden, D., & Ruud, P. (1996). Simulation of multivariate normal rectangle probabilities and their derivatives theoretical and computational results. *Journal of Econometrics*, *72*(1), 85-134.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*(1), 97-109.

Herriges, J. A., & Kling, C. L. (1999). Nonlinear income effects in random utility models. *Review of Economics and Statistics*, *81*(1), 62-72.

Howie, D. (2004). *Interpreting probability: Controversies and developments in the early twentieth century.* New York, NY: Cambridge University Press.

Imai, K., & van Dyk, D. A. (2005). A bayesian analysis of the multinomial probit model using marginal data augmentation. *Journal of Econometrics*, *124*(2), 311-334.

Jondrow, J., C, L., Materov, I. S., & Schmidt, P. (1982). On the estimation of technical inefficiency in the stochastic frontier production function model. *Journal of Econometrics*, *19*(2-3), 233-238.

Keane, M. P. (1994). A computationally practical simulation estimator for panel data. *Econometrica*, *62*(1), 95-116.

Koop, G. (2003). *Bayesian econometrics.* Chichester, UK: John Wiley & Sons.

Lancaster, T. (2004). *An introduction to modern bayesian econometrics.* Oxford, UK: Wiley-Blackwell.

Lewis, S. M., & Raftery, A. E. (1997). Estimating Bayes factors via posterior simulation with the Laplace-Metropolis estimator. *Journal of the American Statistical Association*, *92*(438), 648-655.

Li, K. (1988). Imputation using Markov chains. *Journal of Statistical Computation and Simulation*, *30*(1), 57-79.

Link, W. A., & Eaton, M. J. (2012). On thinning of chains in MCMC. *Methods in Ecology and Evolution*, *3*(1), 112-115.

Liu, C., Rubin, D. B., & Wu, Y. (1998). Parameter expansion to accelerate EM: The PX-EM algorithm. *Biometrika*, *85*(4), 112-115.

Liu, J., & Wu, Y. (1999). Parameter expansion for data augmentation. *Journal of the American Statistical Association*, *94*(448), 1264-1274.

Liu, J. S. (1994). The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. *Journal of the American Statistical Association*, *89*(427), 958-966.

Liu, X., & Daniels, M. J. (2006). A new algorithm for simulating a correlation matrix based on parameter expansion and reparameterization. *Journal of Computational and Graphical Statistics*, *15*(4), 897-914.

Malik, H. J., & Abraham, B. (1973). Multivariate logistic distributions. *The Annals of Statistics*, *1*(3), 588-590.

McCulloch, R., Polson, N. G., & Rossi, P. E. (2000). A Bayesian analysis of the multinomial probit model with fully identified parameters. *Journal of Econometrics*, *99*(1), 173-193.

McCulloch, R., & Rossi, P. E. (1994). An exact likelihood analysis of the multinomial probit model. *Journal of Econometrics*, *64*(1-2), 207-240.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers in econometrics* (p. 105-142). New York, NY: Academic Press.

Meeusen, W., & van den Broeck, J. (1977). Efficiency estimation from Cobb-Douglas production functions with composed error. *International Economic Review*, *18*(2), 435-444.

Meng, X.-L., & van Dyk, D. (1997). The EM algorithm – An old folk-song sung to a fast new tune. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *59*(3), 511-567.

Meng, X.-L., & van Dyk, D. (1999). Seeking efficient data augmentation schemes via conditional and marginal augmentation. *Biometrika*, *86*(2), 301-320.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, *21*(6), 1087-1092.

Mundlak, Y. (1978). On the pooling of time series and cross section data. *Econometrica*, *46*(1), 69-85.

Roberts, G. O., Gelman, A., & Gilks, W. R. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *The Annals of Applied Probability*, *7*(1), 110-120.

Rubin, D. B. (1978). Multiple imputations in sample surveys – A phenomenological Bayesian approach to nonresponse. In *Proceedings of the survey research methods section of the american statistical association* (p. 20-34).

Rubin, D. B. (1980). *Handling nonresponse in sample surveys by multiple imputations*. Monograph: U.S. Department of Commerce, Bureau of the Census.

Rungsuriyawiboon, S., & Stefanou, S. E. (2007). Dynamic efficiency estimation: An application to U.S. electric utilities. *Journal of Business & Economic Statistics*, *25*(2), 226-238.

Schaffner, B., & Ansolabhere, S. (2019). *2017 CCES Common Content* [Data]. Harvard Dataverse. Retrieved from https://doi.org/10.7910/DVN/3STEZY

Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, *82*(398), 528-540.

Thomson, C. J., & Crooke, S. T. (1991). *Results of the Southern California sportfish economic survey* (NOAA Technical Memorandum). National Marine Fisheries Service, Southwest Fisheries Science Center.

Train, K. E. (2009). *Discrete choice methods with simulation* (2nd ed.). Cambridge, MA: Cambridge University Press.

van Dyk, D. A. (2010). Marginal Markov chain Monte Carlo. *Statistica Sinica*, *20*(4), 1423-1454.

van Dyk, D. A., & Park, T. (2008). Partially collapsed Gibbs samplers: Theory and methods. *Journal of the American Statistical Association*, *103*(482), 790-796.

van den Broeck, J., Koop, G., Osiewalski, J., & Steel, M. F. J. (1994). Stochastic frontier models: A Bayesian perspective. *Journal of Econometrics*, *61*(2), 273-303.

Vella, F., & Verbeek, M. (1998). Whose wages do unions raise? A dynamic model of unionism and wage rate determination for young men. *Journal of Applied Econometrics*, *13*(2), 163-183.

Wooldridge, J. M. (2002). *Econometric analysis of cross section and panel data*. Cambridge, MA: MIT Press.

Zellner, A. (1962). An effcient method of estimating seemingly unrelated regressions and tests of aggregation bias. *Journal of the American Statistical Association*, *57*(298), 348-368.

Zellner, A. (1971). *An introduction to Bayesian inference in econometrics*. New York, NY: John Wiley & Sons.

Index