## Contents

1 Econometrics and Bayesian Inference
  1.1 Overview ................................................. 1
  1.2 Econometrics: Frequentist and Bayesian Approaches  1
  1.3 Bayes’ Theorem and Bayesian Inference ............... 3
    1.3.1 The Likelihood Function ......................... 5
    1.3.2 The Prior Density ................................ 6
    1.3.3 The Posterior Density ......................... 9
    1.3.4 Model Comparison and the Marginal Likelihood .... 12
    1.3.5 Prediction and Forecasting ..................... 14
    1.3.6 Discussion ......................................... 15
  1.4 Estimation by Simulation ............................... 15
    1.4.1 The Strong Law of Large Numbers and a Central Limit Theorem ... 16
    1.4.2 Markov-Chain Monte Carlo (MCMC) ................ 19
  1.5 Synopsis .............................................. 29

2 The Linear Model ........................................ 31
  2.1 Overview .............................................. 31
  2.2 Model Setup and Interpretation ....................... 31
  2.3 Likelihood, Priors and Posterior ..................... 33
  2.4 Full Conditionals and Parameter Estimation .......... 35
  2.5 Other Functional Forms and Marginal Effects ........ 38
  2.6 Post-Estimation Inference ............................ 42
    2.6.1 Imposing Parametric Restrictions and Evaluating their Plausibility ... 42
    2.6.2 Model Comparison in the Linear Regression Model .......... 45
    2.6.3 Predicting the Values of the Dependent Variable .... 47
  2.7 Synopsis .............................................. 49

3 Seemingly Unrelated Regressions ......................... 51
  3.1 Overview .............................................. 51
  3.2 The System Approach to Linear Regression ............ 51
  3.3 Likelihood, Priors and Full Conditionals ............. 54
  3.4 Cross-Equation Restrictions and the SUR Model ....... 58
    3.4.1 Demand Systems .................................. 58
    3.4.2 Cost Functions and Cost Share Equations ........... 59
    3.4.3 Imposing Linear Restrictions ..................... 60
  3.5 Synopsis .............................................. 62

4 Data Augmentation ....................................... 63
  4.1 Overview .............................................. 63
  4.2 Data Augmentation in Latent-Data Problems .......... 63
  4.3 Applications of Data Augmentation .................... 65
    4.3.1 The Linear Model with Heteroskedastic Error ....... 65
    4.3.2 The Stochastic Frontier Model .................... 69
  4.4 Marginal Data Augmentation ............................ 74
5 The Linear Model with Panel Data 77
5.1 Overview 77
5.2 Panel Data and Alternative Panel-Data Models 77
5.3 Estimation of the Hierarchical Panel-Data Models 80
  5.3.1 Estimation of the Random-Effects Model 80
  5.3.2 Estimation of the Random-Coefficients Model 83
5.4 Extensions to Other Panel-Data Models 86
  5.4.1 Correlated Random Effects 86
  5.4.2 Models with Group-Specific and Common Coefficients 87
  5.4.3 Random-Coefficients Models with Determinants of the Means 88
5.5 Synopsis 88

6 Models for Binary Choice 91
6.1 Overview 91
6.2 The Nature of Binary-Choice Models 91
  6.2.1 Random Utility: An Underlying Framework for Binary Choice 94
6.3 Estimation of Binary-Response Models 95
  6.3.1 Estimation of the Binary Probit Model 96
  6.3.2 Estimation of the Binary Logit Model 98
6.4 Interpretation of Parameters and Marginal Effects 100
6.5 Binary-Choice Models for Panel Data 102
6.6 Synopsis 105
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 $\text{Prob}(A)$ and $\text{Prob}(B)$, respectively. The probability of both $A$ and $B$ occurring in a repetition of the experiment is denoted by $\text{Prob}(A, B)$. Assuming that $\text{Prob}(B) \neq 0$, the probability of $A$ occurring, given that $B$ has occurred is:

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

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

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

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:

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

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

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

Bayes’ theorem reverses the roles of the two events in conditioning and, by doing so, allows calculation of $\text{Prob}(A|B)$ using knowledge of $\text{Prob}(B|A)$. More precisely and in the context of Bayesian inference, knowledge of $\text{Prob}(B|A)$ and $\text{Prob}(B)$ 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:

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

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 \( \text{Prob}(DL) = 0.3 \). Given the additional information of delayed departure, the probability of delayed landing becomes:

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

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 \( p(x) \) and \( p(y) \), respectively. Using these probability density functions Bayes’ theorem can be expressed as:

\[
p(x|y) = \frac{p(y|x) \cdot p(x)}{p(y)}
\]

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 \( \theta \) and the data in a vector \( y \) and by renaming some of the densities, the theorem becomes:

\[
\pi(\theta|y) = \frac{p(y|\theta) \cdot p(\theta)}{m(y)}
\]

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. \( \pi(\theta|y) \) 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. \( p(y|\theta) \) 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. \( p(\theta) \) 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. \( m(y) \) 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 \( \theta \) 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 \( m(y) \) does not involve \( \theta \), the denominator in the last expression can be viewed as a constant of proportionality for \( \pi(\theta|y) \). This constant is irrelevant for purposes of estimation and can be ignored at this stage. In practice, therefore, it is most often omitted from (1.6) and Bayes’ theorem is expressed as:

\[
\pi(\theta|y) \propto p(y|\theta) \cdot p(\theta)
\]

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 \( p(y|\theta) \), where \( y \) are the data and \( \theta \) the model’s parameters. The meaning of \( \theta \) is straightforward, but the meaning of \( y \) deserves some discussion. With a given dataset at hand, \( y \) 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 \( \theta \) 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 \( y \) and \( \theta \) 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 \( y \) be a vector that contains the observed data. Therefore, \( y \) 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 \( \lambda \) is the rate parameter. In our application \( \lambda \) 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 \( \lambda \): \( y_i \sim \text{Exp}(\lambda) \). The likelihood function is the density of all data points and, because of the independence assumption, this density can be expressed as:

\[
p(y|\lambda) = \prod_{i=1}^{N} \lambda e^{-\lambda y_i} = \lambda^N e^{-\lambda \sum_{i=1}^{N} y_i}
\]
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 $p(\theta)$. 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*.

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 express 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 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

---

3Throughout 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 \( \theta \) should convey the same amount of information as for a monotonic transformation of \( \theta \), 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 \( \infty \). 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, $\theta$, in a model is called **conjugate** if it leads to a posterior for $\theta$ 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 $\lambda$. 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:

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

$a$ 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 $\lambda$ 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 $\lambda$ to deviate from this expectation, if the data suggest so. We can increase the variance of $\lambda$ 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 $\lambda$ 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.

<table>
<thead>
<tr>
<th>Hyperparameters</th>
<th>E ($\lambda$)</th>
<th>V ($\lambda$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a = 8$, $b = 2$</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>$a = 4$, $b = 1$</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>$a = 2$, $b = 0.5$</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>$a = 1$, $b = 0.25$</td>
<td>4</td>
<td>16</td>
</tr>
</tbody>
</table>

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.

```plaintext
// 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 $\pi(\theta|y)$ and it expresses our knowledge about $\theta$ after having seen the data. The posterior density is obtained from an application of Bayes’ theorem:

$$\pi(\theta|y) = \frac{p(y|\theta) \cdot p(\theta)}{m(y)} \propto p(y|\theta) \cdot p(\theta)$$

(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 $\pi(\theta|y)$ 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 $\theta$. For example, if $\theta$ consists of two parameters, $\theta_1$ and $\theta_2$, the results are presented in a table of the following form:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Mean</th>
<th>St.dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta_1$</td>
<td>$E(\theta_1</td>
<td>y)$</td>
</tr>
<tr>
<td>$\theta_2$</td>
<td>$E(\theta_2</td>
<td>y)$</td>
</tr>
</tbody>
</table>

where:

$$E(\theta_1|y) = \int_{\Theta_1} \theta_1 \cdot \pi(\theta_1|y) \, d\theta_1$$

$$V(\theta_1|y) = \int_{\Theta_1} \left( \theta_1 - E(\theta_1|y) \right)^2 \cdot \pi(\theta_1|y) \, d\theta_1$$

(1.9)

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

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

(1.10)

Similar calculations should be performed for $\theta_2$.

If the model contains more than two parameters, calculating the moments involves multidimensional integration. However, these integrals are rarely evaluated analytically in practice. If $\pi(\theta|y)$ 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 $\pi(\theta|y)$ 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 $\theta_1$ being within an interval $[c_1, c_2]$ can be expressed as:

$$\text{Prob}(c_1 \leq \theta_1 \leq c_2) = \int_{c_1}^{c_2} \pi(\theta_1|y) \, d\theta_1$$

(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 \( \theta_1 \) 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 \( \theta_1 \) 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 \( \theta_1 \), as it states that the probability of \( \theta_1 \) 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 \( \theta \) the two concepts, however, become apparent once the meaning of a confidence interval is

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

respectively. Using Bayes’ theorem, the posterior density of \( \lambda \) from this model is:

\[ \pi (\lambda|y) \propto p(y|\lambda) \cdot p(\lambda) \]

\[ = \lambda^N e^{-\lambda \sum_{i=1}^{N} y_i} \times \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b\lambda} \]

\[ \propto \lambda^{(N+a)-1} e^{-\lambda \left( \sum_{i=1}^{N} y_i + b \right)} \]

where \( \frac{b^a}{\Gamma(a)} \) is dropped from the final expression and, because it does not involve \( \lambda \), 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:

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

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: \( c = \frac{b^a}{\Gamma(a)} \). Therefore:

\[ \lambda|y \sim \text{Gamma}(\tilde{a}, \tilde{b}) \]

From the properties of the Gamma distribution we get \( \text{E}(\lambda|y) = \frac{\tilde{a}}{\tilde{b}} \) and \( \text{V}(\lambda|y) = \frac{\tilde{a}}{\tilde{b}^2} \). 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.

<table>
<thead>
<tr>
<th>Hyperparameters</th>
<th>$E(\lambda \mid y)$</th>
<th>$V(\lambda \mid y)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a = 8$, $b = 2$</td>
<td>4.4048</td>
<td>0.1797</td>
</tr>
<tr>
<td>$a = 4$, $b = 1$</td>
<td>4.4220</td>
<td>0.1880</td>
</tr>
<tr>
<td>$a = 2$, $b = 0.5$</td>
<td>4.4312</td>
<td>0.1925</td>
</tr>
<tr>
<td>$a = 1$, $b = 0.25$</td>
<td>4.4359</td>
<td>0.1948</td>
</tr>
</tbody>
</table>

The posterior densities of $\lambda$ 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 $E(\lambda \mid y)$ 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 $\lambda$ as:

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

The only difference between the formulas for $\hat{\lambda}_{\text{MLE}}$ and $E(\lambda \mid y)$ 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, $E(\lambda \mid y)$ converges to the frequentist point estimate. Using the same dataset, the maximum-likelihood estimate of $\lambda$ is $4.4407$.

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

```plaintext
// 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
```
CHAPTER 1. ECONOMETRICS AND BAYESIAN INFERENCE

// 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:

\[
\begin{align*}
\text{Model 0: } & \quad p_0(y_i | \theta_0, \cdot), & \quad p_0(\theta_0) \\
\text{Model 1: } & \quad p_1(y_i | \theta_1, \cdot), & \quad p_1(\theta_1)
\end{align*}
\]

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 “•”, 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.
1.3. BAYES' THEOREM AND BAYESIAN INFERENCE

applying Bayes’ theorem to Model 0 we obtain:

\[
\text{Prob}(M = 0|y) = \frac{m(y|M = 0) \cdot \text{Prob}(M = 0)}{m(y)}
\]  

(1.13)

where \( \text{Prob}(M = 0) \) and \( \text{Prob}(M = 0|y) \) are, respectively, the prior and posterior model probabilities for Model 0. \( m(y|M = 0) \) 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:

\[
m(y|M = 0) = \int_{\theta_0} p_0(y|\theta_0, \bullet) \cdot p_0(\theta_0) \, d\theta_0
\]  

(1.14)

It becomes apparent from this discussion that \( m(y|M = 0) \) 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, \( m(y) \) 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 \( m(y) \), 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:

\[
\text{Prob}(M = 1|y) = \frac{m(y|M = 1) \cdot \text{Prob}(M = 1)}{m(y)}
\]  

(1.15)

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

\[
\frac{\text{Prob}(M = 0|y)}{\text{Prob}(M = 1|y)} = \frac{m(y|M = 0)}{m(y|M = 1)} \cdot \frac{\text{Prob}(M = 0)}{\text{Prob}(M = 1)}
\]  

(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:

\[
\begin{align*}
\text{Prob}(M = 0|y) &= c \cdot \text{Prob}(M = 1|y) \\
\text{Prob}(M = 0|y) + \text{Prob}(M = 1|y) &= 1
\end{align*}
\]  

(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 \( \sum_{j=1}^{J} \text{Prob}(M = j|y) = 1 \)

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 \( m(y|M = j) \) 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 $y$ 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:

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

(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 $p(\mathbf{y}_*|\theta, y)$ with respect to $\theta$ and then by conditioning on $\theta$ inside the integral. $\pi(\theta|y)$ 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 $p(\mathbf{y}_*|\theta, y)$ to $p(\mathbf{y}_*|\theta)$. In such cases, all information contained in the observed data is transmitted to the parameters and $y$ 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 $\lambda$. Using a Gamma prior we expressed the posterior density of $\lambda$ as a Gamma density with shape parameter $a = N + a$ and rate parameter $b = \sum_{i=1}^N y_i + b$, 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:

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

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

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

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:

$$p(y_*|y) = a \cdot b^a \cdot (b + y_*)^{-a-1}$$
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 “choke”. 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 $E(|X|) < \infty$. Then:

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

where “$\xrightarrow{a.s.}$” 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, $\theta$, whose posterior density function is $\pi(\theta | y)$. To summarize the information contained in this posterior density we would like, first of all, to evaluate the expectation:

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

Suppose that we can obtain $G$ random draws from the distribution whose probability density function is $\pi(\theta | y)$ 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:

$$\bar{\theta}_G = \frac{\sum_{g=1}^{G} \theta^{(g)}}{G} \quad (1.20)$$

converges to $E(\theta | y)$. 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 $\theta$ can also be expressed as an expectation:

$$V(\theta | y) = \int_{\Theta} (\theta - E(\theta | y))^2 \cdot \pi(\theta | y) \, d\theta \quad (1.21)$$

---

6 An alternative term used frequently in the context of Bayesian inference instead of simulation methods 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 $\theta$, we can invoke the SLLN a second time to approximate this variance by the sample variance:

$$s_\theta^2 = \frac{\sum_{g=1}^{G} (\theta^{(g)} - \bar{x}_g)^2}{G - 1}$$  \hspace{1cm} (1.22)$$

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

$$\text{Prob}(c_1 \leq \theta \leq c_2 | y) = \int_{c_1}^{c_2} \theta \cdot \pi(\theta | y) \, d\theta = \int_{\Theta} \mathbb{1}(c_1 \leq \theta \leq c_2) \cdot \pi(\theta | y) \, d\theta$$  \hspace{1cm} (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 $\theta$ and the SLLN suggests that the formula:

$$\frac{\sum_{g=1}^{G} \mathbb{1}(c_1 \leq \theta^{(g)} \leq c_2)}{G}$$  \hspace{1cm} (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 $E(X) = \mu$ and $V(X) = \sigma^2$. Then the sample mean, $\hat{\mu}_G = \frac{1}{G} \sum_{g=1}^{G} X_g$, converges in distribution to a Normal:

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

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 $\sigma^2$ is unknown, it is replaced by the sample variance, which is a consistent estimator of the population variance: as $G$ goes to infinity, $s_\theta^2$ converges in probability to $\sigma^2$. 

Example 1.2 Customer Arrival Rate (Continued)

Using a Gamma prior for the arrival-rate parameter, $\lambda$, in the sandwich-store example we expressed the posterior density of $\lambda$ 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 $E(\lambda | y) = \frac{\tilde{a}}{\tilde{b}}$ and $V(\lambda | y) = \frac{\tilde{a} \cdot \tilde{b}^2}{\tilde{b}^2}$. We will now approximate these two posterior moments by drawing random numbers from $\pi(\lambda | y)$. 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: $E(\lambda | y) = 4.4359$ and $V(\lambda | y) = 0.1948$.

| # of draws $(G)$ | $E(\lambda | y)$   | $V(\lambda | y)$   | 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.

```plaintext
// 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, $\theta_1$ and $\theta_2$ and suppose that we can obtain $G$ random draws from this joint density. Denote these draws by $\theta^1, \theta^2, \ldots, \theta^G$, where each $\theta^g$ is a two-dimensional vector. Equations (1.9) and (1.10) suggest that, to approximate $E(\theta_1 | y)$, one needs to first integrate-out $\theta_2$ from $\pi(\theta_1, \theta_2 | y)$:

$$E(\theta_1 | y) = \int_{\theta_1} \theta_1 \cdot \pi(\theta_1 | y) \, d\theta_1 = \int_{\theta_1} \theta_1 \cdot \left[ \int_{\theta_2} \pi(\theta_1, \theta_2 | y) \, d\theta_2 \right] \, d\theta_1$$

(1.25)

Given that the draws are from the joint posterior density of $\theta_1$ and $\theta_2$, simply summarizing the values of $\theta_1$ contained in $\theta^1, \theta^2, \ldots, \theta^G$ would take care of both integrals:

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

(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, \( \pi(\theta | y) \). 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: \[ \sigma^2 = V(X_1) + 2 \sum_{j=2}^{\infty} \text{Cov}(X_1, X_j) \] (1.27)

where \( V(X_1) \) is the variance of a random variable that follows the same distribution as the stationary distribution of the Markov chain, and \( \text{Cov}(X_1, X_j) \) 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:

\[ \hat{\kappa} = \frac{\hat{\sigma}^2}{s^2/G} \] (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, \( \hat{G} \), that would lead to the same Monte Carlo standard error. \( \hat{G} \) 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 $\pi(\theta | y)$. The algorithm starts by fixing an initial value for $\theta$. Based on this initial value it proposes a move to a new value, $\theta^*$, using a proposal density, $q(\theta, \theta^* | y)$. The proposal density is chosen by the researcher and, at least in theory, could be any proper probability density function. Lastly, the proposed value, $\theta^*$, is accepted with probability:

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

(1.29)

This means that the new value of $\theta$ becomes $\theta^*$ and the process is repeated multiple times. If the move to $\theta^*$ 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 $\theta$ and with covariance matrix $C$:

$$
q(\theta, \theta^* | y) = \left\{ C \right\}^{-1/2} \exp \left\{ -\frac{1}{2} (\theta - \theta^*)' C^{-1} (\theta - \theta^*) \right\}
$$

(1.30)

leading to the random-walk Metropolis-Hastings algorithm. This choice is convenient because $q(\theta^*, \theta | y) = q(\theta, \theta^* | y)$ for all $\theta$ and $\theta^*$ 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 $\theta^*$ and $\theta$. 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, $C$, 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 $C$ equal to $\mathcal{T} \cdot I_K$, where $I_K$ 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 $\theta^*$ may be too erratic and, therefore, rarely accepted. When $\mathcal{T}$ is set to a small value, the proposed $\theta^*$ will be frequently accepted, but this will be because $\theta^*$ is very close to the current state of the Markov chain, $\theta$. 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, $\pi(\theta | y)$. 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 $\theta$ 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 \(\theta\) to a reasonable starting value

for \(q = 1 : (D+G)\) do
    draw \(\theta^*\) from the proposal, \(q(\theta, \theta^*|y)\)
    accept the move (set current \(\theta\) equal to \(\theta^*\)) with probability:
    \[
    \alpha(\theta, \theta^*|y) = \min \left\{ \frac{\pi(\theta^*|y)}{\pi(\theta|y)} \frac{q(\theta^*|\theta, y)}{q(\theta|\theta^*, y)}, 1 \right\}
    \]
    if \(g > D\) then
        store the current value of \(\theta\)
    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 \(\lambda\), 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 \(\lambda\) becomes:

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

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 \(\lambda\) must be positive, we will use a log-Normal proposal, with location parameter equal to the logarithm of \(\lambda\) in the current iteration:

\[
q(\lambda, \lambda^*|y) = \frac{\sqrt{T}}{\lambda^* (2\pi)^{1/2}} \exp \left\{ -\frac{(\log \lambda^* - \log \lambda)^2}{2T} \right\}
\]

where \(T\) will be used as the tuning parameter. The logarithm of the Metropolis-Hastings ratio becomes:

\[
\log \text{MH}(\lambda, \lambda^*) = \log \pi(\lambda^*|y) - \log \pi(\lambda|y) + \log q(\lambda^*, \lambda|y) - \log q(\lambda, \lambda^*|y)
\]

\[
= N (\log \lambda^* - \log \lambda) - (\lambda^* - \lambda) \sum_{i=1}^{N} y_i - \frac{(\log \lambda^* - m)^2}{2s^2} + \frac{(\log \lambda - m)^2}{2s^2}
\]

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 \(T = 0.3\), we are ready to implement the algorithm. Note that these values of the hyperparameters lead to a prior density for \(\lambda\) 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 \(T\) is chosen such that approximately 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 \(\lambda\) obtained after running this code are \(E(\lambda|y) = 4.4445\) and \(V(\lambda|y) = 0.1907\), respectively.
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 $\theta$ 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, $\theta$, into $B \geq 2$ blocks, $\theta_1, \theta_2, \ldots, \theta_B$. Next, the density of each block, $\theta_b$, called this block’s full conditional, is obtained from the posterior density of $\theta$ by treating the parameters contained in all other blocks except $\theta_b$ as fixed. The full conditional for block $\theta_b$ is denoted by $\pi(\theta_b|y, \theta_1, \ldots, \theta_{b-1}, \theta_{b+1}, \ldots \theta_B)$, or, more compactly, by $\pi(\theta_b|\cdot)$. In this notation the origin of the term “full conditional” becomes apparent: $\pi(\theta_b|\cdot)$ is the density of $\theta_b$ 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
1.4. ESTIMATION BY SIMULATION

part of the constant of proportionality of $\pi (\theta_b \cdot)$. 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 $\theta_b^*$ are proposed from block $b$’s proposal density and they are accepted or rejected using the acceptance probability in (1.29) and while using $\pi (\theta_b \cdot)$ in place of $\pi (\theta | y)$. 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 $\theta_1, \ldots, \theta_B$ to reasonable starting values

for $g = 1: (D+G)$ do
  for $b = 1:B$ do
    draw $\theta_b^*$ from its proposal, $q(\theta_b, \theta_b^* | y)$
    accept the move (set current $\theta_b$ equal to $\theta_b^*$) with probability:
    $$\alpha_b (\theta_b, \theta_b^* | y) = \min \left\{ \frac{\pi(\theta_b^* \cdot) q(\theta_b^{\ast}, \theta_b | y)}{\pi(\theta_b \cdot) q(\theta_b, \theta_b^* | y)}, 1 \right\}$$
  end for

  if $g > D$ then
    store the current value of $\theta$
  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 $\mu$ and variance $\sigma^2$. 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 $y_i \sim \text{N} (\mu, \tau)$. Thus, the likelihood:

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

We will place a Normal prior on $\mu$, with mean $m$ and precision $t$, and a Gamma prior on $\tau$ with shape and rate parameters $a$ and $b$, respectively. Application of Bayes’ rule leads to:

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

We will treat $\mu$ and $\tau$ as the two blocks of the Metropolis-Hastings algorithm, with full conditionals:

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

$$\pi (\tau | y, \mu) \propto \tau^{a-1} e^{-b\tau}$$
and:
\[
\pi(\tau|\cdot) \propto \tau^{N/2+a-1} \exp\left\{ -\tau \left[ \frac{1}{2} \sum_{i=1}^{N} (y_i - \mu)^2 + b \right] \right\}
\]
respectively. Finally, we will use a Normal distribution as a proposal for \(\mu\) and a log-Normal distribution for \(\tau\). The mean of the proposal for \(\mu\) will be the current value of parameter and its precision parameter, \(T_{\mu}\), will be used as the tuning parameter. The location parameter for \(\tau\)’s proposal is set equal to the logarithm of the current value of \(\tau\) and its scale parameter, \(T_{\tau}\), 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.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Mean</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\mu)</td>
<td>3.26548</td>
<td>3.104e-05</td>
</tr>
<tr>
<td>(\tau)</td>
<td>157.192</td>
<td>278.411</td>
</tr>
<tr>
<td>(\sigma)</td>
<td>0.08010</td>
<td>1.834e-05</td>
</tr>
</tbody>
</table>

```plaintext
// 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
```
1.4. ESTIMATION BY SIMULATION

```matlab
// 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, $\theta$, has been partitioned into $B$ blocks, $\theta_1, \theta_2, \ldots, \theta_B$. Suppose also that the full conditional of block $\theta_b$, $\pi(\theta_b|\bullet)$, 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. Then, the proposal density for $\theta_b$, $q(\theta_b, \theta_b^*|y)$ can be set to be independent of the current values of $\theta$ and equal to $\pi(\theta_b^*|\bullet)$. In this case the Metropolis-Hastings ratio simplifies to one for all values of $\theta_b^*$ and the proposed move is accepted with probability one. This fact simplifies the procedure of updating the values of $\theta_b$ considerably and, if these simplifications can be performed for all blocks of $\theta$, 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

1. set the number of burn-in iterations, $D$
2. set the number of draws to be retained, $G$
3. set $\theta_1, \ldots, \theta_B$ to reasonable starting values
4. for $g = 1:(D+G)$ do
   5. for $b = 1:B$ do
      6. draw $\theta_b^*$ from $\pi(\theta_b^*|\bullet)$ and set the current value of $\theta_b$ to $\theta_b^*$
   7. end for
8. if $g > D$ then
   9. store the current value of $\theta$
10. end if
11. 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, $\pi(\theta_b|\bullet)$ can be derived without missing a constant of proportionality and the term *complete conditional* may be used to refer to $\pi(\theta_b|\bullet)$. 

mean parameter, \( \mu \), of this distribution and a Gamma prior for the precision parameter, \( \tau \), we obtained the following 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:

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

respectively. Let’s start from the full conditional of \( \tau \). This density resembles the probability density function of Gamma-distributed random variable with shape and rate parameters \( \tilde{a} = \frac{N}{2} + a \) and \( \tilde{b} = \frac{1}{2} \sum_{i=1}^{N} (y_i - \mu)^2 + b \), 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:

\[
\tau|y, \mu \sim \text{Gamma}(\tilde{a}, \tilde{b})
\]

Additional algebra is required to show that the full conditional of \( \mu \) is proportional to:

\[
\exp\left\{-\frac{t}{2} (\mu - \tilde{m})^2\right\}, \quad \text{with} \quad \tilde{m} = \frac{1}{\tilde{t}} \left( \tau \sum_{i=1}^{N} y_i + t \cdot m \right).
\]

The steps involved in this derivation are expanding the squares in the original expression for the full conditional of \( \mu \), 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 \( \mu \) resembles a Normal probability density function with mean \( \tilde{m} \) and precision parameter \( \tilde{t} \). With \( \frac{\tilde{m}^2}{(2\pi)^{1/2}} \) as the constant of proportionality, we get:

\[
\mu|y, \tau \sim N\left(\tilde{m}, \frac{1}{\tilde{t}}\right)
\]

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 the following table.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Mean</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \mu )</td>
<td>3.26638</td>
<td>3.713e-05</td>
</tr>
<tr>
<td>( \tau )</td>
<td>156.934</td>
<td>289.268</td>
</tr>
<tr>
<td>( \sigma )</td>
<td>0.08018</td>
<td>1.928e-05</td>
</tr>
</tbody>
</table>

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.
1.4. ESTIMATION BY SIMULATION

---

```matlab
// 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

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, $\theta_b$, 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. This version of the algorithm works by analytically marginalizing a block from one or more full conditionals. For example, consider a model that contains three blocks of parameters, $\theta_1$, $\theta_2$ and $\theta_3$. Suppose also that $\theta_2$ can be
integrated out analytically from the joint density of $\theta_1$ and $\theta_2$ given $\theta_3$, resulting in an expression of the form:

$$
\pi (\theta_1 | y, \theta_3) = \int \pi (\theta_1, \theta_2 | y, \theta_3) \, d\theta_2
$$

(1.31)

If sampling from $\pi (\theta_1 | y, \theta_3)$ can be accomplished easily, a collapsed Gibbs sampler would take draws from $\pi (\theta_1 | y, \theta_3)$ instead of $\pi (\theta_1 | y, \theta_2, \theta_3)$ and from the regular full conditionals of $\theta_2$ and $\theta_3$. In a simple setting, the collapsed Gibbs sampler can be justified by blocking $\theta_1$ and $\theta_2$ together: instead of sampling iteratively from $\pi (\theta_1 | y, \theta_2, \theta_3)$, $\pi (\theta_2 | y, \theta_1, \theta_3)$ and $\pi (\theta_3 | y, \theta_1, \theta_2)$, one could think of $(\theta_1, \theta_2)$ as constituting a single block and sample iteratively from $\pi (\theta_1, \theta_2 | y, \theta_3)$ and $\pi (\theta_3 | y, \theta_1, \theta_2)$. The former density can be expressed as:

$$
\pi (\theta_1, \theta_2 | y, \theta_3) = \pi (\theta_2 | y, \theta_1, \theta_3) \times \pi (\theta_1 | y, \theta_3)
$$

(1.32)

thus justifying sampling from $\pi (\theta_1 | y, \theta_3)$ instead of sampling from $\pi (\theta_1 | y, \theta_1, \theta_3)$. However, 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 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 $x$s 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:

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

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 \(i\). \(\varepsilon_i\) is the disturbance or error term associated with observation \(i\) and it captures any remaining variability in \(y\) that cannot be explained by the \(x_s\) in this model. The error term may be non-zero for multiple reasons. Firstly, even the original functional relationship between \(y\) and the \(x_s\) 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 \(x_s\).

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 \(x_s\) 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:

\[
y_i = x_i'\beta + \varepsilon_i
\]

where \(x_i\) is a \(K \times 1\) vector that stores the values of the independent variables for a potential observation \(i\) and \(\beta\) is a \(K \times 1\) vector that contains the \(\beta\) 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 \(\varepsilon_i\) are random variables and \(x_i\) is a random vector. We will assume in this chapter that the error term is independent of \(x_i\), and that it follows a Normal distribution with mean zero and variance \(\sigma^2\). The variance of the error term is another parameter to be estimated along with the \(\beta\)s and, to simplify the algebra necessary to proceed with estimation, we will express the distribution of \(\varepsilon_i\) in terms of the precision parameter: \(\varepsilon_i \sim N(0, \frac{1}{\tau})\), where \(\tau \equiv \frac{1}{\sigma^2}\). From the properties of the Normal distribution and given the independence assumption, we get that \(y_i|x_i \sim N(x_i'\beta, \frac{1}{\tau})\). This, in turn, implies that:

\[
E(y_i|x_i) = x_i'\beta
\]

In words, the expected value of \(y_i\), 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 \(x_i\), 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 \(x_i\) 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 \(x_i\) 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 \(x_i\), with expected value \(x_i'\beta\) 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 $\beta$ 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:

$$\beta_k = \frac{\partial E(y_i|x_i)}{\partial x_{ik}}$$  \hspace{1cm} (2.4)

Using econometric jargon, $\beta_k$ 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:

$$y = X\beta + \varepsilon$$  \hspace{1cm} (2.5)

where:

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}, \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{bmatrix} \quad \text{and} \quad X = \begin{bmatrix} x_1' \\ x_2' \\ \vdots \\ x_N' \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1K} \\ x_{21} & x_{22} & \cdots & x_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{NK} \end{bmatrix}$$

In this notation, $\varepsilon$ 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 $X$, independent from one another, the covariance of this Normal distribution is $\frac{1}{\tau}I_N$. From the properties of the Normal distribution, this representation implies that $y|X \sim N(X\beta, \frac{1}{\tau}I_N)$ and $E(y|X) = X\beta$. This representation still describes the data-generating process in the population and the term “potential observation” was used multiple times to stress that $y$ and $X$ 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:

$$p(y|\beta, \tau) = \prod_{i=1}^{N} p(y_i|x_i, \beta, \tau) = \prod_{i=1}^{N} \frac{\tau^{1/2}}{(2\pi)^{1/2}} \exp \left\{ -\frac{\tau}{2}(y_i - x_i'\beta)^2 \right\}$$  \hspace{1cm} (2.6)

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

$$p(y|X, \beta, \tau) = \frac{\tau^{N/2}}{(2\pi)^{N/2}} \exp \left\{ -\frac{\tau}{2}(y - X\beta)'(y - X\beta) \right\}$$  \hspace{1cm} (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. $y, X$ and $\varepsilon$ 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 $y|X$. 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, \( y \) and \( X \) 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: \( \beta \) and \( \tau \). This is because the slope coefficients in \( \beta \) define the conditional mean of \( y \) and enter the likelihood function as a group, while \( \tau \) defines the variance of \( y \) and appears in the likelihood function in entirely different places. Additionally, we do not need to impose any restrictions on the values of \( \beta \), except if economic theory requires so, while we need to restrict \( \tau \) 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 \( \tau \) is Gamma with hyperparameters \( a \) and \( b \) and the prior for \( \beta \) conditional on \( \tau \) is multivariate Normal with mean \( m \) and variance matrix \( \frac{1}{\tau}V \), where \( m \) and \( V \) are \( \beta \)’s hyperparameters. Because this prior is conjugate, the joint posterior density of \( \beta \) and \( \tau \) 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 \( \beta \) depends on \( \tau \)’s hyperparameters and this may pose some problems when eliciting the values of \( a \), \( b \), \( m \) and \( V \) 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 \( \tau \) follows a Gamma distribution, but we will assume that \( \beta \) follows a multivariate Normal distribution, marginally with respect to \( \tau \):

\[
p(\beta) = \frac{|P|^{1/2}}{(2\pi)^{K/2}} \exp \left\{ -\frac{1}{2} (\beta - m)' P (\beta - m) \right\} \quad \text{and} \quad p(\tau) = \frac{b^a}{\Gamma(a)} \tau^{a-1} e^{-br} \quad (2.8)
\]

It is stressed that, although we assume that \( \beta \) and \( \tau \) 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 \( \beta \) in terms of the inverse of the variance matrix: \( P \) 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:

\[
\pi (\beta, \tau|y, X) \propto p(y|X, \beta, \tau) \times p(\beta) \times p(\tau)
\]

\[
= \frac{\tau^{N/2}}{(2\pi)^{N/2}} \exp \left\{ -\frac{\tau}{2} (y - X\beta)' (y - X\beta) \right\} \times \frac{|P|^{1/2}}{(2\pi)^{K/2}} \exp \left\{ -\frac{1}{2} (\beta - m)' P (\beta - m) \right\} \times \frac{b^a}{\Gamma(a)} \tau^{a-1} e^{-br} \quad (2.9)
\]

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 \( \beta \) and \( \tau \) 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 $\beta$ and $\tau$ 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 $\tau$. By dropping terms from the posterior that enter multiplicatively and do not involve $\tau$ we get:

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

Next, collecting terms leads to:

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

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

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

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

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

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

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

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

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

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

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

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 $\exp \left\{ -\frac{1}{2} \tilde{m}' \tilde{P} \tilde{m} \right\}$ 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:

$$ y_i = x_i' \beta + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \frac{1}{\tau}) $$

and with a Normal prior for $\beta$ and a Gamma prior for $\tau$:

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

the full conditionals of $\beta$ and $\tau$ are, respectively, Normal and Gamma:

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

where:

$$ \tilde{P} = \tau X'X + P \quad \text{and} \quad \tilde{m} = (\tau X'X + P)^{-1} (\tau X'y + Pm) $$

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, $\beta$ and $\tau$, within each iteration. We will not present an implementation here, but the interested reader is directed to the lm.bs function 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:

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

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

<table>
<thead>
<tr>
<th>Mean</th>
<th>Median</th>
<th>Sd.dev.</th>
<th>5%</th>
<th>95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>constant</td>
<td>-4.72133</td>
<td>-4.73423</td>
<td>5.43244</td>
<td>-13.6764</td>
</tr>
<tr>
<td>pFood</td>
<td>0.0725127</td>
<td>0.0730401</td>
<td>0.0445887</td>
<td>-0.00084805</td>
</tr>
<tr>
<td>pOther</td>
<td>0.0140819</td>
<td>0.0140129</td>
<td>0.0311635</td>
<td>-0.0374249</td>
</tr>
<tr>
<td>expOther</td>
<td>0.0188497</td>
<td>0.0188519</td>
<td>0.00155341</td>
<td>0.016298</td>
</tr>
<tr>
<td>Hsize</td>
<td>0.429952</td>
<td>0.429473</td>
<td>0.0499011</td>
<td>0.348149</td>
</tr>
<tr>
<td>Children</td>
<td>-0.207939</td>
<td>-0.207817</td>
<td>0.0731264</td>
<td>-0.328807</td>
</tr>
<tr>
<td>tau</td>
<td>0.178622</td>
<td>0.178583</td>
<td>0.00411218</td>
<td>0.171855</td>
</tr>
<tr>
<td>sigma_e</td>
<td>2.36657</td>
<td>2.36636</td>
<td>0.0272516</td>
<td>2.3221</td>
</tr>
</tbody>
</table>

The first column of the table contains the names of the variables with which each parameter is associated. For example, $\beta_1$ is associated with a constant variable ($\beta_1$ is always multiplied by...
one in the model), $\beta_2$ 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 mean 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 $\tau$ parameter of the model. The two subplots at the top present a history and a correlogram of the draws for $\tau$ 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.

```R
// 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 not require the relationship between \( y \) and the \( x \)s 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:

\[
y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i
\]  
\[(2.17)\]

Furthermore, the marginal effects of \( x_2 \) and \( x_3 \) are \( \beta_2 \) and \( \beta_3 \), respectively. However, the model:

\[
y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i2}^2 + \beta_5 x_{i2} x_{i3} + \varepsilon_i
\]  
\[(2.18)\]

is linear in the parameters as well, although it is non-linear in the \( x \)s: \( x_2 \) enters this expression once linearly and associated with \( \beta_2 \), once as a quadratic term and associated with \( \beta_4 \) and once as an interaction term with \( x_3 \) and associated with \( \beta_5 \). That is, the \( x \)s 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, \( x_1 \), as a linear function of the parameters and assumes that noise enters the resulting expression additively. This two qualifications do not require the relationship between \( y \) and the \( x \)s to be expected: if \( x_2 \) affects \( y \) in a non-linear fashion, then its effect on \( y \) should vary by the value of \( x_2 \).

Due to the non-linear fashion in which the \( x \)s 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:

\[
\frac{\partial E(y_i|x_i)}{\partial x_{i2}} = \beta_2 + 2\beta_4 x_{i2} + \beta_5 x_{i3} \quad \text{and} \quad \frac{\partial E(y_i|x_i)}{\partial x_{i3}} = \beta_3 + \beta_5 x_{i2}
\]  
\[(2.19)\]

respectively. The marginal effects now depend on the values of the \( x \)s 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 \( x \)s, 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 \( x \)s 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 \( \beta_3 \), \( \beta_5 \) and the sample mean of \( x_{i2}, \bar{x}_2 \). Of course, the uncertainty around the values of \( \beta_3 \) and \( \beta_5 \) is transmitted to the marginal effect and one could get the posterior moments of this marginal effect by evaluating it at all draws of \( \beta_3 \) and \( \beta_5 \) from the posterior and summarizing.

Let’s return to the general problem of representing the possibly non-linear relationship among \( K \) \( x \)s 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:

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

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 \( x \)s and their interactions, respectively. Notice that from Young’s theorem we get:

\[
\frac{\partial^2 f}{\partial x_k \partial x_\ell} = \frac{\partial^2 f}{\partial x_\ell \partial x_k}
\]  
\[(2.21)\]
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:

\[
y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i2}^2 + \beta_5 x_{i2} x_{i3} + \beta_6 x_{i3}^2 + \epsilon_i
\]  

(2.22)

where \(\beta_2\) and \(\beta_3\) are the partial derivatives of \(f(\cdot)\) with respect to \(x_2\) and \(x_3\), \(\beta_5\) is the second cross-derivative and \(\beta_4\) and \(\beta_6\) 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: \(Y = \alpha K^\beta L^{1-\beta}\), where \(\alpha\) is the amount of produced output and \(K\) and \(L\) are the amounts of capital and labor input employed during production. \(\alpha\) is a productivity index and \(\alpha\) and \(\beta\) 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 \(\log Y = \alpha \log K + \beta \log L\). After appending an error term and assuming that \(\log A\) 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:

\[
\log Y_i = \beta_1 + \beta_2 \log K_i + \beta_3 \log L_i + \beta_4 t + \epsilon_i
\]  

(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 \(E(\log Y_i | K_i, L_i, t) = \beta_1 + \beta_2 \log K_i + \beta_3 \log L_i + \beta_4 t\) and the marginal effect of \(\log K_i\) on \(\log Y_i\) is:

\[
\beta_2 = \frac{\partial E(\log Y_i | K_i, L_i, t)}{\partial \log K_i} = \frac{\partial \log E(Y_i | K_i, L_i, t)}{\partial \log K_i} = \frac{\partial \log E(Y_i | K_i, L_i, t)}{\partial K_i} \cdot \frac{K_i}{E(Y_i | K_i, L_i, t)}
\]  

(2.24)

which has the form of an elasticity. Therefore, \(\beta_2\) 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 \(\beta_3\) as an elasticity and extends to all models in double-log form, \(\log y_i = (\log x_i) / \beta + \epsilon_i\), or in forms where only some of the independent variables are in logarithms, \(\log y_i = (\log x_i) / \beta + z_i' \gamma + \epsilon_i\).

It is stressed that, in general:

\[
\frac{\partial \log E(\log y_i | x_i)}{\partial \log x_{ik}} = \frac{\partial \log y_i | x_i)}{\partial \log x_{ik}}
\]  

(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 \(x_i\); an assumption that we are maintaining throughout this chapter. In particular, if \(\epsilon_i | x_i \sim N(0, \frac{1}{\tau})\) in the double-log model \(\log y_i = (\log x_i) / \beta + \epsilon_i\), then \(\log y_i | x_i\) follows a Normal distribution as well. By definition, \(y_i | x_i\) follows a log-Normal distribution and from the properties of this distribution we get:

\[
\log E(y_i | x_i) = (\log x_i) / \beta + \frac{1}{2\tau}
\]  

(2.26)

From this expression we can see that \(\frac{\partial \log E(y_i | x_i)}{\partial \log x_{ik}} = \beta_k = \frac{\partial E(\log y_i | x_i)}{\partial \log x_{ik}}\).
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 \( x_i \). 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 \( x \)s is modeled in logarithms, there may still be scope for using as independent variables squared and interactions terms of the logarithms of the \( x \)s. The translog is another flexible functional form, obtained by taking a second-order Taylor-series expansion of \( \log f(\log x_1, \log x_2, \ldots, \log x_K) \) around a vector of ones (for the \( x \)s):

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

(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
- \( \text{trend} \): 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:

\[
\log Y_i = \beta_1 + \beta_2 \log K_i + \beta_3 \log L_i + \beta_4 \text{trend}_i + \varepsilon_i
\]

The results obtained using BayES’ \texttt{lm()} function are presented in the following table. For this model we use the default priors for the \( \beta \)s and the precision parameters of the error term: the \( 4 \times 1 \) vector \( \beta \) 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 \( \tau \) 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.

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Median</th>
<th>Sd.dev.</th>
<th>5%</th>
<th>95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>constant</td>
<td>3.24514</td>
<td>3.24896</td>
<td>0.279077</td>
<td>2.784</td>
<td>3.69835</td>
</tr>
<tr>
<td>logK</td>
<td>0.583076</td>
<td>0.582767</td>
<td>0.023898</td>
<td>0.544447</td>
<td>0.622634</td>
</tr>
<tr>
<td>logL</td>
<td>0.441425</td>
<td>0.441675</td>
<td>0.022575</td>
<td>0.403951</td>
<td>0.477781</td>
</tr>
<tr>
<td>trend</td>
<td>0.00119668</td>
<td>0.00120461</td>
<td>0.00050872</td>
<td>0.00035365</td>
<td>0.00020245</td>
</tr>
<tr>
<td>tau</td>
<td>72.1455</td>
<td>72.0801</td>
<td>3.9129</td>
<td>65.778</td>
<td>78.7437</td>
</tr>
<tr>
<td>sigma(\varepsilon)</td>
<td>0.117862</td>
<td>0.117787</td>
<td>0.003204</td>
<td>0.112692</td>
<td>0.123302</td>
</tr>
</tbody>
</table>

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:

\[
\log Y_i = \beta_1 + \beta_2 \log K_i + \beta_3 \log L_i + \beta_4 \text{trend}_i + \beta_5 \log K_i \log K_i + \beta_6 \log L_i \log L_i + \beta_7 \log L_i \log K_i + \beta_8 \text{trend}_i \log K_i + \beta_9 \text{trend}_i \log L_i + \epsilon_i
\]

If we proceed to create the interaction terms and estimate the model, we will get parameter estimates for the \(\beta\)s, 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:

\[
\text{log} K_i = \log K_i - \overline{\log K} \quad \text{and} \quad \text{log} L_i = \log L_i - \overline{\log L}
\]

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:

\[
\frac{\partial E (\log Y_i | \log K_i, \log L_i)}{\partial \text{log} K_i} = \beta_2 + 2\beta_5 \overline{\log K} + \beta_6 \overline{\log L} + \beta_8 \text{trend}_i
\]

and if we evaluate this expression at the arithmetic means of the right-hand side variables we will get \(\beta_2\), 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’ \texttt{lm()} function, we obtain the results presented in the following table. Again, we use BayES’ default priors for the \texttt{lm()} function. The results for the parameters associated with the first-order terms (\(\beta_2, \beta_3\) and \(\beta_4\)) 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 \(\text{tlogK}\) is negative with probability greater than 95% and the parameter associated with variable \(\text{tlogL}\) 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.

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Median</th>
<th>Sd.dev.</th>
<th>5%</th>
<th>95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>constant</td>
<td>12.7573</td>
<td>12.7573</td>
<td>0.0064693</td>
<td>12.7467</td>
<td>12.7678</td>
</tr>
<tr>
<td>logK</td>
<td>0.570639</td>
<td>0.570463</td>
<td>0.0260929</td>
<td>0.527734</td>
<td>0.613743</td>
</tr>
<tr>
<td>logL</td>
<td>0.44325</td>
<td>0.443427</td>
<td>0.0247698</td>
<td>0.402362</td>
<td>0.484039</td>
</tr>
<tr>
<td>trend</td>
<td>0.0014026</td>
<td>0.0014039</td>
<td>0.00052461</td>
<td>0.000547946</td>
<td>0.00226372</td>
</tr>
<tr>
<td>logKlogK</td>
<td>0.196592</td>
<td>0.196859</td>
<td>0.10778</td>
<td>0.0194922</td>
<td>0.374709</td>
</tr>
<tr>
<td>logKlogL</td>
<td>-0.29448</td>
<td>-0.294986</td>
<td>0.210733</td>
<td>-0.642433</td>
<td>0.0504442</td>
</tr>
<tr>
<td>logLlogL</td>
<td>0.0963105</td>
<td>0.0962573</td>
<td>0.102469</td>
<td>-0.0717326</td>
<td>0.265435</td>
</tr>
<tr>
<td>tlogK</td>
<td>-0.00700459</td>
<td>-0.00700388</td>
<td>0.00300928</td>
<td>-0.0119623</td>
<td>-0.00205764</td>
</tr>
<tr>
<td>tlogL</td>
<td>0.00491162</td>
<td>0.00492376</td>
<td>0.00288508</td>
<td>0.000153932</td>
<td>0.00963616</td>
</tr>
<tr>
<td>tau</td>
<td>81.4347</td>
<td>81.3652</td>
<td>4.47955</td>
<td>74.1769</td>
<td>88.9722</td>
</tr>
<tr>
<td>sigma_e</td>
<td>0.11094</td>
<td>0.110862</td>
<td>0.00305806</td>
<td>0.106017</td>
<td>0.116109</td>
</tr>
</tbody>
</table>

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

```r
// 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);`
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:

\[ \log Y_i = \beta_1 + \beta_2 \log K_i + \beta_3 \log L_i + \beta_4 t + \epsilon_i \]  

(2.28)

In this model \( \beta_4 \) 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. 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 \( \beta_4 \) 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 \( \beta_4 \) 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 \( \beta_2 \) and \( \beta_3 \) is equal to one, decreasing if the sum is smaller than one and increasing otherwise. The probability of the production function exhibiting, for example,

\footnote{Recall that \( \beta_4 \) 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 $\beta_2$ and $\beta_3$ are draws from the joint distribution of these two parameters, conditional on the data and marginally with respect 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:

$$y_i = x_i'\beta + \varepsilon_i$$  \hspace{1cm} (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 $g(\beta) > r$, where $g$ is a vector-valued function and $r$ 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 $\beta$ is a continuous random variable and the probability of a function of it being exactly equal to a given vector, $r$, is zero. An obvious workaround is to calculate the probability of $g(\beta)$ being within a certain small interval around $r$, 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:

$$R\beta = r$$  \hspace{1cm} (2.30)

where $R$ is a $q \times K$ matrix and $R\beta$ assumes the role of the possibly non-linear function $g(\beta)$, defined above. If this system of equations has multiple solutions, then the dimensionality of $\beta$ can be reduced by expressing one or more of the $\beta$s as functions of the remaining and of the coefficients in $R$ and $r$. 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 $\beta$. If the uncertainty about the value of $\beta$ is reflected in this prior density, then $\beta$ is a random vector and, therefore, $r$ should also be treated as random. However, economic theory provides values for $r$ and to impose the restrictions through the prior we should get the prior density of $\beta$ given $r$. Thus, apart from accommodating the restrictions, $p(\beta|r)$ also expresses how forcefully we want to impose them or, in other words how far away are we willing to allow $\beta$ 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 $\beta$ that does not impose the restrictions or, to put it differently, a prior that disregards economic theory:

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

\(^4\)In the previous example $g(\beta)$ and $r$ are:

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

respectively.
The second step imposes a distribution on $r\beta$:

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

This density suggests, first of all, that the value of $r$ 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 $r$ is $R\beta$) and the variability of $r$ around its mean can be controlled by the value of the precision matrix, $\Xi$. For example, expressing strong prior beliefs that $R\beta$ should be close to $r$ can be achieved by setting $\Xi$ equal to a diagonal matrix with large values on the diagonal. Alternatively, a diagonal $\Xi$ with small values on the diagonal would allow $r$ to be far from what economic theory prescribes. Thus, $\Xi$ 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:

$$p(\beta|r) = \frac{p(r|\beta) \times p(\beta)}{p(r)} = \frac{|\hat{P}|^{1/2}}{(2\pi)^{K/2}} \exp \left\{-\frac{1}{2} (\beta - \hat{m})' \hat{P} (\beta - \hat{m})\right\} \quad (2.33)$$

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

$$\log Y_i = \beta_1 + \beta_2 \log K_i + \beta_3 \log L_i + \beta_4 \text{trend}_i + \epsilon_i$$

We estimated this model in the previous part of this example. Using the results from this model we approximate the probability of $\beta_4$ being positive (technological progress) as:

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

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:

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

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 $\text{Prob}(\beta_4 > 0, \beta_2 + \beta_3 > 1|y)$. 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.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Cases</th>
<th>Successes</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_4 &gt; 0$</td>
<td>20000</td>
<td>19816</td>
<td>0.9908</td>
</tr>
<tr>
<td>$\beta_2 + \beta_3 &gt; 1$</td>
<td>20000</td>
<td>20000</td>
<td>1</td>
</tr>
<tr>
<td>Overall</td>
<td>20000</td>
<td>19816</td>
<td>0.9908</td>
</tr>
</tbody>
</table>

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’ $\text{im}()$ function while imposing this restriction with great conviction: we set the $\Xi$ matrix that defines the precision of the right-hand side of the constraint $\beta_1 + \beta_2 = r$ equal to $10^6$. 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.
### 2.6. POST-ESTIMATION INFERENCE

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Median</th>
<th>Sd.dev.</th>
<th>5%</th>
<th>95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>constant</td>
<td>4.08074</td>
<td>4.08425</td>
<td>0.256233</td>
<td>3.65764</td>
<td>4.49684</td>
</tr>
<tr>
<td>logK</td>
<td>0.516158</td>
<td>0.515854</td>
<td>0.0222997</td>
<td>0.480052</td>
<td>0.553002</td>
</tr>
<tr>
<td>logL</td>
<td>0.483844</td>
<td>0.484147</td>
<td>0.0222999</td>
<td>0.447011</td>
<td>0.519957</td>
</tr>
<tr>
<td>trend</td>
<td>0.00251447</td>
<td>0.00251857</td>
<td>0.000482089</td>
<td>0.00171773</td>
<td>0.0032989</td>
</tr>
</tbody>
</table>

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

```r
// 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:

\[
\begin{align*}
\text{Model 0: } & y_i = x_i' \beta_0 + \varepsilon_i, & \varepsilon_i & \sim N\left(0, \frac{1}{\tau_0}\right), \\
\text{Model 1: } & y_i = z_i' \beta_1 + \xi_i, & \xi_i & \sim N\left(0, \frac{1}{\tau_1}\right).
\end{align*}
\]

along with their priors, \( p_0(\beta_0, \tau_0) \) and \( p_1(\beta_1, \tau_1) \). We will keep using Normal and Gamma priors for the two \( \beta \)s and two \( \tau \)s, 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: \( x \) could contain a subset of the variables in \( z \) or the other way around, the sets of independent variables in the two vectors could overlap only partially, with some of the variables in \( x \) not appearing in \( z \) and some of the variables in \( z \) not appearing in \( x \), 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 \( \beta \)s, 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 \( \text{Prob}(M=0) = \)
we will now compare the following two models 

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:

\[
m(y|M=j) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p_j(y|\beta_j, \tau_j, \bullet) \cdot p_j(\beta_j, \tau_j) \, d\beta_j \, d\tau_j
\]

where \( p_j(y|\beta_j, \tau_j, \bullet) \) and \( p_j(\beta_j, \tau_j) \) 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 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 \( \beta \) and \( \tau \) 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 \( \beta \) and \( \tau \) 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:

\[
\log Y_t = \beta_1 + \beta_2 \log K_t + \beta_3 \log L_t + \beta_4 \text{trend}_t + \varepsilon_t
\]

and:

\[
\log Y_t = \beta_1 + \beta_2 \log K_t + \beta_3 \log L_t + \beta_4 \text{trend}_t
\]

\[+ \beta_5 \log K_t, \log K_t + \beta_6 \log K_t, \log L_t + \beta_7 \log L_t, \log L_t \]

\[+ \beta_8 \text{trend}_t, \log K_t + \beta_9 \text{trend}_t, \log L_t + \varepsilon_t\]

The Cobb-Douglas model can be obtained from the translog model by restricting \( \beta_5 \) to \( \beta_0 \) 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 within 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.
### 2.6. POST-ESTIMATION INFERENCE

<table>
<thead>
<tr>
<th>Model</th>
<th>Log-Marginal Likelihood</th>
<th>Type of log-ML Approximation</th>
<th>Prior Model Probability</th>
<th>Posterior Model Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>CobbDouglas</td>
<td>442.663</td>
<td>Lewis &amp; Raftery</td>
<td>0.5</td>
<td>0.834778</td>
</tr>
<tr>
<td>Translog</td>
<td>441.043</td>
<td>Lewis &amp; Raftery</td>
<td>0.5</td>
<td>0.165222</td>
</tr>
<tr>
<td>CobbDouglas</td>
<td>442.671</td>
<td>Chib &amp; Jeliazkov</td>
<td>0.5</td>
<td>0.833955</td>
</tr>
<tr>
<td>Translog</td>
<td>441.057</td>
<td>Chib &amp; Jeliazkov</td>
<td>0.5</td>
<td>0.166045</td>
</tr>
</tbody>
</table>

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.

```r
// 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 $y_i = x_i'\beta + \varepsilon_i$, along with the distributional assumption $\varepsilon_i \sim N(0, \frac{1}{\tau})$, 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 $y_*|\beta, \tau \sim N(\mathbf{x}'_\beta, \tau)$ 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 $\beta$ and $\tau$ 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 $\beta$ and $\tau$ from the joint density of $y_*$ and the parameters, conditional on the observed data:

$$p(y_*|\mathbf{x}, y) = \int_0^\infty \int_0^\infty p(y_*|\beta, \tau, \mathbf{x}, y) \pi(\beta, \tau) \, d\beta d\tau$$

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

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

Notice that, once we condition on $\beta$ and $\tau$, the density of $y_*$ does not depend on the observed data, $y$, 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 $\beta$ and $\tau$, the observed data have nothing more 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 $\beta$ and $\tau$, 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:

$$E(y_*|\mathbf{x}, y) = \int_{-\infty}^\infty y_* \cdot p(y_*|\mathbf{x}, y) \, dy_*$$

Because $y_*|\beta, \tau, \mathbf{x} \sim N(\mathbf{x}'\beta, \tau)$, samples from $p(y_*|\mathbf{x}, y)$ can be obtained by using the $G$ retained draws from the posterior for $\beta$ and $\tau$ 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:

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

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:

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

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

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

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, \( \tau \), 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 \( y = f(x) \), where \( y \) is a vector of dependent variables, \( x \) is a vector of independent variables and \( f(\cdot) \) 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:

\[
\begin{align*}
  y_{1i} &= x_{1i}' \beta_1 + \varepsilon_{1i} \\
  y_{2i} &= x_{2i}' \beta_2 + \varepsilon_{2i} \\
  &\vdots \\
  y_{Mi} &= x_{Mi}' \beta_M + \varepsilon_{Mi}
\end{align*}
\]

(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, \( x_{mi} \), and a \( K_m \times 1 \) vector of parameters, \( \beta_m \), plus an error term, \( \varepsilon_{mi} \), 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 \( x \)s.

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 \( \varepsilon \)s 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 \( \varepsilon \)s follows a multivariate Normal distribution, with expected value equal to a vector of zeros and precision matrix (inverse covariance matrix) \( \Omega \). 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 \( \Omega \) is not diagonal, the marginal distribution of each \( \varepsilon_{mi} \) 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 \( \beta \)s.

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 \( \Omega \) 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 $\beta$s than allowing $\Omega$ 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 con-
ditional expectation interpretation: $E(y_m|x_m) = x_m \beta_m$, for $m = 1, 2, \ldots, M$. Similarly,
nothing changes with respect to meaning of the $\beta$s in the SUR model: if $\beta_{mk}$ 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 $\Omega$, on the other hand, requires some attention. It may be more
natural to start by interpreting the variance matrix of the error terms, $\Omega^{-1}$, rather than $\Omega$
itsel.f. 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 $\Omega^{-1}$ are
the pairwise covariances of the error terms, again marginally with respect to the remaining $\varepsilon$s.
Mathematically:

$$
\Omega^{-1} = 
\begin{bmatrix}
V(\varepsilon_{11}) & \text{Cov}(\varepsilon_{11}, \varepsilon_{21}) & \cdots & \text{Cov}(\varepsilon_{11}, \varepsilon_{Mi}) \\
\text{Cov}(\varepsilon_{21}, \varepsilon_{11}) & V(\varepsilon_{21}) & \cdots & \text{Cov}(\varepsilon_{21}, \varepsilon_{Mi}) \\
\vdots & \vdots & \ddots & \vdots \\
\text{Cov}(\varepsilon_{Mi}, \varepsilon_{11}) & \text{Cov}(\varepsilon_{Mi}, \varepsilon_{21}) & \cdots & V(\varepsilon_{Mi})
\end{bmatrix}
$$

(3.2)

where all variances and covariances are taken marginally with respect to the error terms that
do not appear as their arguments. If $\Omega^{-1}$ 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 $\Omega$, not its inverse.
It turns out that the precision matrix also has an intuitive interpretation, albeit not as direct.
To start with, $\Omega^{-1}$ will be diagonal if and only if $\Omega$ is diagonal. Therefore, and given that the
$\varepsilon$s jointly follow a multivariate Normal distribution, whether the error terms across equations
are independent or not can be inferred directly from $\Omega$. The off-diagonal elements of the
precision matrix can be used to obtain the partial correlations of the error terms, conditional
on the remaining $\varepsilon$s. In particular, the partial correlation coefficient of $\varepsilon_{mi}$ and $\varepsilon_{\ell i}$ can be
obtained as $-\frac{\text{Cov}(\varepsilon_{mi}, \varepsilon_{\ell i})}{\sqrt{\text{Var}(\varepsilon_{mi}) \text{Var}(\varepsilon_{\ell i})}}$, where $\omega_{mi}$ is the element of $\Omega$ in row $m$, column $\ell$.

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:

$$
y_i = X_i \beta + \varepsilon_i, \quad \varepsilon_i \sim N(0, \Omega^{-1})
$$

(3.3)

where:

$$
y_i = \begin{bmatrix} y_{1i} \\ y_{2i} \\ \vdots \\ y_{Mi} \end{bmatrix}, \quad X_i = \begin{bmatrix} x_{i1} & 0 & \cdots & 0 \\ 0 & x_{i2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & x_{Mi} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_M \end{bmatrix} \quad \text{and} \quad \varepsilon_i = \begin{bmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \\ \vdots \\ \varepsilon_{Mi} \end{bmatrix}
$$

and $K$ is the total number of $\beta$ 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 $y = f(x)$ 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 $y_i$ is also multivariate Normal, with mean $X_i\beta$ and precision matrix $\Omega$. Adding to this an independence assumption on the error terms across observations, the likelihood of the SUR model with $N$ observations is:

$$p(\{y_i\} | \beta, \Omega) = \prod_{i=1}^{N} \frac{|\Omega|^{1/2}}{(2\pi)^{M/2}} \exp \left\{ \frac{1}{2} (y_i - X_i\beta)' \Omega (y_i - X_i\beta) \right\}$$

$$= \frac{|\Omega|^{N/2}}{(2\pi)^{MN/2}} \exp \left\{ -\frac{1}{2} \sum_{i=1}^{N} (y_i - X_i\beta)' \Omega (y_i - X_i\beta) \right\} \quad \text{(3.4)}$$

The parameters to be estimated in the model are the $K \times 1$ vector $\beta$ and the $M \times M$ matrix $\Omega$. A multivariate Normal prior for $\beta$ appears as the natural choice and it turns out to be conjugate. $\Omega$ 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:

$$p(\Omega) = \frac{\Gamma_M(\frac{n-M-1}{2})|\Omega|^{n-M-1}}{\Gamma_M(\frac{n}{2})} \frac{1}{2^n|\Omega|^\frac{M}{2}} \exp \left\{ -\frac{1}{2} \text{tr}(\Omega^{-1}) \right\} \quad \text{(3.5)}$$

where $n$ is the degrees-of-freedom parameter, $V$ is the scale matrix and $\Gamma_M(\cdot)$ is the $M$-dimensional Gamma function. In the context of Bayesian inference, $n$ and $V$ 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, $\Omega$, is $nV$ and the variance of the element in row $m$, column $\ell$ of $\Omega$, marginally with respect to the remaining elements, is $n \left( v_m^2 + v_m v_{m\ell} + v_{\ell\ell} \right)$, where $v_m$ is the element in row $m$, column $\ell$ of $V$. For a given value of $n$ and using the expected-value formula only, a reasonable choice for $V$ would be $\frac{1}{n}Q$, where $Q$ is a prior guess on the value of $\Omega$. The choice of $n$ and $V$, however, also affect the variance of $\Omega$. If we restrict attention to values of the hyperparameters that lead to proper priors and given that $V$ is set to $\frac{1}{n}Q$, the least informative prior for $\Omega$ is obtained by setting $n$ equal to the dimension of the problem. This is because, if $V = \frac{1}{n}Q$, the marginal variances are maximized for large values of $V$ (therefore, small values of $n$), since the values of $V$ 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 $\beta$ and $\Omega$ 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 $\beta$ is similar to what was done for the linear regression model. Derivation of the full conditional of $\Omega$ 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:

$$y_i = X_i \beta + \varepsilon_i, \quad \varepsilon_i \sim N(0, \Omega^{-1})$$

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

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

and:

$$p(\Omega) = |\Omega|^{-\frac{a-M-1}{2}} \left| V^{-1} \right|^{-\frac{n}{2}} \exp \left\{ -\frac{1}{2} \text{tr} \left( V^{-1} \Omega \right) \right\}$$

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

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

and:

$$\pi(\Omega|m) = |\Omega|^{-\frac{a-M-1}{2}} |V|^{-\frac{n}{2}} \exp \left\{ -\frac{1}{2} \text{tr} \left( V^{-1} \Omega \right) \right\}$$

where:

- $\tilde{P} = \sum_{i=1}^{N} X_i' \Omega X_i + P$
- $\tilde{m} = \left( \sum_{i=1}^{N} X_i' \Omega X_i + P \right)^{-1} \left( \sum_{i=1}^{N} X_i' \Omega y_i + Pm \right)$
- $\tilde{n} = N + n$
- $\tilde{V}^{-1} = \sum_{i=1}^{N} (y_i - X_i \beta) (y_i - X_i \beta)' + V^{-1}$

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 $\beta$ 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 $\beta$ from the SUR model converges to the posterior mean obtained by the separate regressions. To see how this works, it is convenient to set $P$ exactly equal to a matrix of zeros, before considering what happens when the prior for $\beta$ is proper.

In the case where all equations contain the same independent variables in a $k \times 1$ vector $x_i$, $X_i$ can be written as $I_M \otimes x_i'$, where $I_M$ 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, $X'_i \Omega X_i$ becomes $\Omega \otimes x_i x_i'$ and $X'_i \Omega y_i$ becomes $(\Omega \otimes I_k) (I_M \otimes x_i) y_i$. With $P = 0$, the expression for $\tilde{m}$ can be written as:

$$\tilde{m} = \left( \Omega \otimes \sum_{i=1}^{N} x_i x_i' \right)^{-1} \left( \Omega \otimes I_k \sum_{i=1}^{N} (I_M \otimes x_i) y_i \right)$$

$$= \left( I_M \otimes \sum_{i=1}^{N} x_i x_i' \right)^{-1} \sum_{i=1}^{N} (I_M \otimes x_i) y_i$$

(3.6)
Carrying-out the multiplications in this expression leads to:

\[ \hat{\mathbf{m}} = \begin{bmatrix} 
\left( \sum_{i=1}^{N} x_i x_i' \right)^{-1} \sum_{i=1}^{N} x_i y_{1i} \\
\left( \sum_{i=1}^{N} x_i x_i' \right)^{-1} \sum_{i=1}^{N} x_i y_{2i} \\
\vdots \\
\left( \sum_{i=1}^{N} x_i x_i' \right)^{-1} \sum_{i=1}^{N} x_i y_{Mi} 
\end{bmatrix} \]

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

\[ (3.7) \]

---

**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:

- \( m_{\text{Grade}}, p_{\text{Grade}} \): final grade obtained by the student in mathematics and Portuguese, respectively, on a 0-20 scale
- \( \text{age} \): age of the student in years
- \( \text{female} \): dummy variable, 1 if female
- \( m_{\text{Study}}, p_{\text{Study}} \): amount of time spent studying for mathematics and Portuguese, respectively, coded such that higher values correspond to greater effort
- \( m_{\text{Fails}}, p_{\text{Fails}} \): number of past class failures for mathematics and Portuguese, respectively
- \( m_{\text{Paid}}, p_{\text{Paid}} \): dummy variable, 1 if the student took extra paid classes in mathematics and Portuguese, respectively
- \( m_{\text{Absent}}, p_{\text{Absent}} \): 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:

\[
\begin{align*}
m_{\text{Grade}}_i &= \beta_{11} + \beta_{12} \text{age}_i + \beta_{13} \text{female}_i \\
&\quad + \beta_{14} m_{\text{Study}}_i + \beta_{15} m_{\text{Fails}}_i + \beta_{16} m_{\text{Paid}}_i + \beta_{17} m_{\text{Absent}}_i + \varepsilon_{1i} \\
p_{\text{Grade}}_i &= \beta_{21} + \beta_{22} \text{age}_i + \beta_{23} \text{female}_i \\
&\quad + \beta_{24} p_{\text{Study}}_i + \beta_{25} p_{\text{Fails}}_i + \beta_{26} p_{\text{Paid}}_i + \beta_{27} p_{\text{Absent}}_i + \varepsilon_{2i}
\end{align*}
\]

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.

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Median</th>
<th>Sd.dev.</th>
<th>5%</th>
<th>95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>constant</td>
<td>20.0977</td>
<td>20.1022</td>
<td>3.14519</td>
<td>14.917</td>
<td>25.242</td>
</tr>
<tr>
<td>age</td>
<td>-0.567237</td>
<td>-0.566791</td>
<td>0.188453</td>
<td>-0.878842</td>
<td>-0.255136</td>
</tr>
<tr>
<td>female</td>
<td>-1.56698</td>
<td>-1.56346</td>
<td>0.460623</td>
<td>-2.32726</td>
<td>-0.808936</td>
</tr>
<tr>
<td>mStudy</td>
<td>0.381196</td>
<td>0.382687</td>
<td>0.278552</td>
<td>-0.0795295</td>
<td>0.841523</td>
</tr>
<tr>
<td>mFails</td>
<td>-2.23463</td>
<td>-2.23236</td>
<td>0.310932</td>
<td>-2.75392</td>
<td>-1.71992</td>
</tr>
<tr>
<td>mPaid</td>
<td>0.346688</td>
<td>0.346526</td>
<td>0.449231</td>
<td>-0.395886</td>
<td>1.09017</td>
</tr>
<tr>
<td>mAbsent</td>
<td>0.0417686</td>
<td>0.0415376</td>
<td>0.0292148</td>
<td>-0.0057057</td>
<td>0.0903097</td>
</tr>
<tr>
<td>tau</td>
<td>0.0555064</td>
<td>0.0554304</td>
<td>0.004053</td>
<td>0.0489308</td>
<td>0.0623213</td>
</tr>
<tr>
<td>sigma_e</td>
<td>4.25305</td>
<td>4.24743</td>
<td>0.156127</td>
<td>4.00583</td>
<td>4.5208</td>
</tr>
</tbody>
</table>

We will first run linear regression models for each equation separately before we compare the results to a SUR model.
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 the 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.

BayES does not present summaries of the draws from the posterior distribution of \(\Omega\), 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 \(\Omega\) is:

\[
E(\Omega|\cdot) = \begin{bmatrix}
0.07169 & -0.05440 \\
-0.05440 & 0.18254
\end{bmatrix}
\]

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

\[
E(\rho_{e_1,e_2}|\cdot) \approx \frac{1}{G} \sum_{g=1}^{G} - \frac{\omega_{12}^{(g)}}{\sqrt{\omega_{11}^{(g)} \omega_{22}^{(g)}}} = 0.47498
\]

Creating a credible interval for \(\rho_{e_1,e_2}\) or calculating the probability of it being above or below a certain threshold is straightforward. For example, \(\text{Prob}(\rho_{e_1,e_2}|\cdot) > 0.5\) 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.

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

// run the two regressions separately
matM = lm(mGrade ~ constant age female mStudy mFails mPaid mAbsent);
portM = lm(pGrade ~ constant age female pStudy pFails pPaid pAbsent);
```
CHAPTER 3. SEEMINGLY UNRELATED REGRESSIONS

// run the two regressions in a SUR system
jointM = surf {
    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:

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

(3.8)

where $u$ is the utility level and $p_1, p_2, \ldots, p_M$ are the prices of the $M$ goods. The $\alpha$s and $\beta$s 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 $\ell$.

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:

\[ h_m(u, p_1, p_2, \ldots, p_M) = \frac{\partial e(u, p_1, p_2, \ldots, p_M)}{\partial p_m}, \quad m = 1, 2, \ldots, M \]  

(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:

\[ s_m = \alpha_m + \sum_{\ell=1}^{M} \alpha_{m\ell} \log p_{\ell} + \beta_m \log \frac{E}{P}, \quad m = 1, 2, \ldots, M \]

(3.10)

where \( E \) is total expenditure in all \( M \) goods and \( P \) is a price index constructed using the original prices and the \( \alpha \) parameters. Due to this index, the \( \alpha \)s 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:

\[ s_{1i} = \alpha_1 + \sum_{\ell=1}^{M} \alpha_{1\ell} \log p_{\ell i} + \beta_1 \log \frac{E_i}{P_i} + \varepsilon_{1i} \]

\[ s_{2i} = \alpha_2 + \sum_{\ell=1}^{M} \alpha_{2\ell} \log p_{\ell i} + \beta_2 \log \frac{E_i}{P_i} + \varepsilon_{2i} \]

\[ \vdots \]

\[ s_{Mi} = \alpha_M + \sum_{\ell=1}^{M} \alpha_{M\ell} \log p_{\ell i} + \beta_M \log \frac{E_i}{P_i} + \varepsilon_{Mi} \]

(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, renders the covariance matrix of \( \varepsilon \) 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, \( \alpha_0 \) and \( \beta_0 \) 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:

\[
\log C_i = \beta_0 + \sum_{m=1}^{M} \beta_m \log w_{mi} + \frac{1}{2} \sum_{m=1}^{M} \sum_{\ell=1}^{M} \beta_{m\ell} \log w_{mi} \log w_{\ell i} + \delta_1 \log y_i + \frac{1}{2} \delta_2 \log^2 y_i + \sum_{m=1}^{M} \gamma_m \log y_i \log w_{mi} + \varepsilon_i
\]

(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 \( w_s \) 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 $\ell$.

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:

$$s_{1i} = \beta_1 + \sum_{\ell=1}^{M} \beta_{1\ell} \log w_{\ell i} + \gamma_1 \log y_i + \varepsilon_{1i}$$
$$s_{2i} = \beta_2 + \sum_{\ell=1}^{M} \beta_{2\ell} \log w_{\ell i} + \gamma_2 \log y_i + \varepsilon_{2i}$$
$$\vdots$$
$$s_{Mi} = \beta_M + \sum_{\ell=1}^{M} \beta_{M\ell} \log w_{\ell i} + \gamma_M \log y_i + \varepsilon_{Mi}$$

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:

$$y_i = X_i \beta + \varepsilon_i, \quad \varepsilon_i \sim N(0, \Omega^{-1})$$

where $y_i$ and $\varepsilon_i$ are $M \times 1$ vectors, $\beta$ is $K \times 1$ and $X_i$ 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 $\beta$ can be incorporated in the prior. If these constraints are linear in the slope parameters, that is if they take the form:

$$R \beta = r$$

then the procedure discussed in Section 2.6 can be used for doing so. In the expression above $R$ is a $q \times K$ matrix and $r$ 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 $\beta$, with mean $\mathbf{m}$ and precision matrix $\mathbf{P}$, and which ignores the constraints altogether. Next, a multivariate Normal prior for
3.4. CROSS-EQUATION RESTRICTIONS AND THE SUR MODEL

\( r|\beta \), with mean \( R\beta \) and precision matrix \( \Xi \), is used to define how far away \( \beta \) is allowed to be from satisfying the restrictions. Finally, an application of Bayes’ theorem updates the prior for \( \beta \) 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:

\[
\log C = \beta_0 + \beta_K \log r + \beta_L \log w + \frac{1}{2} \beta_{KK} \log^2 r + \beta_{KL} \log r \log w + \frac{1}{2} \beta_{LL} \log^2 w + \delta_1 \log Y + \frac{1}{2} \delta_{YY} \log Y,
\]

\[
nK = \beta_K + \beta_{KK} \log r + \beta_{KL} \log w + \gamma_K \log Y + \epsilon_{1K},
\]

\[
nL = \beta_L + \beta_{KL} \log r + \beta_{LL} \log w + \gamma_L \log Y + \epsilon_{1L},
\]

where \( nK \) and \( nL \) are the shares of capital and labor inputs, respectively, in the cost of production. The restrictions that we need to impose on the \( \beta \)s and \( \gamma \)s for the cost function to be homogeneous of degree one in \( r \) and \( w \) are:

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

During estimation we also need to make sure that the values of the parameters that appear in more than 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’ \( \text{sur}() \) 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, \( \text{E}(\delta_1|\cdot) \approx 0.976 \) 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.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Mean</th>
<th>Median</th>
<th>S.d.dev.</th>
<th>5%</th>
<th>95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \log C )</td>
<td>12.7149</td>
<td>12.7148</td>
<td>0.00582343</td>
<td>12.7052</td>
<td>12.7244</td>
</tr>
<tr>
<td>( \log r )</td>
<td>0.397411</td>
<td>0.397414</td>
<td>0.00234058</td>
<td>0.393545</td>
<td>0.401282</td>
</tr>
<tr>
<td>( \log w )</td>
<td>0.602527</td>
<td>0.602525</td>
<td>0.00234095</td>
<td>0.598654</td>
<td>0.604602</td>
</tr>
<tr>
<td>( \log \log r )</td>
<td>0.093034</td>
<td>0.0930325</td>
<td>0.00352821</td>
<td>0.0872569</td>
<td>0.0988219</td>
</tr>
<tr>
<td>( \log \log w )</td>
<td>-0.18606</td>
<td>-0.18606</td>
<td>0.00705639</td>
<td>-0.197617</td>
<td>-0.174525</td>
</tr>
<tr>
<td>( \log Y )</td>
<td>0.97556</td>
<td>0.975545</td>
<td>0.00377648</td>
<td>0.969294</td>
<td>0.981731</td>
</tr>
<tr>
<td>( \log Y \log Y )</td>
<td>0.0127822</td>
<td>0.0127701</td>
<td>0.00169557</td>
<td>0.0099977</td>
<td>0.0155466</td>
</tr>
<tr>
<td>( \log Y \log r )</td>
<td>-0.00747926</td>
<td>-0.00751856</td>
<td>0.00176528</td>
<td>-0.0103636</td>
<td>-0.00453448</td>
</tr>
<tr>
<td>( \log Y \log w )</td>
<td>0.0074797</td>
<td>0.00751737</td>
<td>0.00176483</td>
<td>0.00453253</td>
<td>0.0103668</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Mean</th>
<th>Median</th>
<th>S.d.dev.</th>
<th>5%</th>
<th>95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>( sK )</td>
<td>0.397411</td>
<td>0.397414</td>
<td>0.00234071</td>
<td>0.393538</td>
<td>0.401282</td>
</tr>
<tr>
<td>( \log r )</td>
<td>0.186062</td>
<td>0.186061</td>
<td>0.00705629</td>
<td>0.17452</td>
<td>0.197649</td>
</tr>
<tr>
<td>( \log w )</td>
<td>-0.186059</td>
<td>-0.18606</td>
<td>0.00705653</td>
<td>-0.197616</td>
<td>-0.174531</td>
</tr>
<tr>
<td>( \log Y )</td>
<td>-0.00747984</td>
<td>-0.0075215</td>
<td>0.00176798</td>
<td>-0.0103651</td>
<td>-0.00453182</td>
</tr>
</tbody>
</table>
The results presented above can be obtained in BayES using the code in the following box.

```plaintext
// 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 constraints 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, $\theta$, and likelihood function $p(y|\theta)$, where $y$ is a vector that will store the observed data. Suppose also that $p(\theta)$ is the prior density of $\theta$. Bayesian inference in such a model would proceed by sampling from the posterior distribution of the parameters:

$$\pi(\theta|y) \propto p(y|\theta) \cdot p(\theta) \quad (4.1)$$

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 $z$ be a vector that would store the values of this random variable. Because $z$ is, by assumption, not observable it represents the latent data.
Since \( z \) is not observable, inferences about \( \theta \) cannot be made conditionally on the latent data. The obvious approach would then be to integrate the latent data from the joint density of \( \theta \) and \( z \), given the observed data:

\[
\pi(\theta|y) = \int_z \pi(\theta, z|y) \, dz = \int_z \pi(\theta|y, z) \, p(z|y) \, dz
\] (4.2)

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

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

\[
\pi(\theta|y, z) \propto p(y, z|\theta) \, p(\theta) = p(y|\theta, z) \, p(z|\theta) \, p(\theta)
\] (4.3)

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

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

\[
p(z|y) = \int_{\Theta} \pi(\theta, z|y) \, d\theta = \int_{\Theta} p(z|y, \theta) \, \pi(\theta|y) \, d\theta
\] (4.4)

Expressing the predictive density of the latent data as an integral that involves \( \pi(\theta|y) \) gives the impression that we are going around in circles: to get \( \pi(\theta|y) \) from (4.2) we need \( p(z|y) \) and to get \( p(z|y) \) from (4.4) we need \( \pi(\theta|y) \). 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 \( z^{(1)}, z^{(2)}, \ldots, z^{(Q)} \), \( Q \geq 1 \), from \( p(z|y) \), given the current approximation to \( \pi(\theta|y) \). This step is further broken into the following items:

(a1) sample \( \theta \) from the current approximation of \( \pi(\theta|y) \)

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

(b) update the current approximation to \( \pi(\theta|y) \) using \( \frac{1}{Q} \sum_{q=1}^{Q} \pi(\theta|y, z^{(q)}) \) 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 \( \theta \) and \( z \) is given by:

\[
\pi(\theta, z|y) = \frac{p(y, z|\theta) \, p(\theta)}{p(y)} \propto p(y|\theta, z) \, p(z|\theta) \, p(\theta)
\] (4.5)

\(^1\)One subtle point in the derivation is that we need to distinguish between \( \theta \) as an argument of \( \pi(\bullet|y) \) and \( \pi(\bullet|y, z) \) and the dummy of integration in (4.4). Tanner & Wong use \( \phi \) 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 $\theta$ and $z$, 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 $\theta$, 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 $z$ from $\pi(\theta, z|y)$, 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 $\theta$ corresponds to step (a1) in the Tanner & Wong formulation, but with $Q = 1$. In this case, $z$ is integrated-out from $\pi(\theta, z|y)$ using only a single draw and step (b) becomes degenerate. Sampling for $z$ 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 $\theta$ to a reasonable starting value
- set $z$ to a reasonable starting value

for $q = 1:(D+G)$ do

- draw $\theta$ from $\pi(\theta|y, z)$, either in one or multiple blocks
- draw $z$ from $p(z|y, \theta)$, either in one or multiple blocks

if $g > D$ then

- store the current value of $\theta$
- 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 $z$. 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 $\theta$ and $z$, the draws on $z$ alone are from the posterior density of the latent data, marginally with respect to $\theta$ and conditional only on the observed data. Therefore, these draws can be used to make probabilistic statements regarding the values of $z$.

### 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 $\tau$. In the linear regression model with *heteroskedastic error* we will relax the assumption that the $\varepsilon_i$s 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:

$$y_i = x_i^T \beta + \varepsilon_i, \quad \varepsilon_i \sim N\left(0, \frac{1}{\tau_i}\right)$$

(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 \( K \) \( \beta \)s, 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 \( \tau_i \)s are determined in the population. There are a few approaches for doing so. Koop (2003, pp.124-130) describes a procedure where each \( \tau_i \) 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 \( \tau_i \)s.

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

\[
\log \tau_i \sim N \left( w_i \delta, \frac{1}{\phi} \right) \tag{4.7}
\]

where \( w_i \) is an \( L \times 1 \) vector of observable variables which affect the precision of \( \varepsilon_i \), and \( \delta \) is an \( L \times 1 \) vector of parameters. \( \phi \) is another precision parameter to be estimated. The model allows for \( w_i \) 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, \( \delta \) and \( \phi \) are additional parameters to be estimated, while viewing the model from an incomplete-data perspective leads to a natural interpretation of the \( \tau_i \)s as latent data: if the \( \tau_i \)s 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 \( \tau_i \)s when drawing inferences on the model’s parameters: \( \beta \), \( \delta \) and \( \phi \).

Given that each \( \tau_i \) follows a log-Normal distribution, the complete-data likelihood for this model is:

\[
p(\{y\}, \{\tau\} | X, W, \beta, \delta, \phi) = \prod_{i=1}^{N} p(y_i | x_i, \beta, \tau_i) p(\tau_i | w_i, \delta, \phi)
\]

\[
= \prod_{i=1}^{N} \frac{\tau_i^{1/2}}{(2\pi)^{1/2}} \exp \left\{ \frac{-\tau_i (y_i - x_i \beta)^2}{2} \right\}
\]

\[
\times \prod_{i=1}^{N} \frac{\phi^{1/2}}{\tau_i (2\pi)^{1/2}} \exp \left\{ -\frac{\phi (\log \tau_i - w_i \delta)^2}{2} \right\} \tag{4.8}
\]

where \( W \) 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 \( \tau_i \)s, \( y \) no longer depends on \( W \), \( \delta \) and \( \phi \). Likewise, the \( \tau_i \)s depend on no other parameters or data, once we condition on \( W \), \( \delta \) and \( \phi \). The second equality results from the Normality of the error terms and log-Normality of the \( \tau_i \)s and the conditional independence of the \( y_i \)s, as well as of the \( \tau_i \)s.

As in the linear regression model with homoskedastic error, we will use a Normal prior for \( \beta \), with mean \( m_\beta \) and precision matrix \( P_\beta \). For \( \delta \) we will, again, use a Normal prior, with mean \( m_\delta \) and precision matrix \( P_\delta \), while for \( \phi \) 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:

\[
\pi (\beta, \delta, \phi, \{\tau\} | y, X, W) \propto p(\{\tau\} | X, W, \beta, \delta, \phi) p(\beta) p(\delta) p(\phi) \tag{4.9}
\]

The densities in the right-hand side of last expression are all known: the the first density is given in (4.8), \( p(\beta) \) and \( p(\delta) \) are multivariate-Normal densities and \( p(\phi) \) is a Gamma density.
Before we can implement a Gibbs sampler we need to derive the full conditionals of all unobserved quantities: $\beta$, $\delta$, $\phi$ and the $\tau_i$s (all $N$ of them). It is stressed that the $\tau_i$s 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 $\beta$ in the linear model with heteroskedastic error is very similar to the case of homoskedastic error
- deriving the full conditionals of $\delta$ and $\phi$ requires exactly the same transformations presented above Theorem 2.1, with $\log \tau_i$ assuming the role of $y_i$ and $w_i$ that of $x_i$
- deriving the full conditional of each $\tau_i$ 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 $\tau_i$s 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 $\tau_i$ 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:

$$y_i = x_i' \beta + \epsilon_i, \quad \epsilon_i \sim N \left( 0, \frac{1}{\tau_i} \right)$$

and $L$ independent variables in the precision equation:

$$\log \tau_i = w_i' \delta + v_i, \quad v_i \sim N \left( 0, \frac{1}{\phi} \right)$$

and with Normal priors for $\beta$ and $\delta$ and a Gamma prior for $\phi$:

$$p(\beta) = \frac{|P_{\beta}|^{1/2}}{(2\pi)^{K/2}} \exp \left\{ -\frac{1}{2} (\beta - m_\beta)' P_\beta (\beta - m_\beta) \right\},$$

$$p(\delta) = \frac{|P_\delta|^{1/2}}{(2\pi)^{L/2}} \exp \left\{ -\frac{1}{2} (\delta - m_\delta)' P_\delta (\delta - m_\delta) \right\} \quad \text{and} \quad p(\phi) = \frac{b^a}{\Gamma(a)} \phi^{a-1} e^{-\frac{b}{\phi}}$$

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

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

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

and the full conditional of $\phi$ is Gamma:

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

where:
CHAPTER 4. DATA AUGMENTATION

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:

\[ \log Y_i = \beta_1 + \beta_2 \log K_i + \beta_3 \log L_i + \beta_4 \text{trend}_i + \epsilon_i \]

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, \( \gamma \):

\[ \log \tau_i = \gamma + \nu_i, \quad \nu_i \sim N \left( 0, \frac{1}{\tau_i} \right) \]

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

\[ \log \tau_i = \gamma_1 + \gamma_2 \log K_i + \gamma_3 \log L_i + \gamma_4 \text{trend}_i + v_i, \quad v_i \sim N \left( 0, \frac{1}{\tau_i} \right) \]

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.

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Median</th>
<th>Sd.dev.</th>
<th>5%</th>
<th>95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>constant</td>
<td>3.24514</td>
<td>3.24896</td>
<td>0.27907</td>
<td>2.784</td>
<td>3.69835</td>
</tr>
<tr>
<td>logK</td>
<td>0.583076</td>
<td>0.582767</td>
<td>0.023892</td>
<td>0.544447</td>
<td>0.622634</td>
</tr>
<tr>
<td>logL</td>
<td>0.441425</td>
<td>0.441675</td>
<td>0.022575</td>
<td>0.403951</td>
<td>0.477781</td>
</tr>
<tr>
<td>trend</td>
<td>0.00119668</td>
<td>0.00120461</td>
<td>0.00050872</td>
<td>0.000353659</td>
<td>0.0020245</td>
</tr>
<tr>
<td>tau</td>
<td>72.1455</td>
<td>72.0801</td>
<td>3.9129</td>
<td>65.778</td>
<td>78.7437</td>
</tr>
<tr>
<td>sigma_\tau</td>
<td>0.117862</td>
<td>0.117787</td>
<td>0.00320426</td>
<td>0.112692</td>
<td>0.123302</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Median</th>
<th>Sd.dev.</th>
<th>5%</th>
<th>95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>logY</td>
<td>2.60779</td>
<td>2.60556</td>
<td>0.236055</td>
<td>2.22318</td>
<td>3.00122</td>
</tr>
<tr>
<td>logK</td>
<td>0.638856</td>
<td>0.639117</td>
<td>0.0203548</td>
<td>0.604912</td>
<td>0.672072</td>
</tr>
<tr>
<td>logL</td>
<td>0.390394</td>
<td>0.390249</td>
<td>0.0197409</td>
<td>0.358244</td>
<td>0.423338</td>
</tr>
<tr>
<td>trend</td>
<td>-0.000753112</td>
<td>-0.000761663</td>
<td>0.000426506</td>
<td>-0.00042877</td>
<td>-3.1367e-05</td>
</tr>
<tr>
<td>logtau</td>
<td>4.70099</td>
<td>4.7012</td>
<td>0.099249</td>
<td>4.53613</td>
<td>4.86476</td>
</tr>
<tr>
<td>phi</td>
<td>1.00228</td>
<td>0.943447</td>
<td>0.308244</td>
<td>0.625406</td>
<td>1.615</td>
</tr>
<tr>
<td>sigma_\phi</td>
<td>1.02992</td>
<td>1.02956</td>
<td>0.1422</td>
<td>0.786939</td>
<td>1.26465</td>
</tr>
</tbody>
</table>

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 \( \tau_i \) 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.
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.

<table>
<thead>
<tr>
<th>Model Type</th>
<th>Log-Marginal Likelihood</th>
<th>Type of log-ML Approximation</th>
<th>Prior Model Probability</th>
<th>Posterior Model Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>homosked</td>
<td>442.739</td>
<td>Lewis &amp; Raftery</td>
<td>0.333333</td>
<td>2.25774e-12</td>
</tr>
<tr>
<td>heterosked1</td>
<td>462.762</td>
<td>Lewis &amp; Raftery</td>
<td>0.333333</td>
<td>0.00112074</td>
</tr>
<tr>
<td>heterosked2</td>
<td>469.554</td>
<td>Lewis &amp; Raftery</td>
<td>0.333333</td>
<td>0.998879</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model Type</th>
<th>Log-Marginal Likelihood</th>
<th>Type of log-ML Approximation</th>
<th>Prior Model Probability</th>
<th>Posterior Model Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>homosked</td>
<td>442.746</td>
<td>Chib &amp; Jeliazkov</td>
<td>0.333333</td>
<td>2.555555e-12</td>
</tr>
<tr>
<td>heterosked1</td>
<td>462.417</td>
<td>Chib &amp; Jeliazkov</td>
<td>0.333333</td>
<td>0.000892387</td>
</tr>
<tr>
<td>heterosked2</td>
<td>469.438</td>
<td>Chib &amp; Jeliazkov</td>
<td>0.333333</td>
<td>0.999108</td>
</tr>
</tbody>
</table>

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

```r
// 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 } );
```

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, \( f(x) \), while explicitly recognizing that this production function returns the maximum possible output, \( y \), that can be produced given \( x \) 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, \( i \):

\[
TE_i = \frac{y_i}{f(x_i)}
\]

(4.10)

Technical efficiency is the ratio of observed output, \( y_i \), over maximum possible output, \( f(x_i) \), 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:

\[
\log y_i = \log f(x_i) - u_i + v_i
\]

(4.11)

where \( u_i \equiv -\log TE_i \). Because \( TE_i \in (0,1] \), \( u_i \) 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 \( f(x) \), 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 \( x_i \) 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 \( x_i \) is simply equal to \( \log x_i \). If, on the other hand, the production function is translog then \( x_i \) will contain the logarithms of inputs, as well as their squared terms and their interactions. With these definitions, the statistical model becomes:

\[
y_i = x_i'\beta - u_i + v_i, \quad v_i \sim N\left(0, \frac{1}{\tau}\right), \quad u_i \sim \text{Exp}(\lambda)
\]

(4.12)

The parameters of the model are \( \beta, \tau \) and \( \lambda \). The \( u_i \)'s are unknown and, in a data-augmentation setting, will be treated as latent data. Notice, however, that the \( u_i \)'s now have an interesting interpretation: because we defined \( u_i \) as \(-\log TE_i \), 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:

\[
p(y, u | x, \beta, \tau, \lambda) = \prod_{i=1}^{N} p(y_i | x_i, u_i, \beta, \tau) p(u_i | \lambda)
\]

\[
= \prod_{i=1}^{N} \frac{\tau^{1/2}}{(2\pi)^{1/2}} \exp \left\{ -\frac{\tau (y_i - x_i'\beta + u_i)^2}{2} \right\} \times \prod_{i=1}^{N} \lambda e^{-\lambda u_i}
\]

(4.13)
where, as before, \( y \) and \( X \) are the vector and matrix of the stacked values for the dependent and independent variables, respectively, and \( u \) is the \( N \times 1 \) vector of stacked values for the \( u_i \)s. 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 \( \beta \), with mean \( m \) and precision matrix \( P \), and a Gamma prior for \( \tau \), with shape and rate parameters \( a_\tau \) and \( b_\tau \), respectively. The values of the hyperparameters can be set such that these priors are very vague. For \( \lambda \), 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, \( a_\lambda \), equal to one and rate parameter, \( b_\lambda \), equal to \( -\log r^* \), where \( r^* \) 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:

\[
\pi(\beta, \tau, \lambda, u, y \mid X) \propto p(y, u \mid X, \beta, \tau, \lambda) p(\beta) p(\tau) p(\lambda)
\] (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 \( y^* = y - u \), it is easy to see that the full conditionals of \( \beta \) and \( \tau \) are the same as the ones presented in Theorem 2.1, with \( y^* \) assuming the role of \( y \). The full conditional of \( \lambda \) 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:

\[
y_i = x_i' \beta - u_i + v_i, \quad v_i \sim N(0, \frac{1}{\tau}), \quad u_i \sim \text{Exp}(\lambda)
\]

and with Normal prior for \( \beta \) and Gamma priors for \( \tau \) and \( \lambda \):

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

\[
p(\tau) = \frac{b_\tau^{a_\tau}}{\Gamma(a_\tau)} \tau^{a_\tau - 1} e^{-b_\tau \tau} \quad \text{and} \quad p(\lambda) = \frac{b_\lambda^{a_\lambda}}{\Gamma(a_\lambda)} \lambda^{a_\lambda - 1} e^{-b_\lambda \lambda}
\]

the full conditional of \( \beta \) is Normal:

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

and the full conditionals of \( \tau \) and \( \lambda \) are Gamma:

\[
\pi(\tau \mid \bullet) = \frac{\tilde{b}_{\tau}^{a_{\tau}}}{\Gamma(a_{\tau})} \tau^{a_{\tau} - 1} e^{-\tilde{b}_{\tau} \tau} \quad \text{and} \quad \pi(\lambda \mid \bullet) = \frac{\tilde{b}_{\lambda}^{a_{\lambda}}}{\Gamma(a_{\lambda})} \lambda^{a_{\lambda} - 1} e^{-\tilde{b}_{\lambda} \lambda}
\]

where:

- \( \tilde{P} = X'X + P \) and \( \tilde{m} = (X'X + P)^{-1} (X'y^* + Pm) \)
- \( \tilde{a}_{\tau} = \frac{N}{\tau} + a_{\tau} \) and \( \tilde{b}_{\tau} = \frac{1}{\tau} (y^* - X\beta)' (y^* - X\beta) + b_{\tau} \)
CHAPTER 4. DATA AUGMENTATION

\[ \tilde{a}_\lambda = N + a_\lambda \text{ and } \tilde{b}_\lambda = \sum_i u_i + b_\lambda \]

\[ y^* = y + u \]

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

\[ \pi(u_i|\bullet) = \frac{1}{(2\pi)^{1/2}} \Phi\left(\frac{u_i - \mu_i}{\tau}\right) \]

where \( \mu_i = -(y_i - x'_i\beta) - \frac{\lambda}{\tau} \).

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 \( u_i \sim N^+\left(0, \frac{1}{\phi}\right) \). The full conditionals of \( \beta \) and \( \tau \) are exactly the same as in the Exponential stochastic-frontier model. A Gamma prior for \( \phi \) is conjugate and the full conditional of \( \phi \) is Gamma with shape and rate parameters, \( \frac{N}{2} + a_\phi \) and \( \frac{\tilde{a}_\phi}{2} + \tilde{b}_\phi \), respectively. Finally, the full conditional of \( u_i \) is Normal, truncated from below at zero, but with different location and scale parameters:

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

where \( \mu_i = -\frac{\tau + \phi}{\tau} (y_i - x'_i\beta) \).

Example 4.2 US Electric Utilities

In this example we will use part of the dataset constructed and first used by Rungsiyawiboon & 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
- \( \text{trend} \): 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:

\[ \log q_i = \beta_1 + \beta_2 \log K_i + \beta_3 \log L_i + \beta_4 \log F_i + \beta_5 \text{trend}_i - u_i + v_i \]

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.

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Median</th>
<th>Sd.dev.</th>
<th>5%</th>
<th>95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>constant</td>
<td>5.11508</td>
<td>5.11499</td>
<td>0.130217</td>
<td>4.90232</td>
<td>5.32902</td>
</tr>
<tr>
<td>logK</td>
<td>0.179785</td>
<td>0.179915</td>
<td>0.0234004</td>
<td>0.141448</td>
<td>0.218299</td>
</tr>
<tr>
<td>logL</td>
<td>0.196873</td>
<td>0.197032</td>
<td>0.0191965</td>
<td>0.165351</td>
<td>0.228522</td>
</tr>
<tr>
<td>logF</td>
<td>0.620872</td>
<td>0.620971</td>
<td>0.023638</td>
<td>0.58204</td>
<td>0.659569</td>
</tr>
<tr>
<td>trend_7</td>
<td>0.0151138</td>
<td>0.0151111</td>
<td>0.00251212</td>
<td>0.0109742</td>
<td>0.0192037</td>
</tr>
<tr>
<td>tau</td>
<td>14.584</td>
<td>14.5772</td>
<td>0.660596</td>
<td>13.5131</td>
<td>15.6864</td>
</tr>
<tr>
<td>sigma_e</td>
<td>0.262057</td>
<td>0.261917</td>
<td>0.00595144</td>
<td>0.252489</td>
<td>0.272037</td>
</tr>
</tbody>
</table>

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.
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 $u_i$s, 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).
### 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 algorithm. 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 slightly 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, \( \alpha \), 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, \( \alpha \) 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:

\[
\int p(y, z | \theta, \alpha) \, dz = p(y | \theta)
\]  

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

$$\int_{\mathcal{A}} \left[ \int_{\mathcal{Z}} p(y,z|\theta,\alpha) p(\alpha|\theta) \, d\alpha \right] \, dz = p(y|\theta) \quad (4.16)$$

If we carry out the integration inside the square brackets analytically, we obtain $p(y,z|\theta)$ and, thus, revert back to standard data augmentation, where we iteratively sample from $p(z|y,\theta)$ and $p(\theta|y,z)$. 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 $\int_{\mathcal{A}} p(y,z|\theta,\alpha) \, d\alpha$ is likely more diffuse than the one based on $p(y,z|\theta,\alpha)$. This is desirable because, in a standard data augmentation setting, we could achieve zero autocorrelation in the draws if we could sample iteratively from $p(z|y)$ and $p(\theta|y,z)$. Thus, we should aim at having as diffuse a $p(z|y,\theta)$ as possible, up to the limit of $p(z|y)$. But, because:

$$p(z|y,\theta) = \frac{p(y,z|\theta)}{p(y|\theta)} \quad (4.17)$$

and the denominator is unaffected by the introduction of $\alpha$, inducing a more diffuse numerator would make $p(z|y,\theta)$ 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.\footnote{When this is not the case, improper priors on the working parameter are to be preferred. See, for example, J. Liu & Wu (1999).}

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, $\theta$, 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, $\mathcal{D}_a (w)$. This mapping is indexed by the working parameter and, according to Meng & van Dyk (1997), prominent choices are:

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

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

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

(a1) draw $\alpha$ from $p(\alpha)$
(a2) draw $w$ from $p(w|y,\theta,\alpha)$

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

(b1) draw $\alpha$ from $p(\alpha|y,w)$
(b2) draw $\theta$ from $p(\theta|y,w,\alpha)$

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

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:

\[
y_{it} = x_{it}'\beta + \varepsilon_{it}, \quad \varepsilon_{it} \sim N(0, \frac{1}{\sigma^2})
\]  

(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, \( x \), as well as a set of time-invariant independent variables, \( w \):

\[
y_{it} = x_{it}'\beta + w_{it}'\gamma + \varepsilon_{it}, \quad \varepsilon_{it} \sim N(0, \frac{1}{\sigma^2})
\]  

(5.2)

where \( \beta \) is a \( K \times 1 \) of parameters associated with the time-varying variables and \( \gamma \) is a vector of parameters associated with the time-invariant variables.\(^1\) If both \( x \) and \( w \) are observed, then the pooled model can produce estimates of both \( \beta \) and \( \gamma \). With the availability of panel data, estimates of \( \beta \) can be obtained even if the time invariant variables are unobserved. By defining \( \alpha_i = w_{it}'\gamma \) as the unobserved group effect, the model above becomes:

\[
y_{it} = \alpha_i + x_{it}'\beta + \varepsilon_{it}, \quad \varepsilon_{it} \sim N(0, \frac{1}{\sigma^2})
\]  

(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 characteristics 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 \( \beta \). Second, because typical panel datasets consist of many groups but have a short time dimension, the number of \( \alpha_i \)'s 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 \( \alpha_i \)'s are independent from the variables in \( x \) and that each one of them follows a Normal distribution with a common mean and precision \( \omega \). If the set of independent variables includes a constant term, this model can be expressed as:

\[
y_{it} = \alpha_i + x_{it}'\beta + \varepsilon_{it}, \quad \varepsilon_{it} \sim N(0, \frac{1}{\sigma^2}), \quad \alpha_i \sim N(0, \frac{1}{\omega})
\]  

(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 \( w \) has only an \( i \) subscript because it does not vary over \( t \).
constant term in \( x \). 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 \( \beta, \tau \) and all \( \alpha_i \). In a panel dataset with a short time dimension, the priors placed on each \( \alpha_i \) 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 \( x \).

Estimation of the \( \alpha_i \)s 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 \( \bar{x}_i \) and \( \bar{\varepsilon}_i \), accordingly. If (5.3) holds in the population, then it should also hold:

\[
\bar{y}_i = \alpha_i + \bar{x}_i' \beta + \bar{\varepsilon}_i. \tag{5.5}
\]

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:

\[
(y_{it} - \bar{y}_i) = (x_{it} - \bar{x}_i)' \beta + (\varepsilon_{it} - \bar{\varepsilon}_i). \tag{5.6}
\]

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 \( \tau \) that appears in the distribution of each \( \varepsilon_{it} \). Therefore, \( \beta \) 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 \( \varepsilon_{it} \) 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:

\[
y_{it} = \gamma_i' z_i + \varepsilon_{it}, \quad \varepsilon_{it} \sim N(0, \frac{1}{\tau}) \tag{5.7}
\]

If the number of time observations per group is greater than the number of independent variables in \( z \), then the individual \( \gamma_i \)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 \( \gamma_i \) and having limited information to estimate each \( \gamma_i \) in isolation. On the other hand, the random-coefficients model imposes a hierarchical structure on the \( \gamma_i \)s in a similar fashion the random-effects model does on the \( \alpha_i \)s. Compared to running \( N \) regressions separately, this hierarchical structure allows information about \( \gamma_i \) 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 \( \gamma_i \) is now a \( K \times 1 \) random vector, the typical assumption made is that \( \gamma_i \)s are draws from a multivariate-Normal distribution with common mean and precision matrix:

\[
\gamma_i \sim N(\bar{\gamma}, \Omega^{-1}) \tag{5.8}
\]
where $\tilde{\gamma}$ is a $K \times 1$ vector of parameters to be estimated and $\Omega$ 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 $\beta$ from (5.6) is exactly the same as the one used for the linear regression model. Furthermore, estimating the $\alpha_i$s 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 ($\alpha_i$s or $\gamma_i$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:

$$ y_{it} = \alpha_i + x_{it}'\beta + \varepsilon_{it}, \quad \varepsilon_{it} \sim N \left(0, \frac{1}{\omega}\right), \quad \alpha_i \sim N \left(0, \frac{1}{\tau}\right) \tag{5.9} $$

where $x_{it}$ 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 $\beta$, $\tau$ and $\omega$, while in a data-augmentation setting, the $\alpha_i$s are the latent data.

The complete-data likelihood for the random-effects model is:

$$ p(y, \{\alpha_i\} | X, \beta, \tau, \omega) = \prod_{i=1}^{N} \prod_{t=1}^{T} p(y_{it} | x_{it}, \beta, \alpha_i) \times p(\alpha_i | \omega) $$

$$ = \frac{\tau^{NT/2}}{(2\pi)^{NT/2}} \exp \left\{ -\frac{\tau}{2} \sum_{i=1}^{N} \sum_{t=1}^{T} (y_{it} - \alpha_i - x_{it}'\beta)^2 \right\} \times \frac{\omega^{N/2}}{(2\pi)^{N/2}} \exp \left\{ -\frac{\omega}{2} \sum_{i=1}^{N} \alpha_i^2 \right\} \tag{5.10} $$

where $y$ and $X$ 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 $\varepsilon_{it}$ follows a Normal distribution and the second factor is due to each $\alpha_i$ following a Normal distribution with zero mean.

As in the linear regression model, we will use a multivariate-Normal prior for $\beta$ and a Gamma prior for $\tau$. Furthermore, since $\omega$ is another precision parameter, we will use a Gamma prior for it as well. By letting $y^\ast$ be the $NT \times 1$ vector of stacked values of $y_{it} - \alpha_i$, it becomes apparent that the full conditionals of $\beta$ and $\tau$ are exactly the same as the ones presented in Theorem 2.1, with $y^\ast$ taking the place of $y$. Similar transformations as the ones presented
above the same Theorem can be used to show that the full conditional of \( \omega \) is Gamma and that the full conditional of each \( \alpha_i \) 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:

\[
y_{it} = \alpha_i + x_{it}' \beta + \varepsilon_{it}, \quad \varepsilon_{it} \sim N \left(0, \frac{\omega}{\tau} \right), \quad \alpha_i \sim N \left(0, \frac{\omega}{\tau} \right)
\]

and with a Normal prior for \( \beta \) and Gamma priors for \( \tau \) and \( \omega \):

\[
p(\beta) = \frac{|P|^{1/2}}{(2\pi)^{K/2}} \exp \left\{ -\frac{1}{2} (\beta - m)' P (\beta - m) \right\}, \quad p(\tau) = \frac{b_\tau^{a_\tau}}{\Gamma(a_\tau)} \tau^{a_\tau - 1} e^{-b_\tau \tau} \quad \text{and} \quad p(\omega) = \frac{b_\omega^{a_\omega}}{\Gamma(a_\omega)} \omega^{a_\omega - 1} e^{-b_\omega \omega}
\]

the full conditional of \( \beta \) is Normal:

\[
\pi(\beta | \bullet) = \frac{|P|^{1/2}}{(2\pi)^{K/2}} \exp \left\{ -\frac{1}{2} (\beta - \hat{m})' \hat{P} (\beta - \hat{m}) \right\}
\]

and the full conditionals of \( \tau \) and \( \omega \) are Gamma:

\[
\pi(\tau | \bullet) = \frac{\tilde{b}_\tau^{a_\tau}}{\Gamma(\tilde{a}_\tau)} \tilde{a}_\tau^{a_\tau - 1} e^{-\tilde{b}_\tau \tau} \quad \text{and} \quad \pi(\omega | \bullet) = \frac{\tilde{b}_\omega^{a_\omega}}{\Gamma(\tilde{a}_\omega)} \omega^{a_\omega - 1} e^{-\tilde{b}_\omega \omega}
\]

where:

- \( \hat{P} = X'X + P \) and \( \hat{m}_\beta = (X'X + P)^{-1} (X'y^* + Pm) \)
- \( \tilde{a}_\tau = \frac{NT}{2} + a_\tau \) and \( \tilde{b}_\tau = \frac{1}{2} (y^* - X\beta)' (y^* - X\beta) + b_\tau \)
- \( \tilde{a}_\omega = \frac{N}{2} + a_\omega \) and \( \tilde{b}_\omega = \frac{1}{2} \sum_{i=1}^{N} \alpha_i^2 + b_\omega \)
- \( y^* \) is the \( NT \times 1 \) vector obtained by stacking \( y^*_{it} = y_{it} - \alpha_i \) over \( T \) and \( N \)

The full conditional of \( \alpha_i, i = 1, 2, \ldots, N, \) 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 \( \tilde{m}_i = \frac{\tau}{\tau T + \omega} \sum_{t=1}^{T} (y_{it} - x_{it}'\beta) \)

◆ **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:

\[
\log Y_{it} = \beta_1 + \beta_2 \log K_{it} + \beta_3 \log L_{it} + \beta_4 \text{trend}_{it} + \varepsilon_{it}, \quad \varepsilon_{it} \sim N \left(0, \frac{\omega}{\tau} \right)
\]

The results from the pooled model are presented in the following table.
CHAPTER 5. THE LINEAR MODEL WITH PANEL DATA

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

The slope coefficients in the fixed-effects model:

\[
\log Y_{it} = \alpha_i + \beta_2 \log K_{it} + \beta_3 \log L_{it} + \beta_4 \text{trend}_{it} + \varepsilon_{it}, \quad \varepsilon_{it} \sim N(0, \tau)
\]

can be estimated using BayES’ \texttt{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.

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Median</th>
<th>Sd.dev.</th>
<th>5%</th>
<th>95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>dlogK</td>
<td>0.393942</td>
<td>0.393992</td>
<td>0.0333159</td>
<td>0.338807</td>
<td>0.448848</td>
</tr>
<tr>
<td>dlogL</td>
<td>0.563322</td>
<td>0.563276</td>
<td>0.0327412</td>
<td>0.509985</td>
<td>0.617021</td>
</tr>
<tr>
<td>dtrend</td>
<td>0.00489146</td>
<td>0.00489257</td>
<td>0.000855796</td>
<td>0.00348505</td>
<td>0.00630568</td>
</tr>
<tr>
<td>tau</td>
<td>176.757</td>
<td>176.619</td>
<td>9.69914</td>
<td>161.188</td>
<td>192.928</td>
</tr>
<tr>
<td>sigma_e</td>
<td>0.0753014</td>
<td>0.0752457</td>
<td>0.00207175</td>
<td>0.0719966</td>
<td>0.0787679</td>
</tr>
</tbody>
</table>

Finally, BayES’ \texttt{lm_re()} function can be used to estimate the random-effects model:

\[
\log Y_{it} = \alpha_i + \beta_2 \log K_{it} + \beta_3 \log L_{it} + \beta_4 \text{trend}_{it} + \varepsilon_{it}, \quad \varepsilon_{it} \sim N(0, \tau), \quad \alpha_i \sim N(0, \omega)
\]

These results appear in the following table.

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Median</th>
<th>Sd.dev.</th>
<th>5%</th>
<th>95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>constant</td>
<td>5.21101</td>
<td>5.20963</td>
<td>0.383709</td>
<td>4.58129</td>
<td>5.84385</td>
</tr>
<tr>
<td>logK</td>
<td>0.419055</td>
<td>0.419135</td>
<td>0.0312374</td>
<td>0.36744</td>
<td>0.470369</td>
</tr>
<tr>
<td>logL</td>
<td>0.576326</td>
<td>0.576208</td>
<td>0.0293342</td>
<td>0.528282</td>
<td>0.624983</td>
</tr>
<tr>
<td>trend</td>
<td>0.00396158</td>
<td>0.0039528</td>
<td>0.000652133</td>
<td>0.00290387</td>
<td>0.00504271</td>
</tr>
<tr>
<td>tau</td>
<td>172.605</td>
<td>172.478</td>
<td>9.56803</td>
<td>156.993</td>
<td>188.641</td>
</tr>
<tr>
<td>omega</td>
<td>93.0941</td>
<td>87.6235</td>
<td>37.8641</td>
<td>41.6574</td>
<td>162.728</td>
</tr>
<tr>
<td>sigma_e</td>
<td>0.0762035</td>
<td>0.076144</td>
<td>0.00212013</td>
<td>0.0728087</td>
<td>0.0798106</td>
</tr>
<tr>
<td>sigma_alpha</td>
<td>0.110479</td>
<td>0.106834</td>
<td>0.0240874</td>
<td>0.0783917</td>
<td>0.154945</td>
</tr>
</tbody>
</table>

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 \(\log Y_{it}\) 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.

<table>
<thead>
<tr>
<th>Model</th>
<th>Log-Marginal Likelihood</th>
<th>Type of log-ML Approximation</th>
<th>Prior Model Probability</th>
<th>Posterior Model Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>LM</td>
<td>442.739</td>
<td>Lewis &amp; Raftery</td>
<td>0.333333</td>
<td>1.44989e-136</td>
</tr>
<tr>
<td>FE</td>
<td>755.519</td>
<td>Lewis &amp; Raftery</td>
<td>0.333333</td>
<td>1</td>
</tr>
<tr>
<td>RE</td>
<td>702.433</td>
<td>Lewis &amp; Raftery</td>
<td>0.333333</td>
<td>8.81473e-24</td>
</tr>
</tbody>
</table>
5.3. ESTIMATION OF THE HIERARCHICAL PANEL-DATA MODELS

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

```csharp
// 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:

\[ y_{it} = z_{it}' \gamma_i + \varepsilon_{it}, \quad \varepsilon_{it} \sim N\left(0, \frac{1}{\tau}\right), \quad \gamma_i \sim N\left(\bar{\gamma}, \Omega^{-1}\right) \quad (5.11) \]

where \(z_{it}\) is a \(K \times 1\) vector of independent variables and each \(\gamma_i\) 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 \(\bar{\gamma}\), \(\Omega\) and \(\tau\). The \(\gamma_i\)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:

\[
p(y, \{ \gamma_i \} | Z, \bar{\gamma}, \Omega, \tau) = \prod_{i=1}^{N} \left[ \prod_{t=1}^{T} p(y_{it} | z_{it}, \tau, \gamma_i) \right] \times p(\gamma_i | \bar{\gamma}, \Omega) \\
= \frac{\tau^{NT/2}}{(2\pi)^{NT/2}} \exp \left\{ -\frac{\tau}{2} \sum_{i=1}^{N} \sum_{t=1}^{T} (y_{it} - z_{it}' \gamma_i)^2 \right\} \\
\times \frac{|\Omega|^{N/2}}{(2\pi)^{NK/2}} \exp \left\{ -\frac{1}{2} \sum_{i=1}^{N} (\gamma_i - \bar{\gamma})' \Omega (\gamma_i - \bar{\gamma}) \right\}
\]

where \(y\) and \(Z\) 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 \(\varepsilon_{it}\) follows a Normal distribution and the second factor is due to each \(\gamma_i\) following a multivariate-Normal distribution.

As in the random-effects model, we will use a multivariate-Normal prior for \(\bar{\gamma}\) and a Gamma prior for \(\tau\). Because \(\Omega\) is a precision matrix, we will place a Wishart prior on it, with degrees-of-freedom parameter \(n\) and scale matrix \(V\). All three priors are conjugate, while the derivations of the full conditionals for \(\gamma\) and \(\tau\) follow similar steps as the ones used for the linear regression...
model. Deriving the full conditional of $\Omega$ requires transformations similar to the ones used in SUR model. Finally, the full conditional of each $\gamma_i$ is multivariate Normal and its derivation is similar to the way the full conditional of $\beta$ 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:

$$y_{it} = z_{it}'\gamma_i + \varepsilon_{it}, \quad \varepsilon_{it} \sim N(0, \frac{1}{\tau}), \quad \gamma_i \sim N(\bar{\gamma}, \Omega^{-1})$$  \hspace{1cm} (5.12)

and with a Normal prior for $\bar{\gamma}$, a Wishart prior for $\Omega$ and a Gamma prior for $\tau$:

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

$$p(\Omega) = \frac{\Omega^{n-K-1}}{2^{n/2} \pi^{n/2} \Gamma(K/2)} \exp \left\{ -\frac{1}{2} \text{tr} \left( V^{-1} \Omega \right) \right\},$$

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

the full conditional of $\bar{\gamma}$ is Normal:

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

the full conditionals of $\Omega$ are Wishart and the full conditional of $\tau$ is Gamma:

$$\pi(\Omega \mid \bullet) = \frac{\tilde{n}-K-1}{2^{n/2} \pi^{n/2} \Gamma(K/2)} \exp \left\{ -\frac{1}{2} \text{tr} \left( \tilde{V}^{-1} \Omega \right) \right\},$$

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

where:

- $\tilde{P} = N\Omega + P$ and $\tilde{m} = (N\Omega + P)^{-1} (\Omega \sum_{i=1}^{N} \gamma_i + Pm)$
- $\tilde{n} = N + n$, $\tilde{V}^{-1} = CC' + V^{-1}$ and $C = [\gamma_1 - \bar{\gamma} \quad \gamma_2 - \bar{\gamma} \quad \ldots \quad \gamma_t - \bar{\gamma}]$
- $\tilde{a}_r = \frac{NT}{2} + a_r$ and $\tilde{b}_r = \frac{1}{2} \sum_{i=1}^{N} \sum_{t=1}^{T} \frac{1}{2} (y_{it} - z_{it}'\gamma_i)^2 + b_r$

The full conditional of $\gamma_i$, $i = 1, 2, \ldots, N$, is multivariate Normal:

$$\pi(\gamma_i \mid \bullet) = \frac{|\tilde{\Omega}|^{1/2}}{(2\pi)^{K/2}} \exp \left\{ -\frac{1}{2} (\gamma_i - \tilde{\gamma}_i)' \tilde{\Omega} (\gamma_i - \tilde{\gamma}_i) \right\}$$

where:

- $\tilde{\Omega} = \tau z_i'z_i + \Omega$
- $\tilde{\gamma}_i = (\tau z_i'z_i + \Omega)^{-1} (\tau z_i'y_i + \Omega \gamma)$
- $y_i$ and $z_i$ 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:

$$\log Y_{it} = \gamma_{i1} + \gamma_{i2} \log K_{it} + \gamma_{i3} \log L_{it} + \gamma_{i4} \text{trend}_{it} + \varepsilon_{it}, \quad \varepsilon_{it} \sim N(0, \frac{1}{\tau}), \quad \gamma_i \sim N(\bar{\gamma}, \Omega^{-1})$$

This model can be estimated in BayES using the `lm_re()` function. The results are presented in the following table.
5.3. ESTIMATION OF THE HIERARCHICAL PANEL-DATA MODELS

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Median</th>
<th>Sd.dev.</th>
<th>5%</th>
<th>95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>constant</td>
<td>5.36874</td>
<td>5.64416</td>
<td>0.85051</td>
<td>3.57618</td>
<td>6.20717</td>
</tr>
<tr>
<td>logK</td>
<td>0.418731</td>
<td>0.401558</td>
<td>0.0658398</td>
<td>0.34499</td>
<td>0.555542</td>
</tr>
<tr>
<td>logL</td>
<td>0.49582</td>
<td>0.49558</td>
<td>0.0774272</td>
<td>0.369143</td>
<td>0.623302</td>
</tr>
<tr>
<td>trend</td>
<td>0.00537389</td>
<td>0.00532298</td>
<td>0.0105028</td>
<td>-0.0118546</td>
<td>0.0226852</td>
</tr>
<tr>
<td>tau</td>
<td>630.637</td>
<td>619.458</td>
<td>58.1705</td>
<td>550.6</td>
<td>740.025</td>
</tr>
<tr>
<td>sigma_e</td>
<td>0.0399441</td>
<td>0.0401789</td>
<td>0.00179197</td>
<td>0.0367607</td>
<td>0.0426187</td>
</tr>
</tbody>
</table>

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 $\gamma$, the random-coefficients model contains additional parameters. The posterior mean of $\Omega$ for this model is:

$E(\Omega|\bullet) = \begin{bmatrix} 252.71951 & 28.404319 & -40.135201 & -0.50773612 \\ 28.404319 & 554.45376 & 65.313201 & 20.203856 \\ -40.135201 & 65.313201 & 111.26204 & 14.386714 \\ -0.50773612 & 20.203856 & 14.386714 & 738.18898 \end{bmatrix}$

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.

```csharp
// 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 );
```
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:

\[ y_{it} = x_{it}' \beta + w_i' \gamma + \varepsilon_{it}, \quad \varepsilon_{it} \sim N \left( 0, \frac{1}{\tau} \right) \]  

(5.13)

where \( x_{it} \) is a \( K \times 1 \) vector of time-varying and \( w_i \) 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 \( z_i \) be an \( L \times 1 \) vector of observed time-invariant variables and \( \delta \) an \( L \times 1 \) vector of parameters associated with these variables. \( w_i \) now represents a vector of unobserved time-invariant variables, which can be linearly combined to form the group effect, \( \alpha_i \). Thus, by estimating the model:

\[ y_{it} = \alpha_i + x_{it}' \beta + z_i' \delta + \varepsilon_{it}, \quad \varepsilon_{it} \sim N \left( 0, \frac{1}{\tau} \right), \quad \alpha_i \sim N \left( 0, \frac{1}{\omega} \right) \]  

(5.14)

one can produce estimates for the parameters associated with the time-varying variables, \( \beta \), as well as for the parameters associated with the observed time-invariant variables, \( \delta \), while at the same time controlling for time-invariant unobserved heterogeneity. An additional assumption is made in the background: the \( \alpha_i \)'s are independent of both the time-varying variables in \( x \) and the observed time-invariant in \( z \).

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 \( x \):

\[ \alpha_i = x_{it}' \lambda + \nu_i \]  

(5.15)
5.4. EXTENSIONS TO OTHER PANEL-DATA MODELS

where \( v_i \) is a Normally-distributed error term. Substituting the last expression in (5.14) leads to:

\[
y_{it} = v_i + x_i' \lambda + x_{ii}' \beta + z_i' \delta + \varepsilon_{it}, \quad \varepsilon_{it} \sim N \left(0, \frac{1}{\tau}\right), \quad v_i \sim N \left(0, \frac{1}{\omega}\right) \tag{5.16}
\]

From this expression it becomes apparent that \( v_i \) takes the place of \( \alpha_i \) in the typical random-effects model, while we still need to assume that \( \alpha_i \) is uncorrelated with \( z_i \). Estimation of \( \beta \) and \( \delta \) 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, \( \lambda \). At the same time, any correlation of the form expressed in (5.15) between the time-varying independent variables and the \( \alpha_i \)s 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 \( \beta \). This result, however, does not hold exactly in a Bayesian setting, because the posterior mean of \( \beta \) 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:

\[
\alpha_i = x_{i1}' \lambda_1 + x_{i2}' \lambda_2 + \ldots + x_{iT}' \lambda_T + v_i \tag{5.17}
\]

where \( x_{it} \) is a \( K \times 1 \) vector of the values of \( x \) for group \( i \) and in period \( t \) and \( \lambda_i \) is the associated \( K \times 1 \) vector of coefficients. Substituting this expression in (5.14) and collecting terms leads to:

\[
y_{it} = v_i + \sum_{s \neq t} x_{is}' \lambda_s + x_{it}' (\beta + \lambda_T) + z_i' \delta + \varepsilon_{it}, \quad \varepsilon_{it} \sim N \left(0, \frac{1}{\tau}\right), \quad v_i \sim N \left(0, \frac{1}{\omega}\right) \tag{5.18}
\]

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 \( x \), to severe multicollinearity problems. Additionally, they require that any observed time-invariant variables included in the model are independent of the \( \alpha_i \)s; otherwise \( \delta \) will capture, apart from the effect of the variables in \( w \) 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:

\[
y_{it} = z_{it}' \gamma_i + \varepsilon_{it}, \quad \varepsilon_{it} \sim N \left(0, \frac{1}{\tau}\right), \quad \gamma_i \sim N \left(\hat{\gamma}, \Omega^{-1}\right) \tag{5.19}
\]

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 \( \Omega \) needs to be estimated along with \( \hat{\gamma} \) and the number of unique elements in \( \Omega \) is a quadratic function of the number of independent variables. Furthermore, the resulting group-specific \( \gamma_i \)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 \( \hat{\gamma} \) 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:

\[
y_{it} = z_{it}'\gamma_i + x_{it}'\beta + \varepsilon_{it}, \quad \varepsilon_{it} \sim N(0, \frac{1}{\tau^2}), \quad \gamma_i \sim N(\bar{\gamma}, \Omega^{-1})
\] (5.20)

where \( x_{it} \) is an \( L \times 1 \) vector of independent variables which are associated with a parameter vector, \( \beta \), 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, \( \bar{\gamma}, \Omega \) and \( \tau \), as in the basic random-coefficients model, plus \( \beta \), 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 \( y_{it} - x_{it}'\beta \) in the full conditional for \( \gamma_i \), by \( y_{it} - z_{it}'\gamma_i \) in the full conditional of \( \beta \) in the typical linear regression model, and use the complete residual, \( y_{it} - z_{it}'\gamma_i - x_{it}'\beta \), in the full conditional of \( \tau \).

### 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 \( \gamma_i \)'s. The assumption that we maintained until now in this model is that each \( \gamma_i \) is a draw from a multivariate-Normal distribution with common mean, \( \bar{\gamma} \). 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 \( \gamma_i \) becomes:

\[
\gamma_i \sim N(W_i\xi, \Omega^{-1})
\] (5.21)

where \( W_i \) is a matrix constructed by the variables which affect the mean of the \( \gamma_i \)'s and \( \xi \) 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:

\[
\gamma_i = W_i\xi + u_i, \quad u_i \sim N(0, \Omega^{-1})
\] (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 \( \tau \) is the same as in the typical random-coefficients model and the full conditionals of \( \xi \) and \( \Omega \) are the ones presented in Theorem 3.1, with \( \gamma_i \) assuming the role of \( \bar{y}_i \), \( W_i \) that of \( X_i \) and \( \xi \) that of \( \beta \). Finally, the full conditional of each \( \gamma_i \) is the same as in the typical random-coefficients model, with \( W_i\xi \) replacing \( \bar{\gamma} \).

### 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 Choice

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 choice. 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 models, which always have to rely on numerical data. Instead of modeling the response variable directly, binary-choice 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-choice models, some of which are useful for estimating their parameters, while others facilitate interpretation of the results. After defining formally what a binary-choice model is, the statistical formulation is presented in the following section. Binary-choice models, however, 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. The final section of this chapter extends the binary-choice models to accommodate panel data.

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-Choice Models

Binary-choice or binary-response 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-choice 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

\[ \text{Prob}(y_i = 1 | x_i) \]

where \( x_i \) 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 \( \text{Prob}(y_i = 1 | x_i) \) 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:

\[ p_i = x_i' \beta + \varepsilon_i \]  \hspace{1cm} (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-choice 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, then \( y_i \) follows a Bernoulli distribution with probability mass function:

\[ p(y_i) = p_i^{y_i} \cdot (1 - p_i)^{1 - y_i} \]  \hspace{1cm} (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 \( x_i' \beta \). 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:

\[ p_i = F(x_i' \beta) \]  \hspace{1cm} (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:

\[ p(y | X, \beta) = \prod_{i=1}^{N} [F(x_i' \beta)]^{y_i} [1 - F(x_i' \beta)]^{1 - y_i} \]  \hspace{1cm} (6.4)
where \(y\) is the \(N \times 1\) vector that stores the values (0 or 1) of the response variable for \(N\) potential observations and \(X\) 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 \(x_i'\beta\) 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:

\[
\Phi(x_i'\beta) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{t^2}{2}\right\} dt
\]

(6.5)

and the cumulative density function of a standard-Logistically distributed random variable:

\[
\Lambda(x_i'\beta) = \frac{1}{1 + e^{-x_i'\beta}} = \frac{e^{x_i'\beta}}{1 + e^{x_i'\beta}}
\]

(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 \(\beta\). This made Logit the model of choice when computers were too slow to accurately approximate the value of \(\Phi(x_i'\beta)\), 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, \(\varepsilon_i\). This is because, by modeling the probability of the response variable being true rather than \(y_i\) 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:

\[
y_i^* = x_i'\beta + \varepsilon_i
\]

(6.7)

\[
y_i = \begin{cases} 
1 & \text{if } y_i^* > 0 \\
0 & \text{if } y_i^* \leq 0
\end{cases}
\]

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 \(\varepsilon_i\) 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:

\[
\text{Prob}(y_i = 1|X_i) = \text{Prob}(y_i^* > 0|X_i)
\]

(6.8)

\[
= \text{Prob}(\varepsilon_i > -x_i'\beta|X_i)
\]

\[
= 1 - \text{Prob}(\varepsilon_i \leq -x_i'\beta|X_i)
\]

To proceed, let \(F(\cdot)\) be the cumulative density function of \(\varepsilon_i\). If this function is such that the probability density function of \(\varepsilon_i\), \(F'(\cdot)\), is symmetric around zero, then \(1 - F(z) = F(-z)\) for all real \(z\). Thus:

\[
\text{Prob}(y_i = 1|X_i) = F(x_i'\beta)
\]

(6.9)
which is exactly what was specified in (6.3). Therefore, if we assume that \( \varepsilon_i \) follows a standard-Normal distribution, then the latent-variable representation leads to the Probit model, while if \( \varepsilon_i \) 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, 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 \( \varepsilon_i \) to one? Suppose that we allow this scale parameter, \( \sigma^2 \), to be different from one. Then:

\[
\text{Prob}(y_i = 1|x_i) = \text{Prob}(\varepsilon_i \leq x_i'\beta|x_i) = \text{Prob}\left(\frac{\varepsilon_i}{\sigma} \leq \frac{x_i'\beta}{\sigma} \Bigg| x_i\right) \quad (6.10)
\]

Because \( \varepsilon_i/\sigma \) has a standardized distribution, the conditional probability of success is equal to the standard-Normal or standard-Logistic cumulative density function, evaluated at \( x_i'\beta/\sigma \). This makes apparent that multiplying \( \beta \) and \( \sigma \) by any real number, \( z \), the cumulative density function of the standardized distribution will return exactly the same value for the combination \( (z\beta, z\sigma) \) as for the combination \( (\beta, \sigma) \). In other words, there is an infinite number of parameters that produce exactly the same probability of success and, therefore, \( \beta \) and \( \sigma \) 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 \( \sigma \) to unity. In this way, \( \varepsilon_i \) is restricted to follow either a standard-Normal or standard-Logistic distribution and the two formulations of binary-choice models are equivalent.

### 6.2.1 Random Utility: An Underlying Framework for Binary Choice

The two alternative formulations of binary-choice 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 essentially 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:

\[
u_{1i}^* = x_i'\gamma_1 + \epsilon_{1i} \quad (6.11)
\]

if she buys the product, and:

\[
u_{0i}^* = x_i'\gamma_0 + \epsilon_{0i} \quad (6.12)
\]

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.

\(^2\)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.
6.3. ESTIMATION OF BINARY-RESPONSE MODELS

statements about whether the consumer will purchase the product or not. The probability of the consumer buying the product is:

\[
\text{Prob}(u_{1i}^* > u_{0i}^* | x_i) = \text{Prob}(x_i' (\gamma_1 - \gamma_0) + (\epsilon_{1i} - \epsilon_{0i}) > 0 | x_i) = \text{Prob}(x_i' \beta + \epsilon_i > 0 | x_i) \tag{6.13}
\]

where \( \beta \equiv \gamma_1 - \gamma_0 \) and \( \epsilon_i \equiv \epsilon_{1i} - \epsilon_{0i} \). 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 \( \epsilon_i \) and we obtain the Probit model. To get the Logit model we need to assume that \( \epsilon_{1i} \) and \( \epsilon_{0i} \) are independent 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 or product characteristics may be important in determining the decision to buy or not. These are the characteristics that could affect the utility that the consumer derives from buying. 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 its value such that the scale parameter of \( \epsilon_i \) is equal to one.

6.3 Estimation of Binary-Response Models

There are two alternative approaches for estimating the parameters of binary-choice 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:

\[
g(\mu_i) = x_i' \beta \tag{6.14}
\]

where \( \mu_i \) is the expected value of the dependent variable and \( g(\cdot) \) is the associated link function. In binary-choice 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 \( x_i' \beta \) 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-choice models, this approach cannot be easily extended to accommodate more complex discrete-choice 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

\(^3\)The probability density and cumulative density functions of a random variable that follows a type I extreme-value distribution are, respectively, \( p(z) = e^{-z} e^{-e^{-z}} \) and \( P(z) = e^{-e^{-z}} \). The random variable that is obtained as the difference of two independent type-I extreme value distributed random variables follows a standard-Logistically 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:

\[
\text{Prob}(U < u) = \int_{-\infty}^{u} \text{Prob}(U < u | w) p(w) \, dw = \int_{-\infty}^{\infty} P(u+w) p(w) \, dw = \int_{-\infty}^{\infty} e^{-w} e^{-e^{-(1+e^{-u})}} \, dw = \frac{e^{\frac{u}{1+e^{u}}}}}{1+e^{u}}
\]

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.
importantly, estimation via data augmentation scales well when more complex discrete-choice 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 following two subsections.

6.3.1 Estimation of the Binary Probit Model

Consider the latent-variable representation of the Probit model:

\[ y_i^* = x_i' \beta + \varepsilon_i, \quad \varepsilon_i \sim N(0, 1) \]

\[ y_i = \begin{cases} 
1 & \text{if } y_i^* > 0 \\
0 & \text{if } y_i^* \leq 0 
\end{cases} \tag{6.15} \]

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 \( p(y_i, y_i^*|x_i, \beta) = p(y_i^*|x_i, \beta) \times p(y_i|x_i, \beta) \). Given that \( \varepsilon_i \) follows a standard-Normal distribution, the second factor in this expression is:

\[ p(y_i^*|x_i, \beta) = (2\pi)^{-1/2} \exp \left\{ -\frac{(y_i^* - x_i' \beta)^2}{2} \right\} \tag{6.16} \]

Although the latent-variable representation of the binary Probit model explicitly specifies the relationship between \( y_i \) and \( y_i^* \), expressing \( p(y_i|y_i^*) \) 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:

\[ p(y_i|y_i^*) = \mathbb{I}(y_i^* > 0)^{y_i} \cdot \mathbb{I}(y_i^* \leq 0)^{1-y_i} \tag{6.17} \]

where \( \mathbb{I}(\cdot) \) 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 \( \mathbb{I}(y_i^* > 0) \). 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:

\[ p(y, y^*|X, \beta) = \prod_{i=1}^{N} \left[ \mathbb{I}(y_i^* > 0)^{y_i} \cdot \mathbb{I}(y_i^* \leq 0)^{1-y_i} \times (2\pi)^{-1/2} \exp \left\{ -\frac{1}{2} (y_i^* - x_i' \beta)^2 \right\} \right] \tag{6.18} \]

where \( y \) and \( y^* \) are \( N \times 1 \) vectors that store the values of the observed and latent data, respectively, and \( X \) 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 \( \beta \). As we have done until now for slope coefficients, we will place a multivariate-Normal prior on \( \beta \), with mean vector \( m \) and precision matrix \( P \). A standard application of Bayes’ theorem leads to the following posterior density for \( \beta \) and the latent data:

\[ \pi (\beta, y^*|y, X, \beta) \propto p(y, y^*|X, \beta) \ p(\beta) \]

\[ = \prod_{i=1}^{N} \left[ \mathbb{I}(y_i^* > 0)^{y_i} \cdot \mathbb{I}(y_i^* \leq 0)^{1-y_i} \times (2\pi)^{-1/2} \exp \left\{ -\frac{1}{2} (y_i^* - x_i' \beta)^2 \right\} \right] \times \frac{|P|^{1/2}}{(2\pi)^{K/2}} \exp \left\{ -\frac{1}{2} (\beta - m)' P (\beta - m) \right\} \tag{6.19} \]

In deriving the full conditional of \( \beta \) we first need to drop all terms from the posterior density that do not involve \( \beta \) 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:
\[
\pi (\beta | \bullet) \propto \exp \left\{ -\frac{1}{2} (y^* - X\beta)' (y^* - X\beta) \right\} \times \exp \left\{ -\frac{1}{2} (\beta - m)' P (\beta - m) \right\}
\] (6.20)

The only difference from the expression in (2.13) is that \( \tau \), the precision parameter of the error term, is now restricted to be equal to one. Following exactly the same steps as the ones below equation (2.13) leads to the full conditional of \( \beta \) being a multivariate-Normal density with mean \( (X'X + P)^{-1} (X'y^* + Pm) \) and precision matrix \( X'X + P \).

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:
\[
\pi (y^*_i | \bullet) \propto I(y^*_i > 0) \cdot I(y^*_i \leq 0) (2\pi)^{-1/2} \exp \left\{ -\frac{1}{2} (y^*_i - x'_i\beta)^2 \right\}
\] (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 \( I(y^*_i > 0) \) and the full conditional of \( y^*_i \) is a Normal density with mean \( x'_i\beta \) 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:
\[
y^*_i = x'_i\beta + \varepsilon_i, \quad \varepsilon_i \sim N(0, 1)
\]
\[
y_i = \begin{cases} 1 & \text{if } y^*_i > 0 \\ 0 & \text{if } y^*_i \leq 0 \end{cases}
\]
and with a Normal prior for \( \beta \):
\[
p(\beta) = \frac{|P|^{1/2}}{(2\pi)^{K/2}} \exp \left\{ -\frac{1}{2} (\beta - m)' P (\beta - m) \right\}
\]

the full conditional of \( \beta \) is Normal:
\[
\pi (\beta | \bullet) = \frac{|P|^{1/2}}{(2\pi)^{K/2}} \exp \left\{ -\frac{1}{2} (\beta - \hat{m})' \hat{P} (\beta - \hat{m}) \right\}
\]

where:
\[
\hat{P} = X'X + P \quad \hat{m} = (X'X + P)^{-1} (X'y^* + Pm)
\]

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 \( y_i \):
\[
p(y^*_i | \bullet) = \begin{cases} \frac{1}{(2\pi)^{1/2}} \exp \left\{ -\frac{1}{2} (y^*_i - x'_i\beta)^2 \right\} I(y^*_i > 0) & \text{if } y_i = 1 \\ \frac{1}{(2\pi)^{1/2}} \exp \left\{ -\frac{1}{2} (y^*_i - x'_i\beta)^2 \right\} I(y^*_i \leq 0) & \text{if } y_i = 0 \end{cases}
\]

**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 by 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:

\[ \text{Prob}( \text{union}_i = 1) = \Phi (\beta_1 + \beta_2 \text{hours}_i + \beta_3 \text{married}_i + \beta_4 \text{black}_i + \beta_5 \text{hisp}_i + \beta_6 \text{health}_i) \]

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.

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Median</th>
<th>Sd.dev.</th>
<th>5%</th>
<th>95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>constant</td>
<td>-0.504901</td>
<td>-0.504068</td>
<td>0.0865019</td>
<td>-0.649016</td>
<td>-0.363148</td>
</tr>
<tr>
<td>hours</td>
<td>-0.167877</td>
<td>-0.168039</td>
<td>0.0393891</td>
<td>-0.232094</td>
<td>-0.102869</td>
</tr>
<tr>
<td>married</td>
<td>0.196843</td>
<td>0.197083</td>
<td>0.04372</td>
<td>0.124709</td>
<td>0.268911</td>
</tr>
<tr>
<td>black</td>
<td>0.492032</td>
<td>0.492229</td>
<td>0.062998</td>
<td>0.388421</td>
<td>0.594935</td>
</tr>
<tr>
<td>hisp</td>
<td>0.189786</td>
<td>0.189897</td>
<td>0.0568246</td>
<td>0.0958511</td>
<td>0.283087</td>
</tr>
<tr>
<td>health</td>
<td>-0.458973</td>
<td>-0.45428</td>
<td>0.192687</td>
<td>-0.786547</td>
<td>-0.150931</td>
</tr>
</tbody>
</table>

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 magnitude of the estimates.

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

```plaintext
// 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:

\[
\begin{align*}
    y^*_i &= x'_i \beta + \varepsilon_i, \quad \varepsilon_i \sim \text{Logistic}(0, 1) \\
    y_i &= \begin{cases} 
    1 & \text{if } y^*_i > 0 \\
    0 & \text{if } y^*_i \leq 0 
    \end{cases}
\end{align*}
\]

(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, this function is:

\[
p(y, y^*|X, \beta) = \prod_{i=1}^{N} \left[ \mathbb{1}(y^*_i > 0)^{y_i} \cdot \mathbb{1}(y^*_i \leq 0)^{1-y_i} \times \frac{e^{y^*_i x'_i \beta}}{(1 + e^{y^*_i x'_i \beta})^2} \right]
\]

(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 $x_i'\beta$ and scale parameter equal to one. Placing a multivariate-Normal prior on $\beta$, with mean vector $m$ and precision matrix $P$ and applying Bayes’ theorem leads to the following posterior density:

$$
\begin{align*}
\pi (\beta, y^* | y, X) & \propto p(y, y^* | X, \beta) p(\beta) \\
& = \prod_{i=1}^{N} \frac{1}{2} (y_i^* > 0)^{y_i} \cdot \frac{1}{2} (y_i^* \leq 0)^{1-y_i} \times \frac{e^{y_i^* - x_i'\beta}}{(1 + e^{y_i^* - x_i'\beta})^2} \\
& \quad \times \frac{|P|^{1/2}}{(2\pi)^{K/2}} \exp \left\{ -\frac{1}{2} (\beta - m)' P (\beta - m) \right\} 
\end{align*}
$$

(6.24)

The assumption of a Logistically-distributed error term has severe implications for the full conditional of $\beta$. In particular, after dropping terms that enter the complete-data likelihood multiplicatively and which do not involve $\beta$, this full conditional becomes:

$$
\begin{align*}
\pi (\beta | \bullet) & \propto \prod_{i=1}^{N} \frac{e^{y_i^* - x_i'\beta}}{(1 + e^{y_i^* - x_i'\beta})^2} \times \exp \left\{ -\frac{1}{2} (\beta - m)' P (\beta - m) \right\} 
\end{align*}
$$

(6.25)

and can hardly be simplified any further. The posterior density of $\beta$ 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:

$$
\begin{align*}
y_i^* &= x_i'\beta + \varepsilon_i, \quad \varepsilon_i \sim \text{Logistic} (0, 1) \\
y_i &= \begin{cases} 
1 & \text{if } y_i^* > 0 \\
0 & \text{if } y_i^* \leq 0
\end{cases}
\end{align*}
$$

and with a Normal prior for $\beta$:

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

the full conditional of $\beta$ is:

$$
\begin{align*}
\pi (\beta | \bullet) & \propto \prod_{i=1}^{N} \frac{e^{y_i^* - x_i'\beta}}{(1 + e^{y_i^* - x_i'\beta})^2} \times \exp \left\{ -\frac{1}{2} (\beta - m)' P (\beta - m) \right\}
\end{align*}
$$


**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 by 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:
\[
\text{Prob}(\text{union}_i = 1) = \Lambda (\beta_1 + \beta_2 \text{hours}_i + \beta_3 \text{married}_i + \beta_4 \text{black}_i + \beta_5 \text{hisp}_i + \beta_6 \text{health}_i)
\]
which leads to the Logit model. The results obtained using BayES’ logit() function to estimate this model are presented in the following table.

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Median</th>
<th>Sd.dev.</th>
<th>5%</th>
<th>95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>constant</td>
<td>-0.84562</td>
<td>-0.844747</td>
<td>0.149351</td>
<td>-1.09097</td>
<td>-0.605491</td>
</tr>
<tr>
<td>hours</td>
<td>-0.272184</td>
<td>-0.271297</td>
<td>0.066651</td>
<td>-0.381035</td>
<td>-0.160828</td>
</tr>
<tr>
<td>married</td>
<td>0.334537</td>
<td>0.334365</td>
<td>0.073017</td>
<td>0.216731</td>
<td>0.455367</td>
</tr>
<tr>
<td>black</td>
<td>0.826861</td>
<td>0.828256</td>
<td>0.102698</td>
<td>0.658371</td>
<td>0.992609</td>
</tr>
<tr>
<td>hisp</td>
<td>0.32349</td>
<td>0.322177</td>
<td>0.096794</td>
<td>0.16496</td>
<td>0.485275</td>
</tr>
<tr>
<td>health</td>
<td>-0.847172</td>
<td>-0.843419</td>
<td>0.348166</td>
<td>-1.40942</td>
<td>-0.294031</td>
</tr>
</tbody>
</table>

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.

```r
// 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-choice 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 $x_i'\beta$ from the real line onto the unit interval is monotonically increasing.\(^5\)

This implies that, if $x_i'\beta$ 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:

\[
\frac{\partial \text{Prob}(y_i = 1 | x_i)}{\partial x_{ik}} = f(x_i'\beta) \cdot \beta_k
\]  
\(^{6.26}\)

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:

\[
\frac{\partial \text{Prob}(y_i = 1 | x_i)}{\partial x_{ik}} = \left(2\pi\right)^{-1/2} \exp \left\{ -\frac{(x_i'\beta)^2}{2} \right\} \cdot \beta_k
\]  
\(^{6.27}\)

\(^5F(\cdot)\) is the standard-Normal or standard-Logistic cumulative density function for the Probit and Logit models, respectively.
for the Probit model and:
\[
\frac{\partial \text{Prob}(y_i = 1|x_i)}{\partial x_{ik}} = \left[ \frac{e^{x_i'\beta}}{(1 + e^{x_i'\beta})^2} \right] \cdot \beta_k
\]  
(6.28)
for the Logit model. Due to the first factor in each expression being non-linear in $x_i'\beta$, the marginal effects depend on the point, $x_i$, 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:

\[
\text{Prob}(y_i = 1|x_{i1}) - \text{Prob}(y_i = 1|x_{i0}) = F(x_{i1}'\beta) - F(x_{i0}'\beta)
\]  
(6.29)

where $x_{i0}$ 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. $x_{i1}$ 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 $\beta$s. 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 $\beta$, 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 by a collective bargaining agreement, using a Probit and Logit model, respectively. If the models estimated using the BayES’ \texttt{probit()} and a \texttt{logit()} functions are stored in memory (they are given a left-hand-side value), then the \texttt{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.

<table>
<thead>
<tr>
<th>\text{dProb}(y=1)/dx</th>
<th>\text{Mean}</th>
<th>\text{Median}</th>
<th>\text{Sd.dev.}</th>
<th>\text{5%}</th>
<th>\text{95%}</th>
</tr>
</thead>
<tbody>
<tr>
<td>hours</td>
<td>-0.0521208</td>
<td>-0.052167</td>
<td>0.0122039</td>
<td>-0.0719087</td>
<td>-0.0320055</td>
</tr>
<tr>
<td>*married</td>
<td>0.0615992</td>
<td>0.0616473</td>
<td>0.0137477</td>
<td>0.0389037</td>
<td>0.0844024</td>
</tr>
<tr>
<td>*black</td>
<td>0.170462</td>
<td>0.170389</td>
<td>0.0235062</td>
<td>0.131982</td>
<td>0.209253</td>
</tr>
<tr>
<td>*hisp</td>
<td>0.061734</td>
<td>0.0615518</td>
<td>0.0191363</td>
<td>0.0305068</td>
<td>0.0935144</td>
</tr>
<tr>
<td>*health</td>
<td>-0.114524</td>
<td>-0.117767</td>
<td>0.0396981</td>
<td>-0.174334</td>
<td>-0.0444449</td>
</tr>
</tbody>
</table>

*Marginal effect is calculated for discrete change from 0 to 1.

The following table gives the corresponding marginal effects for the Logit model.

<table>
<thead>
<tr>
<th>\text{dProb}(y=1)/dx</th>
<th>\text{Mean}</th>
<th>\text{Median}</th>
<th>\text{Sd.dev.}</th>
<th>\text{5%}</th>
<th>\text{95%}</th>
</tr>
</thead>
<tbody>
<tr>
<td>hours</td>
<td>-0.0493737</td>
<td>-0.0491959</td>
<td>-0.0121544</td>
<td>-0.069335</td>
<td>-0.0292322</td>
</tr>
<tr>
<td>*married</td>
<td>0.0613089</td>
<td>0.0612926</td>
<td>0.0135126</td>
<td>0.0394384</td>
<td>0.0840309</td>
</tr>
<tr>
<td>*black</td>
<td>0.170462</td>
<td>0.170389</td>
<td>0.0235062</td>
<td>0.131982</td>
<td>0.209253</td>
</tr>
<tr>
<td>*hisp</td>
<td>0.0622516</td>
<td>0.0617431</td>
<td>0.019458</td>
<td>0.0308945</td>
<td>0.0952048</td>
</tr>
<tr>
<td>*health</td>
<td>-0.114524</td>
<td>-0.117767</td>
<td>0.0396981</td>
<td>-0.174334</td>
<td>-0.0444449</td>
</tr>
</tbody>
</table>

The following table gives the corresponding marginal effects for the Logit model.
CHAPTER 6. MODELS FOR BINARY CHOICE

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 0.049 units. Similarly, a married person has higher probability of having his wage set by a collective agreement relative to a single person by 0.062/0.061 units, according to each model.

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

```plaintext
// 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-Choice Models for Panel Data

Extending binary-choice 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-choice 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:

$$\text{Prob}(y_{it} = 1|x_{it}, \alpha_i) = F(\alpha_i + x_{it}'\beta)$$

(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-choice models, estimation of random-effects models is easier to handle using the latent-variable representation. In this formulation the Probit model becomes:

$$y^*_it = \alpha_i + x_{it}'\beta + \varepsilon_{it}, \quad \varepsilon_{it} \sim \text{N}(0, 1), \quad \alpha_i \sim \text{N}(0, 1/\omega)$$

$$y_{it} = \begin{cases} 1 & \text{if } y^*_it > 0 \\ 0 & \text{if } y^*_it \leq 0 \end{cases}$$

where the assumption that each group effect follows a Normal distribution is added. The only difference in a random-effects Logit model is that $\varepsilon_{it}$ 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 $\beta$ 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 extended such that the means of the corresponding densities are $\alpha_i + x_{it}'\beta$ instead of $x_{it}'\beta$. 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 $\alpha_i$s, as they are presented in Theorem 5.1, while no changes are required for the full conditional of $\omega$ from the linear model. For the Logit model, however, the full conditionals of the $\alpha_i$s no longer belong to a parametric family from which random numbers can be drawn directly:

$$\pi(\alpha_i|\bullet) \propto \prod_{t=1}^{T} \left[ \frac{e^{y^*_it - \alpha_i - x_{it}'\beta}}{1 + e^{y^*_it - \alpha_i - x_{it}'\beta}} \right] \times \exp \left\{ -\frac{\omega \alpha_i^2}{2} \right\}$$

(6.31)
6.5. BINARY-CHOICE MODELS FOR PANEL DATA

As always, Metropolis-Hastings updates can be used to sample from the full conditional of each $\alpha_i$.

Although no real conceptual differences appear when moving from simple binary-choice 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:

$$\frac{\partial \text{Prob}(y_{it} = 1|x_{it}, \alpha_i)}{\partial x_{it,k}} = f(\alpha_i + x_{it}'\beta) \cdot \beta_k$$

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 $\alpha_i$ to zero, as this is both its most likely and its expected value according to the assumption $\alpha_i \sim N(0, \frac{1}{\omega})$. This leads to conditional marginal effects.
2. integrate uncertainty with respect to the values of $\alpha_i$ from the marginal effects. This leads to averaged marginal effects or Average Partial Effects (APE), as they are better known in the literature.

The method of calculating the averaged marginal effects is easier to describe using the latent-variable representation of the Probit model and much easier to perform in the Probit model. The latent-variable representation of the Probit model implies:

$$y_{it} = \begin{cases} 1 & \text{if } \alpha_i + \varepsilon_{it} > -x_{it}'\beta \\ 0 & \text{if } \alpha_i + \varepsilon_{it} \leq -x_{it}'\beta \end{cases}$$

(6.33)

Notice that because $\alpha_i$ 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 $\xi = \frac{1}{\omega^2}$. Therefore:

$$\text{Prob}(y_{it} = 1|x_{it}) = \text{Prob}(w_{it} > -x_{it}'\beta | x_{it}) = \text{Prob}(z_{it} > -\sqrt{\xi} \cdot x_{it}'\beta | x_{it})$$

(6.34)

$$= \Phi(\sqrt{\xi} \cdot x_{it}'\beta)$$

where $z_{it} \sim N(0, 1)$. Given this result, the averaged marginal effect for the $k$-th independent variable, evaluated at point $x_{it}$ is:

$$\frac{\partial \text{Prob}(y_{it} = 1)}{\partial x_{it,k}} = (2\pi)^{-1/2} \exp \left\{ -\frac{(\sqrt{\xi} \cdot x_{it}'\beta)^2}{2} \right\} \cdot \sqrt{\xi} \cdot \beta_k$$

(6.35)

This procedure takes care of uncertainty with respect to the value of $\alpha_i$. 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, $(\beta^{(g)}, \omega^{(g)})$, 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 $\varepsilon_{it}$ follows a standard-Logistic distribution, $w_{it} = \text{convolution of a Normally and a Logistically distributed random variable}$ and its cumulative density function has a very complex form (Gupta & Nadarajah, 2008). Instead of approximating this complex function, $\alpha_i$ 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:

$$\frac{\partial \text{Prob}(y_{it} = 1|x_{it})}{\partial x_{it,k}} = \beta_k \int_{-\infty}^{\infty} \frac{\partial \text{Prob}(y_{it} = 1|x_{it}, \alpha)}{\partial x_{it,k}} p(\alpha) \, d\alpha$$

(6.36)

were the fact that $\alpha_i$ 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:

$$\text{Prob}(\text{union}_{it} = 1) = F(\alpha_i + \beta_1 \text{hours}_{it} + \beta_2 \text{married}_{it} + \beta_3 \text{black}_{it} + \beta_4 \text{hisp}_{it} + \beta_5 \text{health}_{it})$$

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 $\omega$ being relatively small, indicates that individual-specific unobserved heterogeneity has a large impact on the results. As in the simple binary-choice models, we can only interpret the signs of the parameter estimates, but not their magnitude.

<table>
<thead>
<tr>
<th>Mean</th>
<th>Median</th>
<th>Sd.dev.</th>
<th>5%</th>
<th>95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>constant</td>
<td>-1.26185</td>
<td>-1.26038</td>
<td>0.179899</td>
<td>-1.56109</td>
</tr>
<tr>
<td>hours</td>
<td>-0.174721</td>
<td>-0.174382</td>
<td>0.0661689</td>
<td>-0.284149</td>
</tr>
<tr>
<td>married</td>
<td>0.147338</td>
<td>0.147481</td>
<td>0.0830975</td>
<td>0.0102956</td>
</tr>
<tr>
<td>black</td>
<td>0.981674</td>
<td>0.981347</td>
<td>0.265875</td>
<td>0.548038</td>
</tr>
<tr>
<td>hisp</td>
<td>0.495902</td>
<td>0.494572</td>
<td>0.232273</td>
<td>0.117523</td>
</tr>
<tr>
<td>health</td>
<td>-0.426287</td>
<td>-0.422745</td>
<td>0.275598</td>
<td>-0.88255</td>
</tr>
<tr>
<td>$\omega$</td>
<td>0.34379</td>
<td>0.341892</td>
<td>0.0398265</td>
<td>0.281654</td>
</tr>
<tr>
<td>$\sigma^2_{\alpha}$</td>
<td>1.71411</td>
<td>1.71024</td>
<td>0.0995942</td>
<td>1.55669</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Mean</th>
<th>Median</th>
<th>Sd.dev.</th>
<th>5%</th>
<th>95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>constant</td>
<td>-2.23299</td>
<td>-2.2333</td>
<td>0.314293</td>
<td>-2.75218</td>
</tr>
<tr>
<td>hours</td>
<td>-0.32676</td>
<td>-0.326845</td>
<td>0.117889</td>
<td>-0.520654</td>
</tr>
<tr>
<td>married</td>
<td>0.285391</td>
<td>0.285187</td>
<td>0.146128</td>
<td>0.0426119</td>
</tr>
<tr>
<td>black</td>
<td>1.78431</td>
<td>1.79204</td>
<td>0.437257</td>
<td>1.05695</td>
</tr>
<tr>
<td>hisp</td>
<td>0.87266</td>
<td>0.875463</td>
<td>0.420684</td>
<td>0.173618</td>
</tr>
<tr>
<td>health</td>
<td>-0.78715</td>
<td>-0.763513</td>
<td>0.495039</td>
<td>-1.0183024</td>
</tr>
<tr>
<td>$\omega$</td>
<td>0.108519</td>
<td>0.10798</td>
<td>0.0130879</td>
<td>0.0879224</td>
</tr>
<tr>
<td>$\sigma^2_{\alpha}$</td>
<td>3.05228</td>
<td>3.04319</td>
<td>0.185364</td>
<td>2.76494</td>
</tr>
</tbody>
</table>

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-choice models without random effects, the posterior moments of the marginal effects from the random-effects Probit and Logit models are very similar.

<table>
<thead>
<tr>
<th>dProb(y=1)/dx</th>
<th>Mean</th>
<th>Median</th>
<th>Sd.dev.</th>
<th>5%</th>
<th>95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>hours</td>
<td>-0.0274208</td>
<td>-0.0273116</td>
<td>0.0104285</td>
<td>-0.0447663</td>
<td>-0.010412</td>
</tr>
<tr>
<td>married</td>
<td>0.0232345</td>
<td>0.0231955</td>
<td>0.0131784</td>
<td>0.00160827</td>
<td>0.0449508</td>
</tr>
<tr>
<td>black</td>
<td>0.172209</td>
<td>0.171637</td>
<td>0.0494562</td>
<td>0.0917155</td>
<td>0.254541</td>
</tr>
<tr>
<td>hisp</td>
<td>0.0829335</td>
<td>0.082032</td>
<td>0.040045</td>
<td>0.0187375</td>
<td>0.150201</td>
</tr>
<tr>
<td>health</td>
<td>-0.0595993</td>
<td>-0.0611832</td>
<td>0.036586</td>
<td>-0.117273</td>
<td>0.00287371</td>
</tr>
</tbody>
</table>

**Marginal effect is calculated for discrete change from 0 to 1.**

<table>
<thead>
<tr>
<th>dProb(y=1)/dx</th>
<th>Mean</th>
<th>Median</th>
<th>Sd.dev.</th>
<th>5%</th>
<th>95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>hours</td>
<td>-0.0290922</td>
<td>-0.0290373</td>
<td>0.0105196</td>
<td>-0.0463726</td>
<td>-0.0119123</td>
</tr>
<tr>
<td>married</td>
<td>0.025339</td>
<td>0.0253305</td>
<td>0.0129942</td>
<td>0.00379623</td>
<td>0.0467248</td>
</tr>
<tr>
<td>black</td>
<td>0.173123</td>
<td>0.173182</td>
<td>0.0454789</td>
<td>0.0990087</td>
<td>0.248134</td>
</tr>
<tr>
<td>hisp</td>
<td>0.0809686</td>
<td>0.0807606</td>
<td>0.0399958</td>
<td>0.0155531</td>
<td>0.147917</td>
</tr>
<tr>
<td>health</td>
<td>-0.0622204</td>
<td>-0.0630041</td>
<td>0.0362704</td>
<td>-0.121379</td>
<td>-0.00161252</td>
</tr>
</tbody>
</table>

**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.
6.6 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.

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.
References


Index

Almost Ideal Demand System, 58
almost sure convergence, 16
autocorrelation time, 19
Average Partial Effects, 103
averaged marginal effects, 103
Bayes factor, 7, 13, 46
binary response, see binary-choice model
binary-choice model, 91
index function, 92
borrowing of strength, 79
burn-in, 20
Central Limit Theorem, 17
Chamberlain’s approach, 87
Cobb-Douglas, 39
complete conditional, 25
complete-data likelihood, 64
composition sampling, 27
conditional marginal effects, 103
conditional probability, 3
confidence interval, 2, 10
confidence level, 2
correlated random effects, 87
credible interval, 10, 14
shortest, 10
data augmentation, 63
data-generating process, 4, 5, 32
dependent variable, 32
diffuse prior, 7
disturbance, 32
econometrics, 1
effect size, 19
elasticity, 39
error term, 32
extension parameter, see data augmentation
Expectation-Maximization algorithm, 63
fixed-effects, 79
flat prior, 7
flexible functional form, 38
full conditional, 22
generalized linear model, 95
Gibbs algorithm, 25
collapsed, 27
Gibbs sampler, see Gibbs algorithm
group effect, 78
heteroskedastic error, 65
Hicksian demand function, 59
hierarchical model, 66, 79, 88
homoskedastic error, 66
hyperparameters, 6, 7
incomplete-data likelihood, 64
independent variable, 32
indicator function, 17, 49
inefficiency factor, 19
interaction term, 38
invariance principal, 7
inverse probability transform, 27
Laplace approximation, 46
latent data, 46, 63
likelihood function, 4, 5, 33
linear regression model, 31
link function, 95
Logit model, 93
marginal data augmentation, see data augmentation
marginal effect, 33
marginal likelihood, 4, 13
Markov-chain Monte Carlo, 16, 19
Metropolis-Hastings, 20
multiple-block, 22
random-walk, 20
ratio, 20
within Gibbs, 25
mixing, 29
Monte Carlo, 16
standard error, 17
Mundlak’s approach, 87
noninformative prior, 7
objective prior, 6
observed data, 63
observed-data likelihood, 64
panel data, 77
balanced, 78
group, 77
unbalanced, 78
parameter-expanded data augmentation, see data augmentation
INDEX

pooled model, 78
post-estimation inference, 42
posterior density, 4, 9
posterior model probability, 13
posterior odds ratio, 13, 46
posterior predictive density, 14, 48
precision matrix, 34
precision parameter, 23, 32
prior density, 4, 6
prior model probability, 12
prior odds ratio, 13
probability
   objective Bayesian, 2
   subjective Bayesian, 2
Probit model, 93
proposal density, 20

quadratic term, 38

random utility, 94
random-coefficients model, 79
random-effects model, 79
reference prior, 7
rejection sampling, 27
relative numerical efficiency, 19

Seemingly Unrelated Regressions, 51
sensitivity analysis, 6
Shephard’s lemma, 59
stochastic frontier, 69
Strong Law of Large Numbers, 16
subjective prior, 6

target distribution, 19
Taylor series expansion, 38
technical-efficiency, 70
thinning parameter, 28
translog specification, 40, 59
type I extreme-value distribution, 95

unobserved heterogeneity, 78

vague prior, 6, 7

Wishart distribution, 54
working parameter, see data augmentation

Young’s theorem, 38, 58