CONTENTS

LIST OF FIGURES

CHAPTER 1

BINARY AND ORDINAL DATA ANALYSIS IN ECONOMICS: MODELING AND ESTIMATION

Ivan Jeliazkov and Mohammad Arshad Rahman

Department of Economics, University of California, Irvine

1.1 INTRODUCTION

This chapter is concerned with the analysis of statistical models for binary and ordinal outcomes. Binary data arise when a particular response variable of interest y_i can take only two values, i.e. $y_i \in \{0, 1\}$, where the index $i = 1, \ldots, n$ refers to units in the sample such as individuals, families, firms, and so on. Such dichotomous outcomes are widespread in the social and natural sciences. For example, to understand socio-economic processes, economists often need to analyze individuals' binary decisions such as whether to make a particular purchase, participate in the labor force, obtain a college degree, see a doctor, migrate to a different country, or vote in an election. By convention, $y_i = 1$ typically indicates the occurrence of the event of interest, whereas the occurrence of its complement is denoted by $y_i = 0$.

Mathematical Modeling with Multidisciplinary Applications.
By Xin-She Yang

We also examine modeling and estimation issues related to another type of data, called ordinal data, where y_i can take one of J ordered values, $j = 1, \ldots, J$. The defining feature of ordinal data is that even though the outcomes are monotone, the scale on which they are measured is not assumed to be cardinal and differences between categories are not directly comparable. For instance, in quantifying survey responses on consumer satisfaction, 1 could be assigned to "very unhappy", 2 to "not too happy", 3 to "happy", and 4 to "very happy", but even though the scale tells us that 4 implies more happiness than 2, this does not mean that 4 implies twice as much happiness as 2, or that the difference in happiness between 1 and 3 is the same as that between 2 and 4. Even though ordinal data models were developed primarily for the analysis of data on rankings, they offer a flexible modeling framework that can also be very useful in the analysis of certain types of count data.

In this chapter we pursue several goals. We briefly review relevant results from the theory of choice which formalize the link between economic theory and empirical practice in binary and ordinal data analysis. We then turn our attention to the topic of estimation and highlight the identification issues that arise in binary and ordinal models. We review both classical and Bayesian approaches to estimation, and introduce a new simulation-based estimation algorithm for logit models based on data augmentation. Even though the theoretical foundations for this algorithm have been available for decades, the approach has remained unexploited until now. Our estimation approach removes important obstacles that have hindered extensions of logistic regression to multivariate and hierarchical model settings.

Another topic that we examine here is covariate effect estimation, which allows us to evaluate the impact of particular covariates on the outcome of interest and gives concrete practical meaning to the parameters of the model. The techniques are illustrated in two applications in economics including women's labor force participation and educational attainment. The methods discussed here form a foundation for studying other more complex recent developments in the literature such as extensions to panel data, multivariate and multinomial outcomes, dynamics, mixed models, and copula models.

1.2 THEORETICAL FOUNDATIONS

There exist a number of statistical models for binary and ordinal data, but they share a common foundation in which the observed discrete outcomes can be represented by the crossing of particular thresholds by an underlying continuous latent variable. This latent variable threshold-crossing formulation can in turn be related to the theory of choice in economics to form an elegant link between behavioral and statistical models. Because of this link, models for discrete data in econometrics are also frequently referred to as discrete choice models. The derivations are important because the latent variable representation turns out to be particularly useful not only in theory, but

also in estimation. It also helps clarify the relationship between empirical models based on different distributional assumptions and provides a basis for the calculation of important quantities in economics, such as consumer surplus or willingness to pay. Note, however, that the econometric techniques are fully general and can be used to represent various phenomena that do not necessarily entail references to utility or choice (e.g., weather patterns, accident probabilities, volcanic eruptions, etc.).

In order for the decision problem to be well-posed, the set of available alternatives, or *choice set*, must be defined so that alternatives are (i) *mutually exclusive*, i.e. they represent distinct non-overlapping outcomes, and (ii) *exhaustive*, so that all possible outcomes are fully accounted for. These criteria are easily satisfied in the context of binary and ordinal data where the dependent variable y_i is simply an indicator variable for the occurrence of a particular event. One should keep in mind, that while in some contexts the dichotomy can be a natural feature of the data (e.g. medical tests, welfare program participation, home ownership, criminal recidivism, etc.), in other cases it can be introduced subjectively by the researcher to study a particular socio-economic phenomenon. For example, in studying market participation, a researcher may set $y_i = 1$ for producers whose sales in a given market are positive and $y_i = 0$ for all others. At first glance this discretization may seem unreasonable as it leads to loss of information on magnitudes (since both small and large sellers are treated alike). However, economic theory suggests that the presence of fixed costs leads firms to treat market entry and exit differently than the problem of how much to produce conditionally on being in the market. For this reason, the delineation of firms into market participants (regardless of sales volume) and non-participants (those with zero sales) can be an important first step in studying market outcomes. In the case of ordinal data, the outcomes will easily satisfy the first criterion if the dependent variable $y_i \in \{1, \ldots, J\}$ is defined as the sum of indicator variables over a particular monotone set of events. The second criterion, on the other hand, can either be satisfied naturally if outcomes are measured on a finite scale (as in surveys, or bond and stock ratings) or may have to be imposed by specifying a composite category that captures all possible outcomes beyond a certain value (as is common in the analysis of count data). Therefore, the nature of the choice set in binary and ordinal data models is in sharp contrast with standard models for continuous dependent variables, such as consumption or growth.

1.2.1 Binary Outcomes

The roots of the random utility framework that underlies discrete choice models in econometrics can be traced back to the pioneering work of [15], [16], and [17]. A detailed recent review with applications to problems in modern econometrics is given in [25]. The basic setup involves utility maximizing decision makers, who choose among competing alternatives associated with

certain levels of utility. The theory is quite general and can handle a variety of possible choices; the same ideas apply in our binary data context where there are only two possible alternatives. Specifically, individual i has two levels of utility, U_{i1} and U_{i0}, that are associated with $y_i = 1$ or $y_i = 0$, respectively. The utility maximizing agent then selects the option providing the higher of the two utilities:

$$y_i = \begin{cases} 1 & \text{if } U_{i1} > U_{i0}, \\ 0 & \text{otherwise.} \end{cases}$$

The utilities U_{i1} and U_{i0} are known to the decision maker but are unknown to the researcher, who can only observe a vector x_i of characteristics of the decision maker that can be related to utility through $U_{ij} = x_i'\beta_j + \varepsilon_{ij}$ for $j = 0, 1$. The term $x_i'\beta_j$ is sometimes referred to as *representative utility*, whereas ε_{ij} captures unobserved factors that affect utility but are not included in $x_i'\beta_j$. In essence, $x_i'\beta_j$ is a systematic component and $\varepsilon_{i,j}$ is a stochastic (from the point of view of the researcher) part of individual utility.

This theoretical setup will be used to make probabilistic statements about the observed choices y_i conditionally on x_i. In the remainder of this chapter, conditioning of one variable on another will be denoted by a vertical bar ' | ', for example, $\Pr(A|B)$ will represent the conditional probability of A given B. Similarly, if s is a continuous random variable $f(s|t)$ will be used to denote the conditional density of s given t. In some contexts, when it is important to make clear the link between a random variable and its density, we may use notation such as $s|t$ to emphasize that we are interested in a random variable with density $f(s|t)$, i.e. $s|t \sim f(s|t)$, as opposed to a random variable s with density $f(s)$, i.e. $s \sim f(s)$.

To develop a model for the observed choices, note that given x_i and the parameters β_0 and β_1, the conditional probability of observing $y_i = 1$ can be expressed as an exceedance probability between the two utility levels

$$\begin{aligned} \Pr(y_i = 1|x_i, \beta_0, \beta_1) &= \Pr(U_{i1} > U_{i0}) \\ &= \Pr(x_i'\beta_1 + \varepsilon_{i1} > x_i'\beta_0 + \varepsilon_{i0}) \\ &= \Pr[(\varepsilon_{i0} - \varepsilon_{i1}) < x_i'(\beta_1 - \beta_0)]. \end{aligned} \tag{1.1}$$

The model is operationalized by specifying a density for the random variable $(\varepsilon_{i0} - \varepsilon_{i1})$, but before we consider specific cases, we need to address the important topic of parameter identification. From equation (1.1) we see that the choice probability depends only on the differences in utilities between alternatives, not on the absolute level of utilities. Specifically, because the probability in (1.1) depends on the difference $(\beta_1 - \beta_0)$, it will not change if we add an arbitrary constant c to both β_0 and β_1, i.e., $x_i'(\beta_1 - \beta_0) = x_i'(\tilde{\beta}_1 - \tilde{\beta}_0)$, where $\tilde{\beta}_1 = \beta_1 + c$ and $\tilde{\beta}_0 = \beta_0 + c$. Second, the scale of utility is not identified because the probability is unchanged if both sides of (1.1) are multiplied by an arbitrary constant $c > 0$, i.e., $\Pr[(\varepsilon_{i0} - \varepsilon_{i1}) < x_i'(\beta_1 - \beta_0)] = \Pr[c(\varepsilon_{i0} - \varepsilon_{i1}) < cx_i'(\beta_1 - \beta_0)]$.

To deal with these problems, we need to fix both the location and scale of utility. The location is fixed by measuring utility relative to that of the baseline category, U_{i0}. In other words, we work with the differenced form

$$z_i = x_i'\beta + \nu_i, \qquad i = 1, \ldots, n, \tag{1.2}$$

where $z_i = U_{i1} - U_{i0}$, $\beta = \beta_1 - \beta_0$, and $\nu_i = \varepsilon_{i1} - \varepsilon_{i0}$. As a result, the relationship between the observed outcome y_i and the latent z_i is given by

$$y_i = \begin{cases} 1 & \text{if } z_i > 0, \\ 0 & \text{otherwise,} \end{cases} \tag{1.3}$$

which can alternatively be written as $y_i = 1\{z_i > 0\}$ using the indicator function $1\{\cdot\}$ that takes the value 1 if its argument is true and 0 otherwise. The scale of utility is normalized by fixing the variance of ν_i and treating it as given rather than as a parameter to be estimated; doing so is only a normalization that does not restrict the underlying flexibility of the model. (One should keep in mind that the value at which the variance of ν_i is fixed will be model specific.) In the following examples, we review the three most common model specifications used in empirical analysis – probit, logit, and t-link or robit.

EXAMPLE 1.1

The probit model is obtained by assuming that the errors in (1.2) follow a standard normal distribution $\nu_i \sim N(0, 1)$ with probability density function (pdf) and cumulative distribution function (cdf) given by

$$\phi(\nu_i) = (2\pi)^{-1/2} e^{-\nu_i^2/2} \quad \text{and} \quad \Phi(\nu_i) = \int_{-\infty}^{\nu_i} \phi(t)dt.$$

Note that the pdf $\phi(\cdot)$ is symmetric and the variance of ν_i is fixed at 1 as a normalization. In addition, even though the expression for the Gaussian cdf $\Phi(\cdot)$ does not have a closed form solution, it is readily available in most statistical software packages.

EXAMPLE 1.2

The logit model is obtained by assuming that the errors in (1.2) follow a logistic distribution whose cdf $F_L(\cdot)$ and pdf $f_L(\cdot)$ are explicitly available (see Exercise 1.1 for a derivation of $f_L(\cdot)$ from $F_L(\cdot)$):

$$F_L(\nu_i) = \left(1 + e^{-\nu_i}\right)^{-1} \quad \text{and} \quad f_L(\nu_i) = F_L(\nu_i)[1 - F_L(\nu_i)].$$

The logistic distribution is symmetric with mean 0, variance $\pi^2/3$, and heavier tails than the normal distribution. The tail mass makes it more

likely to observe "non-conforming" behavior such as choosing $y_i = 0$ for large positive $x_i'\beta$ or $y_i = 1$ for large negative $x_i'\beta$. In Exercise 1.2, we derive another well known result (see [16] and [18]) that the logit choice probabilities are obtained if the errors ε_{i0} and ε_{i1} in (1.1) follow an extreme value type I distribution.

■ **EXAMPLE 1.3**

The t-link or "robit" model is obtained by assuming that the errors in (1.2) follow a standard Student's t distribution with τ degrees of freedom. The distribution is symmetric around 0, has variance $\tau/(\tau-2)$ for $\tau > 2$, and its pdf $f_{T_\tau}(\cdot)$ and cdf $F_{T_\tau}(\cdot)$ are given by

$$f_{T_\tau}(\nu_i) = \frac{\Gamma(\frac{\tau+1}{2})}{\Gamma(\frac{\tau}{2})\sqrt{\tau\pi}}\left(1+\frac{\nu_i^2}{\tau}\right)^{-\frac{\tau+1}{2}} \quad \text{and} \quad F_{T_\tau}(\nu_i) = \int_0^{\nu_i} f_{T_\tau}(s)ds,$$

where $\Gamma(s) = \int_0^\infty t^{s-1}e^{-t}dt$ denotes the gamma function (which equals $(s-1)!$ for positive integer values of s). Note that the variance of the t distribution is larger than in the probit case but approaches 1 for $\tau \to \infty$. Also, the cdf $F_{T_\tau}(\cdot)$ does not have a closed form solution, but is readily available in most statistical software packages.

An appealing feature of the t-link model is its flexibility: low values of τ produce heavier tails than the logistic distribution, setting $\tau \approx 8$ approximates the logit model, and as $\tau \to \infty$, the t distribution approximates the standard normal. Figure 1.1 shows the log-densities for the standard normal, scaled logistic and scaled t with 4 degrees of freedom (the scaling is done so that all three variances are 1). Because the t-link offers a modeling approach that is robust to variations in the tail behavior of the latent z_i, it has also been referred to by the portmanteau word "robit" ("robust" + the suffix "-it" to resemble probit and logit).

Given the three specifications we have just considered, we can now obtain the outcome probabilities $\Pr(y_i = 1|\beta)$ and $\Pr(y_i = 0|\beta) = 1 - \Pr(y_i = 1|\beta)$ (we suppress the dependence of these probabilities on x_i for notational convenience). In particular, from (1.2) and (1.3) and under the assumption that the density of ν_i is symmetric, we have that

$$\begin{aligned}\Pr(y_i = 1|\beta) &= \Pr(z_i > 0)\\ &= \Pr(x_i'\beta + \nu_i > 0)\\ &= 1 - \Pr(\nu_i < -x_i'\beta)\\ &= 1 - [1 - \Pr(\nu_i < x_i'\beta)]\\ &= \Pr(\nu_i < x_i'\beta) = F(x_i'\beta),\end{aligned}$$

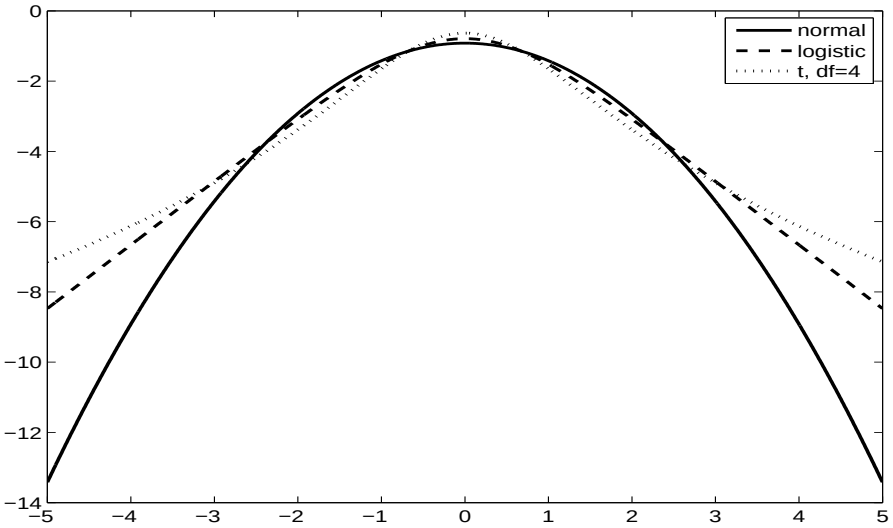

Figure 1.1 Log-densities for the standard normal, scaled logistic and Student's t with 4 degrees of freedom.

where $F(\cdot)$ is the assumed cdf of ν_i – as before, $F(\cdot) = \Phi(\cdot)$ produces the probit model, $F(\cdot) = F_L(\cdot)$ leads to logit, and $F(\cdot) = F_{T_\tau}(\cdot)$ gives the t-link model. Symmetry is used in obtaining the second to last line, and while all models considered here involve symmetric distributions, readers are cautioned to be careful in general.

1.2.2 Ordinal Outcomes

We now turn attention to models for ordinal data, where the alternatives are inherently ordered or ranked. Common applications that involve ordered outcomes include sentiment or opinion surveys, quality tests, health assessment studies, the level of employment (unemployed, part-time, full-time), the level and usage of insurance, and others.

Similarly to the models studied in Section 1.2.1, ordinal data models can be motivated by an underlying latent variable threshold-crossing framework. In particular, as in (1.2) we assume that a continuous latent random variable z_i depends on a k-vector of covariates x_i through the relationship $z_i = x_i'\beta + \nu_i$, $i = 1, \ldots, n$, but with the difference that the observed outcomes $y_i \in \{1, \ldots, J\}$ arise according to

$$y_i = j \quad \text{if} \quad \gamma_{j-1} < z_i \leq \gamma_j, \tag{1.4}$$

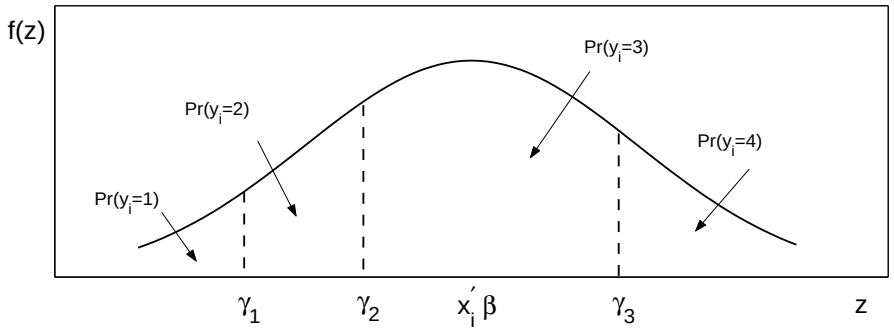

Figure 1.2 Outcome probabilities in an ordinal data model.

where $-\infty = \gamma_0 < \gamma_1 < ... < \gamma_{J-1} < \gamma_J = \infty$ are cutpoint parameters that determine the discretization of the data into J ordered categories. An alternative way of writing (1.4) is to let $y_i = \sum_{j=1}^{J} 1\{z_i > \gamma_{j-1}\}$. Given this representation and a particular cdf $F(\nu_i)$, the probability of observing $y_i = j$, conditional on β and $\gamma = (\gamma_1, \ldots, \gamma_{J-1})'$, is given by

$$\Pr(y_i = j|\beta, \gamma) = F(\gamma_j - x_i'\beta) - F(\gamma_{j-1} - x_i'\beta). \tag{1.5}$$

Figure 1.2 depicts the probabilities of y_i falling in category j as determined by (1.5) in a four-category setting. As before, various choices of the cdf $F(\cdot)$ are possible—e.g. $F(\cdot) = \Phi(\cdot)$, $F(\cdot) = F_L(\cdot)$, $F(\cdot) = F_{T_\tau}(\cdot)$, and so on—but the ordinal probit model is one of the most practical because it is tractable in univariate cases and can be easily generalized to flexible multivariate and hierarchical settings. In contrast, the logistic distribution can not handle correlations in multivariate settings.

As with models for binary data, we require both location and scale restrictions in order to identify the parameters. To see the need for doing so, note that the probabilities in (1.5) are invariant to shifting and rescaling the parameters by some arbitrary constants c and $d > 0$ because

$$F(\gamma_j - x_i'\beta) = F(\gamma_j + c - (x_i'\beta + c))$$

and

$$F(\gamma_j - x_i'\beta) = F\left(\frac{\gamma_j d - x_i'\beta d}{d}\right),$$

which can be applied to both terms in (1.5) without affecting $\Pr(y_i = j|\beta, \gamma)$. The first identification problem is easily corrected by fixing a cutpoint – in particular, letting $\gamma_1 = 0$ removes the possibility for shifting the distribution without changing the probability of observing y_i. As in binary data models, we resolve the possibility for rescaling (our second identification problem) by fixing the variance of ν_i. The variance equals 1 in the probit case, $\pi^2/3$ in the logit case, and $\tau/(\tau - 2)$ in the t-link model.

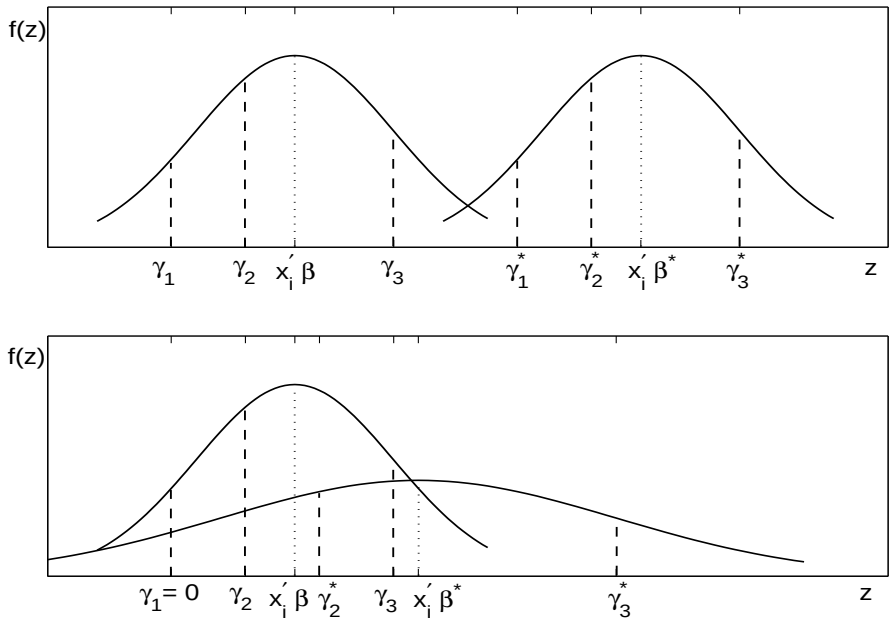

Figure 1.3 Parameter identification in ordinal data models.

Figure 1.3 illustrates these identification considerations. The first panel in the figure illustrates that shifting the density and all cutpoints leaves the probability unaffected; the second panel shows that even if one sets $\gamma_1 = 0$, in the absence of a scale restriction, one can simultaneously rescale $F(\cdot)$, the mean, and the remaining cutpoints without affecting $\Pr(y_i = j|\beta, \gamma)$.

In addition to fixing one cutpoint and the variance of ν_i, there are other possible ways to achieve parameter identification. For example, as an alternative to letting $\gamma_1 = 0$, it is possible to identify the parameters by dropping the intercept term from $x_i'\beta$. Moreover, instead of fixing the variance of ν_i, one can impose a scale restriction by fixing two cutpoints (e.g., $\gamma_1 = 0$ and $\gamma_{J-1} = 1$). The presence and effectiveness of these alternative approaches has been examined in [13] and the references therein, however, these alternatives will not be examined here because they are primarily of interest in multivariate models.

1.3 ESTIMATION

This section reviews both classical and Bayesian methods for estimating the models considered in Section 1.2. Classical estimation in this class of models typically employs the method of maximum likelihood, which requires numerical optimization of the log-likelihood function. Bayesian estimates, on the other hand, are generally obtained by Markov chain Monte Carlo (MCMC)

simulation methods such as Gibbs sampling or the Metropolis-Hastings algorithm.

In addition to reviewing existing estimation methods, this chapter also introduces a new estimation algorithm for logit models that has been overlooked in the literature. The method not only supplements our toolkit for dealing with logistic regression, but also lays a foundation for estimating important extensions of the logit model to multivariate and hierarchical settings.

1.3.1 Maximum Likelihood Estimation

Consider a set of observations $y = (y_1, \ldots, y_n)'$ that comes from some statistical model with sampling density $f(y|\theta)$ written in terms of a parameter vector θ. Because $f(y|\theta)$ provides a mathematical description of the probabilistic phenomenon that generates the observed data sample y given θ, it is called the data generating process. Note that the data generating process is a function of the data conditionally on the parameters, and indeed we can think of it as the mathematical model by which, given θ, nature generates y. In practice, empirical researchers see the sample y generated from $f(y|\theta)$, but do not know the value of θ. When $f(y|\theta)$ is viewed as a function of the parameter vector θ given the sample y, it is called the likelihood function. Although the two functions refer to the same object, $f(y|\theta)$, they emphasize (and take as arguments) its two different components. A thorough review of likelihood inference can be found in standard statistics and econometrics references such as [11].

The maximum likelihood estimator (or MLE) is defined as the value of θ that maximizes the log-likelihood function

$$\hat{\theta}_{MLE} = \arg\max_{\theta} \ln f(y|\theta), \tag{1.6}$$

or heuristically, it is the value of θ that makes the observed sample y as "likely" as possible within the confines of the assumed data generating process. Note that because the logarithmic transformation is monotone, the value $\hat{\theta}_{MLE}$ that maximizes $\ln f(y|\theta)$ also maximizes $f(y|\theta)$, however, it is common to work with $\ln f(y|\theta)$ because it is more stable and easier to evaluate than $f(y|\theta)$, and also because the most important statistical properties of $\hat{\theta}_{MLE}$ are associated with features of $\ln f(y|\theta)$. Specifically, it is known that under mild regularity conditions, the maximum likelihood estimator $\hat{\theta}_{MLE}$ defined in (1.6) is consistent and asymptotically normally distributed. Consistency means that as the sample size $n \to \infty$, the probability limit (or plim) of $\hat{\theta}_{MLE}$ is the true value θ_0, i.e., $\operatorname{plim} \hat{\theta}_{\text{MLE}} = \theta_0$. Asymptotic normality means that in large samples, as $n \to \infty$,

$$\hat{\theta}_{MLE} \sim N\left(\theta_0, V^{-1}\right),$$

where V is the Fisher information defined as the negative of the expected value of the second derivative (or Hessian) matrix of the log-likelihood

$$V = -E\left[\frac{\partial^2 \ln f(y|\theta)}{\partial\theta\partial\theta'}\right]$$

evaluated at θ_0 and the expectation is taken with respect to $f(y|\theta_0)$. Because it is typically impossible to evaluate this expectation, it is common to approximate V by the observed Hessian

$$V = -\frac{\partial^2 \ln f(y|\theta)}{\partial\theta\partial\theta'},$$

which is evaluated at the maximum likelihood value $\theta = \hat{\theta}_{MLE}$. The standard errors of the individual elements of $\hat{\theta}_{MLE}$ are given by the square root of the diagonal entries of V^{-1}, and those can be used in testing and constructing confidence intervals. Next, we consider the likelihood functions for the models studied in Section 1.2

For the binary data models that we considered in Section 1.2.1, the likelihood function can be written as

$$\begin{aligned} f(y|\beta) &= \Pr(y_1, y_2, \ldots, y_n|\beta) \\ &= \prod_{i=1}^{n} \Pr(y_i|\beta) \\ &= \left\{\prod_{i:y_i=1} F(x_i'\beta)\right\}\left\{\prod_{i:y_i=0} [1 - F(x_i'\beta)]\right\} \\ &= \prod_{i=1}^{n} [F(x_i'\beta)]^{y_i} [1 - F(x_i'\beta)]^{(1-y_i)}, \end{aligned} \tag{1.7}$$

where the second line follows by assuming independence among the observations and the last line is simply a convenient expression for picking the relevant probability. This likelihood function captures all three binary data models discussed in Section 1.2.1 – probit, logit, and t-link – which could be obtained by using the appropriate cdf in place of $F(\cdot)$ as discussed in Section 1.2.1.

In order to find the maximum likelihood estimate $\hat{\beta}_{MLE}$, we maximize the log-likelihood function

$$\ln f(y|\beta) = \sum_{i=1}^{n} \{y_i \ln F(x_i'\beta) + (1 - y_i) \ln[1 - F(x_i'\beta)]\},$$

which is typically done iteratively using standard hill climbing algorithms such as Newton-Raphson or BHHH (see [3]) because the first-order condition for

maximization

$$\frac{\partial \ln f(y|\beta)}{\partial \beta} \equiv \sum_{i=n}^{n} \left[\frac{y_i f(x_i'\beta)}{F(x_i'\beta)} - (1-y_i)\frac{f(x_i'\beta)}{1-F(x_i'\beta)} \right] x_i = 0$$

does not admit an explicit analytical solution even though the log-likelihood is typically well behaved (unimodal and concave) in this class of models.

Turning attention to ordinal outcomes, equation (1.5) and the assumption of independent sampling give the following likelihood function for the ordinal data model

$$\begin{aligned} f(y|\beta,\gamma) &= \prod_{i=1}^{n} \Pr(y_i|\beta,\gamma) \\ &= \prod_{i=1}^{n} \left[F(\gamma_j - x_i'\beta) - F(\gamma_{j-1} - x_i'\beta) \right], \end{aligned} \tag{1.8}$$

where the index j on the cutpoints in the second line is determined by the realization of y_i (recall that because y_i takes values in $\{1,\ldots,J\}$, it can be used for indexing the cutpoints, i.e. $\gamma_j = \gamma_{y_i}$ and $\gamma_{j-1} = \gamma_{y_i-1}$).

A minor complication arises in maximizing $\ln f(y|\beta,\gamma)$ because the values of the free cutpoints must satisfy an ordering constraint: $\gamma_1 = 0 < \gamma_2 < \ldots < \gamma_{J-1}$. In order to avoid the computational complexities associated with constrained optimization, it is useful to reparameterize the problem in order to remove those constraints. For example, optimization can be simplified by transforming the cutpoints γ so as to remove the ordering constraint by the one-to-one map

$$\delta_j = \ln(\gamma_j - \gamma_{j-1}),\ 2 \le j \le J-1, \tag{1.9}$$

and rewriting the likelihood as a function of β and $\delta = (\delta_2,\ldots,\delta_{J-1})'$, i.e. drawing inferences from $f(y|\beta,\delta)$. Other transformations have been considered in [4] and comparisons have been drawn in [13], but these transformations relate to alternative identification restrictions of the scale of the model and will not be examined here.

1.3.2 Bayesian Estimation

In contrast to classical (or frequentist) inference, which only involves the likelihood function $f(y|\theta)$, Bayesian analysis rests on Bayes' theorem

$$\pi(\theta|y) = \frac{f(y|\theta)\pi(\theta)}{\int f(y|\theta)\pi(\theta)d\theta},$$

and inference is based on the posterior density $\pi(\theta|y)$, which is proportional to the product of the likelihood and the prior density $\pi(\theta)$. There are important theoretical advantages of Bayesian analysis over classical inference,

which have been carefully reviewed in [10], [14], [21], and [23]. For example, the posterior density allows for finite sample inferences about the unknown parameter vector θ that incorporates information from the observed sample (which enters through the likelihood) and non-sample information (e.g. from previous studies, theoretical considerations, the researcher's experience, etc.), which enters through the prior. In addition to their finite sample properties, Bayesian estimators also have desirable asymptotic properties (as $n \to \infty$).

An important practical benefit of Bayesian estimation is that inference is possible even in models where the likelihood $f(y|\theta)$ is difficult to evaluate and hence maximum likelihood estimation is infeasible. In those cases, progress has been made possible by recent advances in simulation-based estimation and data augmentation which allow sampling from $\pi(\theta|y)$ without requiring evaluation of $f(y|\theta)$. Such simulation methods, based on MCMC theory, have enabled inference in previously intractable applications. Once a sample of draws $\{\theta\}$ from $\pi(\theta|y)$ is available, those draws can be used to summarize features of the posterior (such as mean, variance, quantiles, etc.) and construct point and interval estimates.

For the binary and ordinal data models we have examined in this chapter, Bayes' theorem will lead to a posterior density

$$\pi(\theta|y) \propto f(y|\theta)\pi(\theta)$$

that typically does not belong to a known family of distributions and can not be sampled directly. This is because even if the prior $\pi(\theta)$ is selected from a well-known class of distributions (e.g., Gaussian), the parameters enter the likelihood $f(y|\theta)$ in such a way (note the nonlinearity in equations (1.7) and (1.8)) that the posterior $\pi(\theta|y)$ does not have a recognizable analytical representation.

In this Section we present tools for dealing with this problem in two ways. First, we discuss a general MCMC simulation technique, called the Metropolis-Hastings algorithm, that can be employed to produce draws from intractable distributions. Second, we review a method that circumvents the problem by augmenting the sampling scheme with an additional vector of variables in a way that restores tractability. The benefit of this approach, called data augmentation, is that it enables estimation by Gibbs sampling (another MCMC simulation technique). In the remainder of this Section, we review all of these methods and propose a new data augmentation algorithm for the logit model which has not appeared elsewhere in the literature.

1.3.2.1 Metropolis-Hastings Algorithm. The Metropolis-Hastings (MH) algorithm ([19], [12], [24], [5]) is a versatile Markov chain simulation method for non-standard distributions. Denoting the current value of θ by θ^c, it proceeds by generating a proposed value $\theta^p \sim q(\theta|y)$ from the proposal density $q(\cdot)$. In principle $q(\cdot)$ can depend on θ^c (e.g. in random walk proposal densities), but in this chapter we examine a version of the MH algorithm, called independence chain MH, in which the proposal density does not vary with θ^c. The

proposed draw θ^p is accepted with probability

$$\alpha_{MH}(\theta^c, \theta^p|y) = \min\left\{1, \frac{f(y|\theta^p)\pi(\theta^p)q(\theta^c|y)}{f(y|\theta^c)\pi(\theta^c)q(\theta^p|y)}\right\},$$

and if θ^p is rejected, θ^c is repeated as the next value of θ in the Markov chain. As shown by [12] (also see [24], [5]), the limiting distribution of the draws of θ coming from the MH algorithm is $\pi(\theta|y)$. In practice, this means that after a transient phase (called the burn-in period), draws obtained by MH simulation can be viewed as coming from $\pi(\theta|y)$.

To apply the independence chain MH algorithm in our context, we note that a suitable proposal density can be obtained by employing the MLE results from Section 1.3.1. In particular, for any of the models studied in Sections 1.2.1 and 1.2.2, the proposal density can be constructed as a multivariate t density

$$q(\theta|y) = f_{T_\omega}(\theta|\hat{\theta}, a\Psi),$$

with mean $\hat{\theta} = \hat{\theta}_{MLE}$ and scale matrix $a\Psi$, where Ψ is given by the inverse of the negative Hessian of the log-likelihood

$$\Psi = -\left[\frac{\partial^2 \ln f(y|\theta)}{\partial\theta\partial\theta'}\right]^{-1}.$$

evaluated at $\theta = \hat{\theta}_{MLE}$, a is a scalar tuning parameter, and ω is the degrees of freedom of the proposal density. The tuning parameter a is typically taken to be $a \geq 1$ and ω is usually set at a small value, both of which are intended to ensure that the proposal has sufficiently heavy tails to explore the space more thoroughly. In the examples in this paper, we use $\omega = 10$ and $a = 1.25$.

The independence chain MH algorithm can then be employed to estimate probit, logit and robit models for binary data using the likelihood in (1.7) with parameter vector $\theta = \beta$, or ordinal models using likelihood (1.8) written in terms of the transformed cutpoints in (1.9) whereby the parameter vector θ is given by $\theta = (\beta', \delta')'$.

1.3.2.2 Gibbs Sampling and Data Augmentation. Gibbs sampling (see [9]) is an MCMC method for simulation from a distribution when its full conditional densities have known form. To review the approach, suppose there are three parameter blocks θ_1, θ_2, and θ_3 with joint density $\pi(\theta_1, \theta_2, \theta_3|y)$. The Gibbs sampler produces draws $\{\theta_1, \theta_2, \theta_3\} \sim \pi(\theta_1, \theta_2, \theta_3|y)$ by sequentially drawing from the set of full conditional densities $\pi(\theta_1|y, \theta_2, \theta_3)$, $\pi(\theta_2|y, \theta_1, \theta_3)$ and $\pi(\theta_3|y, \theta_1, \theta_2)$. Under mild conditions, it can be shown that the Markov chain formed by the Gibbs sampler has a limiting invariant distribution that is the distribution of interest $\pi(\theta_1, \theta_2, \theta_3|y)$. This means that draws obtained by Gibbs sampling after the initial burn-in period, can be viewed as coming from $\pi(\theta_1, \theta_2, \theta_3|y)$. Some authors have likened the way in which the Gibbs sampler traverses the parameter space to the way a rook moves in chess. In addition,

the particular order in which the full conditional densities are sampled does not affect the limiting distribution. A thorough review of the method and its applications in econometrics is offered in [6].

The application of Gibbs sampling to models for binary and ordinal data is complicated by the fact that the posterior and its full conditional densities are not of known form. However, a method known as data augmentation can be used to overcome this problem.

The idea behind data augmentation is simple. Instead of focusing on the intractable posterior density

$$\pi(\theta|y) \propto f(y|\theta)\pi(\theta),$$

we choose to work with $\pi(\theta, w|y)$, a density judiciously augmented with w in such a way that the full-conditionals $\pi(\theta|y, w)$ and $\pi(w|y, \theta)$ are tractable and can be sampled directly. As a result, a Gibbs sampler constructed using sequential sampling from $\pi(\theta|y, w)$ and $\pi(w|y, \theta)$ will produce draws $\{\theta, w\} \sim \pi(\theta, w|y)$.

But how do we relate the draws $\{\theta, w\}$ from $\pi(\theta, w|y)$ to our original goal of sampling $\theta \sim \pi(\theta|y)$? This is easily done by only collecting the draws of θ and simply ignoring w. The approach works because by the properties of cdfs, given two vectors of constants a_θ and a_w conformable with θ and w, respectively, the marginal cdf is obtained from the joint cdf as

$$F(a_\theta) \equiv \Pr(\theta \ll a_\theta) = \lim_{a_w \to \infty} F(a_\theta, a_w) \equiv \Pr(\theta \ll a_\theta, w \ll \infty),$$

where '$\ll$' is used to denote element-by-element weak inequality comparison. The condition $w \ll \infty$ always holds and in this sense we "simply ignore" w to obtain draws $\theta \sim \pi(\theta|y)$ from $\{\theta, w\} \sim \pi(\theta, w|y)$.

Having presented the theory behind data augmentation, we now discuss its specific application to the models considered in this chapter.

EXAMPLE 1.4

The binary probit model can be estimated easily, as shown in [1], if we were to introduce the latent $z = (z_1, \ldots, z_n)'$ from equation (1.2) into our MCMC simulation algorithm. Specifically, instead of working with $\pi(\beta|y)$, we specify a Gibbs sampler to simulate the augmented posterior $\pi(\beta, z|y)$, which can be written as

$$\begin{aligned} \pi(\beta, z|y) &\propto f(y|\beta, z)\, f(\beta, z) \\ &= f(y|\beta, z)\, f(z|\beta)\, \pi(\beta) \\ &= \left\{ \prod_{i=1}^{n} f(y_i|z_i) \right\} f(z|\beta)\, \pi(\beta). \end{aligned} \tag{1.10}$$

Note that the last line of (1.10) involves terms that are easy to evaluate. In particular, $f\left(y_i|z_i\right) = 1\{z_i \in \mathcal{B}_i\}$, where

$$\mathcal{B}_i = \begin{cases} (0,\infty) & \text{if} \quad y_i = 1, \\ (-\infty, 0] & \text{if} \quad y_i = 0, \end{cases} \tag{1.11}$$

which follows from the relationship between y_i and z_i in binary data models. Note that conditionally on z_i, y_i does not depend on β. In addition, $f(z|\beta) = f_N(z|X\beta, I_n)$, where $X = (x_1', \ldots, x_n')'$ is the matrix of covariates and I_n denotes the $n \times n$ identity matrix; this follows from the latent variable representation of the probit model, namely $z_i = x_i'\beta + \nu_i$ with $\nu_i \sim N(0,1)$ for $i = 1, \ldots, n$. Finally, $\pi\left(\beta\right)$ is the prior distribution on β which we assume to be $f_N(\beta|\beta_0, B_0)$, i.e. β is assumed to be *a priori* normally distributed, i.e. $\beta \sim N(\beta_0, B_0)$.

A Gibbs sampler now can be easily constructed to explore $\pi\left(\beta, z|y\right)$ because the full conditional densities $\pi\left(\beta|y, z\right)$ and $\pi\left(z|y, \beta\right)$ are of known form. Specifically, $\pi\left(\beta|y, z\right)$ is proportional to the terms in (1.10) that involve β so that $\pi\left(\beta|y, z\right) \propto f\left(z|\beta\right)\pi\left(\beta\right)$, which technically does not depend on y. Because both $f\left(z|\beta\right)$ and $\pi\left(\beta\right)$ are normal, the full conditional is also normal and therefore we draw

$$\beta|y, z \sim N(\hat{\beta}, \hat{B}),$$

where $\hat{B} = (B_0^{-1} + X'X)^{-1}$ and $\hat{\beta} = \hat{B}(B_0^{-1}\beta_0 + X'z)$. Details of the derivation are considered in Exercise 1.3.

The density $\pi\left(z|y, \beta\right)$ is proportional to the terms in (1.10) that involve z so that

$$\begin{aligned} \pi\left(z|y, \beta\right) &\propto \left\{\prod_{i=1}^{n} 1\{z_i \in \mathcal{B}_i\}\right\} f_N(z|X\beta, I_n) \\ &= \prod_{i=1}^{n} \left[1\{z_i \in \mathcal{B}_i\} f_N(z_i|x_i'\beta, 1)\right], \end{aligned}$$

whereby $z|y, \beta$ is easily sampled by drawing z_i, $i = 1, \ldots, n$, from appropriately truncated normal densities

$$z_i|y_i, \beta \sim TN_{\mathcal{B}_i}(x_i'\beta, 1),$$

where the region of truncation $\mathcal{B}_i$ is defined in (1.11).

EXAMPLE 1.5

The t-link (robit) model can be estimated by extending the data augmentation approach presented in Example 1.4. The discussion follows

[1] and rests on the result (see, e.g., [2]) that the t distribution can be represented as a scale mixture of normals. Specifically, if for $i = 1, \ldots, n$, λ_i has a gamma distribution

$$\lambda_i \sim G(\tau/2, \tau/2), \tag{1.12}$$

and conditionally on λ_i, we have

$$z_i|\lambda_i \sim N(x_i'\beta, 1/\lambda_i), \tag{1.13}$$

then marginally of λ_i, z_i is distributed

$$z_i \sim T_\tau(x_i'\beta, 1).$$

Therefore, letting $\lambda = (\lambda_1, \ldots, \lambda_n)'$, we can consider the augmented posterior

$$\begin{aligned} \pi(\beta, z, \lambda|y) &\propto f(y|\beta, z, \lambda) f(\beta, z, \lambda) \\ &= f(y|\beta, z, \lambda) f(z|\beta, \lambda) \pi(\beta)\pi(\lambda) \\ &= \left\{ \prod_{i=1}^{n} f(y_i|z_i) \right\} f(z|\beta, \lambda) \pi(\beta)\pi(\lambda), \end{aligned} \tag{1.14}$$

where $f(y_i|z_i) = 1\{z_i \in \mathcal{B}_i\}$ as before, $f(z|\beta, \lambda) = f_N(z|X\beta, \Lambda^{-1})$ with $\Lambda = \text{diag}(\lambda)$ which follows from (1.13), $\pi(\beta) = f_N(\beta|\beta_0, B_0)$ is the prior on β, and $\pi(\lambda)$ is given by the product of n independent gamma densities stemming from (1.12)

$$\pi(\lambda) = \prod_{i=1}^{n} f_G(\lambda_i|\tau/2, \tau/2).$$

It is then quite straightforward to show that the Gibbs sampler for simulating from $\pi(\beta, z, \lambda|y)$ can be constructed by sequentially drawing from the following full conditionals

$$\beta|z, \lambda \sim N(\hat{\beta}, \hat{B}),$$

with $\hat{B} = (B_0^{-1} + X'\Lambda X)^{-1}$ and $\hat{\beta} = \hat{B}(B_0^{-1}\beta_0 + X'\Lambda z)$,

$$z_i|y, \beta, \lambda_i \sim TN_{\mathcal{B}_i}(x_i'\beta, \lambda_i^{-1}), \qquad i = 1, \ldots, n,$$

and

$$\lambda_i|y, \beta, z \sim G\left(\frac{\tau + 1}{2}, \frac{\tau + (z_i - x_i'\beta)^2}{2} \right), \qquad i = 1, \ldots, n.$$

EXAMPLE 1.6

The logit model can be estimated by pursuing a new data augmentation scheme that has not been exploited in the literature. Because the logistic distribution can be written as a scale mixture of normals with respect to the Kolmogorov distribution ([2], [22]), we can write that

$$f_L(s|\mu) = \int f_N(s|\mu, 4\kappa^2) f_K(\kappa) d\kappa \tag{1.15}$$

where $f_L(s|\mu)$ denotes the density of a random variable that has a logistic distribution around μ and variance $\pi^2/3$, and $f_K(\kappa)$ represents the Kolmogorov density $f_K(\kappa) = 8\kappa \sum_{j=1}^{\infty} (-1)^{j+1} j^2 e^{-2j^2\kappa^2}$. This implies, analogously to Example 1.6, that if κ_i has a Kolmogorov distribution and conditionally on κ_i, $z_i|\kappa_i \sim N(x_i'\beta, 4\kappa_i^2)$, then marginally of κ_i, z_i has logistic density $f_L(z_i|x_i'\beta)$.

Therefore, letting $\kappa = (\kappa_1, \ldots, \kappa_n)'$, we can consider the augmented posterior

$$\begin{aligned} \pi(\beta, z, \kappa|y) &\propto f(y|\beta, z, \kappa) f(\beta, z, \kappa) \\ &= f(y|\beta, z, \kappa) f(z|\beta, \kappa) \pi(\beta)\pi(\kappa) \\ &= \left\{ \prod_{i=1}^{n} f(y_i|z_i) \right\} f(z|\beta, \kappa) \pi(\beta)\pi(\kappa). \end{aligned} \tag{1.16}$$

where $f(y_i|z_i) = 1\{z_i \in \mathcal{B}_i\}$, $f(z|\beta, \kappa) = f_N(z|X\beta, K)$ with $K = \text{diag}(4\kappa^2)$, $\pi(\beta) = f_N(\beta|\beta_0, B_0)$, and $\pi(\kappa) = \prod_{i=1}^{n} f_K(\kappa_i)$.

The resulting Gibbs sampler for simulating from $\pi(\beta, z, \kappa|y)$ is constructed by sequentially drawing from the following full conditionals

$$\beta|z, \kappa \sim N(\hat{\beta}, \hat{B}),$$

with $\hat{B} = (B_0^{-1} + X'K^{-1}X)^{-1}$ and $\hat{\beta} = \hat{B}(B_0^{-1}\beta_0 + X'K^{-1}z)$,

$$z_i|y, \beta, \kappa_i \sim TN_{\mathcal{B}_i}(x_i'\beta, 4\kappa_i^2), \qquad i = 1, \ldots, n,$$

and

$$\kappa_i|y, \beta, z_i \sim f(\kappa_i|z_i, \beta), \qquad i = 1, \ldots, n, \tag{1.17}$$

where $f(\kappa_i|z_i, \beta)$ does not belong to a known family of distributions. However, a very convenient result can be obtained by representing this

distribution in terms of Bayes' formula as

$$\begin{aligned} f(\kappa_i|z_i,\beta) &= \frac{f(z_i|\beta,\kappa_i)f(\kappa_i)}{\int f(z_i|\beta,\kappa_i)f(\kappa_i)d\kappa_i} \\ &= \frac{f_N(z_i|x_i'\beta,4\kappa_i^2)f_K(\kappa_i)}{f_L(z_i|x_i'\beta)}. \end{aligned} \tag{1.18}$$

The last line in (1.18) follows by recognizing that the numerator densities are Gaussian and Kolmogorov, and the denominator, by equation (1.15), is simply the logistic density. Therefore, the unknown $f(\kappa_i|z_i,\beta)$ can now be represented very simply in terms of other well-known densities.

The fact that $f(\kappa_i|z_i,\beta)$ can be evaluated explicitly means that one can also evaluate the corresponding cdf

$$F_{\kappa|z,\beta}(\kappa_i|z_i,\beta) = \int_0^{\kappa_i} f(s|z_i,\beta)ds.$$

In turn, $F_{\kappa|z,\beta}(\kappa_i|z_i,\beta)$ can be utilized to produce the draws needed in (1.17) by solving $\kappa_i = F^{-1}_{\kappa|z,\beta}(u)$, where $u \sim U(0,1)$ is a uniform random variable on the unit interval. The latter technique is known as the inverse cdf method and follows because

$$\Pr(\kappa_i \leq a) = \Pr(F^{-1}_{\kappa|z,\beta}(u) \leq a) = \Pr(u \leq F_{\kappa|z,\beta}(a)) = F_{\kappa|z,\beta}(a).$$

This completes the proposed Gibbs sampling scheme for logit models. However, to provide additional intuition about the behavior of $f(\kappa_i|z_i,\beta)$ and compare it to the Kolmogorov distribution $f_K(\kappa_i)$, Figure 1.4 plots $f(\kappa_i|z_i,\beta)$ for two settings of $z_i - x_i'\beta$. The figure reveals that when z_i is close to the mean $x_i'\beta$ the mass of the distribution is closer to the origin than when z_i is far (in absolute terms) from $x_i'\beta$. This is to be expected because κ_i enters the conditional variance of z_i.

EXAMPLE 1.7

The analysis of the ordinal probit model is similar to the cases considered in the preceding examples. In particular, given the priors $\beta \sim N(\beta_0, B_0)$ and $\delta \sim N(d_0, D_0)$, the augmented posterior distribution is given by

$$\begin{aligned} \pi(\beta,\delta,z|y) &\propto f(y|\beta,\delta,z)\, f(z|\beta)\pi(\beta)\pi(\delta) \\ &= \left\{\prod_{i=1}^{n} f(y|\delta,z_i)\right\} f(z|\beta)\pi(\beta)\pi(\delta), \end{aligned} \tag{1.19}$$

where $f(y_i|\delta,z_i) = 1\{\gamma_{j-1} < z_i \leq \gamma_j\}$, the correspondence between γ and δ is determined by (1.9), and the cutpoint index j is given by

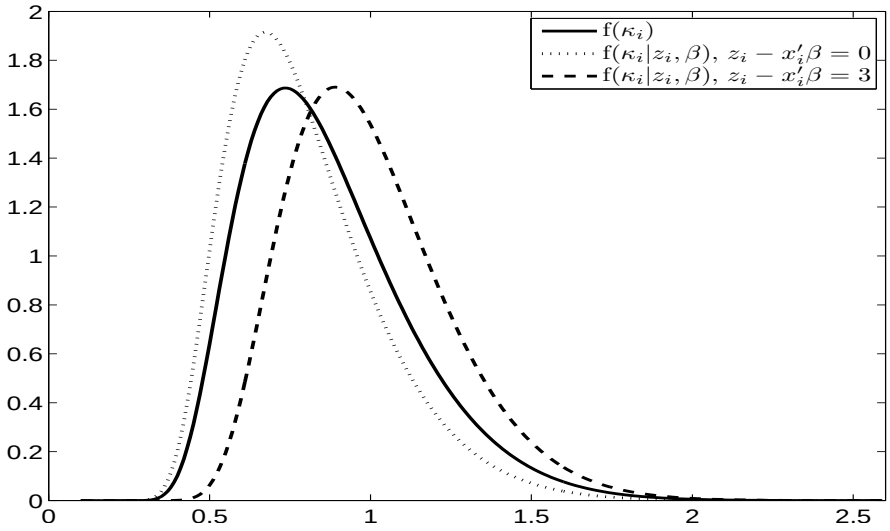

Figure 1.4 Behavior of the density $f(\kappa_i|z_i,\beta)$ relative to $f_K(\kappa_i)$.

the realization of y_i. Furthermore, $f(z|\beta) = f_N(z|X\beta, I_n)$, $\pi(\beta) = f_N(\beta|\beta_0, B_0)$, and $\pi(\delta) = f_N(\delta|d_0, D_0)$.

It has been noted in the literature that in order to design an efficient MCMC sampler for the ordinal probit model, δ and z must be simulated jointly, not conditionally on each other. The reason that conditional sampling does not mix well is that z and δ (which determines the values of γ) constrain each other through the restrictions $\{\gamma_{y[i]-1} < z_i < \gamma_{y[i]}\}$, whereby the sampler can only slowly explore the posterior distribution. Several alternatives for joint sampling are reviewed in [13], and the following simulation scheme is suggested.

1. Sample $\delta, z|y, \beta$ in one block as follows:

   (a) Sample $\delta|y,\beta$ marginally of z by drawing $\delta^p \sim q(\delta|y,\beta)$ from a proposal density $q(\delta|y,\beta) = f_{T_\omega}(\delta|\hat{\delta}, \hat{D})$, where

   $$\hat{\delta} = \arg\max_{\delta} \ln f(y|\beta,\delta) \quad \text{and} \quad \hat{D} = -\left[\frac{\partial^2 \ln f(y|\beta,\delta)}{\partial\delta\partial\delta'}\right]^{-1}\Bigg|_{\delta=\hat{\delta}}.$$

   Accept δ^p with probability

   $$\alpha_{MH}(\delta,\delta^p) = \min\left\{1, \frac{f(y|\beta,\delta^p)\pi(\delta^p)}{f(y|\beta,\delta^c)\pi(\delta^c)}\frac{q(\delta^c|y,\beta)}{q(\delta^p|y,\beta)}\right\},$$

   otherwise repeat the current value δ^c.

(b) Sample $z_i|y, \beta, \delta \sim TN_{(\gamma_{j-1},\gamma_j)}(x_i'\beta, 1)$ for $i = 1, \ldots, n$, where γ is obtained by the one-to-one mapping relating γ and δ.

2. Sample $\beta|z \sim N(\hat{\beta}, \hat{B})$ with

$$\hat{B} = (B_0^{-1} + X'X)^{-1} \quad \text{and} \quad \hat{\beta} = \hat{B}(B_0^{-1}\beta_0 + X'z).$$

In Step 1 of this algorithm, the degrees of freedom parameter ω is taken to be a low number such as 5 or 10 to ensure that the proposal density has sufficiently heavy tails. Grouping δ and z into a single sampling block dramatically improves the mixing of the Markov chain.

We complete the discussion of data augmentation by emphasizing its practical appeal. For instance, data augmentation is often the only viable estimation approach in a variety of multivariate and hierarchical models. Maximum likelihood estimation becomes infeasible in those settings owing to the intractability of the likelihood function. However, data augmentation allows us to circumvent this difficulty by simulating from well-known distributions without having to evaluate the likelihood. This has enabled inference in difficult settings such as multivariate and multinomial probit, mixed logit, multivariate ordinal probit, copula models, panel data models, models with incidental truncation, treatment models, and many others.

1.3.3 Marginal Effects

Having estimated the parameters of a model, one is typically interested in the practical implications of those estimates. However, interpretation of the parameter estimates is complicated by the nonlinearity of the models we have considered. In binary data models, for example, $E(y_i|x_i, \beta) = \Pr(y_i = 1|x_i, \beta) = F(x_i'\beta)$. Therefore, the marginal effect of changing some continuous covariate in x_i, say x_h, is not simply given by β_h. This can be easily seen by taking the derivative of $\Pr(y_i = 1|x_i, \beta)$ with respect to x_h

$$\frac{\partial \Pr(y_i = 1|x_i, \beta)}{\partial x_h} = \frac{\partial F(x_i'\beta)}{\partial x_h} = f(x_i'\beta)\beta_h,$$

and hence the marginal effect of x_h depends on β_h, but also on all of the covariates in x_i, all of the parameters in β, and will differ with the choice of cdf $F(\cdot)$ and respective pdf $f(\cdot)$. Table 1.1 gives the choice probability and marginal effects for the three commonly used binary data models.

Given a specific model, there are several approaches to compute the average marginal effect of covariate x_h. One approach is to evaluate the marginal effect using the sample average of the regressors $\bar{x}_i$ and the point estimate $\hat{\beta}$, i.e. $f(\bar{x}_i'\hat{\beta})\hat{\beta}_h$. However, this average effect may not represent the effect in the population well because $f(\cdot)$ is a non-linear function and by Jensen's inequality $f(\bar{x}_i'\hat{\beta}) \neq \overline{f(x_i'\hat{\beta})}$. Therefore, a more reasonable approach would be

Model	Probability $P(y_i = 1\|x_i, \beta))$	Marginal Effect of x_h
Logit	$F_L(x_i'\beta) = \frac{e^{x_i'\beta}}{1+e^{x_i'\beta}}$	$f_L(x_i'\beta)\beta_h$
Probit	$\Phi(x_i'\beta) = \int_{-\infty}^{x_i'\beta} \phi(z)dz$	$\phi(x_i'\beta)\beta_h$
t-link	$F_{T_\tau}(x_i'\beta) = \int_{-\infty}^{x_i'\beta} f_{T_\tau}(t)dt$	$f_{T_\tau}(x_i'\beta)\beta_h$

Table 1.1 Marginal effects in binary data models.

to calculate the sample average of the marginal effects

$$\overline{f(x_i'\hat{\beta})}\hat{\beta}_h = n^{-1}\sum_{i=1}^{n} f(x_i'\hat{\beta})\hat{\beta}_h.$$

Even though this quantity is better than computing $f(\bar{x}_i'\hat{\beta})\hat{\beta}_h$ as suggested in [26], it has an important drawback: it does not account for the variability in β. For this reason, [8] and [13] suggest that the average covariate effect should be computed by averaging over both the covariates and parameters. If estimation is done by MCMC simulation, one can use draws $\beta^{(m)} \sim \pi(\beta|y)$ to construct the average covariate effect as follows

$$\overline{f(x_i'\beta)\beta_h} = \frac{1}{nM}\sum_{i=1}^{n}\sum_{m=1}^{M} f\left(x_i'\beta^{(m)}\right)\beta_h^{(m)}.$$

Note that unlike the earlier quantities we considered, $\overline{f(x_i'\beta)\beta_h}$ produces an estimate of the average effect that accounts for variability in both x_i and β.

1.4 APPLICATIONS

1.4.1 Women's Labor Force Participation

We apply the techniques of this chapter to study the determinants of women's labor force participation, a topic that has been extensively studied because of the large increases in women's participation and hours of work in the post-war period. For instance there has been a seven-fold increase in the participation rate of married women since the 1920s. Understanding labor force participation and entry and exit decisions is a fundamental prerequisite for understanding wages because wages are not observed for women who do not work.

The data set used in this application has been studied in [20] and [7]. The sample consists of 753 married women, 428 of whom were employed. The variables in the data set are summarized in Table 1.2.

Covariate	Explanation	Mean	SD
KLT6	number of kids under 6 years old	0.28	0.52
KGE6	number of kids 6–18 years old	1.35	1.32
NWINC	estimated nonwife income (1975, in $10,000)	2.01	1.16
MEDU	mother's years of schooling	9.25	3.37
FEDU	father's years of schooling	8.81	3.57
HEDU	husband's years of schooling	12.49	3.02
AGE	woman's age in years	42.54	8.07
EXPER	actual labor market experience in years	10.63	8.07

Table 1.2 Covariates in the women's labor supply example.

We implemented the techniques developed in this chapter to estimate probit, logit, and t-link models of the binary participation decision. Estimation was carried out by the MCMC simulation methods discussed in Section 1.3.2 and our results are summarized in Table 1.3.

	Probit		t-link		Logit	
Covariate	Mean	SD	Mean	SD	Mean	SD
1	1.1758	0.4358	1.1737	0.4586	1.3931	0.6188
KLT6	-0.7964	0.1115	-0.8285	0.1210	-1.2476	0.1847
KGE6	0.0346	0.0415	0.0362	0.0443	0.0763	0.0695
NWINC	-0.0773	0.0484	-0.0817	0.0531	-0.1384	0.0825
MEDU	0.0320	0.0184	0.0339	0.0197	0.0580	0.0306
FEDU	0.0143	0.0175	0.0158	0.0189	0.0250	0.0300
HEDU	0.0251	0.0188	0.0265	0.0207	0.0476	0.0326
AGE	-0.0517	0.0078	-0.0534	0.0083	-0.0769	0.0117
EXPER	0.0745	0.0074	0.0796	0.0084	0.1270	0.0138

Table 1.3 Parameter estimates in the women's labor force participation application.

The estimates in Table 1.3 are consistent with the predictions of economic theory. For example, having young children reduces labor force participation as evidenced by the negative mean and a 95% credibility interval that lies below zero, but older children have little impact on the mother's decision to work. Again, consistent with economic theory, higher non-wife income and lower parents' and husband's schooling reduce participation. The table also shows that age has a strong negative effect, which is consistent with cohort and life-cycle effects, whereas experience has a strong positive effect on probability of working, which is consistent with increases in productivity as experience grows.

1.4.2 An Ordinal Model of Educational Attainment

We now consider the ordinal probit model of educational attainment studied in [13]. Educational attainment has been the subject of a large literature because of its implications for earnings, economic growth, and social well-being. The setting is suitable for ordinal modeling because the dependent variable is naturally categorized by measurable thresholds into a number of distinct groups. This application considers the following four ordered outcomes: (1) less than a high school education, (2) high school degree, (3) some college or associate's degree, and (4) college or graduate degree. The data are obtained from the National Longitudinal Survey of Youth (NLSY79).

In this study it is of interest to examine the effect of family background and individual variables on educational attainment. The family background variables included in the data set are: the highest grade completed by the individual's father and mother, whether the mother worked, square root of family income, an indicator for whether the youth lived in an urban area, and an indicator for whether the youth lived in the South. The individual variables include gender and race, as well as three indicator variables that control for age cohort affects. The sample is restricted to those cohorts that were between 14 and 17 years old in 1979. The sample is restricted to include only individuals whose records have all relevant variables. Additionally, the sample excludes disabled individuals and those who report more than 11 years of education at age 15. The resulting sample consists of 3923 individuals.

The model was estimated by the MCMC simulation techniques discussed in Section 1.3.2. The results are presented in Table 1.4. The coefficient estimates in the table are consistent with other findings in the literature. Parental education and income have a positive effect on educational attainment, as might be expected. *A priori*, the effect of mother's labor force participation is theoretically ambiguous – on the one hand, a mother's work force participation could be detrimental due to reduced parental supervision, but on the other, it provides a positive example for her children to follow. The empirical findings in Table 1.4 indicate that the net effect is positive, although it is not precisely estimated. Conditionally on the remaining covariates, we also see that blacks and individuals from the South have higher educational attainment.

Following [13], we computed the effect of an increase in family income on educational outcomes following the discussion in Section 1.3.3. For the overall sample, the effect of a \$1000 increase in family income is to lower the probability of dropping out of high school by approximately 0.0050, lower the probability of only obtaining a high school degree by 0.0006, but increase the probability of having some college or associate's degree by 0.0020 and increase the probability of getting a college or graduate degree by 0.0036. We also computed these effects for specific subsamples that are of interest. For the subsample of females, the effects of an income increase on the four outcome probabilities were comparable at approximately -0.0048, -0.0009, 0.0019, and 0.0038, respectively. For the subsample of blacks, the effects of

Parameter	Covariate	Mean	SD
β	Intercept	-1.34	0.09
	Family income (sq. rt.)	0.14	0.01
	Mother's education	0.05	0.01
	Father's education	0.07	0.01
	Mother worked	0.03	0.04
	Female	0.16	0.04
	Black	0.15	0.04
	Urban	-0.05	0.04
	South	0.05	0.04
	Age cohort 2	-0.03	0.05
	Age cohort 3	0.00	0.06
	Age cohort 4	0.23	0.06
δ	(transformed cutpoint)	0.08	0.02
	(transformed cutpoint)	-0.28	0.03

Table 1.4 Parameters estimates in the educational attainment application.

income change were somewhat stronger – in that subsample, an increase of \$1000 in family income changed the four educational outcome probabilities by -0.0060, -0.0009, 0.0026, and 0.0043, respectively.

1.5 CONCLUSIONS

This chapter has introduced the theory behind binary and ordinal models in economics, and has examined their estimation by both maximum likelihood and Bayesian simulation methods. We have reviewed several existing MCMC algorithms and have proposed a new data augmentation method for the estimation of logit models. The ability to implement data augmentation techniques makes it possible to extend these techniques and estimate models in which the likelihood function is intractable.

The methods are examined in two applications dealing with labor force participation and educational attainment. The applications illustrate that the models and estimation methods are practical and can uncover interesting features in the data.

EXERCISES

1.1 The cdf of the logistic distribution is given by

$$F_L(\nu) = \left(1 + e^{-\nu}\right)^{-1} = \frac{e^{\nu}}{1 + e^{\nu}}.$$

Show that the logistic pdf, $f_L(\nu)$, can be written as

$$f_L(\nu) = F_L(\nu)\left[1 - F_(\nu)\right].$$

1.2 Suppose the random utility model is given by

$$U_{ij} = x_i'\beta_j + \varepsilon_{ij}, \quad i = 1, \ldots, n, \quad j = 0, 1.$$

Starting with equation (1.1), show that if ε_{i0} and ε_{i1} are independent and identically distributed as extreme value type I with density

$$f_{EV}(\varepsilon) = e^{-\varepsilon}e^{-e^{-\varepsilon}}$$

and cumulative distribution function

$$F_{EV}(\varepsilon) = e^{-e^{-\varepsilon}},$$

then (1.1) gives rise to the logistic outcome probability

$$\Pr(y_i = 1|\beta) = \frac{1}{1 + e^{-x_i'\beta}},$$

where $\beta = \beta_1 - \beta_0$.

1.3 Consider the probit model and assume the prior $\beta \sim N(\beta_0, B_0)$. Show that given the latent data z, the full conditional distribution for β is

$$\beta|y, z \sim N(\hat{\beta}, \hat{B}),$$

where $\hat{B} = (B_0^{-1} + X'X)^{-1}$ and $\hat{\beta} = \hat{B}(B_0^{-1}\beta_0 + X'z)$.

REFERENCES

1. Albert J. and Chib S., "Bayesian Analysis of Binary and Polychotomous Response Data," *Journal of American Statistical Association*, Vol. 88, No. 422, pp. 669-679, (1993).

2. Andrews D.F. and Mallows C.L., "Scale Mixtures of Normal Distributions," *Journal of the Royal Statistical Society, Series B*, Vol. 36, No. 1, pp. 99-102, (1974).

3. Berndt E., Hall B., Hall R., and Hausman J., "Estimation and Inference in Nonlinear Structural Models," *Annals of Economic and Social Measurement*, Vol. 3, pp. 653-665.

4. Chen M.-H. and Dey D. K., "Bayesian Analysis for Correlated Ordinal Data Models," in D. Dey, S. Ghosh and B. Mallick (eds.), *Generalized Linear Models: A Bayesian Perspective*, pp. 133-157. New York: Marcel-Dekker.

5. Chib S. and Greenberg E., "Understanding the Metropolis-Hastings Algorithm," *The American Statistician*, Vol. 49, No. 4, pp. 327-335, (1995).

6. Chib S. and Greenberg E., "Markov Chain Monte Carlo Simulation Methods in Econometrics," *Econometric Theory*, Vol. 12, No. 3, pp. 409-431, (1996).

7. Chib S., Greenberg E., and Jeliazkov I., "Estimation of Semiparametric Models in the Presence of Endogeneity and Sample Selection," *Journal of Computational and Graphical Statistics*, Vol. 18, pp. 321-348, (2009).

8. Chib S. and Jeliazkov I., "Inference in Semiparametric Dynamic Models for Binary Longitudinal Data," *Journal of the American Statistical Association*, Vol. 101, pp. 685–700, (2006).

9. Gelfand, A. E. and Smith A.F.M., "Sampling Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, 85, 398-409, (1990).

10. Greenberg E., *Introduction to Bayesian Econometrics*, Cambridge University Press, (2007).

11. Greene W. H., *Econometric Analysis*, 7th edition, Prentice Hall, (2011).

12. Hastings W.K., "Monte Carlo Sampling Methods using Markov Chains and Their Applications," *Biometrika*, Vol. 57, pp. 97-109, (1970).

13. Jeliazkov I., Graves J., and Kutzbach M., "Fitting and Comparison of Models for Multivariate Ordinal Outcomes," *Advances in Econometrics: Bayesian Econometrics*, Vol. 23, pp. 115-156, (2008).

14. Koop G. and Poirier D.J. and Tobias J.L., *Bayesian Econometric Methods*, Cambridge University Press, (2007).

15. Luce R. D., *Individual Choice Behavior*. John Wiley & Sons, (1959).

16. Luce D. and Suppes P., "Preferences, Utility and Subjective Probability," *Handbook of Mathematical Psychology*, R. D. Luce, R. Bush, and E. Galanter (eds.), John Wiley & Sons, (1965).

17. Marschak, J., "Binary-Choice Constraints and Random Utility Indicators," *Mathematical Methods in the Social Sciences*, K. J. Arrow, S. Karlin, and P. Suppes (eds.), Stanford University Press, pp. 312-329, (1960).

18. McFadden D., "Conditional Logit Analysis of Qualitative Choice Behavior," *Frontiers in Econometrics*, P. Zarembka (ed.), pp. 105-142, New: York, Academic Press, 1974.

19. Metropolis N. and Rosenbluth A.W. and Rosenbluth M.N. and Teller A.H. and Teller E., "Equations of State Calculations by Fast Computing Machines," *Journal of Chemical Physics*, Vol. 21, pp. 1087-1092, (1953).

20. Mroz T.A., "The Sensitivity of an Empirical Model of Married Women's Hours of Work to Economic and Statistical Assumptions," *Econometrica*, Vol. 55, pp. 765–799, (1987).

21. OHagan A., *Kendalls Advanced Theory of Statistics: Bayesian Inference*. John Wiley & Sons, (1994).

22. Poirier D. J., "A Curious Relationship between Probit and Logit Models," *Southern Economic Journal*, Vol. 40, pp. 640-641, (1978).

23. Poirier D. J., *Intermediate Statistics and Econometrics: A Comparative Approach*. Cambridge, MA: MIT Press, (1995).

24. Tierney L., "Markov Chains for Exploring Posterior Distributions," (with discussion), *Annals of Statistics*, Vol. 22, pp. 1701-1762, (1994).

25. Train K., *Discrete Choice Methods with Simulation*, Cambridge University Press, (2009).

26. Verlinda J. A., "A comparison of two common approaches for estimating marginal effects in binary choice models," *Applied Economics Letters*, Vol. 13, pp. 77-80, (2006).

PROBLEM SOLUTIONS

SOLUTIONS FOR CHAPTER 1

1.1 Taking the derivative of $F_L(\nu) = e^\nu/(1+e^\nu)$ and employing the product and chain rules from calculus, we get

$$\begin{aligned} f_L(\nu) &= \frac{dF_L(\nu)}{d\nu} \\ &= \frac{e^\nu}{1+e^\nu} - \frac{e^\nu e^\nu}{(1+e^\nu)^2} \\ &= \left(\frac{e^\nu}{1+e^\nu}\right)\left(1 - \frac{e^\nu}{1+e^\nu}\right) \\ &= F(\nu)\left[1 - F(\nu)\right], \end{aligned}$$

as required.

1.2 Working with expression (1.1) and letting $\beta = \beta_1 - \beta_0$, we have

$$\begin{aligned}
\Pr(y_i = 1|\beta) &= P(U_{i1} > U_{i0}) \\
&= \Pr(\varepsilon_{i0} < \varepsilon_{i1} + x_i'\beta) \\
&= \int_{-\infty}^{\infty} F_{EV}\left(\varepsilon_{i1} + x_i'\beta\right) f_{EV}(\varepsilon_{i1})\, d\varepsilon_{i1} \\
&= \int_{-\infty}^{\infty} \exp\left(-e^{-\left(\varepsilon_{i1}+x_i'\beta\right)}\right) e^{-\varepsilon_{i1}} \exp\left(-e^{-\varepsilon_{i1}}\right) d\varepsilon_{i1} \\
&= \int_{-\infty}^{\infty} \exp\left(-e^{-\varepsilon_{i1}} - e^{-\left(\varepsilon_{i1}+x_i'\beta\right)}\right) e^{-\varepsilon_{i1}}\, d\varepsilon_{i1} \\
&= \int_{-\infty}^{\infty} \exp\left(-e^{-\varepsilon_{i1}}\left(1 + e^{-x_i'\beta}\right)\right) e^{-\varepsilon_{i1}}\, d\varepsilon_{i1}
\end{aligned}$$

Letting $t = e^{-\varepsilon_{i1}}$, we have that $dt = -e^{-\varepsilon_{i1}} d\varepsilon_{i1}$. As $\varepsilon_{i1} \to \infty$, $t \to 0$ and as $\varepsilon_{i1} \to -\infty$, $t \to \infty$. Therefore, we can rewrite the integral as

$$\begin{aligned}
\Pr(y_i = 1|\beta) &= \int_{\infty}^{0} -\exp\left(-t\left(1 + e^{-x_i'\beta}\right)\right) dt \\
&= \int_{0}^{\infty} \exp\left(-t\left(1 + e^{-x_i'\beta}\right)\right) dt \\
&= -\frac{1}{1 + e^{-x_i'\beta}} \exp\left(-t\left(1 + e^{-x_i'\beta}\right)\right)\Big|_{t=0}^{\infty} \\
&= -\frac{1}{1 + e^{-x_i'\beta}}(0 - 1) \\
&= \frac{1}{1 + e^{-x_i'\beta}}
\end{aligned}$$

as required. A more general version of the proof for the case of multinomial outcomes is available in [25].

1.3 The full conditional distribution $\pi(\beta|y, z)$ is proportional to $f(z|\beta)\pi(\beta)$ and its kernel can be written as

$$\begin{aligned}
&\pi(\beta|y, z) \propto \exp\left[-\frac{1}{2}\left\{(z - X\beta)'(z - X\beta) + (\beta - \beta_0)' B_0^{-1}(\beta - \beta_0)\right\}\right] \\
&\propto \exp\left[-\frac{1}{2}\left\{-z'X\beta - \beta'X'z + \beta'X'X\beta + \beta'B_0^{-1}\beta - \beta'B_0^{-1}\beta_0 - \beta_0'B_0^{-1}\beta\right\}\right],
\end{aligned}$$

where we have omitted terms that do not involve β. Collecting terms and using the definitions of $\hat{B}$ and $\hat{\beta}$, we have that $\pi(\beta|y,z)$ is proportional to

$$\exp\left[-\frac{1}{2}\left\{\beta'\left(X'X+B_0^{-1}\right)\beta-\beta'\left(X'z+B_0^{-1}\beta_0\right)-\left(z'X+\beta_0'B_0^{-1}\right)\beta\right\}\right]$$
$$=\exp\left[-\frac{1}{2}\left\{\beta'\hat{B}^{-1}\beta-\beta'\hat{B}^{-1}\hat{\beta}-\hat{\beta}'\hat{B}^{-1}\beta\right\}\right]$$

Adding and subtracting $\hat{\beta}'\hat{B}^{-1}\hat{\beta}$ inside the curly braces, we can complete the square and write

$$\begin{aligned}\pi(\beta|y,z) &\propto \exp\left[-\frac{1}{2}\left\{\left(\beta-\hat{\beta}\right)'\hat{B}^{-1}\left(\beta-\hat{\beta}\right)-\hat{\beta}'\hat{B}^{-1}\hat{\beta}\right\}\right]\\ &\propto \exp\left[-\frac{1}{2}\left\{\left(\beta-\hat{\beta}\right)'\hat{B}^{-1}\left(\beta-\hat{\beta}\right)\right\}\right],\end{aligned}$$

where the last line follows by recognizing that $\hat{\beta}'\hat{B}^{-1}\hat{\beta}$ does not involve β and can therefore be absorbed in the constant of proportionality. The result is the kernel of the Gaussian density and hence we have shown that

$$\beta|y,z\sim N(\hat{\beta},\hat{B})$$

as required.