4

Regression Models for Ordinal Data

In the last chapter, we discussed regression models for binary data. These models were illustrated using grades of students in a statistics course. The grades themselves were not actually dichotomous—that is, they were not binary outcomes. Instead, we created dichotomous data by assigning a failing value of 0 to grades D and F, and a passing value of 1 to grades A, B, and C. By reducing the grades—which were originally recorded in five categories—to a dichotomous outcome, we threw away a significant portion of the information contained in the data. In this chapter, we explore techniques for analyzing this type of data without collapsing response categories by extending the binary regression models of Chapter 3 to the more general setting of ordered, categorical response data, or ordinal data.

Ordinal data are the most frequently encountered type of data in the social sciences. Survey data, in which respondents are asked to characterize their opinions on scales ranging from "strongly disagree" to "strongly agree," are a common example of such data. For our purposes, the defining property of ordinal data is that there exist a clear ordering of the response categories, but no underlying interval scale between them. For example, it is generally reasonable to assume an ordering of the form

$$\text{strongly disagree} < \text{disagree} < \text{don't know} < \text{agree} < \text{strongly agree},$$

but it usually does not make sense to assign integer values to these categories. Thus, statements of the type

$$\text{“disagree”} - \text{“strongly disagree”} = \text{“agree”} - \text{“don't know”}$$

are not assumed.

V.E Johnson (ed.), *Ordinal Data Modeling*

4.1 Ordinal data via latent variables

The most natural way to view ordinal data is to postulate the existence of an underlying latent (unobserved) variable associated with each response. Such variables are often assumed to be drawn from a continuous distribution centered on a mean value that varies from individual to individual. Often, this mean value is modeled as a linear function of the respondent's covariate vector.

This view of ordinal data was illustrated in Chapter 3 for the pass/fail response in the statistics class. In that example, we assumed that the logit of the pass probability could be expressed as a linear function of each student's SAT-M score. From a latent variable perspective, this model is equivalent to assuming that we can associate a latent performance variable with each student and that the distribution of this unobserved variable has a logistic distribution centered on a linear function of the student's SAT-M score. A geometric interpretation of this statement is provided in Figure 4.1. In this figure, a logistic distribution is centered on the linear predictor $\mathbf{x}'\beta$, which in this instance is presumed to be 1. If the latent variable drawn from this distribution is greater than 0, then the student is assumed to pass the course. Otherwise, the student is assumed to fail.

The same geometric interpretation extends immediately to the original grade data if we introduce additional grade cutoffs. In the pass/fail version of this model, the point 0 represented the cutoff for a passing grade. As additional categorical

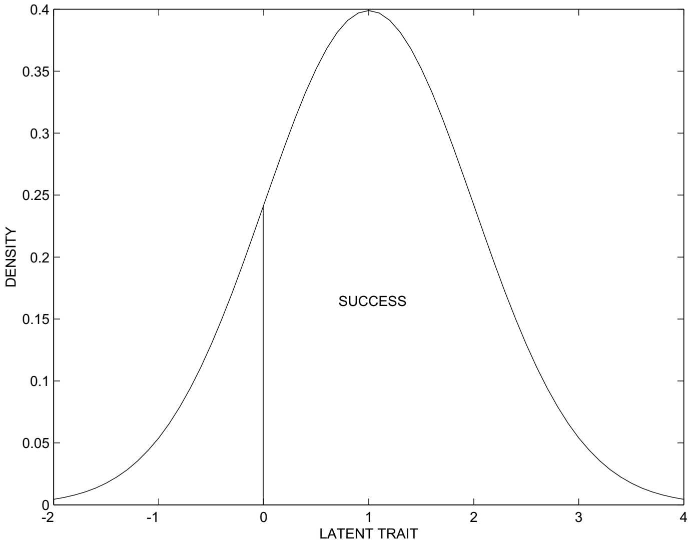

FIGURE 4.1. Latent trait interpretation of success probability. The logistic density represents the distribution of latent traits for a particular individual. In modeling the event that this individual or item is a "success," we assume that a random variable is drawn from this density. If the random variable drawn falls below 0, a failure occurs; if it falls above 0, a success is recorded.

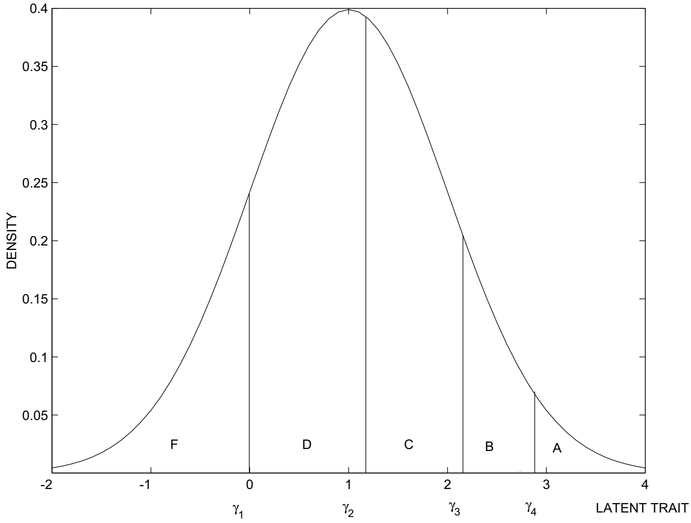

FIGURE 4.2. Latent trait interpretation of ordinal classification. In this plot, the logistic density again represents the distribution of latent traits for a particular individual. It is assumed that a random variable is drawn from this density, and the value of this random variable determines an individual's classification. For example, if a deviate of 0.5 is drawn, the individual receives a D.

responses or grade categories are introduced, we must create additional category or grade cutoffs within the model. For a class with five grades, a total of four additional grade cutoffs must be introduced. Also, because the response categories are ordered, we must impose a constraint on the values of grade cutoffs. Letting the upper grade cutoff for an F be denoted by γ_1, the upper grade cutoff for a D be denoted by γ_2, and so on, this ordering constraint may be stated mathematically as

$$-\infty < \gamma_1 \leq \gamma_2 \leq \gamma_3 \leq \gamma_4 \leq \gamma_5 \equiv \infty.$$

Note that the upper cutoff for an A, γ_5, is assumed to be unbounded. For notational convenience, we define $\gamma_0 = -\infty$.

Graphically, modeling the probability that the students in the statistics class received the grades A–F, rather than simply a pass or fail mark, requires only a minor modification of the situation depicted in Figure 4.1. Using the grade cutoffs $\gamma_1, \ldots, \gamma_4$, the expanded model is illustrated in Figure 4.2.

From Figure 4.2, we can imagine a latent variable Z that underlies the generation of the ordinal response. The distribution of this variable is governed by the equation

$$Z = \mathbf{x}'\beta + \epsilon, \tag{4.1}$$

where ϵ is a random variable drawn from a standard logistic distribution. When Z falls between the grade cutoffs γ_{c-1} and γ_c, the observation is classified into category c. To link this model for the data generation to the probability that an individual receives a particular grade, let f denote the density of the standard logistic distribution and let F denote the logistic distribution function. Denote by p_{ic} the probability that individual i receives a grade of c. Then, from (4.1), it follows that

$$\begin{aligned} p_{ic} &= \int_{\gamma_{c-1}}^{\gamma_c} f(z - x_i'\beta)dz \\ &= \Pr(\gamma_{c-1} < Z_i < \gamma_c) \\ &= F(\gamma_c - \mathbf{x}_i'\beta) - F(\gamma_{c-1} - \mathbf{x}_i'\beta). \qquad (4.2) \end{aligned}$$

The latent variable formulation of the problem thus provides a model for the probability that a student receives a particular grade in the course, or in the more general case, that a response is recorded in a particular category. If we further assume that the responses or grades for a sample of n individuals are independent of one another given these probabilities, the sampling distribution for the observed data is given by a multinomial distribution.

To specify this multinomial distribution, let us assume that there are C possible grades, denoted by $1, \ldots, C$. Also, suppose that n items are observed and that the grades or categories assigned to these n items are denoted by $y_1, \ldots y_n$; y_i denotes the grade observed for the ith individual.[1] Associated with the ith individual's response, we define a continuous latent variable Z_i and, as above, we assume that $Z_i = x_i'\beta + \epsilon_i$, where x_i is the vector of covariates associated with the ith individual and ϵ_i is distributed according to the distribution F. We observe the grade $y_i = c$ if the latent variable Z_i falls in the interval (γ_{c-1}, γ_c). Let $\mathbf{p}_i$ denote the vector of probabilities associated with assignment of the ith item into the categories $1, \ldots, C$; that is, $\mathbf{p}_i = (p_{i1}, ..., p_{iC})$, where each element of p_{ic} denotes the probability that individual i is classified into category c. Let $\mathbf{y} = (y_1, ..., y_n)$ denote the observed vector of responses for all individuals. It then follows that the probability of observing the data $\mathbf{y}$, for a fixed value of the probability vectors $\{\mathbf{p}_i\}$, is given by a multinomial density proportional to

$$\Pr(\mathbf{y} \mid \mathbf{p}_i) \propto \prod_{i=1}^{n} p_{iy_i}. \qquad (4.3)$$

[1] In defining the multinomial sampling density for an ordinal response, we assume that the multinomial denominator associated with each response is 1. For the more general case in which the ordinal responses are grouped by covariate, so that the multinomial denominator, say m_i, for the ith individual is greater than 1, this simply means that the m_i observations associated with the ith individual are considered independently in our model description. Because a multinomial observation with denominator greater than $m_i > 1$ can always be reexpressed as m_i multinomial observations with denominator 1, this distinction is irrelevant for most of the theoretical development discussed in this chapter and it somewhat simplifies notation and exposition. Of course, the likelihood function is unaffected by this change. The distinction only becomes important in defining the deviance statistic and individual deviance contributions; further comments concerning this point are delayed until Section 4.4.

Substituting the value of p_{ic} from (4.2) leads to the following expression for the likelihood function for β:

$$L(\beta,\gamma)=\prod_{i=1}^{n}[F(\gamma_{y_i}-\mathbf{x}_i'\beta)-F(\gamma_{y_i-1}-\mathbf{x}_i'\beta)]. \tag{4.4}$$

If we parameterize the model by use of the latent data $\mathbf{Z}$ along with the parameters β and γ, then the likelihood, as a function of this entire set of unknown parameters and data, may be reexpressed in terms of the latent variables $\mathbf{Z}$ as

$$L(\beta,\gamma,\mathbf{Z})=\prod_{i=1}^{n}f(Z_i-\mathbf{x}_i'\beta)I(\gamma_{y_i-1}\le Z_i<\gamma_{y_i}), \tag{4.5}$$

where $I(\cdot)$ indicates the indicator function and $\gamma=\{\gamma_1,\ldots,\gamma_C\}$ denotes the vector of grade cutoffs. Note that the latent variables Z_i may be integrated out of (4.5) to obtain (4.4).

4.1.1 Cumulative probabilities and model interpretation

Ordinal regression models are often specified in terms of cumulative probabilities rather than individual category probabilities. If we define

$$\theta_{ic}=p_{i1}+p_{i2}+\cdots+p_{ic}$$

to be the probability that the ith individual is placed in category c or below, then the regression component of an ordinal model of the form (4.2) may be rewritten

$$\theta_{ic}=F(\gamma_{ic}-\mathbf{x}_i'\beta). \tag{4.6}$$

For example, if a logistic link function is assumed, Equation (4.6) becomes

$$\log\left(\frac{\theta_{ic}}{1-\theta_{ic}}\right)=\gamma_c-\mathbf{x}_i'\beta. \tag{4.7}$$

Note that the sign of the coefficient of the linear predictor is negative, as opposed to the positive sign of this term in the usual binary regression setting.

An interesting feature of model (4.7) is that the ratio of the odds for the event $y_1\le c$ to the odds of the event $y_2\le c$ is

$$\frac{\theta_{1c}/(1-\theta_{1c})}{\theta_{2c}/(1-\theta_{2c})}=\exp[-(\mathbf{x}_1-\mathbf{x}_2)'\beta], \tag{4.8}$$

independently of the category of response, c. For this reason, (4.7) is often called the *proportional odds model* (e.g., McCullagh, 1980).

Another common regression model for ordinal data, the proportional hazards model, may be obtained by assuming a complementary log-log link in (4.2). In this case,

$$\log[-\log(1-\theta_{ic})]=\gamma_c-\mathbf{x}_i'\beta.$$

If one interprets $1-\theta_{ic}$ as the probability of survival beyond (time) category c, this model may be considered a discrete version of the *proportional hazards model* proposed by Cox (1972). Further details concerning the connection between this model and the proportional hazards model may be found in McCullagh (1980).

Another link function often used to model cumulative probabilities of success is the standard normal distribution. With such a link, (4.2) becomes

$$\Phi^{-1}(\theta_{ic}) = \gamma_c - \mathbf{x}_i'\beta.$$

This model is referred to as the *ordinal probit model*. The ordinal probit model produces predicted probabilities similar to those obtained from the proportional odds model, just as predictions from a probit model for binary data produce predictions similar to those obtained using a logistic model. However, as we will see in Section 4.3.2, the ordinal probit model possesses a property that makes sampling from its posterior distribution particularly efficient. For that reason, it may be preferred to other model links (at least in preliminary studies) if a Bayesian analysis is entertained.

4.2 Parameter constraints and prior models

An ordinal regression model with C categories and $C-1$ unknown cutoff parameters $\gamma_1, ..., \gamma_{C-1}$ is overparameterized. To see this, note that if we add a constant to every cutoff value and subtract the same constant from the intercept in the regression function, the values of $\gamma_c - \mathbf{x}_i'\beta$ used to define the category probabilities are unchanged. There are two approaches that might be taken toward resolving this identifiability problem. The first is to simply fix the value of one cutoff, usually the first. This approach was implicitly taken in Chapter 3 when we considered binary regression and defined a success as an observation for which the latent value exceeded 0. In other words, we assumed that γ_1, the upper cutoff for a failure, was 0. A second approach that can be taken for establishing identifiability of parameters is to specify a proper prior distribution on the vector of category cutoffs, γ. For ordinal data containing more than three categories, a Bayesian approach toward inference requires that a prior distribution be specified for at least one category cutoff, regardless of which approach is taken. For that reason, we now turn our discussion to prior specifications for ordinal regression models.

4.2.1 Noninformative priors

In situations where little prior information is available, the simplest way to construct a prior distribution over the category cutoffs and regression parameter is to fix the value of one cutoff, usually γ_1, at 0.[2] After fixing the value of one cutoff, a uniform prior can then be assumed for the remaining cutoffs, subject to the

[2] When the constraint $\gamma_1 = 0$ is imposed, the values of the remaining cutoffs are defined relative to this constraint, and posterior variances of category cutoffs represent the variances of the contrasts $\gamma_c - \gamma_1$.

constraint that

$$\gamma_1 \leq \cdots \leq \gamma_{C-1}.$$

The components of the category cutoff vector and the regression parameter are assumed to be a priori independent, and a uniform prior is also taken for β.

This choice of prior results in a MAP estimate of the parameter values that is identical to the MLE. In general, these point estimators provide satisfactory estimates of the multinomial cell probabilities when moderate counts are observed in all C categories. However, if there are categories in which no counts are observed or in which the number of observations is small, the MLE and MAP estimates will differ significantly from the posterior mean. Furthermore, the bias and other properties of estimators of the extreme category cutoffs may differ substantially from the corresponding properties of estimators of the interior category cutoffs.

4.2.2 Informative priors

As in the case of binary regression, specifying a prior density on the components of β and γ can be difficult. A naive method for assigning an informative prior to these parameters proceeds by assuming that the vector β of regression parameters and the vector γ of cutpoints are independent and assign independent multivariate prior distributions to each. If the dimension of the regression vector is q and the dimension of the random component of the cutpoint vector γ is b, then a distribution of dimension $q + b$ is required for the specification of such prior. However, direct specification of the joint prior on (β, γ) is problematic because of the indirect effect that these parameters have on the multinomial probabilities of interest. Moreover, the task is made more difficult by the order restriction on the components of the category cutoffs.

A more attractive method of constructing an informative prior distribution generalizes the conditional means approach of Bedrick, Christensen, and Johnson (1996) that was used in the binary regression setting of Chapter 3. In setting the prior density using the conditional means approach, prior estimates of cumulative success probabilities are specified instead of specifying prior distributions on the values of the model parameters themselves. However, to establish identifiability of parameters in the conditional means prior, at least one cumulative probability of success must be specified for each model parameter. In other words, if there are four categories of response and $\gamma_1 = 0$, at least one prior guess must be made of the cumulative probability that a response is observed to be less than or equal to the second category ($y_i \leq 2$), and at least one prior guess must be made of the probability of observing at least one response less than or equal to the third category. In addition, the design matrix selected for the covariate values (including category cutoffs) must be nonsingular.

To illustrate the conditional means approach to specifying a prior, suppose that there are $C - 2$ unknown components of the cutoff vector γ (recall that $\gamma_1 = 0$ and $\gamma_C = \infty$) and q unknown components of the regression vector β. To construct a conditional means prior for $\{\gamma, \beta\}$, we must specify $M = q + C - 2$ values of the covariate vector x—-call these covariate values $x_1, ..., x_M$. For each of the covariate vectors x_j, we specify a prior estimate and prior precision of our estimate of the corresponding cumulative probability $\theta_{(j)}$. Thus, for each covariate value, two items are specified:

1. A guess at the cumulative probability $\theta_{(j)}$ — call this guess g_j.
2. The prior precision of this guess in terms of the number of data equivalent "prior observations." Denote this prior sample size by K_j.

This prior information about $\theta_{(j)}$ can be incorporated into the model specification using a beta density with parameters $K_j g_j$ and $K_j(1-g_j)$. If the prior distributions of the cumulative probabilities $\theta_{(1)}, ..., \theta_{(M)}$ are assumed to be independent, it follows that the joint prior density is given by the product

$$g(\theta_{(1)}, ..., \theta_{(M)}) \propto \prod_{j=1}^{M} \theta_{(j)}^{K_j g_j - 1}(1-\theta_{(j)})^{K_j(1-g_j)-1}.$$

By transforming this prior on the cumulative probabilities back to (β, γ), the induced conditional means prior may be written

$$g(\beta, \gamma) \propto \prod_{j=1}^{M} \{F(\gamma_{(j)} - \mathbf{x}_j'\beta)^{K_j g_j} [1 - F(\gamma_{(j)} - \mathbf{x}_j'\beta)]^{K_j(1-g_j)} \\ \times f(\gamma_{(j)} - \mathbf{x}_j'\beta)\}, \tag{4.9}$$

subject to $\gamma_1 = 0 \leq \gamma_2 \leq \cdots \leq \gamma_{C-1}$. As before, $F(\cdot)$ denotes the link distribution function, and $f(\cdot)$ the link density.

4.3 Estimation strategies

Unlike the binary regression models of the previous chapter, maximum likelihood estimation routines for the ordinal regression models described above are often not supported in standard statistical packages. For this reason, it is useful to discuss iterative solution techniques for both maximum likelihood estimation as well as Markov chain Monte Carlo strategies for sampling from their posterior distributions of (γ, β).

4.3.1 Maximum likelihood estimation

Maximum likelihood estimates for ordinal regression models may be obtained using iteratively reweighted least squares (IRLS). An algorithm that implements IRLS can be found in the appendix at the end of this chapter. A byproduct of the algorithm is an estimate of the asymptotic covariance matrix of the MLE, which can be used to perform classical inference concerning the value of the MLE. If a uniform prior is assumed in a Bayesian analysis of the regression parameters, then the MLE and asymptotic covariance matrix obtained using IRLS also serve as approximations to the posterior mean and posterior covariance matrix.

4.3.2 MCMC sampling

In principle, the Metropolis-Hastings algorithm presented in Chapter 3 for sampling from the posterior distribution over a binary regression parameter can be adapted for sampling from the posterior distribution on the parameters in an ordinal regression model. There are, however, two important distinctions between the ordinal setting ($C > 2$) and the binary setting.

First, multivariate normal proposal densities are not well suited for generating candidate vectors for ordinal regression parameters due to the ordering constraints imposed on the components of the category cutoff vector γ. Because of this ordering constraint, candidate vectors drawn from a multivariate normal density are rejected whenever the constraint on the components of γ is not satisfied. This can make generating candidate points inefficient, and for this reason, more complicated proposal densities are generally employed.

A second problem with standard Metropolis-Hasting schemes is that the probability that a candidate point is accepted often decreases dramatically as the number of classification categories increases. This effect is caused by the small probabilities that must invariably be associated with at least some of the observed categories and by the large relative change in the values of these probabilities assigned by candidate draws to the same categories. To overcome this difficulty, hybrid Metropolis-Hastings/Gibbs algorithms are often used to sample from the posterior distribution on ordinal regression parameters. Several such algorithms have been proposed for the case of ordinal probit models; among the more notable are those of Albert and Chib (1993), Cowles (1996), and Nandram and Chen (1996). Each of these algorithms exploits the latent data formulation described in Section 4.1. Here, we describe the Cowles' algorithm. It has the advantages that it is relatively simple to implement, displays good mixing, and extends to models with arbitrary constraints on the category cutoffs.

We describe the basic algorithm for ordinal probit models with the first cutoff parameter γ_1 fixed at 0 and uniform priors taken on (β, γ), where $\gamma = (\gamma_2, \ldots, \gamma_{C-1})$. Using the latent variable representation for the likelihood function (4.5), and letting ϕ denote the standard normal density, the joint posterior density of the latent variables and model parameters is given by

$$g(\beta, \gamma, \mathbf{Z} \mid \mathbf{y}) \propto \prod_{i=1}^{n} \phi(Z_i - \mathbf{x}_i'\beta) I(\gamma_{y_i-1} \leq Z_i < \gamma_{y_i}), \tag{4.10}$$

for $-\infty < 0 < \gamma_2 < \cdots < \gamma_{C-1} < \infty$.

This representation of the joint posterior density suggests that a simple Gibbs sampling approach for simulating from the joint posterior of ($\mathbf{Z}$, β, γ) might be possible because the full-conditional posterior distributions for ($\mathbf{Z}$, β, γ),

- $g(\mathbf{Z} \mid \mathbf{y}, \beta, \gamma)$,
- $g(\beta \mid \mathbf{y}, \mathbf{Z}, \gamma)$,
- $g(\gamma \mid \mathbf{y}, \mathbf{Z}, \beta)$,

all have analytically tractable forms. The distribution of the components of **Z**, given β and γ, have independent, truncated normal distributions where the truncation points are defined by current values of the category cutoffs. If we condition on **Z** and γ, the posterior density of the regression vector β is the same as the posterior density of the regression parameter for the standard normal model when the observational variance is known to be one. Finally, the conditional distribution of the components of the category cutoffs, for example γ_c, given current values of β and **Z**, is uniformly distributed on the interval $(\max_{y_i=c-1} Z_i, \min_{y_i=c} Z_i)$. Unfortunately, when there are a large number of observations in adjacent categories, this interval tends to be very small, and movement of the components of γ will be minimal. This results in very slow mixing in Gibbs sampling schemes defined using these conditionals.

This difficulty can be resolved by means of an alternative simulation strategy for the category cutoffs. In Cowles' algorithm, this alternative strategy is obtained by partitioning the model parameters into two sets, $\{\mathbf{Z}, \gamma\}$ and β. Gibbs and MH sampling are then used to simulate in turn from the conditional distributions on the parameters in each set; that is, we alternately simulate between

- $g(\beta \mid \mathbf{y}, \mathbf{Z}, \gamma)$ and
- $g(\mathbf{Z}, \gamma \mid \mathbf{y}, \beta)$.

The conditional distribution of the regression parameter β is unchanged by this partitioning scheme and follows the multivariate normal distribution specified above. To simulate from the joint conditional posterior density of $(\mathbf{Z}, \gamma)$, we factor this density into the product

$$g(\mathbf{Z}, \gamma \mid \mathbf{y}, \beta) = g(\mathbf{Z} \mid \mathbf{y}, \gamma, \beta) g(\gamma \mid \mathbf{y}, \beta) .$$

The composition method can then be applied to simulate γ from $g(\gamma \mid \mathbf{y}, \beta)$, and the components of **Z** can be sampled from $g(\mathbf{Z} \mid \mathbf{y}, \gamma, \beta)$. As noted in Section 4.1, $g(\gamma \mid \mathbf{y}, \beta)$ can be obtained analytically by integrating (4.10) over the latent variables to obtain

$$g(\gamma \mid \mathbf{y}, \beta) \propto \prod_{i=1}^{n} [\Phi(\gamma_{y_i} - \mathbf{x}_i'\beta) - \Phi(\gamma_{y_i-1} - \mathbf{x}_i'\beta)]. \tag{4.11}$$

A Metropolis-Hastings step is used to sample from the conditional distribution of γ given **y** and β.

Steps in the hybrid Metropolis-Hastings/Gibbs sampler based on Cowles' algorithm can thus be summarized as follows.

Hybrid Metropolis-Hastings/Gibbs sampler:

0. Initialize $\beta^{(0)}$ and $\gamma^{(0)}$ (possibly to their MLE values); set $k = 1$ and $\sigma_{MH} = 0.05/C$. This value of σ_{MH} is simply a rule of thumb, and adjustments to σ_{MH} may be necessary if appropriate acceptance rates for γ are not obtained.

1. Generate a candidate **g** for updating $\gamma^{(k-1)}$:

a. For $j = 2, ..., C-1$, sample $g_j \sim N(\gamma_j^{(k-1)}, \sigma_{MH}^2)$ truncated to the interval $(g_{j-1}, \gamma_{j+1}^{(k-1)})$ (take $g_0 = -\infty$, $g_1 = 0$, and $g_C = \infty$).

b. Compute the acceptance ratio R according to

$$R = \prod_{i=1}^{n} \frac{\Phi(g_{y_i} - \mathbf{x}_i'\beta^{(k-1)}) - \Phi(g_{y_i-1} - \mathbf{x}_i'\beta^{(k-1)})}{\Phi(\gamma_{y_i}^{(k-1)} - \mathbf{x}_i'\beta^{(k-1)}) - \Phi(\gamma_{y_i-1}^{(k-1)} - \mathbf{x}_i'\beta^{(k-1)})} \times \prod_{j=2}^{C-1} \frac{\Phi((\gamma_{j+1}^{(k-1)} - \gamma_j^{(k-1)})/\sigma_{MH}) - \Phi((g_{j-1} - \gamma_j^{(k-1)})/\sigma_{MH})}{\Phi((g_{j+1} - g_j)/\sigma_{MH}) - \Phi((\gamma_{j-1}^{(k-1)} - g_j)/\sigma_{MH})}. \tag{4.12}$$

c. Set $\gamma^{(k)} = \mathbf{g}$ with probability R. Otherwise, take $\gamma^{(k)} = \gamma^{(k-1)}$. Note that the second term in (4.12) accounts for the difference in the normalization of the proposal densities on the truncated normal intervals from which candidate points are drawn, whereas the first represents the contribution from the likelihood function. If a nonuniform prior is employed for γ, R should be multiplied by the corresponding ratio of the value of the prior at the candidate vector ($\beta^{(k-1)}$, $\mathbf{g}$) to the prior evaluated at ($\beta^{(k-1)}$, $\gamma^{(k-1)}$).

2. For $i = 1, \ldots, n$, sample $Z_i^{(k)}$ from its full conditional density given $\gamma^{(k)}$ and $\beta^{(k-1)}$. The full conditional density for $Z_i^{(k)}$ is a normal density with mean $\mathbf{x}_i'\beta^{(k-1)}$ and variance one, truncated to the interval $(\gamma_{y_i-1}^{(k)}, \gamma_{y_i}^{(k)})$.

3. Sample $\beta^{(k)}$ from a multivariate normal distribution given by

$$\beta^{(k)} \sim N((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}^{(k)}, (\mathbf{X}'\mathbf{X})^{-1})). \tag{4.13}$$

4. Increment $k = k + 1$ and repeat steps (1)–(3) until a sufficient number of sampled values have been obtained.

When implementing this algorithm, the acceptance rate for the category cutoffs should be monitored in Step 1. If this rate falls too much below 0.25 or above 0.5, σ_{MH} should be decreased or increased, respectively. Further details may be found in Cowles (1996).

A similar Metropolis-Hastings/Gibbs algorithm can be specified to accommodate non-Gaussian link functions—like the proportional odds and proportional hazards models—in a straightforward way. In step (1), the function $\Phi(\cdot)$ is replaced with the appropriate link distribution F. In step (2), the latent variables are sampled from the link distribution truncated to the interval bounded by the current estimates of the category cutoffs generated in step (1). For proportional hazards and proportional odds model, this is easily accomplished using the inversion sampling method (see Chapter 2). The most significant change to the algorithm is required in step (3). Without the normal link distribution, the full-conditional distribution of the regression parameter β will generally not be in a recognizable form. However, simple random-walk Metropolis-Hastings updates for β can be made using a multivariate normal proposal density centered on the previously sampled value of β and having covariance matrix $c(\mathbf{X}'\mathbf{X})^{-1}$. Values of the constant c in the interval (0.5,1) usually result in suitable acceptance rates.

4.4 Residual analysis and goodness of fit

Associated with every multinomial observation are C categories, and an individual's response (or absence of a response) in each of these categories can be used to define a residual. For binomial data (C=2), two such residuals are $y_i - n_i p_i$ and $(n_i - y_i) - n_i(1 - p_i)$. Of course, if you know the value of the first residual—that is, if you know p_i—you can figure out the value of the second, which depends only on $(1 - p_i)$ (since y_i and n_i are assumed known). The same is true for ordinal data with C categories; if you know the values of p_{ic} for $C - 1$ of the categories, you can figure out the probability for the last, since the probabilities have to sum to 1. Thus, for ordinal data, we potentially have $C - 1$ residuals for each multinomial observation.

This increase in dimensionality, from 1 to $C - 1$, complicates residual analyses. Not only are there more residual values to be examined, but the $C - 1$ residuals from each observation are correlated. It is therefore not clear how classical residuals (e.g., Pearson, deviance, and adjusted deviance residuals) should be displayed and analyzed. In the case of Bayesian residual analyses, the standard Bayesian residual and posterior-predictive residuals both involve $(C - 1)$-dimensional distributions, which, again, complicates model criticism. One possible solution to this problem is to create a sequence of binary residuals by collapsing response categories. For example, we might redefine a "success" as exceeding the first, second, ..., or $(C - 1)$st category. The resulting binary residuals can then be analyzed using the procedures described in Chapter 3, keeping in mind that the residuals defined for each success threshold are correlated. From a practical viewpoint, the binary residuals formed using exceedence of the extreme categories (categories 1 and $(C - 1)$) are often the most informative in identifying outliers, and so attention might be focused first on these residuals.

In contrast, residuals based on the vector of latent variables $\mathbf{Z}$ do not suffer from the problem of dimensionality, since only a single latent variable is defined for each individual. The latent residual for the ith observation is defined as before by

$$r_{i,L} = Z_i - x_i'\beta.$$

Nominally, the residuals $r_{1,L}, ..., r_{n,L}$ are independently distributed as draws from the distribution F. Deviations from the model structure are reflected by atypical values of these quantities from samples drawn from F. For this reason, case analyses based on latent residuals are generally easier to perform and interpret using scalar-valued latent residuals than using residuals defined directly in terms of observed data.

To judge the overall goodness of fit of an ordinal regression model, we can use the the deviance statistic, defined as

$$D = 2 \sum_{i=1}^{n} \sum_{j=1}^{C} I(y_i = j) \log(I(y_i = j)/\hat{p}_{ij}),$$

where $\hat{p}_{ij}$ denotes the maximum likelihood estimate of the cell probability p_{ij} and I is the indicator function. In this expression, the term $I(\cdot)\log(I(\cdot)/\hat{p}_{ij})$ is assumed to be 0 whenever the indicator function is 0. The degrees of freedom associated with the deviance statistic is $n - q - (C - 2)$, where q is the number of regression parameters in the model including the intercept. Asymptotically, the deviance statistic for ordinal regression models has a χ^2 distribution only when observations are grouped according to covariate values and the expected counts in each cell become large. When only one observation is observed at each covariate value, the deviance statistic is not well approximated by a χ^2 distribution.[3]

Besides its role as a goodness-of-fit statistic, the deviance statistic can be used for model selection. Perhaps surprisingly, the distribution of differences in deviance statistics for nested models is often remarkably close to a χ^2 random variable, even for data in which the expected cell counts are relatively small. The degrees of freedom of the χ^2 random variable that approximates the distribution of the difference in deviances is equal to the number of covariates deleted from the larger model to obtain the smaller model.

Related to the model deviance are the contributions to the deviance that accrue from individual observations. In the case of binary residuals, the signed square root of these terms were used to define the deviance residuals. However, for ordinal data, it is preferable to examine the values of the deviance contribution from individual observations directly, given by

$$d_i = 2\sum_{j=1}^{C} I(y_i = j)\log(I(y_i = j)/\hat{p}_{ij}).$$

Observations that contribute disproportionately to the overall model deviance should be regarded with suspicion.[4]

[3] For grouped ordinal data, a more general definition of the deviance is needed. Letting y_{ij} denote the observed counts in category j for observation i, the deviance statistic can be defined more generally as

$$2\sum_{i=1}^{n}\sum_{j=1}^{C} y_{ij}\log(y_{ij}/\hat{y}_{ij}),$$

where $\hat{y}_{ij}$ denotes the maximum likelihood estimate of the expected cell counts y_{ij}. As the expected number of counts in each cell of every observation approaches infinity (i.e., > 3), the distribution of this form of the deviance statistic approaches a χ^2 distribution, and so should be used for assessing goodness of fit whenever it is possible to group observations.

[4] For grouped ordinal data, an alternative definition of the deviance contribution from an individual observation is

$$\frac{2}{m_i}\sum_{j=1}^{C} y_{ij}\log(y_{ij}/\hat{y}_{ij}),$$

where $m_i = \sum_{j=1}^{C} y_{ij}$.

Turning to Bayesian case analyses, posterior-predictive residuals provide a generally applicable tool by which model adequacy can be judged and outlying observations can be identified.

Posterior-predictive residuals for ordinal data models are defined using the standard prescription discussed for binary models in Section 3.4, and samples from the posterior-predictive distribution of each observation can be obtained from an existing MCMC sample sequence of β and γ values, denoted by $\{\beta^{(j)}, \gamma^{(j)}, j = 1, ..., m\}$, using the following procedure:

For $j = 1, \ldots, m$ (m the MCMC run length):

1. For $i = 1, \ldots, C$, compute $\theta_{ic}^{(j)} = F(\gamma_{ic}^{(j)} - \mathbf{x}_i' \beta^{(j)})$ and set $p_{ic}^{(j)} = \theta_{ic}^{(j)} - \theta_{i-1,c}^{(j)}$.
2. Simulate observations y_i^* from multinomial distributions with success probabilities $p_{i1}^{(j)}, ..., p_{iC}^{(j)}$.

The posterior-predictive residual distribution can then be approximated using sampled values of the quantities

$$r_{i,PP} = y_i - y_i^*.$$

As in the case of binary regression, observations for which the residual posterior-predictive distributions are concentrated away from zero represent possible outliers.

4.5 Examples

4.5.1 *Grades in a statistics class*

We return now to the analysis of grades reported in Chapter 3. For convenience, the data from this example are reproduced in Table 4.1. We begin by illustrating maximum likelihood estimation for a proportional odds model. After discussing classical model checking procedures, we then discuss Bayesian analyses using both informative and noninformative priors.

Maximum likelihood analysis

As a first step in the analysis, we assume that the logit of the probability that a student receives an ordered grade of c or worse is a linear function of his or her SAT-M score; that is, we assume a proportional odds model of the form

$$\log\left(\frac{\theta_{ic}}{1-\theta_{ic}}\right) = \gamma_c - \beta_0 - \beta_1 \times \text{SAT-M}_i. \tag{4.14}$$

Because an intercept is included in this relation, to establish identifiability we fix $\gamma_1 = 0$.

The maximum likelihood estimates and associated standard errors for the parameters γ and β are displayed in Table 4.2. The corresponding estimates of the

Student #	Grade	SAT-M score	Grade in previous statistics course
1	D	525	B
2	D	533	C
3	B	545	B
4	D	582	A
5	C	581	C
6	B	576	D
7	C	572	B
8	A	609	A
9	C	559	C
10	C	543	D
11	B	576	B
12	B	525	A
13	C	574	F
14	C	582	D
15	B	574	C
16	D	471	B
17	B	595	B
18	D	557	C
19	F	557	A
20	B	584	A
21	A	599	B
22	D	517	C
23	A	649	A
24	B	584	C
25	F	463	D
26	C	591	B
27	D	488	C
28	B	563	B
29	B	553	B
30	A	549	A

TABLE 4.1. Grades for a class of statistics students. The first column is student number. The second column lists the grade received in the class by the student, and the third and fourth columns provide the SAT-math score and grade for a prerequisite statistics course.

fitted probabilities that a student receives each of the five possible grades are plotted as a function of SAT-M score in Figure 4.3. In this figure, the white area reflects the probability that a student with a given SAT-M received an A, the lightly shaded area the probability of an B, and so on. From the plot, we see that the probability that a student with a 460 SAT-M score receives a D or F is about 57%, that a student scoring 560 on the SAT-M has approximately a 50% chance of receiving a B, and that a student who scored 660 on their SAT-M has a better than 80% chance of earning an A in the course.

It is interesting to compare the fit of the proportional odds model to the fit of the logistic regression model of Chapter 3, which was obtained by arbitrarily defining a passing mark as a grade of C or better. The maximum likelihood estimate of

Parameter	Estimate	Standard error
γ_2	2.22	0.64
γ_3	3.65	0.78
γ_4	6.51	1.33
β_0	−26.58	6.98
β_1	.0430	0.012

TABLE 4.2. Maximum likelihood estimates and standard errors for proportional odds model for statistics class grades example.

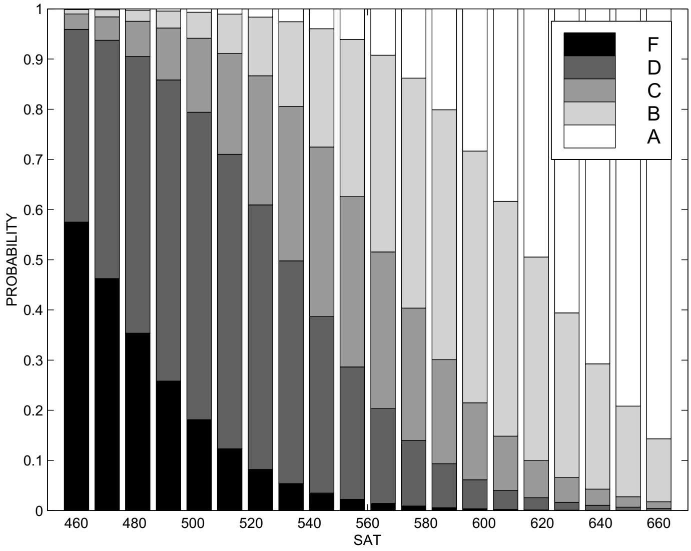

FIGURE 4.3. Fitted multinomial probabilities from the maximum likelihood fit of the proportional odds model. For each SAT value, the five shaded areas of the stacked bar chart represent the fitted probabilities of the five grades.

the probability that a student received a grade of C or higher under the logistic regression model was

$$\log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = -31.11 + 0.058 \times \text{SAT-M}_i.$$

In the proportional odds model, θ_{i2} represents the probability that a student earns a grade of D or lower, so, $1-\theta_{i2}$ is the probability that a student gets a C or higher. Under the proportional odds model, the latter probability may be expressed as

$$\log\left(\frac{1-\theta_{i2}}{\theta_{i2}}\right) = -\gamma_2 + \beta_0 + \beta_1 \times \text{SAT-M}_i.$$

Using the estimates in Table 4.2, we see that the maximum likelihood estimate of this logit is

$$\log\left(\frac{1-\hat{\theta}_{i2}}{\hat{\theta}_{i2}}\right) = -2.22 - 26.58 + 0.043 \times \text{SAT-M}_i .$$

This value is similar to that obtained using the binary regression model. The logit of the probability that a student received a C or better under the logistic model was estimated to increase at the rate of 0.058 units per SAT-M point; under the proportional odds model, this rate is 0.043. The asymptotic standard deviations of these slope estimates are 0.026 and 0.012, respectively.

The similarity of the parameter estimates obtained under the proportional odds model and the logistic model illustrates an important aspect of the ordinal modeling approach taken in this book. Due to the latent variable approach that underlies the model for both ordinal and binary responses, the interpretation of regression parameters is invariant with respect to the number of classification categories used. In the present case, the regression parameter β in the proportional odds model has the same interpretation as the regression parameter appearing in the logistic model, despite the fact that the grade categories A–C and D–F were collapsed to obtain the logistic model for binary responses. Of course, collapsing categories in this way results in some loss of information. This fact is reflected in the larger asymptotic standard deviations reported for the logistic model.

As a cursory check for model fit, we plotted the contributions to the deviance from individual observations against observation number in Figure 4.4. Interestingly, the most extreme deviant observation in the proportional odds fit appears to be student 19, rather than student 4 as it was in the logistic model. The reason for this difference becomes clear upon examining the data; student 19 received an F, while student 4 received only a D. Under the logistic model, both students were classified as failures, although student 19 apparently did much worse than student 4. Neither student's poor performance is well predicted by their SAT-M scores. It is also interesting to note that student 30's grade resulted in the second highest deviance contribution; this student had a slightly below average SAT-M score but received an A in the course. The grade of this student was not regarded as extreme in the logistic model.

For purposes of comparison, we next fit the ordinal probit model to the same data. In this case, the ordinal probit model takes the form

$$\theta_{ic} = \Phi\left(\gamma_c - \beta_0 - \beta_1 \times \text{SAT-M}_i\right). \tag{4.15}$$

As before, an intercept was included in this model since γ_1 was assigned the value 0. The maximum likelihood estimates for the probit model appear in Table 4.3.

As in the proportional odds model, one can plot the deviance contributions from each observation. The appearance of this plot was almost identical to Figure 4.4; thus, comments regarding the fit of the proportional odds model to individual student marks apply to the ordinal probit model as well. The similarity of the two deviance plots is a consequence of the fact that the fitted values under each

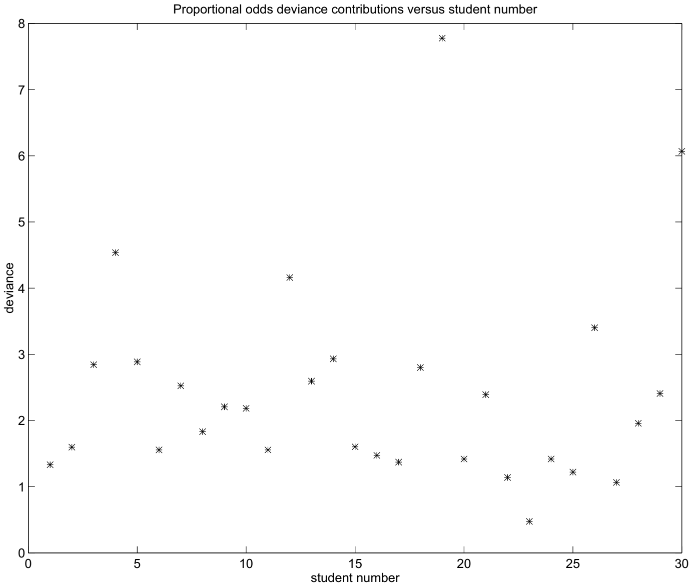

FIGURE 4.4. Deviance contributions in the proportional odds model for student grades. This plot does NOT depict deviance residuals. The square root of the deviance contributions was not taken, nor was there a natural way to attribute a sign to each observation.

Parameter	Estimate	Asy. std. dev.
γ_2	1.29	0.35
γ_3	2.11	0.41
γ_4	3.56	0.63
β_0	-14.78	3.64
β_1	0.0238	0.0063

TABLE 4.3. Maximum likelihood estimates and standard errors for ordinal probit model.

model are nearly identical. This point is illustrated in Figure 4.5, in which the predicted cell probabilities under the two models are plotted against one another. The deviance under the ordinal probit model was 73.5, while it was 72.7 under the proportional odds model.

Bayesian analysis with a noninformative prior

To further investigate the relationship between the student grades and SAT-M score, we next consider a Bayesian model using a vague prior on the parameters γ and β. Because of the similarity of fitted values obtained under the ordinal probit and proportional hazards model and the computational simplicity of sampling from the ordinal probit model using Cowles' algorithm, we restrict attention to the probit link.

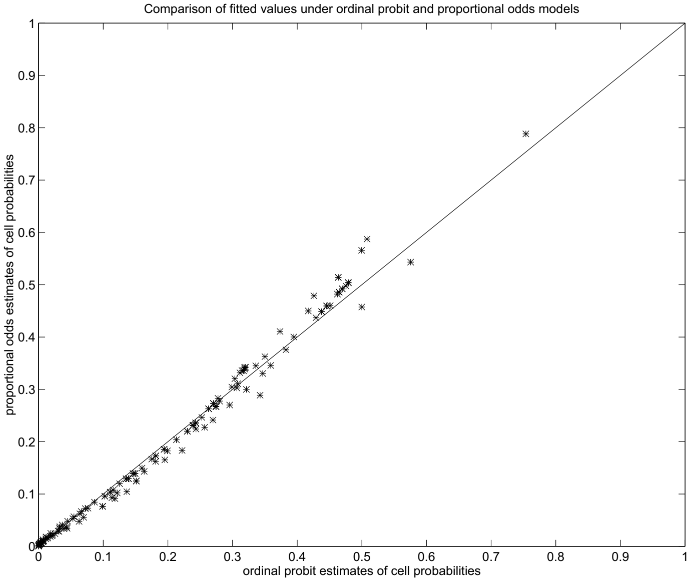

FIGURE 4.5. Fitted probabilities under ordinal probit model versus fitted probabilities for proportional odds model. All 150 predicted cell probabilities from the 30 observations are shown.

In applying Cowles' algorithm to these data, we initialized the parameter vectors with the maximum likelihood values. We then performed 20,000 MCMC iterations. The MCMC sample estimates of the posterior means of the parameter values are displayed in Table 4.4 and indicate that the posterior means agree well with the maximum likelihood estimates provided in Table 4.3. This fact suggests that the posterior distribution of the parameter estimates is approximately normal. The histogram estimates of the marginal posterior distributions displayed in Figure 4.7 support this conclusion.

A byproduct of the MCMC algorithm used to estimate the posterior means of the parameter estimates is the vector of latent variables **Z**. As discussed at the end of Section 4.4, these variables provide a convenient diagnostic for detecting outliers and assessing goodness of fit. A priori, the latent residuals $Z_1 - x_1'\beta$, ..., $Z_n - x_n'\beta$ are a random sample from a $N(0, 1)$ distribution. Thus, deviations in the values of the latent residuals from an independent sample of standard normal deviates are symptomatic of violations of model assumptions.

A normal scores plot of the posterior means of the latent residuals are depicted in Figure 4.6. There are three points that appear to fall off of the 45° line. Recall from Section 3.4 that, since the posterior means are computed by averaging the sorted latent residuals across all iterations of the simulation, the points on the graph generally do not correspond to specific observations. However, from inspection of the latent residuals from the individual iterations, the smallest latent residual did correspond to observation 19 for 91% of the iterations, the next smallest residual

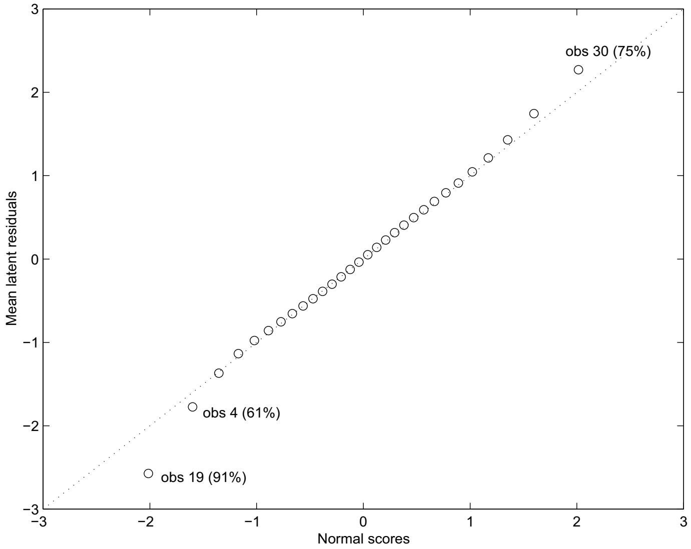

FIGURE 4.6. Normal scores plot of the posterior means of the sorted latent residuals from grades example. The labeled points indicate the percentages that particular observations contributed in the computation of the posterior mean residual.

Parameter	Post. mean	Post. std. dev.
γ_2	1.38	0.37
γ_3	2.26	0.42
γ_4	3.86	0.63
β_0	-12.05	3.73
β_1	.0257	0.0065

TABLE 4.4. Simulation estimates of the posterior means and standard errors for ordinal probit model using vague priors.

corresponded to observation 4 for 61% of the iterations, and observation 30 was the largest latent residual for 75% of the iterations. Thus, the most extreme latent residual posterior means appear to correspond to students 4, 19, and 30. With the exception of these three points, the normal scores plot does not suggest serious violations of model assumptions.

Bayesian analysis with an informative prior

We now illustrate how the methodology described in Section 4.2.2 can be used to specify a prior distribution for the parameters of the ordinal probit model for the statistics grades. Recall that in Chapter 3, we summarized our prior belief regarding the value of the regression parameter β through two prior estimates of the probability that a student received a grade of C or higher for two specified values of SAT-M. These estimates were that a student with a 500 SAT-M score

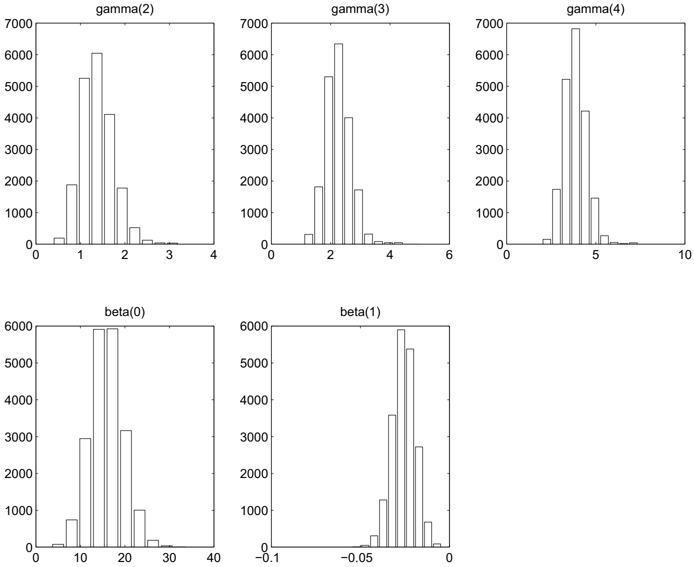

FIGURE 4.7. Histogram estimates of the marginal posterior distributions of the regression and category cutoff parameters in the statistics grades example.

would "pass" with probability 0.3 and that a student with an 600 SAT-M score would pass with probability 0.7. We also assigned a prior precision equivalent to five observations to each of these guesses. Here, we assume less certainty, and assign each guess the weight of one observation.

To specify a proper prior for all parameters in the ordinal probit model, we need to specify three more prior estimates for each of the three additional model parameters, γ_2, γ_3, and γ_4. Our three additional prior estimates are that the probability that a student having a 520 SAT-M score receives an F is 0.2, that the probability that a student with a 540 SAT-M score receives a C or lower is 0.75, and that the probability that a student with a 570 SAT-M score receives a B or lower is 0.85. As discussed above, we assign one observation's worth of information to each guess. The conditional means prior that results from these guesses can be expressed as follows:

$$\begin{aligned} g(\beta, \gamma) \propto\ & \Phi(-\beta_0 - 520\beta_1)^{0.2}(1 - \Phi(-\beta_0 - 520\beta_1))^{0.8} \\ & \times \Phi(\gamma_2 - \beta_0 - 500\beta_1)^{0.7}(1 - \Phi(\gamma_2 - \beta_0 - 500\beta_1))^{0.3} \\ & \times \Phi(\gamma_3 - \beta_0 - 540\beta_1)^{0.75}(1 - \Phi(\gamma_3 - \beta_0 - 540\beta_1))^{0.25} \\ & \times \Phi(\gamma_4 - \beta_0 - 570\beta_1)^{0.85}(1 - \Phi(\gamma_4 - \beta_0 - 570\beta_1))^{0.15} \\ & \times \Phi(\gamma_2 - \beta_0 - 600\beta_1)^{0.3}(1 - \Phi(\gamma_2 - \beta_0 - 600\beta_1))^{0.7} \\ & \times \phi(-\beta_0 - 520\beta_1)\phi(\gamma_2 - \beta_0 - 500\beta_1)\phi(\gamma_3 - \beta_0 - 540\beta_1) \\ & \times \phi(\gamma_4 - \beta_0 - 570\beta_1)\phi(\gamma_2 - \beta_0 - 600\beta_1). \end{aligned} \tag{4.16}$$

Like the analysis with a uniform prior, the posterior density that results from this prior specification is not amenable to closed-form analysis. Thus, we must again resort to MCMC methods to obtain samples from the joint posterior distribution.

The MCMC algorithm described in Section 4.3 assumed a uniform prior for (β, γ). Modifying this algorithm for application with generic priors on γ and β requires the superposition of a Metropolis-Hastings step on the Gibbs sampler in step (3), and modification of the acceptance ratio in step (1). Letting

$$s = \frac{g(\beta^{new}, \gamma^{new})}{g(\beta^{old}, \gamma^{old})}$$

denote the ratio of the prior density at the new parameter value to the old within each updating step, the required changes to Cowles' algorithm are

1. In step (1), the ratio R is multiplied by s. In this case, $\beta^{new} = \beta^{old} = \beta^{(k-1)}$, $\gamma^{old} = \gamma^{(k-1)}$, and $\gamma^{new} = \mathbf{g}$.

2. In step (3), take $\beta^{new} = \beta^{(k)}$, and $\gamma^{new} = \gamma^{old} = \gamma^{(k)}$. With probability equal to $\min(1, s)$, accept β^{new} as the new value of $\beta^{(k)}$. Otherwise, set $\beta^{(k)} = \beta^{(k-1)}$.

Posterior means and standard deviations estimated from a run of 10,000 iterates of this algorithm using the prior in (4.16) are provided in Table 4.5. Because the prior density used in this example was afforded six "prior observations," which is approximately one-fifth the weight of the data, the parameter estimates depicted in Table 4.5 differ noticeably from those obtained using a uniform prior.

To assess the effect of the informative prior information on the fitted probabilities for this model, in Table 4.6 we list the probabilities of various events using both noninformative and informative priors. The first row of this table provides the estimated probabilities that a student with a SAT-M score of 500 received a grade of D or lower in the statistics course under both informative and noninformative models. Note that the prior probability assigned to this event was 0.7. The posterior probability of this event using the noninformative prior analysis is 0.719; the corresponding posterior probability using the informative prior is 0.568. The prior information lowers this probability *below* its expected prior value, although this apparently nonlinear effect can be explained by examining the other probabilities in Table 4.6. In this table, we see that the overall effect of the informative prior is to shift the estimated posterior probabilities obtained under the noninformative posterior toward the prior estimates.

Parameter	Post. mean	Post. std. dev.
γ_2	1.09	0.28
γ_3	1.80	0.34
γ_4	2.85	0.44
β_0	−5.68	2.73
β_1	.0132	0.0048

TABLE 4.5. Simulation estimates of the posterior means and standard errors for ordinal probit model using using an informative prior.

Event	Prior prob.	Noninf. post.	Inf. post.
(500, D or lower)	.7	.719	.568
(520, F)	.2	.094	.118
(540, C or lower)	.75	.667	.638
(570, B or lower)	.85	.896	.843
(600, D or lower)	.3	.023	.125

TABLE 4.6. Prior and posterior probability estimates of particular events using noninformative and informative priors. The notation (540, D or lower) refers to the event that a student with a SAT-M score of 540 receives a grade of D or lower.

4.6 Prediction of essay scores from grammar attributes

A problem faced by large educational testing companies (e.g., ETS, ACT) involves grading thousands of student essays. As a result, there is great interest in automating the grading of student essays or—failing this—determining easily measurable qualities of essays that are associated with their ranking. The purpose of this example is to study the relationships between essay grades and essay attributes. The data in this example consist of grades assigned to 198 essays by 5 experts, each of whom rated all essays on a 10-point scale. A score of 10 indicates an excellent essay. Similar data have also been analyzed by, for example, Page (1994) and Johnson (1996). For present purposes, we examine only the grades assigned by the first expert grader, and the essay characteristics of average word and sentence length, number of words, and the number of prepositions, commas, and spelling errors.

Following a preliminary graphical analysis of the data, we chose to examine the predictive relationships between an expert's grade of an essay and the variables square root of the number of words in the essay (SqW), average word length (WL), percentage of prepositions (PP), number of commas × 100 over number of words in the essay (PC), the percentage of spelling errors (PS), and the average sentence length (SL). Plots of each of these variables versus the essay grades are displayed in Figure 4.8.

Based on the plots in Figure 4.8, we posited a baseline model of the form

$$\Phi^{-1}(\theta_{ic}) = \gamma_c - \beta_0 - \beta_1\text{WL} - \beta_2\text{SqW} - \beta_3\text{PC} - \beta_4\text{PS} - \beta_5\text{PP} - \beta_6\text{SL}, \quad (4.17)$$

where, as before, θ_{ic} denotes the cumulative probability that the ith essay received a score of c or below and Φ denotes the standard normal distribution function. The maximum likelihood estimates for this model are displayed in Table 4.7.

The deviance of model (4.17) was 748.7 on $198 - 15 = 183$ degrees of freedom, using the usual convention that the number of degrees of freedom in a generalized linear model is equal to the number of observations less the number of estimated parameters. The deviance statistic is much larger than the degrees of freedom, suggesting some overdispersion in the model. This confirms our prior intuition that the six explanatory variables in the model cannot accurately predict the grades assigned by any particular human expert. (In fact, we might expect considerable variation between the grades assigned by different experts to the same essay.)

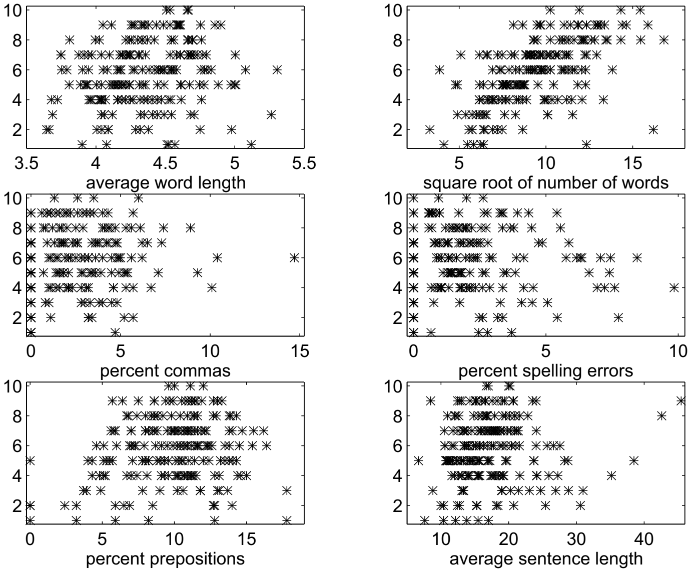

FIGURE 4.8. Plots of essay grades obtained from the first expert grader versus six explanatory variables.

Thus, to interpret the standard errors in the table, it is probably prudent to apply a correction for overdispersion. Because the usual estimate of overdispersion for ordinal regression models is deviance/degrees of freedom, in this case 4.09, each of the standard errors in Table 4.7 should be multiplied by the square root of the estimated overdispersion (≈ 2.0) to obtain a more realistic estimate of the sampling uncertainty associated with each parameter.

To investigate the source of overdispersion, it is interesting to examine the deviance contribution from each essay grade. To this end, a plot of deviance contribution versus the square root of the number of words is provided in Figure 4.9. As illustrated in the figure, the deviance of several observations exceeds 8, and the deviance for two observations exceeds 14. The values 8 and 14 correspond approximately to the 0.995 and 0.9998 points of a χ_1^2 random variable, although it is unlikely that the asymptotic distribution of either the total deviance or the deviance of individual observations is well approximated by a χ^2 random variable. However, the large values of the deviance associated with these observations provides further evidence that the grammatical variables included in the model do not capture all features used by the grader in evaluating the essays.

From Table 4.7 and the preliminary plots of the essay grades versus explanatory variables, it is clear that several of the variables included in the baseline model were not significant in predicting essay grade. To explore which of the variables should be retained in the regression function, we used a backward selection procedure in which variables were excluded sequentially from the model. The results of this procedure are summarized in the analysis of deviance table displayed in Table

Parameter	MLE	Asy. std. dev.
γ_2	0.632	0.18
γ_3	1.05	0.20
γ_4	1.63	0.21
γ_5	2.19	0.22
γ_6	2.71	0.23
γ_7	3.39	0.24
γ_8	3.96	0.26
γ_9	5.09	0.35
β_0	-3.74	1.08
β_1	0.656	0.23
β_2	0.296	0.032
β_3	0.0273	0.032
β_4	−0.0509	0.038
β_5	0.0461	0.023
β_6	0.00449	0.013

TABLE 4.7. Maximum likelihood estimates and asymptotic standard errors for the baseline regression model for essay grades.

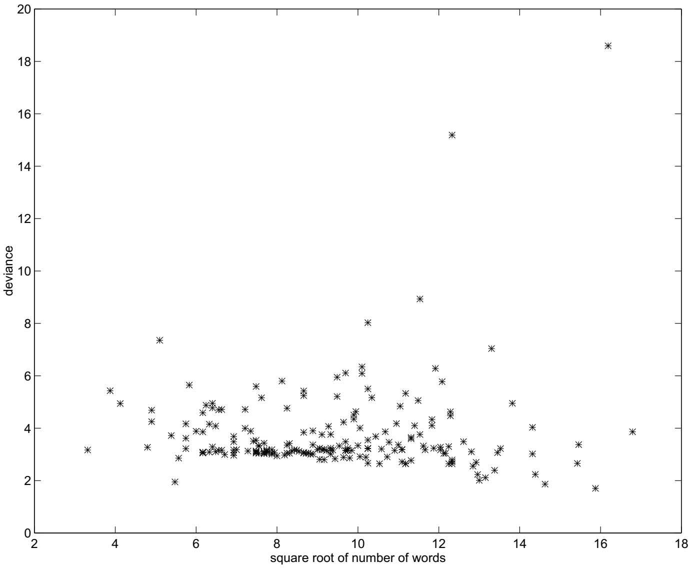

FIGURE 4.9. Deviance contribution from individual essay grades versus the square root of the number of words contained in each essay.

4.8. In this table, the entry on the third line (labeled "-PC") provides the reduction of deviance and adjusted deviance that result from deleting the variable PC from the probit model containing all covariates in the full model, excluding those already deleted, in this case SL. Both the reduction in deviance associated with

Model	Change in deviance	Corrected change in deviance
Full Model	–	–
-SL	0.12	0.03
-PC	0.70	0.17
-PS	1.77	0.43
-PP	3.98	0.97
-WL	7.90	1.93
-SqW	84.08	20.56

TABLE 4.8. AOD table for essay grades. The entries in the second column represent the increase in deviance resulting from deletion of the variable indicated in the first column, as compared to the model on the previous row. For example, the last row provides the difference in model deviance between a model containing only the category cutoffs and SqW and a model with covariates comprised of the category cutoffs, WL, and SqW. Entries in the third column represent the entries in the second columns divided by 4.09, the estimate of the model overdispersion from the full model.

the deletion of each model variable and the reduction in the deviance corrected for overdispersion are provided. As in Chapter 3, the overdispersion of the model was calculated as the model deviance divided by the degrees of freedom. Further motivation for this correction for overdispersion may be found in, for example, McCullagh and Nelder (1989).

By comparing the corrected changes in deviance displayed in the table to the corresponding tail probabilities of a χ^2_1 random variable, it appears that the variables SL, PC, and PS (average sentence length and percentage of commas and spelling errors) were not important in predicting the essay scores assigned by this grader. Likewise, the variable PP (percentage of prepositions) appears to be only marginally significant as a predictor, while the variables WL and SqW (word length and square root of number of words) are significant or highly significant. These results suggest that the variables SL, PC, and PS might be excluded from the model, leaving a predictive model of the form

$$\Phi^{-1}(\theta_{ic}) = \gamma_c - \beta_0 - \beta_1\mathrm{WL} - \beta_2\mathrm{SqW} - \beta_3\mathrm{PP}. \tag{4.18}$$

Turning next to a default Bayesian analysis of these data, if we assume a vague prior on all model parameters, we can use the MCMC algorithm described in Section 4.3.2 to sample from the posterior distribution on the parameters appearing in either the full model (4.17) or the reduced model (4.18). For purposes of illustration, we generated 5,000 iterates from the full model and used these sampled values to estimate the posterior means of the regression parameters. These estimates are provided in Table 4.9 and are quite similar to the maximum likelihood (and, in this case, maximum a posteriori) estimates listed in Table 4.7.

Bayesian case analyses based on output from the MCMC algorithm proceeds as in the previous example. By saving the latent variables values generated in the MCMC scheme, we can construct a normal scores plot of the posterior means of the latent residuals, as depicted in Figure 4.10. Like the deviance plot, this figure

Parameter	Post. mean	Post. std. dev.
γ_2	0.736	0.16
γ_3	1.19	0.18
γ_4	1.79	0.21
γ_5	2.35	0.21
γ_6	2.88	0.22
γ_7	3.59	0.22
γ_8	4.18	0.24
γ_9	5.30	0.30
β_0	−3.76	1.12
β_1	0.670	0.24
β_2	0.305	0.033
β_3	0.0297	0.033
β_4	−0.0520	0.038
β_5	0.0489	0.024
β_6	0.00463	0.013

TABLE 4.9. Posterior means of parameter estimates and standard errors for the full regression model for the essay grades.

suggests that the smallest two residuals are unusually small for this particular model. In the computation of the posterior means for these two smallest residuals, we see from Figure 4.10 that these posterior means were small primarily due to the contributions of observations 3 and 45. There is also evidence that the distribution of the latent residuals is non-Gaussian, due to the snake-like appearance of this graph.

In addition to the latent residuals, we can also examine the posterior-predictive residuals to investigate the overdispersion detected in the likelihood-based analysis. As in Chapter 3, if we let y_i^* denote a new essay grade based on the posterior-predictive distribution for covariate value $\mathbf{x}_i$ and we let y_i denote the observed grade of the ith essay, then the posterior-predictive residual distribution for the ith observation is defined as the distribution of $y_i - y_i^*$.

A plot of the estimated interquartile ranges of the posterior-predictive residuals is provided in Figure 4.11. The appearance of this plot is similar to Figure 3.12, and like Figure 3.12, it indicates model lack of fit. To more formally quantify this lack of fit, we might again posit a random effects model, but in this case there are at least two distinct sources of error which we would like to model. The first is the inability of the regression model to fully explain the nuances of human graders; the regression model clearly cannot account for all of the essay attributes used by the expert in arriving at a grade for an essay. The second is the variability between experts in assigning grades to essays. As we mentioned at the beginning of this example, there were four other experts who also assigned grades to these same essays, and there was considerable disagreement among the experts on the appropriate grade for any particular essay. Thus, a simple random effects model is unlikely to capture both sources of overdispersion, which suggests that more a comprehensive model is needed. We investigate such models in the next chapter.

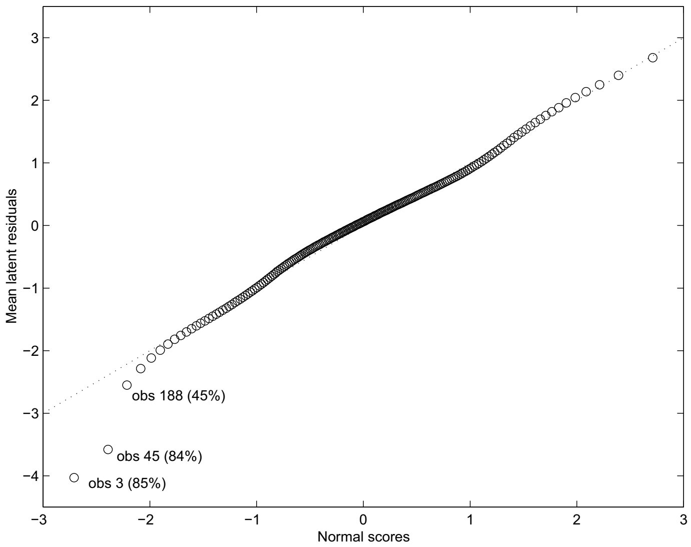

FIGURE 4.10. Normal scores plot of the posterior means of the sorted latent residuals for essay grading example. The labeled points indicate the percentages that particular observations contributed in the computation of the posterior mean residual.

4.7 Further reading

Classical ordinal regression models based on cumulative probabilities are described in McCullagh (1980) and Chapter 3 of Fahrmeir and Tutz (1994). Albert and Chib (1993), Cowles, Carlin, and Connett (1996), and Bradlow and Zaslavsky (1999) illustrate Bayesian fitting of ordinal regression models using latent variables.

4.8 Appendix: iteratively reweighted least squares

The following iteratively reweighted least squares (IRLS) algorithm can be used to find both the maximum likelihood estimate and asymptotic covariance matrix for parameters appearing in the ordinal regression models described in Chapter 4. Algorithmically, implementing IRLS requires definitions of a working dependent variable, a matrix of regressors, and regression weights at each update. To define these variables, begin by letting α denote the vector of model parameters, $(\gamma_2, \ldots, \gamma_C, \beta_0, \ldots, \beta_p)$. Note that γ_1 is not included in this vector because its value is assumed to be 0. For concreteness, assume that there are five response categories.

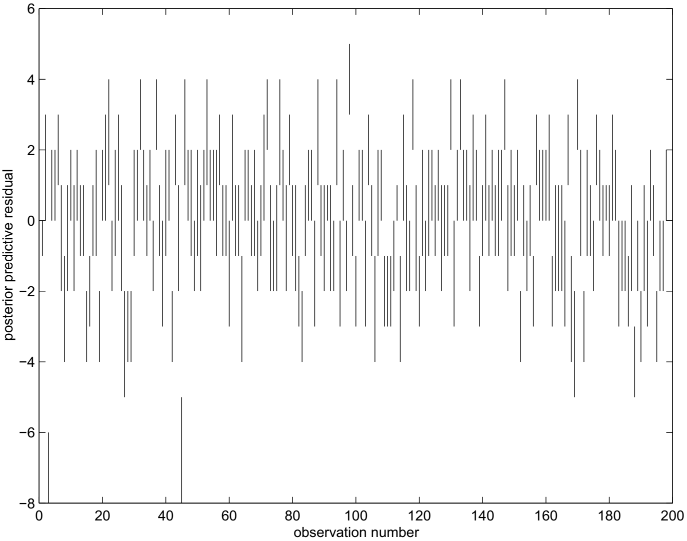

FIGURE 4.11. Interquartile ranges of the posterior predictive residuals. The fact that a high proportion of these ranges do not cover 0 is an indication of overdispersion, or other lack-of-fit.

Assuming five response categories, define

$$X_i = \begin{bmatrix} 0 & 0 & 0 & -\mathbf{x}_i' \\ 1 & 0 & 0 & -\mathbf{x}_i' \\ 0 & 1 & 0 & -\mathbf{x}_i' \\ 0 & 0 & 1 & -\mathbf{x}_i' \end{bmatrix}$$

where the covariate vectors $\mathbf{x}_i$ are preceded by a $(C-1) \times (C-1)$ identity matrix with the first column omitted (corresponding to γ_1, which is 0). Also, take

$$S_i = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & -1 \end{bmatrix}$$

to be a $C \times (C-1)$ array, and define $H_i = \text{diag}(f_{i1}, \ldots, f_{i,C-1})$, where f_{ic} is the derivative of the link distribution evaluated at $\gamma_c - \mathbf{x}_i'\beta$. Finally, take $V_i = \text{diag}(\mu_i)$, where $\mu_i = \mathbf{p}_i$, the vector of response probabilities for the ith observation.

With this admittedly tedious notation in place, the components of the IRLS algorithm can be defined as follows. Let the working dependent vector be $\mathbf{w} = (w_1', \ldots, w_n')'$, where

$$w_i = S_i H_i X_i + (\mathbf{y}_i - \mu_i),$$

and let the regression matrix be defined as

$$\mathbf{R} = (X_1' H_1 S_1', \ldots, X_n' H_n S_n')'.$$

Finally, let the weight matrix be $\mathbf{V} = \text{diag}(V_1^{-1}, \ldots, V_n^{-1})$.

To implement IRLS, each of the components of the algorithm must be initialized. The most reliable way to accomplish this initialization is to base the initial values on the observed counts $\mathbf{y}$. For example, μ_i can be initialized by taking

$$\mu_{ic} = \frac{I(y_i = c) + 1/2}{1 + C/2}.$$

Similar initializations follow for other components of the algorithm.

After initialization, a least squares estimation of the linear equation

$$\mathbf{w} = \mathbf{R}\alpha$$

with weight matrix $\mathbf{V}$ is performed recursively until changes in the regression vector α are negligible. Note that $\mathbf{w}$, $\mathbf{R}$, and $\mathbf{V}$ must all be updated using the new value of α obtained after each least squares update.

Upon convergence of the algorithm, the information matrix is

$$\mathbf{I_F} = \sum_{i=1}^{n} X_i' H_i S_i' V_i^{-1} S_i H_i X_i,$$

and $-\mathbf{I_F}^{-1}$ represents the asymptotic covariance matrix of the MLE $\hat{\alpha}$.

Further details concerning this algorithm may be found in Jansen (1991).

4.9 Exercises

1. Reconsider the educational achievement data of Exercise 5 of Chapter 3. In the dataset `educ.dat`, the response variable "years of education" is coded into the five ordered categories, where the codes 1–5 denote the categories "not completed high school", "completed high school", "had some college", "completed college", and "some graduate education", respectively.

a. Find the MLE and associate standard errors for the full model (probit link), which includes the covariates gender, race, region and parents' education. Constrast your fitted model with the estimated binary regression model found in Exercise 5 of Chapter 3.

b. Refit the ordinal model using a maximum likelihood logistic fit. Construct a scatterplot of the fitted probabilities from the logistic and probit fits. Comment on any observed differences between the two fits.

c. Use the MCMC algorithm described in Section 4.3 to sample from the posterior distribution of the ordinal regression model using a noninformative prior and a probit link. Compute the posterior means and standard deviations of the individual regression coefficients. Compare your answers with the MLE estimates and standard errors found in part (a).

d. Plot the deviance contributions from the MLE fit against the observation number to assess model fit. In addition, construct a normal probability plot of the posterior means of the sorted latent residuals from the Bayesian fit. Comment on any unusual observations that are not well explained by the fitted model.

2. Reanalyze the data of Exercise 6 of Chapter 3 without collapsing the categories of the response variable. Provide both the MLE and associated asymptotic standard errors, and the posterior means and posterior standard errors of all regression coefficients. Be sure to state and justify your prior assumptions on all model parameters. Compare the results of this analysis to the results obtained using the definition of the response specified in Exercise 6 of Chapter 3 (stored in the file `survey.dat` on web locations cited in Preface).

3. (Continuation of Exercise 2.) Fit the same linear predictor used in Exercise 2 to the grade data using a probit link. Using Cowles' algorithm to generate samples from the posterior distribution on the regression parameters in your model, plot histogram estimates of the marginal posterior distribution of each parameter.

4. (Continuation of Exercise 2.) For the final model that you selected in Exercise 2, plot the contribution to the deviance by observation against observation number. Next, compute the posterior-predictive residuals for this model. Plot the interquartile ranges of each of these residuals against their fitted values, and comment on this plot. Finally, using the same linear predictor used in the model you selected in Exercise 2, compute the posterior mean of the regression parameter for this model under a probit link, and plot the interquartile range of each latent residual distribution against fitted value. Construct a normal scores plot of the posterior mean of these latent residuals.

5. (Continuation of Exercise 2.) Provide an analysis of deviance table that includes the final model you selected in Exercise 2 and at least three other competing models. Using a normal approximation to the posterior in conjunction with Bayes theorem (see Section 2.3.1), calculate an approximate Bayes factor for each of these three competing models to the model that you selected. Be sure to specify the prior distributions used for each of the models. Compare model selection based on the AOD table to model selection using Bayes factors. For the model that you selected as "best," compute the fitted values of the response probabilities for several values of the covariates in your model. Comment. Provide an interpretation of your model coefficients and conclusions for a nonstatistician.

6. In a study of the development of Downs syndrome children, Skotko collected survey data from 55 parents concerning the extent and type of language training provided to children, along with pertinent developmental response variables. The following questions were scored on a six-point scale, in which the first category was "Not applicable," and the remaining questions ranged from "never" to "very frequently," in that order.

a. Were speech therapists/pathologists involved in your child's development before the age of 5?

b. Were speech therapists/pathologists involved in your child's development at the age of 5 or later?
c. Was sign language used with your child before the age of 5?
d. Was sign language used with your child at the age of 5 or later?
e. Were nutritional supplements used before the age of 5?
f. Were nutritional supplements used at the age of 5 or later?
g. Did you read books to your child before the age of 5?
h. Did you read books to your child at the age of 5 or later?

Interest in this study focused on the relationship between the answers to the questions above and several outcome variables, including parent response to the question "How well is your child able to maintain a conversation with a friend?" This question was scored on a 5-point scale, with categories ranging from "(1) Completely unable" to "(5) Extremely well." Treating the answer to this question as your response variable, investigate ordinal regression models that predict this response using the questions described in the preceding paragraph. Provide both the MLE and associated asymptotic standards errors, and the posterior means and posterior standard errors of all regression coefficients. Be sure to state and justify any prior assumptions you make for your model parameters and provide a substantive interpretation of your results. Data from the study is contained in the file `skotko.dat` obtained from the web site referenced in the Preface.

7. (Continuation of Exercise 6.) Conduct both classical and Bayesian case analyses for the final model you selected in the previous question. Comment on your results.

8. (Continuation of Exercise 6.) Aside from the "best" model that you selected in Exercise 6, identify at least three other competing models, at least one of which is not nested in your model, and display an AOD table for these models. Next, specify a prior distribution for the parameters in each of these models and provide approximate Bayes factors of each alternative model to the model you selected. Compare the AOD selection procedure to that based on Bayes factors.