CHAPTER 8

Models for Multinomial Responses

In Chapters 5, 6, and 7, we modeled binary response variables with *binomial* GLMs. Multicategory responses use *multinomial* GLMs. In this chapter we generalize logistic regression to handle multinomial response variables, with separate models for nominal and ordinal cases.

In Section 8.1 we present a model for nominal responses. It uses a separate binary logistic equation for each pair of response categories. In Section 8.2 we present a model for ordinal responses, using logits of cumulative response probabilities. In Section 8.3 we use other link functions for those cumulative probabilities and consider alternative ordinal logit models.

In Section 8.4 we present tests of conditional independence with multinomial responses using models and using generalizations of the Cochran–Mantel–Haenszel statistic. In Section 8.5 we introduce a multinomial logit model for *discrete-choice modeling* of a subject's choice from one of several options when values of predictors may depend on the option. The final section discusses Bayesian methods for multinomial response modeling.

8.1 NOMINAL RESPONSES: BASELINE-CATEGORY LOGIT MODELS

For a nominal-scale response variable Y with J categories, multicategory (also called *polytomous*) logistic models simultaneously describe the log odds for all $\binom{J}{2}$ pairs of categories. Given a certain choice of $J-1$ of these, the rest are redundant.

8.1.1 Baseline-Category Logits

Let $\pi_j(\boldsymbol{x}) = P(Y = j|\boldsymbol{x})$ at a fixed setting $\boldsymbol{x}$ for explanatory variables, with $\sum_j \pi_j(\boldsymbol{x}) = 1$. For observations at that setting, we treat the counts at the J categories of Y as a multinomial variate with probabilities $\{\pi_1(\boldsymbol{x}), \ldots, \pi_J(\boldsymbol{x})\}$. Logistic models pair each response category with a baseline category, such as the last one or the most common one. Consider the model

$$\log \frac{\pi_j(\boldsymbol{x})}{\pi_J(\boldsymbol{x})} = \alpha_j + \boldsymbol{\beta}_j^T \boldsymbol{x}, \quad j = 1, \ldots, J-1. \tag{8.1}$$

Categorical Data Analysis, Third Edition. Alan Agresti.

The left-hand side is the logit of a conditional probability, $\text{logit}[P(Y = j | Y = j \text{ or } Y = J)]$. This model simultaneously describes the effects of $\boldsymbol{x}$ on these $J - 1$ logits. The effects vary according to the response paired with the baseline. These $J - 1$ equations determine parameters for logits with other pairs of response categories, since

$$\log \frac{\pi_a(\boldsymbol{x})}{\pi_b(\boldsymbol{x})} = \log \frac{\pi_a(\boldsymbol{x})}{\pi_J(\boldsymbol{x})} - \log \frac{\pi_b(\boldsymbol{x})}{\pi_J(\boldsymbol{x})}.$$

With categorical predictors, X^2 and G^2 goodness-of-fit statistics provide a model check when data are not sparse. When an explanatory variable is continuous or the data are sparse, such statistics are valid only for comparing nested models differing by relatively few terms.

8.1.2 Example: Alligator Food Choice

Table 8.1 is from a study of factors influencing the primary food choice of alligators. The study captured 219 alligators in four Florida lakes. The nominal response variable is the primary food type, in volume, found in an alligator's stomach. This had five categories: fish, invertebrate, reptile, bird, other. The invertebrates included apple snails, aquatic insects, and crayfish. The reptiles were primarily turtles, although one stomach contained the tags of 23 baby alligators released in the lake the previous year! The "other" category consisted of amphibian, mammal, plant material, stones or other debris, or no food or dominant type. Table 8.1 also classifies the alligators according to L = lake of capture (Hancock, Oklawaha, Trafford, George), G = gender (male, female), and S = size ($\leq$2.3 meters long, >2.3 meters long).

Table 8.1 Primary Food Choice of Alligators, by Lake, Gender, and Size of the Alligator

			Primary Food Choice				
Lake	Gender	Size (m)	Fish	Invertebrate	Reptile	Bird	Other
Hancock	Male	≤ 2.3	7	1	0	0	5
		> 2.3	4	0	0	1	2
	Female	≤ 2.3	16	3	2	2	3
		> 2.3	3	0	1	2	3
Oklawaha	Male	≤ 2.3	2	2	0	0	1
		> 2.3	13	7	6	0	0
	Female	≤ 2.3	3	9	1	0	2
		> 2.3	0	1	0	1	0
Trafford	Male	≤ 2.3	3	7	1	0	1
		> 2.3	8	6	6	3	5
	Female	≤ 2.3	2	4	1	1	4
		> 2.3	0	1	0	0	0
George	Male	≤ 2.3	13	10	0	2	2
		> 2.3	9	0	0	1	2
	Female	≤ 2.3	3	9	1	0	1
		> 2.3	8	1	0	0	1

Source: Data courtesy of Clint Moore, from an unpublished manuscript by M. F. Delaney and C. T. Moore.

Table 8.2 Goodness of Fit of Baseline-Category Logit Models for Table 8.1 on Alligator Primary Food Choice

Model[a]	G^2	X^2	df	Collapsed over G	G^2	X^2	df
( )	116.8	106.5	60	( )	81.4	73.1	28
(G)	114.7	101.2	56				
(S)	101.6	86.9	56	(S)	66.2	54.3	24
(L)	73.6	79.6	48	(L)	38.2	32.7	16
$(L + S)$	52.5	58.0	44	$(L + S)$	17.1	15.0	12
$(G + L + S)$	50.3	52.6	40				

[a] G, gender; S, size; L, lake of capture.

Baseline-category logit models can investigate the effects of L, G, and S on primary food type. Table 8.2 contains fit statistics for several models. We denote a model by its predictors: for instance, $(L + S)$ has additive lake and size effects, and ( ) has no predictors. The data are sparse, 219 observations scattered among 80 cells. Thus, G^2 is more reliable for comparing models than for testing a model's fit. The statistics $G^2[( )|(G)] = 2.1$ and $G^2 = [(L + S)|(G + L + S)] = 2.2$, each based on df $= 4$, suggest simplifying by collapsing the table over gender. (Other analyses, not presented here, show that adding interaction terms including G do not improve the fit significantly.) The G^2 and X^2 values for reduced models for the collapsed table indicate that both L and S have effects. Table 8.3 exhibits fitted values for model $(L + S)$ for the collapsed table. Absolute values of standardized residuals comparing observed and fitted values exceed 2 in only two of the 40 cells and exceed 3 in none of the cells. The fit seems adequate.

Fish was the most common food choice. We now estimate the effects of lake and size on the odds that alligators select other primary food types instead of fish. Let $s = 1$ for size

Table 8.3 Observed and Fitted Values for Baseline-Category Logit Model Using Lake and Size of Alligator Main Effects to Predict Primary Food Choice

		Primary Food Choice				
Lake	Size of Alligator (meters)	Fish	Invertebrate	Reptile	Bird	Other
Hancock	≤ 2.3	23	4	2	2	8
		(20.9)	(3.6)	(1.9)	(2.7)	(9.9)
	> 2.3	7	0	1	3	5
		(9.1)	(0.4)	(1.1)	(2.3)	(3.1)
Oklawaha	≤ 2.3	5	11	1	0	3
		(5.2)	(12.0)	(1.5)	(0.2)	(1.1)
	> 2.3	13	8	6	1	0
		(12.8)	(7.0)	(5.5)	(0.8)	(1.9)
Trafford	≤ 2.3	5	11	2	1	5
		(4.4)	(12.4)	(2.1)	(0.9)	(4.2)
	> 2.3	8	7	6	3	5
		(8.6)	(5.6)	(5.9)	(3.1)	(5.8)
George	≤ 2.3	16	19	1	2	3
		(18.5)	(16.9)	(0.5)	(1.2)	(3.8)
	> 2.3	17	1	0	1	3
		(14.5)	(3.1)	(0.5)	(1.8)	(2.2)

Table 8.4 Estimated Parameters in Baseline-Category Logit Model for Alligator Food Choice, Based on Indicator Variable for Size (1 = Small, 0 = Large) and for Each Lake except Lake George[a]

			Lake		
Logit[b]	Intercept	Size ≤ 2.3	Hancock	Oklawaha	Trafford
$\log(\pi_I/\pi_F)$	−1.55	1.46 (0.40)	−1.66 (0.61)	0.94 (0.47)	1.12 (0.49)
$\log(\pi_R/\pi_F)$	−3.31	−0.35 (0.58)	1.24 (1.19)	2.46 (1.12)	2.94 (1.12)
$\log(\pi_B/\pi_F)$	−2.09	−0.63 (0.64)	0.70 (0.78)	−0.65 (1.20)	1.09 (0.84)
$\log(\pi_O/\pi_F)$	−1.90	0.33 (0.45)	0.83 (0.56)	0.01 (0.78)	1.52 (0.62)

[a] *SE* values in parentheses.
[b] Response categories: *I*, invertebrate; *R*, reptile; *B*, bird; *O*, other; *F*, fish.

$\leq$2.3 meters and 0 otherwise, let z_H be an indicator variable for Lake Hancock ($z_H = 1$ for alligators in that lake and 0 otherwise), and let z_O and z_T be indicator variables for Lakes Oklawaha and Trafford. With fish as the baseline category, Table 8.4 contains ML estimates of effect parameters. We use letter subscripts to denote the food choice categories. For example, the prediction equation for the log odds of selecting invertebrates instead of fish is

$$\log(\hat{\pi}_I/\hat{\pi}_F) = -1.55 + 1.46s - 1.66z_H + 0.94z_O + 1.12z_T.$$

Size of alligator has a noticeable effect. For a given lake, for small alligators the estimated odds that primary food choice was invertebrates instead of fish are $\exp(1.46) = 4.3$ times the estimated odds for large alligators; the Wald 95% confidence interval is $\exp[1.46 \pm 1.96(0.396)] = (2.0, 9.3)$. The lake effects indicate that the estimated odds that the primary food choice was invertebrates instead of fish are relatively higher at Lakes Trafford and Oklawaha and relatively lower at Lake Hancock than they are at Lake George.

The equations in Table 8.4 determine those for other food-choice pairs. For instance, for the pair (invertebrate, other),

$$\begin{aligned}\log(\hat{\pi}_I/\hat{\pi}_O) &= \log(\hat{\pi}_I/\hat{\pi}_F) - \log(\hat{\pi}_O/\hat{\pi}_F)\\ &= (-1.55 + 1.46s - 1.66z_H + 0.94z_O + 1.12z_T)\\ &\quad -(-1.90 + 0.33s + 0.83z_H + 0.01z_O + 1.52z_T)\\ &= 0.35 + 1.13s - 2.48z_H + 0.93z_O - 0.39z_T.\end{aligned}$$

Viewing all these, we see that size has its greatest impact in terms of whether invertebrates rather than fish are the primary food choice.

8.1.3 Estimating Response Probabilities

The equation that expresses multinomial logistic models directly in terms of response probabilities $\{\pi_j(\boldsymbol{x})\}$ is

$$\pi_j(\boldsymbol{x}) = \frac{\exp(\alpha_j + \boldsymbol{\beta}_j^T\boldsymbol{x})}{1 + \sum_{h=1}^{J-1}\exp(\alpha_h + \boldsymbol{\beta}_h^T\boldsymbol{x})} \tag{8.2}$$

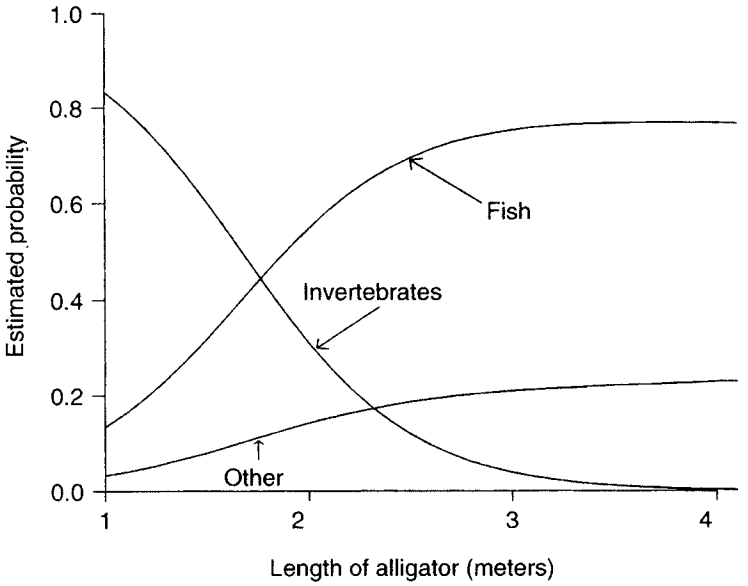

Figure 8.1 Estimated probabilities for primary food choice.

with $\alpha_J = 0$ and $\boldsymbol{\beta}_J = \mathbf{0}$. This follows from (8.1), noting that (8.1) also holds with $j = J$ by setting $\alpha_J = 0$ and $\boldsymbol{\beta}_J = \mathbf{0}$. (The parameters also equal zero for a baseline category for identifiability reasons; see Exercise 8.31.) The denominator of (8.2) is the same for each j. The numerators for various j sum to the denominator, so $\sum_j \pi_j(\boldsymbol{x}) = 1$. For $J = 2$, this formula simplifies to the binary logistic regression probability formula (5.1).

From Table 8.4 the estimated probability that a large alligator in Lake Hancock has invertebrates as the primary food choice is

$$\hat{\pi}_I = \frac{e^{-1.55-1.66}}{1 + e^{-1.55-1.66} + e^{-3.31+1.24} + e^{-2.09+0.70} + e^{-1.90+0.83}} = 0.023.$$

The estimated probabilities for (reptile, bird, other, fish) are (0.072, 0.141, 0.194, 0.570).

This example used qualitative predictors. Multinomial logit models can also contain quantitative predictors. In this study, the biologists used the size indicator variable to distinguish between adult and subadult alligators. However, the alligators' actual length was measured and is quantitative. With quantitative predictors, it is informative to plot the estimated probabilities. To illustrate, for alligators at one lake, Figure 8.1 plots the estimated probabilities that primary food choice is fish, invertebrate, or other (which combines the other, bird, and reptile categories) as a function of length. With more than two response categories, the probability for a given category need not continuously increase or decrease (Exercise 8.32).

8.1.4 Fitting Baseline-Category Logistic Models

ML fitting of multinomial logistic models maximizes the likelihood subject to $\{\pi_j(\boldsymbol{x})\}$ simultaneously satisfying the $J - 1$ equations that specify the model. Let $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iJ})$

represent the multinomial trial for subject i, where $y_{ij} = 1$ when the response is in category j and $y_{ij} = 0$ otherwise, so $\sum_j y_{ij} = 1$. Let $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^T$ denote explanatory variable values for subject i. Let $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jp})^T$ denote parameters for the jth baseline-category logit.

Since $\pi_J = 1 - (\pi_1 + \cdots + \pi_{J-1})$ and $y_{iJ} = 1 - (y_{i1} + \cdots + y_{i,J-1})$, the contribution to the log likelihood by subject i is

$$\log\left[\prod_{j=1}^{J}\pi_j(\boldsymbol{x}_i)^{y_{ij}}\right] = \sum_{j=1}^{J-1} y_{ij}\log\pi_j(\boldsymbol{x}_i) + \left(1 - \sum_{j=1}^{J-1} y_{ij}\right)\log\left[1 - \sum_{j=1}^{J-1}\pi_j(\boldsymbol{x}_i)\right]$$

$$= \sum_{j=1}^{J-1} y_{ij}\log\frac{\pi_j(\boldsymbol{x}_i)}{1 - \sum_{j=1}^{J-1}\pi_j(\boldsymbol{x}_i)} + \log\left[1 - \sum_{j=1}^{J-1}\pi_j(\boldsymbol{x}_i)\right].$$

Thus, the baseline-category logits are the natural parameters for the multinomial distribution.

Next, we construct the likelihood equations, for n independent observations. In the last expression above, we substitute $\alpha_j + \boldsymbol{\beta}_j^T\boldsymbol{x}_i$ for the logit in the first term and

$$\pi_J(\boldsymbol{x}_i) = 1\Big/\left[1 + \sum_{j=1}^{J-1}\exp(\alpha_j + \boldsymbol{\beta}_j^T\boldsymbol{x}_i)\right]$$

in the second term. Then, the log likelihood is

$$\log\prod_{i=1}^{n}\left[\prod_{j=1}^{J}\pi_j(\boldsymbol{x}_i)^{y_{ij}}\right]$$

$$= \sum_{i=1}^{n}\left\{\sum_{j=1}^{J-1} y_{ij}(\alpha_j + \boldsymbol{\beta}_j^T\boldsymbol{x}_i) - \log\left[1 + \sum_{j=1}^{J-1}\exp(\alpha_j + \boldsymbol{\beta}_j^T\boldsymbol{x}_i)\right]\right\}$$

$$= \sum_{j=1}^{J-1}\left[\alpha_j\left(\sum_{i=1}^{n} y_{ij}\right) + \sum_{k=1}^{p}\beta_{jk}\left(\sum_{i=1}^{n} x_{ik}y_{ij}\right)\right]$$

$$- \sum_{i=1}^{n}\log\left[1 + \sum_{j=1}^{J-1}\exp(\alpha_j + \boldsymbol{\beta}_j^T\boldsymbol{x}_i)\right].$$

The sufficient statistic for β_{jk} is $\sum_i x_{ik}y_{ij}$, $j = 1, \ldots, J-1, k = 1, \ldots, p$. The sufficient statistic for α_j is $\sum_i y_{ij} = \sum_i x_{i0}y_{ij}$ for $x_{i0} = 1$; this is the total number of outcomes in category j.

The likelihood equations equate the sufficient statistics to their expected values. The log-likelihood function is concave, and the Newton–Raphson method yields the ML parameter estimates. The exception is when there is a choice of baseline category such that complete or quasi-complete separation occurs for each logit when paired with another category. In that case, some estimates and *SE* values are actually infinite.

The estimators have large-sample normal distributions. As usual, standard errors are square roots of diagonal elements of the inverse information matrix.

8.1.5 Multicategory Logit Model as a Multivariate GLM

For a univariate response variable in the natural exponential family, a GLM has form $g(\mu_i) = \boldsymbol{x}_i^T\boldsymbol{\beta}$ for a link function g, expected response $\mu_i = E(Y_i)$, vector of values $\boldsymbol{x}_i$ of p explanatory variables for observation i, and parameter vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$. This extends to a *multivariate generalized linear model* for distributions in the multivariate exponential family (Exercise 8.29), such as the multinomial.

For response vector $\boldsymbol{y}_i$ for subject i, with $\boldsymbol{\mu}_i = E(\boldsymbol{Y}_i)$, let $\boldsymbol{g}$ be a vector of link functions. The multivariate GLM has the form

$$\boldsymbol{g}(\boldsymbol{\mu}_i) = \boldsymbol{X}_i\boldsymbol{\beta}, \tag{8.3}$$

where row h of the model matrix $\boldsymbol{X}_i$ for observation i contains values of explanatory variables for y_{ih} (Fahrmeir and Tutz 2001, Chap. 3).

The baseline-category logit model is a multivariate GLM. Let $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{i,J-1})^T$, since y_{iJ} is redundant, $\boldsymbol{\mu}_i = (\pi_1(\boldsymbol{x}_i), \ldots, \pi_{J-1}(\boldsymbol{x}_i))^T$ and

$$g_j(\boldsymbol{\mu}_i) = \log\{\mu_{ij}/[1 - (\mu_{i1} + \cdots + \mu_{i,J-1})]\}.$$

The model matrix for observation i is

$$\boldsymbol{X}_i = \begin{pmatrix} 1 & \boldsymbol{x}_i^T & & & & \\ & & 1 & \boldsymbol{x}_i^T & & \\ & & & & \cdots & \\ & & & & 1 & \boldsymbol{x}_i^T \end{pmatrix}$$

with 0 entries in other locations, and $\boldsymbol{\beta}^T = (\alpha_1, \boldsymbol{\beta}_1^T, \ldots, \alpha_{J-1}, \boldsymbol{\beta}_{J-1}^T)$.

8.1.6 Multinomial Probit Models

The multinomial logit model with baseline-category logits results from a latent utility representation that generalizes the one mentioned in Section 7.1.1. Let U_{ij} denote the utility of response outcome j for subject i. Suppose that

$$U_{ij} = \alpha_j + \boldsymbol{\beta}_j^T\boldsymbol{x}_i + \epsilon_{ij}.$$

The response outcome for subject i is the value of j having maximum utility. McFadden (1974) showed that the assumption that $\{\epsilon_{ij}\}$ are independent and have the extreme value distribution (i.e., cdf $F(\epsilon) = \exp[-\exp(-\epsilon)]$) is equivalent to multinomial logit model (8.1) holding. The identifiable parameters for that model are $(\boldsymbol{\beta}_j - \boldsymbol{\beta}_J)$. Likewise, the utilities are identifiable in terms of relative utilities $(U_{ij} - U_{iJ})$.

It may seem more natural to assume that $\{\epsilon_{ij}\}$ have a normal distribution. Aitchison and Bennett (1970) suggested this approach, for independent standard normal variates. The corresponding model, called the *multinomial probit model*, gives a similar fit. For a

particular explanatory variable x_k and pair of categories a and b, $(\beta_{ak} - \beta_{bk})$ describes the effect of a 1-unit increase in x_k on the difference between the mean utilities for those categories. If the normal distribution for $\{\epsilon_{ij}\}$ had instead been scaled to have some fixed standard deviation σ, then $(\beta_{ak} - \beta_{bk})$ would describe the difference in mean utilities in terms of the number of standard deviations of the utility distribution.

Fitting the multinomial probit model is computationally more complex than the corresponding logit model. Finding the likelihood function requires numerical integration, because

$$\pi_j(\boldsymbol{x}_i) = P(U_{ij} > U_{ik}, \text{ for all } k \neq j) = E_{U_{ij}}[P(U_{ik} < u_{ij}, \text{ for all } k \neq j | U_{ij} = u_{ij})]$$
$$= \int \phi(u_{ij} - \alpha_j - \boldsymbol{\beta}_j^T \boldsymbol{x}_i) \prod_{k \neq j} \Phi(u_{ij} - \alpha_k - \boldsymbol{\beta}_k^T \boldsymbol{x}_i) du_{ij},$$

for the standard normal pdf ϕ and cdf Φ.

It often seems unrealistic to expect the errors for different outcomes in the utility latent model to be uncorrelated. A more general model permits an arbitrary covariance matrix for $(\epsilon_{i1}, \ldots, \epsilon_{iJ})$, with $\text{var}(\epsilon_{i1}) = 1$ for identifiability. Fitting is then even more complex. Natarajan et al. (2000) proposed a Monte Carlo EM algorithm for ML estimation that has the advantage of circumventing direct evaluation of the likelihood function by taking advantage of the latent structure. See also Imai and van Dyk (2005) and McCulloch et al. (2000), who utilized a corresponding latent variable model introduced in Section 8.6.3.

8.1.7 Example: Effect of Menu Pricing

Natarajan et al. (2000) described a study to investigate the effect of the pricing of a fish dish in a restaurant on a customer's choice among four popular food choices. On several winter Fridays or Saturdays the fish dish was priced between \$8.95 and \$10.95. Data were collected for 974 orders. Treating the fish dish as the baseline category, the multinomial probit model provided three equations for the difference between the predicted utility for each food item and fish.

For example, the equation relating steak (the first item) to the fish dish had predicted utility difference for subject i of

$$\hat{U}_{i1} - \hat{U}_{i4} = 0.168 - 0.502F_i - 0.072P_i,$$

where $F_i = 1$ for Friday and 0 for Saturday, and P_i is the price of the fish item when subject i ordered. The standard errors were 0.178 for the Friday effect and 0.072 for the fish pricing effect. So, the fish pricing did not have a significant effect on the choice between fish and steak (higher price even having a negative estimated effect on selecting steak). Natarajan et al. (2000) used a general covariance structure for the normal errors, with $\text{var}(U_{i1} - U_{i4}) = 1.0$ for identifiability. Thus, the estimated effect of Friday was to depress the utility for steak relative to fish by half a standard distribution of the normal distribution for the utility difference.

8.2 ORDINAL RESPONSES: CUMULATIVE LOGIT MODELS

We have discussed the benefits of utilizing the ordinality of a variable by focusing inferences on a single parameter (e.g., see Section 5.3.7). These benefits extend to models for ordinal responses. Models with terms that reflect ordinal characteristics such as monotone trend have improved model parsimony and power. In this section we introduce the most popular logistic model for ordinal responses.

8.2.1 Cumulative Logits

We utilize the category ordering by forming logits of cumulative probabilities,

$$P(Y \le j|\boldsymbol{x}) = \pi_1(\boldsymbol{x}) + \cdots + \pi_j(\boldsymbol{x}), \quad j = 1, \ldots, J.$$

The *cumulative logits* are defined as

$$\begin{aligned}\text{logit}[P(Y \le j|\boldsymbol{x})] &= \log \frac{P(Y \le j|\boldsymbol{x})}{1 - P(Y \le j|\boldsymbol{x})} \\ &= \log \frac{\pi_1(\boldsymbol{x}) + \cdots + \pi_j(\boldsymbol{x})}{\pi_{j+1}(\boldsymbol{x}) + \cdots + \pi_J(\boldsymbol{x})}, \quad j = 1, \ldots, J-1. \end{aligned} \tag{8.4}$$

Each cumulative logit uses all J response categories.

8.2.2 Proportional Odds Form of Cumulative Logit Model

A model for $\text{logit}[P(Y \le j)]$ alone is an ordinary logistic model for a binary response in which categories 1 to j form one outcome and categories $j+1$ to J form the second. A model that simultaneously uses all $(J-1)$ cumulative logits in a single parsimonious model is

$$\text{logit}[P(Y \le j|\boldsymbol{x})] = \alpha_j + \boldsymbol{\beta}^T \boldsymbol{x}, \quad j = 1, \ldots, J-1. \tag{8.5}$$

Each cumulative logit has its own intercept. The $\{\alpha_j\}$ are increasing in j, because $P(Y \le j|\boldsymbol{x})$ increases in j for fixed $\boldsymbol{x}$ and the logit is an increasing function of $P(Y \le j|\boldsymbol{x})$.

This model assumes the same effects $\boldsymbol{\beta}$ for each logit. For a single continuous predictor x, Figure 8.2 depicts the model when $J = 4$. For fixed j, the response curve is a logistic

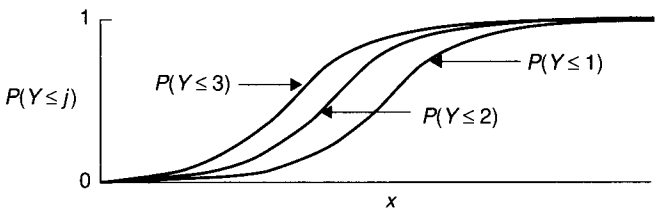

Figure 8.2 Cumulative logit model with the same effect on each of three cumulative probabilities in a four-category response.

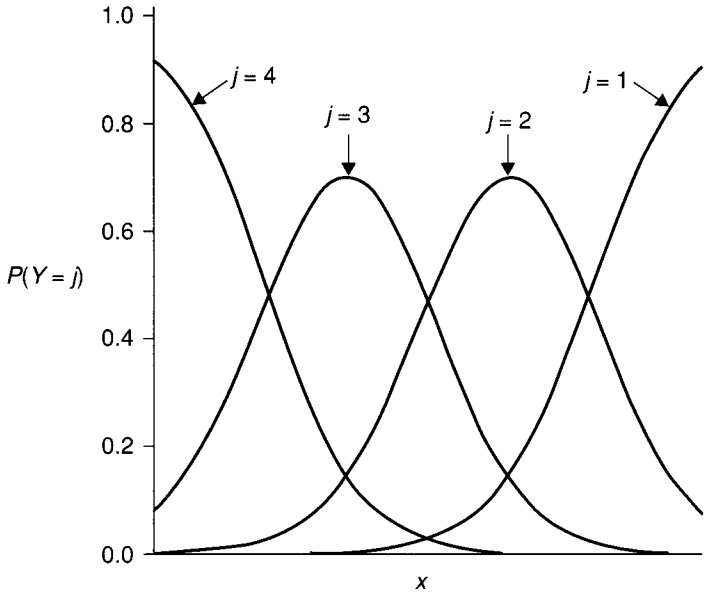

Figure 8.3 Individual category probabilities in cumulative logit model with four response categories.

regression curve for a binary response with outcomes $(Y \leq j)$ and $(Y > j)$. The curves for $j = 1, 2$, and 3 have the same shape. They share exactly the same rate of increase or decrease but are horizontally displaced from each other. Figure 8.3 portrays the corresponding curves for the category probabilities.

The cumulative logit model (8.5) satisfies

$$\begin{aligned}&\text{logit}[P(Y \leq j|\boldsymbol{x}_1)] - \text{logit}[P(Y \leq j|\boldsymbol{x}_2)] \\ &= \log \frac{P(Y \leq j|\boldsymbol{x}_1)/P(Y > j|\boldsymbol{x}_1)}{P(Y \leq j|\boldsymbol{x}_2)/P(Y > j|\boldsymbol{x}_2)} = \boldsymbol{\beta}^T(\boldsymbol{x}_1 - \boldsymbol{x}_2).\end{aligned}$$

An odds ratio of cumulative probabilities is called a *cumulative odds ratio*. The odds of making response $\leq j$ at $\boldsymbol{x} = \boldsymbol{x}_1$ are $\exp[\boldsymbol{\beta}^T(\boldsymbol{x}_1 - \boldsymbol{x}_2)]$ times the odds at $\boldsymbol{x} = \boldsymbol{x}_2$. The log cumulative odds ratio is proportional to the distance between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. The same proportionality constant applies to each logit. Because of this property, model (8.5) is often called a *proportional odds model* (McCullagh 1980).

With a single predictor, the cumulative odds ratio equals e^{β} whenever $x_1 - x_2 = 1$. Figure 8.4 illustrates the constant cumulative odds ratio this model then implies for all j. It shows the J-category response collapsed into the binary outcome $(\leq j, > j)$ and shows the sets of cells that determine the cumulative odds ratio that takes the same value e^{β} for each such collapsing.

Model (8.5) constrains the $J - 1$ response curves to have the same shape. For multicategory indicator $(y_{i1}, \ldots, y_{iJ})$ of the response for subject i, the product multinomial

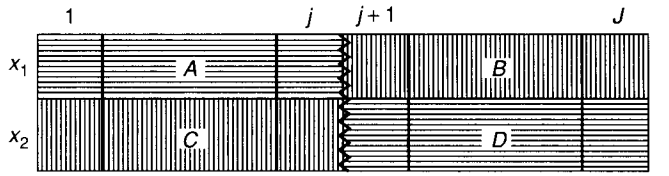

Figure 8.4 Uniform odds ratios AD/BC whenever $x_1 - x_2 = 1$, for all binary collapsings of the response in cumulative logit model of proportional odds form.

likelihood function is

$$\prod_{i=1}^{n}\left[\prod_{j=1}^{J}\pi_j(\boldsymbol{x}_i)^{y_{ij}}\right] = \prod_{i=1}^{n}\left\{\prod_{j=1}^{J}\left[P(Y \le j|\boldsymbol{x}_i) - P(Y \le j-1|\boldsymbol{x}_i)\right]^{y_{ij}}\right\}$$

$$= \prod_{i=1}^{n}\left\{\prod_{j=1}^{J}\left[\frac{\exp(\alpha_j + \boldsymbol{\beta}^T\boldsymbol{x}_i)}{1+\exp(\alpha_j + \boldsymbol{\beta}^T\boldsymbol{x}_i)} - \frac{\exp(\alpha_{j-1} + \boldsymbol{\beta}^T\boldsymbol{x}_i)}{1+\exp(\alpha_{j-1} + \boldsymbol{\beta}^T\boldsymbol{x}_i)}\right]^{y_{ij}}\right\}, \tag{8.6}$$

viewed as a function of $(\{\alpha_j\}, \boldsymbol{\beta})$. This can be maximized to obtain the ML estimates using Fisher scoring (McCullagh 1980, Walker and Duncan 1967) or the Newton–Raphson method. The *SE* values differ somewhat, as the expected information and observed information matrices are not the same for this non-canonical-link model.

8.2.3 Latent Variable Motivation for Proportional Odds Structure

A regression model for a latent continuous variable assumed to underlie Y motivates the common effect $\boldsymbol{\beta}$ for different j in the proportional odds form of the model (Anderson and Philips 1981). Let Y^* denote this underlying latent variable. Suppose that it has cdf $G(y^* - \eta)$, where values of y^* vary around a location parameter η (such as a mean) that depends on $\boldsymbol{x}$ through $\eta(\boldsymbol{x}) = \boldsymbol{\beta}^T\boldsymbol{x}$. Suppose that the thresholds $-\infty = \alpha_0 < \alpha_1 < \cdots < \alpha_J = \infty$ are *cutpoints* of the continuous scale such that the observed response y satisfies

$$y = j \quad \text{if } \alpha_{j-1} < y^* \le \alpha_j.$$

That is, y falls in category j when the latent variable falls in the jth interval of values, as Figure 8.5 depicts. Then

$$P(Y \le j|\boldsymbol{x}) = P(Y^* \le \alpha_j|\boldsymbol{x}) = G(\alpha_j - \boldsymbol{\beta}^T\boldsymbol{x}).$$

The appropriate model for Y implies that the link function G^{-1}, the inverse of the cdf for Y^*, applies to $P(Y \le j|\boldsymbol{x})$. If $Y^* = \boldsymbol{\beta}^T\boldsymbol{x} + \epsilon$, where the cdf G of ε is the standard logistic (Section 4.2.5), then G^{-1} is the logit link and a proportional odds model results. Normality for ε implies a probit link for cumulative probabilities (Section 8.3.2).

In this derivation, the same parameters $\boldsymbol{\beta}$ occur for the effects regardless of how the cutpoints $\{\alpha_j\}$ chop up the scale for the latent variable. The effect parameters are invariant to the choice of categories for Y. If a continuous variable measuring political ideology has a

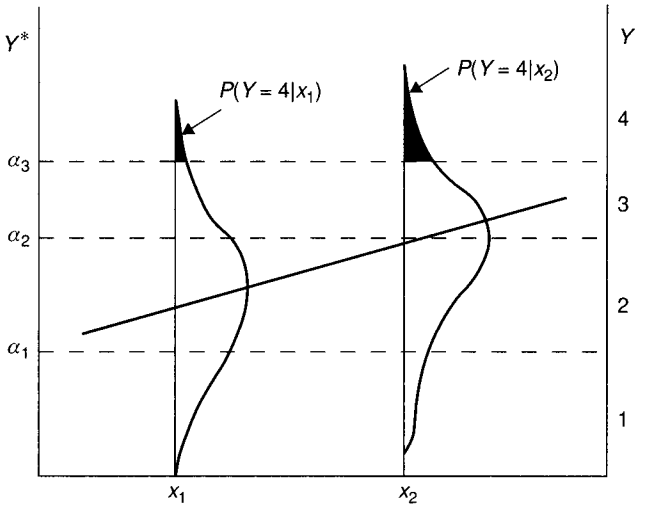

Figure 8.5 Ordinal measurement and underlying regression model for a latent variable.

linear regression with some predictor variables, then the same $\boldsymbol{\beta}$ apply to a discrete version of political ideology with the categories (liberal, moderate, conservative) or (very liberal, slightly liberal, moderate, slightly conservative, very conservative). This feature makes it possible to compare estimates from studies using different response scales.

Using a cdf of form $G(y^* - \eta)$ for the latent variable resulted in linear predictor $\alpha_j - \boldsymbol{\beta}^T \boldsymbol{x}$ rather than $\alpha_j + \boldsymbol{\beta}^T \boldsymbol{x}$. When $\beta_k > 0$, as x_{ik} increases each cumulative logit then decreases, so each cumulative probability decreases and relatively less probability mass falls at the low end of the Y scale. Thus, Y tends to be larger at higher values of x_{ik}. With this parameterization the sign of β_k has the usual meaning. However, some software (e.g., SAS) uses form (8.5).

8.2.4 Example: Happiness and Traumatic Events

Table 8.5 shows GSS data on Y = happiness (categories 1 = very happy, 2 = pretty happy, 3 = not too happy), x_1 = total number of traumatic events that happened to the respondent and his/her relatives in the last year, and x_2 = race (1 = black, 0 = white). We restricted

Table 8.5 Four Observations from Data Set on Happiness, Number of Traumatic Events, and Race

Observation	Happiness	Number of Traumatic Events	Race
1	Pretty happy	2	White
2	Pretty happy	3	Black
3	Very happy	0	White
4	Not too happy	5	White

Source: 1984 General Social Survey; complete data at `www.stat.ufl.edu/~aa/cda/cda.html`.

the age range to 18–22 in order to have a relatively small sample ($n = 97$), to illustrate how certain models may then have infinite ML estimates. In particular, only 13 of the 97 observations were in the black category of race, and of them, none had response in the very happy category. Table 8.5 shows the data for four of the subjects. The complete data set is at the text website.

The main-effects cumulative logit model of proportional odds form (8.5) is

$$\text{logit}[P(Y \leq j|\boldsymbol{x})] = \alpha_j + \beta_1 x_1 + \beta_2 x_2.$$

Table 8.6 shows output. With $J = 3$ response categories, the model has two $\{\alpha_j\}$ intercepts. Usually, these are not of interest except for computing response probabilities. The parameter estimates yield estimated logits and hence estimates of $P(Y \leq j)$, $P(Y > j)$, or $P(Y = j)$. We illustrate for white subjects ($x_2 = 0$) at the mean number of traumatic events score of $x_1 = 1.536$. Since $\hat{\alpha}_1 = -0.518$, the estimated probability of response *very happy* is

$$\hat{P}(Y = 1) = \hat{P}(Y \leq 1) = \frac{\exp[-0.518 - 0.406(1.536)]}{1 + \exp[-0.518 - 0.406(1.536)]} = 0.24.$$

Figure 8.6 plots $\hat{P}(Y \leq 2)$ as a function of the number of traumatic events, at the two levels of race. An alternative way to portray the model is to plot the parallel straight lines for the fit in terms of the underlying latent variable.

The effect estimates $\hat{\beta}_1 = -0.406$ and $\hat{\beta}_2 = -2.036$ suggest that the cumulative probability starting at the very happy end of the happiness scale decreases as the traumatic events score increases and is lower for blacks than for whites. For example, given the traumatic events score, for whites the estimated odds of reporting being very happy were $e^{2.036} = 7.7$ times the estimated odds for blacks. This estimate is imprecise, because relatively few observations were in the black category. The 95% profile likelihood confidence interval for $-\beta_2$ is (0.72, 3.43), corresponding to (2.05, 30.84) for the odds ratio effect. The *SE* values reported are based on the expected information from Fisher scoring. Using observed information (from Newton–Raphson), $\hat{\beta}_1$ and $\hat{\beta}_2$ have *SE* values of 0.183 and 0.686 instead of 0.181 and 0.691.

Descriptions of effects can compare cumulative probabilities rather than use odds ratios. These can make it easier to conceptualize the sizes of effects. We describe effects of quantitative variables by comparing probabilities at their extreme values or at their

Table 8.6 Software Output (Based on SAS) for Fitting Cumulative Logit Model to Data on Happiness

```
Score Test for the Proportional Odds Assumption
Chi-Square   DF          Pr > ChiSq
  0.8668      2              0.6483

                        Std      Like. Ratio 95%      Chi-
Parameter    Estimate   Error      Conf Limits       Square   Pr > ChiSq
Intercept1   -0.5181    0.3382   -1.2020    0.1392     2.35      0.1255
Intercept2    3.4006    0.5648    2.3779    4.6266    36.25      <.0001
traumatic    -0.4056    0.1809   -0.7729   -0.0520     5.03      0.0249
race         -2.0361    0.6911   -3.4287   -0.7156     8.68      0.0032
```

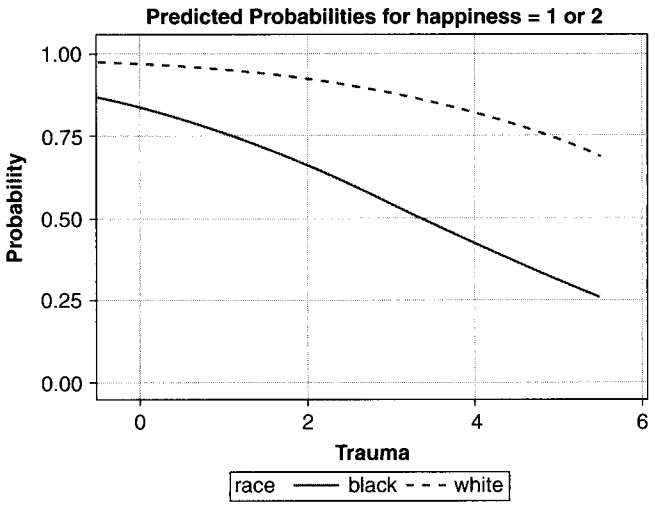

Figure 8.6 Estimated values of $P(Y \leq 2)$ by x_1 = number of traumatic events and x_2 = race.

quartiles. We describe effects of qualitative variables by comparing probabilities for different categories. We fix values of quantitative variables by setting them at their mean or median. For qualitative variables we fix the category, unless there are several, in which case we can set each at their indicator means.

We illustrate again with $P(Y = 1)$, the *very happy* outcome. First, we describe the race effect. At the mean number of traumatic events of 1.536, $\hat{P}(Y = 1) = 0.04$ for blacks (i.e., $x_2 = 1$) and 0.24 for whites ($x_2 = 0$). Next, we describe the number of traumatic events effect. The minimum and maximum values were 0 and 5. For blacks, $\hat{P}(Y = 1)$ changes from 0.07 to 0.01 between these values; for whites, it changes from 0.37 to 0.07. (Note that comparing 0.07 to 0.37 at the minimum and 0.01 to 0.07 at the maximum provides further information about the race effect.) The sample effect is substantial for both predictors. However, these summaries are highly tentative and have large standard errors, because the black sample had only 13 observations, of whom none reported more than 3 traumatic events.

8.2.5 Checking the Proportional Odds Assumption

Models in this section used the proportional odds assumption of the same effects for different cumulative logits. An advantage is that effects are simple to summarize, requiring only a single parameter for each predictor. The models generalize to include separate effects, replacing $\boldsymbol{\beta}$ in (8.5) by $\boldsymbol{\beta}_j$. This implies nonparallelism of curves for different logits. However, curves for different cumulative probabilities may then cross for some $\boldsymbol{x}$ values. Such models then violate the proper order among the cumulative probabilities (Exercise 8.37).

Even if such a model fits better over the observed range of $\boldsymbol{x}$, for reasons of parsimony the simpler model might be preferable. One case is when effects $\{\hat{\boldsymbol{\beta}}_j\}$ with different logits are not substantially different in practical terms. Then the significance in a test of proportional odds may reflect primarily a large value of n. Even with smaller n, although effect estimators

using the simpler model are biased, they may have smaller MSE than estimators from a model having many more parameters. So even if a test of proportional odds has a small P-value, don't discard this model automatically.

The output[1] in Table 8.6 also presents a score test of the proportional odds property. This tests whether the effects are the same for each cumulative logit against the alternative of separate effects. It compares the model with one parameter for x_1 and one for x_2 to the more complex model with two parameters for each, allowing different effects for $\text{logit}[P(Y \leq 1)]$ and $\text{logit}[P(Y \leq 2)]$. Here, the score statistic equals 0.87. It has df = 2, since the more complex model has two additional parameters. The more complex model does not fit significantly better ($P = 0.65$).

When this score test has a small P-value, it's helpful to check whether the violation of the proportional odds property is substantively important, by comparing estimates obtained from separate logistic fits to the binary collapsings of the response. For these data, consider the effect of the number of traumatic events. The model with binary response categories (very happy, pretty happy or not too happy) has $\hat{\beta}_1 = -0.339$ ($SE = 0.213$), whereas the model with binary categories (very happy or pretty happy, not too happy) has $\hat{\beta}_1 = -0.487$ ($SE = 0.276$). The effect has the same direction and a similar magnitude in each case, and it is sensible to use the simpler proportional odds structure. There is less information in the data about the race effect. We obtain $\hat{\beta}_2 = -1.846$ ($SE = 0.763$) for the second collapsing but $\hat{\beta}_2 = -\infty$ for the first collapsing because there were no observations for blacks in the very happy category and there is quasi-complete separation for that logit.

If a proportional odds model fits poorly in terms of practical as well as statistical significance, alternative strategies exist. These include (1) adding additional terms, such as interactions, to the linear predictor; (2) trying a link function for which the response curve is nonsymmetric (e.g., complementary log–log); (3) using an alternative ordinal model for which the more complex non-proportional-odds form is also valid; (4) adding dispersion parameters; (5) permitting separate effects for each logit for some but not all predictors (i.e., *partial proportional odds*); and (6) fitting baseline-category logit models and using the ordinality in an informal way in interpreting the associations.

For approach (1), more complex cumulative logit models are formulated as in ordinary logistic regression. For the example on modeling happiness, permitting interaction yields a model with ML fit

$$\text{logit}[\hat{P}(Y \leq j|\boldsymbol{x})] = \hat{\alpha}_j - 0.469x_1 - 3.057x_2 + 0.608(x_1x_2),$$

where the coefficient of x_1x_2 has $SE = 0.601$. The estimated effect of the number of traumatic events on the cumulative logit is -0.469 for whites and $(-0.469 + 0.608) = 0.139$ for blacks. The impact of the number of traumatic events may be quite different (and possibly nonexistent) for blacks, but recall that the black sample had only 13 observations, and here the difference in effects is not significant.

In the next section we generalize the cumulative logit model to permit extension (2) of alternative link functions. In Sections 8.3.4 and 8.3.6 we introduce models that satisfy option (3). Section 8.3.8 and Note 8.8 discuss extension (4). For approach (5), see Peterson and Harrell (1990), Stokes et al. (2012), and criticism by Cox (1995). Agresti (2010, Chap. 3–5) discussed further these alternative strategies.

[1] Obtained using PROC LOGISTIC in SAS.

8.3 ORDINAL RESPONSES: ALTERNATIVE MODELS

Cumulative logit models use the logit link. As in binary GLMs, other link functions are possible. In this section we introduce models having alternative link functions either for cumulative probabilities or other response probabilities.

8.3.1 Cumulative Link Models

Let G^{-1} denote a link function that is the inverse of the continuous cdf G (recall Section 4.2.5). The *cumulative link* model

$$G^{-1}[P(Y \leq j|\boldsymbol{x})] = \alpha_j + \boldsymbol{\beta}^T\boldsymbol{x} \tag{8.7}$$

links the cumulative probabilities to the linear predictor. The logit link function $G^{-1}(u) = \log[u/(1-u)]$ is the inverse of the standard logistic cdf.

As in the cumulative logit model with proportional odds form (8.5), effects of $\boldsymbol{x}$ in (8.7) are the same for each cumulative probability. In Section 8.2.3 we showed that this assumption holds whenever a latent variable Y^* satisfies a linear regression model with standard cdf G for the error term. Model (8.7) results from discrete measurement of Y^* from a location-parameter family having cdf $G(y^* - \boldsymbol{\beta}^T\boldsymbol{x})$. The parameters $\{\alpha_j\}$ are category cutpoints (or "thresholds") on a standardized version of the latent scale. Thus, we can regard cumulative link models as regression models that use a linear predictor $\boldsymbol{\beta}^T\boldsymbol{x}$ to describe effects of explanatory variables on crude ordinal measurement of Y^*. Using $-\boldsymbol{\beta}$ rather than $+\boldsymbol{\beta}$ in the linear predictor merely results in a change of sign of $\hat{\boldsymbol{\beta}}$.

8.3.2 Cumulative Probit and Log–Log Models

The *cumulative probit model* is the cumulative link model using the standard normal cdf Φ for G. This generalizes the binary probit model (Section 7.1) to ordinal responses. It is appropriate when the conditional distribution for the latent variable Y^* is normal. Parameters in probit models refer to effects on $E(Y^*)$. For instance, consider the model $\Phi^{-1}[P(Y \leq j)] = \alpha_j - \beta x$. From Section 8.2.3, since $Y^* = \beta x + \epsilon$ where $\epsilon \sim N(0, 1)$ has cdf Φ, a 1-unit increase in x corresponds to a β increase in $E(Y^*)$. When ε need not be in standard form with $\sigma = 1$, a 1-unit increase in x corresponds to a β standard deviation increase in $E(Y^*)$.

Cumulative probit models provide fits similar to cumulative logit models. They have smaller estimates and standard errors because the standard normal distribution has standard deviation 1.0 compared with 1.81 for the standard logistic.

An underlying extreme value distribution for Y^* implies the model

$$\log\{-\log[1 - P(Y \leq j|\boldsymbol{x})]\} = \alpha_j + \boldsymbol{\beta}^T\boldsymbol{x}.$$

In Section 7.1 we introduced this *complementary log–log link* for binary data. The ordinal model using this link is sometimes called a *proportional hazards* model since it results from a generalization of the proportional hazards model for survival data to handle grouped survival times (McCullagh 1980, Note 8.6). It has the property

$$P(Y > j|\boldsymbol{x}_1) = [P(Y > j|\boldsymbol{x}_2)]^{\exp[\boldsymbol{\beta}^T(\boldsymbol{x}_1 - \boldsymbol{x}_2)]}.$$

With this link, $P(Y \leq j)$ approaches 1.0 at a faster rate than it approaches 0.0. The related *log–log link* $\log\{-\log[P(Y \leq j)]\}$ is appropriate when the complementary log–log link holds for the categories listed in reverse order.

McCullagh (1980) and Thompson and Baker (1981) treated cumulative link models as multivariate GLMs. McCullagh presented a Fisher scoring algorithm for ML estimation. He showed that sufficiently large n guarantees a unique maximum of the likelihood. Burridge (1981) and Pratt (1981) showed that the log likelihood is concave for many cumulative link models, including the logit, probit, and complementary log–log. Iterative algorithms usually converge rapidly to the ML estimates.

8.3.3 Example: Happiness Revisited with Cumulative Probits

In Section 8.2.4 we modeled Y = happiness in terms of x_1 = total number of traumatic events that happened to the respondent and his/her relatives in the last year, and x_2 = race. The cumulative logit model gave the fit

$$\text{logit}[\hat{P}(Y \leq j)] = \hat{\alpha}_j - 0.406x_1 - 2.036x_2.$$

The corresponding cumulative probit model has fit

$$\Phi^{-1}[\hat{P}(Y \leq j)] = \hat{\alpha}_j - 0.221x_1 - 1.157x_2,$$

with $SE = 0.098$ for $\hat{\beta}_1 = -0.221$ and $SE = 0.382$ for $\hat{\beta}_2 = -1.157$. The nature of the effects and the substantive significance is the same for the two models.

We can interpret parameter estimates in terms of the underlying latent variable model. For example, conditional on the number of traumatic events, the latent distribution on happiness is estimated to have location for whites that is 1.157 standard deviations in the more happy direction compared with that for blacks.

8.3.4 Adjacent-Categories Logit Models

Models for ordinal responses need not use cumulative probabilities. For the logit link, for example, ordinal logits can use pairs of adjacent response probabilities. The *adjacent-categories logits* are

$$\text{logit}[P(Y = j | Y = j \text{ or } j+1)] = \log\frac{\pi_j}{\pi_{j+1}}, \quad j = 1, \ldots, J-1. \tag{8.8}$$

These logits are a basic set equivalent to the baseline-category logits. The connections are

$$\log\frac{\pi_j}{\pi_J} = \log\frac{\pi_j}{\pi_{j+1}} + \log\frac{\pi_{j+1}}{\pi_{j+2}} + \cdots + \log\frac{\pi_{J-1}}{\pi_J} \tag{8.9}$$

and

$$\log\frac{\pi_j}{\pi_{j+1}} = \log\frac{\pi_j}{\pi_J} - \log\frac{\pi_{j+1}}{\pi_J}, \quad j = 1, \ldots, J-1.$$

Either set determines logits for all $\binom{J}{2}$ pairs of response categories.

Models using adjacent-categories logits can be expressed as baseline-category logit models. For instance, consider the adjacent-categories logit model of proportional odds form,

$$\log \frac{\pi_j(\boldsymbol{x})}{\pi_{j+1}(\boldsymbol{x})} = \alpha_j + \boldsymbol{\beta}^T \boldsymbol{x}, \quad j = 1, \ldots, J-1, \tag{8.10}$$

with common effect $\boldsymbol{\beta}$. From adding $(J - j)$ terms as in (8.9), the equivalent baseline-category logit model is

$$\begin{aligned} \log \frac{\pi_j(\boldsymbol{x})}{\pi_J(\boldsymbol{x})} &= \sum_{k=j}^{J-1} \alpha_k + \boldsymbol{\beta}^T (J-j)\boldsymbol{x}, \quad j = 1, \ldots, J-1 \\ &= \alpha_j^* + \boldsymbol{\beta}^T \boldsymbol{u}_j, \qquad\qquad j = 1, \ldots, J-1 \end{aligned}$$

with $\boldsymbol{u}_j = (J - j)\boldsymbol{x}$. The adjacent-categories logit model corresponds to a baseline-category logit model with adjusted model matrix but also a single parameter for each predictor.

The construction of the adjacent-categories logits recognizes the ordering of Y categories. To benefit from this in model parsimony requires appropriate specification of the linear predictor. When an explanatory variable has similar effect for each logit, advantages accrue from using the proportional odds form (8.10) with a single parameter instead of $(J - 1)$ parameters describing that effect. This model fits well in similar situations as the cumulative logit model of proportional odds form. Your choice of model type may reflect whether you prefer effects to refer to individual response categories, as the adjacent-categories logits provide, or instead to groupings of categories using the entire scale or an underlying latent variable, which cumulative logits provide. Since effects in cumulative logit models refer to the entire scale, they are usually larger in magnitude. The ratio of estimate to standard error, however, is usually similar for the two model types.

An advantage of the cumulative logit model is the approximate invariance of effect estimates to the choice and number of response categories. An advantage of the adjacent-categories logit model is that the more general model with $\boldsymbol{\beta}$ replaced by $\boldsymbol{\beta}_j$ is a valid model (i.e., cumulative probabilities will not be out of order), namely, one that is exactly equivalent to an ordinary baseline-category logit model. Also, because of its equivalence with canonical-link (baseline-category logit) models, the model has reduced sufficient statistics and we can use conditional ML estimation for inference with small samples or many parameters. Finally, its effects can be estimated with case–control studies (Mukherjee and Liu 2008).

8.3.5 Example: Happiness Revisited

We return to the example in Sections 8.2.4 and 8.3.3 on modeling happiness in terms of x_1 = total number of traumatic events and x_2 = race. The adjacent-categories logit model of proportional odds form has ML fit

$$\log[\hat{P}(Y = j)/\hat{P}(Y = j + 1)] = \hat{\alpha}_j - 0.357x_1 - 1.842x_2.$$

Conditional on the number of traumatic events, the estimated odds of being very happy instead of pretty happy, and the estimated odds of being pretty happy instead of not too happy, are $e^{1.842} = 6.31$ times as high for whites as for blacks. By contrast, the cumulative logit model had $\hat{\beta}_1 = -0.406$ and $\hat{\beta}_2 = -2.036$. As expected, its estimates are somewhat larger in magnitude. They are not much different for these data, however, because 65 of the 97 observations fall in the middle of the three response categories (pretty happy).

For these data, the more general model having different effects for each adjacent-categories logit has estimate $-\infty$ for the effect of race for the first logit, because there is quasi-complete separation for that logit. The estimates for the effect of number of traumatic events are -0.299 for the first logit and -0.432 for the second logit, suggesting that it is adequate to use the more parsimonious model of proportional odds form with its common estimate of -0.357.

8.3.6 Continuation-Ratio Logit Models

The *continuation-ratio* logits are defined as

$$\log \frac{\pi_j}{\pi_{j+1} + \cdots + \pi_J}, \quad j = 1, \ldots, J-1 \tag{8.11}$$

or as

$$\log \frac{\pi_{j+1}}{\pi_1 + \cdots + \pi_j}, \quad j = 1, \ldots, J-1. \tag{8.12}$$

The continuation-ratio logit model form is useful when a sequential mechanism, such as survival through various age periods, determines the response outcome (e.g., Tutz 1991). Let $\omega_j = P(Y = j | Y \geq j)$. With explanatory variables,

$$\omega_j(\boldsymbol{x}) = \frac{\pi_j(\boldsymbol{x})}{\pi_j(\boldsymbol{x}) + \cdots + \pi_J(\boldsymbol{x})}, \quad j = 1, \ldots, J-1. \tag{8.13}$$

The continuation-ratio logits (8.11) are ordinary logits of these conditional probabilities: namely, $\log[\omega_j(\boldsymbol{x})/(1 - \omega_j(\boldsymbol{x}))]$.

At the ith setting $\boldsymbol{x}_i$ of $\boldsymbol{x}$, let $\{y_{ij}, \ j = 1, \ldots, J\}$ denote the response counts, with $n_i = \sum_j y_{ij}$. When $n_i = 1$, y_{ij} indicates whether the response is in category j, as in Section 8.1.4. Let $b(n, y; \omega)$ denote the binomial probability of y successes in n trials with parameter ω for each trial. From the representation of the multinomial probability of $(y_{i1}, \ldots, y_{iJ})$ in the form $p(y_{i1})p(y_{i2}|y_{i1}) \cdots p(y_{iJ}|y_{i1}, \ldots, y_{i,J-1})$, it follows that the multinomial mass function has factorization

$$b[n_i, y_{i1}; \omega_1(\boldsymbol{x}_i)]b[n_i - y_{i1}, y_{i2}; \omega_2(\boldsymbol{x}_i)] \cdots$$
$$b[n_i - y_{i1} - \cdots - y_{i,J-2}, y_{i,J-1}; \omega_{J-1}(\boldsymbol{x}_i)]. \tag{8.14}$$

The full likelihood is the product of multinomial mass functions from the different $\boldsymbol{x}_i$ values. Thus, the log likelihood is a sum of terms such that different ω_j enter into different terms. When parameters in the model specification for logit(ω_j) are distinct from those

for logit(ω_k) whenever $j \neq k$, maximizing each term separately maximizes the full log likelihood. Thus, separate fitting of models for different continuation-ratio logits gives the same results as simultaneous fitting. The sum of the $J-1$ separate G^2 statistics provides an overall goodness-of-fit statistic pertaining to the simultaneous fitting of $J-1$ models. Because of factorization (8.14), separate fitting can use methods for binary logistic models. Similar remarks apply to continuation-ratio logits (8.12), although those logits and the subsequent analysis do not give equivalent results.

Sometimes, a simpler proportional odds form of the model is plausible in which effects are the same for each logit (McCullagh and Nelder 1989, p. 164; Tutz 1991). Because of the factorization (8.14), it is also possible to fit such a model simply by creating a data file of independent binomials. See Agresti (2010, Sec. 4.2).

8.3.7 Example: Developmental Toxicity Study with Pregnant Mice

Table 8.7 comes from a developmental toxicity study. Such experiments with rodents test substances posing potential danger to developing fetuses. Diethylene glycol dimethyl ether (diEGdiME), one such substance, is an industrial solvent used in the manufacture of protective coatings such as lacquer and metal coatings. This study administered diEGdiME in distilled water to pregnant mice. Each mouse was exposed to one of five concentration levels for 10 days early in the pregnancy. The mice exposed to level 0 formed a control group. Two days later, the uterine contents of the pregnant mice were examined for defects. Each fetus has three possible outcomes (nonlive, malformation, normal). The outcomes are ordered, with nonlive the least desirable result. We use continuation-ratio logits to model (1) the probability π_1 of a nonlive fetus, and (2) the conditional probability $\pi_2/(\pi_2+\pi_3)$ of a malformed fetus, given that the fetus was live.

We fitted the continuation-ratio logit models

$$\log \frac{\pi_1(x_i)}{\pi_2(x_i)+\pi_3(x_i)} = \alpha_1 + \beta_1 x_i, \quad \log \frac{\pi_2(x_i)}{\pi_3(x_i)} = \alpha_2 + \beta_2 x_i,$$

using x_i scores $\{0, 62.5, 125, 250, 500\}$ for concentration level. The ML estimates are $\hat{\beta}_1 = 0.0064$ ($SE = 0.0004$) and $\hat{\beta}_2 = 0.0174$ ($SE = 0.0012$). In each case, the less desirable outcome is more likely as the concentration increases. For instance, given that a fetus was

Table 8.7 Outcomes for Pregnant Mice in Developmental Toxicity Study

Concentration (mg/kg per day)	Response		
	Nonlive	Malformation	Normal
0 (controls)	15	1	281
62.5	17	0	225
125	22	7	283
250	38	59	202
500	144	132	9

[a]Based on results in C. J. Price et al., *Fundam. Appl. Toxicol.* **8**: 115–126, 1987. I thank Louise Ryan for showing me these data.

live, the estimated odds that it was malformed rather than normal multiplies by $\exp(1.74) = 5.7$ for every 100-unit increase in the concentration of diEGdiME. The likelihood-ratio fit statistics are $G^2 = 5.78$ for $j = 1$ and $G^2 = 6.06$ for $j = 2$, each based on df = 3. Their sum, $G^2 = 11.84$ (or similarly $X^2 = 9.76$), with df = 6, summarizes the fit.

This analysis treats pregnancy outcomes for different fetuses as independent, identical observations. In fact, each pregnant mouse had a litter of fetuses, and statistical dependence may exist among different fetuses in the same litter. Different litters at a given concentration level may also have different response probabilities. Heterogeneity of various sorts among the litters (e.g., due to varying physical and/or genetic characteristics among different pregnant mice) would cause these probabilities to vary somewhat. Either statistical dependence or heterogeneous probabilities violates the binomial assumption and causes overdispersion. At a fixed concentration level, the number of fetuses in a litter that die may vary among pregnant mice more than if the counts were independent and identical binomial variates. The total G^2 shows some evidence of lack of fit ($P = 0.07$) but may reflect overdispersion caused by these factors rather than an inappropriate choice of response curve.

To account for overdispersion, we could adjust standard errors using the quasi-likelihood approach (Section 4.7). This multiplies standard errors by $\sqrt{X^2/\text{df}} = \sqrt{9.76/6} = 1.28$. For each logit, strong evidence remains that $\beta_j > 0$. In Chapters 13 and 14 we present other methods that account for the clustering of fetuses in litters.

8.3.8 Stochastic Ordering Location Effects Versus Dispersion Effects

For cumulative link models, settings of the explanatory variables are *stochastically ordered* on the response: For any pair $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, either $P(Y \leq j|\boldsymbol{x}_1) \leq P(Y \leq j|\boldsymbol{x}_2)$ for all j or $P(Y \leq j|\boldsymbol{x}_1) \geq P(Y \leq j|\boldsymbol{x}_2)$ for all j. Figure 8.7a illustrates for underlying continuous density functions and cdf's at two settings of $\boldsymbol{x}$. Likewise, the adjacent-categories and continuation-ratio logit models with proportional odds structure imply stochastically ordered distributions for Y at different predictor values.

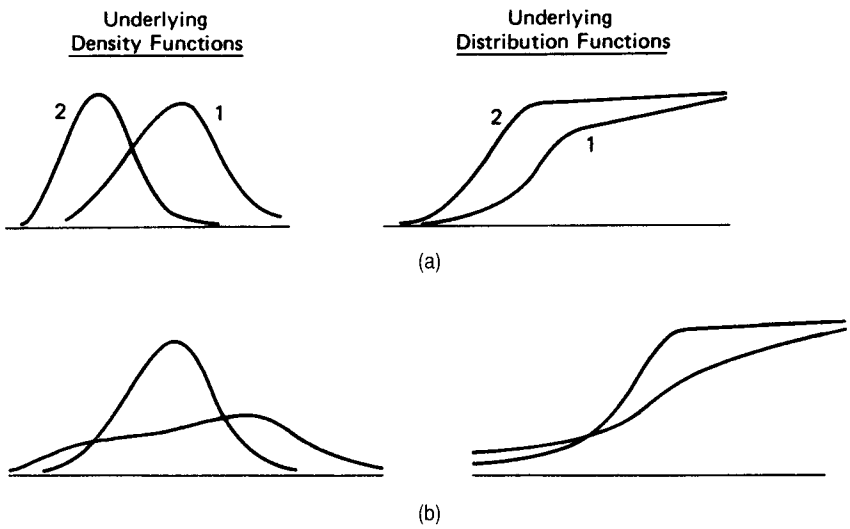

Figure 8.7 (a) Distribution 1 stochastically higher than distribution 2; (b) distributions not stochastically ordered.

When this is violated and such models fit poorly, often it is because the dispersion also varies with $\boldsymbol{x}$. For instance, perhaps responses tend to concentrate around the same location but more dispersion occurs at $\boldsymbol{x}_1$ than at $\boldsymbol{x}_2$. Then perhaps $P(Y \leq j|\boldsymbol{x}_1) > P(Y \leq j|\boldsymbol{x}_2)$ for small j but $P(Y \leq j|\boldsymbol{x}_1) < P(Y \leq j|\boldsymbol{x}_2)$ for large j. In other words, at $\boldsymbol{x}_1$ the responses concentrate more at the extreme categories than at $\boldsymbol{x}_2$. Figure 8.7b illustrates for underlying continuous distributions.

Cumulative link models have been proposed that incorporate dispersion effects, but mainly for relatively simple cases such as with a single predictor that is a factor (Note 8.8). A simpler approach when a cumulative link model fits poorly is to fit the model separately for each cumulative probability to investigate the nature of the lack of fit or to use one of the other options mentioned at the end of Section 8.2.5.

8.3.9 Summarizing Predictive Power of Explanatory Variables

How can we summarize how well the response can be predicted using the fit of the chosen model? One approach estimates a measure such as the multiple correlation or R-squared for the regression model for an underlying latent response variable. McKelvey and Zavoina (1975) suggested this for the cumulative probit model.

Another index of predictive power generalizes the *concordance index* (Section 6.3.4). For all pairs of observations that have different response outcomes, it estimates the probability that the predictions and the outcomes are concordant, that is, that the observation with the larger y-value also has a stochastically higher set of estimated probabilities (and hence, for example, a higher mean for the estimated conditional distribution). The baseline value of no effect is 0.50. A value of 1.0 results when knowing which observation in an untied pair has the stochastically higher estimated distribution enables us to perfectly predict which one has the higher actual response. The higher the value of the concordance index, the better the predictive power.

Such measures are mainly useful for comparing different models. For example, for the happiness data analyzed with a proportional odds type of cumulative logit model in Section 8.2.4, the concordance index is 0.688 for the main-effects model and 0.689 when an interaction term is added. So, the more complex model is not much more useful for predictions, regardless of whether its extra term is statistically significant.

Keep in mind that predictive power is distinct from goodness of fit. A model may fit a particular data set well even if the predictive power the model provides is small. For other approaches to summarizing predictive power, see Agresti (2010, Sec. 3.4.6).

8.4 TESTING CONDITIONAL INDEPENDENCE IN $I \times J \times K$ TABLES

A common statistical analysis in many applications is studying whether an explanatory variable X has an effect on a response variable Y after we adjust for one or more other relevant factors. In Section 6.4 we considered this for binary Y and X using logistic models and the Cochran–Mantel–Haenszel (CMH) test of conditional independence for $2 \times 2 \times K$ tables. This section presents related tests with multicategory variables, in the context of $I \times J \times K$ tables. Likelihood-ratio tests compare the fit of a model specifying XY conditional independence with a model permitting X to have an effect. Generalizations of the CMH statistic are score statistics for certain models.

8.4.1 Testing Conditional Independence Using Multinomial Models

Denote a control factor by Z. Treating Z as nominal scale, we discuss four cases that treat (Y, X) as (nominal, nominal), (nominal, ordinal), (ordinal, nominal), (ordinal, ordinal). When Y is nominal, the baseline-category logit model of XY conditional independence is

$$\log \frac{P(Y = j|X = i, Z = k)}{P(Y = J|X = i, Z = k)} = \alpha_{jk}. \tag{8.15}$$

That is, each logit does not depend on the category of X. For ordinal Y we use cumulative logit models, but other ordinal models yield analogous tests. Then, XY conditional independence is equivalent to the model

$$\text{logit}[P(Y \leq j|X = i, Z = k)] = \alpha_{jk},$$

with $\alpha_{1k} < \alpha_{2k} < \cdots < \alpha_{J-1,k}$ for each k. When the XY association is similar in the partial tables, the power of a test benefits from basing a test statistic on a model of homogeneous association.

1. *Y nominal, X nominal.* An alternative to XY conditional independence that treats X as a factor is

$$\log \frac{P(Y = j|X = i, Z = k)}{P(Y = J|X = i, Z = k)} = \alpha_{jk} + \beta_{ij} \tag{8.16}$$

   with constraint such as $\beta_{Ij} = 0$ for each j. For each outcome category j, X and Z have additive effects of form $\alpha_k + \beta_i$. Conditional independence is H_0: $\beta_{1j} = \cdots = \beta_{Ij}$ for $j = 1, \ldots, J-1$. Large-sample chi-squared tests have df $= (I-1)(J-1)$.
2. *Y nominal, X ordinal.* Let $\{x_i\}$ be ordered scores. A test that is sensitive to the same linear trend alternatives in each partial table compares the conditional independence model to

$$\log \frac{P(Y = j|X = i, Z = k)}{P(Y = J|X = i, Z = k)} = \alpha_{jk} + \beta_j x_i.$$

   Conditional independence is H_0: $\beta_1 = \cdots = \beta_{J-1} = 0$. Large-sample chi-squared tests have df $= J-1$.
3. *Y ordinal, X nominal.* An alternative to XY conditional independence that treats X as a factor is

$$\text{logit}[P(Y \leq j|X = i, Z = k)] = \alpha_{jk} + \beta_i,$$

   with a constraint such as $\beta_I = 0$. A simpler model that also has proportional odds structure for the effects of Z has linear predictor $\alpha_j + \beta_k^Z + \beta_i$. For either model, XY conditional independence is H_0: $\beta_1 = \cdots = \beta_I$. Large-sample chi-squared tests have df $= I-1$.

Table 8.8 Summary of Models for Testing Conditional Independence[a]

Y–*X*	Model	Conditional Independence	df
Ordinal–ordinal	$\text{logit}[P(Y \le j)] = \alpha_{jk} + \beta x_i$	$\beta = 0$	1
Ordinal–nominal	$\text{logit}[P(Y \le j)] = \alpha_{jk} + \beta_i$	$\beta_1 = \cdots = \beta_I$	$I - 1$
Nominal–ordinal	$\log\left[\frac{P(Y=j)}{P(Y=J)}\right] = \alpha_{jk} + \beta_j x_i$	$\beta_1 = \cdots = \beta_{J-1} = 0$	$J - 1$
Nominal–nominal	$\log\left[\frac{P(Y=j)}{P(Y=J)}\right] = \alpha_{jk} + \beta_{ij}$	All $\beta_{ij} = 0$	$(I-1)(J-1)$

[a]The first two cases can also use $\alpha_j + \beta_k^Z$ in place of α_{jk}.

4. *Y ordinal, X ordinal.* For ordered scores $\{x_i\}$, the model

$$\text{logit}[P(Y \le j | X = i, Z = k)] = \alpha_{jk} + \beta x_i \tag{8.17}$$

has the same linear trend for the X effect in each partial table. A simpler model that also has proportional odds structure for the effects of Z has linear predictor $\alpha_j + \beta_k^Z + \beta x_i$. For either model, XY conditional independence is H_0: $\beta = 0$. Large-sample chi-squared tests have df = 1.

Table 8.8 summarizes the four tests. They work well when the model describes a major component of the departure from conditional independence. This does not mean that we must test the fit of the model in order to use the test (see the remarks at the end of Section 6.4.2).

Occasionally, the association may change dramatically across the K partial tables. When Z is ordinal, an alternative by which a log odds ratio changes linearly across levels of Z is sometimes of use. For instance, when Z = age of subject, the association between a risk factor X (e.g., level of smoking) and a response Y (e.g., severity of heart disease) may tend to increase with Z. When Z is nominal, the conditional independence models can be compared with a more general alternative having separate effect parameters at each level of Z. Allowing effects to vary across levels of Z, however, results in the test df being multiplied by K, which handicaps power.

8.4.2 Example: Homosexual Marriage and Religious Fundamentalism

In 2008 the General Social Survey asked whether homosexuals should have the right to marry. One variable with which we'd expect responses to be associated is the fundamentalism/liberalism of a subject's religious beliefs. A subject's attained education is likely associated with both these variables, so is there an association when we condition on education? Table 8.9 shows the relationship between opinion about homosexual marriage (Y) and religious beliefs (X), stratified by Z = attained education, for subjects of age 18–25.

Table 8.10 summarizes the fit of several logistic models and shows the results of related likelihood-ratio tests of conditional independence. Each test compares a model to the model deleting the religious beliefs effect, conditioning on attained education. The models that treat opinion as ordinal use cumulative logits, with linear predictor $\alpha_j + \beta_k^Z + \beta x_i$ to treat

Table 8.9 Opinion About Homosexual Marriage by Religious Beliefs, at Two Education Levels

		Homosexuals Should Be Able to Marry		
Education	Religion	Agree	Neutral	Disagree
High school or less	Fundamentalist	6	2	10
	Moderate	8	3	9
	Liberal	11	5	6
At least some college	Fundamentalist	4	2	11
	Moderate	21	3	5
	Liberal	22	4	1

Source: 2008 General Social Survey, subsample for ages 18–25.

Table 8.10 Summary of Model-Based Likelihood-Ratio Tests of Conditional Independence for Table 8.9

Opinion	Religion	G^2 Fit	df	Test Statistic	df	P-value
Ordinal	Ordinal	10.36	8	16.57	1	<0.0001
	Nominal	9.17	7	17.76	2	0.0001
	Not in model	26.93	9	—	—	—
Nominal	Ordinal	7.33	6	19.53	2	0.0001
	Nominal	6.58	4	20.27	4	0.0004
	Not in model	26.85	8	—	—	—

X as an ordinal predictor using x_i scores (1, 2, 3) and linear predictor $\alpha_j + \beta_k^Z + \beta_i$ to treat X as a nominal factor. The corresponding tests compare these to the model with linear predictor $\alpha_j + \beta_k^Z$. That model is not exactly equivalent to the conditional independence model (8.15), which is the last model listed in the table, with $G^2 = 26.85$ based on df = 8.

Testing conditional independence with the first cumulative logit model yields likelihood-ratio statistic $26.93 - 10.36 = 16.57$ with df $= 9 - 8 = 1$, strong evidence of an effect. Models that treat either or both variables as nominal also provide strong evidence, but not quite as strong. Focusing the test on a linear trend alternative yields a smaller P-value when that model describes reality reasonably well. However, we learn more from estimating model parameters than from these significance tests.

8.4.3 Generalized Cochran–Mantel–Haenszel Tests for $I \times J \times K$ Tables

The CMH statistic generalizes to multiple rows and columns. The tests treat X and Y symmetrically, so the three cases correspond to treating both as nominal, both as ordinal, or one of each. Conditional on row and column totals, each stratum has $(I-1)(J-1)$ nonredundant cell counts. Let

$$\boldsymbol{n}_k = (n_{11k}, n_{12k}, \ldots, n_{1,J-1,k}, \ldots, n_{I-1,J-1,k})^T.$$

Let $\boldsymbol{\mu}_k = E(\boldsymbol{n}_k)$ under H_0: conditional independence, namely,

$$\boldsymbol{\mu}_k = (n_{1+k}n_{+1k}, n_{1+k}n_{+2k}, \ldots, n_{I-1,+,k}n_{+,J-1,k})^T / n_{++k}.$$

Let $\boldsymbol{V}_k$ denote the null covariance matrix of $\boldsymbol{n}_k$, conditional on the margins, where

$$\text{cov}(n_{ijk}, n_{i'j'k}) = \frac{n_{i+k}(\delta_{ii'}n_{++k} - n_{i'+k})n_{+jk}(\delta_{jj'}n_{++k} - n_{+j'k})}{n_{++k}^2(n_{++k} - 1)}$$

with $\delta_{ab} = 1$ when $a = b$ and $\delta_{ab} = 0$ otherwise.

First, suppose the rows and columns are unordered. Let

$$\boldsymbol{n} = \sum_k \boldsymbol{n}_k, \quad \boldsymbol{\mu} = \sum_k \boldsymbol{\mu}_k, \quad \boldsymbol{V} = \sum_k \boldsymbol{V}_k.$$

The generalized CMH statistic for nominal X and Y is

$$\text{CMH} = (\boldsymbol{n} - \boldsymbol{\mu})^T \boldsymbol{V}^{-1}(\boldsymbol{n} - \boldsymbol{\mu}). \tag{8.18}$$

Its large-sample chi-squared distribution has df $= (I - 1)(J - 1)$. The df value equals that for the statistics comparing logistic models (8.15) and (8.16). For $K = 1$ stratum with n observations, CMH $= [(n - 1)/n]X^2$, where X^2 is the Pearson statistic (3.10) for testing independence.

Next, suppose the rows and columns are both ordered. For ordered scores $\{u_i\}$ and $\{v_j\}$, evidence of a positive trend occurs if in each stratum $T_k = \sum_i \sum_j u_i v_j n_{ijk}$ exceeds its null expectation. Given the marginal totals, under conditional independence

$$E(T_k) = \left[\sum_i u_i n_{i+k}\right]\left[\sum_j v_j n_{+jk}\right] \Big/ n_{++k},$$

$$\text{var}(T_k) = \frac{1}{n_{++k} - 1}\left[\sum_i u_i^2 n_{i+k} - \frac{\left(\sum_i u_i n_{i+k}\right)^2}{n_{++k}}\right] \times \left[\sum_j v_j^2 n_{+jk} - \frac{\left(\sum_j v_j n_{+jk}\right)^2}{n_{++k}}\right].$$

The statistic $[T_k - E(T_k)]/\sqrt{\text{var}(T_k)}$ equals the correlation between X and Y in stratum k multiplied by $\sqrt{n_{++k} - 1}$. To summarize across the K strata in a way that is sensitive to a correlation of common sign in each stratum, Mantel (1963) proposed

$$M^2 = \frac{\left\{\sum_k\left[\sum_i \sum_j u_i v_j n_{ijk} - E\left(\sum_i \sum_j u_i v_j n_{ijk}\right)\right]\right\}^2}{\sum_k \text{var}\left(\sum_i \sum_j u_i v_j n_{ijk}\right)}. \tag{8.19}$$

This has a large-sample χ_1^2 null distribution, the same as for testing H_0: $\beta = 0$ in ordinal model (8.17). For $K = 1$, this is the M^2 correlation-based statistic (3.16).

Table 8.11 Output (from SAS, PROC FREQ) for Generalized Cochran–Mantel–Haenszel Tests with Data from Table 8.9

Summary Statistics for opinion by religious fundamentalism
Controlling for education
Cochran--Mantel--Haenszel Statistics (Based on Table Scores)

Statistic	Alternative Hypothesis	DF	Value	Prob
1	Nonzero Correlation	1	16.83	<0.0001
2	Row Mean Scores Differ	2	17.94	0.0001
3	General Association	4	19.76	0.0006

Landis et al. (1978) presented a statistic that has (8.18) and (8.19) as special cases. Their statistic also can treat X as nominal and Y as ordinal, summarizing information about how I row means compare to their null expected values, with df $= I - 1$ (see Note 8.9).

8.4.4 Example: Homosexual Marriage Revisited

Table 8.11 shows output for conducting generalized CMH tests with Table 8.9. Statistics treating a variable as ordinal used scores (1, 2, 3) for opinion and for religious beliefs.

The *general association* alternative treats X and Y as nominal and uses (8.18). It is sensitive to any association that is similar in each category of Z. The *nonzero correlation* alternative treats X and Y as ordinal and uses (8.19). It is sensitive to a similar linear trend in each category of Z. The *row mean scores differ* alternative treats rows as nominal and columns as ordinal. It is sensitive to variation among the I row mean scores on Y, when that variation is similar in each category of Z.

8.4.5 Related Score Tests for Multinomial Logit Models

The generalized CMH tests seem to be non-model-based alternatives to the tests of Section 8.4.1 using multinomial logit models. However, a close connection exists between them. For certain multinomial logit models, the generalized CMH tests are score tests of conditional independence.

The generalized CMH test (8.18) that treats X and Y as nominal is the score test that the $(I-1)(J-1)$ $\{\beta_{ij}\}$ parameters in model (8.16) equal 0. The generalized CMH test using M^2 that treats X and Y as ordinal is the score test of $\beta = 0$ in model (8.17). For the cumulative logit model, the equivalence has the same $\{x_i\}$ scores in the model as in M^2, and the $\{v_j\}$ scores in M^2 are average rank scores. For the adjacent-categories logit model analog of (8.17), the $\{v_j\}$ scores in M^2 are any equally spaced scores.

With large samples in each stratum, the generalized CMH tests give similar results as likelihood-ratio tests comparing the relevant models. An advantage of the model-based approach is providing estimates of effects. An advantage of the generalized CMH tests is maintaining good performance under sparse asymptotics whereby K grows as n does. Also, they are valid under randomization arguments when there is not multinomial sampling from the population of interest but the multivariate hypergeometric distribution applies to each stratum under the null, such as for a volunteer sample of subjects randomly assigned to treatments in a clinical trial.

8.5 DISCRETE-CHOICE MODELS

Many applications of multinomial logit models relate to determining effects of explanatory variables on a subject's choice from a discrete set of options—for instance, transportation system to take to work (driving alone, carpooling, bus, subway, walk, bicycle), housing (buy house, buy condominium, rent), primary shopping location (downtown, mall, catalogs, Internet), or product brand. Models for response variables consisting of a discrete set of choices are called *discrete-choice models*.

8.5.1 Conditional Logits for Characteristics of the Choices

In many discrete-choice applications, an explanatory variable takes different values for different response choices. As predictors of choice of transportation system, the cost and the time to reach the destination take different values for each option. As a predictor of choice of product brand, the price varies according to the option. Explanatory variables of this type are *characteristics of the choices*. They differ from the usual ones, for which values remain constant across the choice set. Such variables, *characteristics of the chooser*, include demographic characteristics such as gender, race, and educational attainment.

McFadden (1974) proposed a discrete-choice model for explanatory variables that are characteristics of the choices. For subject i and response choice j, let $\boldsymbol{x}_{ij} = (x_{ij1}, \ldots, x_{ijp})^T$ denote the values of the p explanatory variables, and let $\boldsymbol{x}_i = (\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{ip})$. The model for the probability of selecting option j is

$$\pi_j(\boldsymbol{x}_i) = \frac{\exp(\boldsymbol{\beta}^T \boldsymbol{x}_{ij})}{\sum_h \exp(\boldsymbol{\beta}^T \boldsymbol{x}_{ih})}. \tag{8.20}$$

For each pair of choices a and b, this model has the logit form for conditional probabilities,

$$\log[\pi_a(\boldsymbol{x}_i)/\pi_b(\boldsymbol{x}_i)] = \boldsymbol{\beta}^T(\boldsymbol{x}_{ia} - \boldsymbol{x}_{ib}). \tag{8.21}$$

Conditional on the choice being a or b, a variable's influence depends on the distance between the subject's values of that variable for those choices. If the values are the same, the model asserts that the variable has no influence on the choice between a and b. Reflecting this property, McFadden originally referred to model (8.20) as a *conditional logit* model.

From (8.21), the odds of choosing a over b do not depend on the other alternatives in the choice set or on their values of the explanatory variables. Luce (1959) called this property *independence from irrelevant alternatives*. It is unrealistic in some applications. For instance, for travel options auto and red bus, suppose that 80% choose auto, corresponding to an odds of 4.0. Now suppose that the options are auto, red bus, and blue bus. According to (8.21), the odds are still 4.0 of choosing auto instead of red bus, but intuitively, we expect them to be about 8.0 (if about 10% choose each bus option), McFadden (1974) stated: "Application of the model should be limited to situations where the alternatives can plausibly be assumed to be distinct and weighed independently in the eyes of each decision-maker."

McFadden's model is actually a bit more general, permitting the choice set to vary among subjects. For instance, some subjects may not have the subway as an option for travel to work. In the denominator of (8.20), the sum is then taken over the choice set for subject i.

8.5.2 Multinomial Logit Model Expressed as Discrete-Choice Model

Discrete-choice models can also incorporate explanatory variables that are characteristics of the chooser. This may seem surprising, since formula (8.20) has a single parameter for each explanatory variable; that is, the parameter vector is the same for each pair of choices. However, multinomial logit model (8.2) has this discrete-choice form when we replace such an explanatory variable by J artificial variables. The jth is the product of the explanatory variable with a indicator variable that equals 1 when the response choice is j. For instance, for a single explanatory variable, let x_i denote its value for subject i. For $j = 1, \ldots, J$, let δ_{jk} equal 1 when $k = j$ and 0 otherwise, and let

$$z_{ij} = (\delta_{j1}, \ldots, \delta_{jJ}, \delta_{j1}x_i, \ldots, \delta_{jJ}x_i)^T.$$

Let $\boldsymbol{\beta} = (\alpha_1, \ldots, \alpha_J, \beta_1, \ldots, \beta_J)^T$. Then $\boldsymbol{\beta}^T z_{ij} = \alpha_j + \beta_j x_i$, and (8.2) is (with $\alpha_J = \beta_J = 0$ for identifiability)

$$\begin{aligned}\pi_j(x_i) &= \frac{\exp(\alpha_j + \beta_j x_i)}{\exp(\alpha_1 + \beta_1 x_i) + \cdots + \exp(\alpha_J + \beta_J x_i)} \\ &= \frac{\exp(\boldsymbol{\beta}^T z_{ij})}{\exp(\boldsymbol{\beta}^T z_{i1}) + \cdots + \exp(\boldsymbol{\beta}^T z_{iJ})}.\end{aligned}$$

This has the discrete-choice model form (8.20).

With this approach, discrete-choice models can contain characteristics of the chooser and of the choices. Thus, the model is very general. The ordinary multinomial logit model using baseline-category logits is a special case.

8.5.3 Example: Shopping Destination Choice

McFadden (1974) used discrete-choice models to describe how residents of Pittsburgh, Pennsylvania, chose a shopping destination. The five possible destinations were different city zones. One explanatory variable measured S = shopping opportunities, defined to be the retail employment in the zone as a percentage of total retail employment in the region. The other explanatory variable was P = price of the trip, defined from a separate analysis using auto in-vehicle time and auto operating cost.

The ML estimates of model parameters were -1.06 ($SE = 0.28$) for price of trip and 0.84 ($SE = 0.23$) for shopping opportunity. From (8.21),

$$\log(\hat{\pi}_a/\hat{\pi}_b) = -1.06(P_a - P_b) + 0.84(S_a - S_b).$$

Not surprisingly, a destination is relatively more attractive as the trip price decreases and as the shopping opportunity increases.

8.5.4 Multinomial Probit Discrete-Choice Models

Let U_{ij} denote the utility of alternative j for subject i. Suppose that

$$U_{ij} = \boldsymbol{\beta}^T \boldsymbol{x}_{ij} + \epsilon_{ij} \tag{8.22}$$

and the response choice is the value of j having maximum utility. McFadden (1974) showed that the assumption that $\{\epsilon_{ij}\}$ are independent and have the standard extreme value distribution is equivalent to discrete-choice model (8.20). (Recall Note 7.2 and Section 8.1.6.)

One way such a construction may be unrealistic is when the error terms partly represent unobserved covariates that are correlated with the response variable. Then, ϵ_{ia} and ϵ_{ib} are unlikely to be independent. However, this utility structure suggests alternative models, in particular, ones that do not have the property of independence from irrelevant alternatives. When we assume that $\boldsymbol{\epsilon} = (\epsilon_{i1}, \ldots, \epsilon_{iJ})$ has a multivariate normal $N(\mathbf{0}, \boldsymbol{\Sigma})$ distribution, this utility model is a *multinomial probit model*, extending the model of Section 8.1.6. Model identifiability requires constraints on $\boldsymbol{\Sigma}$, such as by taking $\text{var}(\epsilon_{i1}) = 1$ (Hausman and Wise 1978). Multinomial probit models are more complex computationally, requiring numerical integration or simulation to obtain the likelihood function.

8.5.5 Extensions: Nested Logit and Mixed Logit Models

In permitting correlated errors among response categories in a model for utilities, we could instead assume that $\boldsymbol{\epsilon}$ has a multivariate form of extreme value distribution. This induces generalized logistic models. For example, McFadden considered applications in which the choice categories are partitioned into groups having a tree-like structure, with each group consisting of similar alternatives and having correlated error terms within groups. This is useful when the choices are naturally nested. An example is a person's choice of where to live: The person first chooses one of several communities to live in, and then within that community chooses a type of dwelling. Such a model for nested choices is called a *nested logit model*. Train (2009, pp. 77–88) gave an overview and multiple references.

Multinomial logit and probit discrete-choice models can be further generalized by treating certain effects as random rather than fixed, in the spirit of models considered later in this text in Chapters 12 and 13. A *mixed logit model* is one in which choice probabilities are obtained by integrating the logistic expression (8.20) for choice probabilities with respect to a distribution for certain model parameters. This allows heterogeneity among subjects in the size of effects. It is useful as a mechanism for inducing positive association among repeated responses with longitudinal data. Estimates of the parameters of the mixing distribution provide information about the average effects and the extent of the heterogeneity. Individual effects can also be predicted. For details, see McFadden (1974), Skrondal and Rabe-Hesketh (2004, Chap. 13), and Train (2009, Chap. 6).

8.5.6 Extensions: Discrete Choice with Ordered Categories

Sometimes the response categories have a natural ordering, such as the choice in renting a car among (subcompact, compact, midsize, large) size levels. Standard discrete-choice models do not account for such ordering. For multinomial logit models, the property of independence from irrelevant alternatives may then be especially unrealistic, as a particular response category is more similar to categories near it than categories further away.

Small (1987) proposed a model related to McFadden's multivariate extreme value model for utilities. In his model, the correlation between utility components for alternatives a and b is a nonincreasing function of $|a - b|$. Another approach uses a multinomial probit model with structure on the covariance matrix for $\boldsymbol{\epsilon}$ that reflects the ordinality. For example, the correlation might have the autoregressive structure whereby $\text{corr}(\epsilon_{ia}, \epsilon_{ib}) = \rho^{|a-b|}$.

Beggs et al. (1981) considered an alternative type of ordered-alternatives problem in which subjects fully rank the outcome categories from best to worst. The categories need not themselves be ordered. They assumed the utility model (8.22), assuming *iid* extreme value errors. Let $(r_{i1}, \ldots, r_{iJ})$ denote the ranking by subject i of the J choices, where r_{i1} is the response category given the highest ranking and r_{iJ} is the response category given the lowest ranking. Based on convenient properties of conditional distributions for extreme value distributions, they showed that that ranking vector for subject i has probability

$$P(U_{r_{i1}} > U_{r_{i2}} > \cdots > U_{r_{iJ}}) = \prod_{h=1}^{J-1} \left[\exp(\boldsymbol{\beta}^T \boldsymbol{x}_{r_{ih}}) \Big/ \sum_{m=h}^{J} \exp(\boldsymbol{\beta}^T \boldsymbol{x}_{r_{im}}) \right].$$

Summing the logs of these terms over the n subjects yields the multinomial log likelihood. It can be maximized using Newton–Raphson methods. Beggs et al. (1981) applied the model to data in which various car types were ranked and the explanatory variables included car choice characteristics such as price, fuel cost, whether gas-powered or electric-powered, and subject socioeconomic family characteristics.

8.6 BAYESIAN MODELING OF MULTINOMIAL RESPONSES

The Bayesian approach for binary regression models extends to multinomial models. We focus here on Bayesian fitting of cumulative link models for ordinal responses and of multinomial (baseline-category) logit and probit models for nominal responses.

8.6.1 Bayesian Fitting of Cumulative Link Models

For an ordinal response Y, many models are special cases of the cumulative link model,

$$G^{-1}[P(Y \leq j|\boldsymbol{x})] = \alpha_j - \boldsymbol{\beta}^T \boldsymbol{x}.$$

From Section 8.2.3, this model is implied by a regression model for a latent variable having cdf G, such as logistic for the logit link. Prior distributions for the cutpoint parameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{c-1})$ should take into account the ordering constraint

$$-\infty < \alpha_1 < \alpha_2 < \cdots < \alpha_{c-1} < \infty.$$

In the cumulative probit case, the latent response for observation i is

$$Y_i^* = \boldsymbol{\beta}^T \boldsymbol{x}_i + \epsilon_i,$$

where $\{\epsilon_i\}$ are independent N(0, 1). Albert and Chib (1993) presented a Bayesian analysis that utilizes the latent variable model and extends the analysis of Section 7.2.6 for binary responses. This model is simpler to handle than the cumulative logit model, because results apply from Bayesian inference for ordinary normal linear regression models, with a multivariate normal prior distribution for the regression parameters and independent normal

latent variables $\boldsymbol{y}^* = (y_1^*, \ldots, y_n^*)^T$. Implementation of MCMC methods is relatively simple because the Monte Carlo sampling is from normal distributions.

A Gibbs sampling scheme determines the posterior distribution by successively sampling from the density of (1) $\boldsymbol{y}^*$ given $\boldsymbol{y}$, $\boldsymbol{\beta}$, and $\boldsymbol{\alpha}$, (2) $\boldsymbol{\beta}$ given $\boldsymbol{y}$, $\boldsymbol{y}^*$, and $\boldsymbol{\alpha}$, and (3) $\boldsymbol{\alpha}$ given $\boldsymbol{y}$, $\boldsymbol{y}^*$, and $\boldsymbol{\beta}$. It uses the fact that if $y_i = j$, then y_i^* is between α_{j-1} and α_j. For example, given $\boldsymbol{y}$, $\boldsymbol{\beta}$, and $\boldsymbol{\alpha}$, the conditional density function of y_i^* is normal with mean $\boldsymbol{\beta}^T \boldsymbol{x}_i$ and variance 1 but truncated between the two cutpoints corresponding to the value of y_i. Since $\boldsymbol{y}^*$ determines $\boldsymbol{y}$, the conditional density of $\boldsymbol{\beta}$ given $\boldsymbol{y}$, $\boldsymbol{y}^*$, and $\boldsymbol{\alpha}$ is proportional to the prior of $\boldsymbol{\beta}$ times the density of $\boldsymbol{y}^*$ given $\boldsymbol{\beta}$, which is normal since both components are normal. The conditional density function of $\boldsymbol{\alpha}$ given $\boldsymbol{y}$, $\boldsymbol{y}^*$, and $\boldsymbol{\beta}$ is proportional to its truncated normal prior density but truncated to reflect that α_j must fall above all y_i^* such that $y_i = j$ but below all y_i^* such that $y_i = j + 1$.

Albert and Chib generalized the model to use link functions based on inverse cdf's for the t distribution. Since the logistic distribution relates closely to the t distribution with df = 8 (as described in Sections 4.2.5 and 7.2.6), this also provides a relatively simple way of fitting corresponding cumulative logit models. Alternatively, these days it is straightforward to use MCMC directly with the product of the chosen prior densities and the multinomial likelihood for the chosen model, regardless of the link function.

8.6.2 Example: Cannabis Use and Mother's Age

Table 8.12 comes from a 21-year follow-up study of mothers and their children who received antenatal care at a public hospital in Brisbane, Australia. At the age of 21, the children were asked "In the last month, how often did you use cannabis, marijuana, pot, etc.?" One explanatory variable was the mother's age at entry to the study.

The deviance statistic for testing goodness of fit of the independence model in this table is $G^2 = 7.71$ (df = 4, $P = 0.10$). There is not much evidence of association, but this statistic ignores the ordinality of the response. Let's consider the cumulative logit model of proportional odds form,

$$\text{logit} P(Y \leq j) = \alpha_j + \beta x,$$

with $x = 1$ for age ≥ 20 and $x = 0$ otherwise. The deviance is now 3.18 (df = 3). The ML estimate $\hat{\beta} = 0.230$ ($SE = 0.107$) and likelihood-ratio statistic of 4.53 (df = 1, $P = 0.033$) for testing H_0: $\beta = 0$ show considerable evidence that cannabis use tended to be lower when mother's age was higher. The 95% profile likelihood confidence interval for β is (0.018, 0.441).

Table 8.12 Cannabis Use at 21 Years by Mother's Age at Study Entry

	Cannabis Use at 21 Years				
Mother's Age	Never Use	Not Last Month	Once Last Month	Every Few Days	Every Day
<20 years	154	91	42	27	30
≥20 years	1078	567	261	157	111

Source: Hayatbakhsh et al. *Am. J. Drug & Alcohol Abuse*, **36**: 350–356, 2010.

For a Bayesian analysis we recode x to take values 0.5 and -0.5 instead of 1 and 0, so the cumulative logits in each row have the same prior variability. For an analysis with uninformative priors, we used independent normal priors for the model parameters (appropriately truncated for $\{\alpha_j\}$) with means of 0 and standard deviations of 10. The posterior distribution of β, based on a Markov chain of length one million, then has a mean of 0.229 and a standard deviation of 0.108. The equal-tail 95% posterior interval for β is (0.018, 0.441), and the posterior $P(\beta < 0) = 0.017$. The sample size is large, so results are similar to those obtained with the frequentist approach. The Bayesian posterior $P(\beta < 0)$ is comparable to a frequentist one-sided P-value for H_a: $\beta > 0$.

8.6.3 Bayesian Fitting of Multinomial Logit and Probit Models

For nominal-scale responses, Albert and Chib (1993) presented Bayesian fitting of the multinomial probit model, using the connection with the latent utility model for maxima of normal random variates outlined in Section 8.1.6. We discuss this in terms of the discrete-choice form of the model outlined in Section 8.5.4, since standard models that do not have characteristics of the choices as explanatory variables are special cases. The underlying model for the utility for subject i making response j is

$$U_{ij} = \boldsymbol{\beta}^T \boldsymbol{x}_{ij} + \epsilon_{ij}$$

and the response choice is the value of j having maximum utility.

Let $\boldsymbol{U}_i = (U_{i1}, \ldots, U_{iJ})$. When the errors are independent standard normal and we use a diffuse normal prior for $\boldsymbol{\beta}$, a Gibbs sampling scheme approximates the posterior distribution by successively sampling from the normal conditional densities of (1) $\boldsymbol{\beta}$ given $\boldsymbol{y}, \boldsymbol{U}_1, \ldots, \boldsymbol{U}_N$, and (2) $\boldsymbol{U}_1, \ldots, \boldsymbol{U}_N$ given $\boldsymbol{y}$ and $\boldsymbol{\beta}$. In case (2) the distribution is truncated to reflect that if $y_i = j$ then component j of $\boldsymbol{U}_i$ is its maximum. The model can also extend to let the utility components be correlated by introducing a parameter $\boldsymbol{\theta}$ for the covariance matrix, such as a common correlation. A third step of the Gibbs sampling then includes sampling from its conditional density.

McCulloch et al. (2000) also dealt with multinomial probit models in terms of the underlying latent model. They noted the difficulty in placing priors on a covariance matrix that incorporate an identifiability constraint such as $\text{var}(U_{i1}) = 1$, and proposed priors that can account for that constraint. See also Imai and van Dyk (2005).

For baseline-category logit models, such routines do not connect with standard ones for normal variables, because the utility construction uses errors with extreme value distributions. However, it is not necessary to base computations on latent variable models or on the general discrete-choice version of the model, and with normal priors for the model parameters, software is widely available. With relatively diffuse priors, substantive results are usually similar to those with corresponding probit models.

Note, however, that if you place simple structure such as a common variance for the priors for $\beta_{1k}, \beta_{2k}, \ldots, \beta_{J-1,k}$ in model (8.1), posterior results then depend somewhat on the choice of baseline category, because an effect relative to a pair of nonbaseline categories, $\beta_{jk} - \beta_{j'k}$, then has twice the prior variance. Alternatively, you can overparameterize by adding β_{Jk} to the model with the same prior but focus on the posterior differences for interpretation. The same remark applies to factors in such models, as results should ideally be invariant to the choice of a baseline category for indicators. One way to do this is to conduct the analysis in terms of corresponding Poisson loglinear models, introduced in

Table 8.13 Estimated Size Effects and Standard Errors in Multinomial Logistic Model for Alligator Food Choice, Using Size and Lake as Predictors

	Maximum Likelihood		Bayes, Prior $\sigma = 100$		Bayes, Prior $\sigma = 1$	
Baseline Logit	$\hat{\beta}_{1j}$	SE	$\hat{\beta}_{1j}$	Std. Dev.	$\hat{\beta}_{1j}$	Std. Dev.
$\log(\pi_I/\pi_F)$	1.46	0.40	1.52	0.40	1.26	0.38
$\log(\pi_R/\pi_F)$	−0.35	0.58	−0.39	0.60	−0.55	0.48
$\log(\pi_B/\pi_F)$	−0.63	0.64	−0.68	0.67	−0.36	0.51
$\log(\pi_O/\pi_F)$	0.33	0.45	0.35	0.46	0.23	0.43

I, invertebrate; R, reptile; B, bird; O, other; F, fish.

the next chapter, which need not identify a baseline category for any categorical variable (Gelman et al. 2004, pp. 431–433). See `www.stat.ufl.edu/~aa/cda/cda.html` for details in terms of the following example. For examples of Bayesian uses of such models, see references cited in Note 8.12.

8.6.4 Example: Alligator Food Choice Revisited

For the alligator food choice data introduced in Section 8.1.2, we found that the probability of selecting a particular food choice was described well by a model with additive effects of size s and indicators contrasting lakes Hancock, Oklawaha, and Trafford with George. For the baseline choice of fish, the model is

$$\log(\pi_j/\pi_F) = \alpha_j + \beta_{1j}s + \beta_{2j}z_H + \beta_{3j}z_O + \beta_{4j}z_T, \quad j = 1, 2, 3, 4.$$

As there was little prior information, especially about the lake effects, we fitted the model using diffuse independent normal prior distributions. Table 8.13 shows posterior means and standard deviations for the size effect, when we parameterize in such a way that the 10 conditional log odds ratios relating size to pairs of food choice categories all have normal distributions with $\mu = 0$ and $\sigma = 100$. Corresponding ML estimates and SE values are also shown. With such uninformative priors, results are quite similar. With either analysis, we conclude that the smaller alligators are relatively more likely to have invertebrates as their primary food choice.

To compare with results from a highly informative Bayesian analysis, we used normal priors for the ten log odds ratios between size and pairs of food choices with each $\sigma = 1$. Table 8.13 shows results. Having more prior information centered at 0 results in shrinkage of posterior estimates and standard deviations toward 0.

NOTES

Section 8.1: Nominal Responses: Baseline-Category Logit Models

8.1 BCL models: Baseline-category logit models were developed in Bock (1970), Haberman (1974a, pp. 352–373), Mantel (1966), Skrondal and Rabe-Hesketh (2003, 2004, Chap. 13), and Theil (1969, 1970). Lesaffre and Albert (1989) presented regression diagnostics. Amemiya

(1981), Haberman (1982), and Theil (1970) presented R-squared measures. Baker (1994), Lang (1996), and Tsodikov and Chefo (2008) showed connections with Poisson models. Kosmidis and Firth (2011) used this connection in giving a penalized likelihood for bias reduction. Tutz and Schauberger (2012) proposed graphics for effects in multinomial response models.

Section 8.2: Ordinal Responses: Cumulative Logit Models

8.2 **Cumulative logits**: Early uses of cumulative logit models include Bock and Jones (1968), Simon (1974), Snell (1964), Walker and Duncan (1967), and Williams and Grizzle (1972). McCullagh (1980) popularized the proportional odds case. Later articles include Agresti and Lang (1993), Hastie and Tibshirani (1987), Peterson and Harrell (1990), and Tutz (1989). See also Note 12.2 and Sections 12.2.3 and 13.4.1.

8.3 **Score test, power, efficiency**: For $2 \times J$ tables and the model $\text{logit}[P(Y \leq j)] = \alpha_j + \beta x$, with x an indicator, McCullagh (1980) noted that the score test of H_0: $\beta = 0$ is equivalent to a discrete version of the Wilcoxon–Mann–Whitney test. Whitehead (1993) gave sample size formulas for this case. The sample size n_J needed for a certain power decreases as J increases: When response categories have equal probabilities, $n_J \approx 0.75 n_2/(1 - 1/J^2)$. The efficiency loss is major in collapsing to $J = 2$. See also Rabbee et al. (2003). Natarajan et al. (2012) extended the score test to complex sample survey data. Edwardes (1997) innovatively adapted the test by treating the cutpoints as random. Rice et al. (2012) discussed ways of dealing with variation in cutpoints.

8.4 **ROC curve**: As a way of evaluating diagnostic tests that have $J > 2$ ordered response categories rather than (positive, negative), an ROC curve can refer to the various possible cutoffs for defining a result to be positive. It plots sensitivity against 1 − specificity for the possible collapsings of the J categories to a (positive, negative) scale (Toledano and Gatsonis 1996).

Section 8.3: Ordinal Responses: Alternative Models

8.5 **Probit, generalized links**: Cumulative probit models were proposed by Aitchison and Silvey (1957) for the one-way layout setting and Gurland et al. (1960) and Bock and Jones (1968, Chap. 8) in a general regression setting. McKelvey and Zavoina (1975) presented the underlying latent normal model. Genter and Farewell (1985) introduced a generalized link function that permits comparison of fits provided by probit, complementary log–log, and other links. Adjacent-categories logit models and models equivalent to them were presented by Goodman (1979a, 1983), Haberman (1974b), and Simon (1974). Greene and Hensher (2010) presented other ordinal modeling strategies.

8.6 **Hazard/survival**: The ratio of a pdf to the complement of the cdf is the *hazard function* (Exercise 4.20). For discrete variables, this is the ratio found in continuation-ratio logits. The model $\log[-\log(1 - \omega_j(\boldsymbol{x}))] = \alpha_j + \boldsymbol{\beta}^T \boldsymbol{x}$ is a discrete-time version of the proportional hazards model (Allison 1982, Aranda-Ordaz 1983, Prentice and Gloeckler 1978, Thompson 1977). Läärä and Matthews (1985) showed this is equivalent to the model using the same link for cumulative probabilities.

8.7 **OLS fitting**: Assigning scores to ordered response categories and using ordinary least-squares regression modeling is not optimal, because the observations do not have constant variance. Instead treating the response as multinomial, with categorical predictors Bhapkar (1968), Grizzle et al. (1969), and Williams and Grizzle (1972) used weighted least squares, and Haber (1985) and Lipsitz (1992) used ML. For large J, such models approximate a regression model for continuous Y. A structural difficulty is that the model can have predicted means outside the range of assigned scores. Also, "floor effects" and "ceiling effects" can occur when a latent

response is categorized and a linear model is fitted to the observed response. See Agresti (2010, Sec. 1.3, 5.6) for details.

8.8 **Dispersion effects**: McCullagh (1980) generalized the cumulative link model to incorporate dispersion effects. With link function g, the model is

$$g[P(Y \leq j)] = \frac{\alpha_j - \boldsymbol{\beta}^T \boldsymbol{x}}{\exp(\boldsymbol{\gamma}^T \boldsymbol{x})}.$$

The denominator contains scale parameters $\boldsymbol{\gamma}$ that describe how the dispersion depends on $\boldsymbol{x}$. This model arises from a latent variable model in which the distribution of Y^* has shape reflected by g, such as normal for the probit link. The latent variable has $E(Y^*) = \boldsymbol{\beta}^T \boldsymbol{x}$ and standard deviation $\exp(\boldsymbol{\gamma}^T \boldsymbol{x})$ that varies as $\boldsymbol{x}$ does. See also Agresti (2010, Sec. 5.4) and Cox (1995). Hamada and Wu (1990) and Nair (1987) presented alternatives models for detecting dispersion effects.

Section 8.4: Testing Conditional Independence in I × J × K Tables

8.9 **Generalized CMH**: Birch (1965), Landis et al. (1978), Mantel (1963), and Mantel and Byar (1978) generalized the CMH statistic. Let the Kronecker product $\boldsymbol{B}_k = \boldsymbol{u}_k \otimes \boldsymbol{v}_k$ denote a matrix of constants based on row scores $\boldsymbol{u}_k$ and column scores $\boldsymbol{v}_k$ for stratum k. The Landis et al. (1978) generalized statistic is

$$L^2 = \Big[\sum_k \boldsymbol{B}_k(\boldsymbol{n}_k - \boldsymbol{\mu}_k)\Big]^T \Big[\sum_k \boldsymbol{B}_k \boldsymbol{V}_k \boldsymbol{B}_k^T\Big]^{-1} \Big[\sum_k \boldsymbol{B}_k(\boldsymbol{n}_k - \boldsymbol{\mu}_k)\Big].$$

When $\boldsymbol{u}_k = (u_1, \ldots, u_I)$ and $\boldsymbol{v}_k = (v_1, \ldots, v_J)$ for all strata, $L^2 = M^2$ in (8.19). When $\boldsymbol{u}_k$ is an $(I-1) \times I$ matrix $(\boldsymbol{I}, -\boldsymbol{1})$, where $\boldsymbol{I}$ is an identity matrix of size $(I-1)$ and $\boldsymbol{1}$ denotes a column vector of $I-1$ ones, and $\boldsymbol{v}_k$ is the analogous matrix of size $(J-1) \times J$, L^2 simplifies to (8.18) with df $= (I-1)(J-1)$. With this $\boldsymbol{u}_k$ and $\boldsymbol{v}_k = (v_1, \ldots, v_J)$, L^2 sums over the strata information about how I row means compare to their null expected values, and it has df $= I-1$. Rank score versions are analogs for ordered categorical responses of stratum-adjusted Spearman correlation and Kruskal–Wallis tests. Kawaguchi et al. (2011) extended the Mantel–Haenszel odds ratio estimate to stratified Mann–Whitney estimators that utilize probability comparisons of two groups [related to Δ in (2.15)]. Landis et al. (2005) and Stokes et al. (2012) reviewed CMH methods. Koch et al. (1982) reviewed related methods.

8.10 **Small-sample tests of conditional independence**: To eliminate nuisance parameters, small-sample tests condition on row and column totals in each partial table. Section 7.3.5 showed this for $2 \times 2 \times K$ tables. When $I > 2$ and/or $J > 2$, the conditional distribution of cell counts in each stratum is the multivariate hypergeometric (Section 16.5.1), and this propagates an exact conditional distribution for the test statistic of interest, such as a generalized CMH statistic (Kim and Agresti 1997).

Section 8.5: Discrete-Choice Models

8.11 **McFadden/Bradley–Terry/Luce**: McFadden's model relates to models proposed by Bradley and Terry (1952) (see Section 11.6) and Luce (1959). Train's (2009) overview text includes many generalized models, and pages 45–50 discuss the independence from irrelevant alternatives assumption and references articles dealing with testing whether that property holds. One

approach uses standard tests to compare it to a more complex nested logit model mentioned in Section 8.5.5.

Section 8.6: Bayesian Modeling of Multinomial Responses

8.12 **Bayes multinomial**: For other discussion of utilizing the connection with an underlying latent variable model, see Hoff (2009, Sec. 12.1) and Johnson and Albert (1999, Chap. 4). See also Congdon (2005, Chap. 7), and many references in Agresti (2010, Chap. 11). For comparing two ordinal categorical distributions, Altham (1969) provided a Bayesian estimate of the probability that one distribution is stochastically higher than the other. For Bayesian inference with baseline-category logit models, see Congdon (2005, Chap. 6), Daniels and Gatsonis (1997), Holmes and Held (2006), Leonard and Hsu (1994), and Sha et al. (2004).

EXERCISES

Applications

8.1 For Table 8.14, let $Y =$ belief in existence of heaven, $x_1 =$ gender ($1 =$ females, $0 =$ males), and $x_2 =$ race ($1 =$ blacks, $0 =$ whites). Table 8.15 shows the fit of the model

$$\log(\pi_j/\pi_3) = \alpha_j + \beta_j^G x_1 + \beta_j^R x_2, \quad j = 1, 2,$$

with *SE* values in parentheses.

a. Find the prediction equation for $\log(\pi_1/\pi_2)$.

b. Using the *yes* and *no* response categories, interpret the conditional gender effect using a 95% confidence interval for an odds ratio.

Table 8.14 Data on Belief in Existence of Heaven for Exercise 8.1

		Belief in Heaven		
Race	Gender	Yes	Unsure	No
Black	Female	88	16	2
	Male	54	7	5
White	Female	397	141	24
	Male	235	189	39

Source: 2008 General Social Survey.

Table 8.15 Fit of Model for Belief in Heaven for Exercise 8.1

	Belief Categories for Logit	
Parameter	Yes/No	Unsure/No
Intercept	1.785 (0.168)	1.554 (0.172)
Gender	1.044 (0.259)	0.254 (0.269)
Race	0.703 (0.411)	−0.106 (0.438)

c. Find $\hat{\pi}_1 = \hat{P}(Y = \text{yes})$ for white females.

d. Without calculating estimated probabilities, explain why the intercept estimates indicate that for white males, $\hat{\pi}_1 > \hat{\pi}_2 > \hat{\pi}_3$. Use the intercept and gender estimates to show that the same ordering applies for black females.

e. Without calculating estimated probabilities, explain why the estimates in the gender row indicate that $\hat{\pi}_1$ is higher for females than for males, for each race.

f. For this fit, $G^2 = 0.69$. Explain why residual df $= 2$. Deleting the gender effect, $G^2 = 47.64$. Conduct a likelihood-ratio test of whether opinion is independent of gender, given race. Interpret.

8.2 A model fit predicting preference for U.S. President (Democrat, Republican, Independent) using $x =$ annual income (in \$10,000) is $\log(\hat{\pi}_D/\hat{\pi}_I) = 3.3 - 0.2x$ and $\log(\hat{\pi}_R/\hat{\pi}_I) = 1.0 + 0.3x$.

a. Find the prediction equation for $\log(\hat{\pi}_R/\hat{\pi}_D)$ and interpret the slope. For what range of x is $\hat{\pi}_R > \hat{\pi}_D$?

b. Find the prediction equation for $\hat{\pi}_I$.

c. Plot $\hat{\pi}_D$, $\hat{\pi}_I$, and $\hat{\pi}_R$ for x between 0 and 10, and interpret.

8.3 Table 8.16 shows recent GSS data for the effect of gender and race on political party identification. Find a baseline-category logit model that fits well. Interpret estimated effects on the odds that party identification is Democrat instead of Republican.

Table 8.16 Data for Exercise 8.3 on Political Party ID

		Political Party Identification		
Gender	Race	Democrat	Republican	Independent
Male	White	132	176	127
	Black	42	6	12
Female	White	172	129	130
	Black	56	4	15

8.4 For 63 alligators caught in Lake George, Florida, Table 8.17 classifies primary food choice as (fish, invertebrate, other) and shows length in meters. Alligators are called subadults if length < 1.83 meters (6 feet) and adults if length > 1.83 meters.

a. Measuring length as (adult, subadult), find a model that adequately describes effects of gender and length on food choice. Interpret the effects. For adult females, find the estimated probabilities of the food choice categories.

b. Using only observations for which primary food choice was fish or invertebrate, find a model that adequately describes effects of gender and binary length. Compare parameter estimates and standard errors for this separate-fitting approach to those obtained with simultaneous fitting, including the other category.

c. Treating length as binary loses information. Adapt the model in part (a) to use the continuous length measurements. Interpret, explaining how the estimated outcome probabilities vary with length. Find the estimated length at which the invertebrate and other categories are equally likely.

Table 8.17 Data for Exercise 8.4[a] on Alligator Food Choice

Males				Females			
Length (m)	Choice	Length (m)	Choice	Length (m)	Choice	Length (m)	Choice
1.30	*I*	1.70	*I*	3.33	*F*	1.78	*O*
1.32	*F*	1.73	*O*	3.56	*F*	1.80	*I*
1.32	*F*	1.78	*F*	3.58	*F*	1.88	*I*
1.40	*F*	1.78	*O*	3.66	*F*	2.16	*F*
1.42	*I*	1.80	*F*	3.68	*O*	2.26	*F*
1.42	*F*	1.85	*F*	3.71	*F*	2.31	*F*
1.47	*I*	1.93	*I*	3.89	*F*	2.36	*F*
1.47	*F*	1.93	*F*	1.24	*I*	2.39	*F*
1.50	*I*	1.98	*I*	1.30	*I*	2.41	*F*
1.52	*I*	2.03	*F*	1.45	*I*	2.44	*F*
1.63	*I*	2.03	*F*	1.45	*O*	2.56	*O*
1.65	*O*	2.31	*F*	1.55	*I*	2.67	*F*
1.65	*O*	2.36	*F*	1.60	*I*	2.72	*I*
1.65	*I*	2.46	*F*	1.60	*I*	2.79	*F*
1.65	*F*	3.25	*O*	1.65	*F*	2.84	*F*
1.68	*F*	3.28	*O*	1.78	*I*		

[a] *F*, fish; *I*, invertebrates; *O*, other.

8.5 Fit the multinomial probit model to the alligator food choice data in Table 8.1 and at the text website, with size and lake as predictors. Compare estimates and *SE* values to those in Table 8.4, and explain why they are larger for the multinomial logit model.

8.6 Fit the baseline-category logit model with main effects to the data in Table 8.5. Describe the effect of the sample having no blacks in the very happy category.

8.7 For recent GSS data, the cumulative logit model (8.5) with Y = political ideology (very liberal, slightly liberal, moderate, slightly conservative, very conservative) and x = party affiliation (1 for the 428 Democrats and 0 for the 407 Republicans) has $\hat{\beta} = 0.975$ ($SE = 0.129$) and $\hat{\alpha}_1 = -2.469$. Interpret $\hat{\beta}$. Find the estimated probability of a very liberal response for each group.

8.8 Table 8.18 is an expanded version of a data set analyzed in Section 9.4.2. The response categories are (1) not injured, (2) injured but not transported by emergency medical services, (3) injured and transported by emergency medical services but not hospitalized, (4) injured and hospitalized but did not die, and (5) injured and died. Table 8.19 shows output for a model of form (8.5).

a. Why are there four intercepts? Explain how they determine the estimated response distribution for males in urban areas wearing seat belts.

b. Construct a confidence interval for the effect of gender, given seat-belt use and location. Interpret.

Table 8.18 Data for Exercise 8.8 on Degree of Injury in Auto Accident

			Response on Injury Outcome				
Gender	Location	Seat Belt	1	2	3	4	5
Female	Urban	No	7,287	175	720	91	10
		Yes	11,587	126	577	48	8
	Rural	No	3,246	73	710	159	31
		Yes	6,134	94	564	82	17
Male	Urban	No	10,381	136	566	96	14
		Yes	10,969	83	259	37	1
	Rural	No	6,123	141	710	188	45
		Yes	6,693	74	353	74	12

Source: Data courtesy of Cristanna Cook, Medical Care Development, Augusta, Maine.

Table 8.19 Output for Exercise 8.8 on Auto Accident Injuries

Parameter			DF	Estimate	Std Error
Intercept1			1	3.3074	0.0351
Intercept2			1	3.4818	0.0355
Intercept3			1	5.3494	0.0470
Intercept4			1	7.2563	0.0914
gender	female		1	−0.5463	0.0272
gender	male		0	0.0000	0.0000
location	rural		1	−0.6988	0.0424
location	urban		0	0.0000	0.0000
seatbelt	no		1	−0.7602	0.0393
seatbelt	yes		0	0.0000	0.0000
location*seatbelt	rural	no	1	−0.1244	0.0548
location*seatbelt	rural	yes	0	0.0000	0.0000
location*seatbelt	urban	no	0	0.0000	0.0000
location*seatbelt	urban	yes	0	0.0000	0.0000

c. Find the estimated cumulative odds ratio between the response and seat-belt use for those in rural locations and for those in urban locations, given gender. Based on this, explain how the effect of seat-belt use varies by region, and explain how to interpret the interaction estimate, −0.1244.

8.9 In a class project, University of Florida students Shahrzad Farshi and Marty Parks used GSS data to study the effect of several explanatory variables on liking for rap music, an ordinal variable with five categories (1 = greatest preference). They found a good fit with the model $\text{logit}[\hat{P}(Y \leq j)] = \hat{\alpha}_j - 1.06r - 0.58a$, where race r was coded 1 for white and 0 for black/other and age a has scores (1, 2, 3, 4) for four successive age categories. Interpret these effects with cumulative odds ratios.

8.10 Table 8.20 refers to a clinical trial for the treatment of small-cell lung cancer. Patients were randomly assigned to two treatment groups. The sequential therapy administered the same combination of chemotherapeutic agents in each treatment

Table 8.20 Data for Exercise 8.10 on Lung Cancer Clinical Trial

		Response to Chemotherapy			
Therapy	Gender	Progressive Disease	No Change	Partial Remission	Complete Remission
Sequential	Male	28	45	29	26
	Female	4	12	5	2
Alternating	Male	41	44	20	20
	Female	12	7	3	1

Source: W. Holtbrugge and M. Schumacher, *Appl. Statist.* **40**: 249–259, 1991.

cycle; the alternating therapy had three different combinations, alternating from cycle to cycle.

a. Fit a cumulative logit model with main effects for therapy and gender. Interpret effect estimates.

b. For the therapy effect, compare $\hat{\beta}_1$ to the estimate obtained when the model is fitted to the binary response obtained by combining the first two response categories and combining the last two response categories. What property of the model does this reflect?

c. For the collapsing in (b), compare $\hat{\beta}_1/SE$ to the ratio obtained for the uncollapsed response. (Usually, a disadvantage of collapsing ordinal responses is that the significance of effects diminishes.)

d. Fit the model to the uncollapsed data that also contains an interaction term. Interpret. Does it fit better? Explain why it is equivalent to using the four gender–therapy combinations as levels of a single factor.

8.11 A study of factors affecting alcohol consumption measures the response variable with the scale (abstinence, a drink a day or less, more than one drink a day). For a comparison of two groups while adjusting for relevant covariates, the researchers hypothesize that the two groups will have about the same prevalence of abstinence, but that one group will have a considerably higher proportion who have more than one drink a day. Even though the response variable is ordinal, explain why a cumulative logit model with proportional odds structure may be inadequate for this study.

8.12 Refer to Table 8.14. Treating belief in heaven as ordinal, fit and interpret a (**a**) cumulative logit model and (**b**) cumulative probit model. Compare results and state interpretations in each case.

8.13 For the cumulative probit model fitted to Table 8.5, find the means and standard deviation for the two normal cdf's that provide the curves for $\hat{P}(Y = 1)$ as a function of x_1 = number of traumatic events, at the two levels of x_2 = race. Interpret the effects.

8.14 For Table 8.5, fit and interpret effects for a (**a**) cumulative link model with complementary log–log link and (**b**) continuation-ratio logit model.

8.15 Refer to Exercise 8.7. With adjacent-categories logit model (8.10), $\hat{\beta} = 0.435$. Interpret using odds ratios for adjacent categories and for the (very liberal, very conservative) pair of categories.

8.16 For the developmental toxicity data in Table 8.7, formulate and fit a continuation-ratio logit model with proportional odds structure. [*Hint*: Create a data file of independent binomials and then construct a model matrix that has the desired model structure.]

8.17 Table 8.21 refers to a study that randomly assigned subjects to a control or treatment group. Daily during the study, treatment subjects ate cereal containing psyllium. The study analyzed the effect on LDL cholesterol.

a. Model the ending cholesterol level as a function of treatment, using the beginning level as a covariate. Interpret the treatment effect.

b. Repeat part (a), now treating the beginning level as qualitative. Compare results.

Table 8.21 Data for Exercise 8.17 on Cholesterol and Cereal

	Ending LDL Cholesterol Level							
	Control				Treatment			
Beginning	≤ 3.4	3.4–4.1	4.1–4.9	> 4.9	3.4	3.4–4.1	4.1–4.9	> 4.9
≤ 3.4	18	8	0	0	21	4	2	0
3.4–4.1	16	30	13	2	17	25	6	0
4.1–4.9	0	14	28	7	11	35	36	6
> 4.9	0	2	15	22	1	5	14	12

Source: Data courtesy of Sallee Anderson, Kellogg Co.

8.18 The book's website (`www.stat.ufl.edu/~aa/cda/cda.html`) has a $3 \times 4 \times 4$ table that cross-classifies dumping severity (Y) and operation (X) for four hospitals (H). The four operations refer to treatments for duodenal ulcer patients and have a natural ordering. Dumping severity describes a possible undesirable side effect of the operation. Its three categories are also ordered. Table 8.22 shows results of generalized CMH tests. For each test, give a pair of models such that a likelihood-ratio test comparing those models would give similar results. Explain how one test can be much more significant than the others.

Table 8.22 Results for Dumping Severity Data of Exercise 8.18

Summary Statistics for dumping by operate Controlling for hospital				
Statistic	Alternative Hypothesis	DF	Value	Prob
1	Nonzero Correlation	1	6.3404	0.0118
2	Row Mean Scores Differ	3	6.5901	0.0862
3	General Association	6	10.5983	0.1016

8.19 A sample of subjects indicate their favorite among four Margarita pizzas characterized by 1 = (thin crust, normal cheese), 2 = (thin crust, extra cheese), 3 = (thick crust, normal cheese), 4 = (thick crust, extra cheese). For the characteristics of the choices x_1 = crust type (1 = thick, 0 = thin) and x_2 = cheese quantity (1 = extra, 0 = normal), the multinomial discrete choice model (8.20) has $\beta_1 = -0.40$ and $\beta_2 = 0.60$. For each pizza type, find the probability that it is the favorite.

8.20 Refer to the previous exercise. For a random sample of 20 pizza lovers, suppose 4 prefer choice 1, 8 prefer choice 2, 3 prefer choice 3, and 5 prefer choice 4. Fit the model and interpret the estimates.

8.21 Describe an application in which a discrete-choice model would be useful. Specify potential explanatory variables, and identify which are characteristics of the chooser and which are characteristics of the choices.

8.22 A cafe has four entrées: chicken, beef, fish, vegetarian. Specify a model of form (8.20) for the selection of an entrée using x = gender (1 = female, 0 = male) and u = cost of entrée, which is a characteristic of the choices. Interpret the model parameters.

8.23 For Table 8.14 on belief in heaven, use Bayesian methods to fit the model of Exercise 8.1. Do this once with uninformative priors (say, $\sigma = 100$) and once with very informative priors (say, $\sigma = 1$). In each case, for the gender effect on the (yes/no) logit, report the posterior mean and standard deviation and the 95% posterior interval. Compare results between them and with the ML estimate, *SE*, and 95% confidence interval.

8.24 In the previous exercise, treat belief in heaven as ordinal and reanalyze with Bayesian methods. Compare results for the gender effect, and interpret.

8.25 Consider the baseline-category logit model of Section 8.6.4 for Bayesian modeling of alligator food choice in terms of size and lake. Try to replicate results in Table 8.13 for $\sigma = 1$. (If your results differ much, for your parameterization the 10 conditional log odds ratios relating size to pairs of food choices may not all have prior $\sigma = 1$.)

8.26 Is political ideology associated with happiness? Conduct a Bayesian analysis for the data in Table 3.7, using a model presented in this chapter. Present a posterior interval and posterior probability that addresses the question, and interpret results.

8.27 Analyze Table 8.5 with two types of model studied in this chapter. Write a report summarizing results and advantages and disadvantages of each modeling strategy.

8.28 This book's website has a $4 \times 2 \times 3 \times 3$ table that cross-classifies a sample of residents of Copenhagen on type of housing (H), degree of contact with other residents (C), feeling of influence on apartment management (I), and satisfaction with housing conditions (S). Treating S as the response variable, analyze these data.

Theory and Methods

8.29 A multivariate generalization of the exponential dispersion family (4.17) is

$$f(\boldsymbol{y}_i;\boldsymbol{\theta}_i,\phi) = \exp\{[\boldsymbol{y}_i^T\boldsymbol{\theta}_i - b(\boldsymbol{\theta}_i)]/a(\phi) + c(\boldsymbol{y}_i,\phi)\},$$

where $\boldsymbol{\theta}_i$ is the natural parameter. Show that a multinomial variate $\boldsymbol{y}_i$ for a single trial with parameters $\{\pi_j, j = 1,\ldots,J-1\}$ is in the $(J-1)$-parameter exponential family, with baseline-category logits as natural parameters.

8.30 Cell counts $\{y_{ij}\}$ in an $I \times J$ contingency table have a multinomial $(n;\{\pi_{ij}\})$ distribution. Show that $\{P(Y_{ij} = n_{ij})\}$ can be expressed as

$$d^n\, n! \prod_i \prod_j (n_{ij}!)^{-1} \exp\left[\sum_{i=1}^{I-1}\sum_{j=1}^{J-1} n_{ij}\log(\alpha_{ij}) + \sum_{i=1}^{I-1} n_{i+}\log(\pi_{iJ}/\pi_{IJ}) + \sum_{j=1}^{J-1} n_{+j}\log(\pi_{Ij}/\pi_{IJ})\right],$$

where $\alpha_{ij} = \pi_{ij}\pi_{IJ}/\pi_{iJ}\pi_{Ij}$ and d is a constant independent of the data. Find an alternative expression using local odds ratios $\{\theta_{ij}\}$, by showing that

$$\sum_i\sum_j n_{ij}\log\alpha_{ij} = \sum_i\sum_j s_{ij}\log\theta_{ij}, \quad \text{where} \quad s_{ij} = \sum_{a\le i}\sum_{b\le j} n_{ab}.$$

(Hence, models for such parameters have reduced sufficient statistics and relatively simple score statistics for testing effects.)

8.31 Consider the baseline-category logit model expressed as

$$\pi_j(\boldsymbol{x}) = \frac{\exp(\alpha_j + \boldsymbol{\beta}_j^T\boldsymbol{x})}{\sum_{h=1}^{J}\exp(\alpha_h + \boldsymbol{\beta}_h^T\boldsymbol{x})}.$$

Show that dividing numerator and denominator by $\exp(\alpha_J + \boldsymbol{\beta}_J^T\boldsymbol{x})$ yields new parameters $\alpha_j^* = \alpha_j - \alpha_J$ and $\beta_j^* = \beta_j - \beta_J$ that satisfy $\alpha_J = 0$ and $\boldsymbol{\beta}_J = \mathbf{0}$. Thus, without loss of generality, we can take $\alpha_J = 0$ and $\boldsymbol{\beta}_J = \mathbf{0}$.

8.32 When there are $J = 3$ outcome categories, suppose that

$$\pi_j(x) = \exp(\alpha_j + \beta_j x)/[1 + \exp(\alpha_1 + \beta_1 x) + \exp(\alpha_2 + \beta_2 x)],$$

$j = 1, 2$. Show that $\pi_3(x)$ is (**a**) decreasing in x if $\beta_1 > 0$ and $\beta_2 > 0$, (**b**) increasing in x if $\beta_1 < 0$ and $\beta_2 < 0$, and (**c**) nonmonotone when β_1 and β_2 have different signs.

8.33 Refer to the log-likelihood function for the baseline-category logit model (Section 8.1.4). Denote the sufficient statistics by $np_j = \sum_i y_{ij}$ and $S_{jk} = \sum_i x_{ik}y_{ij}$,

$j = 1, \ldots, J$, $k = 1, \ldots, p$. Let $\boldsymbol{S} = (S_{11}, \ldots, S_{1p}, \ldots S_{J1}, \ldots, S_{Jp})^T$. Under the null hypothesis that explanatory variables have no effect, conditional on $\sum_i y_{ij}$, $j = 1, \ldots, J$, show that

$$E(\boldsymbol{S}) = n(\boldsymbol{p} \otimes \boldsymbol{m}), \quad \text{var}(\boldsymbol{S}) = n(\boldsymbol{V} \otimes \boldsymbol{\Sigma}),$$

where $\boldsymbol{p} = (p_1, \ldots, p_J)^T$, $\boldsymbol{m} = (\bar{x}_1, \ldots, \bar{x}_p)^T$ with $\bar{x}_k = \left(\sum_i x_{ik}\right)/n$, $\boldsymbol{\Sigma}$ has elements $s_{kv}^2 = \left[\sum_i (x_{ik} - \bar{x}_k)(x_{iv} - \bar{x}_v)\right]/(n-1)$, and $\boldsymbol{V}$ has elements $v_{ii} = p_i(1 - p_i)$ and $v_{ij} = -p_i p_j$ (Zelen 1991).

8.34 An alternative fitting approach for the baseline-category logit model (8.1) fits binary logistic models separately for the $J - 1$ pairings of responses. The estimates have larger *SE* than the ML estimates for simultaneous fitting of the $J - 1$ logits, but Begg and Gray (1984) showed that the efficiency loss is minor when the response category having highest prevalence is the baseline. Illustrate, by showing that the fit using categories *I* and *F* alone of the alligator data is $\log(\hat{\pi}_I/\hat{\pi}_F) = -1.69 + 1.66s - 1.78z_H + 1.05z_O + 1.22z_T$, with *SE* values (0.43, 0.62, 0.49, 0.52) for the effects. Compare with the first row of Table 8.4.

8.35 For explanatory variable k in a baseline-category logit model, suppose the model matrix constrains $\beta_{2k} = \cdots = \beta_{Jk} = 0$, leaving β_{1k} unconstrained. Explain how β_{1k} then describes a contrast for that variable between outcome category 1 and the other categories combined. Explain how to generalize this to contrast one subset of the categories to the other categories.

8.36 Explain why the cumulative logit model of proportional odds form is not a special case of a baseline-category logit model.

8.37 Consider the cumulative logit model, $\text{logit}[P(Y \leq j)] = \alpha_j + \beta_j x$, not having proportional odds form.

a. With continuous x taking values over the real line, show that the model is improper in that cumulative probabilities are misordered for a range of x values.

b. When x is a binary indicator, explain why the model is proper but requires constraints on $(\alpha_j + \beta_j)$ (as well as the usual ordering constraint on $\{\alpha_j\}$) and is then equivalent to the saturated model.

8.38 For an $I \times J$ contingency table with ordinal Y and scores $\{x_i = i\}$, consider the model

$$\text{logit}[P(Y \leq j | X = x_i)] = \alpha_j + \beta x_i. \tag{8.23}$$

a. Show that $\text{logit}[P(Y \leq j | X = x_{i+1})] - \text{logit}[P(Y \leq j | X = x_i)] = \beta$ is a log cumulative odds ratio for the 2×2 table consisting of rows i and $i + 1$ and the binary response having cutpoint following category j. Thus, (8.23) is a *uniform association model* in cumulative odds ratios.

b. Show that (i) residual df $= IJ - I - J$ and (ii) $\beta = 0$ corresponds to independence of X and Y.

c. Using the same linear predictor but with adjacent-categories logits, show that uniform association applies to the local odds ratios (2.10).

8.39 A cumulative link model for an $I \times J$ table with a qualitative predictor is

$$G^{-1}[P(Y \leq j)] = \alpha_j + \mu_i, \quad i = 1, \ldots, I, \ j = 1, \ldots, J-1.$$

Show that (**a**) residual df = $(I-1)(J-2)$, (**b**) independence corresponds to $\mu_1 = \cdots = \mu_I$, (**c**) the test of independence has df = $I-1$, and (**d**) the rows are stochastically ordered on Y.

8.40 Prove factorization (8.14) for the multinomial distribution.

8.41 A response scale has the categories (strongly agree, mildly agree, mildly disagree, strongly disagree, don't know). One model uses a logistic part for P(don't know) and a separate ordinal part for the ordered categories conditional on response in one of those categories. Explain how to construct a likelihood function to do this simultaneously.

8.42 For cumulative link model (8.7), show that for $1 \leq j < k \leq J-1$, $P(Y \leq k|\boldsymbol{x}) = P(Y \leq j|\boldsymbol{x}^*)$, where $\boldsymbol{x}^*$ is obtained by increasing the ith component of $\boldsymbol{x}$ by $(\alpha_k - \alpha_j)/\beta_i$. Interpret.

8.43 When X and Y are ordinal, explain how to test conditional independence by allowing a different trend in each partial table. [*Hint*: Generalize model (8.17) by replacing β by β_k.]

8.44 Consider equation (8.21) and the condition of independence from irrelevant alternatives. Explain why this condition does not hold for the multinomial probit model.

8.45 For a Bayesian analysis, explain why the posterior $P(\beta \leq 0)$ is analogous to the frequentist P-value for H_a: $\beta > 0$.

8.46 After fitting a cumulative logit model of proportional odds form, what might you do to check the model (**a**) as part of a frequentist analysis and (**b**) as part of a Bayesian analysis?