CHAPTER 6

Multinomial Response Models

In Chapter 5 we presented generalized linear models (GLMs) for binary response variables that assume a *binomial* random component. GLMs for multicategory response variables assume a *multinomial* random component. In this chapter we present generalizations of logistic regression for multinomial response variables. Separate models are available for nominal response variables and for ordinal response variables.

In Section 6.1 we present a model for nominal response variables. It uses a separate binary logistic equation for each pair of response categories. An important type of application analyzes effects of explanatory variables on a person's choice from a discrete set of options, such as a choice of product brand to buy. In Section 6.2 we present a model for ordinal response variables. It applies the logit or some other link simultaneously to all the cumulative response probabilities, such as to model whether the importance of religion to a person is below or above some point on a scale (unimportant, slightly important, moderately important, very important). A parsimonious version of the model uses the same effect parameters for each logit. Section 6.3 presents examples and discusses model selection for multicategory responses.

We denote the number of response categories by c. For subject i, let π_{ij} denote the probability of response in category j, with $\sum_{j=1}^{c} \pi_{ij} = 1$. The category choice is the result of a single multinomial trial. Let $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{ic})$ represent the multinomial trial for subject i, $i = 1, \ldots, N$, where $y_{ij} = 1$ when the response is in category j and $y_{ij} = 0$ otherwise. Then $\sum_j y_{ij} = 1$, and the multinomial probability distribution for that subject is

$$p(y_{i1}, \ldots, y_{ic}) = \pi_{i1}^{y_{i1}} \cdots \pi_{ic}^{y_{ic}}.$$

In this chapter we express models in terms of such ungrouped data. As with binary data, however, with discrete explanatory variables it is better to group the N observations according to their multicategory trial indices $\{n_i\}$ before forming the deviance and other goodness-of-fit statistics and residuals.

Foundations of Linear and Generalized Linear Models, First Edition. Alan Agresti.

6.1 NOMINAL RESPONSES: BASELINE-CATEGORY LOGIT MODELS

For nominal-scale response variables having c categories, multicategory logistic models simultaneously describe the log odds for all $c(c-1)/2$ pairs of categories. Given a certain choice of $c-1$ of these, the rest are redundant.

6.1.1 Baseline-Category Logits

We construct a multinomial logistic model by pairing each response category with a baseline category, such as category c, using

$$\log \frac{\pi_{i1}}{\pi_{ic}}, \ \log \frac{\pi_{i2}}{\pi_{ic}}, \ldots, \log \frac{\pi_{i,c-1}}{\pi_{ic}}.$$

The jth *baseline-category logit*, $\log(\pi_{ij}/\pi_{ic})$, is the logit of a conditional probability,

$$\begin{aligned}\text{logit}[P(y_{ij} = 1 \mid y_{ij} = 1 \text{ or } y_{ic} = 1)] \\ = \log\left[\frac{P(y_{ij} = 1 \mid y_{ij} = 1 \text{ or } y_{ic} = 1)}{1 - P(y_{ij} = 1 \mid y_{ij} = 1 \text{ or } y_{ic} = 1)}\right] = \log \frac{\pi_{ij}}{\pi_{ic}}.\end{aligned}$$

Let $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})$ denote explanatory variable values for subject i, and let $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jp})^{\mathrm{T}}$ denote parameters for the jth logit.

Baseline-category logit model:

$$\log \frac{\pi_{ij}}{\pi_{ic}} = \boldsymbol{x}_i\boldsymbol{\beta}_j = \sum_{k=1}^{p} \beta_{jk}x_{ik}, \quad j = 1, \ldots, c-1. \tag{6.1}$$

This model, also often called the *multinomial logit model*, simultaneously describes the effects of $\boldsymbol{x}$ on the $c-1$ logits. The effects vary according to the response paired with the baseline.

These $c-1$ equations determine equations for logits with other pairs of response categories, since

$$\log \frac{\pi_{ia}}{\pi_{ib}} = \log \frac{\pi_{ia}}{\pi_{ic}} - \log \frac{\pi_{ib}}{\pi_{ic}} = \boldsymbol{x}_i(\boldsymbol{\beta}_a - \boldsymbol{\beta}_b).$$

As in other models, typically $x_{i1} = 1$ for the coefficient of an intercept term, which also differs for each logit. The model treats the response variable as nominal scale, in the following sense: if the model holds and the outcome categories are permuted in any way, the model still holds with the corresponding permutation of the effects.

We can express baseline-category logit models directly in terms of response probabilities $\{\pi_{ij}\}$ by

$$\pi_{ij} = \frac{\exp(\boldsymbol{x}_i\boldsymbol{\beta}_j)}{1 + \sum_{h=1}^{c-1}\exp(\boldsymbol{x}_i\boldsymbol{\beta}_h)} \tag{6.2}$$

with $\boldsymbol{\beta}_c = \mathbf{0}$. (The parameters also equal zero for a baseline category for identifiability reasons; see Exercise 6.2.) The numerators in Equation (6.2) for various j sum to the denominator, so $\sum_{j=1}^{c}\pi_{ij} = 1$ for each i. For $c = 2$, this formula simplifies to the binary logistic regression probability formula (5.2).

Interpretation of effects overall rather than conditional on response in category j or c is not simple, because Equation (6.2) shows that all $\{\boldsymbol{\beta}_h\}$ contribute to π_{ij}. The relation $\partial\pi_i/\partial x_{ik} = \beta_k\pi_i(1-\pi_i)$ for binary logistic regression generalizes to

$$\frac{\partial\pi_{ij}}{\partial x_{ik}} = \pi_{ij}\left(\beta_{jk} - \sum_{j'}\pi_{ij'}\beta_{j'k}\right). \tag{6.3}$$

In particular, this rate of change need not have the same sign as β_{jk}, and the curve for π_{ij} as a function of x_{ik} may change direction as the value of x_{ik} increases (see Exercise 6.4).

6.1.2 Baseline-Category Logit Model is a Multivariate GLM

The GLM $g(\mu_i) = \boldsymbol{x}_i\boldsymbol{\beta}$ for a univariate response variable extends to a *multivariate generalized linear model*. The model applies to random components that have distribution in a multivariate generalization of the exponential dispersion family,

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

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

$$\boldsymbol{g}(\boldsymbol{\mu}_i) = \boldsymbol{X}_i\boldsymbol{\beta}, \quad i = 1,\ldots,N, \tag{6.4}$$

where row j of the model matrix $\boldsymbol{X}_i$ for observation i contains values of explanatory variables for response component y_{ij}.

The multinomial distribution is a member of the multivariate exponential dispersion family. The baseline-category logit model is a multivariate GLM. For this representation, we let $\mathbf{y}_i = (y_{i1},\ldots,y_{i,c-1})^{\mathrm{T}}$, since $y_{ic} = 1 - (y_{i1} + \cdots + y_{i,c-1})$ is redundant, $\boldsymbol{\mu}_i = (\pi_{i1},\ldots,\pi_{i,c-1})^{\mathrm{T}}$, and

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

With $(c-1)\times(c-1)p$ model matrix $\boldsymbol{X}_i$ for observation i,

$$\boldsymbol{X}_i\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{x}_i & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{x}_i & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{x}_i \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_{c-1} \end{pmatrix},$$

where $\mathbf{0}$ is a $1\times p$ vector of 0 elements.

6.1.3 Fitting Baseline-Category Logit Models

Maximum likelihood (ML) fitting of baseline-category logit models maximizes the multinomial likelihood subject to $\{\pi_{ij}\}$ simultaneously satisfying the $c-1$ equations that specify the model. The contribution to the log-likelihood from subject i is

$$\begin{aligned} \log\left(\prod_{j=1}^{c}\pi_{ij}^{y_{ij}}\right) &= \sum_{j=1}^{c-1} y_{ij}\log\pi_{ij} + \left(1-\sum_{j=1}^{c-1}y_{ij}\right)\log\pi_{ic} \\ &= \sum_{j=1}^{c-1} y_{ij}\log\frac{\pi_{ij}}{\pi_{ic}} + \log\pi_{ic}. \end{aligned}$$

Thus, the baseline-category logits are the natural parameters for the multinomial distribution. They are the canonical link functions for multinomial GLMs.

Next we construct the likelihood equations for N independent observations. In the last expression above, we substitute $\boldsymbol{x}_i\boldsymbol{\beta}_j$ for $\log(\pi_{ij}/\pi_{ic})$ and

$$\pi_{ic} = 1 \Big/ \left[1+\sum_{j=1}^{c-1}\exp(\boldsymbol{x}_i\boldsymbol{\beta}_j)\right].$$

Then the log-likelihood function is

$$\begin{aligned} L(\boldsymbol{\beta};\boldsymbol{y}) &= \log\left[\prod_{i=1}^{N}\left(\prod_{j=1}^{c}\pi_{ij}^{y_{ij}}\right)\right] \\ &= \sum_{i=1}^{N}\left\{\sum_{j=1}^{c-1} y_{ij}(\boldsymbol{x}_i\boldsymbol{\beta}_j) - \log\left[1+\sum_{j=1}^{c-1}\exp(\boldsymbol{x}_i\boldsymbol{\beta}_j)\right]\right\} \\ &= \sum_{j=1}^{c-1}\left[\sum_{k=1}^{p}\beta_{jk}\left(\sum_{i=1}^{N}x_{ik}y_{ij}\right)\right] - \sum_{i=1}^{N}\log\left[1+\sum_{j=1}^{c-1}\exp(\boldsymbol{x}_i\boldsymbol{\beta}_j)\right]. \end{aligned}$$

The sufficient statistic for β_{jk} is $\sum_i x_{ik}y_{ij}$. When all $x_{i1} = 1$ for an intercept term, the sufficient statistic for β_{j1} is $\sum_i x_{i1}y_{ij} = \sum_i y_{ij}$, which is the total number of observations in category j. Since

$$\frac{\partial L(\boldsymbol{\beta}, \boldsymbol{y})}{\partial \beta_{jk}} = \sum_{i=1}^{N} x_{ik}y_{ij} - \sum_{i=1}^{N}\left[\frac{x_{ik}\exp(\boldsymbol{x}_i\boldsymbol{\beta}_j)}{1 + \sum_{\ell=1}^{c-1}\exp(\boldsymbol{x}_i\boldsymbol{\beta}_\ell)}\right] = \sum_{i=1}^{N} x_{ik}(y_{ij} - \pi_{ij}),$$

the likelihood equations are

$$\sum_{i=1}^{N} x_{ik}y_{ij} = \sum_{i=1}^{N} x_{ik}\pi_{ij},$$

with π_{ij} as expressed in Equation (6.2). As with canonical link functions for univariate GLMs, the likelihood equations equate the sufficient statistics to their expected values.

Differentiating again, you can check that

$$\frac{\partial^2 L(\boldsymbol{\beta}, \boldsymbol{y})}{\partial \beta_{jk}\partial \beta_{jk'}} = -\sum_{i=1}^{N} x_{ik}x_{ik'}\pi_{ij}(1 - \pi_{ij}),$$

and for $j \neq j'$,

$$\frac{\partial^2 L(\boldsymbol{\beta}, \boldsymbol{y})}{\partial \beta_{jk}\partial \beta_{j'k'}} = \sum_{i=1}^{N} x_{ik}x_{ik'}\pi_{ij}\pi_{ij'}.$$

The information matrix consists of $(c-1)^2$ blocks of size $p \times p$,

$$-\frac{\partial^2 L(\boldsymbol{\beta}, \boldsymbol{y})}{\partial \boldsymbol{\beta}_j \partial \boldsymbol{\beta}_{j'}^{\mathrm{T}}} = \sum_{i=1}^{N} \pi_{ij}[I(j = j') - \pi_{ij'}]\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_i,$$

where $I(\cdot)$ is the indicator function. The Hessian is negative-definite, so the log-likelihood function is concave and has a unique maximum. The observed and expected information are identical, so the Newton–Raphson method is equivalent to Fisher scoring for finding the ML parameter estimates, a consequence of the link function being the canonical one. Convergence is usually fast unless at least one estimate is infinite or does not exist (see Note 6.2).

6.1.4 Deviance and Inference for Multinomial Models

For baseline-category logit models, the ML estimator $\hat{\boldsymbol{\beta}}$ has a large-sample normal distribution. As usual, standard errors are square roots of diagonal elements of the inverse information matrix. The $\{\hat{\boldsymbol{\beta}}_j\}$ are correlated. The estimate $(\hat{\boldsymbol{\beta}}_a - \hat{\boldsymbol{\beta}}_b)$ of the

effects in the linear predictor for $\log(\pi_{ia}/\pi_{ib})$ does not depend on which category is the baseline.

Statistical inference can use likelihood-ratio, Wald, and score inference methods for GLMs. For example, the likelihood-ratio test for the effect of explanatory variable k tests H_0: $\beta_{1k} = \beta_{2k} = \cdots = \beta_{c-1,k} = 0$ by treating double the change in the maximized log-likelihood from adding that variable to the model as having a null chi-squared distribution with $df = c - 1$. The likelihood-ratio test statistic equals the difference in the deviance values for comparing the models.

The derivation of the deviance shown in Section 5.5.1 for binomial GLMs generalizes directly to multinomial GLMs. For grouped data with n_i trials for the observations at setting i of the explanatory variables, let y_{ij} now denote the *proportion* of observations in category j. The deviance is the likelihood-ratio statistic comparing double the log of the multinomial likelihood $\prod_i \left(\prod_j \pi_{ij}^{n_i y_{ij}} \right)$ evaluated for the model fitted probabilities $\{\hat{\pi}_{ij}\}$ and the unrestricted (saturated model) alternative $\{\tilde{\pi}_{ij} = y_{ij}\}$. The deviance and the Pearson statistic equal

$$G^2 = 2 \sum_{i=1}^{N} \sum_{j=1}^{c} n_i y_{ij} \log \frac{n_i y_{ij}}{n_i \hat{\pi}_{ij}}, \quad X^2 = \sum_{i=1}^{N} \sum_{j=1}^{c} \frac{(n_i y_{ij} - n_i \hat{\pi}_{ij})^2}{n_i \hat{\pi}_{ij}}. \tag{6.5}$$

These have the form seen in Section 4.4.4 for Poisson GLMs and in Section 5.5.1 for binomial GLMs of

$$G^2 = 2 \sum \text{observed} \log \left(\frac{\text{observed}}{\text{fitted}} \right), \quad X^2 = \sum \frac{(\text{observed} - \text{fitted})^2}{\text{fitted}},$$

with sums taken over all observed counts $\{n_i y_{ij}\}$ and fitted counts $\{n_i \hat{\pi}_{ij}\}$.

As in the binary case, with categorical explanatory variables and the grouped form of the data, G^2 and X^2 are goodness-of-fit statistics that provide a global model check. They have approximate chi-squared null distributions when the expected cell counts mostly exceed about 5. The df equal the number of multinomial probabilities modeled, which is $N(c - 1)$, minus the number of model parameters. The residuals of Section 5.5.3 are useful for follow-up information about poorly fitting models. For ungrouped data (i.e., all $\{n_i = 1\}$), such as when at least one explanatory variable is continuous, formula (6.5) for G^2 remains valid and is used to compare nested unsaturated models.

6.1.5 Discrete-Choice Models

Some 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 (house, condominium, rental, other), 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*.

In most discrete-choice applications, some explanatory variables take different values for different response choices. As predictors of choice of transportation system, the cost and 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 called *characteristics of the choices*. They differ from the usual ones, for which values remain constant across the choice set. Such *characteristics of the chooser* include demographic and socioeconomic variables such as gender, race, annual income, and educational attainment.

We introduce the discrete-choice model for the case that the p explanatory variables are all characteristics of the choices. For subject i and response choice j, let $\boldsymbol{x}_{ij} = (x_{ij1}, \ldots, x_{ijp})$ denote the values of those variables. The discrete choice model for the probability of selecting option j is

$$\pi_{ij} = \frac{\exp(\boldsymbol{x}_{ij}\boldsymbol{\beta})}{\sum_{h=1}^{c} \exp(\boldsymbol{x}_{ih}\boldsymbol{\beta})}. \tag{6.6}$$

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

$$\log(\pi_{ia}/\pi_{ib}) = (\boldsymbol{x}_{ia} - \boldsymbol{x}_{ib})\boldsymbol{\beta}. \tag{6.7}$$

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. The effects $\boldsymbol{\beta}$ are identical for each pair of choices.

From Equation (6.7), 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. This property is referred to as *independence from irrelevant alternatives*. For this to be at all realistic, the model should be used only when the alternatives are distinct and regarded separately by the person making the choice.

A more general version of the model permits the choice set to vary among subjects. For instance, in a study of the choice of transportation system to take to work, some people may not have the subway as an option. In the denominator of Equation (6.6), the sum is then taken over the choice set for subject i.

6.1.6 Baseline-Category Logit Model as a Discrete-Choice Model

Discrete-choice models can also include characteristics of the chooser. A baseline-category logit model (6.2) with such explanatory variables can be expressed in the discrete-choice form (6.6) when we replace each explanatory variable by c artificial variables. The jth is the product of the explanatory variable with an indicator variable that equals 1 when the response choice is j. For instance, for a single explanatory

variable with value x_i for subject i and linear predictor $\beta_{0j} + \beta_{1j}x_i$ for the jth logit, we form the $1 \times 2c$ vectors

$$z_{i1} = (1, 0, \ldots, 0, x_i, 0, \ldots, 0), \ldots, z_{ic} = (0, 0, \ldots, 1, 0, 0, \ldots, x_i).$$

Let $\boldsymbol{\beta} = (\beta_{01}, \ldots, \beta_{0c}, \beta_{11}, \ldots, \beta_{1c})^{\mathrm{T}}$. Then $z_{ij}\boldsymbol{\beta} = \beta_{0j} + \beta_{1j}x_i$, and Equation (6.2) is (with $\beta_{0c} = \beta_{1c} = 0$ for identifiability)

$$\begin{aligned}\pi_{ij} &= \frac{\exp(\beta_{0j} + \beta_{1j}x_i)}{\exp(\beta_{01} + \beta_{11}x_i) + \cdots + \exp(\beta_{0c} + \beta_{1c}x_i)} \\ &= \frac{\exp(z_{ij}\boldsymbol{\beta})}{\exp(z_{i1}\boldsymbol{\beta}) + \cdots + \exp(z_{ic}\boldsymbol{\beta})}.\end{aligned}$$

This has the discrete-choice model form (6.6).

This model extends directly to having multiple explanatory variables of each type. With this approach, the discrete-choice model is very general. The ordinary baseline-category logit model is a special case.

6.2 ORDINAL RESPONSES: CUMULATIVE LOGIT AND PROBIT MODELS

For ordinal response variables, models have terms that reflect ordinal characteristics such as a monotone trend, whereby responses tend to fall in higher (or lower) categories as the value of an explanatory variable increases. Such models are more parsimonious than models for nominal responses, because potentially they have many fewer parameters. In this section we introduce logistic and probit models for ordinal responses.

6.2.1 Cumulative Logit Models: Proportional Odds

Let y_i denote the response outcome category for subject i. That is, $y_i = j$ means that $y_{ij} = 1$ and $y_{ik} = 0$ for $k \neq j$, for the c multinomial indicators. To use the category ordering, we express models in terms of the cumulative probabilities,

$$P(y_i \leq j) = \pi_{i1} + \cdots + \pi_{ij}, \quad j = 1, \ldots, c.$$

The *cumulative logits* are logits of these cumulative probabilities,

$$\begin{aligned}\text{logit}[P(y_i \leq j)] &= \log \frac{P(y_i \leq j)}{1 - P(y_i \leq j)} \\ &= \log \frac{\pi_{i1} + \cdots + \pi_{ij}}{\pi_{i,j+1} + \cdots + \pi_{ic}}, \quad j = 1, \ldots, c-1.\end{aligned}$$

Each cumulative logit uses all c response categories.

A model for logit$[P(y_i \le j)]$ alone is an ordinary logistic model for a binary response in which categories 1 to j represent "success" and categories $j+1$ to c represent "failure." Here is a parsimonious model that simultaneously uses all $(c-1)$ cumulative logits:

Cumulative logit model:

$$\text{logit}[P(y_i \le j)] = \alpha_j + \boldsymbol{x}_i\boldsymbol{\beta}, \quad j = 1, \ldots, c-1. \tag{6.8}$$

Each cumulative logit has its own intercept. The $\{\alpha_j\}$ are increasing in j, because $P(y_i \le j)$ increases in j at any fixed $\boldsymbol{x}_i$, and the logit is an increasing function of $P(y_i \le j)$. We use separate notation α_j and show the intercept terms by themselves in the linear predictor, because they depend on j but the other effects do not. This model states that the effects $\boldsymbol{\beta}$ of the explanatory variables are the same for each cumulative logit. For a single continuous explanatory variable x, Figure 6.1 depicts the model when $c = 4$. The curves for $j = 1$, 2, and 3 have exactly the same shape and do not cross.

This model treats the response variable as ordinal scale, in the following sense: if the model holds and the order of the outcome categories is reversed, the model still holds with a change in the sign of $\boldsymbol{\beta}$; however, the model need not hold if the outcome categories are permuted in any other way.

The cumulative logit model (6.8) satisfies

$$\begin{aligned}&\text{logit}[P(y_i \le j \mid \boldsymbol{x}_i = \boldsymbol{u})] - \text{logit}[P(y_i \le j \mid \boldsymbol{x}_i = \boldsymbol{v})] \\ &= \log \frac{P(y_i \le j \mid \boldsymbol{x}_i = \boldsymbol{u})/P(y_i > j \mid \boldsymbol{x}_i = \boldsymbol{u})}{P(y_i \le j \mid \boldsymbol{x}_i = \boldsymbol{v})/P(y_i > j \mid \boldsymbol{x}_i = \boldsymbol{v})} = (\boldsymbol{u} - \boldsymbol{v})\boldsymbol{\beta}.\end{aligned}$$

The odds that the response $\le j$ at $\boldsymbol{x}_i = \boldsymbol{u}$ are $\exp[(\boldsymbol{u} - \boldsymbol{v})\boldsymbol{\beta}]$ times the odds at $\boldsymbol{x}_i = \boldsymbol{v}$. An odds ratio of cumulative probabilities is called a *cumulative odds ratio*. The log cumulative odds ratio is proportional to the distance between $\boldsymbol{u}$ and $\boldsymbol{v}$. For each j, the odds that $y_i \le j$ multiply by $\exp(\beta_k)$ per 1-unit increase in x_{ik}, adjusting for the other explanatory variables. The same proportionality constant applies to all $c-1$

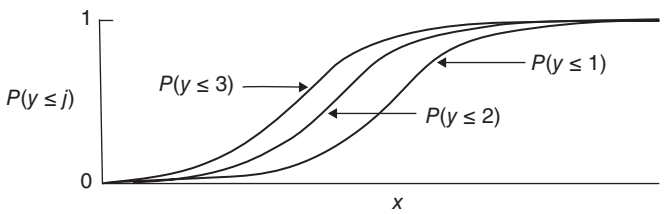

Figure 6.1 Cumulative logit model with the same effect of x on each of three cumulative probabilities, for an ordinal response variable with $c = 4$ categories.

cumulative logits; that is, the effect is β_k, not β_{jk}. This property of a common effect for all the cumulative probabilities is referred to as *proportional odds*.

6.2.2 Latent Variable Motivation for Proportional Odds Structure

A linear 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 cumulative logit model. Let y_i^* denote this underlying latent variable for subject i. Suppose it has cdf $G(y_i^* - \mu_i)$, where values of y^* vary around a mean that depends on $\boldsymbol{x}$ through $\mu_i = \boldsymbol{x}_i\boldsymbol{\beta}$. Suppose that the continuous scale has *cutpoints* $-\infty = \alpha_0 < \alpha_1 < \cdots < \alpha_c = \infty$ such that we observe

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

That is, y_i falls in category j when the latent variable falls in the jth interval of values, as Figure 6.2 depicts. Then

$$P(y_i \leq j) = P\left(y_i^* \leq \alpha_j\right) = G(\alpha_j - \mu_i) = G(\alpha_j - \boldsymbol{x}_i\boldsymbol{\beta}).$$

The model for y implies that the link function G^{-1}, the inverse of the cdf for y^*, applies to $P(y_i \leq j)$. If $y_i^* = \boldsymbol{x}_i\boldsymbol{\beta} + \epsilon_i$, where the cdf G of ϵ_i is the standard logistic (Section 5.1.3), then G^{-1} is the logit link, and the cumulative logit model with proportional odds structure results. Normality for ϵ_i implies a probit link (Section 6.2.3) for the cumulative probabilities.

Using a cdf of the form $G(y_i^* - \mu_i)$ for the latent variable results in the linear predictor $\alpha_j - \boldsymbol{x}_i\boldsymbol{\beta}$ rather than $\alpha_j + \boldsymbol{x}_i\boldsymbol{\beta}$. With this alternate parameterization, when $\beta_k > 0$, as x_{ik} increases, each cumulative logit decreases, so each cumulative probability decreases and relatively less probability mass falls at the low end of the y scale. Thus, y_i tends to be larger at higher values of x_{ik}. Then the sign of β_k has the usual

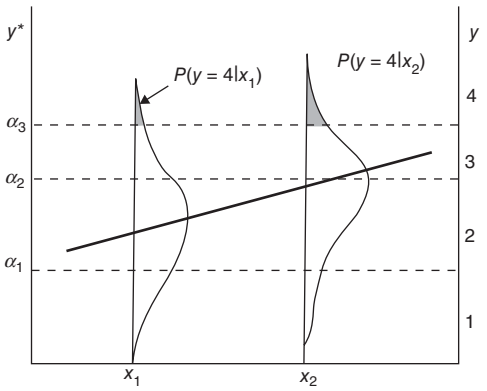

Figure 6.2 Ordinal measurement and underlying linear model for a latent variable.

meaning of a positive or negative effect. When you use software to fit the model, you should check whether it parameterizes the linear predictor as $\alpha_j + \boldsymbol{x}_i\boldsymbol{\beta}$ or as $\alpha_j - \boldsymbol{x}_i\boldsymbol{\beta}$, as signs of estimated effects differ accordingly.

In the latent variable derivation, the same parameters $\boldsymbol{\beta}$ occur for the effects regardless of how the cutpoints $\{\alpha_j\}$ chop up the scale for y^* and regardless of the number of categories. The effect parameters are invariant to the choice of categories for y. This feature makes it possible to compare $\hat{\boldsymbol{\beta}}$ from studies using different response scales.

6.2.3 Cumulative Probit and Other Cumulative Link Models

As in binary GLMs, other link functions are possible for the cumulative probabilities. Let G^{-1} denote a link function that is the inverse of the continuous cdf G. The *cumulative link model*

$$G^{-1}[P(y_i \le j)] = \alpha_j + \boldsymbol{x}_i\boldsymbol{\beta} \tag{6.9}$$

links the cumulative probabilities to the linear predictor. As in the cumulative logit model with proportional odds form (6.8), effects are the same for each cumulative probability. This assumption holds when a latent variable y^* satisfies a linear model with standard cdf G for the error term. Thus, we can regard cumulative link models as linear models that use a linear predictor $\boldsymbol{x}_i\boldsymbol{\beta}$ to describe effects of explanatory variables on a crude ordinal measurement y_i of y_i^*.

The *cumulative probit model* is the cumulative link model that uses the standard normal cdf Φ for G. This generalizes the binary probit model (Section 5.6) to ordinal responses. 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. When we expect an underlying latent variable to be highly skewed, such as an extreme-value distribution, we can generalize the binary model with log–log or complementary log–log link (Section 5.6.3) to ordinal responses.

Effects in cumulative link models can be interpreted in terms of the underlying latent variable model. The cumulative probit model is appropriate when the latent variable model holds with a normal conditional distribution for y^*. Consider the parameterization $\Phi^{-1}[P(y_i \le j)] = \alpha_j - \boldsymbol{x}_i\boldsymbol{\beta}$. From Section 6.2.2, $y_i^* = \boldsymbol{x}_i\boldsymbol{\beta} + \epsilon_i$ where $\epsilon_i \sim N(0, 1)$ has cdf Φ. A 1-unit increase in x_{ik} corresponds to a β_k increase (and thus an increase of β_k standard deviations) in $E(y_i^*)$, adjusting for the other explanatory variables.

To describe predictive power, an analog of the multiple correlation for linear models is the correlation between the latent variable y^* and the fitted linear predictor values. This directly generalizes the measure for binary data presented in Section 5.2.5. Its square is an R^2 analog (McKelvey and Zavoina 1975). This R^2 measure equals the estimated variance of $\hat{y}^*$ divided by the estimated variance of y^*. Here, $\hat{y}_i^* = \sum_j \hat{\beta}_j x_{ij}$ is the same as the estimated linear predictor without the intercept term,

and the estimated variance of y^* equals this plus the variance of ϵ in the latent variable model (1 for the probit link and $\pi^2/3 = 3.29$ for the logit link). It is also helpful to summarize the effect of an explanatory variable in terms of the change in the probability of the highest (or the lowest) category of the ordinal scale over the range or interquartile range of that variable, at mean values of other explanatory variables.

6.2.4 Fitting and Checking Cumulative Link Models

For multicategory indicator $(y_{i1}, \ldots, y_{ic})$ of the response for subject i, the multinomial likelihood function for the cumulative link model $G^{-1}[P(y_i \le j)] = \alpha_j + \boldsymbol{x}_i\boldsymbol{\beta}$ is

$$\prod_{i=1}^{N}\left(\prod_{j=1}^{c}\pi_{ij}^{y_{ij}}\right) = \prod_{i=1}^{N}\left\{\prod_{j=1}^{c}[P(y_i \le j) - P(y_i \le j-1)]^{y_{ij}}\right\}$$

viewed as a function of $(\{\alpha_j\}, \boldsymbol{\beta})$, where $P(y_i \le 0) = 0$. The log-likelihood function is

$$L(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{i=1}^{N}\sum_{j=1}^{c} y_{ij}\log[G(\alpha_j + \boldsymbol{x}_i\boldsymbol{\beta}) - G(\alpha_{j-1} + \boldsymbol{x}_i\boldsymbol{\beta})].$$

Let g denote the derivative of G, that is, the pdf corresponding to the cdf G, and let δ_{jk} denote the Kronecker delta, $\delta_{jk} = 1$ if $j = k$ and $\delta_{jk} = 0$ otherwise. Then the likelihood equations are

$$\frac{\partial L}{\partial \beta_k} = \sum_{i=1}^{N}\sum_{j=1}^{c} y_{ij}x_{ik}\frac{g(\alpha_j + \boldsymbol{x}_i\boldsymbol{\beta}) - g(\alpha_{j-1} + \boldsymbol{x}_i\boldsymbol{\beta})}{G(\alpha_j + \boldsymbol{x}_i\boldsymbol{\beta}) - G(\alpha_{j-1} + \boldsymbol{x}_i\boldsymbol{\beta})} = 0,$$

and

$$\frac{\partial L}{\partial \alpha_k} = \sum_{i=1}^{N}\sum_{j=1}^{c} y_{ij}\frac{\delta_{jk}g(\alpha_j + \boldsymbol{x}_i\boldsymbol{\beta}) - \delta_{j-1,k}g(\alpha_{j-1} + \boldsymbol{x}_i\boldsymbol{\beta})}{G(\alpha_j + \boldsymbol{x}_i\boldsymbol{\beta}) - G(\alpha_{j-1} + \boldsymbol{x}_i\boldsymbol{\beta})} = 0.$$

The Hessian matrix is rather messy[1] and not shown here. The likelihood equations can be solved using Fisher scoring or the Newton–Raphson method. The *SE* values differ somewhat for the two methods, because the expected and observed information matrices are not the same for this noncanonical link model.

Since the latent variable model on which the cumulative link model is based describes location effects while assuming constant variability, settings of the explanatory variables are *stochastically ordered* on the response: for the observed

[1]This is because $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are not orthogonal. See Agresti (2010, Section 5.1.2).

response with any pair $\boldsymbol{u}$ and $\boldsymbol{v}$ of potential explanatory variable values, either $P(y_i \leq j \mid \boldsymbol{x}_i = \boldsymbol{u}) \leq P(y_i \leq j \mid \boldsymbol{x}_i = \boldsymbol{v})$ for all j or $P(y_i \leq j \mid \boldsymbol{x}_i = \boldsymbol{u}) \geq P(y_i \leq j \mid \boldsymbol{x}_i = \boldsymbol{v})$ for all j. When this is violated and such models fit poorly, often it is because the response variability also varies with $\boldsymbol{x}$. For example, with $c = 4$ and response probabilities (0.3, 0.2, 0.2, 0.3) at $\boldsymbol{u}$ and (0.1, 0.4, 0.4, 0.1) at $\boldsymbol{v}$, $P(y_i \leq 1 \mid \boldsymbol{x}_i = \boldsymbol{u}) > P(y_i \leq 1 \mid \boldsymbol{x}_i = \boldsymbol{v})$ but $P(y_i \leq 3 \mid \boldsymbol{x}_i = \boldsymbol{u}) < P(y_i \leq 3 \mid \boldsymbol{x}_i = \boldsymbol{v})$. At $\boldsymbol{u}$ the responses concentrate more in the extreme categories than at $\boldsymbol{v}$.

An advantage of the simple structure of the same effects $\boldsymbol{\beta}$ for different cumulative probabilities is that effects are simple to summarize and are parsimonious, requiring only a single parameter for each explanatory variable. The models generalize to include separate effects, replacing $\boldsymbol{\beta}$ in Equation (6.9) by $\boldsymbol{\beta}_j$. This implies nonparallelism of curves for different cumulative probabilities. Curves may then cross for some $\boldsymbol{x}$ values, violating the proper order among the cumulative probabilities (Exercise 6.13).

When we can fit the more general model with effects $\{\boldsymbol{\beta}_j\}$, a likelihood-ratio test checks whether that model fits significantly better. Often though, convergence fails in fitting the model, because the cumulative probabilities are out of order. We can then use a score test of whether the $\{\boldsymbol{\beta}_j\}$ takes a common value, because the score test uses the likelihood function only at the null (i.e., where common $\boldsymbol{\beta}$ holds). Some software for the model provides this score test. It is often labelled as a "test of the proportional odds assumption," because that is the name for the simple structure when we use the logit link. When there is strong evidence against the common $\boldsymbol{\beta}$, the simpler model is still often useful for describing overall effects, with the fit of the more general model pointing out ways to fine-tune the description of effects. With categorical predictors and grouped data with most expected cell counts exceeding about 5, the deviance and Pearson statistics in Equation (6.5) provide global chi-squared goodness-of-fit tests.

The log-likelihood function is concave for many cumulative link models, including the logit and probit. Iterative algorithms such as Fisher scoring usually converge rapidly to the ML estimates.

6.2.5 Why not Use OLS Regression to Model Ordinal Responses?

Many methodologists analyze ordinal response data by ignoring the categorical nature of y and assigning numerical scores to the ordered categories and using ordinary least squares (OLS) methods such as linear regression and ANOVA. That approach can identify variables that clearly affect y and provide simple description, but it has limitations. First, usually there is no clear-cut choice for the scores. How would you assign scores to categories such as (never, rarely, sometimes, always)? Second, a particular ordinal outcome is consistent with a range of values for some underlying latent variable. Ordinary linear modeling does not allow for the measurement error that results from replacing such a range by a single numerical value. Third, the approach does not yield estimated probabilities for the ordinal categories at fixed settings of the explanatory variables. Fourth, the approach ignores that the variability of y is naturally nonconstant for categorical data: an ordinal response has little

variability at explanatory variable values for which observations fall mainly in the highest category (or mainly in the lowest category), but considerable variability at values for which observations are spread among the categories.

Related to the second and fourth limitations, the ordinary linear modeling approach does not account for "ceiling effects" and "floor effects" that occur because of the upper and lower limits for y. Such effects can cause ordinary linear modeling to give misleading[2] results. To illustrate, we apply a normal linear model to simulated data with an ordinal response variable y based on an underlying normal latent variable y^*. We generated 100 observations as follows: x_i values were independent uniform variates between 0 and 100, z_i values were independent with $P(z_i = 0) = P(z_i = 1) = 0.50$, and the latent y_i^* was a normal variate with mean

$$E\left(y_i^*\right) = 20.0 + 0.6x_i - 40.0z_i$$

and standard deviation 10. The first scatterplot in Figure 6.3 shows the 100 observations on y_i^* and x_i, each data point labelled by the category for z_i. The plot also shows the OLS fit for this model.

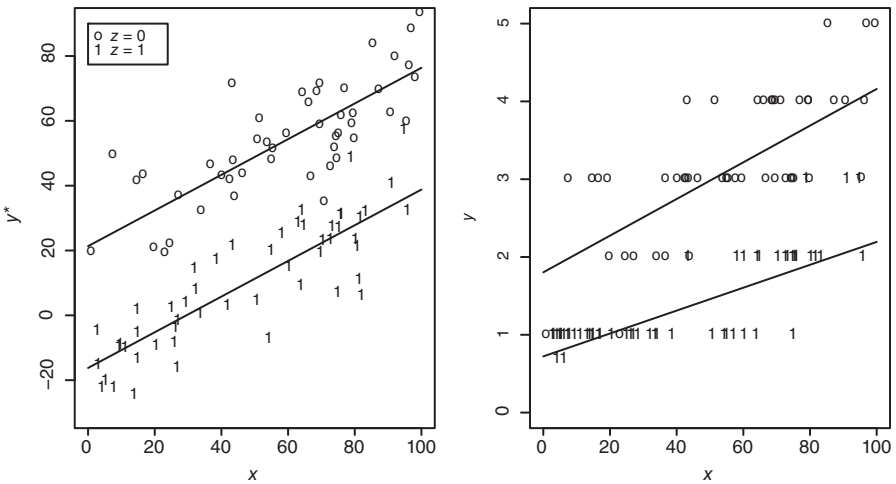

Figure 6.3 Ordinal data (in second panel) for which ordinary linear modeling suggests interaction, because of a floor effect, but ordinal modeling does not. The data were generated (in first panel) from a normal linear model with continuous (x) and binary (z) explanatory variables. When the continuous y^* is categorized and y is measured as (1, 2, 3, 4, 5), the observations labeled "1" for the category of z have a linear x effect with only half the slope of the observations labeled "0."

[2]These effects also result in substantial correlation between the values of residuals and the values of quantitative explanatory variables.

We then categorized the 100 generated values on y^* into five categories to create observations for an ordinal variable y, as follows:

$$y_i = 1 \text{ if } y_i^* \leq 20, \quad y_i = 2 \text{ if } 20 < y_i^* \leq 40, \quad y_i = 3 \text{ if } 40 < y_i^* \leq 60,$$
$$y_i = 4 \text{ if } 60 < y_i^* \leq 80, \quad y_i = 5 \text{ if } y_i^* > 80.$$

The second scatterplot in Figure 6.3 shows the data as they would actually be observed. Using OLS with scores (1, 2, 3, 4, 5) for y suggests a better fit when the model has an interaction term, allowing different slopes relating $E(y_i)$ to x_i when $z_i = 0$ and when $z_i = 1$. The second scatterplot shows the OLS fit of the linear model $E(y_i) = \beta_0 + \beta_1 x_i + \beta_2 z_i + \beta_3(x_i \times z_i)$. The slope of the line is about twice as high when $z_i = 0$ as when $z_i = 1$. This interaction effect is caused by the observations when $z_i = 1$ tending to fall in category $y_i = 1$ whenever x_i takes a relatively low value. As x_i gets lower, the underlying value y_i^* can continue to tend to get lower, but y_i cannot fall below 1. So at low x_i values there is a floor effect, particularly for observations with $z_i = 1$.

Standard ordinal models such as the cumulative logit model with proportional odds structure fit these data well without the need for an interaction term. In fact, by the latent variable model of Section 6.2.2, the true structure is that of a cumulative probit model with common $\boldsymbol{\beta}$. Such models allow for underlying values of y^* when $z = 1$ to be below those when $z = 0$ even if x is so low that y is very likely to be in the first category at both levels of z.

6.3 EXAMPLES: NOMINAL AND ORDINAL RESPONSES

In this chapter we have presented the most popular models for multinomial response variables. End-of-chapter exercises briefly introduce other models for nominal and ordinal responses that are beyond our scope.

6.3.1 Issues in Selecting Multinomial Models

With a nominal-scale response variable, the baseline category logit form of model is usually the default choice. It is not sensible to force similar effects for different logits, so the model will necessarily contain a large number of parameters unless c and p are small.

With ordinal response variables, the choice is not so clear, because we can treat y as nominal when cumulative link models that assume a common value $\boldsymbol{\beta}$ for $\{\boldsymbol{\beta}_j\}$ fit poorly. Especially with large samples, it is not surprising to obtain a small P-value in comparing a model with common $\boldsymbol{\beta}$ to one with separate $\{\boldsymbol{\beta}_j\}$. However, it is poor practice to be guided merely by statistical significance in selecting a model. Even if a more complex model fits significantly better, for reasons of parsimony the simpler model might be preferable when $\{\hat{\boldsymbol{\beta}}_j\}$ are not substantially different in practical terms. It is helpful to check whether the violation of the common $\boldsymbol{\beta}$ property is substantively important, by comparing $\hat{\boldsymbol{\beta}}$ to $\{\hat{\boldsymbol{\beta}}_j\}$ obtained from fitting the more general model (when possible) or from separate fits to the binary collapsings of the response.

Although effect estimators using the simpler model are biased, because of the bias–variance tradeoff (Section 4.6.2) they may have smaller overall mean squared error for estimating the category probabilities. This is especially true when c is large, as the difference is then quite large between the numbers of parameters in the two models.

If a model with common $\boldsymbol{\beta}$ fits poorly in practical terms, alternative strategies exist. For example, you can report $\{\hat{\boldsymbol{\beta}}_j\}$ from the more general model, to investigate the nature of the lack of fit and to describe effects separately for each cumulative probability. Or, you could use an alternative ordinal logit model for which the more complex nonproportional odds form is also valid[3]. Or you can fit baseline-category logit models and use the ordinality in an informal way in interpreting the associations.

6.3.2 Example: Baseline-Category Logit Model for Alligator Food Choice

Table 6.1 comes from a study of factors influencing the primary food choice of alligators. The study captured 219 alligators in four Florida lakes. The nominal-scale response variable is the primary food type, in volume, found in an alligator's stomach. Table 6.1 classifies the primary food choice according to the lake of capture and the size of the alligator. Here, size is binary, distinguishing between nonadult (length $\leq$ 2.3 meters) and adult alligators.

We use baseline-category logit models to investigate the effects of size and lake on the primary food choice. Fish was the most frequent primary food choice, and we use it as the baseline category. We estimate the effects on the odds that alligators select other primary food types instead of fish. Let $s = 1$ for alligator size ≤ 2.3 meters and 0 otherwise, and let z^H, z^O, z^{T}, and z^G be indicator variables for the lakes ($z = 1$ for alligators in that lake and 0 otherwise). The model with main effects is

$$\log(\pi_{ij}/\pi_{i1}) = \beta_{j0} + \beta_{j1}s_i + \beta_{j2}z_i^O + \beta_{j3}z_i^{\mathrm{T}} + \beta_{j4}z_i^G, \quad \text{for } j = 2, 3, 4, 5.$$

In R the `vglm` function in the `VGAM` package can fit this model[4].

Table 6.1 Primary Food Choice of Alligators, by Lake and Size of the Alligator

Lake	Size (meters)	Primary Food Choice				
		Fish	Invertebrate	Reptile	Bird	Other
Hancock	≤ 2.3	23	4	2	2	8
	> 2.3	7	0	1	3	5
Ocklawaha	≤ 2.3	5	11	1	0	3
	> 2.3	13	8	6	1	0
Trafford	≤ 2.3	5	11	2	1	5
	> 2.3	8	7	6	3	5
George	≤ 2.3	16	19	1	2	3
	> 2.3	17	1	0	1	3

Source: Data courtesy of Mike Delany and Clint Moore. For details, see Delany et al. (1999).

[3]See Exercise 6.8 and Agresti (2010, Chapter 4).

[4]The `multinom` function in the `nnet` R package also fits it. The VGAM package fits many models for discrete data. See `www.stat.auckland.ac.nz/~yee/VGAM/doc/VGAM.pdf`.

```
------------------------------------------------------------------------------
> Alligators # file Alligators.dat at www.stat.ufl.edu/~aa/glm/data
  lake size y1 y2 y3 y4 y5
1    1    1 23  4  2  2  8
2    1    0  7  0  1  3  5
...
8    4    0 17  1  0  1  3
> attach(Alligators)
> library(VGAM)
> fit <- vglm(formula = cbind(y2,y3,y4,y5,y1) ~ size + factor(lake),
+   family=multinomial, data=Alligators) # fish=1 is baseline category
> summary(fit)
                  Estimate  Std. Error   z value
(Intercept):1     -3.2074      0.6387   -5.0215
(Intercept):2     -2.0718      0.7067   -2.9315
(Intercept):3     -1.3980      0.6085   -2.2973
(Intercept):4     -1.0781      0.4709   -2.2893
size:1             1.4582      0.3959    3.6828
size:2            -0.3513      0.5800   -0.6056
size:3            -0.6307      0.6425   -0.9816
size:4             0.3315      0.4482    0.7397
factor(lake)2:1    2.5956      0.6597    3.9344
factor(lake)2:2    1.2161      0.7860    1.5472
factor(lake)2:3   -1.3483      1.1635   -1.1588
factor(lake)2:4   -0.8205      0.7296   -1.1247
factor(lake)3:1    2.7803      0.6712    4.1422
factor(lake)3:2    1.6925      0.7804    2.1686
factor(lake)3:3    0.3926      0.7818    0.5023
factor(lake)3:4    0.6902      0.5597    1.2332
factor(lake)4:1    1.6584      0.6129    2.7059
factor(lake)4:2   -1.2428      1.1854   -1.0484
factor(lake)4:3   -0.6951      0.7813   -0.8897
factor(lake)4:4   -0.8262      0.5575   -1.4819
---
Residual deviance: 17.0798 on 12 degrees of freedom
Log-likelihood: -47.5138 on 12 degrees of freedom
> 1 - pchisq(17.0798, df=12)
[1] 0.146619  # P-value for deviance goodness-of-fit test
------------------------------------------------------------------------------
```

The data are a bit sparse, but the deviance of 17.08 ($df = 12$) does not give much evidence against the main-effects model. The df value reflects that we have modeled 32 multinomial probabilities (4 at each combination of size and lake) using 20 parameters (5 for each logit). The more complex model allowing interaction between size and lake has 12 more parameters and is the saturated model. Removing size or lake from the main-effects model results in a significantly poorer fit: the deviance increases by 21.09 ($df = 4$) in removing size and 49.13 ($df = 12$) in removing lake.

The equations shown for fish as the baseline determine those for other primary food choice pairs. Viewing all these, we see that size has its greatest impact on

whether invertebrates rather than fish are the primary food choice. The prediction equation for the log odds of selecting invertebrates instead of fish is

$$\log(\hat{\pi}_{i2}/\hat{\pi}_{i1}) = -3.207 + 1.458s_i + 2.596z_i^O + 2.780z_i^T + 1.658z_i^G.$$

For a given lake, for small alligators the estimated odds that primary food choice was invertebrates instead of fish are $\exp(1.458) = 4.30$ times the estimated odds for large alligators. The estimated effect is imprecise, as the Wald 95% confidence interval is $\exp[1.458 \pm 1.96(0.396)] = (1.98, 9.34)$. The lake effects indicate that the estimated odds that the primary food choice was invertebrates instead of fish are relatively higher at lakes Ocklawaha, Trafford and George than they are at Lake Hancock.

The model parameter estimates yield fitted probabilities. For example, the estimated probability that a large alligator in Lake George has invertebrates as the primary food choice is

$$\hat{\pi}_{i2} = \frac{e^{-3.207+1.658}}{1 + e^{-3.207+1.658} + e^{-2.072-1.243} + e^{-1.398-0.695} + e^{-1.078-0.826}} = 0.14.$$

The estimated probabilities of (invertebrates, reptiles, birds, other, fish) for large alligators in that lake are (0.14, 0.02, 0.08, 0.10, 0.66).

```
------------------------------------------------------------------------
> fitted(fit)
       y2      y3      y4      y5      y1
1  0.0931  0.0475  0.0704  0.2537  0.5353
2  0.0231  0.0718  0.1409  0.1940  0.5702
...
8  0.1397  0.0239  0.0811  0.0979  0.6574
------------------------------------------------------------------------
```

6.3.3 Example: Cumulative Link Models for Mental Impairment

The data in Table 6.2 are based on a study of mental health for a random sample of adult residents of Alachua County, Florida[5]. Mental impairment is ordinal, with categories (1 = well, 2 = mild symptom formation, 3 = moderate symptom formation, 4 = impaired). The study related y = mental impairment to several explanatory variables, two of which are used here. The life events index x_1 is a composite measure of the number and severity of important life events that occurred to the subject within the past 3 years, such as the birth of a child, a new job, a divorce, or a death in the family. In this sample, x_1 has a mean of 4.3 and standard deviation of 2.7. Socioeconomic status (x_2 = SES) is measured here as binary (1 = high, 0 = low).

The cumulative logit model of the proportional odds form with main effects has ML fit

$$\text{logit}[\hat{P}(y_i \le j)] = \hat{\alpha}_j - 0.319x_{i1} + 1.111x_{i2}.$$

[5]Thanks to Charles Holzer for the background for this study; the 40 observations analyzed here are merely reflective of patterns found with his much larger sample.

Table 6.2 Mental Impairment, Life Events Index, and Socioeconomic Status (SES), for 40 Adults in Alachua County, Florida

Subject	Mental Impairment	Life Events	SES	Subject	Mental Impairment	Life Events	SES
1	Well	1	1	21	Mild	9	1
2	Well	9	1	22	Mild	3	0
3	Well	4	1	23	Mild	3	1
4	Well	3	1	24	Mild	1	1
5	Well	2	0	25	Moderate	0	0
6	Well	0	1	26	Moderate	4	1
7	Well	1	0	27	Moderate	3	0
8	Well	3	1	28	Moderate	9	0
9	Well	3	1	29	Moderate	6	1
10	Well	7	1	30	Moderate	4	0
11	Well	1	0	31	Moderate	3	0
12	Well	2	0	32	Impaired	8	1
13	Mild	5	1	33	Impaired	2	1
14	Mild	6	0	34	Impaired	7	1
15	Mild	3	1	35	Impaired	5	0
16	Mild	1	0	36	Impaired	4	0
17	Mild	8	1	37	Impaired	4	0
18	Mild	2	1	38	Impaired	8	1
19	Mild	5	0	39	Impaired	8	0
20	Mild	5	1	40	Impaired	9	0

The estimated cumulative probability, starting at the "well" end of the mental impairment scale, decreases as life events increases and is higher at the higher level of SES, adjusted for the other variable. Given the life events score, at the high SES level the estimated odds of mental impairment below any fixed level are $e^{1.111} = 3.0$ times the estimated odds at the low SES level. The 95% Wald confidence interval[6] for this effect is $\exp[1.111 \pm 1.96(0.614)] = (0.91, 10.12)$. A null SES effect is plausible, but the SES effect could also be very strong. The Wald test shows strong evidence of a life events effect.

```
------------------------------------------------------------------------------
> Mental # file Mental.dat at www.stat.ufl.edu/~aa/glm/data
   impair life ses  # impair has well=1, ... , impaired=4
1       1    1   1
2       1    9   1
...
40      4    9   0
> attach(Mental)
> library(VGAM) # Alternative is polr function in MASS package
> fit <- vglm(impair ~ life + ses, family=cumulative(parallel=TRUE),
```

[6]The `ProfileLikelihood` package in R has a function for profile likelihood intervals for this model.

```
+      data=Mental)  # parallel=TRUE imposes proportional odds structure
> summary(fit)
                Estimate  Std. Error  z value
(Intercept):1   -0.2818      0.6230  -0.4522 # c-1 = 3 intercepts for
(Intercept):2    1.2129      0.6512   1.8626 # c=4 response categories
(Intercept):3    2.2095      0.7172   3.0807
life            -0.3189      0.1194  -2.6697
ses              1.1111      0.6143   1.8088
Residual deviance: 99.0979 on 115 degrees of freedom
Log-likelihood: -49.54895 on 115 degrees of freedom
--------------------------------------------------------------------------
```

To help us interpret the effects, we can estimate response category probabilities. First, consider the SES effect. At the mean life events of 4.3, $\hat{P}(y = 1) = 0.37$ at high SES (i.e., $x_2 = 1$) and $\hat{P}(y = 1) = 0.16$ at low SES ($x_2 = 0$). Next, consider the life events effect. For high SES, $\hat{P}(y = 1)$ changes from 0.70 to 0.12 between the sample minimum of 0 and maximum of 9 life events; for low SES, it changes from 0.43 to 0.04. Comparing 0.70 to 0.43 at the minimum life events and 0.12 to 0.04 at the maximum provides a further description of the SES effect. The sample effect is substantial for each predictor. The following output shows estimated response category probabilities for a few subjects in the sample. Subjects (such as subject 40) having low SES and relatively high life events have a relatively high estimated probability of being mentally impaired.

```
--------------------------------------------------------------------------
> fitted(fit)
         1       2       3       4
1   0.6249  0.2564  0.0713  0.0473 # (for life=1, ses=1)
2   0.1150  0.2518  0.2440  0.3892 # (for life=9, ses=1)
...
40  0.0410  0.1191  0.1805  0.6593 # (for life=9, ses=0)
--------------------------------------------------------------------------
```

To check the proportional odds structure, we can fit a more-complex model that permits effects to vary for the three cumulative logits. The fit is not significantly better, the likelihood-ratio test having P-value = 0.67. Estimated effects are similar for each cumulative logit, taking into account sampling error, with positive effects for SES and negative effects for life events.

```
--------------------------------------------------------------------------
> fit.nonpo <- vglm(impair ~ life + ses, family=cumulative, data=Mental)
> summary(fit.nonpo) # not using parallel=true option for propor. odds
               Estimate  Std. Error  z value
(Intercept):1  -0.1929      0.7387  -0.2611 # first cumulative logit
(Intercept):2   0.8281      0.7037   1.1768 # second cumulative logit
(Intercept):3   2.8037      0.9615   2.9160 # third cumulative logit
life:1         -0.3182      0.1597  -1.9928
life:2         -0.2740      0.1372  -1.9972
life:3         -0.3962      0.1592  -2.4883
```

```
ses:1           0.9732      0.7720   1.2605
ses:2           1.4960      0.7460   2.0055
ses:3           0.7522      0.8358   0.8999
Residual deviance: 96.7486 on 111 degrees of freedom
Log-likelihood: -48.3743 on 111 degrees of freedom
> 1 - pchisq(2*(logLik(fit.nonpo)-logLik(fit)),
+            df=df.residual(fit)-df.residual(fit.nonpo))
[1] 0.6718083 # P-value comparing to model assuming proportional odds
```

When we add an interaction term with the proportional odds structure, the fit suggests that the life events effect may be weaker at higher SES. However, the fit is not significantly better (P-value = 0.44).

```
> fit.interaction <- vglm(impair ~ life + ses + life:ses,
+                   family=cumulative(parallel=TRUE), data=Mental)
> summary(fit.interaction)
               Estimate  Std. Error  z value
(Intercept):1    0.0981      0.8110   0.1209
(Intercept):2    1.5925      0.8372   1.9022
(Intercept):3    2.6066      0.9097   2.8655
life            -0.4204      0.1903  -2.2093
ses              0.3709      1.1302   0.3282
life:ses         0.1813      0.2361   0.7679
Residual deviance: 98.5044 on 114 degrees of freedom
Log-likelihood: -49.2522 on 114 degrees of freedom
> 1 - pchisq(2*(logLik(fit.interaction)-logLik(fit)),
+     df=df.residual(fit)-df.residual(fit.interaction))
[1] 0.44108 # P-value for LR test comparing to model without interaction
```

We obtain similar substantive results with a cumulative probit model. In the underlying latent variable model, we estimate the difference between mean mental impairment at low and high levels of SES to be 0.68 standard deviations, adjusting for life events. The estimated multiple correlation for the underlying latent variable model equals 0.513. With the addition of an interaction term (not shown), this increases only to 0.531.

```
> fit.probit <- vglm(impair ~ life + ses,
+            family=cumulative(link=probit,parallel=TRUE),data=Mental)
> summary(fit.probit) # cumulative probit model
              Estimate  Std. Error  z value
(Intercept):1  -0.1612      0.3755  -0.4293
(Intercept):2   0.7456      0.3864   1.9299
(Intercept):3   1.3392      0.4123   3.2484
life           -0.1953      0.0692  -2.8236
ses             0.6834      0.3627   1.8843
Residual deviance: 98.8397 on 115 degrees of freedom
```

```
Log-likelihood:   -49.4198 on 115 degrees of freedom

> lp <- -0.1953*life + 0.6834*ses # linear predictor for latent model
> sqrt(var(lp)/(var(lp) + 1)) # corr(y*, fitted) for latent variable
[1] 0.513
```

CHAPTER NOTES

Section 6.1: Nominal Responses: Baseline-Category Logit Models

6.1 **BCL and multivariate GLM**: After Mantel (1966), early development and application of baseline-category logit models were primarily in the econometrics literature (e.g., Theil 1969). For details about multivariate GLMs, see Fahrmeir and Tutz (2001).

6.2 **Infinite estimates**: When a choice of baseline category causes complete or quasi-complete separation (Section 5.4.2) to occur for each logit when paired with that category, some ML estimates and *SE* values are infinite or do not exist. Approaches to produce finite estimates include the Bayesian (Note 10.7) and a generalization of a penalized likelihood approach (Kosmidis and Firth 2011) presented in Section 11.1.7 for binary data.

6.3 **Discrete choice**: Daniel McFadden (1974) proposed the discrete-choice model, incorporating explanatory variables that are characteristics of the choices. In 2000, McFadden won the Nobel Prize in Economic Sciences, partly for this work. Greene (2011, Chapter 17–18) and Train (2009) surveyed many generalizations of the model since then, such as to handle nested choice structure.

Section 6.2: Ordinal Responses: Cumulative Logit and Probit Models

6.4 **Proportional odds**: Although not the first to use the proportional odds form of cumulative logit model, the landmark article on modeling ordinal data by McCullagh (1980) popularized it. Peterson and Harrell (1990) proposed a *partial proportional odds model* in which a subset of the explanatory variables have that structure[7]. McKelvey and Zavoina (1975) presented the latent variable motivation for the cumulative probit model. Agresti (2010) reviewed ways of modeling ordinal responses.

6.5 **Infinite estimates**: When at least some ML estimates are infinite in an ordinal model, approaches to produce finite estimates include the Bayesian (Note 10.7) and a reduced-bias solution that corresponds to a parameter-dependent adjustment of the multinomial counts (Kosmidis 2014).

EXERCISES

6.1 Show that the multinomial variate $\boldsymbol{y} = (y_1, \ldots, y_{c-1})^{\mathrm{T}}$ (with $y_j = 1$ if outcome j occurred and 0 otherwise) for a single trial with parameters $(\pi_1, \ldots, \pi_{c-1})$

[7]This model can be fitted with the `vglm` function in the `VGAM` R package.

has distribution in the $(c-1)$-parameter exponential dispersion family, with baseline-category logits as natural parameters.

6.2 For the baseline-category logit model without constraints on parameters,

$$\pi_{ij} = \frac{\exp(\boldsymbol{x}_i\boldsymbol{\beta}_j)}{\sum_{h=1}^{c}\exp(\boldsymbol{x}_i\boldsymbol{\beta}_h)},$$

show that dividing numerator and denominator by $\exp(\boldsymbol{x}_i\boldsymbol{\beta}_c)$ yields new parameters $\boldsymbol{\beta}_j^* = \boldsymbol{\beta}_j - \boldsymbol{\beta}_c$ that satisfy $\boldsymbol{\beta}_c^* = \mathbf{0}$. Thus, without loss of generality, we can take $\boldsymbol{\beta}_c = \mathbf{0}$.

6.3 Derive Equation (6.3) for the rate of change. Show how the equation for binary models is a special case.

6.4 With three outcome categories and a single explanatory variable, suppose

$$\pi_{ij} = \exp(\beta_{j0} + \beta_j x_i)/[1 + \exp(\beta_{10} + \beta_1 x_i) + \exp(\beta_{20} + \beta_2 x_i)],$$

$j = 1, 2$. Show that π_{i3} is (**a**) decreasing in x_i if $\beta_1 > 0$ and $\beta_2 > 0$, (**b**) increasing in x_i if $\beta_1 < 0$ and $\beta_2 < 0$, and (**c**) nonmonotone when β_1 and β_2 have opposite signs.

6.5 Derive the deviance expression in Equation (6.5) by deriving the corresponding likelihood-ratio test.

6.6 For a multinomial response, let u_{ij} denote the utility of response outcome j for subject i. Suppose that

$$u_{ij} = \boldsymbol{x}_i\boldsymbol{\beta}_j + \epsilon_{ij},$$

and the response outcome for subject i is the value of j having maximum utility. When $\{\epsilon_{ij}\}$ are assumed to be iid standard normal, this model is the simplest form of the *multinomial probit model* (Aitchison and Bennett 1970).

a. Explain why $(\beta_{ak} - \beta_{bk})$ describes the effect of a 1-unit increase in explanatory variable k on the difference in mean utilities, as measured in terms of the number of standard deviations of the utility distribution.

b. For observation i, explain why the probability of outcome in category j is

$$\pi_{ij} = \int \phi(u_{ij} - \boldsymbol{x}_i\boldsymbol{\beta}_j)\prod_{k\neq j}\Phi(u_{ij} - \boldsymbol{x}_i\boldsymbol{\beta}_k)du_{ij},$$

for the standard normal pdf ϕ and cdf Φ. Explain how to form the likelihood function.

6.7 Derive the likelihood equations and the information matrix for the discrete-choice model (6.6).

6.8 Consider the baseline-category logit model (6.1).

a. Suppose we impose the structure $\boldsymbol{\beta}_j = j\boldsymbol{\beta}$, for $j = 1, \ldots, c-1$. Does this model treat the response as ordinal or nominal? Explain.

b. Show that the model in (**a**) has proportional odds structure when the $c-1$ logits are formed using pairs of adjacent categories.

6.9 Section 5.3.4 introduced Fisher's exact test for 2×2 contingency tables. For testing independence in a $r \times c$ table in which the data are c independent multinomials, derive a conditional distribution that does not depend on unknown parameters. Explain a way to use it to conduct a small-sample exact test.

6.10 Does it make sense to use the cumulative logit model of proportional odds form with a nominal-scale response variable? Why or why not? Is the model a special case of a baseline-category logit model? Explain.

6.11 Show how to express the cumulative logit model of proportional odds form as a multivariate GLM (6.4).

6.12 For a binary explanatory variable, explain why the cumulative logit model with proportional odds structure is unlikely to fit well if, for an underlying latent response, the two groups have similar location but very different dispersion.

6.13 Consider the cumulative logit model, $\text{logit}[P(y_i \le j)] = \alpha_j + \beta_j x_i$.

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

b. When x_i 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.

6.14 For the cumulative link model, $G^{-1}[P(y_i \le j)] = \alpha_j + \boldsymbol{x}_i\boldsymbol{\beta}$, show that for $1 \le j < k \le c-1$, $P(y_i \le k)$ equals $P(y_i \le j)$ at $\boldsymbol{x}^*$, where $\boldsymbol{x}^*$ is obtained by increasing component h of $\boldsymbol{x}_i$ by $(\alpha_k - \alpha_j)/\beta_h$ for each h. Interpret.

6.15 For an ordinal multinomial response with c categories, let

$$\omega_{ij} = P(y_i = j \mid y_i \ge j) = \frac{\pi_{ij}}{\pi_{ij} + \cdots + \pi_{ic}}, \quad j = 1, \ldots, c-1.$$

The *continuation-ratio logit model* is

$$\text{logit}(\omega_{ij}) = \alpha_j + \boldsymbol{x}_i\boldsymbol{\beta}_j. \quad j = 1, \ldots, c-1.$$

a. Interpret (i) $\boldsymbol{\beta}_j$, (ii) $\boldsymbol{\beta}$ for the simpler model with proportional odds structure. Describe a survival application for which such sequential formation of logits might be natural.

b. Express the multinomial probability for $(y_{i1}, \ldots, y_{ic})$ in the form $p(y_{i1})p(y_{i2} \mid y_{i1}) \cdots p(y_{ic} \mid y_{i1}, \ldots, y_{i,c-1})$. Using this, explain why the $\{\hat{\boldsymbol{\beta}}_j\}$ are independent and how it is possible to fit the model using binary logistic GLMs.

6.16 Consider the null multinomial model, having the same probabilities $\{\pi_j\}$ for every observation. Let $\gamma = \sum_j b_j \pi_j$, and suppose that $\pi_j = f_j(\theta) > 0$, $j = 1, \ldots, c$. For sample proportions $\{p_j = n_j/N\}$, let $S = \sum_j b_j p_j$. Let $T = \sum_j b_j \hat{\pi}_j$, where $\hat{\pi}_j = f_j(\hat{\theta})$, for the ML estimator $\hat{\theta}$ of θ.

a. Show that $\text{var}(S) = [\sum_j b_j^2 \pi_j - (\sum_j b_j \pi_j)^2]/N$.

b. Using the delta method, show $\text{var}(T) \approx [\text{var}(\hat{\theta})][\sum_j b_j f_j'(\theta)]^2$.

c. By computing the information for $L(\theta) = \sum_j n_j \log[f_j(\theta)]$, show that $\text{var}(\hat{\theta})$ is approximately $[N \sum_j (f_j'(\theta))^2 / f_j(\theta)]^{-1}$.

d. Asymptotically, show that a consequence of model parsimony is that $\text{var}[\sqrt{N}(T - \gamma)] \leq \text{var}[\sqrt{N}(S - \gamma)]$.

6.17 A response scale has the categories (strongly agree, mildly agree, mildly disagree, strongly disagree, do not know). A two-part model uses a logistic regression model for the probability of a don't know response and a separate ordinal model for the ordered categories conditional on response in one of those categories. Explain how to construct a likelihood function to fit the two parts simultaneously.

6.18 The file `Alligators2.dat` at the text website is an expanded version of Table 6.1 that also includes the alligator's gender. Using all the explanatory variables, use model-building methods to select a model for predicting primary food choice. Conduct inference and interpret effects in that model.

6.19 For 63 alligators caught in Lake George, Florida, the file `Alligators3.dat` at the text website classifies primary food choice as (fish, invertebrate, other) and shows alligator length in meters. Analyze these data.

6.20 The following R output shows output from fitting a cumulative logit model to data from the US 2008 General Social Survey. For subject i let y_i = belief in existence of heaven (1 = yes, 2 = unsure, 3 = no), x_{i1} = gender (1 = female, 0 = male) and x_{i2} = race (1 = black, 0 = white). State the model fitted here, and interpret the race and gender effects. Test goodness-of-fit and construct confidence intervals for the effects.

```
------------------------------------------------------------------------------
> cbind(race, gender, y1, y2, y3)
     race gender  y1  y2 y3
[1,]    1      1  88  16  2
[2,]    1      0  54   7  5
[3,]    0      1 397 141 24
[4,]    0      0 235 189 39
> summary(vglm(cbind(y1,y2,y3)~gender+race,family=cumulative(parallel=T)))
              Estimate  Std. Error   z value
(Intercept):1   0.0763      0.0896   0.8515
(Intercept):2   2.3224      0.1352  17.1749
gender          0.7696      0.1225   6.2808
race            1.0165      0.2106   4.8266
Residual deviance: 9.2542 on 4 degrees of freedom
Log-likelihood: -23.3814 on 4 degrees of freedom
------------------------------------------------------------------------------
```

6.21 Refer to the previous exercise. Consider the model

$$\log(\pi_{ij}/\pi_{i3}) = \alpha_j + \beta_j^G x_{i1} + \beta_j^R x_{i2}, \quad j = 1, 2.$$

a. Fit the model and report prediction equations for $\log(\pi_{i1}/\pi_{i3})$, $\log(\pi_{i2}/\pi_{i3})$, and $\log(\pi_{i1}/\pi_{i2})$.

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

c. Conduct a likelihood-ratio test of the hypothesis that opinion is independent of gender, given race. Interpret.

6.22 Refer to Exercise 5.33. The color of the female crab is a surrogate for age, with older crabs being darker. Analyze whether any characteristics or combinations of characteristics of the attached male crab can help to predict a female crab's color. Prepare a short report that summarizes your analyses and findings.

6.23 A 1976 article by M. Madsen (*Scand. J. Stat.* **3**: 97–106) showed a $4 \times 2 \times 3 \times 3$ contingency table (the file `Satisfaction.dat` at the text website) that cross classifies a sample of residents of Copenhagen on the type of housing, degree of contact with other residents, feeling of influence on apartment management, and satisfaction with housing conditions. Treating satisfaction as the response variable, analyze these data.

6.24 At the website `sda.berkeley.edu/GSS` for the General Social Survey, download a contingency table relating the variable GRNTAXES (about paying higher taxes to help the environment) to two other variables, using the survey results from 2010 by specifying *year(2010)* in the "Selection Filter." Model the data, and summarize your analysis and interpretations.