Chapter 6

Multinomial Response Models

We now turn our attention to regression models for the analysis of categorical dependent variables with more than two response categories. Several of the models that we will study may be considered generalizations of logistic regression analysis to polychotomous data. We first consider models that may be used with purely qualitative or *nominal* data, and then move on to models for *ordinal* data, where the response categories are ordered.

6.1 The Nature of Multinomial Data

Let me start by introducing a simple dataset that will be used to illustrate the multinomial distribution and multinomial response models.

6.1.1 The Contraceptive Use Data

Table 6.1 was reconstructed from weighted percents found in Table 4.7 of the final report of the Demographic and Health Survey conducted in El Salvador in 1985 (FESAL-1985). The table shows 3165 currently married women classified by age, grouped in five-year intervals, and current use of contraception, classified as sterilization, other methods, and no method.

A fairly standard approach to the analysis of data of this type could treat the two variables as responses and proceed to investigate the question of independence. For these data the hypothesis of independence is soundly rejected, with a likelihood ratio χ^2 of 521.1 on 12 d.f.

TABLE 6.1: Current Use of Contraception By Age
Currently Married Women. El Salvador, 1985

Age	Contraceptive Method			All
	Ster.	Other	None	
15–19	3	61	232	296
20–24	80	137	400	617
25–29	216	131	301	648
30–34	268	76	203	547
35–39	197	50	188	435
40–44	150	24	164	338
45–49	91	10	183	284
All	1005	489	1671	3165

In this chapter we will view contraceptive use as the response and age as a predictor. Instead of looking at the joint distribution of the two variables, we will look at the conditional distribution of the response, contraceptive use, given the predictor, age. As it turns out, the two approaches are intimately related.

6.1.2 The Multinomial Distribution

Let us review briefly the multinomial distribution that we first encountered in Chapter 5. Consider a random variable Y_i that may take one of several discrete values, which we index $1, 2, \ldots, J$. In the example the response is contraceptive use and it takes the values 'sterilization', 'other method' and 'no method', which we index 1, 2 and 3. Let

$$\pi_{ij} = \Pr\{Y_i = j\} \tag{6.1}$$

denote the probability that the i-th response falls in the j-th category. In the example π_{i1} is the probability that the i-th respondent is 'sterilized'.

Assuming that the response categories are mutually exclusive and exhaustive, we have $\sum_{j=1}^{J} \pi_{ij} = 1$ for each i, i.e. the probabilities add up to one for each individual, and we have only $J - 1$ parameters. In the example, once we know the probability of 'sterilized' and of 'other method' we automatically know by subtraction the probability of 'no method'.

For *grouped data* it will be convenient to introduce auxiliary random variables representing counts of responses in the various categories. Let n_i denote the number of cases in the i-th group and let Y_{ij} denote the number

of responses from the i-th group that fall in the j-th category, with observed value y_{ij}.

In our example i represents age groups, n_i is the number of women in the i-th age group, and y_{i1}, y_{i2}, and y_{i3} are the numbers of women sterilized, using another method, and using no method, respectively, in the i-th age group. Note that $\sum_j y_{ij} = n_i$, i.e. the counts in the various response categories add up to the number of cases in each age group.

For *individual data* $n_i = 1$ and Y_{ij} becomes an indicator (or dummy) variable that takes the value 1 if the i-th response falls in the j-th category and 0 otherwise, and $\sum_j y_{ij} = 1$, since one and only one of the indicators y_{ij} can be 'on' for each case. In our example we could work with the 3165 records in the individual data file and let y_{i1} be one if the i-th woman is sterilized and 0 otherwise.

The probability distribution of the counts Y_{ij} given the total n_i is given by the *multinomial* distribution

$$\Pr\{Y_{i1} = y_{i1}, \ldots, Y_{iJ} = y_{iJ}\} = \binom{n_i}{y_{i1}, \ldots, y_{iJ}} \pi_{i1}^{y_{i1}} \ldots \pi_{iJ}^{y_{iJ}} \qquad (6.2)$$

The special case where $J = 2$ and we have only two response categories is the binomial distribution of Chapter 3. To verify this fact equate $y_{i1} = y_i$, $y_{i2} = n_i - y_i$, $\pi_{i1} = \pi_i$, and $\pi_{i2} = 1 - \pi_i$.

6.2 The Multinomial Logit Model

We now consider models for the probabilities π_{ij}. In particular, we would like to consider models where these probabilities depend on a vector $\mathbf{x}_i$ of covariates associated with the i-th individual or group. In terms of our example, we would like to model how the probabilities of being sterilized, using another method or using no method at all depend on the woman's age.

6.2.1 Multinomial Logits

Perhaps the simplest approach to multinomial data is to nominate one of the response categories as a baseline or reference cell, calculate log-odds for all other categories relative to the baseline, and then let the log-odds be a linear function of the predictors.

Typically we pick the *last* category as a baseline and calculate the odds that a member of group i falls in category j as opposed to the baseline as π_{i1}/π_{iJ}. In our example we could look at the odds of being sterilized rather

than using no method, and the odds of using another method rather than no method. For women aged 45–49 these odds are 91:183 (or roughly 1 to 2) and 10:183 (or 1 to 18).

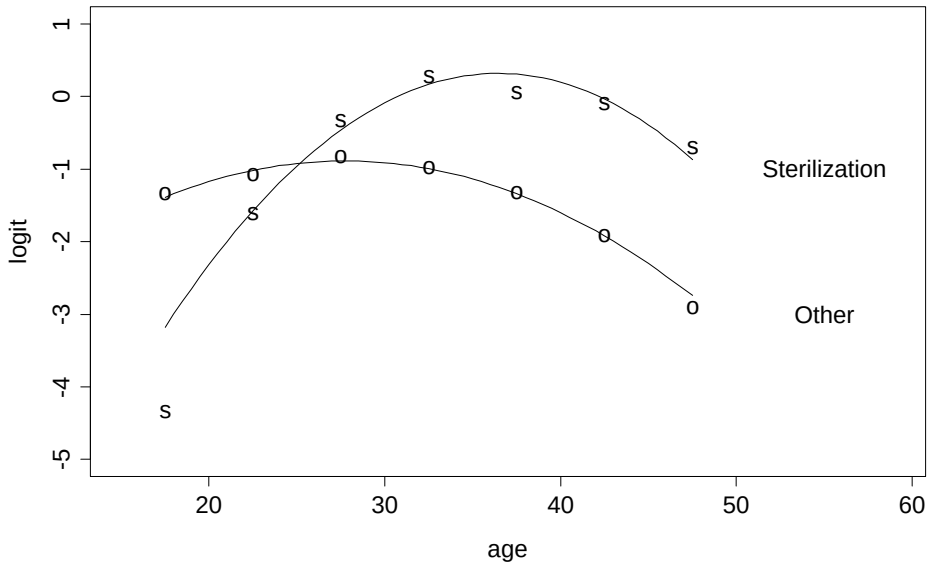

FIGURE 6.1: Log-Odds of Sterilization vs. No Method and Other Method vs. No Method, by Age

Figure 6.1 shows the empirical log-odds of sterilization and other method (using no method as the reference category) plotted against the mid-points of the age groups. (Ignore for now the solid lines.) Note how the log-odds of sterilization increase rapidly with age to reach a maximum at 30–34 and then decline slightly. The log-odds of using other methods rise gently up to age 25–29 and then decline rapidly.

6.2.2 Modeling the Logits

In the multinomial logit model we assume that the log-odds of each response follow a linear model

$$\eta_{ij} = \log \frac{\pi_{ij}}{\pi_{iJ}} = \alpha_j + \mathbf{x}_i' \boldsymbol{\beta}_j, \tag{6.3}$$

where α_j is a constant and $\boldsymbol{\beta}_j$ is a vector of regression coefficients, for $j = 1, 2, \ldots, J-1$. Note that we have written the constant explicitly, so we will

assume henceforth that the model matrix $\mathbf{X}$ does not include a column of ones.

This model is analogous to a logistic regression model, except that the probability distribution of the response is multinomial instead of binomial and we have $J-1$ equations instead of one. The $J-1$ multinomial logit equations contrast each of categories $1, 2, \ldots J-1$ with category J, whereas the single logistic regression equation is a contrast between successes and failures. If $J = 2$ the multinomial logit model reduces to the usual logistic regression model.

Note that we need only $J-1$ equations to describe a variable with J response categories and that it really makes no difference which category we pick as the reference cell, because we can always convert from one formulation to another. In our example with $J = 3$ categories we contrast categories 1 versus 3 and 2 versus 3. The missing contrast between categories 1 and 2 can easily be obtained in terms of the other two, since $\log(\pi_{i1}/\pi_{i2}) = \log(\pi_{i1}/\pi_{i3}) - \log(\pi_{i2}/\pi_{i3})$.

Looking at Figure 6.1, it would appear that the logits are a quadratic function of age. We will therefore entertain the model

$$\eta_{ij} = \alpha_j + \beta_j a_i + \gamma_j a_i^2, \tag{6.4}$$

where a_i is the midpoint of the i-th age group and $j = 1, 2$ for sterilization and other method, respectively.

6.2.3 Modeling the Probabilities

The multinomial logit model may also be written in terms of the original probabilities π_{ij} rather than the log-odds. Starting from Equation 6.3 and adopting the convention that $\eta_{iJ} = 0$, we can write

$$\pi_{ij} = \frac{\exp\{\eta_{ij}\}}{\sum_{k=1}^{J} \exp\{\eta_{ik}\}}. \tag{6.5}$$

for $j = 1, \ldots, J$. To verify this result exponentiate Equation 6.3 to obtain $\pi_{ij} = \pi_{iJ} \exp\{\eta_{ij}\}$, and note that the convention $\eta_{iJ} = 0$ makes this formula valid for all j. Next sum over j and use the fact that $\sum_j \pi_{ij} = 1$ to obtain $\pi_{iJ} = 1/\sum_j \exp\{\eta_{ij}\}$. Finally, use this result on the formula for π_{ij}.

Note that Equation 6.5 will automatically yield probabilities that add up to one for each i.

6.2.4 Maximum Likelihood Estimation

Estimation of the parameters of this model by maximum likelihood proceeds by maximization of the multinomial likelihood (6.2) with the probabilities π_{ij} viewed as functions of the α_j and $\boldsymbol{\beta}_j$ parameters in Equation 6.3. This usually requires numerical procedures, and Fisher scoring or Newton-Raphson often work rather well. Most statistical packages include a multinomial logit procedure.

In terms of our example, fitting the quadratic multinomial logit model of Equation 6.4 leads to a deviance of 20.5 on 8 d.f. The associated P-value is 0.009, so we have significant lack of fit.

The quadratic age effect has an associated likelihood-ratio χ^2 of 500.6 on four d.f. ($521.1 - 20.5 = 500.6$ and $12 - 8 = 4$), and is highly significant. Note that we have accounted for 96% of the association between age and method choice ($500.6/521.1 = 0.96$) using only four parameters.

TABLE 6.2: Parameter Estimates for Multinomial Logit Model Fitted to Contraceptive Use Data

Parameter	Contrast	
	Ster. vs. None	Other vs. None
Constant	-12.62	-4.552
Linear	0.7097	0.2641
Quadratic	-0.009733	-0.004758

Table 6.2 shows the parameter estimates for the two multinomial logit equations. I used these values to calculate fitted logits for each age from 17.5 to 47.5, and plotted these together with the empirical logits in Figure 6.1. The figure suggests that the lack of fit, though significant, is not a serious problem, except possibly for the 15–19 age group, where we overestimate the probability of sterilization.

Under these circumstances, I would probably stick with the quadratic model because it does a reasonable job using very few parameters. However, I urge you to go the extra mile and try a cubic term. The model should pass the goodness of fit test. Are the fitted values reasonable?

6.2.5 The Equivalent Log-Linear Model*

Multinomial logit models may also be fit by maximum likelihood working with an equivalent log-linear model and the Poisson likelihood. (This section

will only be of interest to readers interested in the equivalence between these models and may be omitted at first reading.)

Specifically, we treat the random counts Y_{ij} as Poisson random variables with means μ_{ij} satisfying the following log-linear model

$$\log \mu_{ij} = \eta + \theta_i + \alpha_j^* + \mathbf{x}_i' \boldsymbol{\beta}_j^*, \tag{6.6}$$

where the parameters satisfy the usual constraints for identifiability. There are three important features of this model:

First, the model includes a separate parameter θ_i for each multinomial observation, i.e. each individual or group. This assures exact reproduction of the multinomial denominators n_i. Note that these denominators are fixed known quantities in the multinomial likelihood, but are treated as random in the Poisson likelihood. Making sure we get them right makes the issue of conditioning moot.

Second, the model includes a separate parameter α_j^* for each response category. This allows the counts to vary by response category, permitting non-uniform margins.

Third, the model uses interaction terms $\mathbf{x}_i' \boldsymbol{\beta}_j^*$ to represent the effects of the covariates $\mathbf{x}_i$ on the log-odds of response j. Once again we have a 'step-up' situation, where main effects in a logistic model become interactions in the equivalent log-linear model.

The log-odds that observation i will fall in response category j relative to the last response category J can be calculated from Equation 6.6 as

$$\log(\mu_{ij}/\mu_{iJ}) = (\alpha_j^* - \alpha_J^*) + \mathbf{x}_i'(\boldsymbol{\beta}_j^* - \boldsymbol{\beta}_J^*). \tag{6.7}$$

This equation is identical to the multinomial logit Equation 6.3 with $\alpha_j = \alpha_j^* - \alpha_J^*$ and $\boldsymbol{\beta}_j = \boldsymbol{\beta}_j^* - \boldsymbol{\beta}_J^*$. Thus, the parameters in the multinomial logit model may be obtained as differences between the parameters in the corresponding log-linear model. Note that the θ_i cancel out, and the restrictions needed for identification, namely $\eta_{iJ} = 0$, are satisfied automatically.

In terms of our example, we can treat the counts in the original 7×3 table as 21 independent Poisson observations, and fit a log-linear model including the main effect of age (treated as a factor), the main effect of contraceptive use (treated as a factor) and the interactions between contraceptive use (a factor) and the linear and quadratic components of age:

$$\log \mu_{ij} = \eta + \theta_i + \alpha_j^* + \beta_j^* a_i + \gamma_j^* a_i^2 \tag{6.8}$$

In practical terms this requires including six dummy variables representing the age groups, two dummy variables representing the method choice categories, and a total of four interaction terms, obtained as the products of

the method choice dummies by the mid-point a_i and the square of the mid-point a_i^2 of each age group. Details are left as an exercise. (But see the Stata notes.)

6.3 The Conditional Logit Model

In this section I will describe an extension of the multinomial logit model that is particularly appropriate in models of choice behavior, where the explanatory variables may include attributes of the choice alternatives (for example cost) as well as characteristics of the individuals making the choices (such as income). To motivate the extension I will first reintroduce the multinomial logit model in terms of an underlying latent variable.

6.3.1 A General Model of Choice

Suppose that Y_i represents a discrete choice among J alternatives. Let U_{ij} represent the value or *utility* of the j-th choice to the i-th individual. We will treat the U_{ij} as independent random variables with a systematic component η_{ij} and a random component ϵ_{ij} such that

$$U_{ij} = \eta_{ij} + \epsilon_{ij}. \tag{6.9}$$

We assume that individuals act in a rational way, maximizing their utility. Thus, subject i will choose alternative j if U_{ij} is the largest of $U_{i1}, \ldots, U_{iJ}$. Note that the choice has a random component, since it depends on random utilities. The *probability* that subject i will choose alternative j is

$$\pi_{ij} = \Pr\{Y_i = j\} = \Pr\{\max(U_{i1}, \ldots, U_{iJ}) = U_{ij}\}. \tag{6.10}$$

It can be shown that if the error terms ϵ_{ij} have standard Type I extreme value distributions with density

$$f(\epsilon) = \exp\{-\epsilon - \exp\{-\epsilon\}\} \tag{6.11}$$

then (see for example Maddala, 1983, pp 60–61)

$$\pi_{ij} = \frac{\exp\{\eta_{ij}\}}{\sum \exp\{\eta_{ik}\}}, \tag{6.12}$$

which is the basic equation defining the multinomial logit model.

In the special case where $J = 2$, individual i will choose the first alternative if $U_{i1} - U_{i2} > 0$. If the random utilities U_{ij} have independent

extreme value distributions, their difference can be shown to have a logistic distribution, and we obtain the standard logistic regression model.

Luce (1959) derived Equation 6.12 starting from a simple requirement that the odds of choosing alternative j over alternative k should be independent of the choice set for all pairs j, k. This property is often referred to as the axiom of *independence from irrelevant alternatives*. Whether or not this assumption is reasonable (and other alternatives are indeed irrelevant) depends very much on the nature of the choices.

A classical example where the multinomial logit model does not work well is the so-called "red/blue bus" problem. Suppose you have a choice of transportation between a train, a red bus and a blue bus. Suppose half the people take the train and half take the bus. Suppose further that people who take the bus are indifferent to the color, so they distribute themselves equally between the red and the blue buses. The choice probabilities of $\pi = (.50, .25, .25)$ would be consistent with expected utilities of $\eta = (\log 2, 0, 0)$.

Suppose now the blue bus service is discontinued. You might expect that all the people who used to take the blue bus would take the red bus instead, leading to a 1:1 split between train and bus. On the basis of the expected utilities of $\log 2$ and 0, however, the multinomial logit model would predict a 2:1 split.

Keep this caveat in mind as we consider modeling the expected utilities.

6.3.2 Multinomial Logits

In the usual multinomial logit model, the expected utilities η_{ij} are modeled in terms of characteristics of the individuals, so that

$$\eta_{ij} = \mathbf{x}_i'\boldsymbol{\beta}_j.$$

Here the regression coefficients $\boldsymbol{\beta}_j$ may be interpreted as reflecting the effects of the covariates on the odds of making a given choice (as we did in the previous section) or on the underlying utilities of the various choices.

A somewhat restrictive feature of the model is that the same attributes $\mathbf{x}_i$ are used to model the utilities of all J choices.

6.3.3 Conditional Logits

McFadden (1973) proposed modeling the expected utilities η_{ij} in terms of characteristics of the alternatives rather than attributes of the individuals. If $\mathbf{z}_j$ represents a vector of characteristics of the j-th alternative, then he

postulated the model

$$\eta_{ij} = \mathbf{z}_j'\boldsymbol{\gamma}.$$

This model is called the *conditional logit* model, and turns out to be equivalent to a log-linear model where the main effect of the response is represented in terms of the covariates $\mathbf{z}_j$.

Note that with J response categories the response margin may be reproduced exactly using any $J-1$ linearly independent attributes of the choices. Generally one would want the dimensionality of z_j to be substantially less than J. Consequently, conditional logit models are often used when the number of possible choices is large.

6.3.4 Multinomial/Conditional Logits

A more general model may be obtained by combining the multinomial and conditional logit formulations, so the underlying utilities η_{ij} depend on characteristics of the individuals as well as attributes of the choices, or even variables defined for combinations of individuals and choices (such as an individual's perception of the value of a choice). The general model is usually written as

$$\eta_{ij} = \mathbf{x}_i'\boldsymbol{\beta}_j + \mathbf{z}_{ij}'\boldsymbol{\gamma} \tag{6.13}$$

where $\mathbf{x}_i$ represents characteristics of the individuals that are constant across choices, and $\mathbf{z}_{ij}$ represents characteristics that vary across choices (whether they vary by individual or not).

Some statistical packages have procedures for fitting conditional logit models to datasets where each combination of individual and possible choice is treated as a separate observation. These models may also be fit using any package that does Poisson regression. If the last response category is used as the baseline or reference cell, so that $\eta_{iJ} = 0$ for all i, then the $\mathbf{z}_{ij}$ should be entered in the model as differences from the last category. In other words, you should use $\mathbf{z}^*_{ij} = \mathbf{z}_{ij} - \mathbf{z}_{iJ}$ as the predictor.

6.3.5 Multinomial/Conditional Probits

Changing the distribution of the error term in Equation 6.9 leads to alternative models. A popular alternative to the logit models considered so far is to assume that the ϵ_{ij} have independent standard normal distributions for all i, j. The resulting model is called the multinomial/conditional *probit* model, and produces results very similar to the multinomial/conditional logit model after standardization.

A more attractive alternative is to retain independence across subjects but allow dependence across alternatives, assuming that the vector $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{iJ})'$ has a *multivariate* normal distribution with mean vector $\mathbf{0}$ and arbitrary correlation matrix $\mathbf{R}$. (As usual with latent variable formulations of binary or discrete response models, the variance of the error term cannot be separated from the regression coefficients. Setting the variances to one means that we work with a correlation matrix rather than a covariance matrix.)

The main advantage of this model is that it allows correlation between the utilities that an individual assigns to the various alternatives. The main difficulty is that fitting the model requires evaluating probabilities given by multidimensional normal integrals, a limitation that effectively restricts routine practical application of the model to problems involving no more than three or four alternatives.

For further details on discrete choice models see Chapter 3 in Maddala (1983).

6.4 The Hierarchical Logit Model

The strategy used in Section 6.2.1 to define logits for multinomial response data, namely nominating one of the response categories as a baseline, is only one of many possible approaches.

6.4.1 Nested Comparisons

An alternative strategy is to define a hierarchy of *nested* comparisons between two subsets of responses, using an ordinary logit model for each comparison. In terms of the contraceptive use example, we could consider (1) the odds of using some form of contraception, as opposed to none, and (2) the odds of being sterilized among users of contraception. For women aged 15–49 these odds are 1494:1671 (or roughly one to one) and 1005:489 (or roughly two to one).

The hierarchical or nested approach is very attractive if you assume that individuals make their decisions in a sequential fashion. In terms of contraceptive use, for example, women may first decide whether or nor they will use contraception. Those who decide to use then face the choice of a method. This sequential approach may also provide a satisfactory model for the "red/blue bus" choice.

Of course it is also possible that the decision to use contraception would

be affected by the types of methods available. If that is the case, a multinomial logit model may be more appropriate.

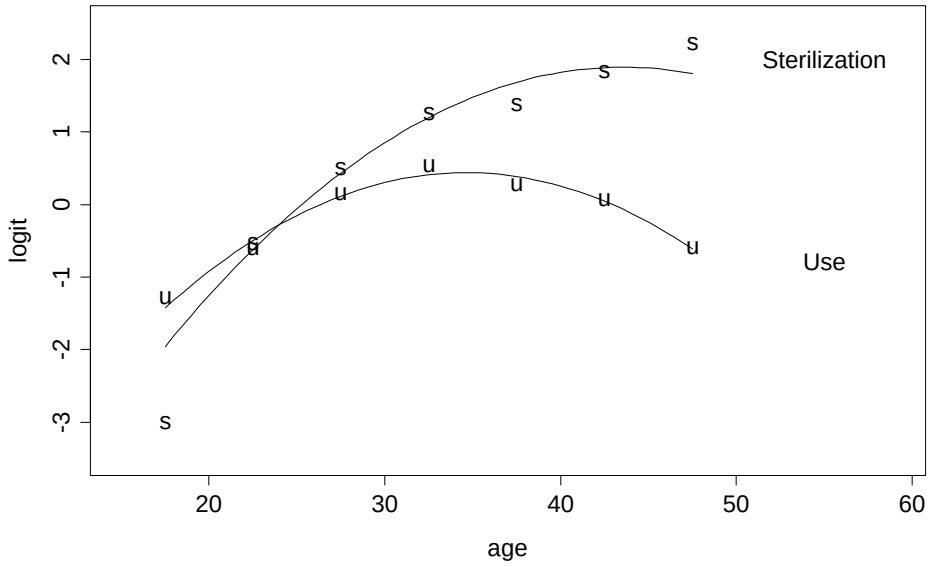

FIGURE 6.2: Log-Odds of Contraceptive Use vs. No Use and Sterilization vs. Other Method, by Age.

Figure 6.2 shows the empirical log-odds of using any method rather than no method, and of being sterilized rather than using another method among users, by age. Note that contraceptive use increases up to age 35–39 and then declines, whereas the odds of being sterilized among users increase almost monotonically with age.

The data suggest that the hierarchical logits may be modeled as quadratic functions of age, just as we did for the multinomial logits. We will therefore consider the model

$$\eta_{ij} = \alpha_j + \beta_j a_i + \gamma_j a_i^2, \tag{6.14}$$

where a_i is the mid-point of the i-th age group, $j = 1$ for the contraceptive use equation and $j = 2$ for the method choice equation.

6.4.2 Maximum Likelihood Estimation

An important practical feature of the hierarchical logit model is that the multinomial likelihood factors out into a product of binomial likelihoods, which may then be maximized separately.

I will illustrate using the contraceptive use data with 3 response categories, but the idea is obviously more general. The contribution of the i-th individual or group to the multinomial likelihood (ignoring constants) has the form

$$L_i = \pi_{i1}^{y_{i1}} \pi_{i2}^{y_{i2}} \pi_{i3}^{y_{i3}}, \tag{6.15}$$

where the π_{ij} are the probabilities and the y_{ij} are the corresponding counts of women sterilized, using other methods, and using no methods, respectively.

Multiply and divide this equation by $(\pi_{i1}+\pi_{i2})^{y_{i1}+y_{i2}}$, which is the probability of using contraception raised to the total number of users of contraception, to obtain

$$L_i = \left(\frac{\pi_{i1}}{\pi_{i1}+\pi_{i2}}\right)^{y_{i1}} \left(\frac{\pi_{i2}}{\pi_{i1}+\pi_{i2}}\right)^{y_{i2}} (\pi_{i1}+\pi_{i2})^{y_{i1}+y_{i2}} \pi_{i3}^{y_{i3}}. \tag{6.16}$$

Let $\rho_{i1} = \pi_{i1} + \pi_{i2}$ denote the probability of using contraception in age group i, and let $\rho_{i2} = \pi_{i1}/(\pi_{i1} + \pi_{i2})$ denote the *conditional* probability of being sterilized given that a woman is using contraception. Using this notation we can rewrite the above equation as

$$L_i = \rho_{i2}^{y_{i1}} (1-\rho_{i2})^{y_{i2}} \rho_{i1}^{y_{i1}+y_{i2}} (1-\rho_{i1})^{y_{i3}}. \tag{6.17}$$

The two right-most terms involving the probability of using contraception ρ_{i1} may be recognized, except for constants, as a standard binomial likelihood contrasting users and non-users. The two terms involving the conditional probability of using sterilization ρ_{i2} form, except for constants, a standard binomial likelihood contrasting sterilized women with users of other methods. As long as the parameters involved in the two equations are distinct, we can maximize the two likelihoods separately.

In view of this result we turn to Table 6.1 and fit two separate models. Fitting a standard logit model to the contraceptive use contrast (sterilization or other method vs. no method) using linear and quadratic terms on age gives a deviance of 6.12 on four d.f. and the parameter estimates shown in the middle column of Table 6.3. Fitting a similar model to the method choice contrast (sterilization vs. other method, restricted to users) gives a deviance of 10.77 on four d.f. and the parameter estimates shown in the rightmost column of Table 6.3.

The combined deviance is 16.89 on 8 d.f. (6.12+10.77 = 16.89 and 4+4 = 8). The associated P-value is 0.031, indicating lack of fit significant at the 5% level. Note, however, that the hierarchical logit model provides a somewhat better fit to these data than the multinomial logit model considered earlier, which had a deviance of 20.5 on the same 8 d.f.

TABLE 6.3: Parameter Estimates for Hierarchical Logit Model Fitted to Contraceptive Use Data

Parameter	Contrast	
	Use vs. No Use	Ster. vs. Other
Constant	-7.180	-8.869
Linear	0.4397	0.4942
Quadratic	-0.006345	-0.005674

To look more closely at goodness of fit I used the parameter estimates shown on Table 6.3 to calculate fitted logits and plotted these in Figure 6.2 against the observed logits. The quadratic model seems to do a reasonable job with very few parameters, particularly for overall contraceptive use. The method choice equation overestimates the odds of choosing sterilization for the age group 15–19, a problem shared by the multinomial logit model.

The parameter estimates may also be used to calculate illustrative odds of using contraception or sterilization at various ages. Going through these calculations you will discover that the odds of using some form of contraception increase 80% between ages 25 and 35. On the other hand, the odds of being sterilized among contraceptors increase three and a half times between ages 25 and 35.

6.4.3 Choice of Contrasts

With three response categories the only possible set of nested comparisons (aside from a simple reordering of the categories) is

{1,2} versus {3}, and
{1} versus {2}.

With four response categories there are two main alternatives. One is to contrast

{1, 2} versus {3, 4},
{1} versus {2}, and
{3} versus {4}.

The other compares

{1} versus {2, 3, 4},
{2} versus {3, 4}, and
{3} versus {4}.

The latter type of model, where one considers the odds of response $Y = j$ relative to responses $Y \geq j$, is known as a *continuation ratio* model (see Fienberg, 1980), and may be appropriate when the response categories are ordered.

More generally, any set of $J - 1$ linearly independent contrasts can be selected for modeling, but only orthogonal contrasts lead to a factorization of the likelihood function. The choice of contrasts should in general be based on the logic of the situation.

6.5 Models for Ordinal Response Data

Most of the models discussed so far are appropriate for the analysis of nominal responses. They may be applied to *ordinal* data as well, but the models make no explicit use of the fact that the response categories are ordered. We now consider models designed specifically for the analysis of responses measured on an ordinal scale. Our discussion follows closely McCullagh (1980).

6.5.1 Housing Conditions in Copenhagen

We will illustrate the application of models for ordinal data using the data in Table 6.4, which was first published by Madsen (1976) and was reproduced in Agresti (1990, p. 341). The table classifies 1681 residents of twelve areas in Copenhagen in terms of the type of housing they had, their feeling of influence on apartment management, their degree of contact with other residents, and their satisfaction with housing conditions.

In our analysis of these data we will treat housing satisfaction as an ordered response, with categories low, medium and high, and the other three factors as explanatory variables.

6.5.2 Cumulative Link Models

All of the models to be considered in this section arise from focusing on the *cumulative* distribution of the response. Let $\pi_{ij} = \Pr\{Y_i = j\}$ denote the probability that the response of an individual with characteristics $\mathbf{x}_i$ falls in the j-th category, and let γ_{ij} denote the corresponding cumulative probability

$$\gamma_{ij} = \Pr\{Y_i \leq j\} \tag{6.18}$$

TABLE 6.4: Housing Condition in Copenhagen

Housing Type	Influence	Contact	Satisfaction		
			low	medium	high
Tower block	low	low	21	21	28
		high	14	19	37
	medium	low	34	22	36
		high	17	23	40
	high	low	10	11	36
		high	3	5	23
Apartments	low	low	61	23	17
		high	78	46	43
	medium	low	43	35	40
		high	48	45	86
	high	low	26	18	54
		high	15	25	62
Atrium houses	low	low	13	9	10
		high	20	23	20
	medium	low	8	8	12
		high	10	22	24
	high	low	6	7	9
		high	7	10	21
Terraced houses	low	low	18	6	7
		high	57	23	13
	medium	low	15	13	13
		high	31	21	13
	high	low	7	5	11
		high	5	6	13

that the response falls in the j-th category *or below*, so

$$\gamma_{ij} = \pi_{i1} + \pi_{i2} + \ldots + \pi_{ij}. \tag{6.19}$$

Let $g(.)$ denote a link function mapping probabilities to the real line. Then the class of models that we will consider assumes that the transformed *cumulative* probabilities are a linear function of the predictors, of the form

$$g(\gamma_{ij}) = \theta_j + \mathbf{x}_i'\boldsymbol{\beta}. \tag{6.20}$$

In this formulation θ_j is a constant representing the baseline value of the transformed cumulative probability for category j, and $\boldsymbol{\beta}$ represents the

effect of the covariates on the transformed cumulative probabilities. Since we write the constant explicitly, we assume that the predictors do not include a column of ones. Note that there is just one equation: if x_{ik} increases by one, then *all* transformed cumulative probabilities increase by β_k. Thus, this model is more parsimonious than a multinomial logit or a hierarchical logit model; by focusing on the cumulative probabilities we can postulate a single effect. We will return to the issue of interpretation when we consider specific link functions.

These models can also be interpreted in terms of a *latent variable*. Specifically, suppose that the manifest response Y_i results from grouping an underlying continuous variable Y_i^* using cut-points $\theta_1 < \theta_2 < \ldots < \theta_{J-1}$, so that Y_i takes the value 1 if Y_i^* is below θ_1, the value 2 if Y_i^* is between θ_1 and θ_2, and so on, taking the value J if Y_i^* is above θ_{J-1}. Figure 6.3 illustrates this idea for the case of five response categories.

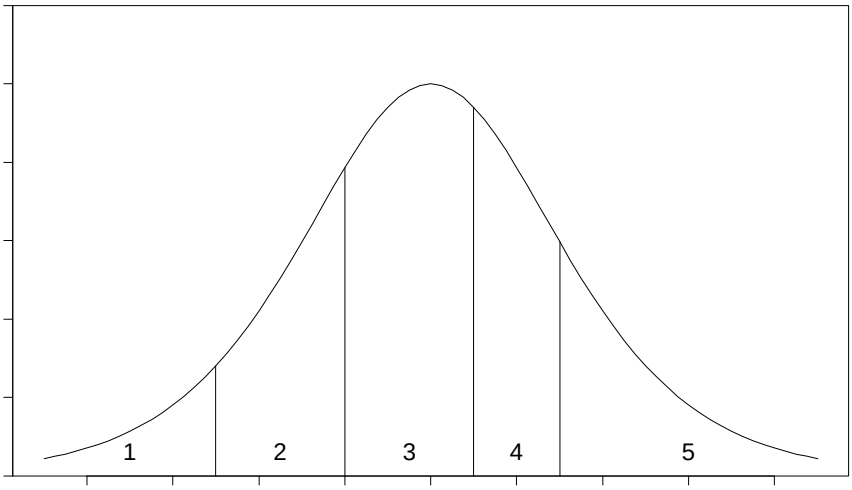

FIGURE 6.3: An Ordered Response and its Latent Variable

Suppose further that the underlying continuous variable follows a linear model of the form

$$Y_i^* = \mathbf{x}_i'\boldsymbol{\beta}^* + \epsilon_i, \tag{6.21}$$

where the error term ϵ_i has c.d.f. $F(\epsilon_i)$. Then, the probability that the response of the i-th individual will fall in the j-th category *or below*, given

$\mathbf{x}_i$, satisfies the equation

$$\gamma_{ij} = \Pr\{Y_i^* < \theta_j\} = \Pr\{\epsilon_i < \theta_j - \mathbf{x}_i'\boldsymbol{\beta}^*\} = F(\theta_j - \mathbf{x}_i'\boldsymbol{\beta}^*) \tag{6.22}$$

and therefore follows the general form in Equation (6.20) with link given by the inverse of the c.d.f. of the error term

$$g(\gamma_{ij}) = F^{-1}(\gamma_{ij}) = \theta_j - \mathbf{x}_i'\boldsymbol{\beta}^* \tag{6.23}$$

and coefficients $\boldsymbol{\beta}^* = -\boldsymbol{\beta}$ differing only in sign from the coefficients in the cumulative link model. Note that in both formulations we assume that the predictors $\mathbf{x}_i$ do not include a column of ones because the constant is absorbed in the cutpoints.

With grouped data the underlying continuous variable Y^* will have real existence and the cutpoints θ_j will usually be known. For example income data are often collected in broad categories, and all we known is the interval where an observation falls, i.e. < \$25,000, between \$25,000 and \$50,000, and so on.

With ordinal categorical data the underlying continuous variable will often represent a latent or unobservable trait, and the cutpoints will not be known. This would be the case, for example, if respondents are asked whether they support a balance budget amendment, and the response categories are strongly against, against, neutral, in favor, and strongly in favor. We could imagine an underlying degree of support Y_i^* and thresholds θ_1 to θ_4, such that when the support is below θ_1 one is strongly against, when the support exceeds θ_1 but not θ_2 one is against, and so on, until the case where the support exceeds θ_4 and one is strongly for the amendment.

While the latent variable interpretation is convenient, it is not always necessary, since some of the models can be interpreted directly in terms of the transformation $g(.)$ of the cumulative probabilities.

6.5.3 The Proportional Odds Model

The first model we will consider is a direct extension of the usual logistic regression model. Instead of applying the logit transformation to the response probabilities π_{ij}, however, we apply it to the *cumulative* response probabilities γ_{ij}, so that

$$\text{logit}(\gamma_{ij}) = \log \frac{\gamma_{ij}}{1-\gamma_{ij}} = \theta_j + \mathbf{x}_i'\boldsymbol{\beta}. \tag{6.24}$$

Some authors refer to this model as the ordered logit model, because it is a generalization of the logit model to ordered response categories. McCullagh

(1980) calls it the *proportional odds* model, for reasons that will be apparent presently. Exponentiating (6.24) we find that the odds of $Y_{ij} \leq j$, in words, the odds of a response in category j or below, are

$$\frac{\gamma_{ij}}{1-\gamma_{ij}} = \lambda_j \exp\{\mathbf{x}_i'\boldsymbol{\beta}\} \tag{6.25}$$

where $\lambda_j = \exp\{\theta_j\}$. The λ_j may be interpreted as the *baseline* odds of a response in category j or below when $x = 0$. The effect of the covariates x is to raise or lower the odds of a response in category j or below by the factor $\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}$. Note that the effect is a proportionate change in the odds of $Y_i \leq j$ for all response categories j. If a certain combination of covariate values doubles the odds of being in category 1, it also doubles the odds of being in category 2 or below, or in category 3 or below. Hence the name proportional odds.

This model may also be obtained from the latent variable formulation assuming that the error term ϵ_i has a standard logistic distribution. In this case the cdf is

$$F(\eta) = \frac{\exp\{\eta\}}{1+\exp\{\eta\}} \tag{6.26}$$

and the inverse cdf is the logit transformation. The $\boldsymbol{\beta}^*$ coefficients may then be interpreted as linear effects on the underlying continuous variable Y_i^*.

The proportional odds model is not a log-linear model, and therefore it can not be fit using the standard Poisson trick. It is possible, however, to use an iteratively re-weighted least squares algorithm very similar to the standard algorithm for generalized linear models, for details see McCullagh (1980).

We will illustrate this model applying it to the housing satisfaction data in Table 6.4. Let us start by noting that the log-likelihood for a saturated multinomial model that treats each of the 24 covariate patterns as a different group is -1715.71. Fitting a proportional odds model with additive effects of housing type, influence in management and contact with neighbors, yields a log-likelihood of -1739.57, which corresponds to a deviance (compared to the saturated multinomial model) of 47.73 on 40 d.f. To calculate the degrees of freedom note that the saturated multinomial model has 48 parameters (2 for each of 24 groups), while the additive proportional odds model has only 8 (2 threshold parameters, 3 for housing type, 2 for influence and one for contact). The 95% critical point of the χ^2_{40} distribution is 55.8, so you might think that this model fits the data.

To be thorough, however, we should investigate interaction effects. The models with one two-factor interaction have log-likelihoods of -1739.47 (in-

cluding contact × influence), -1735.24 (including housing × contact), and -1728.32 (including housing × influence), with corresponding deviance reductions of 0.21, 8.67 and 22.51, at the expense of 2, 3 and 6 degrees of freedom, respectively. Clearly the only interaction of note is that of housing × influence, which has a P-value of 0.001. Adding this term gives a model deviance of 25.22 on 34 d.f. and an excellent fit to the data.

Table 6.5 shows parameter estimates for the final model with all three predictors and a housing × influence interaction. The table lists the cutpoints and the regression coefficients.

TABLE 6.5: Parameter Estimates for Ordered Logit Model (Latent Variable Formulation)

Parameter	Estimate	Std. Error	z-ratio
Apartments	-1.1885	.1972	-6.026
Atrium house	-.6067	.2446	-2.481
Terraced house	-1.6062	.2410	-6.665
Influence medium	-.1390	.2125	-0.654
Influence high	.8689	.2743	3.167
Contact high	.3721	.0960	3.876
Apart × Influence med	1.0809	.2658	4.066
Apart × Influence high	.7198	.3287	2.190
Atrium × Influence med	.6511	.3450	1.887
Atrium × Influence high	-.1556	.4105	-0.379
Terrace × Influence med	.8210	.3307	2.483
Terrace × Influence high	.8446	.4303	1.963
Cutpoint 1	-.8881	.1672	
Cutpoint 2	.3126	.1657	

Note first the cutpoints: -.89 and .31, corresponding to cumulative odds of 0.41 and 1.37, or to cumulative probabilities of 0.29 and 0.58, for the reference cell. Considering residents of tower blocks with low influence in management and low contact with neighbors, we estimate that 29% have low satisfaction, 29% (58-29) have medium satisfaction, and 42% (100-58) have high satisfaction. (These values are fairly close to the observed proportions.)

Before we discuss interpretation of the remaining coefficients, we must note that I have reported the coefficients corresponding to the latent variable formulation (the $\boldsymbol{\beta}^*$'s) rather than the cumulative link coefficients (the $\boldsymbol{\beta}$'s), which have opposite sign. Thus, a positive coefficient is equivalent to a shift

to the right on the latent scale, which increases the odds of being to the *right* of a cutpoint. Equation (6.24) models the odds of being to the *left* of a cutpoint, which would then decrease. I prefer the sign used here because the interpretation is more straightforward. A positive coefficient increases one's underlying satisfaction, which makes a 'high' response more likely.

The coefficient of contact indicates that residents who have high contact with their neighbors are generally more satisfied than those who have low contact. The odds of high satisfaction (as opposed to medium or low), are 45% higher for high contact than for low contact, as are the odds of medium or high satisfaction (as opposed to low). The fact that the effect of contact on the odds is the same 45% for the two comparisons is a feature of the model.

To interpret the effects of the next two variables, type of housing and degree of influence, we have to allow for their interaction effect. One way to do this is to consider the effect of type of housing when the residents feel that they have low influence on management; then residents of apartments and houses (particularly terraced houses) are *less* satisfied than residents of tower blocks. Feeling that one has some influence on management generally increases satisfaction; the effect of having high rather than low influence is to increase the odds of medium or high satisfaction (as opposed to low) by 138% for residents of tower blocks, 390% for apartment dwellers, 104% for residents of atrium houses and 455% for those who live in terraced houses. Having medium influence is generally better than having low influence (except for tower clock residents), but not quite as good as having high influence (except possibly for residents of atrium houses).

Although we have interpreted the results in terms of odds, we can also interpret the coefficients in terms of a latent variable representing degree of satisfaction. The effect of having high contact with the neighbors, as compared to low contact, is to shift one's position on the latent satisfaction scale by 0.37 points. Similarly, having high influence on management, as compared to low influence, shifts one's position by an amount that varies from 0.71 for residents of atrium houses to 1.71 for residents of terraced houses. Interpretation of these numbers must be done by reference to the standard logistic distribution, which is depicted in Figure 6.3. This symmetric distribution has mean 0 and standard deviation $\pi/\sqrt{3} = 1.81$. The quartiles are ± 1.1, and just over 90% of the area lies between -3 and 3.

6.5.4 The Ordered Probit Model

The ordered probit model, first considered by Aitchison and Silvey (1957), results from modeling the *probit* of the cumulative probabilities as a linear function of the covariates, so that

$$\Phi^{-1}(\gamma_{ij}) = \theta_j + \mathbf{x}_i'\boldsymbol{\beta} \tag{6.27}$$

where $\Phi()$ is the standard normal cdf. The model can also be obtained from the latent-variable formulation assuming that the error term has a standard normal distribution, and this is usually the way one would interpret the parameters.

Estimates from the ordered probit model are usually very similar to estimates from the ordered logit model—as one would expect from the similarity of the normal and the logistic distributions—provided one remembers to standardize the coefficients to correct for the fact that the standard normal distribution has variance one, whereas the standard logistic has variance $\pi^2/3$.

For the Copenhagen data, the ordered probit model with an interaction between housing type and influence has a log-likelihood of -1728.67, corresponding to a deviance of 25.9 on 34 d.f., almost indistinguishable from the deviance for the ordered logit model with the same terms. Table 6.6 shows parameter estimates for this model.

The cutpoints can be interpreted in terms of z-scores: the boundary between low and medium satisfaction is at $z = -0.54$ and the boundary between medium and high satisfaction is at $z = 0.19$. These values leave $\Phi(-.54) = 0.29$ or 29% of the reference group in the low satisfaction category, $\Phi(0.19) - \Phi(-0.54) = 0.28$ or 28% in the medium satisfaction category, and $1 - \Phi(0.19) = 0.42$ or 42% in the high satisfaction category.

The remaining coefficients can be interpreted as in a linear regression model. For example, having high contact with the neighbors, compared to low contact, increases one's position in the latent satisfaction scale by 0.23 standard deviations (or increases one's z-score by 0.23), everything else being equal.

Note that this coefficient is very similar to the equivalent value obtained in the ordered logit model. A shift of 0.37 in a standard logistic distribution, where $\sigma = \pi/\sqrt{3} = 1.81$, is equivalent to a shift of $0.37/1.81 = 0.21$ standard deviations, which in turn is very similar to the ordered probit estimate of 0.23 standard deviations. A similar comment applies to the other coefficients. You may also wish to compare the Wald tests for the individual coefficients in Tables 6.5 and 6.6, which are practically identical.

TABLE 6.6: Parameter Estimates for Ordered Probit Model (Latent Variable Formulation)

Parameter	Estimate	Std. Error	z-ratio
Apartments	-.7281	.1205	-6.042
Atrium house	-.3721	.1510	-2.464
Terraced house	-.9790	.1456	-6.725
Influence medium	-.0864	.1303	-0.663
Influence high	.5165	.1639	3.150
Contact high	.2285	.0583	3.918
Apart × Influence med	.6600	.1626	4.060
Apart × Influence high	.4479	.1971	2.273
Atrium × Influence med	.4109	.2134	1.925
Atrium × Influence high	-.0780	.2496	-0.312
Terrace × Influence med	.4964	.2016	2.462
Terrace × Influence high	.5217	.2587	2.016
Cutpoint 1	-.5440	.1023	
Cutpoint 2	.1892	.1018	

6.5.5 Proportional Hazards

A third possible choice of link is the complementary log-log link, which leads to the model

$$\log(-\log(1-\gamma_{ij})) = \theta_j + \mathbf{x}_i'\boldsymbol{\beta} \tag{6.28}$$

This model may be interpreted in terms of a latent variable having a (reversed) extreme value (log Weibull) distribution, with cdf

$$F(\eta) = 1 - \exp\{-\exp\{\eta\}\} \tag{6.29}$$

This distribution is asymmetric, it has mean equal to negative Euler's constant -0.57722 and variance $\pi^2/6 = 1.6449$. The median is $\log\log 2 = -0.3665$ and the quartiles are -1.2459 and 0.3266. Note that the inverse cdf is indeed, the complementary log-log transformation in Equation (6.28).

This model can also be interpreted in terms of a proportional hazards model. The hazard function plays a central role in survival analysis, and will be discussed in detail in the next Chapter.

6.5.6 Extensions and Other Approaches

The general cumulative link model of Section 6.5.2 will work with any monotone link function mapping probabilities to the real line, but the three choices mentioned here, namely the logit, probit, and complementary log-log, are by far the most popular ones. McCullagh (1980) has extended the basic model by relaxing the assumption of constant variance for the latent continuous variable. His most general model allows a separate scale parameter for each multinomial observation, so that

$$g(\gamma_{ij}) = \frac{\theta_j + \mathbf{x}_i'\boldsymbol{\beta}}{\tau_i} \tag{6.30}$$

where the τ_i are unknown scale parameters. A constraint, such as $\tau_1 = 0$, is required for identification. More generally, τ_i may be allowed to depend on a vector of covariates.

An alternative approach to the analysis of ordinal data is to assign scores to the response categories and then use linear regression to model the mean score. Ordinary least squares procedures are not appropriate in this case, but Grizzle et al. (1969) have proposed weighted least-squares procedures that make proper allowances for the underlying independent multinomial sampling scheme. For an excellent discussion of these models see Agresti (1990, Section 9.6).

A similar approach, used often in two-way contingency tables corresponding to one predictor and one response, is to assign scores to the rows and columns of the table and to use these scores to model the interaction term in the usual log-linear model. Often the scores assigned to the columns are the integers $1, 2, \ldots, J-1$, but other choices are possible. If integer scores are used for both rows and columns, then the resulting model has an interesting property, which has been referred to as *uniform association*. Consider calculating an odds ratio for adjacent rows i and $i+1$, across *adjacent* columns or response categories j and $j+1$, that is

$$\rho_{ij} = \frac{\pi_{i,j}/\pi_{i,j+1}}{\pi_{i+1,j}/\pi_{i+1,j+1}} \tag{6.31}$$

Under the additive log-linear model of independence, this ratio is unity for all i and j. Introducing an interaction term based on integer scores, of the form $(\alpha\beta)_{ij} = \gamma ij$, makes the odds ratio constant across adjacent categories. This model often produces fits similar to the proportional odds model, but the parameters are not so easily interpreted. For further details see Haberman (1974), Goodman (1979) or Agresti (1990, Section 8.1).

A fourth family of models for ordinal responses follows from introducing constraints in the multinomial logit model. Let $\boldsymbol{\beta}_j$ denote the vector of coefficients for the j-th equation, comparing the j-th category with the last category, for $j = 1, 2, \ldots, J-1$. The most restrictive model assumes that these coefficients are the same for all contrasts, so $\boldsymbol{\beta}_j = \boldsymbol{\beta}$ for all j. A less restrictive assumption is that the coefficients have a linear trend over the categories, so that $\boldsymbol{\beta}_j = j\boldsymbol{\beta}$. Anderson (1984) has proposed a model termed the *stereotype* model where the coefficients are proportional across categories, so $\boldsymbol{\beta}_j = \gamma_j \boldsymbol{\beta}$, with unknown proportionality factors given by scalars γ_j.

One advantage of the cumulative link models considered here is that the parameter estimates refer to the cumulative distribution of the manifest response (or the distribution of the underlying latent variable) and therefore are not heavily dependent on the actual categories (or cutpoints) used. In particular, we would not expect the results to change much if we were to combine two adjacent categories, or if we recoded the response using fewer categories. If the cumulative odds are indeed proportional before collapsing categories, the argument goes, they should continue to be so afterwards.

In contrast, inferences based on log-linear or multinomial logit models apply only to the actual categories used. It is quite possible, for example, that odds ratios that are relatively constant across adjacent categories will no longer exhibit this property if some of the categories are combined. These considerations are particularly relevant if the categories used are somewhat arbitrary.