CHAPTER 5

Models for polytomous data

5.1 Introduction

If the response of an individual or item in a study is restricted to one of a fixed set of possible values, we say that the response is polytomous. The k possible values of Y are called the *response categories*. Often the categories are defined in a qualitative or non-numerical way. A familiar example is the classification of blood types, with unambiguous but qualitative categories O, A, B, AB. Another example is the ILO scale, 0/0, 0/1,...,3/3, used for classifying chest X-ray images according to apparent severity of lung disease. These categories, defined rather arbitrarily using 'standard' reproductions, are not devoid of ambiguity. Other instances of the type of response considered in this chapter are rating scales used in food testing, measures of mental and physical well-being, and many variables arising in social science research which are, of necessity, not capable of precise measurement.

We need to develop satisfactory statistical models that distinguish several types of polytomous response or measurement scale. For instance, if the categories are ordered, there is no compelling reason for treating the extreme categories in the same way as the intermediate ones. However, if the categories are simply an unstructured collection of labels, there is no reason a priori to select a subset of the categories for special treatment. Considerations such as these lead us to consider qualitatively different classes of link functions for different types of response scale. Whatever the nature of the scale, we may talk without ambiguity about the response probabilities $\pi_1, \ldots, \pi_k$. If the categories are ordered, however, we may prefer to work with the cumulative response probabilities

$$\gamma_1 = \pi_1, \quad \gamma_2 = \pi_1 + \pi_2, \ldots, \gamma_k \equiv 1.$$

149

Obviously, cumulative probabilities are not defined unless the category order is unambiguous. It makes little sense to work with a model specified in terms of γ_j if the response categories are not ordered.

5.2   Measurement scales

5.2.1   *General points*

Measurement scales can be classified at a number of levels. At one level, we may distinguish between *pure scales* and *compound scales*. Bivariate responses are perhaps the simplest instances of compound measurement scales. One can contemplate bivariate responses in which one response is ordinal and the other binary or even continuous. Other examples of compound scales are discussed in section 5.2.5. Among the spectrum of pure measurement scales, we may identify the following major types:

1. *nominal scales* in which the categories are regarded as exchangeable and totally devoid of structure.
2. *ordinal scales* in which the categories are ordered much like the ordinal numbers, 'first', 'second',.... In this context it does not ordinarily make sense to talk of 'distance' or 'spacing' between 'first' and 'second' nor to compare 'spacings' between pairs of response categories.
3. *interval scales* in which the categories are ordered and numerical labels or scores are attached. The scores are treated as category averages, medians or mid-points. Differences between scores are therefore interpreted as a measure of separation of the categories.

Cardinal scales require quite different kinds of models, such as those discussed in Chapters 3,6 and 8, and are not considered here. Binary measurements are special cases of all of the above in which $k = 2$. The distinction between ordinal, interval and nominal does not then arise.

In applications, the distinction between nominal and ordinal scales is usually but not always clear. For instance responses relating to perception of food quality — excellent, good, ..., bad, appalling — are clearly ordinal. Responses concerning preferences for newspaper or television programme would usually be treated

as nominal, at least initially. Political hue and perceived quality may well be sufficient grounds for the subsequent examination of particular contrasts. Hair colour and eye colour can be ordered to a large extent on the grey-scale from light to dark and are therefore ordinal, although the relevance of the order may well depend on the context. Otherwise, unless there is a clear connection with the electromagnetic spectrum or a grey-scale, colours are best regarded as nominal.

5.2.2 *Models for ordinal scales*

We consider ordinal scales first, mainly because these occur more frequently in applications than the other types. In many of these applications such as food-testing, classification of radiographs, determination of physical or mental well-being and so on, the choice and definition of response categories is either arbitrary or subjective. It is essential, therefore, if we are to arrive at valid conclusions, that the nature of those conclusions should not be affected by the number or choice of response categories. As a consequence, if a new category is formed by combining adjacent categories of the old scale, the form of the conclusions should be unaffected. Of course, the amalgamation of response categories in this way will normally reduce the available information, change the estimate, the attained significance level and so on. The important point is that the same parameter is being measured however many categories are combined. This is an important non-mathematical point that is difficult to make mathematically rigorous: it is not simply a matter of retaining the same Greek letter after category combination.

Such considerations lead fairly directly to models based on the cumulative response probabilities $\gamma_j = \mathrm{pr}(Y \leq j)$ rather than the category probabilities π_j. The two sets of probabilities are equivalent, but simple models for the cumulative probabilities are likely to have better properties for ordinal response scales than equally simple models based on the category probabilities. In particular, linear models using the logistic scale, $\log\{\gamma_j/(1-\gamma_j)\}$, or the complementary log-log scale, $\log\{-\log(1-\gamma_j)\}$ are found to work well in practice (McCullagh, 1980).

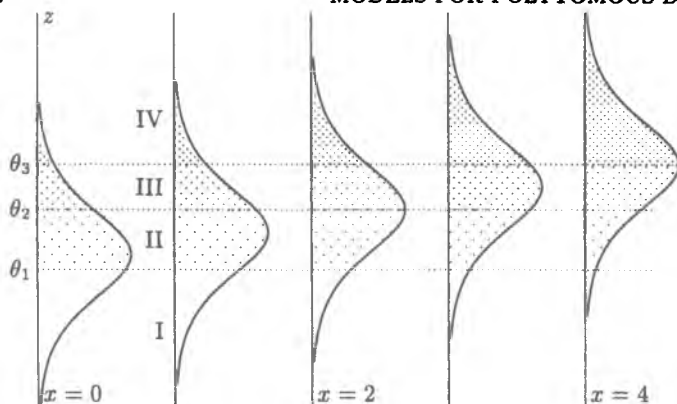The simplest models in this class involve parallel regressions on

Fig. 5.1a. *Diagram showing how the response probabilities for the logistic model (5.1) vary with x when $\beta > 0$. Response categories are represented as four contiguous intervals of the z-axis. Higher-numbered categories have greater shade density.*
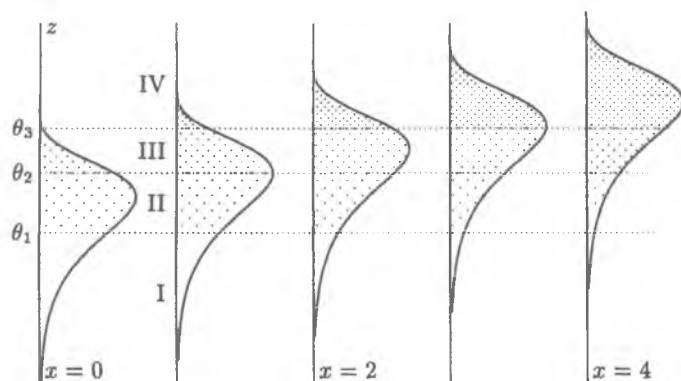

Fig. 5.1b. *Diagram showing how the probabilities for the four response categories in the complementary-log-log model (5.3) vary with x when $\beta > 0$. $\pi_1(x)$ and $\pi_4(x)$ each change by a factor of 10 or more, whereas $\pi_3(x)$ is almost constant over $1 \leq x \leq 4$.*

the chosen scale, such as

$$\log\{\gamma_j(\mathbf{x})/(1 - \gamma_j(\mathbf{x}))\} = \theta_j - \boldsymbol{\beta}^T\mathbf{x}, \qquad j = 1, \ldots, k-1 \qquad (5.1)$$

where $\gamma_j(\mathbf{x}) = \mathrm{pr}(Y \leq j \,|\, \mathbf{x})$ is the cumulative probability up to and including category j, when the covariate vector is $\mathbf{x}$. Model

(5.1) is known as the proportional-odds model because the ratio of the odds of the event $Y \leq j$ at $\mathbf{x} = \mathbf{x}_1$ and $\mathbf{x} = \mathbf{x}_2$ is

$$\frac{\gamma_j(\mathbf{x}_1)/(1 - \gamma_j(\mathbf{x}_1))}{\gamma_j(\mathbf{x}_2)/(1 - \gamma_j(\mathbf{x}_2))} = \exp\{-\boldsymbol{\beta}^T(\mathbf{x}_1 - \mathbf{x}_2)\}, \qquad (5.2)$$

which is independent of the choice of category (j). In particular, if $\mathbf{x}$ is an indicator variable for two treatment groups, T_1 and T_2, (5.2) may be written as

$$\frac{\text{odds}(Y \leq j \,|\, T_1)}{\text{odds}(Y \leq j \,|\, T_2)} = \exp(-\Delta), \qquad j = 1, \ldots, k - 1,$$

where Δ measures the treatment effect. The negative sign in (5.1) is a convention ensuring that large values of $\boldsymbol{\beta}^T\mathbf{x}$ lead to an increase of probability in the higher-numbered categories. Both $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ in (5.1) are treated as unknown and $\boldsymbol{\theta}$ must satisfy $\theta_1 \leq \theta_2 \leq \ldots \leq \theta_{k-1}$.

For the complementary log-log link, the model corresponding to (5.1) is

$$\log[-\log\{1 - \gamma_j(\mathbf{x})\}] = \theta_j - \boldsymbol{\beta}^T\mathbf{x}, \qquad j = 1, \ldots, k - 1 \qquad (5.3)$$

which is known as the proportional-hazards model (Cox, 1972a; McCullagh, 1980). In all of these models we must have $\theta_1 \leq \theta_2 \leq \ldots \leq \theta_{k-1}$ to ensure that the probabilities are non-negative.

Both (5.1) and (5.3) correspond to the same model formula and the same response variable, but with different choice of link function. The response is the set of cumulative observed proportions (or totals). Apart from choice of sign, the model formula in both cases is

$$R + \mathbf{x}$$

where R is the response factor having $k - 1$ levels and $\mathbf{x}$ is itself a model formula, not involving R, for the covariates used. The observations in this instance are not independent, but that is an aspect of the random part of the model and is considered in section 5.4.

Model (5.1) may be derived from the notion of a tolerance distribution or an underlying unobserved continuous random variable Z, such that $Z - \boldsymbol{\beta}^T\mathbf{x}$ has the standard logistic distribution. If the

unobserved variable lies in the interval $\theta_{j-1} < Z \leq \theta_j$ then $y = j$ is recorded. Thus we find

$$\text{pr}(Y \leq j) = \text{pr}(Z \leq \theta_j) = \text{pr}(Z - \boldsymbol{\beta}^T\mathbf{x} \leq \theta_j - \boldsymbol{\beta}^T\mathbf{x})$$
$$= \frac{\exp(\theta_j - \boldsymbol{\beta}^T\mathbf{x})}{1 + \exp(\theta_j - \boldsymbol{\beta}^T\mathbf{x})}.$$

Model (5.3) has a similar derivation based on the extreme-value distribution.

Figure 5.1 illustrates the way in which the response probabilities $\pi_j(x)$ vary with x for the single-variable case in which $\beta > 0$. In that case the larger the value of x the greater the probability of falling in the highest-numbered category. The probability for the lowest-numbered category decreases with x. For intermediate categories, the probability increases with x up to a certain point and thereafter decreases. Over certain ranges of x, the probability for some of the intermediate categories is almost constant: over the same range the probabilities for the extreme categories may change quite appreciably.

It is sometimes claimed that (5.1), (5.3) and related models are appropriate only if there exists a latent variable Z. This claim seems to be too strong and, in any case, the existence of Z is usually unverifiable in practice.

Suppose by way of extension that Z has the logistic distribution with mean $\boldsymbol{\beta}^T\mathbf{x}$ and scale parameter $\exp(\boldsymbol{\tau}^T\mathbf{x})$. In other words, $(Z - \boldsymbol{\beta}^T\mathbf{x})/\exp(\boldsymbol{\tau}^T\mathbf{x})$ has the standard logistic distribution. We are then led by the same argument to consider non-linear models of the particular form

$$\text{logit}\,\gamma_j(\mathbf{x}) = \frac{\theta_j - \boldsymbol{\beta}^T\mathbf{x}}{\exp(\boldsymbol{\tau}^T\mathbf{x})}. \tag{5.4}$$

This model is not of the generalized linear type, but nonetheless it is worthy of serious consideration. In the numerator, $\boldsymbol{\beta}^T\mathbf{x}$ plays the role of linear predictor for the mean and in the denominator $\boldsymbol{\tau}^T\mathbf{x}$ plays the role of linear predictor for the dispersion or variance. Two model formulae are required to specify (5.4) in its most general form. The numerator corresponds to the formula $R + \mathbf{x}$ as in (5.1) and (5.3). The denominator corresponds to an arbitrary model formula not involving R, which may differ from the formula in the numerator.

If, as in (5.2), $\mathbf{x}$ is an indicator variable for treatment, then we have

$$\frac{\text{odds}(Y \le j \mid T_1)}{\text{odds}(Y \le j \mid T_2)} = \exp\left(\frac{\theta_j - \beta_1}{\sigma_1} - \frac{\theta_j - \beta_2}{\sigma_2}\right)$$

$$= \exp\left(\frac{\beta_2}{\sigma_2} - \frac{\beta_1}{\sigma_1}\right) \times \exp\left\{\theta_j\left(\frac{1}{\sigma_1} - \frac{1}{\sigma_2}\right)\right\},$$

where $\sigma_i = \exp(\tau x_i)$ is the scale parameter for the ith treatment group. Thus the odds ratio is increasing in j if $\sigma_1 < \sigma_2$ and decreasing otherwise. Model (5.4) is useful for testing the proportional-odds assumption against the alternative that the odds ratio is systematically increasing or systematically decreasing in j.

Other link functions can be used in (5.4) in place of the logistic function.

Models in which the $k - 1$ regression lines are not parallel can be specified by writing

$$\theta_j + \boldsymbol{\beta}_j^T \mathbf{x}$$

in place of the right side of (5.1) and (5.3). The corresponding model formula is $R + R.\mathbf{x}$, meaning that the slopes vary, though not necessarily in any systematic way, with the levels of R. The usefulness of non-parallel regression models is limited to some extent by the fact that the lines eventually must intersect. Negative fitted values are then unavoidable for some values of $\mathbf{x}$, though perhaps not in the observed range. If such intersections occur in a sufficiently remote region of the $\mathbf{x}$-space, this flaw in the model need not be serious.

5.2.3 *Models for interval scales*

We now turn our attention to measurement scales of a slightly different type where the categories are ordered, but in a stronger or more rigid sense than that discussed in the previous section. Interval scales are distinguished by the following properties:

1. The categories are of interest in themselves and are not chosen arbitrarily.
2. It does not normally make sense to form a new category by amalgamating adjacent categories.
3. Attached to the jth category is a cardinal number or score, s_j, such that the difference between scores is a measure of distance between or separation of categories.

Property 2 is essential because if we were to combine two categories, we would need an algorithm for calculating the score for the new category. Further, the derived model with the new scores should be consistent with the old model in much the same way that the proportional-odds model (5.1) behaves consistently when categories are combined. In particular, the scores for the remaining categories should be unaffected. These properties are difficult to achieve.

Genuine interval scales having these three properties are rare in practice because, although properties 1 and 2 may be satisfied, it is rare to find a response scale having well-determined cardinal scores attached to the categories. Grouped continuous measurements, on the other hand, may satisfy property 3, but usually not 1 or 2. Nevertheless, it may occasionally be helpful to use artificial scores — usually the first k integers — and to treat these as cardinal rather than ordinal.

At this stage, we have three options for model construction. The first is to work with the cumulative response probabilities and, if necessary, to make suitable adaptations of the proportional-odds and related models. For instance, in (5.1) we might consider modelling the cut-points θ_j as functions of the scores. To begin, we might consider expressing θ_j as

$$\theta_j = \zeta_0 + \zeta_1 \left(\frac{s_j + s_{j+1}}{2} \right)$$

for unknown coefficients ζ_0 and $\zeta_1 > 0$. On balance, this seems unhelpful because the 'cut-points' θ_j are ordinarily considered to be incidental parameters of little interest in themselves. More interesting is the possibility of modelling departures from the proportional-odds assumption by allowing certain systematic deviations from parallelism. The most obvious way to achieve this is to replace $\boldsymbol{\beta}^T \mathbf{x}$ in (5.1) by

$$\boldsymbol{\beta}^T \mathbf{x} + \boldsymbol{\zeta}^T \mathbf{x}(c_j - \bar{c}) \qquad\qquad (5.5)$$

where c_j is a suitable function of the scores. Two possibilities are

$$c_j = \frac{s_j + s_{j+1}}{2} \quad \text{and} \quad c_j = \text{logit}\left(\frac{s_j + s_{j+1}}{2s_\bullet} \right).$$

Different choices may be more suitable for other link functions. There is a certain qualitative similarity between (5.5) and the effect achieved in (5.4) without using scores.

The second option is to examine the matrix of probabilities $\{\pi_j(\mathbf{x}_i)\}$ or the matrix of log probabilities

$$\eta_j(\mathbf{x}_i) = \log \pi_j(\mathbf{x}_i), \qquad j = 1, \ldots, k; \ i = 1, \ldots, n$$

and to decompose $\eta_j(\mathbf{x}_i)$ into a small number of effects or contrasts in much the same way as is done in regression and analysis-of-variance problems. In going from the log probabilities to the probabilities it must be borne in mind that $\sum_j \pi_j(\mathbf{x}_i) = 1$ for each i. The inverse transformation is best written in the form

$$\pi_j(\mathbf{x}_i) = \frac{\exp\{\eta_j(\mathbf{x}_i)\}}{\sum_j \exp\{\eta_j(\mathbf{x}_i)\}}.$$

As a consequence, $\{\eta_j(\mathbf{x}_i)\}$ and $\{\eta_j(\mathbf{x}_i) + \alpha_i\}$ represent the same set of probabilities and fitted values.

The simplest model, that of 'independence' or 'no covariate effect' may be written as $\pi_j(\mathbf{x}_i) = \pi_j$ or $\eta_j(\mathbf{x}_i) = \eta_j$ or

$$\eta_j(\mathbf{x}_i) = \eta_j + \alpha_i. \tag{5.6}$$

In purely formal terms, (5.6) is equivalent to the model formula

$$\text{column} + \text{row}$$

for the log probabilities, where 'column' is a k-level response factor indexed by j, and 'row' is a factor with n levels indexed by i.

In order to model departures from (5.6), we may suppose that the effect of the covariate is to increase the probability or log probability in those categories for which the scores are highest. Perhaps the simplest model that achieves this effect is

$$\eta_j(\mathbf{x}_i) = \eta_j + (\boldsymbol{\beta}^T \mathbf{x}_i) s_j + \alpha_i. \tag{5.7}$$

The corresponding model formula is

$$\text{column} + \text{score}.\mathbf{x} + \text{row}, \tag{5.8}$$

where score.$\mathbf{x}$ represents p covariates whose i, j components are $\mathbf{x}_i s_j$. In most applications, $\mathbf{x}$ is itself a model formula not involving the response factor 'column'.

One interpretation of (5.7) is that a unit change in $\boldsymbol{\beta}^T\mathbf{x}$ changes the log probabilities from η_j to $\eta_j + s_j$. Consequently, the relative odds for category j over category j' are changed from

$$\frac{\pi_j}{\pi_{j'}} = \exp(\eta_j - \eta_{j'}) \quad \text{to} \quad \exp(\eta_j - \eta_{j'} + s_j - s_{j'}).$$

In other words, the relative odds are increased multiplicatively by the factor

$$\exp(s_j - s_{j'})$$

per unit increase in the combination $\boldsymbol{\beta}^T\mathbf{x}$.

For a two-way table with one response variable and one explanatory factor (5.7) reduces to

$$\eta_{ij} = \eta_j + \alpha_i + \zeta_i s_j.$$

If, further, the explanatory factor has ordered levels, we may select for special consideration the linear contrast $r_i = i - (n+1)/2$, or other suitable contrast. The reduced model is then the same as above with ζ_i replaced by βr_i, namely

$$\eta_{ij} = \eta_j + \alpha_i + \beta r_i s_j.$$

This is also known as the linear×linear-interaction model, first proposed by Birch (1963).

The third option for model building and significance testing is to reverse the roles of the vector of scores $s = (s_1, \ldots, s_k)$ and the vector of counts $y = (y_1, \ldots, y_k)$. Instead of regarding y as the response and s as a contrast of special interest, we may regard the observed score as the response and y as the set of observed multiplicities or weights. Thus, if $k = 4$, $y = (5, 7, 10, 3)$ is equivalent to 25 observations on S, namely $S = s_1$ five times, $S = s_2$ seven times, $S = s_3$ ten times and $S = s_4$ three times. On the assumption that the mean observed score is linearly related to the covariates, we have

$$E(S \mid \mathbf{x}_i) = \sum_j \pi_j(\mathbf{x}_i)s_j = \boldsymbol{\beta}^T\mathbf{x}_i.$$

This is an incompletely specified model because the parameters $\boldsymbol{\beta}$ determine only the linear combination $\sum_j \pi_j s_j$ and not the individual cell probabilities themselves. Note also the unsatisfactory

property that $E(S \,|\, \mathbf{x}_i)$ must lie between s_1 and s_k whereas $\boldsymbol{\beta}^T \mathbf{x}_i$ is not similarly restricted. Despite these drawbacks, useful and interesting conclusions can frequently be drawn from an analysis of the observed mean scores

$$\bar{S}_i = \sum_j s_j y_{ij} / m_i.$$

In particular, if there are only two treatment groups, with observed counts $\{y_{1j}, y_{2j}\}$, we may use as test statistic the standardized difference

$$T = \frac{\bar{S}_1 - \bar{S}_2}{\sqrt{\left(\sum \tilde{\pi}_j s_j^2 - \left(\sum \tilde{\pi}_j s_j \right)^2 \right) \left(\frac{1}{m_1} + \frac{1}{m_2} \right)}}$$

where $\tilde{\pi}_j = y_{.j} / m_{.}$. Under the null hypothesis of no treatment effect, and provided that the observations have independent multinomial distributions, T is approximately standard Normal. This statistic is due to Yates (1948) and Armitage (1955).

5.2.4 *Models for nominal scales*

If the scale is purely nominal, we are forced to work with the category probabilities, π_j, directly. By the same argument used in the previous section, it is more convenient to work with logarithmic probabilities η_j given by

$$\pi_j = \exp(\eta_j) \Big/ \sum_j \exp(\eta_j), \quad \text{for} \quad j = 1, \ldots, k.$$

The aim is to describe how the vector $(\eta_1, \ldots, \eta_k)$ is affected by changes in the covariates. In doing so, we must bear in mind that $\boldsymbol{\eta}$ and $\boldsymbol{\eta} + c$ represent the same probabilities and fitted values.

In the absence of scores, the most general log-linear model has the form

$$\eta_j(\mathbf{x}_i) = \eta_j(\mathbf{x}_0) + \boldsymbol{\beta}_j^T(\mathbf{x}_i - \mathbf{x}_0) + \alpha_i \tag{5.9}$$

for $j = 1, \ldots, k$. In this expression, $\eta_j(\mathbf{x}_0)$ is the set of base-line log probabilities and $\boldsymbol{\beta}_j$ is the change in the jth log probability per unit change in each of the components of $\mathbf{x}$. To be more precise,

the odds in favour of category j over category j' are increased by the factor

$$\frac{\pi_j(\mathbf{x})}{\pi_{j'}(\mathbf{x})} = \frac{\pi_j(\mathbf{x}_0)}{\pi_{j'}(\mathbf{x}_0)} \times \exp\{(\boldsymbol{\beta}_j - \boldsymbol{\beta}_{j'})^T(\mathbf{x} - \mathbf{x}_0)\}.$$

Thus contrasts among the vectors $\boldsymbol{\beta}_j$ are of interest rather than the vectors themselves.

Note that in (5.7), the use of response category scores enables us to model the change in the log probabilities for all k response cells using a single covariate vector $\boldsymbol{\beta}$. In the absence of scores, it is necessary in (5.9) to use k covariate vectors, $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k$. Since only contrasts among the $\boldsymbol{\beta}_j$ are estimable, we may set $\boldsymbol{\beta}_1 = 0$. The net result is that the nominal response model (5.9) contains many more parameters than (5.7) in order to achieve a similar effect.

The model formula for (5.9) is

$$\text{column} + \text{column}.\mathbf{x} + \text{row},$$

which is the same as (5.8) with the quantitative variable 'score' replaced by the response factor 'column'. As before, $\mathbf{x}$ may itself be a complicated model formula not involving 'column'.

5.2.5 *Nested or hierarchical response scales*

It is difficult to identify precisely the characteristics that distinguish a nested or hierarchical response scale from the types previously discussed. The following examples serve that purpose and show that nested classifications occur in a large number of diverse applications.

Example 1 : *A study of mortality due to radiation*. Suppose that, in a study of the effects of radiation, exposed and non-exposed individuals are classified at the end of the study period as dead or alive. Further information is available regarding the cause of death, at least to the extent that death can be attributed to a single cause. The nature of the study requires that deaths be classified as 'due to cancer' or 'due to other causes'. At a third stage, cancer deaths are sub-divided into 'leukaemia deaths' and 'deaths from other cancers'. The four mutually exclusive response categories are therefore

1. alive,
2. death from causes other than cancer,
3. death from cancers other than leukaemia,
4. death from leukaemia.

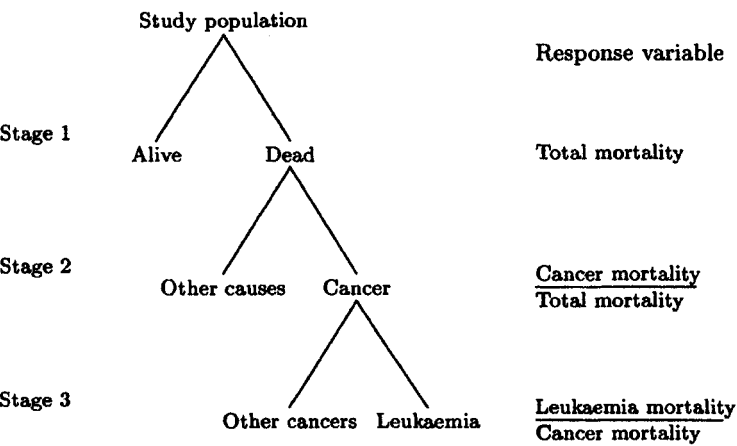Figure 5.2 emphasizes the nested structure of the responses.

Fig. 5.2. *Hierarchical classification used in the study of radiation effects.*

Here it is probably most appropriate to make a separate study of each of the three response variables, one corresponding to the dichotomy at each level of the hierarchy:

1. total mortality,
2. cancer mortality as a proportion of total mortality, and
3. leukaemia mortality as a percentage of cancer mortality.

Each of these variables may be affected by exposure, but perhaps in quite different ways. For example exposure might have a marked effect on the incidence of leukaemia (or thyroid cancer) without having much effect on total mortality or on the incidence of all cancers.

Example 2 : *Fertility of lactating cows.* There is some evidence to support the claim that a winter diet containing a high proportion of red clover has the effect of reducing the fertility of milch cows. In order to test this hypothesis, we begin with, say, 80 cows assigned at random to one of the two diets. Most cows become pregnant at the

first insemination but a few require a second or third insemination or occasionally more. After the first insemination, the most fertile cows have become pregnant. The success rate for those that require subsequent insemination is noticeably less than the initial success rate. In this instance there is an indefinite number of stages corresponding to first attempt, second attempt and so on. Three stages are depicted in Fig. 5.3. The variable measured at each stage is the pregnancy success rate. In that respect, this example differs from the previous one, where the variables measured at each stage were scientifically distinct. If indeed, red clover reduces fertility, this reduction should be apparent at all stages, even though the mean fertility of the remaining cows is reduced at each successive stage. Information concerning the treatment effect must be collected from the pregnancy rates observed at each stage.
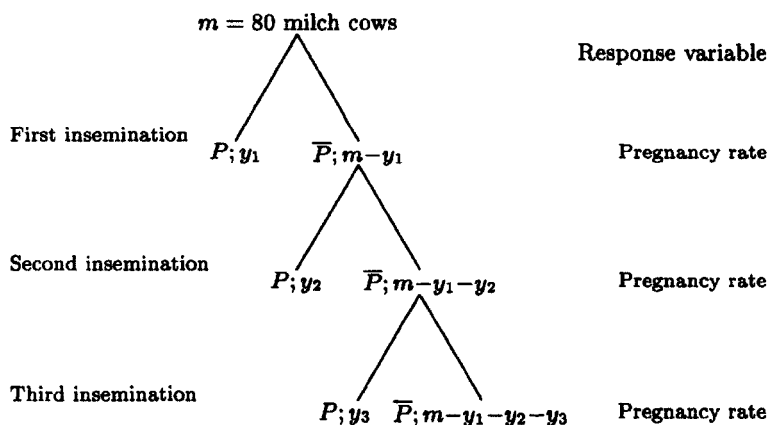

$m = 80$ milch cows

Response variable

First insemination $P; y_1$ $\overline{P}; m - y_1$ Pregnancy rate

Second insemination $P; y_2$ $\overline{P}; m - y_1 - y_2$ Pregnancy rate

Third insemination $P; y_3$ $\overline{P}; m - y_1 - y_2 - y_3$ Pregnancy rate

Fig. 5.3. *Hierarchical response in an insemination experiment:*
P = pregnant, $\overline{P}$ = not pregnant.

There is, of course, the possibility of a more complicated treatment effect whereby only the less fertile cows are affected. The observed treatment effect would then be expected to increase at successive stages. In the analysis, one should be aware that such an interaction might occur and what its symptoms would be.

In order to build up a model for either kind of hierarchical response, it is best to consider separately $k - 1$ responses, one

for each level of the hierarchy. The m subjects available at stage 1 respond positively with probability π_1, or negatively with probability $1 - \gamma_1$. The observed proportions are necessarily slightly different from the theoretical proportions so that, at stage 2, the experimental set or 'risk set' is reduced to $m - y_1$. Among these, the probability of a positive response is $\pi_2/(1 - \gamma_1)$, and the probability of a negative response is $(1 - \gamma_2)/(1 - \gamma_1)$. By the third stage, the risk set is further reduced to $m - y_1 - y_2$. Among these, the probability of a positive response is $\pi_3/(1 - \gamma_2)$, and the probability of a negative response is $(1 - \gamma_3)/(1 - \gamma_2)$. The response is thus broken down into the sequence of conditional factors:

Stage	Response	Probability	Odds
1	$Y_1 \mid m$	π_1	$\pi_1/(1 - \gamma_1)$
2	$Y_2 \mid m - y_1$	$\pi_2/(1 - \gamma_1)$	$\pi_2/(1 - \gamma_2)$
3	$Y_3 \mid m - y_1 - y_2$	$\pi_3/(1 - \gamma_2)$	$\pi_3/(1 - \gamma_3)$

In the particular examples considered, each stage of the hierarchy corresponds to a simple dichotomy. It is natural therefore, to consider binary regression models of the type discussed in Chapter 4. Thus, in the radiation mortality example,

$$g(\pi_1) = \beta_1^T \mathbf{x}$$

relates total mortality to exposure $\mathbf{x}$ via the link function $g(\cdot)$. By extension,

$$g\left(\frac{\pi_2}{1 - \gamma_1}\right) = \beta_2^T \mathbf{x}$$

relates cancer mortality as a proportion of total mortality to exposure. Similarly,

$$g\left(\frac{\pi_3}{1 - \gamma_2}\right) = \beta_3^T \mathbf{x}$$

relates leukaemia cancer mortality as a proportion of total cancer mortality to the exposure variables. There is no good reason here to expect that the coefficients $\beta_1, \beta_2, \beta_3$ might be equal or even comparable. In addition, there is no strong argument for using the same link function in the three regressions. If the identified cancer types at stage three were 'leukaemia', 'thyroid' and 'other', the trichotomy would be regarded as a nominal response scale and

the methods of section 5.2.4 could be used. It would then become impossible to insist on the same link function for each stage.

The insemination example has many of the same features but differs in the important respect that the response is the same at each stage. In constructing a model, however, we must make allowance for the expected decline in fertility at successive stages. A simple sequence of models having a constant treatment effect is as follows:

$$g(\pi_1) = \alpha_1 + \beta^T \mathbf{x},$$
$$g\{\pi_2/(1 - \gamma_1)\} = \alpha_2 + \beta^T \mathbf{x},$$
$$g\{\pi_3/(1 - \gamma_2)\} = \alpha_3 + \beta^T \mathbf{x}.$$

It is essential here to use same link function for each stage. In particular, if the logistic link function is used, we have

$$\log\left(\frac{\pi_j}{1 - \gamma_j}\right) = \alpha_j + \beta^T \mathbf{x}. \tag{5.10}$$

The incidental parameters $\alpha_1, \ldots, \alpha_{k-1}$ make allowance for the expected decline in fertility. If $\mathbf{x}$ is an indicator variable for treatment, model (5.10) asserts that treatment increases the odds of success by a factor $\exp(\beta^T \mathbf{x})$ uniformly at each stage of the experiment. Constancy of the effect can be tested in the usual way by the addition of an interaction term between treatment and stage.

5.3 The multinomial distribution

5.3.1 *Genesis*

The multinomial distribution is in many ways the most natural distribution to consider in the context of a polytomous response variable. It arises in a number of contexts, some apparently artificial, others a consequence of simple random sampling.

Suppose that individuals in some population of interest possess one and only one of the k attributes $A_1, \ldots, A_k$. The attributes might be 'colour of hair', 'socio-economic status', 'family size', 'cause of death' and so on depending on the context. If the population is effectively infinitely large and if a simple random sample of size m is taken, how many individuals will be observed

to have attribute A_j? The answer is given by the multinomial distribution

$$\mathrm{pr}(Y_1 = y_1, \ldots, Y_k = y_k; m, \boldsymbol{\pi}) = \binom{m}{\mathbf{y}} \pi_1^{y_1} \ldots \pi_k^{y_k}, \qquad (5.11)$$

where $\pi_1, \ldots, \pi_k$ are the attribute frequencies in the infinite population and

$$\binom{m}{\mathbf{y}} = \frac{m!}{y_1! \ldots y_k!}.$$

The multinomial distribution arises here simply as a consequence of the method of sampling. A different method of sampling such as cluster sampling or quota sampling would give rise to a different frequency distribution from (5.11).

The sample space or set of all possible values of the vector $\mathbf{y}$ is the set of all integer-valued k-vectors satisfying $0 \le y_j \le m$, $\sum y_j = m$ and comprises $\binom{m+k-1}{k-1}$ points. The sample space is a triangular lattice bounded by a regular simplex: see Fig. 5.4 for a diagram of the trinomial distribution.

Another derivation of the multinomial distribution is as follows. Suppose that $Y_1, \ldots, Y_k$ are independent Poisson random variables with means $\mu_1, \ldots, \mu_k$. Then the conditional joint distribution of $Y_1, \ldots, Y_k$ given that $Y. = m$ is given by (5.11) with $\pi_j = \mu_j/\mu.$.

The multinomial distribution for which $\pi_j = 1/k$ is called the uniform multinomial distribution.

5.3.2 *Moments and cumulants*

The moment generating function of the multinomial distribution, $M(m, \boldsymbol{\pi})$ is

$$M_Y(t) = E \exp\left(\sum t_j Y_j\right) = \left\{\sum \pi_j \exp(t_j)\right\}^m.$$

Thus the cumulant generating function is

$$K_Y(t) = m \log\left\{\sum \pi_j \exp(t_j)\right\}.$$

All cumulants have the form $m \times$ polynomial in $\boldsymbol{\pi}$. In particular, the first four joint cumulants are

$$E(Y_r) = m\pi_r$$

$$\mathrm{cov}(Y_r, Y_s) = \begin{cases} m\pi_r(1 - \pi_r) & r = s \\ -m\pi_r\pi_s & r \ne s \end{cases} \qquad (5.12)$$
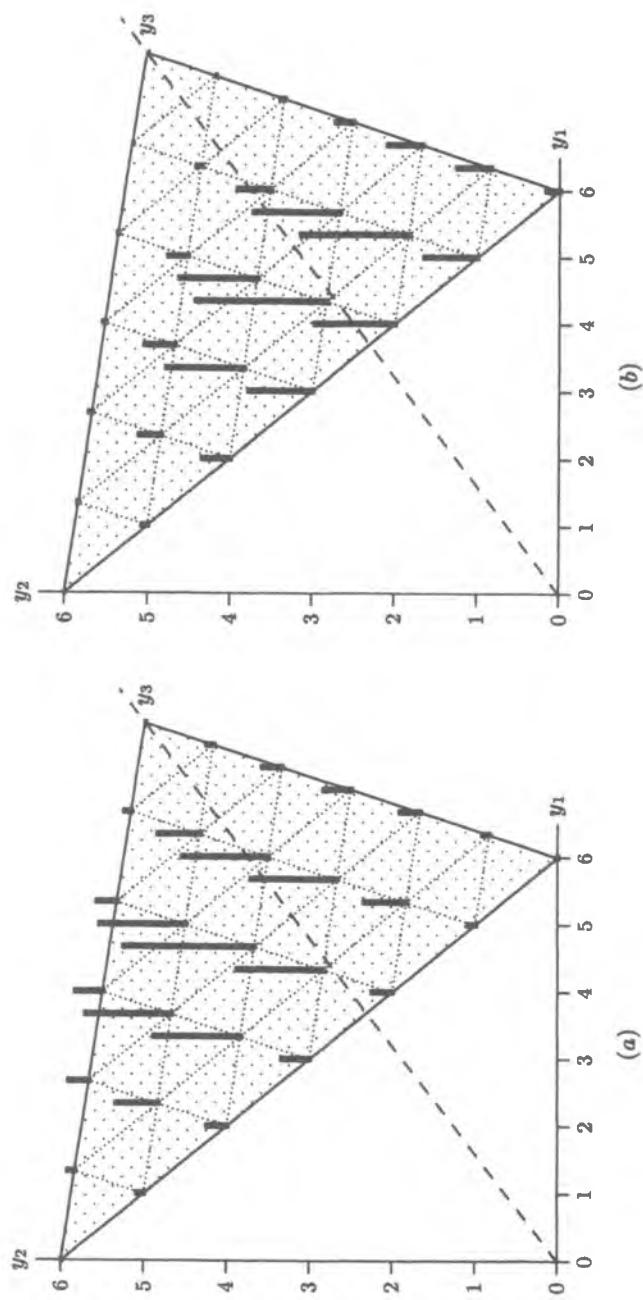
Fig. 5.4. *Trinomial sample space and probabilities for m = 6. The sample points form a triangular lattice on the equiangular plane $y_1 + y_2 + y_3 = 6$ (shaded) in R^3: the y_3-axis recedes into the plane of the page. In (a), $\pi = (1/3, 1/3, 1/3)$. In (b), $\pi = (0.5, 0.3, 0.2)$.*

$$\kappa_3(Y_r, Y_s, Y_t) = \begin{cases} m\pi_r(1-\pi_r)(1-2\pi_r) & r = s = t \\ -m\pi_r\pi_t(1-2\pi_r) & r = s \neq t \\ 2m\pi_r\pi_s\pi_t & r, s, t \text{ distinct} \end{cases}$$

$$\kappa_4(Y_r, Y_s, Y_t, Y_u) = \begin{cases} m\pi_r(1-\pi_r)(1-6\pi_r(1-\pi_r)) & r = s = t = u \\ -m\pi_r\pi_u(1-6\pi_r(1-\pi_r)) & r = s = t \neq u \\ -m\pi_r\pi_t(1-2\pi_r-2\pi_t+6\pi_r\pi_t) & r = s \neq t = u \\ 2m\pi_r\pi_s\pi_u(1-3\pi_r) & r = s \neq t \neq u \\ -6m\pi_r\pi_s\pi_t\pi_u & r, s, t, u \text{ distinct} \end{cases}$$

Frequently, however, it is more convenient to work with the vector of cumulative totals rather than with the cell counts. If we write $Z = LY$ where L is a lower-triangular matrix containing unit values, we see that the vector of cumulative totals is a linear function of Y. The first four cumulants of Z are

$$E(Z_r) = m\gamma_r,$$

$$\gamma_{rs} = \text{cov}(Z_r, Z_s) = m\gamma_r(1-\gamma_s) \quad \text{for} \quad r \leq s, \tag{5.13}$$

$$\kappa_3(Z_r, Z_s, Z_t) = m\gamma_r(1-2\gamma_s)(1-\gamma_t) \quad \text{for} \quad r \leq s \leq t,$$

$$\kappa_4(Z_r, Z_s, Z_t, Z_u) = m\gamma_r(1-\gamma_u)\{1 - 2(\gamma_t - \gamma_s) - 6\gamma_s(1-\gamma_t)\}$$
$$\text{for} \quad r \leq s \leq t \leq u.$$

In other respects as well, the cumulative multinomial vector has simpler properties than the original vector. For instance, it is easily seen that for $r < s < t$, Z_r and Z_t are conditionally independent given Z_s. To be specific, given $Z_s = z_s$,

$$Z_r \sim B(z_s, \gamma_r/\gamma_s)$$

$$Z_t - z_s \sim B\{m - z_s, (\gamma_t - \gamma_s)/(1 - \gamma_s)\}.$$

Linear combinations $\sum s_j Y_j$ with fixed coefficients s_j arise naturally in calculations related to models of the type discussed in section 5.2.3. The cumulants of such a combination are easily obtained either from the expressions given above or by observing that for $m = 1$, $\sum s_j Y_j$ takes the values $s_1, \ldots, s_k$ with probabilities $\pi_1, \ldots, \pi_k$. Consequently if we write

$$\mu_s = E\{\sum s_j Y_j/m\} = \sum \pi_j s_j$$

we have that

$$\text{var}(\sum s_j Y_j) = m\sum \pi_j(s_j - \mu_s)^2 = m\{\sum \pi_j s_j^2 - (\sum \pi_j s_j)^2\}$$

$$\kappa_3(\sum s_j Y_j) = m\sum \pi_j(s_j - \mu_s)^3.$$

Similar expressions may be derived for higher-order cumulants should these be required for Edgeworth approximation.

5.3.3 *Generalized inverse matrices*

Provided only that the cell probabilities π_j are positive, the multinomial covariance matrix $\Sigma_Y = m\{\text{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^T\}$ has rank $k-1$. The simplest generalized inverse is

$$\Sigma_Y^- = \text{diag}\{1/(m\pi_j)\},$$

which has rank k. This is not the Moore-Penrose inverse, but for most statistical calculations the choice of generalized inverse is immaterial and Σ^- given above is perhaps the simplest such inverse. It is easily verified that

$$\Sigma\Sigma^-\Sigma = \Sigma,$$

which is the defining property of a generalized inverse. In fact all generalized inverses have the form $\Sigma^- - c\mathbf{1}\mathbf{1}^T$ for some c. The Moore-Penrose inverse has $c = 1$.

The vector of cumulative totals, $\mathbf{Z}$, may be regarded either as a vector having k components, the last of which is fixed, or alternatively as a vector having $k-1$ components, the last component being ignored. In either case the covariance matrix has rank $k-1$. The covariance matrix $\Gamma = \{\gamma_{rs}\}$ in (5.13) is a particular instance of a Green's matrix, whose inverse is a symmetric Jacobi or tri-diagonal matrix. The particular form of inverse for $k = 5$ is as follows:

$$\Gamma^- = \frac{1}{m}\begin{pmatrix} \pi_1^{-1} + \pi_2^{-1} & -\pi_2^{-1} & 0 & 0 & 0 \\ -\pi_2^{-1} & \pi_2^{-1} + \pi_3^{-1} & -\pi_3^{-1} & 0 & 0 \\ 0 & -\pi_3^{-1} & \pi_3^{-1} + \pi_4^{-1} & -\pi_4^{-1} & 0 \\ 0 & 0 & -\pi_4^{-1} & \pi_4^{-1} + \pi_5^{-1} & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

This is the Moore-Penrose inverse of Γ in (5.13). All generalized inverses have this form, but with arbitrary values in the final row and column.

For a discussion of the geometry of generalized inverse matrices, see Kruskal (1975) or Stone (1987).

5.3.4 *Quadratic forms*

In order to test the simple null hypothesis $H_0 : \pi = \pi^{(0)}$, it is natural to construct, as a test statistic, a quadratic form in the residuals, $R_j = Y_j - m\pi_j^{(0)}$. Since $\sum R_j = 0$, it follows that

$$X^2 = \mathbf{R}^T \mathbf{\Sigma}_Y^- \mathbf{R}$$

is independent of the choice of generalized inverse. Taking the particular inverse given in the previous section, we see that

$$X^2 = \sum_j R_j^2/(m\pi_j^{(0)}) = \sum_j (Y_j - \mu_j^{(0)})^2/\mu_j^{(0)},$$

which is the familiar statistic due to Pearson (1900).

Equally well, if we choose to work with the cumulative multinomial vector and the corresponding generalized inverse, we obtain

$$\sum_1^{k-1} \frac{(Z_j - m\gamma_j)^2}{m} \left(\frac{1}{\pi_j} + \frac{1}{\pi_{j+1}} \right) - 2\sum_{j=1}^{k-2} \frac{(Z_j - m\gamma_j)(Z_{j+1} - m\gamma_{j+1})}{m\pi_{j+1}}$$

with π and γ computed under H_0. It is an elementary if rather tedious exercise to show that the above quadratic form is identical to X^2. Quadratic forms such as these are invariant under nonsingular linear transformation of the original variables.

The first three null cumulants of X^2 are

$$E(X^2) = k - 1,$$
$$\mathrm{var}(X^2) = 2(k-1)\frac{m-1}{m} + (S_{-1} - k^2)/m,$$
$$\kappa_3(X^2) = 8(k-1)\frac{m-1}{m} + 4(k-1)(k-6)\frac{m-1}{m^2}$$
$$+ (S_{-1} - k^2)\big(22(m-1) - 3k\big)/m^2 + (S_{-2} - k^3)/m^2,$$

where $S_r = \sum \pi_j^r$. In the uniform case, $S_{-1} = k^2$, $S_{-2} = k^3$.

If m is large, X^2 is approximately distributed as χ_{k-1}^2 with cumulants $k-1, 2(k-1), 8(k-1), \ldots$. The above exact calculations give a measure of the departure in finite samples of X^2 from its limiting distribution.

Similar moment calculations for X^2 for two-way tables are given in Exercise 6.16.

5.3.5 *Marginal and conditional distributions*

The marginal distribution of each multinomial component of Y is binomial: $Y_j \sim B(m, \pi_j)$. Also, the joint marginal distribution of $(Y_1, Y_2, m - Y_1 - Y_2)$ is multinomial on three categories with index m and parameter $(\pi_1, \pi_2, 1 - \pi_1 - \pi_2)$. This latter property extends to any number of components.

The conditional joint distribution of $Y_1, \ldots, Y_k$, given that $Y_i = y_i$, is multinomial on the reduced set of categories, with reduced index $m \rightarrow m - y_i$ and probabilities renormalized to

$$\pi_j \rightarrow \pi_j / (1 - \pi_i).$$

Analogous results are available for the cumulative multinomial vector, Z. The marginal distribution of Z_j is $B(m, \gamma_j)$. The conditional distribution of Z_i given $Z_j = z_j$, is $B(z_j, \gamma_i / \gamma_j)$ for $i < j$. Also, the conditional distribution of Y_{j+1} given $Z_j = z_j$, is $B(m - z_j, \pi_{j+1}/(1 - \gamma_j))$, which is the basis for the decomposition in section 5.2.5. In fact, the multinomial distribution can be expressed as a product of $k - 1$ binomial factors

$$\mathrm{pr}(Y = y) = p(y_1 \mid z_0)\, p(y_2 \mid z_1)\, p(y_3 \mid z_2) \ldots p(y_{k-1} \mid z_{k-2}),$$

where

$$p(y_j \mid z_{j-1}) = \left(\frac{\pi_j}{1 - \gamma_{j-1}}\right)^{y_j} \left(\frac{1 - \gamma_j}{1 - \gamma_{j-1}}\right)^{m - z_{j-1} - y_j} \binom{m - z_{j-1}}{y_j}$$

and $z_0 = \gamma_0 = 0$.

Evidently, the sequence $Z_1, \ldots, Z_k$ has the Markov property, namely that the conditional distribution of Z_j given the entire sequence $Z_1, \ldots, Z_{j-1}$ up to $j - 1$, depends only on the most recent value, namely Z_{j-1}. Also, the 'past' $Z_1, \ldots, Z_{j-1}$, and the 'future' $Z_{j+1}, \ldots, Z_k$ are conditionally independent given the 'present', Z_j.

5.4   Likelihood functions

5.4.1   *Log likelihood for multinomial responses*

We suppose that there are available n independent multinomial vectors, each with k categories. These observations are denoted by $\mathbf{y}_1, \ldots, \mathbf{y}_n$, where $\mathbf{y}_i = (y_{i1}, \ldots, y_{ik})$ and $\sum_j y_{ij} = m_i$ is fixed for each i. As usual, it is more convenient to consider the log likelihood initially as a function of the n probability vectors $\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_n$. Subsequently, when we contemplate a specific model such as (5.1) or (5.5), we may express the probabilities in terms of the parameters that appear in that model.

From the ith observation $\mathbf{y}_i$, the contribution to the log likelihood is

$$l(\boldsymbol{\pi}_i; \mathbf{y}_i) = \sum_j y_{ij} \log \pi_{ij}.$$

It is understood here that the observations and the probabilities are subject to the linear constraints

$$\sum_j y_{ij} = m_i \quad \text{and} \quad \sum_j \pi_{ij} = 1$$

for each i. Since the n observations are independent by assumption, the total log likelihood is a sum of contributions, one from each of the n observations. Thus,

$$l(\boldsymbol{\pi}; \mathbf{y}) = \sum_{ij} y_{ij} \log \pi_{ij}. \tag{5.14}$$

Differentiation of the log likelihood with respect to π_{ij} subject to the constraint $\sum_j \pi_{ij} = 1$ gives

$$\frac{\partial l(\boldsymbol{\pi}; \mathbf{y})}{\partial \pi_{ij}} = \frac{y_{ij} - m_i \pi_{ij}}{\pi_{ij}}.$$

Equivalently, introducing matrix notation,

$$\frac{\partial l(\boldsymbol{\pi}; \mathbf{y})}{\partial \boldsymbol{\pi}_i} = m_i \Sigma_i^{-} (\mathbf{y}_i - m_i \boldsymbol{\pi}_i)$$

$$= m_i \Sigma_i^{-} (\mathbf{y}_i - \boldsymbol{\mu}_i). \tag{5.15}$$

This set of n derivative vectors can be collected into a single matrix equation

$$\frac{\partial l(\boldsymbol{\pi}; \mathbf{y})}{\partial \boldsymbol{\pi}} = \mathbf{M}\boldsymbol{\Sigma}^{-}(\mathbf{y} - \boldsymbol{\mu}),$$

where $\boldsymbol{\Sigma} = \mathrm{diag}\{\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_n\}$ is $nk \times nk$ of rank $n(k-1)$ and $\mathbf{M}$ is a diagonal matrix of order $nk \times nk$ containing the multinomial indices m_i each repeated k times. The choice of generalized inverse in (5.15) is immaterial because $\mathbf{1}^T(\mathbf{y}_i - \boldsymbol{\mu}_i) = 0$ for each i.

In the above calculations, we have chosen to work with the response vectors $\mathbf{y}_i$ and the probability vectors $\boldsymbol{\pi}_i$. This turns out to be a convenient but arbitrary choice. We could choose to work instead with the cumulative response vectors $\mathbf{z}_i$ together with the cumulative probability vectors $\boldsymbol{\gamma}_i$. Analogous calculations then give

$$\frac{\partial l(\boldsymbol{\gamma}; \mathbf{y})}{\partial \boldsymbol{\gamma}_i} = m_i \boldsymbol{\Gamma}_i^{-}(\mathbf{z}_i - m_i \boldsymbol{\gamma}_i), \qquad (5.16)$$

which can be obtained from (5.15) using the chain rule. In fact,

$$\frac{\partial l}{\partial \gamma_{ij}} = \frac{\partial l}{\partial \pi_{ij}} - \frac{\partial l}{\partial \pi_{i\,j-1}} \quad \text{for} \quad 1 < j < k,$$

which is the same as (5.16).

5.4.2 Parameter estimation

The likelihood equations for the parameters are entirely straightforward to obtain at least in principle. We simply multiply (5.15) by the derivative of π_{ij} with respect to each parameter in turn and sum over i and j. Alternatively, and equivalently, we multiply (5.16) by the derivative of γ_{ij} with respect to the parameters and sum over i and j. Obviously, the form of the resulting equations depends heavily on the particular choice of model. We now consider some of the details of two particular choices.

Suppose that the model chosen has the form

$$\mathrm{logit}\,\gamma_{ij} = \sum_r x_{ijr}^* \beta_r^*$$

for some fixed coefficients x_{ijr}^* and unknown parameters β_r^*. It is helpful here to think of x_{ijr}^* as the components of a matrix $\mathbf{X}^*$

of order $nk \times p^*$ where p^* is the dimension of $\boldsymbol{\beta}^*$. In the case of model (5.1), $\boldsymbol{\beta}^*$ has dimension $p^* = p + k - 1$ with components

$$\boldsymbol{\beta}^* = (\theta_1, \ldots, \theta_{k-1}, \beta_1, \ldots, \beta_p).$$

The (i, j) row of $\mathbf{X}^*$ has components $(0, .., 1, .., 0, \mathbf{x}_i)$, with the unit value in position j. Consequently, the ith block of $k - 1$ rows is

$$\left[\mathbf{I}_{k-1} : \mathbf{1}\mathbf{x}_i\right].$$

Differentiation with respect to $\boldsymbol{\beta}^*$ gives

$$\frac{\partial l}{\partial \beta_r^*} = \sum_{ij} \frac{\partial l}{\partial \gamma_{ij}} \frac{\partial \gamma_{ij}}{\partial \beta_r^*}$$

$$= \sum_{ij} x_{ijr}^* \gamma_{ij}(1 - \gamma_{ij}) \frac{\partial l}{\partial \gamma_{ij}},$$

where $\partial l / \partial \gamma_{ij}$ is given by (5.16). In fact,

$$\frac{\partial l}{\partial \gamma_{ij}} = \frac{y_{ij} - m_i \pi_{ij}}{\pi_{ij}} - \frac{y_{ij-1} - m_i \pi_{ij-1}}{\pi_{ij-1}} \quad \text{for} \quad 1 < j < k.$$

For the proportional-odds model (5.1), these calculations can be simplified to some extent by exploiting the structure of the array x_{ijr}^* but these details will not be pursued here.

For log-linear models such as (5.5)–(5.9), we have

$$\log \pi_{ij} = \sum_r x_{ijr}^* \beta_r^*$$

for various choices of coefficients x_{ijr}^* dependent on the choice of model. In all of these cases, the likelihood equations take on a particularly simple form, namely

$$\sum_{ij} x_{ijr}^*(y_{ij} - \hat{\mu}_{ij}) = 0 \quad \text{for} \quad r = 1, \ldots, p^*.$$

In other words, in this case maximum likelihood is equivalent to the method of moments in which specific linear combinations $\sum x_{ijr}^* y_{ij}$ are equated to their expectations as a function of the parameters. The actual combinations depend on the choice of model. For instance if model (5.8) is used, the combinations are the 'row' and 'column' totals as well as the 'interaction combinations'

$$\sum_{ij} x_{ir} s_j y_{ij} \quad \text{for} \quad r = 1, \ldots, p.$$

5.4.3 Deviance function

The residual deviance function is twice the difference between the maximum achievable log likelihood and that attained under the fitted model. The maximum achievable log likelihood occurs at the point $\tilde{\pi}_{ij} = y_{ij}/m_i$. The deviance function is therefore

$$
\begin{aligned}
D(\mathbf{y}; \hat{\boldsymbol{\pi}}) &= 2l(\tilde{\boldsymbol{\pi}}; \mathbf{y}) - 2l(\hat{\boldsymbol{\pi}}; \mathbf{y}) \\
&= 2\sum y_{ij} \log \tilde{\pi}_{ij} - 2\sum y_{ij} \log \hat{\pi}_{ij} \\
&= 2\sum_{ij} y_{ij} \log(y_{ij}/\hat{\mu}_{ij}).
\end{aligned}
$$

Under the conditions described in section 4.4.3, namely that $\hat{\mu}_{ij}$ are sufficiently large and that there is no over-dispersion, $D(Y; \hat{\boldsymbol{\pi}})$ has an approximate χ^2 distribution. Its use as a goodness-of-fit statistic, however is open to the objections raised in sections 4.4.3, 4.4.5 and 4.5.

5.5 Over-dispersion

Over-dispersion for polytomous responses can occur in exactly the same way as over-dispersion for binary responses. Details are given in section 4.5.1 and will not be repeated here. Under the cluster-sampling model, the covariance matrix of the observed response vector is the sum of the within-cluster covariance matrix and the between-cluster covariance matrix. Provided that these two matrices are proportional, we have

$$
\begin{aligned}
E(Y) &= m\boldsymbol{\pi}, \\
\mathrm{cov}(Y) &= \sigma^2 \boldsymbol{\Sigma},
\end{aligned}
\tag{5.17}
$$

where $\boldsymbol{\Sigma}$ is the usual multinomial covariance matrix. The dispersion parameter σ^2 has the same interpretation given in section 4.5.

Parameter estimation (other than σ^2) is unaffected by over-dispersion and proceeds along the lines described in section 5.4.2 as if the multinomial distribution continued to apply. However, the covariance matrix of $\hat{\beta}$, obtained from the multinomial log likelihood, needs to be inflated by the dispersion factor σ^2. The

only additional step therefore is the estimation of this dispersion factor. For the reasons given in section 4.5.2, we use

$$\tilde{\sigma}^2 = X^2/\{n(k-1)-p\}$$
$$= X^2/\{\text{residual d.f.}\}, \tag{5.18}$$

where X^2 is Pearson's statistic. This estimate is approximately unbiased for σ^2, is consistent for large n regardless of whether the data are sparse and moreover is approximately independent of the estimated $\hat{\beta}$.

For further details see Chapter 9.

5.6 Examples

5.6.1 *A cheese-tasting experiment*

The following data, kindly provided by Dr Graeme Newell, were obtained from an experiment concerning the effect on taste of various cheese additives. The so-called hedonic scale has nine response categories, ranging from 'strong dislike' (1) to 'excellent taste' (9). In this instance, four additives labelled A, B, C and D were tested. The data are given in Table 5.1.

Here the effects are so great that the qualitative ordering (D, A, C, B) can easily be deduced from visual inspection. Nevertheless it is of some interest to check whether the models described earlier are capable of describing these differences and of evaluating the statistical significance of the differences observed.

Table 5.1 *Response frequencies in a cheese-tasting experiment*

Cheese	Response category									Total
	I*	II	III	IV	V	VI	VII	VIII	IX†	
A	0	0	1	7	8	8	19	8	1	52
B	6	9	12	11	7	6	1	0	0	52
C	1	1	6	8	23	7	5	1	0	52
D	0	0	0	1	3	7	14	16	11	52
Total	7	10	19	27	41	28	39	25	12	208

*I = strong dislike; †IX = excellent taste.
Data courtesy of Dr Graeme Newell, Hawkesbury Agricultural College.

The nature of the response is such that a model of the form (5.1) or (5.3) is most obviously appealing. We first try the logistic model with intercept parameters $\theta_1, \ldots, \theta_8$ and treatment effects $\beta_1, \ldots, \beta_4$. In this instance, (5.1) can be written in the form

$$\text{logit } \gamma_{ij} = \theta_j - \beta_i$$

for $j = 1, \ldots, 8$ and $i = 1, \ldots, 4$. As usual, only contrasts among the treatment effects β_i are estimable. We adopt the convention whereby $\hat{\beta}_4 = 0$. The resulting estimates, standard errors and correlation matrix of the $\hat{\beta}$s are given below.

Logistic treatment effects for cheese-tasting data

Additive	Estimate	SE	Correlations		
A	$\hat{\beta}_1 = -1.613$	0.378	1.0		
B	$\hat{\beta}_2 = -4.965$	0.474	0.525	1.0	
C	$\hat{\beta}_3 = -3.323$	0.425	0.574	0.659	1.0
D	$\hat{\beta}_4 = 0.0$	—	—	—	—

Positive values of β represent a tendency towards the higher-numbered categories relative to the chosen baseline — in this case, the probabilities for cheese D. Negative values indicate the reverse effect. The observed estimates quantify and confirm the ordering (D, A, C, B) from best to worst. The quoted standard errors are based on the assumption that $\sigma^2 = 1$, namely that no over-dispersion is present. The correlations, unlike the covariances, are unaffected by the choice or estimate of σ^2.

The deviance for these data is reduced from 168.8 on 24 de-grees of freedom under the model of zero additive effect ($\beta = 0$) to 20.31 on 21 degrees of freedom under the proportional-odds model. Because of the small numbers in the extreme cells the chi-squared approximation for the deviance is not very good here. Residual analysis is awkward partly for the same reason and partly because row sums are fixed. Using the crude standardization $(y_{ij} - \hat{y}_{ij})/[m_i\hat{\pi}_{ij}(1 - \hat{\pi}_{ij})]^{1/2}$, we find two cell residuals exceeding the value 2.0. The values are 2.23 and 2.30 corresponding to cells $(1, 4)$ and $(2, 6)$ with fitted values 3.16 and 2.47 respectively. How-ever, if residual calculations were based on the cumulative totals $z_{ij} = y_{i1} + \ldots + y_{ij}$, arguably a more appropriate procedure here,

the apparently large discrepancies would disappear. At the very least, residuals based on z_{ij} have the strong conceptual advantage that only $k - 1$ of them are defined for each multinomial observation. Correlation among the residuals is a problem regardless of definition but the problem seems more acute for residuals based on z_{ij}. However, these extreme residuals are hardly sufficiently large to refute the model which is, at best, an approximation to reality.

As a further check on the adequacy of the proportional-odds model, we fitted the generalized rational model (5.4) with the treatment factor as the model formula in both numerator and denominator. In other words

$$\text{logit}\,\gamma_{ij} = \frac{\theta_j - \beta_i}{\exp(\tau_i)}.$$

This gives a reduction in deviance of 3.3 on 3 degrees of freedom, so that there is no evidence that the variability of responses is affected by the cheese additive.

So far we have assumed $\sigma^2 = 1$ without justification. However, the estimate for σ^2 from (5.18) is 20.9/21 or almost exactly unity. Here $p = 11$ is the total number of parameters including the θs.

A more serious problem that we have so far ignored is that observations corresponding to different treatments are not independent. The same 52 panellists are involved in all four tests. This is likely to induce some positive correlation ρ between the ratings for the different treatments. Variances of contrasts would then be reduced by a factor $1 - \rho^2$ relative to independent measurements. Inferences based on supposing that $\rho = 0$ are therefore conservative. In other words the general qualitative and quantitative conclusions remain valid with the computed variances being regarded as approximate upper limits.

Finally we examine the effect on $\hat{\beta}$ of reducing the number of response categories. Various combinations are possible: here we combine categories $1, \ldots, 4$ and 7, 8, 9, thus reducing the original nine categories to four. This arrangement makes all cell counts positive. The new estimates for β are $(-1.34, -4.57, -3.07, 0)$ corresponding to an average reduction of about 0.7 standard errors compared with the previous analysis. Reduction of the number of categories does not always have this effect. Estimated variances are increased by an average of about 19%. Correlations are virtually unaffected.

The available evidence suggests that, when the data are sparse, the estimate of $\hat{\beta}_j$ may be too large in magnitude. Grouping of the tail categories has the effect of reducing this bias. A very small-scale simulation experiment based on 25 repetitions using the values of $\hat{\theta}$ and $\hat{\beta}$ obtained from the data in Table 5.1 and the same row totals indicates the following:

1. the bias in the estimates $\hat{\beta}_j$ is no more than 5%.
2. the deviance or likelihood-ratio goodness-of-fit statistic is approximately distributed as χ^2_{21}: at least the first two moments do not differ appreciably from those of this reference distribution.
3. the standard errors obtained from the diagonal elements of (5.11) are, if anything, a little too large — by about 10%.

The first claim is buttressed to some extent by the findings in section 7.5.3, where the nuisance parameters are eliminated by suitable conditioning. Because of the small scale of the simulation, the remaining conclusions are tentative. Nonetheless the conclusions are positive and they show that, even with data as sparse as those in Table 5.1 and where the number of parameters (11) is moderately large in comparison to the number of observations (32), the usual asymptotic results are quite reliable at least for the parameters of primary interest.

5.6.2 *Pneumoconiosis among coalminers*

The following example illustrates the use of a quantitative covariate in an ordinal regression model. For comparative purposes we apply both (5.1) and (5.10). Difficulties associated with residual plots are also illustrated.

The data, taken from Ashford (1959), concern the degree of pneumoconiosis in coalface workers as a function of exposure t measured in years. Severity of disease is measured radiologically and is, of necessity, qualitative. A four-category version of the ILO rating scale was used initially, but the two most severe categories were subsequently combined.

A preliminary plot of the transformed variables

$$\log\left(\frac{y_{i1} + \frac{1}{2}}{m_i - y_{i1} + \frac{1}{2}}\right) \quad \text{and} \quad \log\left(\frac{y_{i1} + y_{i2} + \frac{1}{2}}{m_i - y_{i1} - y_{i2} + \frac{1}{2}}\right) \quad (5.19)$$

Table 5.2 *Period of exposure and prevalence of pneumoconiosis amongst a group of coalminers*

	Number of men		
Period spent (yr)	Category I: normal	Category II	Category III: severe pneumoconiosis
5.8	98	0	0
15.0	51	2	1
21.5	34	6	3
27.5	35	5	8
33.5	32	10	9
39.5	23	7	8
46.0	12	6	10
51.5	4	2	5

against t_i reveals approximately parallel but non-linear relationships. Further investigation shows that the transformed variables (5.19) are approximately linear in $\log t_i$. We are thus led to consider the model

$$\log[\gamma_{ij}/(1 - \gamma_{ij})] = \theta_j - \beta \log t_i, \qquad j = 1, 2; \ i = 1, \ldots, 8. \quad (5.20)$$

We might expect that the non-linearity of (5.20) in t could have been detected by an appropriate analysis of the residuals after fitting the model linear in t. This is indeed so but some care is required. When the 24 cell residuals, appropriately standardized, are plotted against t_i, no strong curvilinear pattern is discernible. On the other hand, a plot against t_i of the cumulative residuals, $y_{i1} - \hat{y}_{i1}$ and $y_{i1} + y_{i2} - \hat{y}_{i1} - \hat{y}_{i2}$, appropriately standardized, clearly reveals the non-linearity. When $k = 3$ this is equivalent to ignoring the residuals associated with category 2 and changing the sign of the category-3 residuals. The two plots are displayed in Figs. 5.5a and 5.5b respectively. The simplified standardization used here takes no account of the errors involved in using estimated values of the parameters.

The analysis using (5.20) gives a value of $\hat{\beta}$ of 2.60 with standard error 0.38, while the values of $\hat{\theta}_1$ and $\hat{\theta}_2$ are 9.68 and 10.58 respectively. No pattern is discernible among the residuals and the fit is good. The conclusions, therefore, are that for a miner with, say, five years of exposure, the odds of having pneumoconiosis are
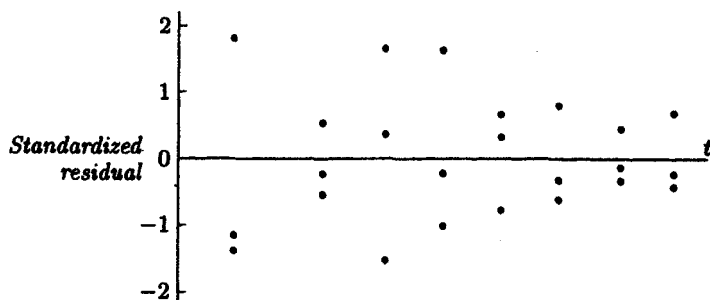
Fig. 5.5a. *Plot of cell residuals against t for the pneumoconiosis data.*
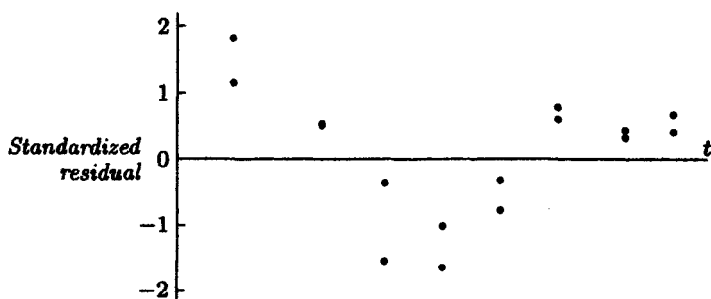

Fig. 5.5b. *Plot of cumulative residuals against t for pneumoconiosis data.*

one in $\exp(9.68 - 2.60\log 5)$, i.e. one in 243. Doubling the exposure increases the risk by a factor of $2^{2.60} = 6.0$, so that after 10 years the risk rises to just under one in 40, and after 20 years to just over one in seven. For severe pneumoconiosis the five-year risk is one in $\exp(10.58 - 2.60\log 5)$, i.e. about one in 600. This risk increases at the same rate as before so that after 10 years exposure the estimated risk is one in 100 and after 20 years is one in 17.

Ashford's analysis of these data proceeds along the same lines as that given here except that he uses the probit function in place of the logit. His conclusions give similar fitted values but his parameter estimates are different, partly because of his use of the probit function and partly because of numerical inaccuracies.

We now proceed to illustrate the use of the alternative model (5.10) in the context of the pneumoconiosis data. Imagine the response categories arranged in the hierarchical format illustrated in Fig. 5.3. The response at stage 1 is the dichotomy between

disease and no disease whereas the response at stage 2 is the dichotomy between the mild form of the disease and the severe form. These responses are very much alike but technically distinct. However, for comparative purposes, we assume here that the effect of continued exposure is comparable for the two responses so that (5.10) can be used. We take model (5.10) in the form

$$\log\left(\frac{\pi_{ij}}{1 - \gamma_{ij}}\right) = \alpha_j - \beta x_i, \qquad (5.21)$$

reversing the sign of the coefficient in order to make the results at least qualitatively comparable to those obtained from (5.20).

The odds in favour of category 1 are $\exp(\alpha_1 - \beta x_i)$ so that the odds or risk of having the disease in the first place is $\exp(-\alpha_1 + \beta x_i)$. Here x_i is a general measure of exposure — in this case t_i or $\log t_i$. Thus the risk of disease increases by the factor e^β per unit increase in x. Among those who have the disease, the risk, or odds of having severe symptoms, is $\exp(-\alpha_2 + \beta x_i)$, so that again the risk increases by the factor e^β per unit increase in x. There is clearly the possibility that a different β might be involved in the second expression.

Because of the special structure of the model (5.10) each trinomial observation can be broken into two binomial components. The first component specifies the number of diseased individuals as a proportion of the total number at risk, while the second component gives the number of severely diseased as a proportion of those with the disease. Thus

$$y_{11}/m_{11} = 0/98, \qquad y_{12}/m_{12} = 0/0,$$
$$y_{21}/m_{21} = 3/54, \qquad y_{22}/m_{22} = 1/3,$$
$$y_{31}/m_{31} = 9/43, \qquad y_{32}/m_{32} = 3/9,$$

and so on with $m_{ij} = y_{ij} + y_{i\,j+1} + \ldots + y_{ik}$ giving the total in categories j through k inclusive.

These binomial observations y_{ij}/m_{ij} can be regarded as independent observations with probabilities π_{ij} satisfying

$$\log[\pi_{ij}/(1 - \pi_{ij})] = -\alpha_j + \beta x_i, \qquad j = 1, 2; \; i = 1, \ldots, 8, \quad (5.22)$$

If the relationships are not parallel it may be necessary to write β_j instead of β in (5.22). The binomial log likelihood for the logistic

model (5.22) is identical to the multinomial log likelihood for the
model (5.21).

For these data, the covariate $\log t_i$ is strongly preferred to t_i;
there is inconclusive evidence on whether $\beta_1 \neq \beta_2$. As measured
by the deviance or likelihood-ratio statistic, the fit of (5.22) is
comparable to that of (5.20). The goodness-of-fit statistics are
5.1 on 13 d.f. for (5.20) and 7.6 on 12 d.f. for (5.22). One degree
of freedom is lost because y_{12} is degenerate or non-random when
(5.22) is used. For model (5.22) the residuals give some indication
of a faint pattern; for this reason the former model (5.20) might be
preferred. In any case the estimate of β is 2.32 with approximate
standard error 0.33, these values being similar to those obtained
earlier despite the slight difference of interpretation. Thus we are
led to the estimate of $2^{2.32} = 5.0$ as the increase in risk associated
with doubling the exposure time.

To summarize we can say that (5.1) and (5.10), or equivalently
(5.20) and (5.22), are different ways of describing the risk associated
with increasing exposure. The conclusions from either analysis
support the claim that doubling the exposure increases the risk
by an estimated factor of between 5 and 6. Approximate 95%
confidence limits for this factor are (3.2, 10.2). It would be of
interest to know (i) whether the risk would continue to increase
if exposure were to cease, and (ii) whether the risk would increase
more slowly if dust levels were reduced. The data given here do not
allow us to investigate these questions; indeed as the data stand,
such effects would be likely to be confounded with age.

5.7 Bibliographic notes

Many of the methods and models discussed in Chapter 4 for bi-
nary data carry over to polytomous responses with only minor
alterations. Consequently, most of the references listed in section
4.7 are also relevant here although there is enormous variation in
emphasis and coverage. Agresti (1984) concentrates almost entirely
on methods for ordinal response variables, including measures of as-
sociation, which are not covered here. Haberman (1978, 1979) em-
phasizes methods for fitting a variety of log-linear models, mainly
with social science applications in mind. Fienberg (1980) p.110 con-
siders a variety of link functions, all of which are variations on the

logit. Both Haberman and Fienberg devote considerable attention to algorithmic details and the computation of maximum-likelihood estimates for two and three-way tables. Details of the iterative proportional fitting algorithm (Deming and Stephan, 1940; Darroch and Ratcliff, 1972) for log-linear models can be found in the book by Bishop *et al.* (1975). This algorithm forms the core of several log-linear computer packages, but is not sufficiently general to cover the range of models considered here.

Aickin (1983) makes a distinction between nominal and nested response scales similar to the distinction made in section 5.2, but does not consider proportional-odds or proportional-hazards models for ordinal responses. For further discussion of measurement scales see Stevens (1951, 1958, 1968).

The idea of representing ordered categories as contiguous intervals on a continuous scale goes back at least to Pearson (1901), who investigated coat-colour inheritance in thoroughbred horses. An extension of this idea to two variables led to the development of the tetrachoric and polychoric correlation coefficients and to the quarrel with Yule (Yule, 1912; Pearson and Heron, 1913).

The proportional-odds model described in section 5.2 was previously used by Hewlett and Plackett (1956), Snell (1964), Walker and Duncan (1967), Clayton (1974), Simon (1974), Bock (1975) and others. Ashford (1959), Gurland *et al.* (1960) and Finney (1971) used the probit link in place of the logistic.

Williams and Grizzle (1972) discuss a number of methods including the proportional-odds model as well as scoring methods in the log-linear context. See also Haberman (1974a,b). McCullagh (1980) compares the use of scores in log-linear models with direct application of the proportional-odds model. He concludes that the proportional-odds and related models based on transforming the cumulative proportions are to be preferred to scoring methods because they are invariant under the grouping of adjacent response categories.

Graubard and Korn (1987) discuss the effect of the choice of scores in testing for independence in two-way tables.

Goodhardt, Ehrenberg and Chatfield (1984) use the Dirichletmultinomial model to accommodate over-dispersion in brand-choice data. This is a natural extension of the beta-binomial model discussed in Chapter 4. For further discussion of specific forms of over-dispersion in this context, see Engel (1987).

Existence and uniqueness of maximum-likelihood estimates for a large subset of the models discussed here has been investigated by Pratt (1981) and by Burridge (1982).

Numerical methods for dealing with composite link functions such as (5.1) are discussed by Thompson and Baker (1981).

5.8   Further results and exercises 5

5.1   Show that

$$\sum_{j=0}^{m} \binom{j+k-1}{k-1} = \binom{m+k}{k}.$$

Hence deduce that the number of integer-valued sample points $\mathbf{y}$ satisfying $0 \le y_j \le m$ and $\sum y_j = m$ is $\binom{m+k-1}{k-1}$. [Hint: consider the series expansion of $(1-x)^{-1}(1-x)^{-m}$.]

5.2   Suppose that $Y \sim M(1, \boldsymbol{\pi})$ and $Z = LY$ is the vector of cumulative totals of Y. Show that

$$E(Z_r Z_s Z_t \ldots) = \gamma_r \quad \text{for} \quad r \le s \le t \le \ldots .$$

Hence show that

$$\text{cov}(Z_r, Z_s) = \gamma_r(1 - \gamma_s) \quad \text{for} \quad r \le s.$$

Derive the third- and fourth-order cumulants of Z.

5.3   Show that the following expressions are equivalent:

$$\sum \gamma_j(1 - \gamma_j)(\pi_j + \pi_{j+1})$$
$$\sum \pi_j(1 - \gamma_j - \gamma_{j-1})^2$$
$$\sum \gamma_j \gamma_{j+1} \pi_{j+1}$$
$$\sum (1 - \gamma_j)(1 - \gamma_{j-1})\pi_j$$
$$\tfrac{1}{3}\{1 - \sum \pi_j^3\}.$$

All sums run from 1 to k with the convention that $\gamma_k = 1$ and

$$\pi_0 = \gamma_0 = \pi_{k+1} = \gamma_{k+1} = 0.$$

Find the minimum and maximum values for fixed $k \ge 2$.

5.4 By considering the case in which the differences $\theta_i - \theta_j$ are known, show that the asymptotic covariance matrix for $(\hat{\theta}_1, \hat{\beta})$ in (5.1) is given by $(\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1}$. The model matrix $\mathbf{X}$ is $n \times (p+1)$, with the constant vector in column 1, and $\mathbf{W}$ is diagonal with components

$$w_i = \tfrac{1}{3} m_i \{ 1 - \textstyle\sum_j \pi_{ij}^3 \}.$$

Deduce that this approximation gives a lower bound for $\text{cov}(\hat{\beta})$ when no assumptions are made about $\boldsymbol{\theta}$. Show that the approximation is accurate if $k = 2$ or if the log odds-ratios are small.

Hence show that in the two-sample problem following (5.2), the approximate variance of $\hat{\Delta}$ is $1/w_1 + 1/w_2$ (Clayton, 1974). This approximation appears to be satisfactory for $|\Delta| \leq 1$, which is usually the most interesting range in applications.

5.5 The data in Table 5.3, taken from Lowe *et al.* 1971, give the frequencies of anencephalus, spina bifida and other malformations of the central nervous system among live births in various South Wales communities. The principal object of the study was to investigate the effect, if any, of water hardness on the incidence of such diseases. What type of response scale is involved here? Analyse the data paying particular attention to the following points.

1. possible effects of water hardness on the incidence of CNS disorders.
2. differences between manual and non-manual workers in the incidence of CNS disorders.
3. anomalous geographical effects.
4. any systematic differences in the distribution of types of CNS disorders.

Give a brief non-technical summary of your conclusions.

5.6 Show that the complementary log-log model (5.3) is equivalent to the continuation-ratio or nested response model

$$g\{\pi_j(\mathbf{x})/(1 - \gamma_{j-1}(\mathbf{x}))\} = \alpha_j - \boldsymbol{\beta}^T \mathbf{x}$$

if $g(\cdot)$ is the complementary log-log function. Express α_j in terms of the 'cut-points' $\theta_1, \ldots, \theta_{k-1}$ appearing in (5.3). [Läärä and Matthews, 1985].

Check this claim numerically by replacing the logistic link in (5.20) and (5.21) with the complementary log-log link. Why are

Table 5.3 *Frequencies of central nervous system malformations in live births in eight South Wales communities (1964–66).*

| | Non-manual | | | | Manual | | | | Water |
| | No CNS malformation | CNS malformation | | | No CNS malformation | CNS malformation | | | hardness |
Area		An.†	Sp.‡	Other		An.	Sp.	Other	
Cardiff	4091	5	9	5	9424	31	33	14	110
Newport	1515	1	7	0	4610	3	15	6	100
Swansea	2394	9	5	0	5526	19	30	4	95
Glamorgan E.	3163	9	14	3	13217	55	71	19	42
Glamorgan W.	1979	5	10	1	8195	30	44	10	39
Glamorgan C.	4838	11	12	2	7803	25	28	12	161
Monmouth V.	2362	6	8	4	9962	36	37	13	83
Monmouth other	1604	3	6	0	3172	8	13	3	122

†Anencephalus, the most serious malformation of the central nervous system.
‡Spina bifida without anencephalus.
Where both malformations are present, anencephalus is recorded.
Data taken from Lowe et al. (1971) are reconstructed from totals and rates.
Water hardness, which is measured in parts per million, varies to some extent within communities. For details of the within-community variability, see Appendix A of Lowe et al. (1971).

the fitted values for category II different for the two models? Show also that the corresponding logit models (5.1) and (5.10) are not equivalent.

5.7 Consider the multinomial response model (5.7) with scores $s = (1, 0, \ldots, 0)$. Show that, with these scores, the log-linear model is equivalent to the nested response model

$$\text{logit } \pi_1(\mathbf{x}_i) = \eta_1 + \boldsymbol{\beta}^T \mathbf{x}_i$$

$$\text{logit}\left(\frac{\pi_j(\mathbf{x}_i)}{1 - \gamma_{j-1}(\mathbf{x}_i)}\right) = \eta_j \qquad j \geq 2.$$

5.8 Let Y_{ij} be the observations in a two-way table with n independent rows such that $\mathbf{Y}_i \sim M(m_i, \boldsymbol{\pi})$ on k categories. Consider the statistic

$$T = \sum_{ij} r_i s_j Y_{ij},$$

with given scores r_i, s_j, as a possible statistic for testing the hypothesis of independence or no row effect. Show that, under the hypothesis

$$E(T) = m_. \mu_r \mu_s$$

$$\text{var}(T) = \sum_i m_i r_i^2 \sigma_s^2$$

where

$$\mu_r = \sum m_i r_i / m_.,$$

$$\mu_s = \sum \pi_j s_j, \qquad \tilde{\mu}_s = \sum y_{.j} s_j / m_.,$$

$$\sigma_s^2 = \sum \pi_j (s_j - \mu_s)^2, \qquad \tilde{\sigma}_s^2 = \sum y_{.j} (s_j - \tilde{\mu}_s)^2 / m_..$$

Explain why, for fixed n, k, the 'standardized statistic'

$$\frac{T - m_. \mu_r \tilde{\mu}_s}{\sqrt{\tilde{\sigma}_s^2 \sum m_i r_i^2}}$$

is approximately Normally distributed but not with unit variance in the limit as $m_i \to \infty$. Show that the 'correct' standardization is

$$\frac{T - m_. \mu_r \tilde{\mu}_s}{\sigma_r \tilde{\sigma}_s \sqrt{m_.}}$$

where $\sigma_r^2 = \sum m_i (r_i - \mu_r)^2 / m_.$. [Yates, 1948].

5.9 Consider the proportional-odds model

$$\text{logit}\,\gamma_j(x_i) = \theta_j - \beta x_i$$

with x and β both scalars. Denote by $\hat{\theta}_j$, $\hat{\pi}_j$ the fitted parameters and probabilities under the hypothesis that $\beta = 0$. Show that the derivative of the log likelihood with respect to β at $\beta = 0$, $\theta_j = \hat{\theta}_j$, is given by

$$T = \sum R_{ij} x_i s_j$$

where $R_{ij} = Y_{ij} - m_i \hat{\pi}_j$ is the residual under independence and $s_j = \hat{\gamma}_j + \hat{\gamma}_{j-1}$. [Tests based on the log-likelihood derivative are sometimes called 'score tests'.]

5.10 Use the results of Exercises 5.2 and 5.8 to find the approximate mean and variance of T in the previous exercise. Hence construct a simple test of the hypothesis that $\beta = 0$. Show that in the two-sample problem T is equivalent to Wilcoxon's statistic.

5.11 Repeat the calculations of the previous two exercises replacing the proportional-odds model with the complementary log-log model. Which non-parametric test does T now correspond to?

5.12 Show that the score test based on the log-linear model (5.7) is identical to the score test based on the linear logistic model (5.1) provided that ridit scores are used for the response categories in (5.7). [Ridit scores (Bross, 1958) are proportional to the average category rank.]

5.13 Table 5.4, taken from Yates (1948), gives teachers' ratings for homework, together with an assessment of homework facilities, for 1019 schoolchildren. In both cases, A denotes the highest or best rating and subsequent letters denote lower grades.

1. Which variable is the response?
2. Fit the model of independence and look for patterns among the residuals. Compute X^2 and D and show that these are approximately equal to their degrees of freedom.
3. Using integer-valued scores for both ratings, compute the statistic T as described in Exercise 5.7. Show that the standardized statistic is 1.527, corresponding to an approximate one-sided p-value of 6.3%.
4. Fit the linear complementary log-log model (5.3) using a quantitative integer-valued covariate for homework conditions.

Show that $\hat{\beta} = 0.0476$, s.e.$(\hat{\beta}) = 0.027$, corresponding to a one-sided p-value of 3.9%. Comment on the direction and magnitude of the estimated effect.

5. Fit the log-linear model (5.7) using integer-valued scores for both rows and columns. Show that $\hat{\beta}/\text{s.e.}(\hat{\beta}) = 1.525$, and that the reduction in deviance is 2.33 on one degree of freedom. Why are these values so remarkably similar to Yates's statistic in part 3 above?

Table 5.4 *Relation between conditions under which homework was carried out, and teacher's assessment of homework quality.*

Homework conditions	Teacher's rating			Total
	A	B	C	
A	141	131	36	308
B	67	66	14	147
C	114	143	38	295
D	79	72	28	179
E	39	35	16	90

5.14 In Example 2 of section 5.2.5, what modifications to the design of the study would you make if the available test animals comprised 60 milch cows and 20 heifers? What modifications would be required in the analysis of the data from this experiment?

5.15 *Logistic discrimination:* Suppose that a population of individuals is partitioned into k sub-populations or groups, $G_1, \ldots, G_k$, say, with relative frequencies $\pi_1, \ldots, \pi_k$. It may be helpful to think of the groups as species or distinct populations of the same genus. Multivariate measurements Z made on individuals have the following distributions for the k groups:

$$G_j: \quad Z \sim N_p(\boldsymbol{\mu}_j, \Sigma), \qquad j = 1, \ldots, k.$$

Let $\mathbf{z}^*$ be an observation made on an individual drawn at random from the combined population. The prior odds that the individual belongs to G_j are $\pi_j/(1 - \pi_j)$. Show that the posterior odds for G_j given $\mathbf{z}^*$ are

$$\text{odds}(Y = j \mid \mathbf{z}^*) = \frac{\pi_j}{1 - \pi_j} \times \exp(\alpha_j + \boldsymbol{\beta}_j^T \mathbf{z}^*).$$

Find expressions for α_j and β_j in terms of μ_j and Σ.

What simplifications can be made if the k Normal means μ_j lie on a straight line in R^p?

Comment briefly on the differences between maximum likelihood estimation of α_j and β_j via the Normal-theory likelihood and estimation via logistic regression.

5.16 Show that the quadratic form

$$\sum_1^{k-1} \frac{(Z_j - m\gamma_j)^2}{m}\left(\frac{1}{\pi_j} + \frac{1}{\pi_{j+1}}\right) - 2\sum_{j=1}^{k-2} \frac{(Z_j - m\gamma_j)(Z_{j+1} - m\gamma_{j+1})}{m\pi_{j+1}},$$

given in section 5.3.4, is identical to Pearson's X^2 statistic. In the above expression Z_j are the components of the cumulative multinomial vector and $E(Z_j) = m\gamma_j$.

Table 5.5 *Effect of mother's age and smoking habits on gestation period and perinatal mortality*

Gestation period (days)	Mother's age	Cigarettes smoked	Perinatal mortality mortality/total births
197–260	< 30	≤ 5	50/365
		> 5	9/49
	30+	≤ 5	41/188
		> 5	4/15
261+	< 30	≤ 5	24/4036
		> 5	6/465
	30+	≤ 5	14/1508
		> 5	1/125

Source: Wermuth (1976).

5.17 Table 5.5, taken from Wermuth (1976), gives the gestation period and perinatal mortality rates for a group of German women, many of whom were pregnant for the first time or had complications with previous pregnancies. Which are the response variables? Examine first how the gestation period or probability of premature birth is related to mother's age and smoking habits. Second, examine how the perinatal mortality rate is related to gestation period, mother's age and smoking habits. Summarize your conclusions in non-technical language.

Table 5.6 *Frequency table for the 54 numbers used in the Illinois lottery for the 12-month period ending 12 Nov 1988. Six numbers are drawn each week.*

Tens	Units										Total
	0	1	2	3	4	5	6	7	8	9	
0+		9	8	5	6	9	4	7	7	6	61
10+	5	5	6	3	5	4	5	5	3	2	43
20+	6	3	7	10	9	10	9	7	1	7	69
30+	11	6	2	9	10	4	8	4	9	6	69
40+	5	10	3	7	8	3	4	4	1	4	49
50+	3	5	4	4	5						21
Total	30	38	30	38	43	30	30	27	21	25	312

Source: Chicago Tribune, 14 Nov 1988.

5.18 Pick-6 is the weekly Illinois lottery in which the winning ticket comprises six unordered numbers in the range 1–54. The winning numbers are chosen by a physical randomizing device in which 54 numbered ping-pong balls are mixed by a draught of air in a closed transparent container. Six balls carrying the winning numbers are permitted to escape, one at a time, through a hole in the top of the apparatus. The frequency of occurrence of each the 54 numbers in a 12-month period is shown in Table 5.6.

By fitting a log-linear model or otherwise, test the following hypotheses, all of which refer to the uniformity of the numbers generated by the randomizing device.

1. that the variation of the frequencies in Table 5.6 is consistent with the hypothesis of uniformity.
2. that the variation of the column totals in Table 5.6 is consistent with the hypothesis of uniformity.
3. that the variation of the row totals in Table 5.6 is consistent with the hypothesis of uniformity.
4. that the frequency of occurrence of the numbers 45–54 is the same as that of the remaining numbers.

You are now given the additional information that for the first 24 weeks only the numbers 1–44 were used: thereafter all 54 numbers were used. Test hypotheses (1.) and (2.) above making due allowance for this change of regime after 24 weeks.

5.19 Repeat the calculations of the previous exercise making due allowance for the fact that the six balls are chosen each week without replacement from the pool of 44 or 54 balls. Show that for

large m, and provided that the number of balls remains constant from week to week,

$$\frac{(k-1)X^2}{k-n} \sim \chi^2_{k-1}$$

where in this case, $k = 44$ or 54, $n = 6$, $m = 52$ and X^2 is Pearson's statistic computed in the usual way as if the counts were multinomial variables on k categories. This finite population correction, which is not asymptotically negligible, should also be used to correct deviance statistics derived from Poisson models. Alternatively the counts may be taken as binomial variables with index 52 for numbers 1–44 and 28 for numbers 45–54. It is then necessary to include an offset in all models.

5.20 Under the conditions described in the previous exercise show that

$$E(X^2) = k - n,$$

$$\text{var}(X^2) = 2\frac{(k-n)^2}{k-1}\frac{m-1}{m}.$$

Check that these calculations are correct for $n = 1$ and $n = k - 1$.